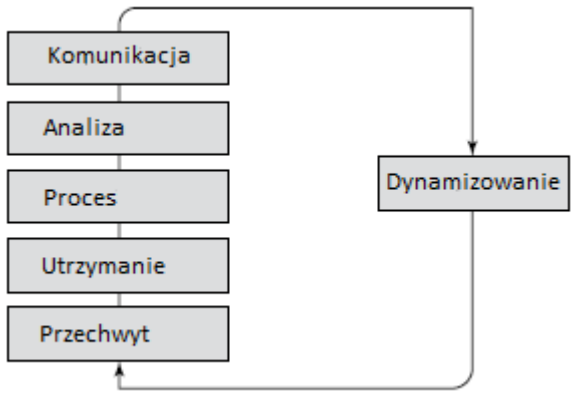
Ramowanie strategii analizy danych

Zamierzam uporządkować podstawy tego, czym jest nauka o danych, ale muszę cię ostrzec, że nauka o danych to termin, który wymyka się jednej pełnej definicji - co oczywiście sprawia, że nauka o danych jest trudna do zrozumienia i aplikować w organizacji. Wiele artykułów i publikacji używa tego terminu dość swobodnie, zakładając, że jest on powszechnie rozumiany. Jednak nauka o danych – w tym jej metody, cele i zastosowania – ewoluuje wraz z czasem i technologią i obecnie znacznie różni się od tego, co mogło być 25 lat temu. Mimo wszystko jestem gotów przedstawić wstępną definicję: Data science to badanie, skąd pochodzą dane, co reprezentują i jak można je przekształcić w wartościowy zasób w tworzeniu strategii biznesowych. Można powiedzieć, że nauka o danych jest dziedziną multidyscyplinarną, która wykorzystuje naukowe metody, procesy, algorytmy i systemy do wydobywania spostrzeżeń z danych w różnych formach, zarówno ustrukturyzowanych, jak i nieustrukturyzowanych. Wydobywanie dużych ilości ustrukturyzowanych i nieustrukturyzowanych danych w celu zidentyfikowania wzorców i odchyleń, które mogą pomóc organizacji w ograniczeniu kosztów, zwiększeniu wydajności, rozpoznaniu nowych możliwości rynkowych i zwiększeniu przewagi konkurencyjnej organizacji. Nauka o danych to pojęcie, które można wykorzystać do ujednolicenia statystyki, analityki, uczenia maszynowego oraz powiązanych z nimi metod i technik w celu zrozumienia i analizy rzeczywistych zjawisk za pomocą danych. Wykorzystuje techniki i teorie zaczerpnięte z wielu dziedzin w kontekście matematyki, statystyki, informatyki i informatyki. Za tym typem definicji kryje się jednak definicja tego, w jaki sposób podchodzi się i jak wykonuje się naukę o danych. A ponieważ ambicją tej części jest sformułowanie strategii analizy danych, muszę najpierw bardziej właściwie określić ten multidyscyplinarny obszar nauki o danych i jego cykl życia.

Ustanawianie narracji nauki o danych

Nie zaszkodzi mieć obraz podczas wyjaśniania skomplikowanego procesu, więc spójrz na rysunek, na którym możesz zobaczyć główne kroki lub fazy cyklu życia nauki o danych.



Pamiętaj jednak, że model przedstawiony na rysunku zakłada, że jako punkt wyjścia zidentyfikowałeś już problem biznesowy wysokiego poziomu lub okazję biznesową. Ta wczesna ambicja jest zwykle wyprowadzana z perspektywy biznesowej, ale musi zostać przeanalizowana i szczegółowo opisana wraz z zespołem ds. analityki danych. Ten dialog jest niezbędny, jeśli chodzi o zrozumienie, jakie dane są dostępne i co można z nimi zrobić, aby można było skupić się na dalszej pracy. Nie jest dobrym pomysłem, aby po prostu zacząć przechwytywać wszystkie dane, które wyglądają wystarczająco interesująco, aby je przeanalizować. Dlatego pierwszym etapem cyklu życia nauki o danych, przechwytywanie, jest sformułowanie potrzebnych danych poprzez przełożenie potrzeby biznesowej na konkretny i dobrze zdefiniowany problem lub możliwość biznesową. Początkowy problem biznesowy i/lub szansa nie są statyczne i będą się zmieniać w miarę dojrzewania wiedzy opartej na

danych. Zachowanie elastyczności w zakresie tego, które dane są przechwytywane, a także jaki problem i/lub szansa jest najważniejsza w danym momencie, jest zatem niezbędne do osiągnięcia celów biznesowych. Model pokazany na rysunku ma na celu przedstawienie widoku różnych etapów cyklu życia nauki o danych, od uchwycenia potrzeb biznesowych i danych, przez przygotowanie, badanie i analizę danych, do uzyskania wglądu i działania na ich podstawie. Wyjście każdego pełnego cyklu generuje nowe dane, które stanowią wynik poprzedniego cyklu. Obejmuje to nie tylko nowe dane lub wyniki, które można wykorzystać do optymalizacji modelu, ale mogą również generować nowe potrzeby biznesowe, problemy, a nawet nowe zrozumienie tego, jaki powinien być priorytet biznesowy. Te etapy cyklu życia nauki o danych można również postrzegać jako nie tylko kroki opisujące zakres nauki o danych, ale także warstwy w architekturze. Więcej o tym później; zacznę od wyjaśnienia różnych etapów.

Przechwyt

Istnieją dwie różne części pierwszego etapu cyklu życia, ponieważ przechwytywanie odnosi się zarówno do przechwytywania potrzeb biznesowych, jak i pozyskiwania i pozyskiwania danych. Ten etap ma kluczowe znaczenie dla reszty procesu. Zacznę od wyjaśnienia, co to znaczy uchwycić potrzebę biznesową. Punktem wyjścia do uszczegółowienia potrzeby biznesowej jest wniosek biznesowy wysokiego poziomu lub problem biznesowy wyrażony przez kierownictwo lub podobne podmioty i powinien obejmować takie zadania, jak:

* Przekładanie niejednoznacznych próśb biznesowych na konkretne, dobrze zdefiniowane problemy lub możliwości

* Zagłębianie się w kontekst próśb, aby lepiej zrozumieć, jak może wyglądać potencjalne rozwiązanie, w tym jakie dane będą potrzebne

* Nakreślenie (jeśli to możliwe) strategicznych priorytetów biznesowych ustalonych przez firmę, które mogą mieć wpływ na pracę w zakresie analizy danych

Teraz, gdy wyjaśniłem, jak ważne jest przechwytywanie i zrozumienie żądań biznesowych oraz wstępnego określania potrzebnych danych, chcę przejść do opisywania aspektów samego procesu przechwytywania danych. Jest to główny interfejs do źródła danych, do którego należy sięgnąć i obejmuje obszary takie jak

* Zarządzanie własnością danych i zabezpieczenie praw do gromadzenia i wykorzystywania danych

* Obsługa danych osobowych i ochrona prywatności danych za pomocą różnych technik anonimizacji

* Korzystanie ze sprzętu i oprogramowania do pozyskiwania danych poprzez przesyłanie wsadowe lub przesyłanie strumieniowe danych w czasie rzeczywistym

* Określanie, jak często dane będą musiały być pozyskiwane, ponieważ częstotliwość zwykle różni się w zależności od typu danych i kategorii

* Wymaganie, aby wstępne przetwarzanie danych miało miejsce w punkcie ich zbierania lub nawet przed ich zebraniem (na przykład na krawędzi urządzenia IoT). Obejmuje to podstawowe przetwarzanie, takie jak czyszczenie i agregowanie danych, ale może również obejmować bardziej zaawansowane czynności, takie jak anonimizacja danych w celu usunięcia poufnych informacji. (Anonimizacja oznacza usuwanie poufnych informacji, takich jak imię i nazwisko osoby, numer telefonu, adres itd.) z zestawu danych.

W większości przypadków dane muszą zostać zanonimizowane przed przesłaniem ich ze źródła danych. Zwykle istnieje również procedura sprawdzania kompletności zbiorów danych. Jeśli dane nie są kompletne, może być konieczne kilkakrotne zbieranie, aby osiągnąć pożądany zakres danych. Przeprowadzenie tego typu walidacji na wczesnym etapie ma pozytywny wpływ zarówno na szybkość procesu, jak i na koszty.

* Zarządzanie procesem przesyłania danych do potrzebnego punktu przechowywania (lokalnego i/lub globalnego). W ramach transferu danych może być konieczne przekształcenie danych - na przykład agregacja, aby były mniejsze. Być może będziesz musiał to zrobić, jeśli masz do czynienia z ograniczeniami przepustowości łączy transferowych, z których korzystasz.

Utrzymanie

Działania związane z utrzymaniem danych obejmują zarówno przechowywanie, jak i utrzymywanie danych. Należy pamiętać, że dane są zwykle przetwarzane na wielu różnych etapach w całym cyklu życia. Konieczność ochrony integralności danych podczas cyklu życia elementu danych jest szczególnie ważna podczas czynności przetwarzania danych. Podczas ręcznego przetwarzania danych łatwo jest przypadkowo uszkodzić zbiór danych z powodu błędu ludzkiego, co powoduje, że zbiór danych jest bezużyteczny do analizy w następnym kroku. Najlepszym sposobem ochrony integralności danych jest zautomatyzowanie jak największej liczby kroków działań związanych z zarządzaniem danymi, prowadzących do punktu analizy danych. Utrzymanie zaufania biznesowego do bazy danych ma kluczowe znaczenie dla zaufania użytkowników biznesowych i korzystania z uzyskanych informacji. Jeśli chodzi o przechowywanie danych, dwa ważne aspekty to:

* Przechowywanie danych: pomyśl o tym jako o wszystkim, co jest związane z tym, co dzieje się w jeziorze danych. Działania związane z przechowywaniem danych obejmują zarządzanie różnymi okresami przechowywania różnych typów danych, a także prawidłowe katalogowanie danych w celu zapewnienia łatwego dostępu do danych i ich wykorzystania.

* Przygotowanie danych: w kontekście utrzymywania danych przygotowanie danych obejmuje podstawowe zadania przetwarzania, takie jak czyszczenie danych drugiego poziomu, etapowanie danych i agregacja danych, z których wszystkie zwykle obejmują stosowanie filtra bezpośrednio podczas przechowywania danych. Nie chcesz umieszczać danych o niskiej jakości w swoim jeziorze danych.

Okresy przechowywania danych mogą być różne dla tego samego typu danych, w zależności od poziomu ich agregacji. Na przykład, surowe dane mogą być interesujące do zaoszczędzenia tylko przez krótki czas, ponieważ zwykle mają bardzo dużą objętość, a zatem są kosztowne w przechowywaniu. Z drugiej strony dane zagregowane są często mniejsze, tańsze i łatwiejsze do przechowywania, a zatem mogą być zapisywane przez dłuższy czas, w zależności od docelowych przypadków użycia.

Proces

Przetwarzanie danych jest główną warstwą przetwarzania danych skoncentrowaną na przygotowaniu danych do analizy i odnosi się do stosowania bardziej zaawansowanych metodologii inżynierii danych, takich jak:

* Klasyfikacja danych: Odnosi się to do procesu organizowania danych w kategorie w celu jeszcze bardziej efektywnego i wydajnego wykorzystania, w tym czynności takich jak etykietowanie i tagowanie danych. Dobrze zaplanowany system klasyfikacji danych ułatwia znajdowanie i pobieranie istotnych danych. Może to mieć również szczególne znaczenie w takich obszarach, jak prawo i zgodność.

* Modelowanie danych: pomaga w wizualnej reprezentacji danych i egzekwuje ustalone reguły biznesowe dotyczące danych. Można również zbudować modele danych w celu wymuszenia zasad dotyczących spójnego korelowania różnych typów danych. Modele danych zapewniają również spójność konwencji nazewnictwa, wartości domyślnych, semantyki i procedur bezpieczeństwa, zapewniając w ten sposób jakość danych.

* Podsumowanie danych: Twoim celem jest wykorzystanie różnych sposobów podsumowywania danych, takich jak użycie różnych technik grupowania.

* Eksploracja danych: jest to proces analizy dużych zbiorów danych w celu zidentyfikowania wzorców lub odchyleń, a także ustanowienia relacji w celu umożliwienia rozwiązywania problemów poprzez analizę danych w dalszej części drogi. Eksploracja danych to rodzaj analizy danych, skoncentrowanej na lepszym zrozumieniu danych, określanej również jako umiejętność korzystania z danych. Budowanie umiejętności korzystania z danych w zespołach zajmujących się analizą danych jest kluczowym składnikiem sukcesu w dziedzinie analizy danych. Przy niskiej znajomości danych i bez prawdziwego zrozumienia danych, które przygotowujesz, analizujesz i uzyskujesz spostrzeżenia, istnieje wysokie ryzyko niepowodzenia, jeśli chodzi o inwestycje w naukę danych

Analiza

Analiza danych to etap, na którym dane ożywają i w końcu możesz uzyskać wgląd w zastosowanie różnych technik analitycznych. Wglądy mogą być skoncentrowane na zrozumieniu i wyjaśnieniu tego, co się wydarzyło, co oznacza, że analiza ma charakter opisowy i bardziej reaktywny. Tak jest również w przypadku analizy w czasie rzeczywistym: nadal jest reaktywna, nawet jeśli dzieje się tu i teraz. Następnie istnieją metody analizy danych, które mają na celu wyjaśnienie nie tylko, dlaczego coś się wydarzyło, ale także co się stało. Tego typu analizy danych są zwykle nazywane analizami diagnostycznymi.

Zarówno metody opisowe, jak i diagnostyczne są zwykle zgrupowane w obszarze raportowania lub analizy biznesowej (BI). Aby móc przewidzieć, co się stanie, musisz użyć innego zestawu technik i metod analitycznych. Prognozy dotyczące przyszłości można przeprowadzać strategicznie lub w czasie rzeczywistym. Aby uzyskać prognozę w czasie rzeczywistym, musisz opracować, przeszkolić i zweryfikować model przed wdrożeniem go na danych w czasie rzeczywistym. Model może następnie wyszukiwać określone wzorce danych i warunki, które zostały wytrenowane przez model, aby pomóc przewidzieć problem, zanim się pojawi.

Ta lista zawiera przykłady rodzajów pytań, które można zadać przy użyciu różnych technik raportowania i BI:

* Raporty standardowe: Jaki był wskaźnik rezygnacji klientów?

* Raporty ad hoc: Jak naprawa kodu przeprowadzona w określonym dniu wpłynęła na wydajność produktu?

* Analiza zapytania: czy podobne problemy z jakością produktów są zgłaszane we wszystkich lokalizacjach geograficznych?

* Alerty: rotacja klientów wzrosła. Jakie działanie jest zalecane? A ta lista zawiera przykłady rodzajów pytań, które możesz zadać przy użyciu różnych technik analitycznych:

* Analiza statystyczna: Jakie czynniki najbardziej przyczyniają się do problemów z jakością produktu?

* Prognozowanie: Jakie będzie zapotrzebowanie na przepustowość za 6 miesięcy?

* Modelowanie predykcyjne: który segment klientów najprawdopodobniej zareaguje na tę kampanię marketingową?

* Optymalizacja. Jaka jest optymalna mieszanka klienta, oferty, ceny i kanału sprzedaży?

Analitykę można również podzielić na dwie kategorie: analitykę podstawową i analitykę zaawansowaną. Podstawowa analityka wykorzystuje podstawowe techniki i metody statystyczne, aby uzyskać wartość z danych, zwykle w sposób ręczny, podczas gdy w zaawansowanej analityce celem jest uzyskanie głębszego wglądu, dokonanie prognoz lub wygenerowanie rekomendacji poprzez autonomiczne lub półautonomiczne badanie danych lub treści przy użyciu bardziej zaawansowanych i wyrafinowanych metod i technik statystycznych. Niektóre przykłady różnic są opisane na tej liście:

* Eksploracyjna analiza danych to statystyczne podejście do analizy zbiorów danych w celu podsumowania ich głównych cech, często za pomocą metod wizualnych. Możesz zdecydować się na użycie modelu statystycznego lub nie, ale jeśli jest używany, taki model służy przede wszystkim do wizualizacji tego, co dane mogą Ci powiedzieć poza formalnym modelowaniem lub testowaniem hipotez. Jest to klasyfikowane jako podstawowe analizy.

* Analityka predykcyjna to wykorzystanie danych, algorytmów statystycznych i technik uczenia maszynowego w celu określenia prawdopodobieństwa przyszłych wyników na podstawie danych historycznych. Jest to klasyfikowane jako zaawansowane analizy.

* Analiza regresji to sposób matematycznego sortowania, które zmienne mają wpływ. Odpowiada na następujące pytania: Które czynniki mają największe znaczenie? Co można zignorować? Jak te czynniki współdziałają ze sobą? A co najważniejsze, na ile jestem pewna co do wszystkich tych czynników? Jest to klasyfikowane jako zaawansowane analizy.

* Eksploracja tekstu lub analiza tekstu to proces eksploracji i analizowania dużych ilości nieustrukturyzowanego tekstu wspomagany przez oprogramowanie, które może identyfikować pojęcia, wzorce, tematy, słowa kluczowe i inne atrybuty w danych. Nadrzędnym celem eksploracji tekstu jest przekształcenie tekstu w dane do analizy za pomocą przetwarzania języka naturalnego (NLP) i różnych metod analitycznych. Eksplorację tekstu można przeprowadzić z bardziej podstawowej perspektywy lub z bardziej zaawansowanej perspektywy, w zależności od przypadku użycia.

Komunikowanie się

Etap komunikacji w nauce o danych polega na upewnieniu się, że spostrzeżenia i wnioski z analizy danych są rozumiane i przekazywane za pomocą różnych środków w celu efektywnego wykorzystania. Obejmuje obszary takie jak

* Raportowanie danych: Proces zbierania i przesyłania danych w celu umożliwienia dokładnej analizy faktów w terenie. Jest to istotna część komunikacji, ponieważ niedokładne raportowanie danych może prowadzić do bardzo nieświadomego podejmowania decyzji w oparciu o niedokładne dowody.

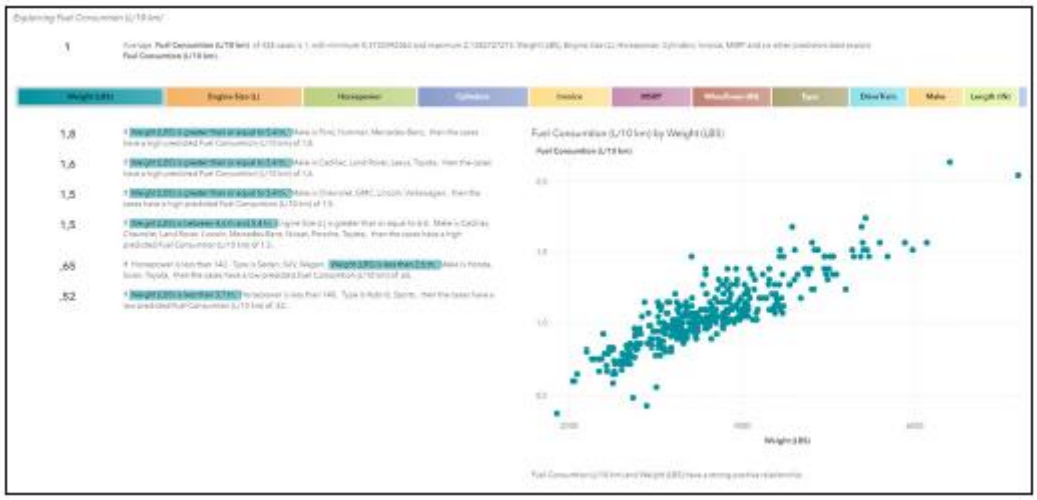
* Wizualizacja danych: można to również postrzegać jako komunikację wizualną, ponieważ obejmuje tworzenie i badanie wizualnej reprezentacji danych i spostrzeżeń. Aby ułatwić przekazywanie wyników analizy danych w sposób jasny i wydajny, wizualizacja danych wykorzystuje grafikę statystyczną, wykresy, grafikę informacyjną i inne narzędzia. Skuteczna wizualizacja pomaga użytkownikom analizować i uzasadniać dane i dowody, ponieważ sprawia, że złożone dane są bardziej dostępne, zrozumiałe i użyteczne.

Użytkownikom mogły zostać przydzielone określone zadania analityczne, takie jak dokonywanie porównań lub zrozumienie przyczynowości, a zasada projektowania wizualizacji graficznej (w tym

przykładzie pokazująca porównania lub pokazując przyczynowość) jest następująca. Tabele są zwykle używane tam, gdzie użytkownicy mogą wyszukać określony pomiar, a wykresy różnych typów służą do pokazywania wzorców lub relacji w danych dla jednej lub większej liczby zmiennych. Rysunek poniższy poniżej ilustruje, jak eksploracja danych może działać przy użyciu formatu tabeli.

Make	Model	Origin	DriveTrain	Type	Cylinders	Engine Size (L)	Frequency	Fuel Consumption (L/10 km)	Horsepower	Invoice	Length (IN)	MSRP
Mercedes-B...	B500	Eur...	All	Wa...	8	5	1	1,17605	302	56 47...	190	60 67...
Isuzu	Ascender S	Asia	All	SUV	6	4,2	1	1,3440571429	275	29 97...	208	31 84...
Toyota	Tundra Access Cab V6 SR5	Asia	All	Truck	6	3,4	1	1,517483871	190	23 52...	218	25 93...
Audi	A6 3.0 Quattro 4dr	Eur...	All	Sedan	6	3	1	1,094	220	35 99...	192	39 64...
Volvo	S60 R 4dr	Eur...	All	Sedan	5	2,5	1	1,094	300	35 38...	181	37 56...
Acura	MDX	Asia	All	SUV	6	3,5	1	1,17605	265	33 33...	189	36 94...
Nissan	Titan King Cab XE	Asia	All	Truck	8	5,6	1	1,4700625	305	24 92...	224	26 65...
Toyota	Sequoia SR5	Asia	All	SUV	8	4,7	1	1,517483871	240	31 82...	204	35 69...
Audi	S4 Quattro 4dr	Eur...	All	Sedan	8	4,2	1	1,3835882353	340	43 55...	179	48 04...
Audi	A8 L Quattro 4dr	Eur...	All	Sedan	8	4,2	1	1,1473658537	330	64 74...	204	69 19...
Subaru	Impreza WRX STi 4dr	Asia	All	Sports	4	2,5	1	1,12004762	300	29 13...	174	31 54...
BMW	330xi 4dr	Eur...	All	Sedan	6	3	1	0,9600408163	225	34 11...	176	37 24...
Infiniti	FX45	Asia	All	Wa...	8	4,5	1	1,3835882353	315	33 12...	189	36 39...
Toyota	Land Cruiser	Asia	All	SUV	8	4,7	1	1,5680666667	325	47 98...	193	54 76...
Subaru	Forester X	Asia	All	Wa...	4	2,5	1	0,9600408163	165	19 64...	175	21 44...
GMC	Sierra HD 2500	USA	All	Truck	8	6	1	1,517483871	300	25 75...	222	29 32...
Mercedes-B...	C240 4dr	Eur...	All	Sedan	6	2,6	1	1,0891363636	168	31 18...	178	33 48...
Jaguar	X-Type 3.0 4dr	Eur...	All	Sedan	6	3	1	1,094	227	30 99...	184	33 99...
Volvo	S80 2.5T 4dr	Eur...	All	Sedan	5	2,5	1	1,000893617	194	35 68...	190	37 88...
Volvo	XC70	Eur...	All	Wa...	5	2,5	1	1,000893617	208	33 11...	186	35 14...
Audi	A6 2.7 Turbo Quattro 4dr	Eur...	All	Sedan	6	2,7	1	1,094	250	38 84...	192	42 84...
Dodge	Grand Caravan SXT	USA	All	Sedan	6	3,8	1	1,094	215	29 81...	201	32 66...

W tym konkretnym przypadku badane dane dotyczą samochodów, a testowana hipoteza dotyczy tego, który atrybut samochodu ma największy wpływ na zużycie paliwa. Czy jest to na przykład marka samochodu, wielkość silnika, moc w koniach mechanicznych, a może waga samochodu? Jak widać eksploracja danych za pomocą tabel ma swoje ograniczenia i nie daje natychmiastowy przegląd. Wymaga szczegółowego przejrzenia danych, aby odkryć relacje i wzorce. Porównaj to z wykresem przedstawionym na rysunku poniżej, gdzie te same dane są wizualizowane w zupełnie inny sposób.



Na rysunku wygenerowano wizualizację w postaci wykresu regresji liniowej dla każdego atrybutu samochodu wraz z tekstem wyjaśniającym siłę każdego związku ze zużyciem paliwa. (Regresja liniowa polega na dopasowaniu linii prostej do zestawu danych, próbując zminimalizować błąd między punktami a dopasowaną linią.) Wykres na rysunku pokazuje bardzo silny dodatni związek między masą samochodu a zużyciem paliwa. Badając zależność między innymi atrybutami a zużyciem paliwa za pomocą wykresu generowanego dla każdej zakładki, dość łatwo będzie znaleźć najsilniejszy związek w porównaniu z tabelą na rysunku wcześniejszym. Jednak w eksploracji danych kluczowe jest zachowanie elastyczności w zakresie stosowanych metod eksploracji. W tym przypadku łatwiej i szybciej było znaleźć zależność za pomocą regresji liniowej, ale w innym przypadku wystarczy tabela lub żadne z wymienionych podejść nie działa. Jeśli masz na przykład dane geograficzne, najlepszym sposobem ich

eksploracji może być użycie mapy geograficznej, na której dane są dystrybuowane na podstawie lokalizacji geograficznej. Ale o tym później.

Dynamizowanie

Ostatnim etapem cyklu życia nauki o danych jest aktywacja spostrzeżeń pochodzących ze wszystkich poprzednich etapów. Ten etap nie zawsze był postrzegany jako część cyklu życia nauki o danych, ale im bardziej społeczeństwo zmierza w kierunku automatyzacji, tym bardziej wzrasta zainteresowanie tym obszarem. . Podejmowanie decyzji o uruchomieniu odnosi się do łączenia wglądu pochodzącego z analizy danych w celu uruchomienia procesu podejmowania decyzji kierowanego przez człowieka lub maszynę, polegającego na identyfikowaniu i decydowaniu o alternatywach dla właściwego działania w oparciu o wartości, zasady, preferencje lub przekonania związane z biznes lub zakres zadania. W rzeczywistości dzieje się tak, że człowiek lub maszyna porównuje wgląd z wcześniej zdefiniowanym zestawem zasad dotyczących tego, co należy zrobić, gdy zostanie spełniony określony zestaw kryteriów. Jeśli kryteria są spełnione, powoduje to podjęcie decyzji lub działania. Wyzwalacz aktywacji może być skierowany do człowieka (na przykład kierownika) w celu podjęcia dalszych decyzji w szerszym kontekście lub do maszyny, gdy wgląd mieści się w zakresie wstępnie zdefiniowanych zasad aktywacji. Automatyzacja zadań lub decyzji przyspiesza i obniża koszty, a odpowiednio skonfigurowana daje również ciągłe i wiarygodne dane o wyniku wdrożonego działania. Etap, na którym decyzje są uruchamiane – przez ludzką rękę lub maszynę – jest jednym z najważniejszych obszarów nauki o danych. Jest to fundamentalne, ponieważ dostarczy specjalistom zajmującym się analizą danych (znanych również jako naukowcy danych) nowe dane oparte na wynikach działania (na przykład rozwiązywanie lub zapobieganie problemowi), które informują naukowców zajmujących się danymi, czy ich modele i algorytmy działają zgodnie z oczekiwaniami po wdrożeniu lub czy należy je poprawić lub ulepszyć. Działania następcze dotyczące wydajności modelu i algorytmu również wspierają koncepcję ciągłego doskonalenia.

AUTOMATYZACJA W KONTEKŚCIE DATA SCIENCE DATA

Jaki jest właściwie związek między nauką o danych a automatyzacją? I czy automatyzacja może przyspieszyć produkcję i wydajność analizy danych? Cóż, zakładając, że ewolucja technologii w społeczeństwie coraz bardziej zmierza w kierunku automatyzacji, nie tylko w przypadku prostych etapów procesów wykonywanych wcześniej przez ludzi, ale także w przypadku złożonych działań zidentyfikowanych i decydowanych przez inteligentne maszyny napędzane algorytmami opracowanymi przez uczenie maszynowe, zależność będzie silny, a produkcja i wydajność analizy danych znacznie przyspieszy dzięki automatyzacji. Decyzje nie będą oczywiście tak naprawdę podejmowane przez maszyny, ale będą oparte na zasadach wstępnie zatwierdzonych przez człowieka, zgodnie z którymi następnie będzie działać maszyna. Uczenie maszynowe nie oznacza, że maszyna może uczyć się bez ograniczeń, ale raczej, że zawsze napotyka granice uczenia się ustanowione przez analityka danych – granice regulowane przez ustalone zasady. Jednak w ramach tych zasad maszyna może nauczyć się optymalizować analizę i wykonywanie przypisanych jej zadań. Pomimo nałożonych na nią granic, automatyzacja napędzana przez maszyny będzie stawać się coraz ważniejsza w data science, nie tylko jako sposób na zwiększenie szybkości (od wykrywania do korekty lub zapobiegania), ale także na obniżenie kosztów oraz zapewnienie jakości i spójności zarządzania danymi , uruchamianie wniosków i generowanie danych na podstawie wyniku. Stosując naukę o danych w swojej firmie, pamiętaj, że nauka o danych ma charakter transformacyjny. Aby w pełni wzmocnić Twój biznes, nie wystarczy po prostu wyjechać i zatrudnić kilku analityków danych (jeśli możesz ich znaleźć) i umieścić ich w tradycyjnym dziale rozwoju oprogramowania i oczekiwać cudów. Aby nauka o danych mogła się rozwijać i generować pełną wartość, musisz być przygotowany na przekształcenie swojej firmy w organizację opartą na danych.

Porządkowanie koncepcji organizacji opartej na danych

Dane to nowy węgiel! Albo nowy olej! Albo nowe złoto! Bez względu na to, z czym porównujesz dane, prawdopodobnie jest to prawda z perspektywy wartości koncepcyjnej. Jako społeczeństwo wkroczyliśmy w nową erę danych i inteligentnych maszyn. I nie jest to przemijający trend ani coś, czego możesz lub powinieneś unikać. Zamiast tego powinieneś to zaakceptować i zadać sobie pytanie, czy rozumiesz go wystarczająco, aby wykorzystać go w swojej firmie. Bądź otwarty i ciekawy! Odważ się zadać sobie pytanie, czy naprawdę rozumiesz, co oznacza kierowanie się danymi. Pojęcie bycia opartym na danych jest kamieniem węgielnym, który musisz zrozumieć, aby poprawnie wykonywać jakąkolwiek strategiczną pracę w nauce o danych i jest omówiony w kilku częściach tej książki. W tym rozdziale postaram się przedstawić ogólny obraz tego, jak myśleć i uzasadniać ideę bycia opartym na danych. Jeśli zaczniesz od umieszczenia zmian zachodzących w społeczeństwie w szerszym kontekście, to powszechnie wiadomo, że my, ludzie, doświadczamy teraz czwartej rewolucji przemysłowej, napędzanej przez dostęp do danych i zaawansowaną technologię. Jest również określany jako rewolucja cyfrowa. Ale bądź świadomy! Digitalizacja lub cyfryzacja firmy to nie to samo, co bycie opartym na danych. Digitalizacja to szeroko stosowana koncepcja, która zasadniczo odnosi się do przejścia z technologii analogowej na cyfrową, podobnie jak konwersja danych do formatu cyfrowego. W związku z tym digitalizacja odnosi się do tego, aby zdigitalizowane informacje działały w Twojej firmie. Koncepcja cyfryzacji firmy jest czasami mylona z napędzaniem danych. Należy jednak pamiętać, że cyfryzacja danych to nie tylko dobra rzecz — to podstawa funkcjonowania przedsiębiorstwa opartego na danych. Bez cyfryzacji po prostu nie można stać się napędzanym danymi.

Podejście oparte na danych

W organizacji opartej na danych punktem wyjścia są dane. To naprawdę podstawa wszystkiego. Ale co to właściwie oznacza? Cóż, bycie opartym na danych oznacza, że musisz być gotowy do poważnego traktowania danych. I co to znaczy? Cóż, w praktyce oznacza to, że dane są punktem wyjścia i używasz danych do analizy i zrozumienia, jaki rodzaj biznesu powinieneś robić. Musisz potraktować wynik analizy na tyle poważnie, aby być przygotowanym na odpowiednią zmianę modeli biznesowych. Musisz być gotowy, aby zaufać i wykorzystać dane do rozwoju swojej firmy. To powinno być Twoim głównym zmartwieniem w firmie. Musisz mieć „obsesję na punkcie danych". Zanim wyjaśnię, co to znaczy mieć obsesję na punkcie danych, zastanów się, jak dzisiaj robisz rzeczy w swojej firmie. Czy jest to trochę oparte na danych? A może wcale? Gdzie jest punkt wyjścia w różnych obszarach biznesowych? Warto porównywać podejścia stosowane w tradycyjnym biznesie i organizacji opartej na danych. Wielu liderów firm myśli, że ich firmy są oparte na danych tylko dlatego, że zbierają i analizują dane. Ale wszystko sprowadza się do tego, w jaki sposób dane napędzają (lub nie napędzają) priorytetów biznesowych, decyzji i realizacji, które mówią Ci, jak naprawdę jest oparta na danych Twoja firma. Zrozumienie, jaki jest punkt wyjścia, pomoże Ci zdefiniować punkt zerowy i określić, które obszary wymagają większej uwagi, aby dokonać zmian.

Obsesja na punkcie danych

Co więc właściwie oznacza termin „obsesja na punkcie danych"? To naprawdę bardzo proste: oznacza to, że zawsze powinieneś zakładać, że dostęp i wykorzystanie danych może usprawnić Twój biznes - we wszystkich aspektach. Skorzystaj z poniższej listy pytań, aby określić, jak bardzo Twoja organizacja ma obsesję na punkcie danych:

* Jakie dane musisz wykorzystać jako firma, w oparciu o swoje cele strategiczne? Czy zbierasz już te dane? Jeśli nie, to jak to zdobyć?

* Czy posiadasz wszystkie potrzebne dane? Jeśli nie, w jaki sposób możesz zabezpieczyć prawa do korzystania z niego dla swoich potrzeb (wewnętrzna wydajność lub możliwości biznesowe)?

* Czy dane są rozmieszczone geograficznie w różnych krajach? Jeśli tak, co musi się stać z Twoją infrastrukturą, aby umożliwić Ci jej efektywne wykorzystanie?

* Czy dane są wrażliwe? To znaczy, czy zawiera dane osobowe? Jeśli tak, jakie są obowiązujące przepisy ustawowe i wykonawcze dotyczące danych? (Pamiętaj, czy te przepisy i regulacje ulegają zmianie w zależności od kraju, w którym znajduje się konkretna placówka do przechowywania danych.) Jak zamierzasz korzystać z danych wrażliwych?

* Czy potrzebujesz dostępu do danych w czasie rzeczywistym, aby analizować i realizować swoje przypadki użycia? Jeśli tak, jakiego rodzaju architektury danych potrzebujesz?

* Jakie okresy przechowywania danych musisz ustalić dla różnych typów danych wykorzystywanych przez Twoją organizację? Do czego wykorzystasz wybrane typy danych? Czy masz kontrolę, jeśli chodzi o oczekiwane ilości danych i koszty przechowywania danych w zależności od typu danych?

* Czy możesz zautomatyzować większość czynności związanych z akwizycją danych i zarządzaniem danymi? Jeśli tak, jakie jest najlepsze rozwiązanie w zakresie architektury danych?

* Czy musisz uwzględnić eksploracyjne środowisko programistyczne, a także wydajne i wysoce zautomatyzowane środowisko produkcyjne w tej samej architekturze? Jeśli tak, jak to sobie uświadomisz?

* Czy pracownicy są gotowi do działania w oparciu o dane? Czy potencjał, wartość i zakres zmiany zostały jasno określone i zakomunikowane? Jeśli tak, czy pracownicy są gotowi na tę zmianę?

* Czy menedżerowie i liderzy są zaangażowani w to, co to znaczy stać się opartym na danych? Czy w pełni rozumieją, co musi się fundamentalnie zmienić? Jeśli tak, to czy menedżerowie i liderzy są gotowi do podjęcia ważnych decyzji w oparciu o dane?

Pytania, które tutaj zamieszczam, nie stanowią wyczerpującej listy, ale obejmują niektóre z głównych obszarów, którymi należy się zająć z perspektywy opartej na danych. Zauważ, że te pytania nie obejmują niczego związanego z wykorzystaniem technik uczenia maszynowego lub sztucznej inteligencji. Powodem, który nie jest uwzględniony, jest to, że w praktyce firma może być oparta na danych w oparciu o dane, analizy i automatyzację. Jednak firmy, które również skutecznie integrują wykorzystanie technologii, takich jak uczenie maszynowe i sztuczna inteligencja, mają lepsze podstawy do reagowania na ewolucję napędzaną przez maszyny w społeczeństwie.

Uporządkowanie koncepcji uczenia maszynowego

Ludzie często proszą mnie o wyjaśnienie różnicy między zaawansowaną analityką a uczeniem maszynowym oraz o wskazanie, kiedy warto wybrać jedno, a drugie podejście. Zawsze zaczynam od zdefiniowania uczenia maszynowego. Uczenie maszynowe (ML) to naukowe badanie algorytmów i modeli statystycznych, których systemy komputerowe używają do stopniowego zwiększania wydajności określonego zadania. Algorytmy uczenia maszynowego budują model matematyczny na podstawie danych przykładowych, znanych jako dane szkoleniowe, w celu przewidywania lub podejmowania decyzji bez wyraźnego zaprogramowania do wykonania zadania. Oto, w jaki sposób zaawansowana analityka i ML mają pewne cechy wspólne:

* Zarówno zaawansowane techniki analityczne, jak i techniki uczenia maszynowego są wykorzystywane do budowania i wykonywania zaawansowanych modeli matematycznych i

statystycznych, a także budowania zoptymalizowanych modeli, które można wykorzystać do przewidywania zdarzeń przed ich wystąpieniem.

* Obie metody wykorzystują dane do opracowania modeli i obie wymagają zdefiniowanych zasad dotyczących modeli.

* Automatyzacji można używać do uruchamiania zarówno modeli analitycznych, jak i modeli uczenia maszynowego po ich wprowadzeniu do produkcji.

A co z różnicami między zaawansowaną analityką a uczeniem maszynowym?

* Istnieje różnica w tym, kim jest aktor podczas tworzenia modelu. W modelu zaawansowanej analizy aktor jest człowiekiem; w modelu uczenia maszynowego aktor jest (oczywiście) maszyną.

* Istnieje również różnica w formacie modelu. Modele analityczne są opracowywane i wdrażane zgodnie z projektem zdefiniowanym przez człowieka, podczas gdy modele ML są dynamiczne i zmieniają projekt i podejście w miarę ich uczenia się na podstawie danych, optymalizując projekt po drodze. Modele uczenia maszynowego można również wdrażać jako dynamiczne, co oznacza, że nadal szkolą, uczą się i optymalizują projekt, gdy są wystawione na rzeczywiste dane i ich kontekst na żywo.

* Kolejna różnica między modelami analitycznymi a modelami uczenia maszynowego dotyczy różnicy w sposobie testowania modeli przy użyciu danych (dla analityki) i trenowania przy użyciu danych (dla uczenia maszynowego). W analityce dane są wykorzystywane do testowania, czy zdefiniowany wynik został osiągnięty zgodnie z oczekiwaniami, podczas gdy w uczeniu maszynowym dane są wykorzystywane do trenowania modelu w celu optymalizacji jego projektu w zależności od charakteru danych.

* Wreszcie, różnią się techniki i narzędzia wykorzystywane do opracowywania zaawansowanych modeli analitycznych i modeli ML. Techniki modelowania uczenia maszynowego są znacznie bardziej zaawansowane i opierają się na innych zasadach związanych z tym, jak maszyna nauczy się optymalizować wydajność modelu.

Modele są zawsze opracowywane i testowane w sposób statyczny, gdzie człowiek decyduje, których metod statystycznych użyć i jak przetestować model przy użyciu zdefiniowanego zestawu danych próbki w celu osiągnięcia optymalnej wydajności modelu. I niezależnie od tego, ile danych (lub które dane) przetworzysz przez model analityczny, pozostaje on taki sam, dopóki człowiek nie zdecyduje się poprawić lub rozwinąć modelu. W rozwoju ML aktor ludzki również decyduje, którą technikę lub metodę zastosować. Metody szkoleniowe w ML różnią się w zależności od używanej techniki - możesz na przykład zastosować uczenie nadzorowane lub uczenie bez nadzoru, uczenie częściowo nadzorowane, uczenie ze wzmocnieniem, a nawet uczenie głębokie, które jest bardziej złożoną metodą. Możliwe jest nawet połączenie dwóch metod, takich jak łączenie uczenia się ze wzmocnieniem z głębokim uczeniem się z tak zwanym uczeniem się ze wzmocnieniem. Zamiast statycznego podejścia używanego w tradycyjnym testowaniu modeli, w modelach ML najpierw trenujesz model przy użyciu wybranego zestawu danych szkoleniowych, który powinien reprezentować środowisko docelowe, w którym zamierzasz wdrożyć model ML. Podczas szkolenia wydajność modelu jest testowana w celu monitorowania postępu uczenia się, a także pomiaru dokładności modelu. W ramach wybranej metody ML, pozwalasz algorytmowi (aktorowi maszyny) szkolić się na zbiorze danych treningowych, aby osiągnąć wyznaczony cel. Następnie maszyna kontynuuje trenowanie modelu ML, aby ewoluować i znajdować najbardziej zoptymalizowaną wydajność modelu, o ile na to pozwolisz. Nadejdzie czas, kiedy nie będzie można poprawić dokładności

modelu przy użyciu zestawu uczącego. Na tym etapie musisz ocenić, czy dokładność modelu jest wystarczająca do wdrożenia.

Jeśli zdecydujesz, że aktor maszyny osiągnął wystarczający poziom szkolenia, musisz zdecydować, jak wdrożyć model w środowisku docelowym, innymi słowy wdrożyć do produkcji. W tym momencie masz dwie opcje. Możesz zdecydować, że model jest wystarczająco przeszkolony, aby osiągnąć swój cel i że możesz go wdrożyć jako model statyczny - co oznacza, że nie będzie się już uczył i optymalizował wydajność opartą na danych, niezależnie od zmian zachodzących w środowisku docelowym. Możesz też zdecydować się na wdrożenie modelu ML do środowiska produkcyjnego jako modelu dynamicznego, co oznacza, że będzie on nadal ewoluował i optymalizował swoją wydajność na podstawie danych i zachowań, które wypełniają model w środowisku produkcyjnym.

Czasami jest to również nazywane szkoleniem online. Kiedy więc należy wybrać model i podejście do wdrażania? Cóż, to zależy od wielu czynników. Zasadniczo nigdy nie powinieneś używać ML, jeśli możesz wykonać zadanie, korzystając z podejścia analitycznego. . Dlaczego? Z tego samego powodu nie używasz młota kowalskiego do wbijania gwoździa. Może ci się to uda, ale równie łatwo możesz zniszczyć gwóźdź i zranić się, powodując stratę czasu i pieniędzy. Jeśli chodzi o wdrożenie statyczne lub dynamiczne, zależy to od modelu biznesowego i tego, czy środowisko docelowe jest statyczne (zmiany zdarzają się rzadko i zwykle są niewielkie) czy dynamiczne (zmiany zachodzą często i na dużą skalę). Jeśli na przykład opracowujesz algorytm do tworzenia rekomendacji online na podstawie wcześniejszych zachowań użytkowników, konieczne jest wdrożenie dynamicznego modelu ML; w przeciwnym razie nie możesz osiągnąć swojego celu. Jeśli z drugiej strony celem modelu ML jest umożliwienie maszynie znalezienia optymalnego sposobu automatyzacji zestawu złożonych zadań, które mają pozostać niezmienione w czasie, zaleca się wdrożenie modelu ML jako statycznego modelu w środowisku docelowym. Pamiętaj, że wdrażanie modeli ML w środowiskach na żywo wymaga od Ciebie więcej zasobów. Szkolenie w zakresie uczenia maszynowego jest złożone i wymaga dużej wydajności przetwarzania, a także większego monitorowania modelu ML. Musisz upewnić się, że model ML nadal działa zgodnie z oczekiwaniami i nie pogarsza się ani nie odbiega od celu w ramach szkolenia na żywo. Innym aspektem do rozważenia jest potrzeba zapewnienia, że model może wchodzić w interakcje z innymi dynamicznymi modelami ML w środowisku docelowym bez zakłócania sobie nawzajem celów lub działania w sposób, który prowadzi do wzajemnego znoszenia się modeli. (To, co tutaj robisz, jest często określane jako zapewnienie interoperacyjności modelu.)

Definiowanie i określanie zakresu strategii analizy danych

Aby zrozumieć elementy składowe strategii analizy danych, a także jej obecne i przyszłe znaczenie, warto przyjrzeć się niektórym z głównych komponentów na wysokim poziomie. Następnie szczegółowo omawiam każdą z tych różnych części w tej książce. Ale wcześniej muszę krótko wyjaśnić różnicę między strategią analizy danych a strategią dotyczącą danych. Na wysokim poziomie strategia analizy danych odnosi się do strategii, którą definiujesz w odniesieniu do całej inwestycji w naukę danych w naszej firmie. Obejmuje takie obszary, jak ogólne cele nauki o danych i wybory strategiczne, strategie regulacyjne, zapotrzebowanie na dane, kompetencje i umiejętności, architektura danych, a także sposób pomiaru wyników. Z drugiej strony strategia danych stanowi podzbiór strategii analizy danych i koncentruje się na wytyczeniu kierunku strategicznego bezpośrednio związanego z danymi. Obejmuje to takie obszary, jak zakres danych, zgoda na dane, względy prawne, regulacyjne i etyczne, częstotliwość gromadzenia danych, okresy przechowywania danych, proces i zasady zarządzania danymi, a także, co nie mniej ważne; zarządzanie danymi. Obie strategie są potrzebne, aby odnieść sukces z inwestycją w naukę danych i powinny się wzajemnie uzupełniać, aby działały.

Cele

Jeśli pytam o cele strategii data science, pytam, czy istnieją jasne cele firmy ustalone i uzgodnione dla którejkolwiek z inwestycji dokonanych w data science. Czy cele są sformułowane w sposób umożliwiający ich realizację i mierzenie sukcesu? Jeśli nie, to cele należy przeformułować; jest to krytycznie ważny punkt wyjścia, który musi zostać poprawnie wypełniony, aby odnieść sukces. Nauka o danych to nowa dziedzina, która daje firmom niesamowite możliwości przeprowadzenia fundamentalnej transformacji, ale jest złożona i często nie jest w pełni rozumiana przez najwyższe kierownictwo. Należy zastanowić się, czy zrozumienie przez zespół wykonawczy na temat nauki o danych jest wystarczające do wyznaczenia właściwych celów, czy też należy ich edukować, a następnie kierować w ustalaniu celów. Niezależnie od tego, czy jesteś menedżerem, czy pracownikiem małej czy dużej firmy, jeśli chcesz, aby Twoja firma odniosła sukces dzięki inwestycjom w analitykę danych, nie siedź i miej nadzieję, że kierownictwo Twojej firmy zrozumie, co należy zrobić. Jeśli masz wiedzę w tej dziedzinie, spraw, aby Twój głos był słyszalny, a jeśli nie, nie wahaj się przyjąć pomocy od osób, które mają doświadczenie w tej dziedzinie. Jeśli zdecydujesz się na zatrudnienie zewnętrznych ekspertów, którzy pomogą Ci w opracowywaniu strategii w zakresie analizy danych, najpierw zapoznaj się z tym obszarem, aby móc ocenić trafność ich rekomendacji dla Twojej firmy - miejsca, w którym jesteś ekspertem.

Podejście

Przyjęcie odpowiedniego początkowego podejścia jest fundamentalną częścią Twojej strategii w zakresie nauki o danych - określi, czy Twoja firma przyjmie odpowiednie podejście do wdrażania i transformacji inwestycji w naukę danych. Na przykład, czy podejście jest wystarczająco ambitne - czy też zbyt ambitne, biorąc pod uwagę szacunki czasowe związane z dostępnymi kompetencjami? Czy istnieje jasna strategia biznesowa i oczekiwana wartość, do której może się odnosić strategia analizy danych? Poświęcenie czasu na przemyślenie tego podejścia z pewnością się opłaci, ponieważ jeśli nie wiesz, dokąd zmierzasz, jest mało prawdopodobne, że tam trafisz.

Wybory

Termin wybory odnosi się tutaj do strategicznych wyborów niezbędnych do przyspieszenia transformacji nauki o danych. Strategia, którą tworzysz, nie może polegać na robieniu wszystkiego. Równie ważne jest dokonywanie strategicznych wyborów dotyczących tego, co robić, jak podejmowanie decyzji o tym, czego nie robić. Decyzje mogą być również rozłożone w czasie w różny sposób, ponieważ wybory mogą dotyczyć rozpoczęcia od określonego obszaru biznesowego lub grupy klientów, uczenia się na podstawie tego doświadczenia, a następnie kontynuowania włączania innych obszarów lub klientów. Ta sama strategia dotyczy wyboru kategorii lub typów danych, na których należy skupić się na wczesnym etapie, a nie później, w miarę dojrzewania firmy i rozszerzania możliwości.

Dane

Zdefiniowanie strategii danych jest podstawą strategii analizy danych - obejmuje wszystkie aspekty związane z danymi, takie jak zrozumienie różnych typów danych, do których potrzebujesz dostępu, aby osiągnąć cele biznesowe. Czy dane są dostępne? Jak podejdziesz do zarządzania danymi i przechowywania danych? Czy ustaliłeś priorytety danych? Czy określiłeś i ustaliłeś cele w zakresie jakości danych? Inny ważny aspekt danych dotyczy zarządzania i bezpieczeństwa danych. Dane będą jednym z Twoich najcenniejszych zasobów w przyszłości; to, jak ją traktujesz, ma fundamentalne znaczenie dla sukcesu Twojej firmy.

Prawo

Zrozumienie konsekwencji prawnych dla potrzebnych danych w zakresie praw dostępu, własności i modeli użytkowania ma kluczowe znaczenie. Jeśli nie jesteś na bieżąco z tym aspektem na początku, możesz znaleźć się w sytuacji, w której nie będziesz mógł uzyskać danych potrzebnych dla Twojej firmy bez złamania prawa lub, nawet jeśli możesz je zdobyć, możesz zdać sobie sprawę, że nie możesz jej używać w sposób, w jaki potrzebujesz, aby zrealizować swoje cele biznesowe. Przepisy i regulacje związane z prywatnością danych sięgają dalej, niż wielu ludziom się wydaje, i ciągle się zmieniają, aby chronić integralność danych ludzi. . Jest to dobre z punktu widzenia prywatności, ale nie zawsze działa dobrze w przypadku innowacji w zakresie danych. Dlatego, jako dobra inwestycja, zawsze powinieneś być na bieżąco z przepisami i regulacjami dotyczącymi danych potrzebnych dla Twojej firmy.

Etyka

Etyka, obszar o rosnącym znaczeniu, odnosi się do tworzenia jasnych wytycznych etycznych dotyczących podejścia do data science w firmie. Wewnętrznie termin ten odnosi się do zapewnienia odpowiedzialnego podejścia do wykorzystania danych i zarządzania nimi, jeśli chodzi o ochronę prywatności danych klientów lub innych interesariuszy. Jednym ze sposobów ochrony prywatności jest anonimizacja danych osobowych w zbiorach danych. Zewnętrznie naleganie na etykę nauki o danych ma kluczowe znaczenie, jeśli chodzi o zdobycie zaufania klientów do sposobu, w jaki obchodzisz się z danymi. Gdy uczenie maszynowe lub sztuczna inteligencja jest wprowadzana - zwłaszcza gdy mamy do czynienia z automatyzacją decyzji i działaniami prewencyjnymi - dotyka innej perspektywy etycznej: „wyjaśnialności" algorytmów. Odnosi się do idei, że musi być możliwe wyjaśnienie decyzji lub działania podjętego przez maszynę. Uczenie maszynowe lub sztuczna inteligencja nie mogą stać się automatycznym, czarnoskrzynkowym wykonaniem przez maszynę. Ludzie muszą… zachować kontrolę, aby zapewnić przejrzystość algorytmów sztucznej inteligencji i zapewnić zachowanie granic etycznych.

Kompetencja

W oparciu o wyznaczone cele, dokonane wybory i wybrane podejście, musisz upewnić się, że posiadasz odpowiednie kompetencje do realizacji swoich celów. Stworzenie doświadczonego i kompetentnego zespołu zajmującego się analizą danych jest łatwiejsze do powiedzenia niż do zrobienia. Dlaczego? Cóż, naprawdę potrzebujesz trzech głównych kategorii kompetencji, a dostępność doświadczonych analityków danych na rynku jest obecnie bardzo niska, po prostu dlatego, że niewielu analityków danych ma wystarczające doświadczenie, a zapotrzebowanie na tego typu kompetencje jest bardzo wysokie. Nie poradzisz sobie z zatrudnieniem tylko analityków danych. Inżynierowie danych z prawdziwym zrozumieniem danych w centrum uwagi mają fundamentalne znaczenie. Bez dobrego zarządzania danymi naukowcy zajmujący się danymi nie mogą wykonywać swojej magii algorytmicznej. To takie proste. Wreszcie, musisz zabezpieczyć wiedzę specjalistyczną w dziedzinie domeny dla docelowego obszaru, niezależnie od tego, czy jest to rozległe zrozumienie biznesowe, czy wyjątkowe zrozumienie operacyjne. Absolutnie kluczowe jest, aby eksperci dziedzinowi ściśle współpracowali z inżynierami danych i naukowcami danych, aby osiągnąć wydajne zespoły zajmujące się analizą danych w Twojej organizacji.

Infrastruktura

Mówiąc o infrastrukturze, chodzi o zrozumienie, co jest potrzebne w zakresie architektury danych i aplikacji, aby zapewnić wydajne i innowacyjne środowisko dla zespołów zajmujących się analizą danych. Obejmuje to rozważenie zarówno środowiska programistycznego (obszar roboczy, w którym wprowadzasz innowacje, rozwijasz, szkolisz i testujesz nowe możliwości) oraz środowisko produkcyjne (środowisko wykonawcze, w którym wdrażasz i uruchamiasz swoje rozwiązania). Infrastruktura obejmuje wszystkie aspekty, od sposobu, w jaki skonfigurujesz gromadzenie danych/pozyskiwanie danych, anonimizację, przechowywanie danych, zarządzanie danymi i warstwę aplikacji z narzędziami

do analizy oraz środowiska rozwoju i produkcji ML/AI. Nie da się zidentyfikować i skonfigurować idealnego środowiska, zwłaszcza dlatego, że ewolucja technologii w tej dziedzinie postępuje bardzo szybko. Jednak istotną częścią konfiguracji infrastruktury jest uniknięcie sytuacji, w której stajesz się całkowicie zależny od określonego dostawcy infrastruktury (na przykład sprzętu, oprogramowania lub chmury). Nie chodzi mi o to, że powinieneś sięgać tylko po produkty open source, ale chodzi mi o to, że musisz dokładnie przemyśleć, jakich bloków budulcowych używasz, a następnie upewnić się, że są one wymienialne na dłuższą metę, jeśli zajdzie taka potrzeba.

Zarządzanie i bezpieczeństwo

Aktywna praca z zarządzaniem danymi i bezpieczeństwem zapewni stałą kontrolę nad wykorzystaniem danych. Jest to ważne nie tylko z punktu widzenia zdobycia zaufania klientów, ale w wielu przypadkach jest to również konieczność przestrzegania prawa. Śledzenie, które dane są gromadzone, przechowywane i wykorzystywane w jakich przypadkach użycia jest minimalnym wymogiem dla większości typów danych. Przepracowanie obszaru zarządzania i bezpieczeństwa będzie miało wpływ na wydajność i innowacyjność w zakresie nauki o danych. Częstym błędem jest nadopiekuńczość w odniesieniu do korzystania z danych, zamykanie wszystkich danych w takim stopniu, że nikt nie ma dostępu do tego, czego potrzebują do wykonywania swojej pracy. Dlatego do konfiguracji zarządzania danymi i bezpieczeństwa należy podchodzić z nastawieniem na otwartość do udostępniania danych pracownikom wewnątrz organizacji. Zamknij bramy przed osobami z zewnątrz, ale dąż do wewnętrznego podejścia do otwartych danych, zwiększając współpracę, ponowne wykorzystanie i innowacyjność.

Modele komercyjne/biznesowe

W ramach strategii analizy danych firmy musisz rozważyć, czy chcesz skoncentrować swoje wysiłki tylko wewnętrznie, aby poprawić wydajność operacyjną, czy też masz ambicje, aby wykorzystać naukę o danych do ulepszenia komercyjnych modeli biznesowych. Usprawnienie firmy za pomocą analityki danych pozwoli absolutnie poszerzyć swoje możliwości, zarówno w usprawnianiu obecnego biznesu, jak i pomaganiu w znajdowaniu nowych możliwości. Podczas komercjalizacji danych postępuj ostrożnie. Jeśli nie dokonałeś najpierw wewnętrznej transformacji, wdrażając operacje oparte na danych, prawdopodobnie nie będziesz w stanie w pełni wykorzystać zewnętrznego podejścia do analizy danych z perspektywy biznesowej. Nie oznacza to, że musisz wdrażać i prowadzić operacje oparte na danych w całej firmie, ale takie operacje będą potrzebne w obszarach związanych z nowymi modelami biznesowymi opartymi na nauce danych i ofertami komercyjnymi, które zamierzasz zrealizować.

Pomiary

Bez mierzenia swojego sukcesu, skąd będziesz wiedzieć, czy rzeczywiście osiągnąłeś swoje cele? Albo być w stanie to udowodnić. Mimo to wiele firm nie myśli o pomiarach na początku. Pomiary są potrzebne nie tylko z perspektywy wewnętrznej efektywności operacyjnej, ale także w celu sprawdzenia, czy udało się dotrzymać obietnic złożonych klientom. Jest to ważne niezależnie od tego, czy uzgodnione cele klientów zostały zakontraktowane, czy nie. Zawsze powinno być dla Ciebie priorytetem, aby wiedzieć, jak Twoja firma radzi sobie z Twoimi celami. Informacje zwrotne dadzą ci wszystkie informacje, których potrzebujesz, aby określić, na czym stoi firma, co należy poprawić, a co być może już zostało osiągnięte. Tak, ustalenie pomiarów na wczesnym etapie ma fundamentalne znaczenie, jeśli chodzi o zapewnienie ciągłej nauki w Twojej firmie, ale pokazuje również klientom, że zależy Ci na osiągnięciu swoich celów. Nie zapomnij jednak przemyśleć struktury metryk, której planujesz użyć. Identyfikacja i zdefiniowanie właściwego zestawu metryk od samego początku nie jest łatwym zadaniem. Jest to również coś, co należy z czasem ponownie ocenić, w oparciu o to, jakie

pomiary faktycznie dają wgląd i informacje zwrotne potrzebne na temat tego, co idzie dobrze – a co nie.

Biorąc pod uwagę nieodłączną złożoność w nauce o danych

Miasta są złożonymi systemami, a polityki miejskie są zazwyczaj tworzone w złożonych środowiskach, w których należy wziąć pod uwagę wiele czynników obejmujących całe spektrum czynników społecznych, środowiskowych, ekonomicznych i technologicznych. Jednak w ostatnich latach złożoność miast była lepiej zarządzana przez ewolucję nauki o danych. Możliwość wykonywania modelowania urbanistycznego i symulowania różnych przyszłych scenariuszy na podstawie rzeczywistych danych otworzyła wiele nowych możliwości związanych z planowaniem urbanistycznym i inwestycjami. Te zmiany w nauce o danych umożliwiły agencjom rządowym lepsze zrozumienie złożonych problemów miejskich, przewidywanie możliwych scenariuszy i podejmowanie najlepszych decyzji politycznych i inwestycyjnych. Ale co tak naprawdę oznacza i do czego odnosi się złożoność? Cóż, moim zdaniem społeczeństwo ma ogólne błędne przekonanie, że złożoność jest zawsze zła. Tak, często najprostsze rozwiązanie jest najlepsze – truizm, który istnieje od XIII wieku, kiedy franciszkanin William z Ockham wymyślił oryginalne sformułowanie, znane obecnie powszechnie jako brzytwa Ockhama. Ale w niektórych przypadkach to faktyczna złożoność sprawy sprawia, że jest ona interesująca dla określonego rozwiązania technicznego. Tak jest w przypadku nauki o danych. Jeśli dany problem jest prosty i można go rozwiązać za pomocą prostego rozwiązania (na przykład przy użyciu zwykłego kodu). Nie ma sensu używać technik uczenia maszynowego do rozwiązania problemu i rzucać na niego oprogramowania samouczącego. W rzeczywistości, jeśli Twoja firma jest prosta i nieskomplikowana, a chcesz, aby tak pozostało i nie wykraczała poza obecne modele biznesowe, możesz zyskać bardzo niewielką wartość, dodając uczenie maszynowe i sztuczną inteligencję do mieszania. Jeśli jednak interesuje Cię rozwijanie swojej firmy oraz możliwości produktów lub usług poza to, co jest obecnie możliwe, bardziej zaawansowana analiza danych może być sposobem na osiągnięcie tego celu. Jednak podróż nie będzie prosta. Nauka o danych jest zdecydowanie aktywatorem i możesz jej użyć, aby zacząć prosto i od tego momentu rozwijać się. Jednak praca z nauką o danych w centrum Twojej firmy wymaga wykwalifikowanych naukowców i architektów danych. Nauka o danych to rzemiosło, które wymaga umiejętności w kilku dyscyplinach, w tym dobrej znajomości architektury. Nie jest to kompetencja, którą zdobywasz po prostu uczestnicząc w kursie R lub Python; to znacznie więcej i nie należy lekceważyć wymaganego poziomu wiedzy. Bardziej zaawansowana nauka o danych, w której wykorzystuje się uczenie maszynowe i sztuczną inteligencję, to złożona sprawa, która służy do rozwiązywania złożonych problemów. Dlatego ten rozdział ma na celu pomóc Ci zrozumieć podstawy tego, dlaczego nauka o danych jest złożona, a także dlaczego potencjał tkwi w tej samej złożoności.

Diagnozowanie złożoności w nauce o danych

Co to znaczy, gdy ludzie mówią, że nauka o danych jest z natury złożona? Cóż, ze swej natury nauka o danych – a zwłaszcza techniki, takie jak uczenie maszynowe i sztuczna inteligencja – są zbudowane w celu rozwiązywania złożonych problemów, których nie mogą rozwiązać nawet najzdolniejsi ludzie. Nie oznacza to, że maszyna pokona ludzi pierwszego dnia, ale z czasem tak się stanie – przynajmniej na tyle, na ile chcemy, aby maszyny nas przewyższały, co jest regulowane przez zasady, których używamy, aby ograniczyć zdolność maszyn do poprawy uczenia się. Ostatecznie nie chodzi o to, jak inteligentne maszyny mogą stać się, ale raczej o to, jak inteligentne mogą uczynić nas ludźmi. W uczeniu maszynowym algorytmy są budowane, aby nauczyć się optymalizować swoje realizacje najpierw na zestawie danych treningowych w środowisku laboratoryjnym, a następnie na danych rzeczywistych. Algorytmy można budować, aby uczyć się z wielu różnych źródeł danych i parametrów, znacznie więcej niż jest to możliwe do szybkiego i ciągłego przetwarzania przez ludzki mózg. Spójrzmy prawdzie w oczy; tak długo, jak działają w elastycznej i skalowalnej architekturze, maszyny nie potrzebują uśpienia, odpoczynku i w zasadzie nie mają ograniczeń pod względem pojemności, jeśli chodzi o przesyłanie większej ilości danych lub inne polityki i ograniczenia. Ludzie po prostu nie mogą się z tym równać.

Automatyzacja powtarzalnego zadania w czasie rzeczywistym przy użyciu wielu różnych źródeł danych niekoniecznie musi być rozwiązywana przez uczenie maszynowe. Wiele zadań automatyzacji można wykonać przy użyciu statycznego modelu statystycznego; jeśli model nie musi się zmieniać i optymalizować w czasie, nie ma potrzeby uczenia maszynowego. Powinno to wchodzić w grę tylko wtedy, gdy dany problem dynamicznie się zmienia i jest złożony. Dopiero wtedy algorytm uczenia maszynowego jest potrzebny do zarządzania złożonością, z którą ludzki mózg nie może sobie poradzić (w czasie rzeczywistym lub nie), poprawiając realizację wystarczająco szybko i w tylu wymiarach, ile jest to wymagane. Ponieważ nauka o danych służy jako podstawa w zarządzaniu rosnącą złożonością naszego wkrótce w pełni zdigitalizowanego i połączonego społeczeństwa, jest ona podstawą rozwiązań potrzebnych do dalszej ewolucji technicznej. Atrakcyjność data science jest w dużej mierze związana z szybko rosnącym dostępem do danych i większą dostępnością technologii, takich jak uczenie maszynowe i sztuczna inteligencja. Jednak zarządzanie złożonością jest często nie tylko trudne do zarządzania za pomocą prostego rozwiązania – czasami jest to niemożliwe. Biorąc pod uwagę ten fakt, niezwykle ważne jest, aby zrozumieć, że chociaż nauka o danych jest dyscypliną naukową, która odegra kluczową rolę w posuwaniu naszego społeczeństwa w kierunku przyszłości większej automatyzacji, robotyki i samozarobkowego oprogramowania, nigdy nie twierdziła, że opiera się na prosta nauka. Zamiast tego nauka o danych jest z natury złożona.

Rozpoznawanie złożoności jako potencjału

Jeśli założymy, że data science jest złożona, jak możemy przekształcić to w potencjał biznesowy? Cóż, ze względu na złożony charakter nauki o danych będzie to wymagało umiejętności, które nie są łatwe do zdobycia, a zatem nie są czymś, co każda firma ma pod ręką. Właściwe podejście do niej we właściwym czasie może zatem przełożyć się na przewagę konkurencyjną Twojej firmy. Ponieważ nauka o danych jest złożona, oznacza to również, że aby zrozumieć, o co chodzi, musisz zainwestować dużo czasu i pieniędzy, aby lepiej zrozumieć, od czego zacząć, co to oznacza dla Twojej firmy i jakie wyniki biznesowe możesz oczekiwać. Jednym z kluczowych elementów, aby to zrobić dobrze, jest poświęcenie czasu na zbudowanie naprawdę dobrej i użytecznej strategii analizy danych. Pośpiech w inwestowaniu w analitykę danych bez jasnego celu lub zrozumienia, w jaki sposób firma musi się fundamentalnie zmienić, aby uchwycić pożądaną wartość biznesową, może przynieść odwrotny skutek. Istnieje wiele sposobów, w jakie Twoja inwestycja w analitykę danych może się nie powieść. Niektóre z tych problemów lub pułapek są łatwiejsze do uniknięcia niż inne. Niektórych nie da się uniknąć, ale można nimi zarządzać poprzez większą świadomość tego, jak do nich podejść.

Zapisywanie się w pułapkach Data Science

Częścią pogodzenia się ze złożonością rozwiązania jest uświadomienie sobie, że – pomimo naszych najlepszych myśli i intencji – wciąż pociągają nas proste rozwiązania złożonych problemów. Rozwiązania data science nie są wyjątkiem od tej reguły. Opracowując strategię analizy danych, na pewno spotkasz się z wieloma „rozsądnymi" twierdzeniami, które w rzeczywistości są dalekie od rozsądnych i mogą potencjalnie zagrozić powodzeniu Twojej inicjatywy w zakresie analizy danych. (Odnoszę się do tych „rozsądnych" twierdzeń jako „pułapek", ponieważ jeśli pozwolisz im się ustabilizować, ty i twoje inicjatywy związane z nauką danych wpadniesz w otchłań bez żadnej strategii wyjścia.) Musisz stale pracować wbrew założeniu, że „ najprostszym rozwiązaniem jest best one", podkreślając raz po raz, że opanowanie złożoności to naprawdę jedyny sposób, w jaki można zapewnić powodzenie każdej strategii analizy danych. Niektórych wyzwań można uniknąć, podczas gdy inne są nieuniknione i należy nimi zarządzać. Aby pomóc Ci w Twoich wysiłkach, przeprowadzę Cię przez przegląd niektórych typowych pułapek, których musisz unikać (oraz wyjaśnienie, dlaczego), aby Twoja strategia analizy danych odniosła sukces. Wiele można wygrać, jeśli potrafisz skoncentrować swoje wysiłki na efektywnym radzeniu sobie z wyzwaniami, których nie możesz uniknąć.

Wierząc, że wszystkie dane są potrzebne

Chłonność danych to usterka powszechna w wielu firmach. Poświęcają dużo czasu i pieniędzy na inwestycje w komponenty infrastruktury, dzięki czemu mogą gromadzić i przechowywać wszystkie dane dostępne w określonym segmencie istotnym dla ich działalności. Dane są pozyskiwane bez strategicznego myślenia o tym, co jest rzeczywiście potrzebne i kiedy. Co się stanie, gdy wprowadzisz wszystkie dane? Czas i pieniądze są wydawane na pozyskiwanie danych, sortowanie ich i upewnianie się, że infrastruktura poradzi sobie z ogromną ilością wprowadzanych danych. Oznacza to, że nie ma już nic na inwestowanie w zadanie wykorzystania danych. Czasami dochodzi nawet do punktu, w którym zarządzanie danymi poświęca się tak wiele wysiłku, że prawie nie pozostaje już czasu na spostrzeżenia, które miały zapewnić dane. Co więcej, spostrzeżenia wynikające z danych często nigdy nie są wprowadzane w życie – skupiamy się gdzie indziej, tkwiąc w zarządzaniu przeciążeniem danych napływających do firmy.

Myślenie, że inwestycja w jezioro danych rozwiąże wszystkie Twoje problemy

Wiele firm poświęciło znaczną ilość czasu i pieniędzy na inwestycje w jeziora danych, wierząc, że poprzez zastąpienie rozproszonych repozytoriów danych (zazwyczaj rozproszonych w różnych aplikacjach i tradycyjnych systemach baz danych) nowym i wspólnym repozytorium danych (zwykle w chmurze), wszystko problemy zostały rozwiązane. Musisz jednak uważać, aby nie postrzegać jeziora danych jako srebrnej kuli — tej części infrastruktury, która rozwiąże każdy problem. Należy pamiętać, że jezioro danych powinno być postrzegane jako tymczasowy punkt przechowywania danych, a nie stały. Pamiętaj, że dodaje wartości tylko tak długo, jak dane w nim przechowywane są wykorzystywane. Oczywiście firma może mieć inne powody przechowywania danych – np. przepisy, które wymagają przechowywania danych przez określony czas. Należy jednak pamiętać, że jezioro danych powinno być postrzegane przede wszystkim jako warstwa w infrastrukturze, która powinna koncentrować się na umożliwieniu bezpiecznego i wydajnego wykorzystania danych przez następną warstwę. Unikaj myślenia o jeziorze danych jako „magazynu", w którym wrzucasz wszystkie zebrane dane i zamykasz drzwi do użytku tylko przez nieznaną osobę, w nieznanym celu, później w niejasnym momencie w przyszłości. (Nazywa się to jeziorem danych, a nie otchłanią danych.) Zamiast tego myśl o przyszłości; Niezwykle ważne jest, abyś jasno określił w swojej strategii danych, które dane będą przechowywane gdzie, w jakim celu i z jakim priorytetem. Musisz także pomyśleć o tym, jak długo chcesz mieć okresy przechowywania dla każdego typu danych, w oparciu o to, co chcesz osiągnąć za pomocą danych. Innym ważnym aspektem, który należy wziąć pod uwagę strategicznie, są koszty związane z przechowywaniem danych. Jeśli regularnie lub w czasie rzeczywistym zbierasz ogromne ilości danych, co oznacza, że przewidujesz stały napływ nowych danych, który z czasem zwiększy całkowity wolumen danych, możesz spodziewać się wykładniczego wzrostu kosztów przechowywania danych w ciągu krótko- i długoterminowe. Zanim dane trafią do Data Lake, musisz również zastanowić się nad strukturą Data Lake, aby móc szybko i wydajnie znajdować dane. Musisz oddzielić dane wrażliwe i niewrażliwe, a także dane, które posiadasz, od danych, których sam nie jesteś właścicielem, ale masz prawo do korzystania z nich. Bardzo ważne jest również wewnętrzne przemyślenie praw dostępu do danych z perspektywy zarządzania danymi. Może nie każdy powinien mieć dostęp do wszystkiego? Po prostu bądź ostrożny w tym względzie. Nie przesadzaj z ograniczaniem dostępu do danych w Twojej firmie. Blokowanie danych tylko po to, aby być po bezpiecznej stronie, nie jest dobrym pomysłem, ponieważ zmniejszy wydajność wykorzystania jeziora danych. Ogranicz tylko to, co jest absolutnie konieczne z punktu widzenia prawa lub polityki firmy w odniesieniu do prywatności danych, zastrzeżonych danych klientów, danych finansowych lub innych danych wrażliwych. Bez tej podstawowej struktury jeziora danych z kategoryzacją danych, tagowaniem, zdefiniowanymi okresami

przechowywania, prawami dostępu itd. istnieje ryzyko, że masz do dyspozycji mnóstwo danych, ale nie możesz ich efektywnie używać, ponieważ są one zgubione lub zablokowane w jeziorze .

Skupienie się na sztucznej inteligencji, gdy wystarczy analityka

Ambicja pozostawania w zgodzie z ewolucją branży za wszelką cenę to kolejna typowa pułapka, którą można znaleźć w coraz większej liczbie firm. Wynika to z faktu, że firmy chcą być na bieżąco z najnowszą ewolucją technologii na rynku, ale nie mają prawdziwego zrozumienia, co to właściwie oznacza. Jeśli chodzi o sztuczną inteligencję (AI), nie ma pewności, czym jest sztuczna inteligencja i co może zrobić, ale jednocześnie nie docenia się tego, co może zrobić analityka bez dodawania złożoności sztucznej inteligencji. Mówienie, że sztuczna inteligencja jest nadmiernie przereklamowana, nie oznacza, że sztuczna inteligencja nie może być istotna. Wręcz przeciwnie, sztuczna inteligencja najprawdopodobniej może usprawnić większość firm pod wieloma względami. Nie powinieneś jednak próbować rozwiązywać prostego problemu za pomocą złożonej technologii, takiej jak sztuczna inteligencja, jeśli w rzeczywistości można go rozwiązać za pomocą analityki. Analytics pomaga eksplorować dane przy użyciu różnych technik, umożliwiając znajdowanie zależności i korelacji, a także budowanie modeli do prognozowania lub przewidywania określonego wyniku lub zachowania. Analytics nie spełnia oczekiwań, gdy występuje jeden z tych czynników:

* Problemy są zbyt złożone, aby ludzie mogli zrozumieć i zaprojektować zoptymalizowane rozwiązanie.

* Środowisko danych, w którym algorytm musi działać, jest dynamiczne i stale się zmienia.

W takich sytuacjach potrzebujesz czegoś innego. Model analityczny wprowadzony do produkcji ma ustaloną konstrukcję; jego zachowanie jest statyczne i nie może się dostosowywać ani zmieniać, co oznacza, że pozostanie takie samo w czasie, nawet jeśli zmienią się dane i warunki. Zamiast natychmiast wskoczyć na modę AI, poświęć trochę czasu na myślenie przez jaki rodzaj środowiska kierujesz swoje rozwiązania i w jakim kontekście potrzebujesz tych rozwiązań do działania. Zadaj sobie pytanie, które z nich są bardziej statyczne, a tym samym z większym prawdopodobieństwem pozostaną takie same w czasie pod względem danych i zachowania, a które stale się zmieniają. Zrozumienie tego na pewnym poziomie szczegółowości pomoże ci uzyskać lepszy przegląd tego, jakiego podejścia użyć do danego problemu.

Wiara w podejście 1-narzędziowe

Wiele firm jest zdania, że zharmonizowane podejście narzędziowe oferuje najbardziej wydajne środowisko IT. I być może dzieje się tak wiele razy, zwłaszcza gdy chcesz dążyć do ujednoliconych sposobów pracy w firmie, harmonizacji wprowadzania danych i tak dalej. Ale jeśli chodzi o analitykę danych, musisz rozważyć, które części muszą być zgodne, a które są bardziej wydajne, jeśli pozostają różnorodne i elastyczne. W ogólnym podejściu należy dążyć do jak największego wyrównania w podstawowych warstwach przechwytywania, przechowywania i zarządzania danymi. Ale kiedy zbliżasz się do wyższych warstw analizowania i komunikowania spostrzeżeń i decyzji, musisz zapewnić znacznie wyższy poziom swobody dla analityków danych, analityków biznesowych i innych zainteresowanych stron. Przystępując do eksploracji i analizy danych oraz opracowywania algorytmów, potrzebujesz różnych technik i narzędzi dostępnych dla swoich zespołów. To samo dotyczy komunikowania się lub wykorzystywania wyników z analizy: Potrzebujesz podejść dostosowanych do potrzeb, które najlepiej pasują do informacji, które chcesz przekazać. Wymuszanie tego w jednym i tym samym środowisku raczej utrudni, niż pobudzi innowacyjność i zdecydowanie ograniczy wpływ nauki o danych na ogólną działalność firmy. . Aby strategia analizy danych odniosła sukces, musi być zintegrowana ze wszystkimi niezbędnymi aspektami Twojej firmy. A do tego potrzebujesz różnorodnych narzędzi i aplikacji do różnych celów. Liderzy wielu firm mają tendencję do patrzenia na konfigurację nauki o danych z

perspektywy kosztów i często uważają, że główny koszt środowiska jest związany z warstwą aplikacji, a nie z częścią infrastruktury, która umożliwia korzystanie z danych — przechwytywanie, przechowywanie i zarządzanie danymi , innymi słowy. Prawidłowe wyrównanie i zoptymalizowanie podstawowych warstw w infrastrukturze nie tylko zapewni kontrolę kosztów inwestycji w infrastrukturę do analizy danych, ale także zapewni zespołom zajmującym się analizą danych większą swobodę w warstwach na górze. Dzięki takiemu podejściu masz większą szansę na ogólne zwiększenie produktywności w zakresie analizy danych.

Inwestowanie tylko w określonych obszarach

Liderzy większych firm często myślą, że można wybrać jeden lub dwa obszary inwestycji w analitykę danych, zamiast decydować się na pełne wdrożenie w całej firmie. Jest to zrozumiałe, ponieważ taka implementacja jest nie tylko kosztowna, ale także fundamentalnie zmienia sposób podejścia i realizacji zadań. Oczywiście zmiana na tak masową skalę jest postrzegana jako poważne ryzyko z perspektywy całej firmy.

Aby naprawdę skorzystać z inwestycji w analitykę danych, musisz podejść do niej z perspektywy kompleksowej. Tak długo, jak poświęcasz czas na przemyślenie swojej inwestycji z długoterminowej perspektywy, nie tylko jest możliwe, ale także wskazane, aby zacząć od małego, a następnie rozwijać się z czasem, obszar biznesowy po obszarze biznesowym. Jednak aby to osiągnąć, potrzebujesz planu włączenia ludzi biznesu w inwestycje w naukę danych. Wszystkie części muszą z czasem ulec transformacji – a ta transformacja może sięgać znacznie dalej, niż myślisz. Jeśli Twoja firma jest duża, być może będziesz musiał nawet rozważyć zmianę relacji z poddostawcami i dostawcami. Jeśli Twoja firma będzie oparta na danych i wartościach we wszystkich aspektach, czy możesz naprawdę współpracować z poddostawcą, który kieruje się kosztami? Planując małe kroki w kierunku posiadania firmy, która jest w pełni skoncentrowana na nauce danych, nie można liczyć na to, że takie podejście będzie najlepsze z punktu widzenia kosztów. Zamiast tego jest bardziej prawdopodobne, że dopóki nauka o danych nie zostanie wdrożona jako siła napędowa w całej firmie, zobaczysz tylko niewielkie korzyści z perspektywy całej firmy. Pamiętaj, że kierowanie tylko częściami organizacji na dane może nawet w krótkim okresie zwiększyć ogólne koszty, ponieważ oznacza to, że musisz równolegle utrzymywać dwa lub więcej typów konfiguracji (infrastruktura, procesy, kompetencje itd.).

Wykorzystanie infrastruktury do raportowania, a nie eksploracji

Powszechny problem skoncentrowania się na raportach jest zwykle związany z sytuacją, w której najwyższe kierownictwo ma błędne wyobrażenie o tym, co data science może wnieść do firmy. Taka sytuacja zwykle ma miejsce, ponieważ niektórzy liderzy firm uważają, że głównym celem nauki o danych jest uzyskanie zestawu odpowiedzi na pewne predefiniowane pytania zadawane przez kierownictwo. Odpowiedzi na te konkretne pytania powinny zatem być głównym motorem wdrażania nauki o danych, a zatem powinny skutkować raportem zwrotnym dla kierownictwa. Być może zadajesz sobie pytanie, co jest złego w próbie spełnienia żądań pochodzących z korporacji – czy nie jest punktem wyjścia wszystkich analiz zestaw pytań biznesowych, na które chcesz uzyskać odpowiedzi? W pewnym sensie to prawda. Jednak równie ważne jest, aby pamiętać, że pytania, które zadajesz, mogą nie być właściwymi do zadawania. Dlaczego? Po prostu dlatego, że ten z góry ustalony zestaw pytań opiera się na aktualnym zrozumieniu Twojej firmy, rynku i bazy klientów. Jeśli Twoja firma opiera się głównie na doświadczeniu, a nie na danych, pytania mogą być poprawne – lub nie. Po prostu nie wiesz, czy podchodzisz do określonego problemu lub okazji pod niewłaściwym kątem. Traktuj swoją inwestycję w naukę danych jako okazję, aby Twoja firma opierała się na danych, spostrzeżeniach i faktach, które pomogą Ci prawidłowo poprowadzić Cię w społeczeństwie opartym na danych. Podobnie jak w firmie Husqvarna, która zajmuje się produktami zasilającymi przeznaczonymi do użytku na zewnątrz, firma

zaczęła udostępniać łączność dla swoich pilarek łańcuchowych. Firma robi to, aby zebrać dane o tym, w jaki sposób są używane lub nie są używane podczas ścinania drzew, aby móc lepiej zrozumieć swój biznes. Po prostu zbadanie danych pod kątem wzorców lub anomalii, które mogą wskazywać na nowe (i być może nieoczekiwane) pytania, które warto zadać, to dobry sposób na rozpoczęcie.

Niedocenianie zapotrzebowania na wykwalifikowanych analityków danych

Zostanie naukowcem danych jest kompetencją nabytą; można na nią zarobić, innymi słowy, za pomocą książek oraz kursów szkoleniowych i warsztatów. Bycie doświadczonym analitykiem danych wymaga jednak czasu i pewnych umiejętności, które nie są tak łatwe do zdobycia. Ważne jest, aby szanować różnicę między podstawowym analitykiem danych a doświadczonym. Równie ważne jest, aby zdać sobie sprawę, że seniorzy są trudni do zdobycia, więc jeśli masz ich w swojej firmie, upewnij się, że się ich trzymasz. Starszy specjalista ds. danych w przestrzeni AI to ktoś, kto pracował w tej dziedzinie od pięciu do dziesięciu lat, zna kilka języków programowania, ale co najważniejsze, ma duże umiejętności w korzystaniu z różnych technik ML/AI podczas budowania algorytmów. Aby być postrzeganym jako senior, obejmuje to również doświadczenie w opracowywaniu i wdrażaniu algorytmów opartych na różnych przypadkach użycia i w różnych typach środowisk docelowych. . Jednak kluczem do stworzenia skutecznych zespołów zajmujących się analizą danych nie jest pozyskanie jak największej liczby starszych analityków danych. W rzeczywistości lepiej jest mieć mniej starszych i rozdzielić ich na wiele zespołów, co pozwoli im działać jako mentorzy dla młodszych analityków danych, a tym samym w lepszy sposób przyczynić się do ogólnej dojrzałości nauki o danych w firmie. Chociaż zatrudnienie starszych analityków danych jest drogie, warto pomyśleć o nich pod kątem wkładu, jakiego można od nich oczekiwać. Przy wsparciu ekspertów dziedzinowych starsi specjaliści ds. danych mogą pracować w dowolnej dziedzinie dyscypliny i, biorąc pod uwagę odpowiednie warunki wstępne w zakresie danych i wydajnej infrastruktury, pomagają w podejściu do praktycznie każdego problemu lub możliwości w wydajny i innowacyjny sposób.

Poruszanie się po złożoności

Uzbrojenie się w przekonujące argumenty, mające na celu przeciwstawienie się zwolennikom filozofii „proste jest lepsze” w Twojej firmie, jest dobrym punktem wyjścia. Rozpoznanie wszelkich wyzwań, które mogą pojawić się na drodze Twojej firmy do pełnego wykorzystania danych, ma kluczowe znaczenie, ale sama świadomość wyzwań nie usuwa ich automatycznie. Wymaga to nie tylko stałej świadomości konieczności niemyślenia o rzeczach w niewłaściwy sposób, ale także strategicznego nastawienia - i planu - aby poruszać się po potencjalnych problemach, gdy się pojawiają. Warto poświęcić czas na zapoznanie się z różnymi scenariuszami i rozwiązaniami dla nich. Kiedy się pojawią (i na pewno się pojawią), będziesz już miał pewien poziom zrozumienia, jak sobie z nimi radzić. Co więcej, biorąc pod uwagę to, co wiesz, możesz działać proaktywnie, aby upewnić się, że nigdy nie znajdziesz się w jednej z tych mniej niż korzystnych sytuacji dla Twojej firmy. Sporządź listę zidentyfikowanych zagrożeń i proponowany plan łagodzenia skutków dla wszystkich scenariuszy i dodaj je do strategii firmy w zakresie analizy danych, aby mieć uzgodniony pogląd na to, co robić – a czego nie – kiedy coś się wydarzy.

Radzenie sobie z trudnymi wyzwaniami

Omówimy szereg skomplikowanych wyzwań, których trudno uniknąć, a które będą wymagały odpowiedniego zestawu taktyk, aby skutecznie sobie z nimi radzić. Mówiąc dokładniej, pokażę Ci, co musisz zrobić, aby podejmować właściwe decyzje, jeśli chodzi o skuteczne i spójne pozyskiwanie danych i zarządzanie nimi, konfigurowanie środowiska nauki o danych, zarządzanie ograniczeniami prawnymi związanymi z danymi i algorytmami, których potrzebujesz dla Twojej firmy, a także przygotowanie się na gwałtowne zmiany w obszarze data science jako całości, które z pewnością nadejdą.

Pobieranie danych stamtąd do tego miejsca

Kiedy firma decyduje się rozpocząć podróż, aby stać się napędzaną danymi, koncentruje się naturalnie na samych danych, co nieuchronnie prowadzi do większej świadomości rzeczywistej różnorodności danych potrzebnych do uzyskania pełnej proaktywnej i opartej na danych kontroli nad bieżącą działalnością firmy . Co więcej, firmy szybko zdają sobie sprawę, że aby wyjść poza to, co jest obecnie możliwe, zbiory danych muszą stać się jeszcze bardziej zróżnicowane. W tym momencie wiele firm zaczyna zdawać sobie sprawę, że dane, które mają zasadnicze znaczenie dla rzeczywistego wykorzystania danych, mogą w rzeczywistości należeć do kogoś innego lub znajdować się w innym kraju, z innymi przepisami dotyczącymi danych. W tej sekcji wyjaśniono, jak strategicznie podejść do takich praktycznych wyzwań w ramach pozyskiwania danych.

Obsługa zależności na danych należących do innych

Radzenie sobie z zastrzeżonymi danymi jest nieuniknionym, ale możliwym do opanowania wyzwaniem, przed którym stoi każda firma dążąca do pełnego wykorzystania danych. Zwykle dzieje się tak, że zidentyfikowałeś i dokładnie określiłeś wszystkie potrzebne dane w swojej strategii danych, a kiedy zaczniesz zastanawiać się, jak strategicznie podejść do przechwytywania danych, zdajesz sobie sprawę, że masz problem z własnością danych. Jeśli korzystasz tylko z danych generowanych z Twojego wewnętrznego środowiska IT, masz oczywiście mniejszy problem. Jeśli jednak tak jest, to prawdopodobnie Twoja firma nie jest tak naprawdę oparta na danych we właściwym tego słowa znaczeniu. Firma oparta na danych wyjaśnia, w jaki sposób wykorzystywane są jej produkty i/lub usługi oraz jak działa w prawdziwym życiu ustawienia, nie tylko w środowisku laboratoryjnym. I za każdym razem, gdy zaczynasz korzystać z danych generowanych przez życie w prawdziwym świecie, napotykasz problem z własnością danych.

O jakich danych mówię? Przede wszystkim dotyczy to danych należących do Twoich klientów, ale może również obejmować dane należące do klientów Twoich klientów, w zależności od tego, w jakiej firmie się znajdujesz. Musisz poświęcić trochę czasu, aby naprawdę zrozumieć szczegółowy kontekst danych, które potrzeba. Może odnosić się do kwestii prywatności danych, ale nie musi. Może po prostu być tak, że dane, których potrzebujesz, aby lepiej zrozumieć wyniki lub potencjał Twojej firmy, należą do kogoś innego.

Nie zniechęcaj się, jeśli chodzi o kwestie własności. Większość sytuacji można rozwiązać z prawnego punktu widzenia, jeśli chcesz otwarcie zająć się nimi z właścicielami danych, wyjaśniając, dlaczego potrzebujesz danych i jak będziesz je traktować po ich posiadaniu. Wszystko sprowadza się do zdobycia zaufania co do tego, w jaki sposób i w jakim celu dane będą wykorzystywane. (Nie zaszkodzi również sprecyzować, w jaki sposób Twoja praca może, jeśli to możliwe, wnieść wkład z powrotem do właścicieli danych.) Pod koniec dnia musisz mieć absolutną pewność, że rozumiesz (i przestrzegasz) ograniczenia prawne mające zastosowanie do każdego rodzaju danych, z których zamierzasz korzystać. Korzystanie z danych musi być również regulowane w drodze umowy ze stroną będącą właścicielem

danych, w tym praw, jakie Twoja firma ma w związku z dostępem do danych, przechowywaniem i użytkowaniem w czasie.

Prawa i przepisy mają w zwyczaju zmieniać się w czasie. Ostatnio obserwuje się tendencję do dalszego zwiększania ograniczeń w celu ochrony prawa jednostki do własnych danych. Jednym z ostatnich przykładów jest dość restrykcyjne Ogólne Rozporządzenie o Ochronie Danych i Rozporządzenia (RODO) przyjęte przez Unię Europejską (UE), które weszło w życie w maju 2018 r. Biorąc pod uwagę niedawne informacje o niewłaściwym wykorzystywaniu danych przez podmioty takie jak Cambridge Analytica i Facebook, USA a Kanada zdecydowanie przygląda się przepisom podobnym do unijnego RODO. Wszystko, co pomaga chronić prawo jednostki do prywatności, jest najlepsze, ale pamiętaj tylko, że sposób, w jaki dziś traktujesz przepisy dotyczące prywatności, najprawdopodobniej będzie zupełnie inny w najbliższej przyszłości. Dlatego należy strategicznie i proaktywnie przemyśleć konfigurację infrastruktury i potrzeby dotyczące danych, aby zapewnić uwzględnienie tego typu ograniczeń w bieżącym i zmieniającym się środowisku nauki o danych.

Zarządzanie transferem danych i obliczeniami ponad granicami kraju

Jeśli Twoja firma ma oddziały w wielu różnych krajach lub prowadzi działalność (a tym samym ma wielu klientów) w wielu krajach, jednym z głównych wyzwań, z jakimi możesz się zmierzyć, jest zarządzanie danymi, które muszą przekraczać granice międzynarodowe. Musisz dokładnie rozważyć szereg różnych aspektów układanki danych, jeśli Twoja firma ma komponent międzynarodowy. Oto lista głównych obaw:

* Legalność: Ograniczenia prawne dotyczące przenoszenia danych przez granice to kwestia, której firma musi przestrzegać. Prawa i przepisy różnią się w zależności od kraju, więc możliwe są różne rozwiązania w zależności od kraju, w którym prowadzisz działalność. Ograniczenia są również różne w zależności od rodzaju danych, które wyprowadzasz z kraju. Dane zawierające dane osobowe są zwykle znacznie trudniejsze do przeniesienia niż dane niewrażliwe. Łamanie przepisów związanych z przesyłaniem danych może być dość kosztowne i może poważnie wpłynąć na markę firmy, jeśli zostanie stwierdzone, że naruszyłeś zaufanie klientów.

* Podejście do przesyłania danych: odnosi się do tego, w jaki sposób faktycznie wykonujesz transfer danych. Jest to zazwyczaj dość kosztowne, a także różni się w zależności od kraju. W zależności od ilości przesyłanych danych i częstotliwości przesyłania danych, możesz wynająć przestrzeń w istniejącej infrastrukturze łączności i łączach danych lub - jeśli nie możesz spełnić swoich wymagań dotyczących takich aspektów jak pojemność, bezpieczeństwo lub wyłączność - zainwestować w własne linki.

* Możliwości lokalnych obliczeń i przechowywania: Jeśli możesz przechowywać dane i przeprowadzać analizę w kraju, w którym dane zostały przechwycone, możesz obniżyć koszty i zwiększyć szybkość dostawy. Jednak, aby ta konfiguracja działała wydajnie, musisz odpowiednio przemyśleć, jak będzie wyglądać Twoja rozproszona architektura obliczeniowa. Co zostanie zrobione gdzie? a gdzie będą przechowywane np. dane źródłowe? Czy będzie centralny punkt przechowywania danych i globalnej analizy, czy tylko rozproszone konfiguracje? To, jak odpowiesz na te pytania, zależy w dużej mierze od rodzaju prowadzonej działalności i tego, jak wygląda konfiguracja w różnych krajach.

Zarządzanie spójnością danych w całym środowisku nauki o danych

Zapewnienie spójności danych w różnych częściach środowiska nauki o danych może wydawać się prostym zadaniem, ale jest znacznie trudniejsze, niż się wydaje. Po pierwsze, ten obszar wydaje się być bardziej złożony niż powinien, pochłaniając więcej czasu i zasobów niż pierwotnie szacowano. Potrzeba spójności obejmuje takie aspekty, jak zarządzanie danymi i formaty danych, ale także spójne

etykietowanie danych - na przykład przy użyciu identyfikatorów klientów z wielu różnych źródeł, aby umożliwić korelację różnych typów danych powiązanych z tym samym klientem. Wyzwanie polega na tym, że istnieje wbudowana sprzeczność w zakresie infrastruktury między umożliwieniem korzystania ze specjalnych narzędzi, aby umożliwić naukowcom i inżynierom danych bycie innowacyjnym i wydajnym, a jednocześnie zapewniającym spójność danych. Dzieje się tak, ponieważ wyspecjalizowane narzędzia są zoptymalizowane pod kątem rozwiązywania określonych problemów, ale albo nie utrzymują spójnego formatu, albo nie współpracują dobrze z innymi narzędziami potrzebnymi w przepływie end-to-end. Zoptymalizowane, wyspecjalizowane narzędzia do uczenia maszynowego po prostu nie są dobre w graniu razem z innymi, podobnymi wyspecjalizowanymi narzędziami, które rozwiązują porównywalne lub inne sąsiadujące problemy. Ale czy to naprawdę takie złe? No cóż, może to prowadzić do prawdziwych problemów, w zależności od tego, jak wiele swobody w realizacji architektonicznej i między zespołami zostanie dopuszczone. Oto kilka przykładów problemów, które mogą wynikać z braku spójności w środowisku AI:

* Rozwiązania ad hoc: każda sprawa jest traktowana jako odosobniony problem, który musi zostać rozwiązany w tej chwili, aby zespół mógł iść do przodu. Wynik? Bez długoterminowego rozwiązania i bez uczenia się między zespołami.

* Zwiększony koszt: gdy musisz powielić możliwości narzędzia, aby poradzić sobie z brakiem spójności, lub gdy musisz wbudować możliwości w zakupione narzędzia, aby zapewnić tylko podstawową spójność, koszty te sumują się.

* Nie działa kompleksowo: niespójności mogą wystąpić, gdy infrastruktura jest wdrażana u kilku dostawców chmury, co utrudnia lub uniemożliwia przesyłanie danych i zachowanie spójności danych w różnych środowiskach zwirtualizowanych.

Ponieważ kierownictwo firmy nie może wymusić i może nie chcieć wymusić spójności danych w całej organizacji jako zasad firmy, musi użyć innych środków, aby zachować spójność danych od początku do końca. Jednym ze sposobów jest upewnienie się, że wszystkie zespoły przestrzegają właściwych i odpowiednich wytycznych dotyczących oceny i zakupu nowych narzędzi, które zawierają konkretne dyrektywy związane z spójnością danych. Wyraźnie motywujące, dlaczego jest to kluczem do pomyślnej realizacji strategii analizy danych. Ważne jest również, aby zastanowić się, jakie limity są potrzebne dla każdej firmy, w zależności od rodzaju działalności, jej celów i tak dalej. Trzymaj się linii, jeśli chodzi o spójność danych: w przeciwnym razie możesz skończyć z kłopotliwą i kosztowną implementacją nauki o danych, daleko odbiegającą od produktywnego środowiska nauki o danych, na które liczyłeś.

Zapewnienie wyjaśnialności w AI

Wyjaśnialna sztuczna inteligencja (XAI), zwana również przezroczystą sztuczną inteligencją, obejmuje zdolność do wyjaśnienia, w jaki sposób algorytm osiągnął konkretny wgląd lub wniosek, który skutkuje podjęciem określonej decyzji o podjęciu działania. Chociaż jest to ważny aspekt, który należy wziąć pod uwagę w ramach ewolucji sztucznej inteligencji, nie jest to łatwe do rozwiązania technicznie, zwłaszcza jeśli sztuczna inteligencja działa w czasie rzeczywistym, a tym samym wykorzystuje dane strumieniowe, które nie zostały zapisane. Aby przybliżyć ten punkt do domu, wyobraź sobie, jeśli chcesz, że nie możesz tego wyjaśnić twojemu klientowi , dlaczego maszyna podjęła określoną decyzję - decyzję, której nie podjąłbyś na podstawie własnego doświadczenia. Co wtedy mówisz klientowi? Zajmowanie się sztuczną inteligencją, którą można wyjaśnić, staje się coraz ważniejsze z punktu widzenia naszej ludzkiej zdolności do lepszego zrozumienia, dlaczego i jak sztuczna inteligencja działa w określony sposób. Innymi słowy, co można zrozumieć, badając, w jaki sposób maszyna uczy się, przetwarzając te ogromne ilości danych z wielu wymiarów, szukając określonych wzorców lub odchyleń? Co takiego

maszyna wykrywa i rozumie, że przeoczyłeś lub zinterpretowałeś inaczej lub po prostu nie byłeś w stanie wykryć? Jakie wnioski można z tego wyciągnąć? Z etycznego punktu widzenia wyjaśnialność AI będzie jeszcze ważniejsza, gdy analitycy danych zaczną budować bardziej zaawansowaną sztuczną inteligencję, w której działa wiele różnych algorytmów. Będzie to klucz do dokładnego zrozumienia, co maszyny interpretują, a także jak przebiega proces podejmowania decyzji przez maszynę. Znajomość tych informacji ma kluczowe znaczenie dla utrzymania się na szczycie ram polityki niezbędnych do ustalenia granic tego, co maszyna powinna, a czego nie powinna robić, a także tego, w jaki sposób te polityki należy rozszerzyć lub być może ograniczyć w przyszłości. Z jednej strony z czysto egzystencjalnej perspektywy, az drugiej z potrzeby utrzymania przez ludzi kontroli nad inteligentnymi maszynami, które są budowane, nie można po prostu postrzegać sztucznej inteligencji jako czarnej skrzynki. (Wyzwanie związane z czarną skrzynką w sztucznej inteligencji odnosi się do potrzeby zapewnienia, że gdy algorytm podejmuje decyzję w oparciu o techniki zastosowane do trenowania algorytmu, proces podejmowania decyzji musi być przejrzysty dla ludzi. Przejrzystość algorytmu jest możliwa, gdy wiele z bardziej podstawowych technik uczenia maszynowego – na przykład uczenie nadzorowane – jest używanych, ale jak dotąd nikt nie znalazł jeszcze sposobu na uzyskanie przejrzystości, jeśli chodzi o algorytmy oparte na technikach głębokiego uczenia. wyjaśnij, dlaczego podjęto określoną decyzję, gdy coś poszło nie tak. Właściwym przykładem jest autonomiczny samochód, w którym działa kilka algorytmów, które współpracują ze sobą i (miejmy nadzieję) przestrzegają wstępnie zdefiniowanych zasad postępowania w określonych okolicznościach. Wszystko działa. zgodnie z planem, ale wtedy następuje zupełnie nieznane i nieoczekiwane zdarzenie i samochód podejmuje nieoczekiwaną akcję, która powoduje wypadek.W takich sytuacjach ludzie na ogół naturalnie oczekiwaliby, że nie Będzie jakiś sposób na wydobycie informacji z autonomicznego samochodu o tym, dlaczego podjęto tę konkretną decyzję - stąd oczekują wyjaśnień w sztucznej inteligencji.

Poza technicznymi, etycznymi i egzystencjalnymi przyczynami zapewnienia wyjaśnialności sztucznej inteligencji, istnieje teraz również powód prawny. Ogólne rozporządzenie o ochronie danych (RODO) UE zawiera klauzulę, która wymaga algorytmicznej interpretacji. W tej chwili te wymagania nie są zbyt surowe, ale z czasem prawdopodobnie zmieni się to dramatycznie. Żądanie RODO wymaga teraz możliwości wyjaśnienia działania algorytmu na podstawie następujących pytań:

* Jakie dane są używane?

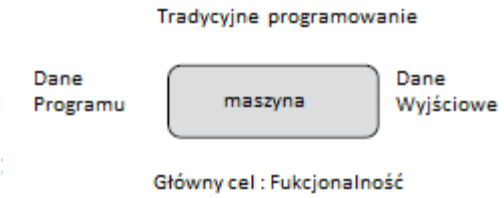
* Jaka logika jest używana w algorytmie?

* Jaki proces jest używany?

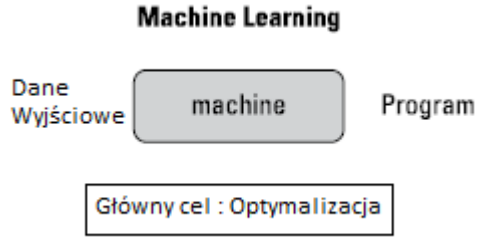
* Jaki wpływ ma decyzja podjęta przez algorytm?

Radzenie sobie z różnicą między uczeniem maszynowym a tradycyjnym programowaniem oprogramowania.

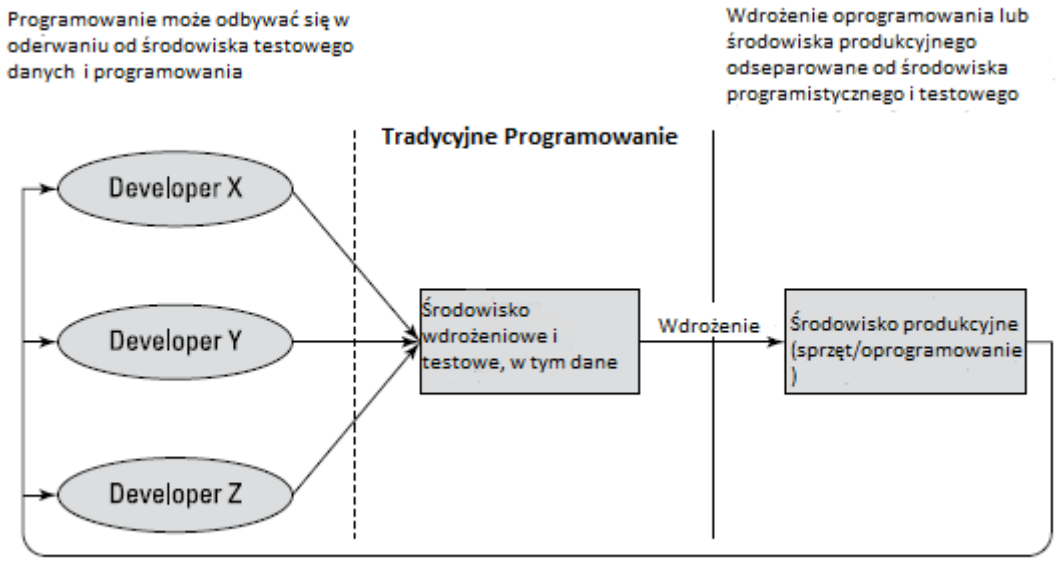
Jest dość dobrze ugruntowane i powszechnie przyjęte w branży oprogramowania, jaka jest rzeczywista różnica między tradycyjnym programowaniem a uczeniem maszynowym. Jednak jeśli chodzi o to, jak należy poradzić sobie z tą różnicą, nie ma zgody. Biorąc pod uwagę ten podział, chcę poświęcić czas na wyjaśnienie, co należy wziąć pod uwagę, jeśli chodzi o podejścia do wdrażania, a także jak radzić sobie z tymi różnymi punktami widzenia pod względem aspektów rozwoju, a także środowiska produkcyjnego. Ale najpierw pozwól, że zacznę od przyjrzenia się, o co chodzi w kłótni. Tradycyjne podejście programistyczne, pokazane na rysunku , wymaga wcześniejszego podjęcia decyzji, jak rozwiązać określony problem za pomocą opracowywanego programu. Głównym celem programisty jest zbudowanie wymaganej funkcjonalności.



Na podstawie danych i programu maszyna wykonuje analizę dokładnie tak, jak chcesz, niezależnie od tego, czy jest to najbardziej zoptymalizowany sposób rozwiązania problemu. Założenie jest takie, że programista (a nie maszyna) najlepiej wie, jak rozwiązać problem. Z drugiej strony, jeśli chodzi o rozwój uczenia maszynowego, punktem wyjścia jest umożliwienie maszynie znalezienia najlepszego rozwiązania, gdy ustalisz granice, których danych użyć i jaki wynik osiągnąć - i nic więcej. Zakłada się, że w tych warunkach maszyna znajdzie najbardziej zoptymalizowany program do rozwiązania problemu.

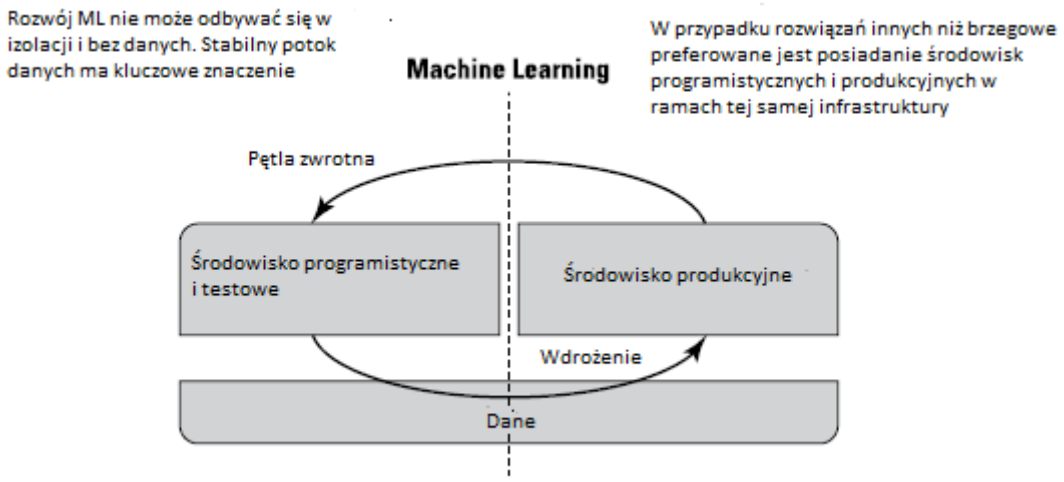


Co więc oznaczają te odrębne podejścia w kontekście środowiska programistycznego i produkcyjnego? Jednym z głównych aspektów do rozważenia jest to, że tradycyjne programowanie obejmuje znacznie bardziej rygorystyczny proces. Jest oparty na regułach i zgodny z predefiniowanymi zasadami projektowania. Z drugiej strony, punkt wyjścia do rozwoju uczenia maszynowego jest znacznie bardziej odkrywczy i otwarty. Jak można się domyślić, będzie to miało dość znaczący wpływ na to, jak należy skonfigurować środowisko programistyczne. Niektóre firmy mają tendencję do bagatelizowania wpływu konfiguracji środowiska programistycznego i tego, jaki będzie to miało wpływ na produktywność analizy danych. Jeśli zaczniesz od tego punktu widzenia, możesz dojść do wniosku, że możesz użyć tej samej (lub podobnej) konfiguracji infrastruktury zarówno dla tradycyjnego środowiska programistycznego, jak i środowiska nauki o danych. Nic nie może być dalsze od prawdy – przyjęcie takiego podejścia oznacza, że stawiasz poważne bariery na drodze do osiągnięcia celu, jakim jest pożyteczna sztuczna inteligencja/uczenie maszynowe. Tradycyjne programowanie jest znacznie bardziej restrykcyjne, jeśli chodzi o to, jakich języków programowania użyć i jakie zasady zastosować do danego zadania. Ma to oczywiście wpływ na sposób konfiguracji zarówno środowiska programistycznego, jak i produkcyjnego. Rysunek przedstawia graficzną reprezentację tego, jak przebiega tradycyjne programowanie.



Pętla sprzężenia zwrotnego

Jak widać po lewej stronie rysunku, tradycyjne programowanie może odbywać się niezależnie od danych oraz środowiska programistycznego i testowego. Nie musi się to odbywać osobno, ale faktem jest, że można to zrobić w izolacji – nawet na laptopie w kawiarni - a następnie zintegrować z innym kodem w środowisku deweloperskim i testowym. W tym momencie do modelu można dodać dane w celu uzyskania pożądanego wyniku. Rysunek pokazuje również, że wdrażanie oprogramowania odbywa się w oddzielnym środowisku (na przykład w produkcie oprogramowania/sprzętu lub podobnym środowisku produkcyjnym) poza środowiskiem programistycznym i testowym. Przechodząc ponownie do podejścia do uczenia maszynowego, musisz pamiętać, że eksploracyjny i uczący się charakter rozwoju uczenia maszynowego wymaga, aby konfiguracja dostępna dla analityków danych była niezwykle elastyczna. Wydajne zarządzanie danymi, łatwy dostęp do danych i różnorodne specjalistyczne narzędzia do uczenia maszynowego muszą być łatwo dostępne. Nikt nie wchodzi w proces z predefiniowanymi pojęciami, której dokładnie techniki uczenia maszynowego użyć, ponieważ wszystko to musi zostać zbadane, a najbardziej zoptymalizowane rozwiązanie może stać się jasne dopiero po rozpoczęciu procesu. Jak pokazuje poniższy rysunek, rozwój uczenia maszynowego nie może odbywać się w izolacji i bez danych. Wszystko zaczyna się i kończy na danych w przepływie rozwoju uczenia maszynowego, ponieważ same dane szkolą model pod kątem zoptymalizowanego projektu. Aby to zadziałało, oczywiście musisz mieć stały przepływ danych, co oznacza, że potrzebujesz stabilnego potoku danych - najlepiej zwirtualizowanego, który zapewnia większą elastyczność infrastruktury w czasie.



W przypadku zwirtualizowanych środowisk produkcyjnych uczenia maszynowego, które nie są zaimplementowane na urządzeniach brzegowych (wewnątrz urządzeń IoT, takich jak telefon komórkowy, samochód, zegarek, lodówka lub inne rodzaje urządzeń, które są połączone i na których może działać algorytm ML), staraj się, aby środowisko programistyczne i produkcyjne było blisko lub w ramach tej samej konfiguracji infrastruktury. Ułatwia to produktywność uczenia maszynowego podczas przechodzenia między rozwojem a produkcją, z szybszymi i wydajniejszymi pętlami sprzężenia zwrotnego w ramach korzyści. Zyskujesz również korzyści w zakresie efektywności kosztowej, gdy nie musisz duplikować infrastruktury, ponieważ obie są oparte na tym samym potoku danych.

Zarządzanie szybką ewolucją technologii AI i brakiem standaryzacji

Technologie AI/ML nieustannie ewoluują i stają się coraz bardziej zaawansowane. Wraz ze wzrostem wydajności obliczeniowej, takie technologie mogą teraz dostosowywać się do działania na mniejszej powierzchni sprzętowej. Te postępy przesuwają analitykę, ML i realizację sztucznej inteligencji również na brzeg, co oznacza, że algorytm ma wsparcie obliczeniowe do działania wewnątrz urządzenia, a nie że urządzenie po prostu dostarcza dane do algorytmu działający zdalnie w chmurze. To dobry trend, ponieważ pozwoli społeczeństwu na szersze wykorzystanie inteligencji maszynowej w środowiskach systemowych oraz miliardach urządzeń mobilnych i innych połączonych podmiotach. Jednak jeden obszar nie nadąża za wszystkimi szybkimi zmianami: standaryzacja ML/AI. Brak standaryzacji nie jest oczywiście czymś, co Ty lub pojedyncza firma możesz rozwiązać, ale ważne jest, aby być świadomym tej sytuacji w ramach swojej strategii analizy danych. I oczywiście pod koniec dnia wszyscy naukowcy zajmujący się danymi mają obowiązek dążyć do większej standaryzacji w uczeniu maszynowym i sztucznej inteligencji. Ale tylko dlatego, że nie istnieje jeszcze żadna oficjalna, międzynarodowa standaryzacja, nie oznacza to, że nie istnieją żadne inicjatywy. Standardy, które są dostępne, opierają się w większości na czymś, co często określa się mianem standaryzacji de facto, wywodzącej się z wpływowych inicjatyw open source koordynowanych przez uniwersytety, takie jak UC-Berkeley (laboratorium AMP i laboratorium RISE) oraz firmy takie jak Google (Google Beam) i AT&T (Acumos). Inny trend, który można wykryć, dotyczy rosnących obaw, jeśli chodzi o dostęp do danych osobowych (i korzystanie z nich) z nieprzejrzystych lub nawet ukrytych powodów. Doprowadziło to do zaostrzenia przepisów w różnych krajach, ale doprowadziło również do bardziej toczących się dyskusji na temat potrzeby zwiększenia regulacji i narzucenia standaryzacji związanych z etyką AI. Miejmy nadzieję, że ten pozytywny trend będzie nadal skłaniał ludzkie społeczeństwo do lepszego wyobrażania sobie – jako grupy - jak powinna wyglądać przyszłość wykorzystania sztucznej inteligencji. Oczywiście ten trend ma swoje minusy. Ponieważ teraz jest tak mało standaryzacji dostępnej, aby oprzeć się na inwestycjach w

naukę danych, musisz wziąć pod uwagę możliwość, że będziesz musiał wprowadzić poważne zmiany w swojej infrastrukturze, gdy nowe standardy w końcu pojawią się w najbliższej przyszłości. Najgorszy scenariusz? Być może będziesz musiał powtórzyć ten proces kilka razy, a nawet całkowicie przemodelować całą infrastrukturę. Moja rada dla ciebie? Nieustannie śledź trendy i wypatruj wszelkich oznak, że nowe przepisy lub regulacje lub inicjatywy standaryzacyjne spływają na dno - zwłaszcza te o otwartym kodzie źródłowym. Przygotuj się na dostosowanie swojego podejścia do analizy danych do nadchodzących zmian.

Zarządzanie zmianą w nauce o danych

Inwestowanie w naukę o danych i podejście oparte na danych oznacza zrozumienie i radzenie sobie ze zmianą, która musi nastąpić. Chociaż nieunikniona transformacja nauki o danych w społeczeństwie może jeszcze nie zaszła w pełni, organizacje wciąż muszą się na to przygotować. Skończył się czas stania z boku i czekania na to, co robią inne firmy. Nadszedł czas działania. Firmy najlepiej przygotowane do zarządzania potrzebną zmianą napędzaną przez analitykę danych w następnej dekadzie będą tymi, które już teraz zaczną się przygotowywać. Nadszedł dzień, w którym firmy zainwestują czas w strategiczne budowanie zrozumienia potrzeb i uchwycenie intencji w strategii analizy danych — nie tylko w jednym obszarze lub funkcji, ale w całej firmie.

Zrozumienie zarządzania zmianą w nauce o danych

W badaniu przeprowadzonym przez PricewaterhouseCoopers (PwC) i Iron Mountain 1800 starszych liderów biznesu w Ameryce Północnej i Europie w firmach średniej wielkości i organizacjach na poziomie przedsiębiorstwa odpowiedziało na ankietę, która wykazała, że tylko niewielki procent firm rzeczywiście rozważa praktyki zarządzania danymi. Badanie wykazało, że chociaż 75 procent liderów biznesu z firm różnej wielkości, lokalizacji i sektorów uważa, że „najlepiej wykorzystują swoje informacje o aktywach", w rzeczywistości tylko niewielka część wydaje się strategicznie podchodzić do tych poważnych zmian we właściwy sposób. Ogólnie rzecz biorąc, aż 43% liderów firm odpowiedziało, że „uzyskują niewielkie namacalne korzyści ze swoich informacji", a 23% „nie czerpie żadnych korzyści", zgodnie z badaniem. Co więc firmy robią źle? Jedną z lekcji, jaką można wyciągnąć z ankiety, jest to, że inwestowanie w technologię, która ma być oparta na danych, to dopiero początek. Aby zapewnić sukces, firmy muszą robić znacznie więcej niż skupiać się na narzędziach potrzebnych do zarządzania danymi. Transformacja nauki o danych dotyczy złożonych i wzajemnie połączonych danych, zarówno małych, jak i dużych zbiorów danych, co wpływa na cały zakres operacji biznesowych i ma wpływ na ludzi, kultury, organizacje, procesy i zestawy umiejętności w nauce o danych. Klejem, który łączy i utrzymuje wszystkie te elementy razem, są ludzie. A kluczem jest zmotywowanie ludzi. Można to osiągnąć na wiele sposobów, ale dobrym sposobem na rozpoczęcie jest wykorzystanie danych do przekazania odpowiednich przykładów i punktów dowodowych w połączeniu z przywództwem firmy. Silne przywództwo napędzające zmianę obejmuje nie tylko wsparcie kierownictwa liniowego, ale jest również w dużym stopniu zależne od silnych liderów i motorów zmian, którzy mogą wzbudzać zaufanie, że zmiana przyniesie rezultaty. Bez tych dedykowanych sterowników zmian w całej firmie nie ma znaczenia, czy masz idealny plan - tego typu całkowicie transformująca zmiana nie nastąpi, a przynajmniej nie w pełnym zakresie. W nauce o danych metody i techniki wykorzystywane do wszystkiego, od wiedzy o tym, jak przechwytywać i przetwarzać dane, po budowanie modeli i uzyskiwanie spostrzeżeń, wciąż ewoluują, tworząc ciągłą potrzebę zarządzania zmianami. Ta zmiana ma również miejsce w takich obszarach, jak praktyki regulacyjne, bezpieczeństwo i prywatność, stale zmieniając podstawę i ramy podejścia do nauki o danych. Aby strategia nauki o danych odniosła sukces, organizacje muszą zrozumieć i zaakceptować fakt, że zestawy umiejętności potrzebne do obsługi różnych aspektów nauki o danych będą się nadal zmieniać. Aby zarządzać tą ciągłą zmianą, musisz mieć otwarty umysł i chcieć wykorzystywać i odkrywać nowe technologie i metodologie, gdy tylko staną się dostępne. W praktyce oznacza to, że jednostki muszą przyjąć sposób myślenia oparty na danych i zaangażowanie w uczenie się przez całe życie jako przedłużenie ich pracy, jeśli kiedykolwiek mają nadzieję zarządzać zmianą. Tylko wtedy, gdy aktywnie wykorzystujesz dane do odkrywania nowych możliwości i rozwiązywania rzeczywistych problemów, możesz uzasadnić inwestycję w naukę o danych. Zdefiniowanie odpowiedniego i odpowiedniego procesu zarządzania zmianą powinno być wspólnym wysiłkiem organizacyjnym, podchodzącym poprzez burzę mózgów i udoskonalanie

pomysłów. Zwykle uzgodnienie, że zmiana jest potrzebna, jest łatwiejsze niż decydowanie, jak zmienić należy podejście.

Zbliżanie się do zmian w nauce o danych

Skuteczne zarządzanie zmianą to wieloetapowy proces, który wymaga znacznych inwestycji czasu i pieniędzy. Mogę ci polecić ogólne podejście do zarządzania zmianą, ale musisz również wziąć pod uwagę kilka specyficznych cech. Poniżej ilustruję graficznie, co ma się wydarzyć, a kilka następnych rozdziałów szczegółowo opisuje zalecane kroki.

Motywująca zmiana : Jaka jest historia, dlaczego ta zmiana jest potrzebna?

↓

Zrozumienie zmiany : Co ta zmiana oznacza w kontekście mojej roli i zestawu umiejętności?

↓

Obejmowanie działań opartych na danych : Jak mogę zastosować proces decyzyjny oparty na danych jako część procesu zmiany?

↓

Zabezpieczanie zmiany własności : Jak mogę zapewnić sobie prawo własności interesariuszy do długoterminowego zaangażowania?

↓

Kształcenie pracowników : Jaki jest najlepszy sposób na edukowanie i szkolenie ludzi na temat zmiany i ich nowej roli?

↓

Ciągła nauka : Jak mogę wprowadzić kulturę ciągłego uczenia się jako część zmiany?

Motywująca zmiana

Stworzenie przekonującego argumentu za zmianą jest niezbędnym punktem wyjścia. Ta fascynująca historia powinna określać, co inwestycja w naukę danych umożliwi firmie i organizacji w odniesieniu nie tylko do wewnętrznych polityk, procesów i pracowników, ale także konkurencji i klientów. Opierając się na podejściu opartym na historii, które wykorzystuje odpowiednie przykłady biznesowe jako część swojej argumentacji, będziesz w stanie pomóc swoim organizacjom zrozumieć pełny wpływ zmian nadchodzących na samym początku procesu. Aby móc jasno motywować zmianę, organizacja musi dokładnie zrozumieć, co oznacza każda zmiana i gdzie nastąpią zmiany w całym spektrum operacji biznesowych i informatycznych.

Zrozumienie zmian

Kolejnym krokiem w gotowości opartej na danych jest zdefiniowanie zmian pod względem operacyjnym w taki sposób, aby pracownicy mogli się do nich odnieść. Obejmuje to takie aspekty, jak wyjaśnienie celu zmiany lub tego, jak nadchodzące zmiany mogą wpłynąć na strukturę, procesy, umiejętności i cele związane z wydajnością. Zmiana nigdy nie jest łatwa. Pracownicy będą narażeni na nowe role, możliwości, kompetencje i sposoby pracy, więc sposób, w jaki firmy przygotowują pracowników do tej fundamentalnej zmiany, ma kluczowe znaczenie. Przede wszystkim musisz skoncentrować się na edukowaniu pracowników dzięki odpowiednim informacjom opartym na rolach i przygotowaniu ich do ewangelizacji danych w organizacji. . To spersonalizowane podejście do

urzeczywistniania i urzeczywistniania zmian napędza gotowość niezbędną do pomyślnego wprowadzenia nauki o danych.

Obejmowanie działań opartych na danych

Nauka o danych powoduje zmianę kulturową - taką, która jest najbardziej widoczna, jeśli chodzi o sposób podejmowania decyzji. W nauce o danych podejmowanie decyzji opiera się na podejściu opartym na danych o wiele bardziej niż na podejściu opartym na doświadczeniu lub przeczuciu. Zakłada również kulturę współpracy w organizacji, ponieważ tylko pracując razem, ludzie w całej organizacji mogą odkryć pełną wartość spostrzeżeń w odpowiednim kontekście biznesowym, które wspierają trwałą zmianę w kierunku decyzji i działań opartych na danych. Poleganie na metodach zwinnych i zespołach DevOps (operacje programistyczne) szybko staje się najlepszą praktyką podczas zarządzania transformacjami nauki o danych. W zwinnym podejściu organizacja umożliwia swoim pracownikom pracę tam, gdzie, kiedy i jak chcą, z maksymalną elastycznością i minimalnymi ograniczeniami, aby zoptymalizować swoją wydajność i zapewnić najlepszą w swojej klasie wartość i obsługę klienta. Podejście zespołowe DevOps łączy tworzenie oprogramowania (Dev) z operacjami informatycznymi (Ops). Celem DevOps jest skrócenie cyklu życia oprogramowania przy jednoczesnym częstym dostarczaniu w ścisłej zgodności z celami biznesowymi. Te związane z zespołem zmiany dodają również pracownikom warstwę złożoności, w której tradycyjne mury między zespołami organizacyjnymi zostają przesunięte, tworząc zespoły współpracujące. Organizacje potrzebują zatem swoich liderów, aby:

* Chętnie przyjmowali zmiany

* Komunikowali się z pracownikami o zachodzących zmianach

* Poświęcali trochę czasu na słuchanie i uczenie się od pracowników

W świecie nauki o danych budowanie tymczasowych zespołów wielofunkcyjnych, takich jak zespoły zadaniowe, nie wystarcza do rozwiązywania złożonych problemów biznesowych lub tworzenia innowacyjnych rozwiązań: organizacje muszą chcieć wspierać nieformalne grupy, w których osoby są zachęcane do poszukiwania i odkrywania ukrytych możliwości lub problemów mogą się zająć. Dla przywództwa równie ważne jest uznanie wkładu wnoszonego przez takie grupy w celu wzmocnienia ich pozycji i utrzymania ich na dłuższą metę.

Zabezpieczanie zmiany własności

Bez wątpienia najlepszym sposobem zarządzania złożonością transformacji jest stworzenie własności wśród interesariuszy, którzy ostatecznie spełnią obietnicę nowych technologii i możliwości. Idea powoływania tradycyjnych liderów zmian jest oldschoolowa. Zamiast tego nowy innowacyjny model uznaje, że liderzy biznesu, którzy w ramach zmiany przyjmują bardziej operacyjną rolę, są najbardziej zaufanymi źródłami informacji i wiarygodności w organizacji, a zatem powinni wdrożyć nową technologię i być właścicielem zmiany jako takiej. Stwórz historię, którą przywódcy mogą ogarnąć. Jednym ze sposobów, aby to zrobić, jest skuteczne wykorzystanie ukierunkowanych warsztatów, które pokazują, w jaki sposób przewidywane zmiany znacznie poprawią procesy biznesowe, systemy i praktyki w różnych segmentach biznesowych. Podczas tych warsztatów możesz umożliwić początkującym użytkownikom wśród kierownictwa współpracę z innymi interesariuszami w całej firmie. Korzystając z tego rozszerzającego się modelu do zarządzania zmianami w nauce o danych, możesz dotknąć wszystkich interesariuszy, których będziesz potrzebować, aby spełnić obietnicę dotyczącą danych, a także zmniejszyć ryzyko, że pracownicy poczują się zdemotywowani i wyobcowani przez zmianę.

Kształcenie pracowników

Zalecam połączenie niezbędnych programów edukacyjnych i szkoleniowych z elementami innymi niż standardowe szkolenie umiejętności, które będą (oczywiście) potrzebne. Umieść na górze listy obszary, takie jak psychologia, gry i komunikacja. Dodanie tych elementów do miksu pomaga skoncentrować wysiłki pracowników w zakresie uczenia się i rozwoju oraz umożliwia pracownikom zdobywanie nowych kompetencji i umiejętności wykraczających poza techniczne aspekty cyfrowych lub opartych na chmurze rozwiązań do analizy danych. Zastanów się, że może być konieczne wymaganie uczenia się na szerszą skalę, aby upewnić się, że podstawowe zasady nauki o danych są rozumiane przez znaczną część Twoich pracowników. Na przykład Google opracował kurs uczenia maszynowego, który jest obowiązkowy dla każdego pracownika technicznego.

Ciągła nauka

Transformacja nauki o danych wymaga nowego sposobu myślenia o tym, jak zmiana wpływa na ludzi, kultury, organizacje, procesy i nie tylko. Nie postrzegaj programu nauki danych jako niewielkiej części procesu; musisz raczej postrzegać to jako część całej podróży do cyfrowej transformacji Twojej firmy. Na przykład liderzy powinni prowadzić ciągłe programy mieszanego uczenia się i rozwoju, które angażują pracowników poprzez opisywanie praktycznych zastosowań nauki o danych, tak aby z czasem wśród pracowników narastało zrozumienie i zażyłość. Stałe wsparcie pomaga pracownikom przyjąć kulturę Agile i tworzy praktyków, którzy stale uczą się małymi krokami, iteracyjnie budują wiedzę i doświadczenie. Wybór podejścia do ciągłego uczenia się uwzględnia preferencje uczenia się wielopokoleniowej siły roboczej i jest skuteczny tam, gdzie występuje znaczna rotacja siły roboczej, niezależnie od tego, czy jest to planowana, czy nieplanowana. W każdym podejściu (tradycyjne lub ciągłe uczenie się) zarządzanie wpływem zmian powinno być postrzegane jako rdzeń dobrze zaplanowanego programu, wspieranego treścią opartą na faktach oraz odpowiednią i terminową komunikacją

Rozpoznanie, czego należy unikać, wprowadzając zmiany w nauce o danych

Na całym świecie wiele firm dokonało znacznych inwestycji w naukę o danych, uświadamiając sobie (prawidłowo) jej rewolucyjny potencjał. . Jednak nie odrobiwszy pracy domowej z właściwej analizy sytuacyjnej, wiele z tych firm poniosło ogromne straty, a nie oczekiwane korzyści. Niepowodzenie inwestycji w analitykę danych jest szczególnie powszechne wśród małych i średnich przedsiębiorstw. Dlaczego? Dlaczego średnie i małe firmy nie są w stanie uzyskać wystarczającej wartości poprzez wdrożenie nauki o danych? Jakie przeszkody stoją na ich drodze? Próbując znaleźć odpowiedzi na te pytania, Computer Associates przeprowadziło wywiady z 1000 menedżerami IT w firmach o przychodach przekraczających pół miliarda dolarów w wielu różnych branżach, od handlu detalicznego przez usługi finansowe po farmację. Wyniki ich badań wykazały, że zdecydowanie największą przeszkodą jest niewystarczająca infrastruktura. Wiele razy firmy utknęły w swoim dotychczasowym środowisku z powodu wcześniejszych kosztownych inwestycji, których „nie można wyrzucić". Dlatego zamiast tworzyć nową, nowoczesną architekturę danych, która skupia się na danych, firmy mają tendencję do dodawania aplikacji i elementów systemu do swojego starego środowiska, co sprawia, że nauka o danych jest nieefektywna i jeszcze bardziej kosztowna. Drugą największą przeszkodą jest złożoność organizacyjna. Zwykle staje się to problemem, gdy kierownictwo firmy nie docenia, jak transformująca jest nauka o danych. Wszystkie aspekty firmy muszą się zmienić, aby stać się zorientowane na dane, co oznacza, że wszyscy menedżerowie w firmie muszą rozumieć i wykorzystywać dane oraz nowe spostrzeżenia oparte na danych do podejmowania decyzji związanych z finansami, marketingiem, sprzedażą, rozwojem produktów i usług itd. . Jednak w rzeczywistości wiele firm traktuje analitykę danych jak działalność poboczną, dodając nowe role i funkcje do pracy z danymi, zamiast przekształcać istniejące funkcje i role. Trzecią najważniejszą przeszkodą są kwestie bezpieczeństwa i inne kwestie związane ze zgodnością. Nie jest to zaskakujące, biorąc pod uwagę

rosnącą świadomość wagi przetwarzania danych w sposób bezpieczny i etycznie poprawny. . Nowe przepisy i regulacje stają się coraz bardziej rygorystyczne, aby chronić prawo ludzi do prywatności, i dopóki standaryzacja w naukach o danych jest bardzo niewielka, wymagania będą się zmieniać. Ogólnym odkryciem w badaniu było to, że w zależności od wybranego rodzaju podejścia analitycznego poziom oporu był różny. Warto przyjrzeć się temu bliżej, dlatego w kolejnych akcjach przeprowadzę Cię przez różne typy projektów analitycznych wysokiego poziomu. Następnie możesz lepiej poznać główne czynniki leżące u podstaw sukcesu (lub niepowodzenia) projektu transformacji nauki o danych.

Projekty transformacji analiz opisowych

Projekty z zakresu analityki opisowej obejmują zadania mające na celu wykorzystanie danych do opisania tego, co się wydarzyło lub jak jest teraz – dlaczego na przykład sprzedaliśmy w tym miesiącu x produktów tego konkretnego typu. Obejmuje czynności, takie jak tworzenie wykresów, wykresów i kokpitów menedżerskich, którym towarzyszy brak (lub stosunkowo proste) funkcje analizy danych. Nacisk kładziony jest na zidentyfikowanie odpowiedniego zestawu wskaźników i przedstawienie informacji w skuteczny sposób.

Rozwiązania analityczne opisowe zazwyczaj napotykają mniejsze wyzwania związane z odpornością podczas ich implementacji. Powody są oczywiste – wyniki są łatwo zrozumiałe dla interesariuszy. Jednak czasami trudno uzasadnić biznesową wartość projektów z zakresu analityki opisowej. Podsumowując, przy ograniczonej analizie w analityce opisowej, jaka jest naprawdę wartość inwestowania w zrozumienie tego, co wydarzyło się wczoraj, kiedy tak naprawdę chcesz być przygotowany na jutro?

Projekty transformacji analityki diagnostycznej

Celem projektów z zakresu analityki diagnostycznej jest zrozumienie przyczyn danego zjawiska i przeprowadzenie analizy przyczyn źródłowych. Projekty z zakresu analityki diagnostycznej mogą zakończyć się opracowaniem modeli statystycznych (modeli wyjaśniających, modeli przyczynowych itd.) oraz pulpitów nawigacyjnych. Jednak wyniki muszą zawierać spostrzeżenia i zalecenia mające pomóc interesariuszom zrozumieć przyczyny tego, co się dzieje i zainicjować odpowiednie działania. Organizacje są zwykle podatne na wnioski analityczne i spostrzeżenia oparte na wynikach analityki diagnostycznej, ale istnieje nieco wyższy opór, jeśli chodzi o wdrażanie zaleceń. Wynika to głównie z faktu, że użytkownicy biznesowi są świadomi, że niektóre rekomendacje nie są wykonalne, ponieważ wymagają zbyt wielu zmian lub mają zbyt wiele ograniczeń.

Projekty transformacji analiz predykcyjnych

Projekty z zakresu analityki predykcyjnej obejmują prognozowanie określonej metryki lub przewidywanie określonego zjawiska. Modelowanie predykcyjne to proces stosowania modelu statystycznego lub algorytmu eksploracji danych na danych w celu przewidywania nowych lub przyszłych obserwacji. Modele predykcyjne mogą być wykorzystywane nie tylko do prognozowania, ale także do celów symulacyjnych. Przykłady obejmują badania kliniczne, prognozy sprzedaży, awarie produkcji i prognozy pogody. Jak można się spodziewać, rozwiązania do analizy predykcyjnej napotykają na największy opór. Rozwiązania diagnostyczne i opisowe w dużej mierze dotyczą tego, co już się wydarzyło, a rozwiązania predykcyjne dotyczą czegoś, co jeszcze się wydarzyło. Dlatego użytkownicy biznesowi mają zastrzeżenia do rozwiązań predykcyjnych. Ten sceptycyzm nie jest bezpodstawny, ponieważ koszt błędnych prognoz może być zdumiewający.

Wykorzystywanie technik Data Science do prowadzenia skutecznych zmian

Aby inwestycja w naukę danych powiodła się, przyjęta strategia analizy danych powinna obejmować dobrze przemyślane strategie zarządzania fundamentalną zmianą, którą rozwiązania do nauki danych narzucają organizacji. Jednym ze skutecznych i wydajnych sposobów radzenia sobie z tymi wyzwaniami jest wykorzystanie technik zarządzania zmianą opartych na danych, aby napędzać samą transformację - innymi słowy, napędzaj zmianę poprzez „praktykowanie tego, co głosisz". Przeprowadzę Cię przez kilka przykładów, jak to zrobić w praktyce.

Korzystanie z cyfrowych narzędzi angażujących

Dla firm pojawiła się nowa generacja narzędzi do badania opinii pracowników w czasie rzeczywistym, które zaczynają zastępować przestarzałe badania opinii pracowników. Te narzędzia mogą powiedzieć znacznie więcej niż tylko to, o czym pracownicy myślą raz w roku. W niektórych firmach pracownicy są ankietowani co tydzień za pomocą ograniczonej liczby pytań. Pytania i modele są skonstruowane w taki sposób, aby kierownictwo mogło śledzić wahania ważnych wskaźników w miarę ich pojawiania się, a nie zwykle raz lub dwa razy w roku. Narzędzia te mają oczywiste znaczenie dla zarządzania zmianą i mogą pomóc odpowiedzieć na takie pytania:

* Czy zmiana jest równie dobrze odbierana w różnych lokalizacjach?

* Czy niektórzy menedżerowie są lepsi od innych w dostarczaniu wiadomości do pracowników?

Załóżmy, że masz dużą firmę zajmującą się podróżami i turystyką, która używa jednego z tych narzędzi do uzyskiwania informacji zwrotnych od pracowników w czasie rzeczywistym. Jednym z podejść opartych na danych do zastosowania w takiej sytuacji jest eksperymentowanie z różnymi strategiami zarządzania zmianą w wybranych populacjach w firmie. Po kilku zmianach w organizacji możesz wykorzystać zebrane dane do zidentyfikowania, którzy menedżerowie okazują się bardziej skuteczni w prowadzeniu zmian niż inni. Po ustaleniu tego możesz obserwować tych menedżerów, aby określić, co robią inaczej. Następnie możesz podzielić się skutecznymi technikami z innymi menedżerami. Ten rodzaj informacji zwrotnej w czasie rzeczywistym daje możliwość szybkiego uczenia się, w jaki sposób odbierane są zdarzenia komunikacyjne lub taktyki zaangażowania, optymalizując w ten sposób działania w ciągu dni (a nie tygodni, co jest typowe dla tradycyjnych metod). Dane te mogą następnie zostać wprowadzone do modelu predykcyjnego, pomagając precyzyjnie określić, które działania pomogą przyspieszyć przyjęcie nowej praktyki, procesu lub zachowania przez daną grupę pracowników. . Można tam znaleźć kilka komercyjnych narzędzi - na przykład sondaże IQ kultury - które wspierają ten rodzaj zbierania danych. Tego rodzaju ankiety obejmują codziennie lub co tydzień grupy pracowników za pośrednictwem aplikacji na smartfony, aby generować w czasie rzeczywistym informacje zgodnie z określonym przez Ciebie zakresem. Inne narzędzie, Waggl.com (www.waggl.com), ma bardziej zaawansowaną funkcjonalność, pozwalającą na stałą rozmowę z pracownikami na temat wysiłku związanego ze zmianą, a także umożliwiającą menedżerom zmian powiązanie tego dialogu z postępem inicjatyw, które ponowne przedsięwzięcie. Te różne rodzaje narzędzi cyfrowego zaangażowania mogą mieć ogromny wpływ na programy zmian, ale strumień danych, który tworzą, może być jeszcze ważniejszy. Wygenerowane dane można wykorzystać do budowy modeli predykcyjnych zmian. Używanie i wdrażanie tych modeli w rzeczywistych projektach transformacji, a następnie dzielenie się swoimi odkryciami pomaga zapewnić wyższy wskaźnik sukcesu dzięki inicjatywom zmian opartych na danych w przyszłości.

Zastosowanie analityki mediów społecznościowych do identyfikacji nastrojów interesariuszy Menedżerowie ds. zmian mogą również spojrzeć poza granice przedsiębiorstwa, aby uzyskać wgląd w wpływ programów zmian. Klienci, partnerzy handlowi, dostawcy i inwestorzy są kluczowymi interesariuszami, jeśli chodzi o programy zmian. Są również bardziej skłonni niż pracownicy do komentowania w mediach społecznościowych zmian wprowadzanych przez firmę, co daje potencjalnie

istotny wgląd w to, jak reagują. Ernst & Young (obecnie znany jako EY) korzysta z narzędzia do analizy mediów społecznościowych o nazwie SMAART, które może interpretować nastroje wśród grup konsumentów i influencerów. W projekcie dla firmy farmaceutycznej EY udało się wyodrębnić konkretne źródła informacji, które wzbudziły pozytywne i negatywne nastawienie do marki klienta. Firma zaczyna teraz stosować te techniki, aby zrozumieć zewnętrzny wpływ wysiłków na rzecz zmian, a rozszerzenie tych technik w przedsiębiorstwie jest prostym krokiem. Postępy w lingwistycznej analizie tekstów oznaczają że wskazówki dotyczące zachowania można teraz uchwycić na podstawie doboru słów danej osoby; nawet użycie rodzajników i zaimków może pomóc ujawnić, jak ktoś się czuje. Zastosowanie narzędzi do analizy sentymentu do danych w zanonimizowanych firmowych wiadomościach e-mail lub dialogu w narzędziach takich jak Waggl.com może dać świeży wgląd w gotowość organizacji do zmian i reakcje pracowników na różne inicjatywy. Spostrzeżenia uzyskane z analizy komunikacji wewnętrznej będą silniejsze w połączeniu z zewnętrznymi danymi z mediów społecznościowych.

Zbieranie danych referencyjnych w projektach zmian

Czy kiedykolwiek pracowałeś w organizacji, w której różne programy lub projekty zmian były ze sobą porównywane pod kątem skuteczności ich wprowadzenia? Albo taki, w którym w różnych inicjatywach zmian zastosowano standardowy zestaw pomiarów? Nie? Ja też nie. Dlaczego organizacje często wydają się mieć obsesję na punkcie mierzenia ułamkowych zmian w wydajności operacyjnej i przechwytywania danych dotyczących sprzedaży, obrotów zapasów i wydajności produkcji, ale nie wykazują zainteresowania śledzeniem wydajności od projektu zmiany do projektu zmiany, poza wiedzą, które z nich spełniły swoje zadanie cele? Niektórzy ludzie mogą twierdzić, że nie można porównywać projektów zmian w organizacji; to byłoby jak porównywanie jabłek do pomarańczy. Nie zgadzam się: różne projekty mogą mieć unikalne cechy, ale znajdziesz więcej podobieństw niż różnic między różnymi typami projektów. Dobrym pomysłem jest przechwytywanie informacji o zaangażowanym zespole, populacji zaangażowanej w zmianę, czasie potrzebnym na wdrożenie, zastosowanych taktykach i tak dalej. Umożliwia zbudowanie zestawu danych referencyjnych do przyszłego uczenia się, ponownego wykorzystania i testowania wydajności. Należy jednak pamiętać, że chociaż może nie przynieść natychmiastowych korzyści, wraz ze wzrostem ogólnego zbioru danych, ułatwi budowanie dokładnych modeli predykcyjnych przyszłych zmian organizacyjnych. . Wykorzystywanie danych do selekcji osób do zmiany ról Od dłuższego czasu firmy stosują metody oparte na danych przy selekcji kandydatów na wyższe stanowiska kierownicze. A dzisiaj niektóre firmy, takie jak detaliści, zaczynają używać analiz predykcyjnych do zatrudniania pracowników pierwszej linii. Zastosowanie tych narzędzi podczas budowania zespołu ds. zmian może zarówno znacząco poprawić wydajność projektu, jak i pomóc w stworzeniu kolejnego nowego zestawu danych. Gdyby każdy lider zmiany i członek zespołu przeszedł testy i ocenę przed rozpoczęciem projektu zmiany, dane te mogą stać się ważnymi zmiennymi, które należy uwzględnić podczas wyszukiwania modelu bazowego, który prowadzi do udanych projektów zmian. Można to nawet rozszerzyć na bardziej nieformalne role, takie jak liderzy zmian, umożliwiając organizacjom optymalizację wyboru w oparciu o to, co wiedzą o sukcesie osobowości dla tego typu ról. W tym duchu kalifornijski start-up LEDR Technologies jest pionierem w zakresie technik przewidywania wydajności zespołu. . Integruje źródła danych i wykorzystuje je, aby pomóc zespołom przewidywać wyzwania, z jakimi mogą się zmierzyć, dzięki dynamice zespołu, tak aby zespół mógł im zapobiec, zanim się pojawią

Automatyzacja metryk zmian

Wyobraź sobie firmę lub organizację, która ma spersonalizowany pulpit nawigacyjny opracowany we współpracy z zespołem kierowniczym firmy - taki, który odzwierciedla priorytety firmy, pozycję konkurencyjną i plany na przyszłość. Te pulpity nawigacyjne powinny również służyć do oferowania

wglądu w różne dokonane inwestycje w transformację. Należy pamiętać, że wiele danych, które mogą pełnić funkcję interesujących wskaźników zmian, jest już dostępnych – po prostu nie są one gromadzone. Kiedy firma buduje pulpit nawigacyjny do identyfikowania rekrutacji i odpływu, uczy zespół wykonawczy wykorzystywania danych do podejmowania decyzji związanych z ludźmi. Jednak prawidłowe skonfigurowanie i usunięcie błędów może zająć trochę czasu. Moja sugestia? Nie czekaj. Zacznij budować tego typu kokpity tak szybko, jak to możliwe i tam, gdzie to możliwe, zautomatyzuj je. Dlaczego automatyzacja? Pulpity nawigacyjne zmian są podatne na problemy z kontrolą wersji, błędy ludzkie i politykę wewnętrzną. Automatyzacja zarządzania danymi i generowania pulpitów nawigacyjnych może sprawić, że będzie on bardziej przejrzysty i pomoże zachować integralność danych.

Pierwsze kroki

W miarę gromadzenia przez organizacje większej ilości danych i budowania dokładniejszych modeli, menedżerowie zmian będą mogli bez obaw wykorzystywać je do określania strategii umożliwiających organizacjom osiągnięcie ich celów. Będą w stanie odpowiedzieć na ważne pytania, takie jak:

* Którzy interesariusze są zaangażowani? Jakie podejście do zmiany sprawdza się w przypadku grup, które mają te cechy?

* Jakie zagrożenia wiążą się z programami, które udostępniają te funkcje?

* Jakie są techniki przyspieszające dostarczanie korzyści biznesowych i jakie są ich względne koszty?

* Jaka jest przyczyna i skutek poszczególnych rodzajów inwestycji?

Na wszystkie te pytania można odpowiedzieć za pomocą danych i będą one stanowić podstawę planów transformacji opartej na danych. Opracowywanie tego rodzaju wskaźników nie jest ani szybkie, ani łatwe. Nie są to jednorazowe instalacje, ale wieloletnie zobowiązania do przechwytywania danych, budowania modeli i udoskonalania pulpitów nawigacyjnych. Stworzenie stabilnych i wiarygodnych zbiorów danych wymaga czasu. Jakość danych jest problemem wszędzie, podobnie jak potrzeba wspólnego języka danych, który pozwoli organizacjom wiedzieć, że mierzą to, co zamierzają mierzyć. Stanowiło to problem dla analityki danych w innych dziedzinach; nie ma powodu sądzić, że zarządzanie zmianą będzie inne. Chociaż zajmie to trochę czasu, w końcu będziesz w stanie zamknąć pętlę przyczynową i dokonać wiarygodnych prognoz dotyczących tego, jak działanie lub inicjatywa w programie zmian wpłynie na daną metrykę. To sprawi, że inwestycja w zmianę zmieni się z aktu wiary w decyzję opartą na danych. Zarządzanie zmianą przejdzie z dyscypliny opartej na projektach, która ma problemy z uzasadnieniem odpowiednich inwestycji, do takiej, która doradza w zakresie wyników biznesowych i sposobów ich dostarczania. Doprowadzi to do spadku jedynej miary, która jest dobrze znana w programach zmian - wskaźnika niepowodzeń. W ramach wprowadzenia zarządzania zmianą opartego na danych powinno wreszcie być możliwe rozwiązanie wielkiej zagadki, dlaczego tak wiele wysiłków transformacyjnych kończy się niepowodzeniem.

Zrozumienie przeszłości, teraźniejszości i przyszłości danych

Decyzje biznesowe zmuszają Cię do skupienia się, alokacji ograniczonych zasobów i zastanowienia się nad tym, jak być wyjątkowym w porównaniu z konkurencją. Pamiętaj, że nie możesz być wszystkim dla wszystkich. Strategicznie powinieneś myśleć o prostocie nad złożonością, ponieważ jasna i prosta strategia jest dużo łatwiejsza do wyjaśnienia i wdrożenia. Ale co z dokonywaniem strategicznych wyborów dotyczących Twoich danych? Jasne, dane mogą pomóc Ci zrozumieć opcje strategiczne, a także potencjalny wpływ różnych wyborów z perspektywy biznesowej, ale jak wykorzystać dane, aby lepiej zrozumieć same dane? Cóż, aby dokonywać wyborów, musisz dokonywać wyborów. Prawdziwych wyborów dotyczących Twoich danych nie można dokonać, jeśli nie wiesz, jakie masz opcje i jakie opcje zdecydowanie odrzucasz. Decydując się na realną strategię, zbyt często alternatywne opcje strategiczne są rozważane tylko powierzchownie i potrzebujesz więcej, aby dokonać właściwych wyborów.

Porządkowanie podstaw danych

Terminy dane i informacje są często używane zamiennie; jest jednak między nimi różnica. Na przykład dane można opisać jako surowe, niezorganizowane fakty, które wymagają przetworzenia - zbiór liczb, symboli lub znaków, zanim zostaną oczyszczone i poprawione. Surowe dane muszą zostać poprawione, aby usunąć wady, takie jak wartości odstające i błędy wprowadzania danych. Można generować surowe dane na wiele różnych sposobów. Na przykład dane terenowe to surowe dane, które zostały zebrane w niekontrolowanym środowisku na żywo. Dane eksperymentalne zostały wygenerowane w kontekście badań naukowych poprzez obserwację i rejestrację. Dane mogą być proste i pozornie losowe i bezużyteczne, dopóki nie zostaną zorganizowane, ale gdy dane zostaną przetworzone, zorganizowane, ustrukturyzowane lub zaprezentowane w określonym kontekście, który czyni je użytecznymi, nazywane są informacją. Historycznie pojęcie danych było najściślej związane z badaniami naukowymi, ale obecnie dane są gromadzone, przechowywane i wykorzystywane przez coraz większą liczbę firm, organizacji i instytucji. W przypadku firm przykładami interesujących danych mogą być dane klientów, dane produktów, dane sprzedaży, przychody i zyski; w przypadku rządów może obejmować takie dane, jak wskaźniki przestępczości i stopy bezrobocia. W drugiej połowie XX wieku podjęto kilka prób standaryzacji kategoryzacji i struktury danych w celu zrozumienia ich różnych form. Jednym z dobrze znanych modeli jest piramida DIKW (dane, informacja, wiedza i mądrość), opisana na poniższej liście; pierwsza wersja tego modelu powstała już w połowie lat 50., ale po raz pierwszy pojawiła się w obecnym stanie w połowie lat 90., jako próba zrozumienia rosnącej ilości danych (surowych lub przetworzonych), które były generowane z różnych systemów komputerowych:

* Dane są surowe. Po prostu istnieje i nie ma znaczenia poza swoim istnieniem (samo w sobie). Może istnieć w dowolnej formie, użytecznej lub nie. Dane reprezentują fakt lub stwierdzenie zdarzenia bez związku z innymi czynnikami - na przykład pada deszcz.

* Informacja to dane, którym nadano znaczenie w wyniku jakiegoś związku. To znaczenie może być przydatne, ale nie musi. Zależność informacyjna może mieć związek przyczynowo-skutkowy - temperatura spadła o 15 stopni, a potem zaczął padać np. deszcz.

* Wiedza to zbieranie informacji, które mają być przydatne. Reprezentuje wzór, który łączy dyskretne elementy i ogólnie zapewnia wysoki poziom przewidywalności tego, co jest opisane lub co się stanie dalej: Jeśli wilgotność jest bardzo wysoka, a temperatura znacznie spada, atmosfera często nie jest w stanie utrzymać wilgoci i tak na przykład pada.

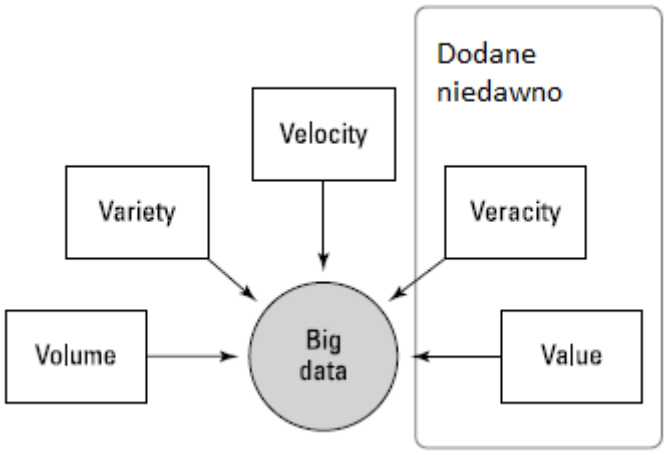
* Mądrość jest przykładem zrozumienia podstawowych zasad w wiedzy, które zasadniczo tworzą podstawę wiedzy będącej tym, czym jest. Mądrość jest zasadniczo jak wspólne zrozumienie, które nie

jest kwestionowane; Pada, bo na przykład pada. Obejmuje to zrozumienie wszystkich interakcji zachodzących między deszczem, parowaniem, prądami powietrza, gradientami temperatury, zmianami i deszczem.

Piramida DIKW zaoferowała nowy sposób kategoryzowania danych przechodzących przez różne etapy swojego cyklu życia i przez lata przykuwała uwagę. Jednak został również skrytykowany i pojawiły się warianty, które miały ulepszyć oryginał. Jedną z głównych krytyki jest to, że chociaż dość łatwo jest zrozumieć przejście od danych do informacji, znacznie trudniej jest wytyczyć jasną i słuszną granicę od informacji do wiedzy i od wiedzy do mądrości, co utrudnia zastosowanie w praktyce. Modele koncepcyjne to narzędzia heurystyczne: są przydatne tylko wtedy, gdy oferują sposób na nauczenie się czegoś nowego. Ten czy inny model może być dla Ciebie bardziej atrakcyjny, ale z perspektywy wdrożenia data science najważniejszą rzeczą do rozważenia jest takie pytanie: Czy moja firma zyska wartość dzięki posiadaniu czterech poziomów piramidy DIKW, czy po prostu sprawi, że wdrożenie będzie trudniejsze i bardziej złożone? (Osobiście lubię piramidę z tylko dwoma poziomami: danymi i spostrzeżeniami. Ta jak dotąd działała dobrze i jest o wiele łatwiejsza do wyjaśnienia i zdobycia poparcia.)

Wyjaśnienie tradycyjnych danych kontra big data

Tradycyjne dane to dane w objętości i formacie, który ułatwia dostęp, pracę i działanie. Big data to inne zwierzę, jednak definiowane bardziej przez ilość danych, różnorodność ich typów i szybkość ich przetwarzania. Jeśli wszystkie trzy z tych cech są spełnione, jeśli chodzi o to, że są zbyt duże, aby można je było obsłużyć w zwykłym środowisku przetwarzania, można założyć że masz do czynienia z big data, a nie tradycyjnymi danymi (obecnie znanymi, po pojawieniu się big data na scenie, jako small data). Ponadto termin big data jest używany w odniesieniu do zestawów danych, które są zbyt duże lub zbyt złożone, aby mogły być obsługiwane przez tradycyjne aplikacje do przetwarzania danych. Wyzwania związane z dużymi danymi obejmują takie zadania, jak przechwytywanie danych, przesyłanie danych, przechowywanie danych, czyszczenie i przygotowywanie danych, eksploracja i analiza danych, wyszukiwanie danych, udostępnianie i ponowne wykorzystywanie danych, wizualizacja danych, aktualizacja danych oraz zarządzanie prywatnością, własnością danych i zarządzaniem. Chociaż pierwotnie big data opisywano trzema kluczowymi pojęciami – objętością, różnorodnością i szybkością – ostatnio dodano dwa inne pojęcia: prawdziwość (innymi słowy jakość danych) i wartość. Te dodatkowe cechy zostały dodane, aby opisać dwa inne ważne aspekty big data, które należy wziąć pod uwagę podczas szacowania potencjalnych korzyści z zestawu danych big data. Co jest tak ważnego w prawdziwości? Jeśli zestaw danych, który dana firma chce eksplorować, spełnia kryteria big data w oparciu o ilość, różnorodność i szybkość, może to być nadal bezużyteczne dla firmy, jeśli jakość danych jest słaba i nie można jej poprawić. Niska jakość danych (zestaw danych jest na przykład niekompletny, uszkodzony lub stronniczy) bezpośrednio wpływa na zaufanie do samych danych, ostatecznie wpływając na postrzeganą wartość całego zestawu danych. Rysunek przedstawia graficznie kluczowe pojęcia, znane jako „pięć V”, które definiują duże zbiory danych.



Dodaję własne podejście do tych pojęć na poniższej liście:

* Volume (Objętość): Odnosi się do ilości generowanych i przechowywanych danych. Rozmiar danych określa wartość i potencjalny wgląd oraz to, czy można je uznać za duże zbiory danych. Gdy zaczniesz uwalniać moc big data, wkrótce odkryjesz, że strumienie danych połączone z Twoją firmą będą rosły wykładniczo. Ten gwałtowny wzrost ilości danych może spowodować znaczne trudności w Twojej organizacji, jeśli nie zaplanujesz ich odpowiednio, ponieważ każdy nowy zestaw danych znacznie obciąża Twoje obecne przechowywanie danych i konfigurację obliczeniową.

* Velocity (Prędkość): odnosi się do szybkości, z jaką dane są generowane i przetwarzane w celu spełnienia celów biznesowych Twojej organizacji. Duże zbiory danych są często dostępne w czasie rzeczywistym. W porównaniu z małymi danymi, duże zbiory danych są tworzone w sposób bardziej ciągły. Te dwa rodzaje prędkości są związane z Big Data:

• Częstotliwość generowania

• Częstotliwość przechwytywania

Jeśli masz trudności z pogodzeniem się z wykładniczo rosnącą ilością danych w Twojej firmie, fakt, że prędkość również wzrośnie, prawdopodobnie wydaje się onieśmielający. Aby jednak w pełni wykorzystać big data, musisz skupić się nie tylko na ilości gromadzonych informacji, ale także na tym, jak szybko możesz wykorzystać dane do podejmowania decyzji biznesowych lub włączania i wykorzystywania ich jako części usługi lub oferty produktowej. Twoja firma.

* Variety (Różnorodność) : Odnosi się do rodzaju i kontekstu danych. Różnorodność pomaga osobom analizującym dane efektywnie wykorzystać uzyskany wgląd w dane. Historycznie dane były w dobrze ustrukturyzowanym formacie w miejscu ich gromadzenia. Teraz, gdy korzystanie z big data staje się standardową praktyką, to nieustrukturyzowane dane szybko stają się normą w świecie korporacji. Jeśli chcesz rozszerzyć wykorzystanie big data, musisz przyzwyczaić się do braku uporządkowanych danych. Jednak przy odpowiedniej konfiguracji analitycznej te nowe rodzaje danych mogą przyspieszyć rozwój Twojej firmy, umożliwiając eksplorację nowych możliwości na nieznanym wcześniej terytorium.

* Veracity (Wiarygodność): odnosi się do poziomu szumu w danych. Jakość przechwyconych danych może się znacznie różnić, wpływając na dokładność analizy. Tak, big data zapewnia Twojej firmie możliwość gromadzenia informacji z miejsc, o których nigdy nie myślałeś, że są możliwe, ale jeśli dane nie są dokładne lub aktualne, nie ma znaczenia, co zdecydujesz się z nimi zrobić.

* Value (Wartość): odnosi się do wartości biznesowej uzyskanej dzięki dużym zbiorom danych.

A więc masz to - pięć V big data: objętość, prędkość, różnorodność, wiarygodność i wartość. Nie wszystkie te cechy big data są jednak równie ważne. Cztery z nich (objętość, prędkość, różnorodność i prawdziwość) można postrzegać jako czynniki umożliwiające. Osiągnięcie piątego V (wartości) wymaga zrozumienia tego, co organizacja stara się osiągnąć. Dlatego podkreślam potrzebę jasnego określenia ogólnych strategicznych celów biznesowych, zanim będziesz mógł wykorzystać wolumen, szybkość, różnorodność i prawdziwość, aby osiągnąć wartość w dużych zbiorach danych.

Znajomość wartości danych

Stwierdzenie „Dane to nowa ropa" to takie, które wypowiada wiele osób, ale co to znaczy? Pod pewnymi względami analogia pasuje: łatwo jest narysować paralele ze względu na sposób, w jaki informacje (dane) są wykorzystywane do napędzania większości dostępnej obecnie technologii transformacyjnej poprzez sztuczną inteligencję, uczenie maszynowe, automatyzację i zaawansowaną analitykę – podobnie jak napędy naftowe światowa gospodarka przemysłowa. Tak więc, jako podejście marketingowe i opis wysokiego poziomu, wyrażenie spełnia swoje zadanie, ale jeśli potraktujesz je jako wskazówkę, jak strategicznie zająć się wartością danych, może to prowadzić do inwestycji, których nie można przekształcić w wartość. Na przykład przechowywanie danych nie ma gwarantowanej przyszłej wartości, tak jak ma to miejsce w przypadku ropy naftowej. Przechowywanie jeszcze większej ilości danych ma jeszcze mniejszą wartość, ponieważ jeszcze trudniej jest je znaleźć, aby można było z nich korzystać. Wartość danych nie polega na ich zapisywaniu czy przechowywaniu, ale na używaniu ich w nieskończoność. Wtedy realizowana jest wartość w danych. Jeśli zaczniesz od spojrzenia na sedno analogii, zobaczysz, że odnosi się ona do aspektów wartości danych jako czynnika umożliwiającego fundamentalną transformację społeczeństwa – tak jak ropa naftowa udowodniła, że jest w historii. Z tej perspektywy zdecydowanie pokazuje podobieństwa między ropą naftową a danymi. Innym podobieństwem jest to, że chociaż z natury są cenne, dane wymagają przetworzenia – tak jak ropa wymaga rafinacji – zanim będzie można odkryć ich prawdziwą wartość. Jednak dane mają również wiele innych aspektów, które powodują, że analogia się rozpada przy bliższym przyjrzeniu się. Aby zobaczyć, co mam na myśli, sprawdź niektóre różnice, które widzę między tymi dwoma czynnikami umożliwiającymi transformację:

* Dostępność: chociaż ropa jest ograniczonym zasobem, dane są nieskończonym i stale rosnącym zasobem. Oznacza to, że traktowanie danych jak ropa (na przykład gromadzenie jej i przechowywanie w silosach) przynosi niewielkie korzyści i zmniejsza ich użyteczność. Niemniej jednak, z powodu błędnego przekonania, że dane są podobne do ropy (niedobór), często właśnie to robi się z danymi, kierując inwestycje i zachowania w złym kierunku.

* Możliwość ponownego wykorzystania: dane stają się tym bardziej przydatne, im częściej są używane, co jest dokładnym przeciwieństwem tego, co dzieje się z olejem. Kiedy olej jest używany do wytwarzania energii, takiej jak ciepło lub światło, lub gdy olej jest trwale przekształcany w inną formę, taką jak plastik, olej znika i nie można go ponownie wykorzystać. Dlatego traktowanie danych jak oleju – użycie go raz, a potem założenie, że jego przydatność się wyczerpała i wyrzucenie – jest zdecydowanie błędem.

* Przechwytywanie: wszyscy wiedzą, że w miarę zmniejszania się światowych rezerw ropy naftowej wydobycie staje się coraz trudniejsze i bardziej kosztowne. Z drugiej strony dane stają się coraz bardziej dostępne wraz ze wzrostem cyfryzacji społeczeństwa.

* Różnorodność: Dane są również znacznie bardziej zróżnicowane niż olej. Surowy olej, który jest wiercony z ziemi, jest oczywiście przetwarzany na różne sposoby w wiele różnych produktów, ale w

stanie surowym wszystko jest takie samo. Dane w swoim surowym formacie mogą reprezentować słowa, obrazy, dźwięki, pomysły, fakty, pomiary, statystyki lub inne cechy, które mogą być przetwarzane przez komputery.

Niemniej jednak pozostaje faktem, że dostępne dziś ilości danych stanowią zupełnie nowy towar, chociaż wciąż trwają prace nad zasadami gromadzenia, przechowywania, przetwarzania i wykorzystywania danych. Podkreślam jednak, że dane, podobnie jak ropa, są istotnym źródłem siły, a firmy, które wykorzystują dostępne dane w najbardziej zoptymalizowany sposób (a tym samym kontrolują rynek) stają się liderami światowej gospodarki, po prostu jak baronowie naftowi sto lat temu.

Badanie aktualnych trendów w danych

Zaledwie kilka lat temu zdecydowanie istniały Big Data, ale teraz jest znacznie więcej szumu wokół idei wartości danych - a dokładniej, w jaki sposób analiza może przekształcić dane w wartość. W następnych kilku sekcjach przyjrzymy się niektórym trendom związanym z wykorzystywaniem danych do uchwycenia nowej wartości.

Monetyzacja danych

Dane dotyczące zarabiania odnoszą się do tego, w jaki sposób firmy mogą wykorzystać swoją wiedzę specjalistyczną w dziedzinie domeny, aby przekształcić dane, które posiadają lub do których mają dostęp, w rzeczywistą, namacalną wartość biznesową lub nowe możliwości biznesowe. Monetyzacja danych może odnosić się do aktu generowania wymiernych korzyści ekonomicznych z dostępnych źródeł danych za pomocą analityki lub, rzadziej, może odnosić się do aktu monetyzacji usług danych. W przypadku analityki korzyści te zwykle pojawiają się jako przychody lub oszczędności kosztów, ale mogą również obejmować udział w rynku lub wzrost wartości rynkowej firmy. Można argumentować, że monetyzacja danych w celu zwiększenia przychodów firmy lub oszczędności kosztów jest po prostu wynikiem bycia organizacją opartą na danych. Chociaż ten argument nie jest całkowicie błędny, liderzy firm coraz bardziej interesują się rynkiem, aby zbadać, w jaki sposób monetyzacja danych może napędzać innowacje całkowicie nowych modeli biznesowych w różnych segmentach biznesowych. Dobrym przykładem tego, jak ten proces może działać, jest sprzedaż przez operatorów telekomunikacyjnych danych o pozycjach szybko tworzących się grup użytkowników (wyobraźcie sobie zakończenie wydarzenia sportowego lub koncertu najnowszej sensacji YouTube) firmom taksówkarskim. Dzięki temu taksówki mogą być proaktywnie dostępne w odpowiednim miejscu, kiedy najprawdopodobniej będzie potrzebna taksówka. To zupełnie nowy model biznesowy i baza klientów dla tradycyjnego operatora telekomunikacyjnego, otwierający nowe rodzaje działalności i przychodów w oparciu o dostępne dane.

Odpowiedzialna sztuczna inteligencja

Odpowiedzialne systemy AI charakteryzują się przejrzystością, odpowiedzialnością i uczciwością, gdzie użytkownicy mają pełny wgląd w to, jakie dane są wykorzystywane i w jaki sposób. Zakłada również, że firmy informują o możliwych konsekwencjach wykorzystania danych. Obejmuje to zarówno potencjalny pozytywny, jak i negatywny wpływ. Odpowiedzialna sztuczna inteligencja to także budowanie zaufania klientów i interesariuszy w oparciu o przestrzeganie komunikowanych polityk i zasad w czasie, w tym zdolność do utrzymania kontroli nad samym środowiskiem systemu sztucznej inteligencji. Strategiczne projektowanie infrastruktury i rozwiązań analitycznych w Twojej firmie z myślą o odpowiedzialnej sztucznej inteligencji jest nie tylko mądre, ale może również okazać się prawdziwym wyróżnikiem biznesowym w przyszłości. Wystarczy spojrzeć, jak odwrotne podejście, przyjęte przez Facebooka i Cambridge Analytica, przerodziło się w skandal, który zakończył się

wykluczeniem Cambridge Analytica z rynku. Być może pamiętasz, że Cambridge Analytica uzyskała dostęp do prywatnych i osobistych informacji ponad 50 milionów użytkowników Facebooka w Stanach Zjednoczonych, a następnie zaoferowała narzędzia, które mogłyby następnie wykorzystać te dane do identyfikacji osobowości amerykańskich wyborców i wpływać na ich zachowanie. Facebook, zamiast zostać zhakowanym, chętnie pozwalał na wykorzystywanie danych swoich użytkowników do innych celów bez wyraźnej zgody użytkownika. Dane zawierały szczegóły dotyczące tożsamości użytkowników, sieci znajomych i „polubień". Pomysł polegał na zmapowaniu cech osobowości na podstawie tego, co ludzie polubili na Facebooku, a następnie wykorzystaniu tych informacji do kierowania reklam cyfrowych do odbiorców. Facebook został również oskarżony o rozpowszechnianie rosyjskiej propagandy i fake newsów, które wraz z incydentem z Cambridge Analytica poważnie wpłynęły na markę Facebooka w ciągu ostatnich kilku lat. Ten rodzaj poważnego naruszenia prywatności nie tylko otworzył oczy wielu osobom pod względem wykorzystania ich danych, ale także wpłynął na marki firmy.

ROLA OPEN SOURCE W NAUCE O DANYCH

Architektury danych o otwartym kodzie źródłowym nie są już analogiczne do projektów badawczych prowadzonych w środowiskach laboratoryjnych na potrzeby prób i eksperymentów. Obecnie uważane za główny nurt w środowiskach IT, architektury te są szeroko wdrażane w produkcji na żywo w kilku branżach. W rzeczywistości stało się to tak powszechne, że jeśli budujesz nowoczesną architekturę danych, prawdopodobnie używasz stosu open source. Niektóre firmy odkryły nawet, że korzystanie z architektur open source jest jedyną opłacalną drogą do wykonania czegoś. Punkt krytyczny jest mniej więcej tutaj, co oznacza, że nadszedł czas, aby zdecydować, jak strategicznie zareagować na możliwości związane z open source. To już punkt, w którym wystarczyło wprowadzanie niewielkich zmian przyrostowych lub bezpieczne korzystanie z tradycyjnych, zastrzeżonych infrastruktur. Teraz, jeśli nadal będziesz grać bezpiecznie lub będziesz trzymać się małych kroków, zostawisz swoją firmę na ryzyko zostania w tyle, podczas gdy konkurenci będą iść do przodu. Należy również pamiętać, że podejmowanie decyzji o wykorzystaniu oprogramowania open source nie jest przyrostowe; raczej wymaga to pełnego, destrukcyjnego podejścia architektonicznego.

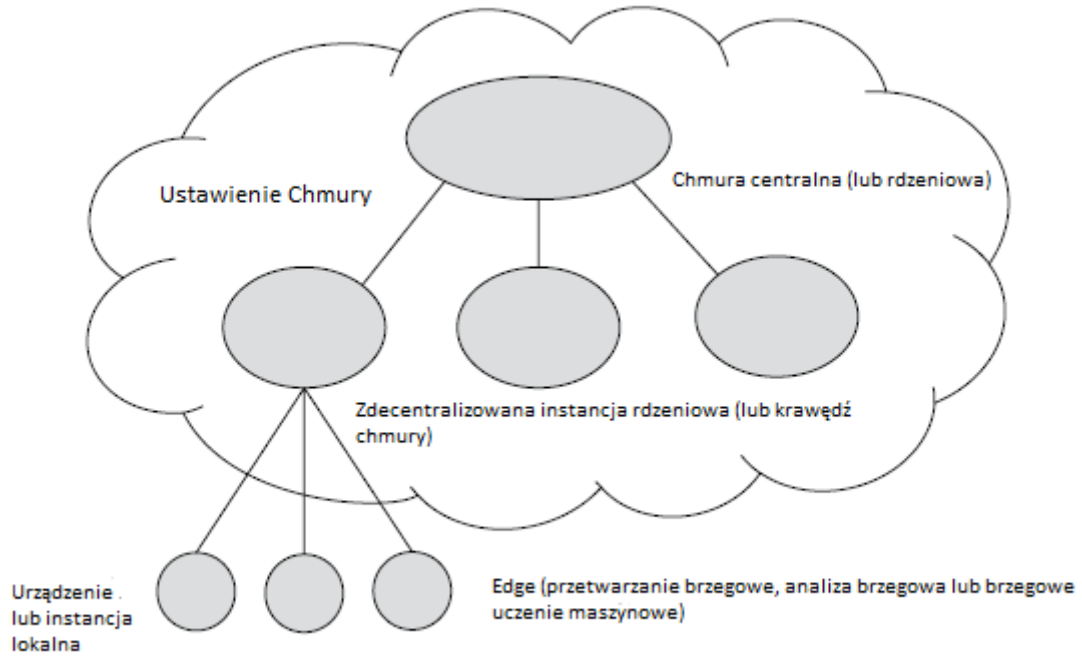
Architektury danych oparte na chmurze

Coraz więcej firm odchodzi od inwestycji w lokalną infrastrukturę danych w kierunku zwirtualizowanych i opartych na chmurze architektur danych. Siłą napędową tego ruchu jest to, że tradycyjne środowiska danych odczuwają presję zwiększania ilości danych i nie są w stanie skalować się w górę iw dół, aby sprostać stale zmieniającym się wymaganiom. Infrastruktury lokalnej po prostu brakuje elastyczności umożliwiającej dynamiczną optymalizację i sprostanie wyzwaniom nowych cyfrowych wymagań biznesowych. Zmiana architektury tych tradycyjnych, lokalnych środowisk danych w celu zapewnienia większego dostępu i skalowalności zapewnia architektury platform danych, które bezproblemowo integrują dane i aplikacje z różnych źródeł. Wykorzystanie pojemności obliczeniowej i pamięci masowej w chmurze umożliwia dodanie elastycznej warstwy sztucznej inteligencji i narzędzi uczenia maszynowego jako najwyższej warstwy w architekturze, dzięki czemu można przyspieszyć wartość, którą można uzyskać z dużych ilości danych.

Obliczenia i inteligencja na krawędzi

Edge computing opisuje architekturę obliczeniową, w której przetwarzanie danych odbywa się bliżej miejsca, w którym dane są tworzone – na przykład urządzeń Internetu rzeczy (IoT), takich jak podłączony bagaż, drony i połączone pojazdy, takie jak samochody i rowery. Istnieje różnica między wypychaniem obliczeń na brzeg (obliczanie brzegowe) a wypychaniem analiz lub uczenia maszynowego na brzeg (analiza brzegowa lub uczenie maszynowe). Obliczenia brzegowe mogą być

wykonywane jako oddzielne zadanie na brzegu, co pozwala na wstępne przetwarzanie danych w sposób rozproszony przed ich zebraniem i przesłaniem do centralnego lub częściowo scentralizowanego środowiska, w którym w celu uzyskania wglądu stosowane są metody analityczne lub technologie uczenia maszynowego/sztucznej inteligencji . Pamiętaj tylko, że prowadzenie analiz i uczenia maszynowego na brzegu wymaga pewnej formy obliczeń brzegowych, aby umożliwić wgląd i działanie bezpośrednio na brzegu. Przyczyna tendencji do wykonywania większej liczby operacji na brzegu sieci zależy głównie od takich czynników, jak ograniczenia łączności, przypadki użycia o niskim opóźnieniu, w których do przeprowadzenia natychmiastowej analizy i podjęcia decyzji potrzebne są milisekundy czasu reakcji (w przypadku samochodów autonomicznych, na przykład). Ostatnim powodem wykonywania większej liczby operacji na brzegu są ograniczenia przepustowości związane z przesyłaniem danych do centralnego punktu w celu analizy. Ze strategicznego punktu widzenia, przetwarzanie na brzegu sieci jest ważnym aspektem do rozważenia z perspektywy projektowania infrastruktury, szczególnie w przypadku firm z istotnymi elementami IoT. Jeśli chodzi o projektowanie infrastruktury, warto również zastanowić się, jak rozwiązania obliczeniowe i analityczne na brzegu sieci będą współpracować ze scentralizowaną (zwykle opartą na chmurze) architekturą. Wielu postrzega chmurę i brzeg jako konkurencyjne podejścia, ale chmura to styl przetwarzania, w którym elastycznie skalowalne możliwości technologiczne są dostarczane jako usługa, oferując środowisko wspierające dla części brzegowej infrastruktury. Nie wszystko jednak da się rozwiązać na krawędzi; wiele przypadków użycia i potrzeb dotyczy całego systemu lub sieci i dlatego do przeprowadzenia analizy potrzebna jest agregacja wyższego poziomu. Samo wykonanie analizy na krawędzi może nie dać wystarczającego kontekstu do podjęcia właściwej decyzji. Tego typu wyzwania obliczeniowe i spostrzeżenia najlepiej rozwiązywać za pomocą scentralizowanego modelu opartego na chmurze, jak pokazano na rysunku

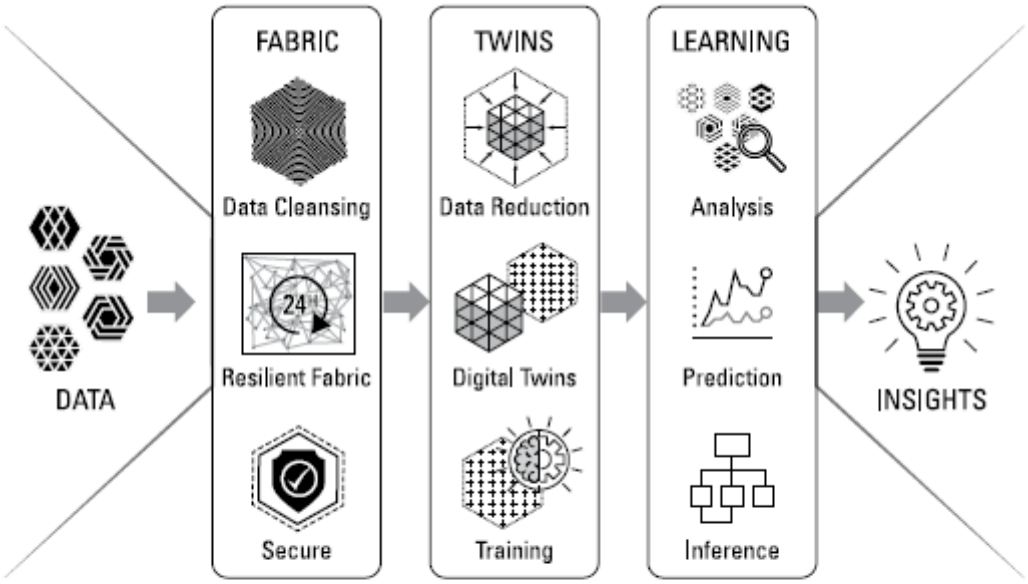


Jak widać, konfigurację chmury można również wykonać w sposób zdecentralizowany, a te zdecentralizowane instancje są określane jako krawędź chmury. W przypadku większej konfiguracji na skalę regionalną lub globalną, zdecentralizowany model może być wykorzystywany do obsługi implementacji brzegowych na poziomie urządzeń IoT w określonym kraju lub do wspierania operatora

telekomunikacyjnego w jego staraniach o włączenie wszystkich podłączonych urządzeń do sieci. Jest to przydatne, aby utrzymać krótki czas odpowiedzi i nie przenosić nieprzetworzonych danych poza granice krajów.

Cyfrowe bliźniaki

Cyfrowy bliźniak odnosi się do cyfrowej reprezentacji podmiotu lub systemu w świecie rzeczywistym – na przykład cyfrowego widoku miejskiej sieci telekomunikacyjnej zbudowanej na podstawie rzeczywistych danych. Cyfrowe bliźniaki w kontekście projektów IoT to obiecujący obszar, który obecnie prowadzi zainteresowanie cyfrowymi bliźniakami. Najprawdopodobniej jest to obszar, który znacznie się rozwinie w ciągu najbliższych trzech do pięciu lat. Dobrze zaprojektowane cyfrowe bliźniaki to zasoby, które mogą znacznie poprawić kontrolę nad przedsiębiorstwem i podejmowanie decyzji w przyszłości. Cyfrowi bliźniacy integrują sztuczną inteligencję, uczenie maszynowe i analizy z danymi, aby tworzyć żywe cyfrowe modele symulacji, które aktualizują się i zmieniają wraz ze zmianą ich fizycznych odpowiedników. Cyfrowy bliźniak nieustannie uczy się i aktualizuje z wielu źródeł, aby reprezentować swój status, warunki pracy lub pozycję w czasie zbliżonym do rzeczywistego. (Zobacz rysunek , aby zapoznać się z przeglądem tego procesu.)

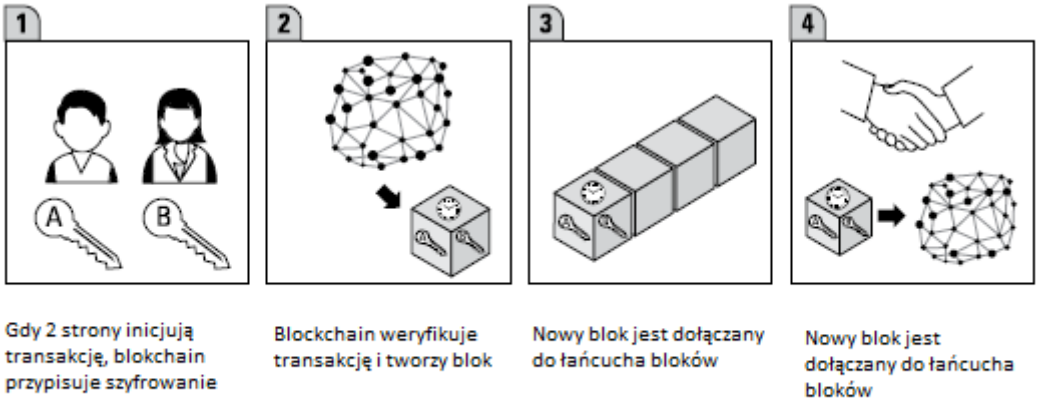


Cyfrowe bliźniaki są połączone ze swoimi odpowiednikami w świecie rzeczywistym i służą do zrozumienia stanu systemu, reagowania na zmiany, ulepszania operacji i dodawania wartości. Cyfrowi bliźniacy zaczynają od prostych cyfrowych widoków rzeczywistego systemu, a następnie ewoluują z czasem, poprawiając ich zdolność do gromadzenia i wizualizacji właściwych danych, stosowania odpowiednich analiz i reguł oraz reagowania w sposób, który sprzyja realizacji celów biznesowych Twojej organizacji. Ale możesz również użyć cyfrowego bliźniaka do uruchomienia modeli predykcyjnych lub symulacji, które można wykorzystać do znalezienia pewnych wzorców w danych tworzących cyfrowego bliźniaka, które mogą prowadzić do problemów. Te spostrzeżenia można następnie wykorzystać do aktywnego zapobiegania problemom. Dodanie zautomatyzowanych możliwości podejmowania decyzji w oparciu o koncepcję „cyfrowych bliźniaków” wstępnie

zdefiniowanych i wstępnie zatwierdzonych zasad byłoby świetną możliwością dodania do dowolnej perspektywy operacyjnej – na przykład zarządzania systemem IoT, takim jak inteligentne miasto.

Blockchain

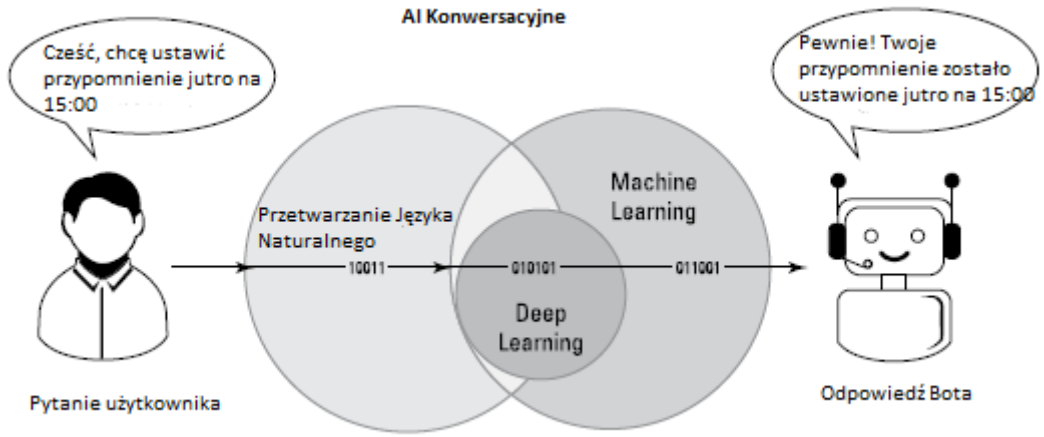
Koncepcja blockchain ewoluowała z infrastruktury walut cyfrowych w platformę do transakcji cyfrowych. Blockchain to rosnąca lista rekordów (bloków), które są połączone za pomocą kryptografii. Każdy blok zawiera skrót kryptograficzny poprzedniego bloku, znacznik czasu i dane transakcji. Z założenia blockchain jest odporny na modyfikację danych. Jest to otwarta i publiczna księga, która może skutecznie rejestrować transakcje między dwiema stronami w sposób weryfikowalny i trwały. Blockchain to również zdecentralizowana i rozproszona cyfrowa księga, która służy do rejestrowania transakcji na wielu komputerach, dzięki czemu żaden zapis nie może zostać zmieniony wstecznie bez zmiany wszystkich kolejnych bloków. Technologie blockchain oferują znaczący krok od obecnych scentralizowanych mechanizmów opartych na transakcjach i mogą stanowić podstawę nowych cyfrowych modeli biznesowych zarówno dla przedsiębiorstw o ugruntowanej pozycji, jak i start-upów. Rysunek pokazuje, jak wykorzystać blockchain do przeprowadzenia transakcji blockchain.



Chociaż szum wokół blockchainów początkowo koncentrował się na branży usług finansowych, blockchainy mają wiele potencjalnych obszarów zastosowania, w tym rząd, opiekę zdrowotną, produkcję, weryfikację tożsamości i łańcuch dostaw. Chociaż blockchain ma długoterminową obietnicę i niewątpliwie spowoduje zakłócenia, jej obietnica nie została jeszcze udowodniona w rzeczywistości: wiele powiązanych technologii jest zbyt niedojrzałych, aby można je było wykorzystać w środowisku produkcyjnym i tak pozostanie przez następne dwa do trzech lat.

Platformy konwersacyjne

Konwersacyjna sztuczna inteligencja to forma sztucznej inteligencji, która pozwala ludziom komunikować się z aplikacjami, witrynami i urządzeniami w codziennym, ludzkim języku naturalnym za pomocą głosu, tekstu, dotyku lub gestów. Użytkownikom umożliwia szybką interakcję przy użyciu własnych słów i terminologii. Dla przedsiębiorstw oferuje sposób na budowanie bliższego kontaktu z klientami poprzez spersonalizowaną interakcję i otrzymywanie w zamian ogromnej ilości ważnych informacji biznesowych. Rysunek przedstawia interakcję między człowiekiem a botem.



Ten rodzaj platformy najprawdopodobniej doprowadzi do kolejnej zmiany paradygmatu w sposobie interakcji ludzi ze światem cyfrowym. Odpowiedzialność za tłumaczenie zmian intencji z ludzi na maszyny. Platforma przyjmuje pytanie lub polecenie od użytkownika, a następnie odpowiada, wykonując jakąś funkcję, prezentując jakąś treść lub prosząc o dodatkowe dane wejściowe. W ciągu najbliższych kilku lat interfejsy konwersacyjne staną się głównym celem projektowania dla interakcji użytkownika i będzie dostarczany na dedykowanym sprzęcie, z podstawowymi funkcjami systemu operacyjnego, platformami i aplikacjami. Sprawdź poniższą listę niektórych potencjalnych obszarów, w których można skorzystać z zastosowania platform konwersacyjnych za pomocą botów:

* Informacyjne: Chatboty, które pomagają w badaniach, prośbach o informacje i prośbach o status różnych typów

* Produktywność: boty, które mogą łączyć klientów z usługami handlowymi, pomocniczymi, doradczymi lub konsultacyjnymi

* B2E (business-to-employee): boty umożliwiające pracownikom dostęp do danych, aplikacji, zasobów i działań

* Internet rzeczy (IoT): Boty, które umożliwiają interfejsy konwersacyjne dla różnych interakcji urządzeń, takich jak drony, urządzenia, pojazdy i wyświetlacze

Korzystając z tych różnych typów platform konwersacyjnych, możesz spodziewać się zwiększonej produktywności botów (ponieważ mogą skoncentrować się na najcenniejszych interakcjach), zautomatyzowanej siły roboczej 24/7, zwiększonej lojalności i satysfakcji klientów, nowego wglądu w interakcje z klientami i zmniejszonych kosztów operacyjnych. Platformy konwersacyjne osiągnęły teraz punkt zwrotny pod względem zrozumienia języka i podstawowych intencji użytkownika, ale nadal nie są wystarczająco dobre, aby w pełni wystartować. Wyzwanie, przed którym stoją platformy konwersacyjne, polega na tym, że użytkownicy muszą komunikować się w ustrukturyzowany sposób, a w prawdziwym życiu jest to często frustrujące doświadczenie. Podstawowym wyróżnikiem wśród platform konwersacyjnych jest niezawodność ich modeli oraz interfejsów programowania aplikacji (API) i modeli zdarzeń używanych do uzyskiwania dostępu, przyciągania i organizowania usług innych firm w celu dostarczania złożonych wyników.

Opracowanie niektórych przyszłych scenariuszy

Chociaż wokół nas ma miejsce eksplozja nowych przypadków użycia (tych specyficznych sytuacji, w których nauka o danych może być potencjalnie wykorzystana) i aplikacji w nauce o danych, wciąż

istnieją scenariusze, które mają dopiero nadejść. W tej sekcji opiszę niektóre potencjalne przyszłe scenariusze w przestrzeni data science, w tym wyzwania i motywacje.

Standaryzacja pod kątem produktywności w badaniach danych

Standaryzacja generalnie zapewnia sprawne działanie procesów i buduje wiarygodność w czasie. Najlepsze praktyki zapewniają wydajność i redundancję. Obecnie ilość danych generowanych na świecie rośnie z sekundy na sekundę. Ponieważ te dane są gromadzone i przechowywane, kluczowe znaczenie ma ich standaryzacja i normalizacja w celu optymalnego wykorzystania. W przeciwnym razie może być bardzo głośno. Jednym z największych wyzwań w budowaniu zoptymalizowanego rozwiązania do zarządzania danymi jest brak standaryzacji podczas zbierania danych z całego Internetu – a nawet z różnych części dużej, globalnej firmy. Standaryzacja ma kluczowe znaczenie, gdy próbujemy uniknąć nadmiarowości i zwiększyć dokładność dopasowywania typów danych. Chociaż jest to konieczne, nadal jest to trudny problem do rozwiązania. Brak standaryzacji okazuje się być przeszkodą dla wielu systemów biznesowych. Wszystkie dane należy przekonwertować do wstępnie zdefiniowanego formatu, co wymaga wiedzy specjalistycznej w danej dziedzinie, a także uzgodnienia (zarówno wewnętrznie, jak i zewnętrznie) definicji i struktury danych. Czy są więc powody, dla których standaryzacja w przestrzeni ML/AI jest szczególnie ważna? Cieszę się, że pytałeś. Poniższa lista przedstawia niektóre powody, dla których jest to szczególnie potrzebne:

* Interoperacyjność modelu: standardy interoperacyjności są ważne nie tylko dla globalnej społeczności, ale także dla każdej firmy wdrażającej uczenie maszynowe i sztuczną inteligencję. Powodem jest to, że w celu skalowania rozwoju algorytmów przez firmę potrzebna jest interoperacyjność między modelami w produkcji, aby zapewnić nie tylko możliwość współpracy modeli, ale także maksymalizację wydajności modelu w środowisku wielomodelowym.

* Standaryzacja procesu: nacisk na kontrolę procesu stawia również pod znakiem zapytania, które standaryzacji są potrzebne w odniesieniu do bezpieczeństwa, wydajności, opóźnień, niezawodności, stronniczości, a nawet prywatności. Dzięki standaryzacji najlepszych praktyk opracowywania, na przykład, złożonych technik, takich jak głębokie uczenie, nie tylko więcej zespołów może przyspieszyć ich rozwój, ale także bardziej innowacyjne rozwiązania mogą być opracowywane niezależnie i być podłączane w celu przyspieszenia znacznie większego procesu.

* Kompatybilność człowiek-maszyna: tutaj kompatybilność odnosi się do standaryzacji interakcji między człowiekiem a maszyną – gdy człowiek używa pewnych słów lub fraz, na przykład maszyna używa ustandaryzowanego zestawu odpowiedzi lub działań. Wiele błędów może wystąpić z powodu różnych celów i metod interakcji. W przypadku systemów o znaczeniu krytycznym wyobraźmy sobie zarządzaną maszynowo wieżę kontroli ruchu lotniczego; może to być katastrofa, jeśli interakcja nie jest ustandaryzowana między systemami lub przynajmniej niewygodna, aby nauczyć się wszystkich różnych wersji tego, jak osiągnąć cel we współpracy z maszyną, w zależności od tego, która implementacja jest używana i gdzie.

* Etyka: Wyzwania związane ze standaryzacją sztucznej inteligencji obejmują wiele poziomów obaw, ale ten jest kluczowy – każda forma standaryzacji sztucznej inteligencji powinna obejmować metody, jak najlepiej napędzać sztuczną inteligencję z maksymalnymi korzyściami dla ludzkości. Całkowitą porażką byłoby, gdyby standaryzacja prowadziła do bardziej zaawansowanej autonomicznej broni lub bardziej udoskonalonych metod przewidywania i manipulowania ludzkim zachowaniem.

Od scenariuszy monetyzacji danych do gospodarki opartej na danych

Ogromny postęp technologiczny doprowadził do eksplozji tempa tworzenia nowych danych. Gospodarka oparta na danych zapewnia to, czego chcą firmy i rządy na całym świecie: tworzenie wysokiej jakości miejsc pracy, generowanie wzrostu gospodarczego i umożliwianie organizacjom we wszystkich sektorach pomyślnego rozwoju i obsługi klientów. Ale czy wraz ze wzrostem ilości danych liderzy firm zdają sobie sprawę z jej prawdziwego potencjału? Wraz z pojawieniem się gospodarki opartej na danych, zmiany oczekiwań klientów i postęp technologiczny przekształcą łańcuchy dostaw w złożone ekosystemy. Strategie produkcyjne ulegną zmianie, a współpraca między organizacjami i ekosystemami stworzy bardziej otwarty przepływ informacji i pomysłów. Firmy będą musiały wymyślić się na nowo, określając swoje pożądane role w gospodarce opartej na danych w drodze oceny ich zaangażowania w te ekosystemy. . Pozwoli to organizacjom ocenić, czy potrzebne będą nowe jednostki biznesowe, wspólne przedsięwzięcia i przejęcia.

Eksplozja hybrydowych systemów człowiek/maszyna

Hybrydowe systemy człowiek/maszyna łączą inteligencję maszynową i ludzką, aby przezwyciężyć wady istniejących systemów sztucznej inteligencji. Konieczność zaangażowania człowieka w przezwyciężenie błędów i ograniczeń systemów sztucznej inteligencji jest już uznana w krytycznych dziedzinach, takich jak medycyna i prowadzenie pojazdów. (Oczekuje się, że kierowca półautonomicznego samochodu będzie stale czuwał nad decyzjami maszyny i korygował je w razie potrzeby, aby na przykład zapobiegać wypadkom). Jednak pomyślna integracja inteligencji człowieka i maszyny ma swoje wyzwania. Inteligencja ludzka jest cennym zasobem związanym z wyższymi kosztami i ograniczeniami, takimi jak na przykład 8-godzinny dzień pracy. Jakość i dostępność informacji wprowadzanych przez człowieka może się również różnić w zależności od innych czynników, w tym stanu człowieka, takiego jak choroba lub zmęczenie. Jednym ze sposobów przezwyciężenia wyzwań związanych z hybrydowymi systemami człowiek/maszyna jest zmiana sposobu, w jaki maszyny uzyskują dostęp do ludzkiej inteligencji. Aby tak się stało, systemy sztucznej inteligencji musiałyby być wyposażone w zdolności rozumowania, które mogą podejmować skuteczne decyzje dotyczące tego, w jaki sposób powinny uzyskać dostęp do ludzkiej inteligencji. W tym przypadku ostatnie postępy w dziedzinie obliczeń ludzkich mogą dostarczyć pewnych wskazówek na temat tego, jak systemy AI mogą to osiągnąć. Platformy crowdsourcingowe zapewniają łatwy dostęp do ludzkiej inteligencji na żądanie w skalowalny i elastyczny sposób. Mówiąc najprościej, crowdsourcing ma miejsce, gdy firma lub instytucja zleca funkcję, którą kiedyś pełniła ograniczona liczba pracowników, do nieokreślonej (i generalnie dużej) sieci osób w formie otwartego zaproszenia. Takie podejście może przybrać formę wzajemnej produkcji (kiedy praca jest wykonywana wspólnie), ale często jest również podejmowane przez pojedyncze osoby. W przypadku systemów AI, w których użytkownik nie jest zaangażowany w udzielanie pomocy, potrzebna przez system pomoc ludzka może być zapewniona przez tłum. Dla wielu prac badawczych, w tym tych przedstawionych tutaj, platformy crowdsourcingowe funkcjonują jako stanowiska testowe do zbierania danych i eksperymentowania związanego z wyzwaniami dostępu i pracy z ludzką inteligencją.

Obliczenia kwantowe rozwiążą nierozwiązywalne problemy

Jeśli spędzasz więcej niż pięć minut w Internecie, oglądając wiadomości i w inny sposób pozostając na bieżąco ze światem, słyszałeś ekscytację towarzyszącą ostatnim postępom w rozwoju systemów komputerów kwantowych. To nie przesada – to naprawdę wszystko zmieni. Komputery kwantowe mają potencjał, by przebić się przez przeszkody, które ograniczają moc klasycznych komputerów, rozwiązując problemy w ciągu kilku sekund, których rozwiązanie zajęłoby klasycznemu komputerowi całe życie Wszechświata - na przykład szyfrowanie i badania nad nową, zaawansowaną medycyną . Kiedy chemicy badają nowe leki, większość ich pracy polega na testowaniu setek możliwych zmiennych w formule chemicznej w celu znalezienia pożądanych cech potrzebnych do leczenia różnych chorób. Ten proces eksperymentowania i odkrywania często prowadzi do powstania ponad 10 lat, zanim nowy

lek zostanie wprowadzony na rynek – często kosztem miliardów dolarów. Obecnie obliczenia są wykonywane na komputerach, które muszą łączyć i ponownie łączyć elementy, aby przetestować wyniki.

Nie trzeba dodawać, że trwa wyścig, aby komputery kwantowe stały się praktycznymi narzędziami codziennego użytku dla biznesu, przemysłu i nauki w celu uzyskania przewagi konkurencyjnej. Przetwarzanie kwantowe jest tutaj, aby pozostać, rośnie, a jeśli nie rozwiąże wszystkich problemów świata, może potencjalnie rozwiązać wiele komputerów kwantowych różniących się jakościowo od standardowych komputerów pod względem sposobu obliczania danych. Z jednej strony masz standardowy binarny cyfrowy komputer elektroniczny, w którym dane muszą być zakodowane w cyfry binarne (bity), z których każda jest zawsze w jednym z dwóch określonych stanów (0 lub 1). Obliczenia kwantowe wykorzystują bity kwantowe (kubity), które mogą znajdować się w superpozycji stanów – to znaczy, tak jak kot Schrödingera może być zarówno żywy, jak i martwy, kubit może mieć zarówno 0, jak i 1.

Znając swoje dane

Właściwe podejście do strategii dotyczącej danych ma fundamentalne znaczenie dla zapewnienia stabilnej podstawy dla reszty inwestycji w naukę danych. . I nie chodzi tylko o zabezpieczenie integralności danych; musisz również upewnić się, że typy danych wybrane do celów biznesowych są właściwe i zostały wybrane z właściwych powodów. Aby tak się stało, musisz zrozumieć dane, na które kierujesz reklamy. Aby to zrozumieć, musisz pomyślnie wykonać cztery główne kroki: wybrać dane, opisać dane, zbadać dane i ocenić jakość danych.

Wybór Twoich danych

Selekcja danych to proces określania odpowiedniego typu i źródła danych – a także odpowiednich metod – do zbierania danych. Wybór danych poprzedza właściwe zadanie zbierania danych. Głównym celem selekcji danych jest określenie odpowiedniego typu danych, źródła i metody niezbędnych do udzielenia odpowiedzi na zadane pytania. Wybór jest często powiązany z określonym obszarem - na przykład finansami, sprzedażą, produktem lub konsumentem - i jest głównie uzależniony od rodzaju analizy, której zamierzasz użyć, a także od możliwości uzyskania dostępu do niezbędnych źródeł danych. Problemy z integralnością mogą pojawić się, gdy decyzje o wyborze odpowiednich danych do zebrania opierają się przede wszystkim na rozważaniach dotyczących kosztów i wygody, a nie na zdolności danych do skutecznej odpowiedzi na postawione pytania. Z pewnością koszt i wygoda są ważnymi czynnikami w procesie podejmowania decyzji. Musisz jednak ocenić, w jakim stopniu te czynniki mogą wpłynąć na integralność analizy. W tej pierwszej części procesu wyboru danych rozważ swoje odpowiedzi na te pytania:

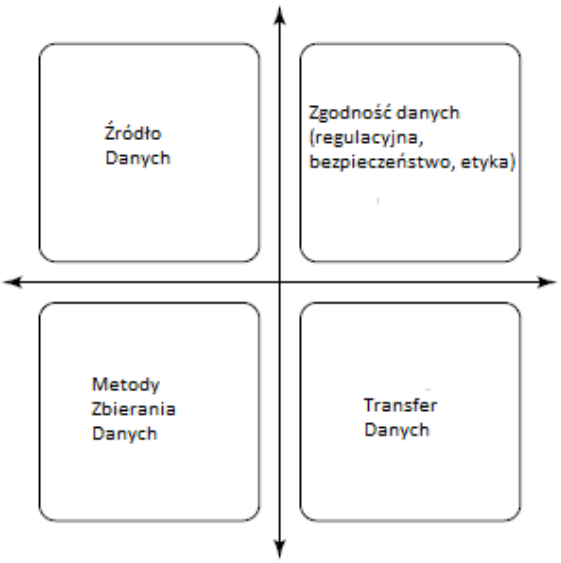
* Na jakie pytania próbujesz odpowiedzieć?

* Jaki jest zakres analizy?

* W dziedzinie, którą zamierzasz przeanalizować, jakiego rodzaju dane są zazwyczaj kierowane przez branżę?

* Jaki format danych jest potrzebny, aby odpowiedzieć na pytania biznesowe: ilościowe, jakościowe czy oba?

Rysunek przedstawia główne obszary zainteresowania związane z gromadzeniem danych.



W ramach procesu zbierania danych rozważ swoje odpowiedzi na następujące pytania:

* Gdzie są źródła danych, z których musisz korzystać? Czy istnieją, czy musisz stworzyć nowe dane (np. za pomocą ankiety)? Jeśli dane istnieją, masz je już w swojej firmie, czy musisz je pozyskać?

* Czy istnieją jakieś ograniczenia prawne, etyczne lub związane z bezpieczeństwem związane z potrzebnymi danymi, takie jak prywatność danych, własność danych lub okresy przechowywania danych?

* Jak musisz zbierać dane? Czy częste przesyłanie wsadowe jest wystarczające, czy musisz przesyłać strumieniowo dane na żywo? Czy Twoja infrastruktura będzie w stanie zarządzać różnymi sposobami pozyskiwania danych?

* Czy istnieje potrzeba przenoszenia danych poza granice kraju? Czy istnieją ograniczenia prawne związane z konkretnymi danymi, na które kierujesz reklamy? Jeśli takie ograniczenia istnieją, jak zostaną rozwiązane?

Opisywanie danych

Opisywanie danych odnosi się do zadania polegającego na zbadaniu i udokumentowaniu zebranych danych, tak aby elementy takie jak format danych, ilość danych i metadane mogły być odnotowane i zarejestrowane. Element metadanych to typ danych używany do opisu innych danych w celu zwiększenia użyteczności oryginalnych danych. Tworzenie i utrzymywanie metadanych jest istotną częścią środowiska nauki o danych. Pamiętaj, że podczas konfigurowania środowiska nauki o danych napotkasz wiele różnych typów metadanych. Krótko opiszę różne rodzaje na tej liście:

* Opisowe: używane do opisania głównych cech elementu danych w celu odkrycia i identyfikacji. Może zawierać takie elementy, jak tytuł, streszczenie, autor i słowa kluczowe.

* Strukturalny: zajmuje się metadanymi dotyczącymi grup danych i wskazuje, jak wiele obiektów jest ze sobą połączonych - na przykład jak strony są uporządkowane, tworząc rozdziały. Opisuje kategorie danych, wersje, relacje i inne cechy materiałów cyfrowych.

* Administracyjne: Zawiera informacje pomocne w zarządzaniu elementem danych, takie jak czas i sposób jego utworzenia, typ pliku i inne informacje techniczne, a także kto ma do niego dostęp.

* Odniesienie: opisuje zawartość i jakość danych statystycznych.

* Statystyczne: opisuje procesy, które zbierają, przetwarzają lub generują statystyki danych (nazywane również danymi procesowymi).

Opisanie danych jest strategicznie ważnym zadaniem w celu oceny, czy zebrane dane spełniają zidentyfikowane wymagania biznesowe — i obejmuje kilka odrębnych kroków. Ta lista zawiera przegląd działań, które należy wykonać:

* Przeanalizuj ilość danych i spróbuj oszacować poziom złożoności.

* Opisz różne potrzebne tabele i ich wzajemne relacje. Tabele danych pomagają uporządkować informacje. Jeśli zbierasz dane z eksperymentu lub badań naukowych, zapisanie ich w tabeli danych ułatwi późniejsze wyszukiwanie. Tabele danych mogą również pomóc w tworzeniu wykresów i innych wykresów na podstawie informacji.

* Sprawdź dostępność atrybutów, co pomaga opisać kontekst każdego typu danych – np. lokalizację geograficzną, z której zostały zebrane lub datę ich zebrania.

* Określ, czy potrzebne są różne typy atrybutów. Typy mogą obejmować:

-Nominalne: numery identyfikacyjne, kolor oczu, kody pocztowe

- Porządkowe: Rankingi (np. smak chipsów ziemniaczanych w skali od 1-10), oceny, wzrost w kategoriach (wysokie, średnie, niskie)

- Interwał: daty kalendarza, temperatury w stopniach Celsjusza lub Fahrenheita

- Ratio: Dokładna temperatura w kelwinach, długość, czas, liczby. Zastanów się, w jaki sposób zamierzasz wykorzystać dane, aby określić, które atrybuty będą wymagane.

* Opisz zakres wartości wybranych atrybutów (jeśli dotyczy). Pamiętaj, że ten sam atrybut można odwzorować na różne wartości atrybutów. (Wysokość można na przykład zmierzyć w stopach lub metrach.)

* Przeanalizuj potencjalne korelacje atrybutów, takie jak płeć i wzrost lub data i temperatura.

* Zrozum znaczenie każdego atrybutu i opisz wartość w kategoriach biznesowych.

* Dla każdego atrybutu oblicz podstawowe statystyki (na przykład rozkład, średnia, maksymalna, minimalna, odchylenie standardowe, wariancja, tryb i skośność) i powiąż wyniki z ich znaczeniem biznesowym.

* Zdecyduj o trafności atrybutu związanej z konkretnym celem biznesowym, angażując ekspertów dziedzinowych.

* Określ, czy znaczenie każdego atrybutu jest używane konsekwentnie.

* Zdecyduj, czy konieczne jest zrównoważenie danych, jeśli dystrybucja danych jest zniekształcona.

* Analizuj i dokumentuj kluczowe relacje w danych.

Bez właściwości danych dołączonych do Twoich danych, wartość i użyteczność Twoich danych ulegną znacznemu zmniejszeniu. Dokumentacja właściwości danych jest podstawą dalszej analizy danych i

musi być aktualna w ramach działań związanych z zarządzaniem danymi, o ile zamierzasz w jakikolwiek sposób wykorzystać dane.

Eksploracja danych

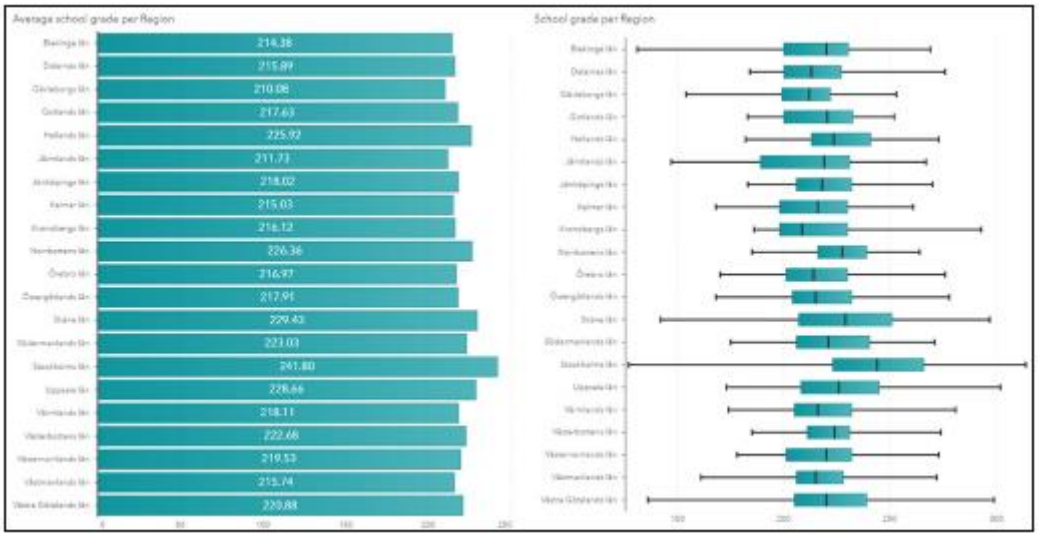
Ważnym krokiem w lepszym poznaniu danych jest ich eksploracja. Eksploracja danych to podejście podobne do wstępnej analizy danych, w której analityk danych wykorzystuje eksplorację wizualną, aby zrozumieć, co znajduje się w zestawie danych, a także cechy danych. Cechy te mogą obejmować na przykład rozmiar lub ilość danych, kompletność danych, poprawność danych lub możliwe relacje lub spostrzeżenia, które mogą być ukryte w danych. Wizualna eksploracja danych to czynność polegająca na wyszukiwaniu i dowiadywaniu się więcej o danych przy użyciu różnych modeli statystycznych wizualizowanych poprzez graficzną reprezentację danych - na przykład mapy cieplne, mapy geograficzne, wykresy pudełkowe i chmury słów. Wystarczy spojrzeć na ten sam zestaw danych, używając różnych reprezentacji graficznych, aby wykryć korelacje i zależności danych, a także nowe wglądy w dane. Eksploracja danych jest zwykle prowadzona przy użyciu kombinacji tych rodzajów metod:

* Zautomatyzowane: Może obejmować profilowanie danych lub wizualizację danych, aby dać analitykowi wstępny widok danych, a także zrozumienie kluczowych cech.

* Ręczny: często następuje automatyczne działanie poprzez ręczne drążenie lub filtrowanie danych w celu identyfikacji anomalii lub wzorców zidentyfikowanych automatycznie. Eksploracja danych zwykle wymaga również ręcznego tworzenia skryptów i zapytań do danych (na przykład przy użyciu języków takich jak Python lub R) lub używania programu Excel (w przypadku mniejszych zestawów danych) lub podobnych narzędzi do przeglądania nieprzetworzonych danych.

Rzeczywiste zadanie eksploracji danych to półautomatyczna lub automatyczna analiza dużych ilości danych w celu wyodrębnienia wcześniej nieznanych i/lub interesujących wzorców, takich jak grupy rekordów danych (analiza skupień), nietypowe rekordy (wykrywanie anomalii) i zależności (eksploracja reguł asocjacji , sekwencyjne wydobywanie wzorców).

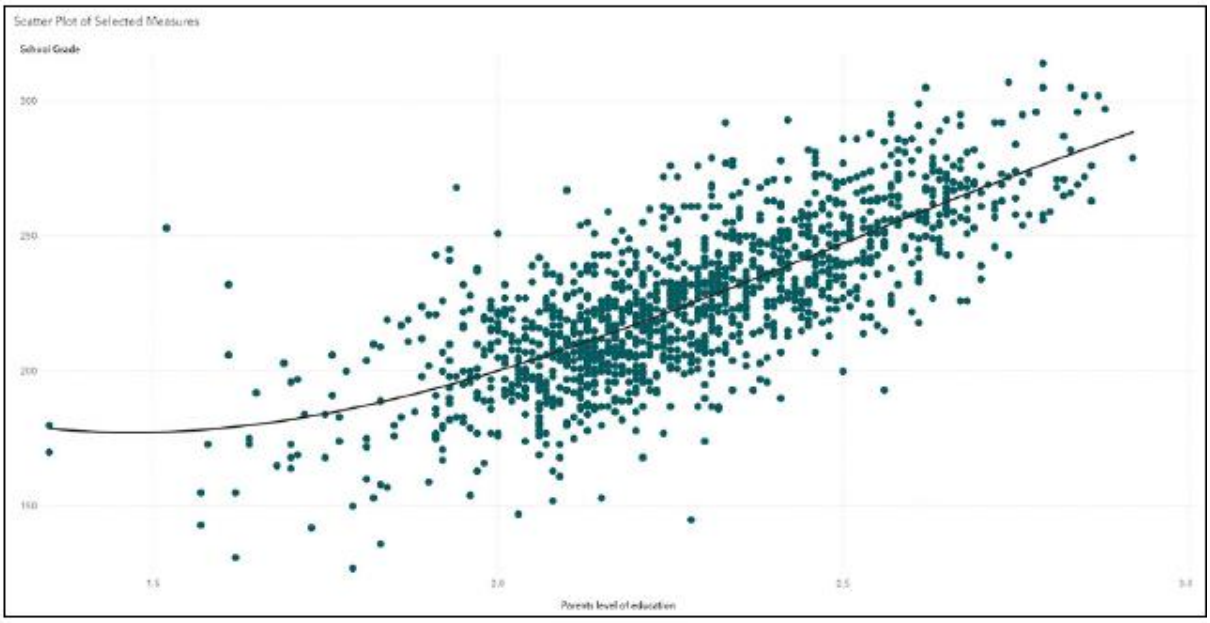
Rysunki 3, 4, 5 i 6 przedstawiają różne sposoby eksploracji danych. Rysunek 3 zaczyna się od przyjrzenia się ocenom szkolnym w Szwecji.



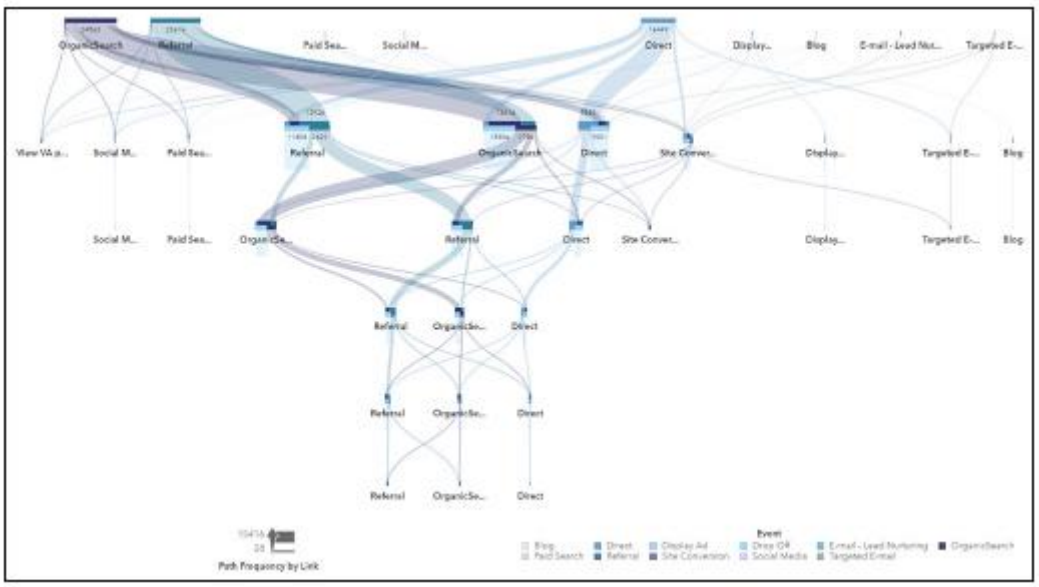
Wykres słupkowy po lewej przedstawia klasyczną średnią wartość na region; z tego punktu widzenia najlepiej wypada Sztokholm. Nie można jednak ufać średniej wartości na wykresie słupkowym.

Dlaczego? Bliższe przyjrzenie się regionowi Sztokholmu na wykresie pudełkowym pokazuje, że w tym obszarze występują problemy. Istnieje bardzo duża rozpiętość ocen – w rzeczywistości największa rozpiętość w Szwecji występuje w Sztokholmie – a kiedy przestudiujesz to bardziej szczegółowo, zobaczysz, że istnieje bardzo duża różnica między szkołami. Nazywa się to segregacją. Jeśli porównasz te liczby z regionem Norrbotten, zobaczysz znacznie mniejszy rozrzut ocen, co wskazuje na wyższą ogólną wydajność.

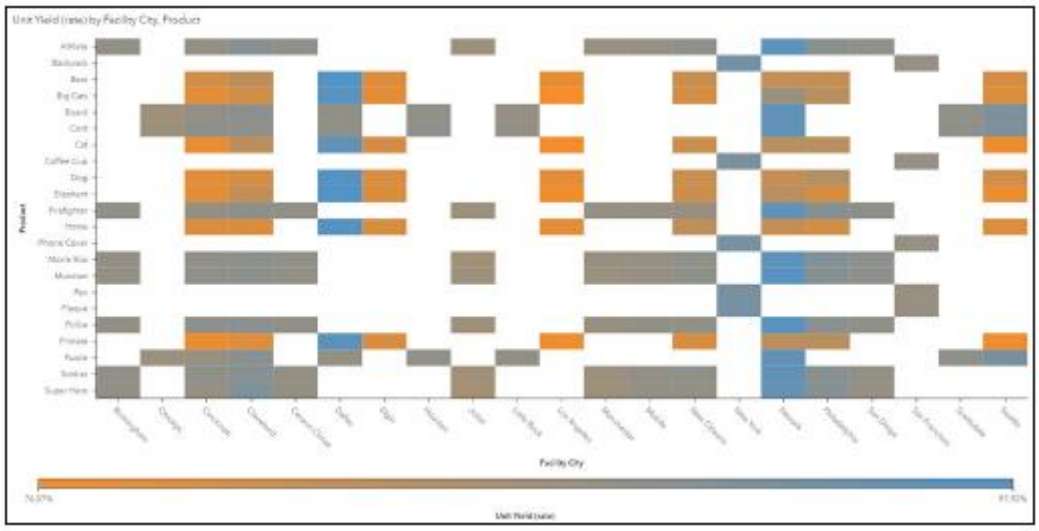
Na rysunku 4 widać przykład, w którym wykres rozrzutu służy do zrozumienia, czy poziom wykształcenia rodziców faktycznie wpływa na oceny w szkole. A jak można wyczytać z wykresu, zdecydowanie istnieje silna korelacja, na którą wskazuje rozkład wartości zgrupowanych wzdłuż ukośnej 45-stopniowej linii od lewego dolnego rogu do prawego górnego rogu.



Rysunek 5 wykorzystuje dane, aby pokazać, w jaki sposób odwiedzający witrynę poruszają się między różnymi częściami witryny. Dzięki danym można określić, gdzie większość ludzi zaczyna przeglądać (wyszukiwarki, odsyłacze, linki bezpośrednie i inne) oraz dokąd zmierzają z punktu wejścia. Uzyskanie lepszego zrozumienia tego typu wzorców może być pomocne w zrozumieniu wrażenia użytkownika i czy istnieje wzorzec, w którym mają tendencję do „odpadania” i opuszczania witryny internetowej.



Rysunek 6 przedstawia inny sposób eksploracji danych za pomocą mapy cieplnej — graficzną reprezentację danych, która wykorzystuje system kodowania kolorami do reprezentowania różnych wartości. Ten przykład analizuje dane dotyczące wydajności produkcji dla różnych produktów, aby znaleźć zależności lub korelację między określonym produktem a miastem, w którym jest wytwarzany. Wydajność produkcyjna odnosi się do jakości produktu z punktu widzenia częstotliwości wymiany produktu (poziom wydajności). Na mapie cieplnej na rysunku 6 widać, że kolorystyka sięga od jaśniejszego (co jest złe) do ciemniejszego (co jest dobre). Korzystając z mapy ciepła, możesz szybko uzyskać przegląd sytuacji i zobaczyć, które miasta mają problemy z produkcją niektórych produktów, na podstawie ich wskaźników wydajności.



Wszystkie te zadania eksploracyjne mają na celu stworzenie jaśniejszego widoku danych, dzięki czemu można lepiej poznać swoje dane, a wszystko to w nadziei uzyskania pierwszych spostrzeżeń na podstawie eksploracji, które możesz chcieć dalej analizować. Jest to kluczowy pierwszy krok analityka i jest tak samo ważny jak zdefiniowanie podstawowych metadanych (statystyki, struktury, relacji) dla zbioru danych, które można następnie wykorzystać w dalszej analizie.

Ocena jakości danych

Kolejna podstawowa część zrozumienia danych polega na jak najszybszym uzyskaniu szczegółowego obrazu jakości danych. Wiele firm uważa, że jakość danych i ich wpływ jest zbyt późno – dawno temu, kiedy mogło to mieć znaczący wpływ na powodzenie projektu. Integrując jakość danych z aplikacjami operacyjnymi, organizacje mogą uzgadniać różne dane, usuwać nieścisłości, standaryzować wspólne metryki i tworzyć strategiczne, wiarygodne i cenne zasoby danych, które usprawniają podejmowanie decyzji. Ponadto, jeśli wstępna analiza wskazuje, że jakość danych jest niewystarczająca, można podjąć kroki w celu wprowadzenia ulepszeń. Jednym ze sposobów udoskonalenia danych jest usunięcie nieużytecznych części; innym sposobem jest poprawienie źle sformatowanych części. Zacznij od zadania sobie pytań takich jak: Czy dane są kompletne? Czy obejmuje wszystkie wymagane przypadki? Czy jest poprawny, czy zawiera błędy? Jeśli występują błędy, jak powszechne są one? Czy w danych brakuje wartości? Jeśli tak, to w jaki sposób są przedstawiane, gdzie występują i jak często występują? Bardziej uporządkowana lista punktów kontrolnych jakości danych obejmuje następujące kroki:

* Sprawdź pokrycie danych (czy na przykład reprezentowane są wszystkie możliwe wartości).

* Weryfikacja, czy znaczenie atrybutów i dostępnych wartości pasują do siebie. Na przykład, jeśli analizujesz dane dotyczące lokalizacji geograficznej sklepów detalicznych, czy wartość jest ujęta w szerokości i długości geograficznej, a nie w nazwie obszaru regionalnego, w którym jest umieszczona?

* Identyfikowanie brakujących atrybutów i pustych pól.

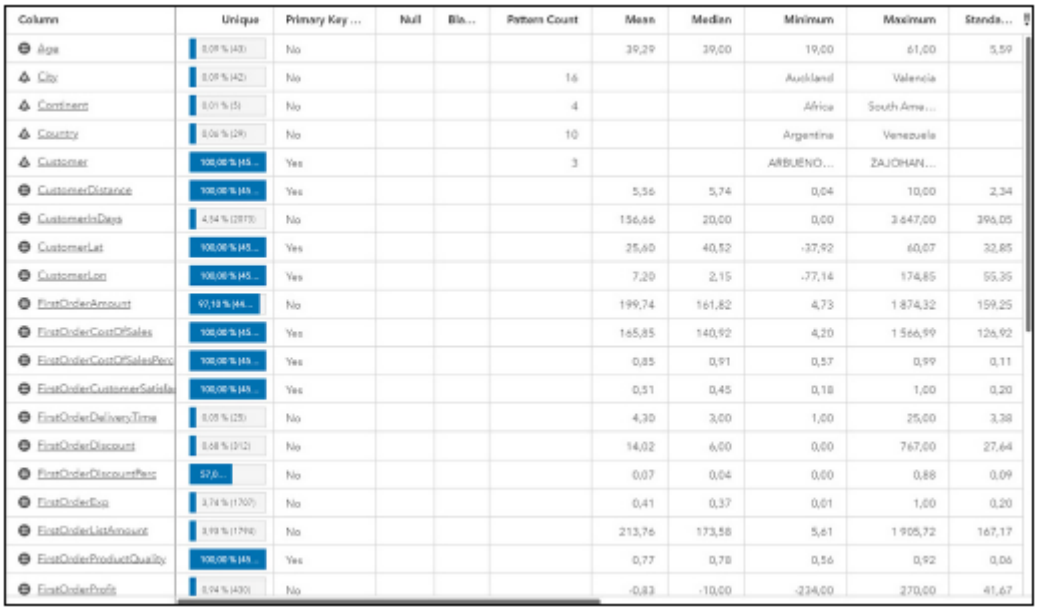
* Klasyfikowanie znaczenia brakujących lub błędnych danych i podwójne sprawdzanie atrybutów o różnych wartościach, ale podobnych znaczeniach.

* Sprawdzanie niespójności w pisowni i formatowaniu wartości (sytuacje, w których ta sama wartość czasami zaczyna się od małej litery, a czasami od dużej). Bez spójnych konwencji nazewnictwa i formatu liczbowego korelacja i analiza danych nie będą możliwe w różnych zestawach danych.

* Przeglądanie odchyleń i decydowanie, czy którekolwiek z nich kwalifikuje się jako zwykły szum (odstające) lub wskazuje na interesujące zjawisko (wzór).

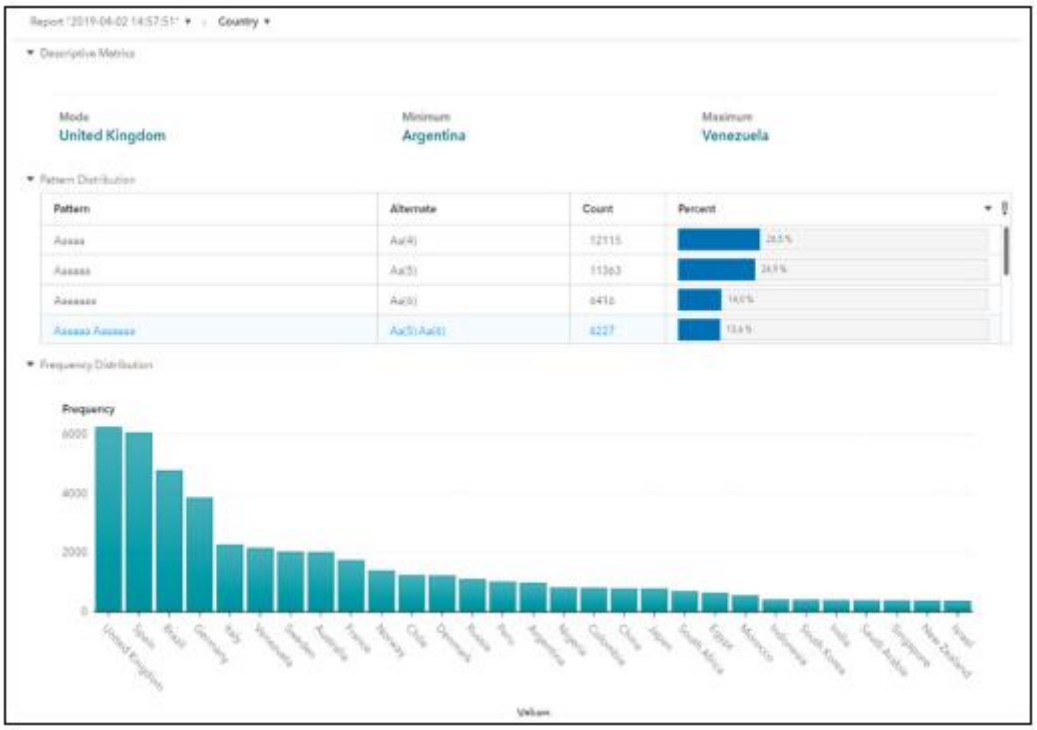
* Sprawdź, czy między źródłami danych występują szumy i niespójności.

Jeśli w ramach kontroli jakości wykryjesz problemy z jakością danych, musisz zdefiniować i udokumentować możliwe rozwiązania. Skoncentruj się na atrybutach, które wydają się być sprzeczne ze zdrowym rozsądkiem; wykresy wizualizacji, histogramy i inne sposoby wizualizacji i eksploracji danych to świetne sposoby na ujawnienie możliwych niespójności danych. Konieczne może być również całkowite wykluczenie danych o niskiej jakości lub bezużytecznych w celu przeprowadzenia potrzebnej analizy. Poniższy rysunek przedstawia tabelę sformatowaną w celu przedstawienia przeglądu zestawu danych. Takie tabele są dobrym sposobem na uzyskanie pierwszego przeglądu danych z perspektywy jakości, ponieważ wykorzystują statystyki opisowe do szybkiego wykrywania wartości ekstremalnych pod względem wartości minimalnych, maksymalnych, mediany, średniej i odchylenia standardowego. Tabela pozwala nam również przeanalizować kluczowe wartości, aby upewnić się, że są one w 100% niepowtarzalne i nie zawierają zduplikowanych lub brakujących wartości. Jeśli na przykład studiujesz dane dotyczące swoich klientów, chcesz mieć pewność, że klient nie wystąpi dwa razy z powodu błędu ortograficznego - lub w ogóle go nie ma na liście!



Column	Unique	Primary Key ...	Null	Bla...	Pattern Count	Mean	Median	Minimum	Maximum	Standa...
Age	0,08 % (43)	No				39,29	39,00	19,00	61,00	5,59
City	0,08 % (42)	No			16			Auckland	Valencia	
Continent	0,01 % (5)	No			4			Africa	South Ame...	
Country	0,06 % (29)	No			10			Argentina	Venezuela	
Customer	100,00 % (45...	Yes			3			ARBUENO...	ZAJOHAN...	
CustomerDistance	100,00 % (45...	Yes				5,56	5,74	0,04	10,00	2,34
CustomerInDays	4,54 % (2073)	No				156,66	20,00	0,00	3 647,00	396,05
CustomerLat	100,00 % (45...	Yes				25,60	40,52	-37,92	60,07	32,85
CustomerLon	100,00 % (45...	Yes				7,20	2,15	-77,14	174,85	55,35
FirstOrderAmount	97,13 % (44...	No				199,74	161,82	4,73	1 874,32	159,25
FirstOrderCostOfSales	100,00 % (45...	Yes				165,85	140,92	4,20	1 566,99	126,92
FirstOrderCostOfSalesPerc	100,00 % (45...	Yes				0,85	0,91	0,57	0,99	0,11
FirstOrderCustomerSatisfa	100,00 % (45...	Yes				0,51	0,45	0,18	1,00	0,20
FirstOrderDeliveryTime	0,08 % (35)	No				4,30	3,00	1,00	25,00	3,38
FirstOrderDiscount	0,68 % (312)	No				14,02	6,00	0,00	767,00	27,64
FirstOrderDiscountPerc	57,0...	No				0,07	0,04	0,00	0,88	0,09
FirstOrderExp	4,74 % (1707)	No				0,41	0,37	0,01	1,00	0,20
FirstOrderListAmount	8,99 % (1794)	No				213,76	173,58	5,61	1 905,72	167,17
FirstOrderProductQuality	100,00 % (45...	Yes				0,77	0,78	0,56	0,92	0,06
FirstOrderProfit	8,94 % (400)	No				-0,83	-10,00	-234,00	270,00	41,67

Rysunek przedstawia graficzną wizualizację tych samych danych, ale ten wykres skupia się tylko na jednej kolumnie; kraj. Patrząc na dane z perspektywy kraju, można zweryfikować dystrybucję danych w inny sposób i ewentualnie wykryć niespójności lub brakujące wartości, które były trudne do wykrycia na podstawie przeglądu. W tym konkretnym przykładzie narzędzie faktycznie posiada funkcję o nazwie Pattern, która wskazuje, kiedy wartości danych odbiegają od normy.



Zapytaj dowolną dobrze prosperującą organizację o sekret sukcesu, a otrzymasz wiele odpowiedzi, solidną strategię dotyczącą danych lub skalkulowane ryzyko w połączeniu z ostrożnym budżetowaniem. Wszystkie są dobrymi praktykami biznesowymi pochodzącymi z tego samego miejsca: solidną podstawą wysokiej jakości danych. Gdy dysponujesz dokładnymi, aktualnymi danymi jazdy

Twoja firma, nie tylko bijesz rachunki – bijesz rekordy. Proces oceny jakości danych jest niezbędny do zapewnienia wiarygodnych wyników analitycznych. Proces ten zależy od podejścia opartego na ludzkim nadzorze, ponieważ niemożliwe jest określenie defektu wyłącznie na podstawie danych.

Poprawa jakości danych

Co więc właściwie robisz, jeśli zdajesz sobie sprawę, że jakość Twoich danych jest naprawdę zła? Moją rekomendacją jest zastosowanie opisanego poniżej czteroetapowego podejścia, aby rozpocząć od podkreślenia luk i zdefiniowania mapy drogowej w celu wdrożenia potrzebnych ulepszeń w jakości danych.

1. Zakres: Zdefiniuj problem z jakością danych i opisz problem biznesowy związany z problemem jakościowym. Określ szczegóły danych oraz procesy biznesowe, produkty i/lub usługi, na które mają wpływ problemy z jakością danych.

2. Eksploruj: Przeprowadź wywiady z kluczowymi interesariuszami na temat potrzeb i oczekiwań dotyczących jakości danych, a także problemów z jakością danych. Przejrzyj procesy jakości danych i wsparcie narzędzi (jeśli istnieje) w ramach funkcji biznesowych i zidentyfikuj potrzebne zasoby dla działań związanych z poprawą jakości.

3. Analizuj: Oceń obecne praktyki w porównaniu z najlepszymi praktykami branżowymi w zakresie jakości danych i dostosuj je do ustaleń z fazy eksploracji. .

4. Zalecenie: Opracuj mapę drogową poprawy procesu jakości danych i zdefiniuj, jak musi wyglądać architektura techniczna procesu jakości danych, uwzględniając Twoje ustalenia dotyczące tego, co obecnie nie działa i co jest istotne z perspektywy biznesowej.

Uwzględnienie etycznych aspektów nauki o danych

Według ankiety Gartnera CIO Agenda Survey z 2018 r., 85% projektów AI do 2020 r. przyniesie błędne wyniki z powodu stronniczości danych, algorytmów lub zespołów programistycznych. To poważna postać, którą należy się zająć. Pomyśl o tym: jeśli coraz więcej firm i organizacji staje się opartych na danych i sztucznej inteligencji, a także automatyzuje swoje działania w oparciu o technologie sztucznej inteligencji, którym nie można ufać, oznacza to, że na horyzoncie pojawiają się kłopoty - nie tylko z ewolucją sztucznej inteligencji , ale dla całego społeczeństwa. W tym kontekście zajęcie się etycznymi aspektami sztucznej inteligencji ma fundamentalne znaczenie i będzie moim celem w tym rozdziale. Ale chcę też wyjaśnić, że nie powinieneś zaczynać myśleć o etyce dopiero wtedy, gdy zaczniesz wdrażać swoją strategię analizy danych. Perspektywa etyczna jest w rzeczywistości niezwykle ważna do rozważenia od samego początku – to znaczy od momentu rozpoczęcia projektowania modeli biznesowych, architektury, infrastruktury i sposobów pracy oraz budowania samych zespołów. W tym rozdziale wyjaśniono podstawy, które należy wziąć pod uwagę zarówno z perspektywy strategicznej, jak i praktycznej.

Wyjaśnienie etyki AI

Do czego więc właściwie odnosi się etyka AI i które obszary są ważne, aby wzbudzić zaufanie do danych i algorytmów? Cóż, ta koncepcja ma wiele aspektów, ale istnieje pięć podstaw, na których można polegać;

* Bezstronne dane, zespoły i algorytmy. Odnosi się to do znaczenia zarządzania nieodłącznymi uprzedzeniami, które mogą wynikać ze składu zespołu programistów, jeśli nie ma dobrej reprezentacji płci, rasy i płci. Dane i metody szkoleniowe muszą być jasno określone i uwzględnione w projekcie AI. Zdobywanie spostrzeżeń i potencjalne podejmowanie decyzji w oparciu o model, który jest w jakiś sposób stronniczy (na przykład tendencja do nierówności płci lub postaw rasistowskich), nie jest czymś, co chcesz osiągnąć.

* Wydajność algorytmu. Wyniki decyzji AI powinny być zgodne z oczekiwaniami interesariuszy, że algorytm działa na pożądanym poziomie precyzji i spójności oraz nie odbiega od celu modelu. Gdy modele są następnie wdrażane w środowisku docelowym w sposób dynamiczny i nadal trenują i optymalizują wydajność modelu, model dostosuje się do potencjalnych nowych wzorców danych i preferencji i może zacząć odbiegać od pierwotnego celu. Dlatego niezbędne jest ustalenie wystarczających polityk, aby szkolenie modelowe było zgodne z celami.

* Odporna infrastruktura. Upewnij się, że dane wykorzystywane przez komponenty systemu AI oraz sam algorytm są zabezpieczone przed nieautoryzowanym dostępem, uszkodzeniem i/lub atakiem adwersarza.

* Przejrzystość użytkowania i zgoda użytkownika. Użytkownik musi zostać wyraźnie powiadomiony o interakcji z AI i musi mieć możliwość wybrania poziomu interakcji lub całkowitego odrzucenia tej interakcji. Odnosi się również do znaczenia uzyskania zgody użytkownika na gromadzone i wykorzystywane dane. Wprowadzenie ogólnego rozporządzenia o ochronie danych (RODO) w UE wywołało dyskusje w USA wzywające do podobnych środków, co oznacza, że świadomość wagi danych osobowych, a także potrzeba ochrony tych informacji powoli się poprawia. Tak więc, nawet jeśli dane są gromadzone w sposób bezstronny, a modele są budowane w sposób bezstronny, nadal możesz spotkać się z etycznie trudnymi sytuacjami (lub nawet złamać prawo), jeśli używasz danych osobowych bez odpowiednich uprawnień.

* Modele do wyjaśnienia. Odnosi się to do potrzeby, aby metody szkoleniowe i kryteria decyzyjne AI były łatwe do zrozumienia, udokumentowane i łatwo dostępne do oceny i walidacji przez ludzi. Odnosi się do sytuacji, w których zadbano o to, aby algorytm, będący częścią inteligentnej maszyny, wytwarzał działania, którym ludzie mogą zaufać i które są łatwe do zrozumienia. Przeciwieństwem wyjaśnialności AI jest traktowanie algorytmu jako czarnej skrzynki, w której nawet projektant algorytmu nie jest w stanie wyjaśnić, dlaczego sztuczna inteligencja doszła do określonego spostrzeżenia lub decyzji.

Dodatkowa kwestia etyczna, która ma bardziej techniczny charakter, dotyczy odtwarzalności wyników poza środowiskiem laboratoryjnym. Sztuczna inteligencja jest wciąż niedojrzała, a większość prac badawczo-rozwojowych ma charakter eksploracyjny. Nadal istnieje niewielka standaryzacja dotycząca uczenia maszynowego/sztucznej inteligencji. Pojawiają się de facto zasady rozwoju AI, ale powoli i nadal są one w dużej mierze napędzane przez społeczność. Dlatego musisz upewnić się, że wszelkie wyniki algorytmu są rzeczywiście odtwarzalne – co oznacza, że uzyskasz takie same wyniki w rzeczywistym, docelowym środowisku, jak nie tylko w środowisku laboratoryjnym, ale także między różnymi środowiskami docelowymi (pomiędzy różnymi operatorami w sektorze telekomunikacyjnym , na przykład.)

Adresowanie godnej zaufania sztucznej inteligencji

Jeśli dane, do których potrzebujesz dostępu, aby zrealizować swoje cele biznesowe, można uznać za niepoprawne etycznie, jak sobie z tym radzisz? Łatwo powiedzieć, że aplikacje nie powinny zbierać danych o rasie, płci, niepełnosprawności lub innych chronionych klasach. Ale faktem jest, że jeśli nie zbierzesz tego typu danych, będziesz miał problem ze sprawdzeniem, czy Twoje aplikacje są rzeczywiście uczciwe wobec mniejszości. Algorytmy uczenia maszynowego, które uczą się na podstawie danych, staną się tak dobre, jak dane, na których działają. Niestety, wiele algorytmów okazało się całkiem dobrych w ustalaniu własnych odpowiedników dla rasy i innych klas, w sposób sprzeczny z tym, co wielu uważa za właściwe ludzkie myślenie etyczne. Twoja aplikacja nie byłaby pierwszym systemem, który mógłby okazać się niesprawiedliwy, pomimo najlepszych intencji jego twórców. Ale żeby było jasne, ostatecznie Twoja firma będzie odpowiedzialna za działanie swoich algorytmów, a (miejmy nadzieję) przepisy dotyczące uprzedzeń w przyszłości będą bardziej rygorystyczne niż obecnie. Jeśli firma nie przestrzega praw i przepisów lub granic etycznych, koszty finansowe mogą być znaczne – a być może nawet gorzej, ludzie mogą całkowicie stracić zaufanie do firmy. Może to mieć poważne konsekwencje, od klientów porzucających markę, przez pracowników tracących pracę, po ludzi idących do więzienia. Aby uniknąć tego typu scenariuszy, należy wcielić w życie zasady etyczne, a aby tak się stało, pracownicy muszą mieć możliwość i zachęcać pracowników do etycznego postępowania w codziennej pracy. Powinni umieć rozmawiać o tym, co tak naprawdę oznacza etyka w kontekście celów biznesowych i jakie koszty dla firmy mogą być w ich imieniu znoszone. Muszą też być w stanie przynajmniej przedyskutować, co by się stało, gdyby rozwiązanie nie mogło zostać wdrożone w etycznie poprawny sposób. Czy taka realizacja wystarczyłaby do jej zakończenia? Ogólnie rzecz biorąc, naukowcy zajmujący się danymi uważają, że ważne jest dzielenie się najlepszymi praktykami i artykułami naukowymi na konferencjach, pisanie postów na blogach oraz opracowywanie technologii i algorytmów open source. Jednak problemy, takie jak uzyskanie świadomej zgody, nie są omawiane tak często. Nie jest tak, że problemy nie są rozpoznawane lub rozumiane; są po prostu postrzegane jako mniej warte dyskusji. Zamiast pozwalać, aby takie nastawienie trwało, firmy powinny aktywnie zachęcać (a nie tylko pozwalać) na więcej dyskusji na temat uczciwości, właściwego wykorzystania danych i szkód, jakie może wyrządzić niewłaściwe wykorzystanie danych. Niedawne skandale związane z naruszeniami bezpieczeństwa komputerowego pokazały konsekwencje chowania głowy w piasek: wiele firm, które nigdy nie poświęciły czasu na wdrożenie dobrych praktyk i zabezpieczeń, teraz płaci za to zaniedbanie szkodą na reputacji i finansach.

Ważne jest, aby dochować takiej samej należytej staranności, jaka jest obecnie stosowana w kwestiach bezpieczeństwa, gdy myślimy o kwestiach takich jak uczciwość, odpowiedzialność i niezamierzone konsekwencje wykorzystania danych. Nigdy nie będzie możliwe przewidzenie wszystkich niezamierzonych konsekwencji takiego użycia i tak, możliwość przewidywania przyszłości jest ograniczona. Ale można było łatwo przewidzieć wiele niezamierzonych konsekwencji. (Funkcja przeglądu roku na Facebooku, która wydawała się robić wszystko, aby przypomnieć użytkownikom Facebooka o śmierci w rodzinie i innych bolesnych wydarzeniach, jest doskonałym przykładem.) Słynne motto Marka Zuckerberga: „Ruszaj się szybko i niszcz rzeczy" jest niedopuszczalne, jeśli nie zostało to przemyślane pod kątem tego, co może się zepsuć. Liderzy firm powinni nalegać, aby mogli rozważać takie aspekty – i zatrzymywać linię produkcyjną, gdy coś pójdzie nie tak. Pomysł ten wywodzi się z metody produkcyjnej Andona Toyoty: każdy pracownik linii montażowej może zatrzymać linię, jeśli zobaczy, że coś jest nie tak. Linia nie uruchamia się ponownie, dopóki problem nie zostanie rozwiązany. Pracownicy nie muszą obawiać się konsekwencji ze strony kierownictwa za zatrzymanie linii; cieszą się zaufaniem i oczekuje się od nich odpowiedzialnego zachowania. Co by to znaczyło, gdybyś mógł to zrobić za pomocą funkcji produktu lub algorytmów AI/ML? Gdyby ktoś na Facebooku mógł powiedzieć: „Czekaj, otrzymujemy skargi na przegląd roku" i wycofał go z produkcji, Facebook byłby teraz w znacznie lepszej sytuacji z etycznego punktu widzenia. Oczywiście to duża, skomplikowana firma, z dużym, skomplikowanym produktem. Ale to samo dotyczy Toyoty i tam zadziałało. Kwestią kryjącą się za wszystkimi tymi obawami jest oczywiście kultura korporacyjna. Środowiska korporacyjne mogą być wrogo nastawione do wszystkiego innego niż krótkoterminowa rentowność. Jednak w czasach, gdy publiczna nieufność i rozczarowanie są na najwyższym poziomie, etyka zamienia się w dobrą inwestycję korporacyjną. Kierownictwo wyższego szczebla dopiero zaczyna to dostrzegać, a zmiany w kulturze korporacyjnej nie nastąpią szybko, ale jasne jest, że użytkownicy chcą współpracować z firmami, które traktują ich i ich dane w sposób odpowiedzialny, a nie tylko jako potencjalny zysk lub zaangażowanie, zmaksymalizowany. Firmy, które odniosą sukces w zakresie etyki AI, to te, które tworzą przestrzeń dla etyki w swoich organizacjach. Oznacza to umożliwienie naukowcom zajmującym się danymi, inżynierom danych, programistom i innym specjalistom ds. danych „zajęcie się etyką" w praktyce. Nie chodzi o zatrudnianie wyszkolonych etyków i przydzielanie ich do swoich zespołów; chodzi o to, by każdego dnia żyć wartościami etycznymi, a nie tylko o nich mówić. To właśnie oznacza „robić dobrą analizę danych".

Przedstawiamy etykę według projektu

Jaki jest najlepszy sposób podejścia do wdrażania etyki AI już w fazie projektowania? Czy może być dostępna lista kontrolna? Teraz, kiedy o tym wspomniałeś, jest jeden i znajdziesz go w Wielkiej Brytanii. Tamtejszy rząd uruchomił ramy etyki danych, zawierające podręcznik etyki danych. W ramach inicjatywy wyodrębnili siedem odrębnych zasad dotyczących etyki AI. Zeszyt ćwiczeń, który wymyślili, składa się z szeregu pytań otwartych, które mają na celu zbadanie, czy przestrzegasz tych zasad. Trzeba przyznać, że jest wiele pytań – a dokładnie 46, co jest zbyt dużą liczbą dla analityka danych, aby stale śledzić i skutecznie włączać do codziennej rutyny. Aby takie pytania były naprawdę przydatne, muszą być osadzone nie tylko w rozwojowych sposobach pracy, ale także jako część infrastruktury i wsparcia systemów data science. Nie chodzi tylko o umożliwienie praktycznego przestrzegania zasad etycznych w codziennej pracy i udowodnienie, że firma działa zgodnie z zasadami etyki – firma musi również stać za tymi ambicjami i uwzględniać je jako część swojego kodeksu postępowania . Jednak gdy firma mówi o dodaniu etyki AI do swojego kodeksu postępowania, wartość nie wynika z samej obietnicy, ale raczej wynika z procesu, jaki ludzie przechodzą podczas jej opracowywania. Ludzie, którzy pracują z danymi, zaczynają teraz prowadzić dyskusje na szeroką skalę, które nigdy nie miałyby miejsca jeszcze dziesięć lat temu. Ale same dyskusje nie zakończą ciężkiej pracy. Istotne jest, aby nie tylko mówić o tym, jak korzystać z danych w sposób etyczny, ale także o etycznym korzystaniu z danych. Zasady muszą zostać

wprowadzone w życie! Oto krótsza lista pytań, które należy rozważyć, gdy Ty i Twoje zespoły ds. analityki danych współpracujecie, aby uzyskać wspólne i ogólne zrozumienie tego, co jest potrzebne do rozwiązania problemów etycznych związanych z AI:

* Hakowanie: W jakim stopniu zamierzona technologia AI jest podatna na hakowanie, a tym samym potencjalnie podatna na nadużycia?

* Dane treningowe: Czy przetestowałeś swoje dane treningowe, aby upewnić się, że są uczciwe i reprezentatywne?

* Stronniczość: Czy Twoje dane zawierają możliwe źródła uprzedzeń?

* Skład zespołu: Czy skład zespołu odzwierciedla różnorodność opinii i środowisk?

* Zgoda: Czy potrzebujesz zgody użytkownika na zbieranie i wykorzystywanie danych? Czy masz mechanizm zbierania zgód od użytkowników? Czy jasno wyjaśniłeś, na co użytkownicy wyrażają zgodę?

* Odszkodowanie: Czy oferujesz zwrot kosztów, jeśli ludzie ucierpią z powodu wyników Twojej technologii AI?

* Hamulec awaryjny: czy możesz wyłączyć to oprogramowanie w środowisku produkcyjnym, jeśli zachowuje się źle?

* Przejrzystość i uczciwość: czy wykorzystywane dane i algorytmy sztucznej inteligencji są zgodne z wartościami korporacyjnymi dotyczącymi technologii, takimi jak zachowanie moralne, szacunek, uczciwość i przejrzystość? Czy przetestowałeś uczciwość w odniesieniu do różnych grup użytkowników?

* Wskaźniki błędów: czy przetestowałeś różne poziomy błędów wśród różnych grup użytkowników?

* Wydajność modelu: Czy monitorujesz wydajność modelu, aby zapewnić, że oprogramowanie pozostanie sprawiedliwe w czasie? Czy można mu zaufać, że będzie działał zgodnie z zamierzeniami, nie tylko podczas wstępnego szkolenia lub modelowania, ale także podczas jego ciągłego „uczenia się" i ewolucji?

* Bezpieczeństwo: Czy masz plan ochrony i zabezpieczenia danych użytkownika?

* Odpowiedzialność: Czy istnieje wyraźna linia odpowiedzialności wobec jednostki i jasność dotycząca sposobu działania sztucznej inteligencji, używanych przez nią danych i stosowanych ram decyzyjnych?

* Projekt: Czy projekt AI uwzględniał lokalny i makro-społeczny wpływ, w tym jego wpływ na finansowe, fizyczne i psychiczne samopoczucie ludzi i naszego środowiska naturalnego?

Stawanie się opartym na danych

Jeśli firma w dzisiejszym klimacie biznesowym nie zainwestuje dużych pieniędzy w rozwój oparty na danych, w końcu zginie. Firmy, które nie wierzą, że ich dane są atutem (i dlatego powinny być odpowiednio zarządzane), będą miały poważne kłopoty w ciągu najbliższych pięciu lat. W tym rozdziale wyjaśniono, dlaczego konieczne jest, aby Twoja firma stała się oparta na danych, i oferuje porady na temat kroków, które firma musi podjąć, aby stać się opartą na danych.

Zrozumienie, dlaczego sterowanie danymi jest koniecznością

Firmy i organizacje z wielu obszarów biznesowych rozpoczęły swoją podróż do przechwytywania, tworzenia i wykorzystywania danych w sposób, który zasadniczo zmienia sposób pracy i życia ludzi. Jak zapewne już wiesz, punktem wyjścia dla każdej organizacji opartej na danych jest po prostu uświadomienie sobie, że dane są podstawą wszystkiego, co robi. To naprawdę podstawa wszystkiego, a organizacje z całego spektrum biznesowego stają się teraz świadome transformacyjnej mocy danych, analiz i sztucznej inteligencji. Firmy zaczynają również rozumieć prawdziwe wyzwania, które stoją przed nimi. Dla wielu wystarczy skatalogowanie i kategoryzowanie wszystkich dostępnych danych; określenie i dodanie zasad przetwarzania i wykorzystywania danych w celu przełożenia danych na wartość materialną wydaje się zadaniem niemal niewykonalnym. Ale chociaż jest to trudne, nie jest niemożliwe i kilka firm zaczyna się teraz adresować to wyzwanie bardziej strategicznie i bardziej praktycznie. „Miło wiedzieć", możesz powiedzieć, „ale od czego zaczynamy?" Zamiast zaczynać od zatrudnienia zewnętrznej osoby zajmującej się danymi (jak ta osoba jest powszechnie znana), radzę Ci, zainwestować trochę czasu i wysiłku w znalezienie kluczowej osoby w Twojej organizacji, która jest chętna poprowadź swoją inicjatywę opartą na danych. Ta osoba powinna być kimś, kto widzi szerszy obraz i pomaga stworzyć strategię danych opartą na dokładnym wglądzie w sposób funkcjonowania firmy, a także jej przyszłe cele biznesowe. Ta kluczowa osoba powinna również posiadać umiejętności ludzkie i komunikacyjne, aby przekształcić całą organizację - oczywiście z wystarczającym wsparciem - i powinna być gotowa (i wystarczająco cierpliwa), aby być na pierwszej linii, aby przenieść firmę z prostego integratora danych przez cały drogę do innowatora na rynku. Kiedy dziś firmy twierdzą, że uznają, że dane mają wartość, ich twierdzenia niekoniecznie są poprawne pod względem prawdziwej wartości danych. Łatwo zrozumieć, że dane transakcyjne można wykorzystać do raportowania lub analizy danych, co może następnie prowadzić do lepszego podejmowania decyzji. Jednak mimo że postrzegana wartość danych wzrosła w ciągu ostatnich dwóch dekad, wiele firm wciąż pozostaje w tyle, jeśli chodzi o wydajne przechwytywanie i udostępnianie danych oraz zarządzanie nimi. Dzieje się tak głównie dlatego, że ich systemy i procesy odzwierciedlają przestarzałe przekonanie, że dane są po prostu produktem ubocznym jakiejś innej działalności, a nie kluczem do ich sukcesu biznesowego. Aby wyjść poza takie podejście i faktycznie wejść w XXI wiek, takie organizacje muszą zainwestować znacznie więcej czasu i wysiłku w stworzenie strategii danych. Wszystko dobrze, ale co to właściwie znaczy, gdy mówię, że firma musi stworzyć strategię dotyczącą danych? Przede wszystkim opracowanie strategii dotyczącej danych oznacza upewnienie się, że wszystkie zasoby danych są rozmieszczone w taki sposób, aby można było z nich łatwo i wydajnie korzystać, udostępniać i przenosić. Innymi słowy, posiadanie strategii danych gwarantuje, że dane są zarządzane i wykorzystywane jako zasób, a nie tylko jako produkt uboczny innej aplikacji. Poprzez ustanowienie wspólnych metod, praktyk i procesów zarządzania, manipulowania i udostępniania danych w całej firmie w powtarzalny sposób, strategia danych zapewnia, że cele i zadania efektywnego i wydajnego wykorzystania danych są dopasowane w świadomy i strategiczny sposób. Niestety, podobnie jak wiele firm nadal wykorzystuje dane jako produkt uboczny, a nie jako podstawową wartość swojej działalności, wiele z nich nie rozwiązuje swoich problemów z danymi poprzez tworzenie i realizowanie własnych strategii dotyczących danych, ale raczej zatrudnia analityków danych, których zadaniem jest

ta praca. „znalezienia rzeczy w danych”. W rezultacie zamiast zatrudniać kogoś z wyraźnym celem biznesowym, jakim jest przekształcenie danych w spostrzeżenia, dajesz grupie analityków danych dostęp do bazy danych i każesz im uruchamiać zapytania, które każde podstawowe narzędzie analityczne może dostarczyć w ciągu kilku sekund. Jaki jest w tym sens? Bez strategii danych nie ma znaczenia, czy osoby, które zatrudniasz, nazywają się analitykami danych, naukowcami danych, inżynierami danych czy inżynierami uczenia maszynowego/sztucznej inteligencji. Jeśli zatrudniasz ludzi z fantazyjnymi tytułami tylko dlatego, że wszyscy na rynku wydają się to robić, nie oznacza to, że nagle zrozumiałeś wartość strategii danych. Bez jasno określonych celów, które zobowiązałeś się realizować w strategiczny sposób, wszystkie zatrudnienie, które zatrudnisz, będzie gorsza niż garderoba, ponieważ zainwestujesz czas i pieniądze, a otrzymasz niewiele w zamian. Zasadniczo wydajesz ogromne pieniądze, aby przyciągnąć i zatrzymać analityków danych (lub naukowców lub inżynierów), którzy spędzają większość czasu na wyodrębnianiu, czyszczeniu i modelowaniu danych, nie wiedząc a) na których problemach się skoncentrować i b) w jaki sposób może stworzyć nową okazję biznesową, która przyniesie firmie przychody lub zysk.

Przejście na model oparty na danych

Kierowanie się na dane jest zarówno ambicją organizacyjną, jak i absolutną koniecznością. Obejmuje zarówno aspekty kulturowe, jak i technologiczne. Chodzi o wykorzystywanie danych do podejmowania bezpośrednich działań, a także budowanie relacji i zaufania wokół danych. Ale chodzi również o to, jak patrzysz na dane i jak korzystasz z nich w codziennej działalności. Doświadczone technologicznie zespoły zarządzające rozumieją, że próba „zagotowania oceanu” podczas przyjmowania strategii zorientowanej na dane jest głupotą. Niektórzy już nauczyli się tego na własnej skórze, podejmując niezbyt udane projekty transformacji typu „big bang”. Biorąc pod uwagę zakres i złożoność dynamicznej natury cyfrowego świata, firmy zaczynają rozumieć, że zmiany będą nadal przyspieszać, nawet jeśli osiągną lub wyjdą poza stan docelowy. Kierownictwo musi wyznaczyć wizjonerskie cele dotyczące skoncentrowania się na danych, ale muszą również umożliwić zmiany podczas wdrażania — a nawet być może zmiany w samej wizji. Biorąc pod uwagę te realia, konieczne jest, aby kierownictwo podchodziło zarówno do zarządzania danymi, jak i do analityki oraz inicjatyw uczenia maszynowego/sztucznej inteligencji z perspektywy ciągłego doskonalenia, napędzając postęp w kierunku celów, jednocześnie podążając za mapą drogową, która jest dostosowywana w miarę rozwoju potrzeb biznesowych i organizacyjnych. Kierowanie się danymi to nowy sposób podejścia do firmy i zrozumiałe jest, że wiele firm gubi się we wszystkich swoich danych i ambicjach. Co więcej, wszystko toczy się bardzo szybko, jeśli chodzi o stale ewoluujące techniki wykorzystywane w obszarach nauki o danych, sztucznej inteligencji i infrastruktur zwirtualizowanych. (I nawet nie zaczynaj od nowych i rozszerzonych zasad w zakresie praktyk regulacyjnych, bezpieczeństwa i etycznych). Kiedy wprowadzasz do swojej organizacji podejście oparte na danych, nie wystarczy mieć jasne cele dotyczące tego, co musi być osiągnięty – potrzebujesz również sposobu mierzenia swoich osiągnięć w realizacji tych celów. Oprócz śledzenia postępów musisz być w stanie zmierzyć i udowodnić wartość i wpływ, jaki analityka danych i uczenie maszynowe mają na Twoją firmę.

Zapewnienie poparcia dla kierownictwa i wyznaczenie dyrektora ds. danych (CDO)

Najważniejszą decyzją, jaką musi podjąć kierownictwo firmy, jest sprawienie, by ktoś w pełni odpowiadał za dane. Takie postępowanie wysyła właściwy komunikat do całej organizacji nie tylko wewnętrznie, ale także na zewnątrz, w kierunku rynku i jego klientów. Chcesz, aby wszyscy wiedzieli, że poważnie traktujesz zadanie polegające na kierowaniu się danymi i że to właśnie napędza firmę w przyszłość. Dużą część odpowiedzialności CDO powinno należeć do zarządzania całą firmą i wykorzystywania danych jako zasobu organizacyjnego i strategicznego. Oznacza to współpracę z każdym działem w celu zaprojektowania wspólnego sposobu pozyskiwania, przechowywania,

zarządzania, udostępniania i wykorzystywania danych. Równie ważne jest, aby upewnić się, że kultura przyjmuje sposób myślenia oparty na danych, tak aby proces podejmowania decyzji był oparty na dyskusjach, które umożliwiają udostępnianie i ponowne wykorzystywanie danych, modeli i spostrzeżeń. W ostatecznym rozrachunku chodzi o wykorzystanie danych w rzeczywistych decyzjach podejmowanych w całej firmie oraz w ramach jej portfolio produktów i/lub usług. Sponsoring i zatwierdzenie przez najwyższe kierownictwo jest oczywiście niezbędne dla powodzenia każdej strategii dotyczącej danych, ale kierownictwo musi przejąć coś więcej niż tylko zarządzanie lub egzekwowanie. Przywództwo musi również „walk the talk", obejmujące podejmowanie decyzji w oparciu o fakty, dążenie do większej ilości lepszych danych oraz uznawanie osiągnięć, gdy wysiłki się powiodą. Kierownictwo musi również przedstawić jasną wizję, priorytetyzować aplikacje analityczne, rozumieć zwrot z inwestycji, przydzielać odpowiednie zasoby, zarządzać talentami, zapewniać koordynację między funkcjami i usuwać niektóre bariery, które nieuchronnie pojawią się podczas wdrażania. Wreszcie, kierownictwo musi nalegać na przestrzeganie wymogów prawnych i regulacyjnych dotyczących danych objętych zakresem.

Identyfikacja kluczowej wartości biznesowej dostosowanej do dojrzałości biznesowej

Punktem wyjścia każdej analizy danych powinno być zrozumienie najważniejszych możliwości biznesowych i/lub problemów Twojej firmy. Biorąc pod uwagę ten punkt wyjścia, możesz skoncentrować swoją analizę na identyfikacji i opisaniu, w jaki sposób podejście oparte na danych może wnieść wkład i zapewnić wartość w tej perspektywie. W obu przypadkach celem podejścia opartego na danych jest zapewnienie wartości tam, gdzie wcześniej jej nie było. Nigdy nie daj się złapać w potencjał konkretnej technologii; zawsze skup się na jego zastosowaniu. Zamiast pytać: „Co ta nowa technologia może dla nas zrobić?" powinieneś zapytać: „Jakie problemy muszę rozwiązać?" Sytuacja nasiliła się ze względu na szybką ewolucję technologii, powodując niedobór umiejętności w większości firm. Oczywiście jest miejsce na eksperymenty dotyczące tego, co mogą umożliwić nowe technologie, ale tylko pod warunkiem, że skupisz się na głównym celu, którym jest napędzanie biznesu do przodu. Skuteczne wykorzystanie danych, analiz i uczenia maszynowego/sztucznej inteligencji zapewnia, że będziesz w stanie zwiększyć korzyści z przyjęcia podejścia opartego na danych, co z kolei zwiększy chęć do dalszego wdrażania. Wszystkie firmy przechodzą przez różne etapy biznesowego cyklu życia – czyli przechodzenie firmy przez różne fazy w czasie, najczęściej podzielone na pięć etapów: uruchomienia, wzrostu, wstrząsu, dojrzałości i schyłku. Warto jednak zauważyć, że w dużych, globalnych firmach różne obszary biznesowe w ramach tej samej firmy mogą w danym momencie znajdować się na różnych etapach cyklu życia biznesowego. Dzieje się tak na przykład, gdy nowy segment lub obszar biznesowy zostanie dodany do konfiguracji z bardziej tradycyjnym zestawem obszarów biznesowych – na przykład nowy segment dotyczący biznesu cyfrowego. Ta mieszanka nowych i bardziej tradycyjnych obszarów biznesowych w dużym przedsiębiorstwie może być trudna w obsłudze, jeśli chodzi o dostosowanie celów biznesowych lub przeprowadzenie planu transformacji biznesowej. W obu przypadkach jednak ważne jest dopasowanie celów biznesowych do dojrzałości biznesowej, ponieważ sukces wdrożenia podejścia opartego na danych zależy w dużej mierze od gotowości firmy. Niedawno założona firma w fazie uruchamiania lub rozwoju może już opierać się na w pełni zdigitalizowanym modelu biznesowym i być już w połowie drogi do stania się już w fazie opartej na danych. Ustalenie celów biznesowych dla takiej firmy powinno być ambitne, a wdrożenie dość proste, biorąc pod uwagę korzystne warunki, takie jak dostęp do danych, prawa do danych, konfiguracja infrastruktury i podejście do zarządzania. Dla firmy będącej w fazie dojrzałości, a może nawet wchodzącej w fazę schyłku, ustalenie zrównoważonych i osiągalnych celów dla fundamentalnej zmiany będzie trudniejsze. Biorąc pod uwagę tę problematyczną mieszankę starego i nowego, podejście firmy podczas przekształcania się w organizację opartą na danych musi być dostosowane do podstawowych potrzeb organizacji dla całej firmy lub do potrzeb określonego obszaru docelowego. Możliwa jest transformacja

różnych obszarów biznesowych z różną prędkością i różnymi ambicjami, a nawet w różnym czasie. Jednak w zależności od integracji i zależności między różnymi obszarami biznesowymi, może to powodować problemy związane z zależnościami danych i infrastruktury, a także ofertami portfelowymi i komunikacją z klientami. Ważnym priorytetem przydzielonego CDO powinno zatem być określenie, które podejście należy przyjąć, aby możliwe było dostosowanie go do ogólnej strategii analizy danych. Poniższa lista przedstawia kilka przykładów podejść, które można zastosować, gdy staje się oparty na danych:

* Integrator danych: firma powinna skupić się na wdrażaniu opartym na danych przede wszystkim na nowoczesnej, zintegrowanej, wewnętrznej infrastrukturze danych zaprojektowanej w celu wprowadzenia nowych i większej ilości danych, które może wykorzystać do osiągnięcia celów biznesowych związanych z różnymi sposobami monetyzacji swoich danych w całej firmie.

* Optymalizator biznesowy: Firma zaangażowana w optymalizację bieżącej działalności powinna skupić się przede wszystkim na wykorzystaniu aktualnie dostępnych danych, aby wewnętrzne i zorientowane na klienta procesy biznesowe były jak najbardziej efektywne i wydajne.

* Zakłócacz rynku/innowator: w przypadku firmy, która ma ambicję stać się innowatorem rynkowym, należy skupić się na zwiększaniu ludzkich możliwości za pomocą uczenia maszynowego i technik sztucznej inteligencji. To położy podwaliny pod to, że firma stanie się przełomem na rynku cyfrowym.

Opracowanie strategii danych

Po tym, jak cele Twojej firmy staną się jaśniejsze, Twój CDO, w ramach ogólnej strategii analizy danych, musi stworzyć strategię danych opartą na biznesie, opartą na znacznym poziomie szczegółowości. Ponadto osoba ta musi zdefiniować zakres pożądanej kultury opartej na danych i sposobu myślenia dla Twojej firmy i kontynuować rozwój tej kultury. W tej sekcji wyjaśniam, o czym CDO musi pamiętać, aby wykonać te zadania, a także przedstawiam przykład zakresu strategii danych.

Dbanie o Twoje dane

Jednym z kluczowych aspektów każdej strategii dotyczącej danych jest dbanie o dane tak, jakby były one twoją siłą napędową - ponieważ tak jest. Musisz zająć się kwestiami jakości i integracji danych jako kluczowymi czynnikami swojej strategii dotyczącej danych, a także musisz dostosować swoje programy zarządzania danymi do celów organizacji, upewniając się, że zdefiniowałeś wszystkie strategie, zasady, procesy i standardy wspierające te cele. Organizacje powinny ocenić swój obecny stan i opracować plany osiągnięcia odpowiedniego poziomu dojrzałości w zakresie zarządzania danymi w określonym okresie. Należy pamiętać, że zarządzanie danymi nigdy nie jest kompletne; z konieczności ewoluuje, podobnie jak potrzeby i cele korporacyjne, technologia oraz aspekty prawne i regulacyjne. Programy zarządzania mogą obejmować tworzenie programów opartych na danych i informacji na poziomie firmy dla integratorów danych, a także tworzenie dostosowanych, opartych na segmentach programów dla optymalizatorów biznesowych i osób zakłócających rynek/innowatorów. Jednak nawet najlepsza strategia może się załamać, jeśli kultura biznesowa nie będzie skłonna do zmian. Integratorzy danych rozwijają się w opartym na dowodach środowisku operacyjnym, w którym dane i badania są wykorzystywane do tworzenia kultury opartej na danych, podczas gdy optymalizatorzy biznesowi i podmioty zakłócające rynek/innowatorzy muszą przyjąć „szybkie niepowodzenie" zwinnej kultury tworzenia oprogramowania w celu zwiększenia szybkości -rynek i innowacje.

Demokratyzacja danych

Równie ważne, jak zrozumienie wartości danych, do których Twoja firma ma dostęp, jest upewnienie się, że dane są łatwo dostępne dla tych, którzy muszą z nimi pracować. To właśnie oznacza

demokratyzacja danych. Biorąc pod uwagę jego znaczenie, powinieneś dążyć do tego, aby demokratyzacja zachodziła w całej Twojej organizacji. Faktem jest, że każdy w Twojej firmie każdego dnia podejmuje decyzje biznesowe, a decyzje te muszą być oparte na dokładnym zrozumieniu wszystkich dostępnych danych. Wiemy, że decyzje oparte na danych są lepszymi decyzjami, więc dlaczego nie miałbyś zapewnić ludziom dostępu do danych, których potrzebują do podejmowania lepszych decyzji? Chociaż większość ludzi rozumie potrzebę demokratyzacji danych, wcale nie jest niczym niezwykłym, że strategia firmy w zakresie danych koncentruje się na blokowaniu danych – tylko po to, aby zachować bezpieczną stronę. Jednak nic nie może być bardziej niszczycielskie dla realizacji wartości danych dla Twojej firmy niż przyjęcie mentalności bunkra na temat danych. Sposobem na rozpoczęcie generowania wewnętrznej i zewnętrznej wartości danych jest ich użycie, a nie blokowanie. Nawet przyjęcie radykalnego podejścia do całkowicie otwartego środowiska danych wewnętrznie jest lepsze niż zbytnie ograniczanie sposobu udostępniania i udostępniania danych w firmie.

Standaryzacja danych

Trzecim kluczowym elementem każdej strategii danych jest standaryzacja w celu szybkiego i wydajnego skalowania. Standaryzacja danych jest ważnym elementem sukcesu – takim, którego nie należy lekceważyć. Firma nie może liczyć na osiągnięcie celów, które zakładają 360-stopniowy widok wszystkich klientów poparty poprawnymi danymi bez wspólnego zestawu definicji i struktur danych w firmie i klientach. TM Forum, stowarzyszenie branżowe non-profit dla dostawców usług i ich dostawców w branży telekomunikacyjnej, opracowało coś, co nazywają ramami informacyjnymi (SID) we współpracy z profesjonalistami z branży telekomunikacyjnej i informacyjnej współpracującymi w celu zapewnienia uniwersalnych informacji i modeli danych. (Część nazwy dotycząca identyfikatora SID pochodzi z modelu Shared Information Data). Korzyści tego wspólnego modelu wynikają z jego zdolności do znacznego wspierania zwiększonej standaryzacji wokół danych w przestrzeni telekomunikacyjnej i obejmują takie aspekty, jak:

* Szybszy czas wprowadzania na rynek nowych produktów i usług

* Tańsza integracja danych i systemów

* Mniej czasu na zarządzanie danymi

* Zmniejszony koszt i wsparcie przy wdrażaniu wielu technologii

Organizacje od dawna dostrzegają potrzebę dążenia do standaryzacji w swoich strukturach danych transakcyjnych, ale muszą także zdawać sobie sprawę z wagi dążenia do standaryzacji w swoich strukturach danych analitycznych. Tradycyjne konfiguracje analityki i analizy biznesowej nadal wykorzystują hurtownie danych i hurtownie danych jako podstawowe repozytoria danych i tak, nadal są one bardzo cenne dla organizacji opartych na danych, ale umożliwienie dynamicznej analizy big data i rozwiązań z zakresu uczenia maszynowego/sztucznej inteligencji wymaga innej struktury aby być skutecznym.

Strukturyzacja strategii danych

Akt tworzenia strategii danych jest szansą na generowanie rozmów na temat danych, edukowanie kadry kierowniczej i identyfikowanie ekscytujących nowych możliwości dla organizacji. W rzeczywistości proces tworzenia strategii danych może generować wsparcie polityczne, zmiany w kulturze i sposobie myślenia oraz nowe cele i priorytety biznesowe, które są nawet bardziej wartościowe niż sama strategia danych. Ale co właściwie powinna obejmować strategia dotycząca danych? Poniższa lista daje pewien pomysł.

* Wizja zorientowana na dane i cele biznesowe, w tym scenariusze użytkownika

* Strategiczne zasady dotyczące danych, w tym traktowanie danych jako zasobu

* Wytyczne dotyczące bezpieczeństwa danych, praw do danych i względów etycznych

* Zasady zarządzania danymi, w tym zarządzanie danymi i jakość danych

* Zasady infrastruktury danych dotyczące architektury danych, pozyskiwania danych, przechowywania danych i przetwarzania danych

* Zakres danych, w tym priorytety w czasie

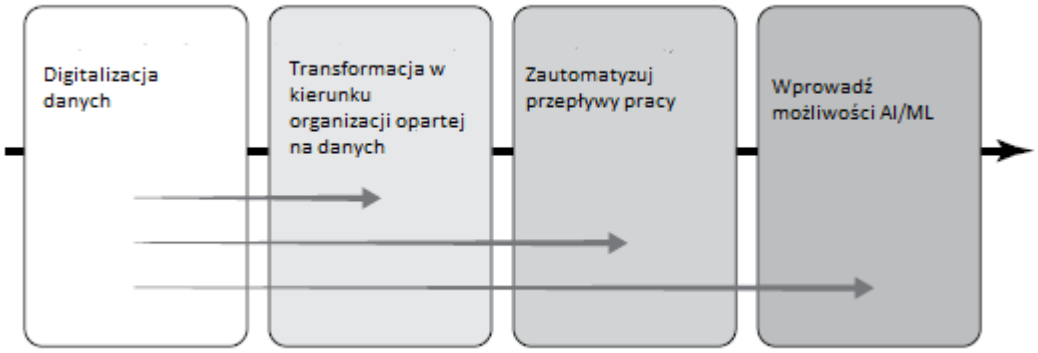
Nie mieszaj strategii dotyczącej danych ze strategią analizy danych. Główna różnica polega na tym, że strategia danych koncentruje się na kierunku strategicznym i zasadach dotyczących danych i stanowi podzbiór strategii nauki o danych. Strategia nauki o danych obejmuje strategię dotyczącą danych, ale także takie aspekty, jak organizacja, ludzie, kultura i sposób myślenia, kompetencje i role w zakresie nauki o danych, zarządzanie zmianą, pomiary i komercyjne konsekwencje biznesowe dla portfela firmy.

Ustanowienie kultury i sposobu myślenia opartego na danych

Oczywistym, ale ważnym krokiem w kierunku opierania się na danych jest poświęcenie czasu na zaangażowanie pracowników w to, co ta fundamentalna zmiana naprawdę oznacza w ich codziennej działalności biznesowej. Dotarcie tam zajmie trochę czasu i wysiłku, ale nie tylko jest warte zainwestowanego czasu, ale jest to również główny warunek wstępny, aby zmiana nastąpiła i trwała w czasie. Na wczesnych etapach wprowadzania sposobu myślenia opartego na danych skoncentruj się na wyjaśnianiu tego, co się dzieje, za pomocą przykładów zbliżonych do tego, co już jest robione. Zacznij od konkretnych przykładów tego, co faktycznie będą oznaczać zmiany w ich codziennej konfiguracji. Na przykład, jeśli obecne sposoby pracy w firmie są silnie reaktywne – co oznacza, że proces zaczyna się od reklamacji klienta – jaki byłby nowy punkt wyjścia? W jaki sposób podejście oparte na danych i proaktywne wpłynęłoby praktycznie na bieżące przepływy pracy, gdy przepływ zaczyna się od predykcyjnej analizy danych i ambicji zapobiegania reklamacjom? Na co dzień, w ramach zwykłego podejmowania decyzji, liderzy muszą aktywnie zachęcać pracowników do a) wyrobienia sobie nawyku proszenia o potrzebne im dane oraz b) korzystania z danych, którymi dysponują. Praktyczna i jasna prośba kierownictwa o faktyczne wykorzystanie dostępnych danych przeniknie do organizacji i będzie miała znacznie większy wpływ, niż mogłoby się wydawać. To naprawdę ważny krok, aby wprowadzić w firmie myślenie zorientowane na dane. Co więcej, wysiłek ten można by jeszcze silniej podkreślić, ustanawiając system nagród dla tych pracowników, którzy promują i napędzają kulturę, wykorzystując dane jako główny wkład w podejmowanie decyzji. Ostatnio krąży plotka, że „analityka staje się łatwiejsza” i pod pewnymi względami to prawda. Dostępne są dość wydajne, gotowe narzędzia analityczne, a niektóre nowe samoobsługowe aplikacje analityczne usunęły część złożoności, udostępniając analizę danych większej liczbie osób o różnych typach ról – a tym samym wspierając kulturę opartą na danych. Jednak chociaż analityka może być coraz łatwiejsza, zarządzanie danymi staje się coraz trudniejsze. Wynika to głównie z rosnącej różnorodności źródeł i struktur oraz coraz większej prędkości, z jaką jest generowany. Dlatego ważne jest, aby rozważyć posiadanie odpowiednich talentów i umiejętności wspierających aspekty inżynierii danych, a także naukę danych jako całość, a także należy pamiętać, że ten stan rzeczy pozostanie taki w przewidywalnej przyszłości.

Ewoluowanie od opartego na danych do napędzanego maszynowo

Kierowanie się danymi i kierowanie się maszyną to nie to samo. Możesz je nazwać stanami powiązanymi w nauce o danych, jeśli chcesz, ale zdecydowanie istnieją na różnych etapach dojrzałości. Kierowanie się danymi przede wszystkim odnosi się do idei, że każdy postęp w działaniu jest związany raczej z danymi niż z intuicją lub osobistym doświadczeniem. Z drugiej strony bycie sterowanym przez maszynę nie odnosi się do granic działania, ale raczej do sposobu, w jaki ta czynność jest wykonywana – dokładniej mówiąc, że jest to czynność połączona, zautomatyzowana i kontrolowana przez maszynę oraz jej wdrożenie pewien model lub algorytm. Napęd maszynowy to ostatni etap industrializacji i automatyzacji nauki o danych od początku do końca, napędzany zaprojektowaną inteligencją maszyny. Obecnie niewiele firm i organizacji znajduje się na etapie, na którym można powiedzieć, że są w pełni oparte na danych lub zdecydowanie w pełni napędzane przez maszyny. Jednak kilka firm zaczyna teraz zastanawiać się, jak sprawić, by część ich działalności była napędzana maszynowo, co jest pierwszym krokiem w kierunku pełnej transformacji na późniejszym etapie. . Takie podejście jest nie tylko możliwe, ale również wskazane w przypadku wczesnych eksperymentów dotyczących potencjalnych wysiłków, korzyści i wpływu. Kiedy zaczynasz eksperymentować z konsolidacją części swojej firmy w podejście bardziej oparte na maszynach, musisz mieć świadomość czterech głównych kroków – wszystkie ze współzależnościami, ale każdy z nich wymaga tego samego punktu wyjścia: digitalizacji danych.



Rysunek prowadzi do domu — zawsze najpierw digitalizuj dane. Bez tego kroku nie można podjąć żadnych innych kroków. Jednak po zdigitalizowaniu danych i ich udostępnieniu możesz wypróbować dowolne z poniższych kroków w dowolnej kolejności (chociaż zalecam przestrzeganie sekwencji zalecanej na rysunku zarówno dla pełnego zakresu obejmującego całą firmę, jak i wybranych części działalności) .

Digitalizacja danych

Gdy digitalizujesz dane, przekształcasz zasoby danych w cyfrowe treści do odczytu maszynowego. Mówiąc najprościej, bez cyfryzacji danych w firmie nie możesz zrobić kolejnego kroku w kierunku napędzania maszyn. Digitalizacja danych to proces przekształcania informacji na format cyfrowy (czytelny dla komputera). Wyjście jest cyfrową reprezentacją obiektu, obrazu, dźwięku, dokumentu lub sygnału (zazwyczaj sygnału analogowego) i obejmuje zmianę analogowego materiału źródłowego na format liczbowy. Rezultatem są zdigitalizowane dane w postaci liczb binarnych, które ułatwiają przetwarzanie komputerowe i inne operacje wykonywane maszynowo. Digitalizacja danych ma kluczowe znaczenie dla przetwarzania, przechowywania i transmisji danych, ponieważ umożliwia przesyłanie informacji wszelkiego rodzaju i we wszystkich formatach z taką samą wydajnością. Dane cyfrowe, w przeciwieństwie do danych analogowych (które zazwyczaj tracą jakość za każdym razem, gdy są kopiowane lub przesyłane), mogą (teoretycznie) być rozpowszechniane nieskończoną liczbę razy

bez żadnego pogorszenia jakości. Jakie zasoby, o których mówię, muszą zostać zdigitalizowane? A co ze wszystkim? Oznacza to wszystko, od danych klientów firmy, wewnętrznych procedur, procesów i przepływów pracy po informacje o pracownikach, informacje o produktach i inne rodzaje informacji. Digitalizacja wszystkich tych danych otwiera możliwość wykorzystania dostępnych danych jako danych wejściowych do analizy danych, co prowadzi do poprawy wewnętrznej wydajności i nowych możliwości biznesowych w ramach podejścia opartego na danych.

Stosowanie podejścia opartego na danych

Po przekształceniu danych w format cyfrowy i czytelny dla komputerów możesz zacząć podejmować kroki w kierunku pełnego wykorzystania danych biznesowych. Jak opisano wcześniej, wprowadzenie podejścia opartego na danych nie jest drobnym ulepszeniem bieżących operacji; aby odnieść sukces, musisz przeprowadzić fundamentalną transformację swojej organizacji od początku do końca, zaczynając od poparcia najwyższego kierownictwa. Po podjęciu decyzji o przeprowadzeniu tej transformacji i powszechnie wiadomo, co to będzie oznaczać pod względem nie tylko oczekiwanych korzyści biznesowych, ale także wymaganych wysiłków, pierwszym krokiem jest stworzenie solidnych i przemyślanych danych. strategia naukowa. Ta strategia pomoże Ci zaplanować i wykonać wszystkie niezbędne działania, aby osiągnąć ogólne cele biznesowe. Jeśli długoterminowy cel obejmuje napędzanie maszynami, to automatyzacja, uczenie maszynowe i sztuczna inteligencja muszą być już uwzględnione w strategii analizy danych na tym etapie. Po zdefiniowaniu i uzgodnieniu strategii analizy danych, ciężka praca nad przekształceniem firmy w opartą na danych zaczyna się na dobre, obejmując takie aspekty, jak udostępnianie i wykorzystywanie danych, ustalanie sposobu myślenia i kultury opartej na danych w całej firmie i upewniając się, że dane są wymagane i wykorzystywane w decyzjach i pomiarach w całej firmie. Aby przygotować organizację nie tylko na lepsze zrozumienie danych, ale także na konieczną zmianę sposobu myślenia i kultury, poświęć trochę czasu na badanie i eksperymentowanie z różnymi rozwiązaniami opartymi na maszynach dla wybranych obszarów. Takie postępowanie pomaga przygotować organizację na to, co nadchodzi, a także stanowi przykład tego, co to oznacza pod względem potrzebnych kompetencji, wpływu na obecne sposoby pracy, wpływu architektury i infrastruktury – a także uzyskanych korzyści. Tak, punktem wyjścia dla organizacji opartej na danych jest to, że wszystko zaczyna się i kończy na danych, ale aby te dane dostarczały wartość, pamiętaj, że należy to zrozumieć i wdrożyć w rzeczywistym kontekście firmy.

Automatyzacja przepływów pracy

Krok do automatyzacji części lub wszystkich procesów lub przepływów pracy można wykonać bezpośrednio po zdigitalizowaniu treści; nie wymaga to, abyś najpierw przekształcił swoją firmę w opartą na danych, nawet jeśli jest to wskazane, aby osiągnąć swój długoterminowy cel, jakim jest napędzanie maszyn. Automatyzacja przepływów pracy to pierwszy konkretny krok w przenoszeniu kontroli z ludzi na maszyny. Pamiętaj, że automatyzując przepływy pracy, ludzie wciąż decydują o podejściu i krokach, które maszyna musi podjąć; po prostu przenosisz odpowiedzialność za wykonanie kroków na maszynę. Ta metoda (zwana również automatyzacją procesów) polega na wykorzystaniu technologii komputerowej i inżynierii oprogramowania, aby usprawnić działanie systemów i procesów. (Najlepszym przykładem jest pomoc zakładom i fabrykom w bardziej wydajnym, bezpieczniejszym i niższym koszcie w branżach tak różnych, jak papiernicza, górnicza i cementowa). Automatyzację procesów można również zastosować na poziomie biznesowym, gdzie jest ona następnie kierowana jako automatyzacja procesów biznesowych (BPA). Stosowany tutaj odnosi się do technologii umożliwiającej automatyzację złożonych procesów biznesowych. BPA można wykorzystać do usprawnienia wielu aspektów działalności, w tym zwiększenia efektywności kosztowej, osiągnięcia transformacji cyfrowej, zwiększenia jakości usług i poprawy świadczenia usług. Pełne wdrożenie BPA zazwyczaj obejmuje takie czynności jak udostępnienie danych, integracja aplikacji, restrukturyzacja

pracowników oraz zastosowanie oprogramowania (maszyny) do automatyzacji zadań w całej organizacji. Robotic Process Automation (RPA) to rozwijająca się dziedzina w ramach BPA, która robi kolejny krok w kierunku napędzania maszyn i może (w bardziej zaawansowanych wersjach) dodać możliwości sztucznej inteligencji do zakresu maszyn.

Przedstawiamy możliwości AI/ML

Wprowadzając w firmie możliwości sztucznej inteligencji / uczenia maszynowego, robisz znaczący krok w kierunku napędzania maszyn. Ten krok obejmuje skoncentrowanie się na celu biznesowym i odpuszczenie części ludzkiej kontroli nad tym, kiedy i jak należy wykonać określone zadania. Dodanie funkcji inteligencji do maszyny oznacza, że modele i algorytmy są zaprojektowane tak, aby wykorzystywać dane do znajdowania i osiągania najlepszego możliwego sposobu realizacji określonego celu. Praktycznie rzecz biorąc, musisz skonfigurować i uprzemysłowić infrastrukturę w firmie, która jest zorientowana zarówno na dane, jak i na maszyny. Musi być miejsce na eksploracyjne środowiska programistyczne, w których analitycy danych będą mogli identyfikować nowe możliwości i tworzyć nowe modele, ale należy również zapewnić stabilne środowisko produkcyjne, aby umożliwić działanie algorytmów w ramach rzeczywistości operacyjnej firmy. W ostatecznym rozrachunku bycie napędzanym maszyną polega na umożliwieniu działania algorytmom, a następnie znalezieniu i wykonaniu najlepszego sposobu rozwiązania lub przewidywania i zapobiegania problemowi.

ROLA LUDZI W AI

Mimo że bycie napędzanym maszyną polega na maszynach, nie zapominaj o ludzkim zaangażowaniu poza opracowywaniem algorytmów. Twoim głównym celem nie jest tworzenie najmądrzejszych maszyn na świecie, ale raczej wykorzystanie inteligentnych maszyn w celu zwiększenia ludzkich możliwości Twojej organizacji, abyś mógł osiągnąć więcej. Dlatego chociaż rola, jaką ludzie odgrywają w biznesie napędzanym maszynami, może być inna, nadal są one dość ważne. Rola obejmuje zadania, których można oczekiwać, takie jak monitorowanie wydajności modelu i algorytmu, aby upewnić się, że maszyna robi to, czego potrzebuje, z oczekiwanym poziomem jakości, ale obejmuje również bezpośrednią interakcję z maszyną. Chodzi o to, że wczesne działania prowadzone przez maszyny koncentrują się zwykle na dostarczaniu nowych spostrzeżeń i zalecaniu działań, a nie na uruchamianiu w pełni zamkniętej pętli, w której decyzje są zarówno podejmowane, jak i podejmowane w całości przez maszynę. Jednak w miarę dojrzewania technologii w organizacji coraz więcej takich zamkniętych, opartych na maszynach scenariuszy zacznie działać w różnych firmach, zwłaszcza tam, gdzie można ograniczyć zakres i złożoność algorytmu. Przykładami tego typu działań prowadzonych przez maszyny są systemy zamknięte zaprojektowane z myślą o konkretnych wynikach w branżach takich jak górnictwo, produkcja i mniejsze systemy IoT. . Innym ważnym aspektem z perspektywy człowieka jest wymagany poziom kompetencji i zróżnicowane zestawy umiejętności wymagane nie tylko do wdrożenia podejścia opartego na maszynach, ale także do zaprojektowania strategii nauki o danych, aby uwzględnić wszystkie niezbędne aspekty i wymiary. . Niestety, dostęp do doświadczonej wiedzy z zakresu analityki danych w obszarze pełnej automatyzacji biznesowej opartej na maszynach typu end-to-end jest co najmniej ograniczony, zwłaszcza jeśli chodzi o bardziej złożone techniki sztucznej inteligencji w przestrzeni rozumowania kognitywnego. Jednak jeśli chodzi o dostępność analityków danych w sensie ogólnym, to rośnie ona powoli. Zainteresowanie różnymi uniwersytetami na całym świecie rośnie, podobnie jak dostępność programów uniwersyteckich z zakresu nauki o danych i innych rodzajów programów edukacyjnych z zakresu nauki o danych. Jednakże, ponieważ niewiele działających, działających, kompleksowych firm napędzanych maszynami jest obecnie w grze, jeszcze ważniejsze jest zbadanie obszaru biznesu opartego na maszynach w małych i łatwych do zarządzania krokach. Pozwoli to na powolny wzrost kompetencji i rozprzestrzenienie się ich w firmie w czasie - i pozwoli ciekawości kierować zmianą zamiast narzucać ją odgórnie. Egzekwowanie nauki o danych w

firmie bez pełnego zrozumienia wpływu wymaganych wysiłków lub oczekiwanych korzyści nie jest dobrym sposobem na rozpoczęcie pracy.

Budowanie skutecznych zespołów Data Science

Data scientist stał się najbardziej atrakcyjną rolą na dzisiejszym konkurencyjnym rynku pracy. Wynagrodzenia na poziomie podstawowym mogą wynosić sześciocyfrowe, a do 2020 r. przewiduje się około 700 000 wakatów. Co napędza ten popyt? Proste - wartość biznesowa. Zadaniem analityka danych jest wydobywanie spostrzeżeń ukrytych w górach danych — spostrzeżeń, które można następnie wykorzystać do osiągnięcia różnych celów biznesowych, od wykrywania oszustw po rozpoznawanie twarzy. Uznanie tej różnorodności u podstaw nauki o danych jest kluczem do budowania wydajnych zespołów zajmujących się analizą danych, które muszą składać się z osób o wysoce wyspecjalizowanych i uzupełniających się zestawach umiejętności, aby odnieść sukces. Ale zanim będziesz mógł wyruszyć w podróż w celu stworzenia idealnego zespołu zajmującego się analizą danych, musisz upewnić się, że masz odpowiednie kierownictwo.

Zaczynając od lidera zespołu Data Science

Zatrudniając kierownika zespołu zajmującego się analizą danych, pamiętaj, że nie zatrudniasz naukowca danych - zatrudniasz lidera w dziedzinie analityki danych. Jest różnica. Po pierwsze, pamiętaj, że umiejętności przywódcze są ważniejsze niż umiejętności eksperckie w nauce o danych. Nie oznacza to, że menedżer nie może być byłym analitykiem danych, ale oznacza to, że dana osoba musi być gotowa do objęcia roli lidera. Jeśli wyznaczysz lidera, który nie rozumie podstawowego obszaru nauki o danych i tego, jak różni się on od zwykłego tworzenia oprogramowania, budowanie wydajnego i odnoszącego sukcesy zespołu zajmującego się badaniem danych będzie przebiegać powoli, zarówno w perspektywie krótko-, jak i długoterminowej. Pamiętaj, że przywództwo jest umiejętnością samą w sobie. To, że ktoś w przeszłości udowodnił, że jest skutecznym współpracownikiem zespołu, nie oznacza, że ma umiejętności niezbędne do utrzymania i rozwijania wielkiego talentu, jednocześnie dostarczając cenne spostrzeżenia, produkty i wyniki swoim klientom i z powrotem do organizacji. Świetni analitycy danych mają mnóstwo opcji kariery i nie będą długo tolerować złych menedżerów. Jeśli chcesz zatrzymać świetnych analityków danych, dobrym punktem wyjścia jest zaangażowanie świetnych menedżerów. Jeśli lider jest byłym menedżerem ds. rozwoju oprogramowania, ważne jest, aby zrozumieć, że bycie liderem w dziedzinie nauki o danych to coś innego. Minimalnym wymaganiem w takiej sytuacji jest to, aby kierownik zespołu ds. nauki danych wiedział, że nie ma wystarczającej wiedzy w zakresie nauki o danych, a zatem musi zachować otwarty umysł i chce dowiedzieć się więcej. Ta sytuacja korzysta również z posiadania w zespole doświadczonych analityków danych, którzy mogą wesprzeć lidera na początku.

Przyjmowanie różnych podejść do przywództwa

Zrozumienie, jak kierować zespołami zajmującymi się analizą danych w porównaniu z zespołami zajmującymi się wyłącznie tworzeniem oprogramowania, ma podstawowe znaczenie dla odniesienia sukcesu w tej dziedzinie. W tej sekcji przedstawiam listę obszarów i aspektów wskazujących na główne różnice pomiędzy zespołami programistycznymi a zespołami data science – różnice które należy wziąć pod uwagę w ramach kierowania tymi zespołami:

* Metody pracy zespołowej: w nauce o danych metodologia różni się od tej, którą można znaleźć w tradycyjnym tworzeniu oprogramowania. Nauka o danych wymaga znacznie dłuższych cykli rozwojowych, aby uzyskać wynik i zweryfikować ten wynik, ze względu na jego zależność od pozyskania i przygotowania danych do odpowiedniego poziomu jakości, zanim będzie można przeprowadzić jakikolwiek rodzaj analizy. Ponieważ nauka o danych jest nadal we wczesnej fazie swojej ewolucji, wciąż jest w stanie zmian. Oznacza to, że techniki i metodologie stale się rozwijają, co wymaga wysokiego stopnia eksperymentowania w podejściu.

* Stosowane technologie i techniki: w nauce o danych musisz używać technologii rzadko używanych w tworzeniu oprogramowania (na przykład statystyki, uczenie maszynowe i sztuczna inteligencja) w celu tworzenia dynamicznych, samouczących się algorytmów i dynamicznych środowisk implementacji. Obszary takie jak robotyka stają się również coraz ważniejsze w obszarze data science. Technologie te wymagają innych technik, języków programowania i zestawów narzędzi oprócz tego, co jest używane w tradycyjnej inżynierii oprogramowania, w zależności od strategii i ukierunkowania na naukę danych. Bardziej ambitne i złożone modele i rozwiązania uczenia maszynowego/sztucznej inteligencji wymagają jeszcze bardziej zaawansowanych zestawów narzędzi.

* Wymagane kompetencje: Nauka o danych jest odrębnym obszarem kompetencji, z własnymi rolami i zestawami narzędzi. Były programista może zostać dobrym naukowcem zajmującym się danymi, ale wymaga to innego zestawu umiejętności – na przykład związanych z modelami statystycznymi, z bardziej zaawansowanymi kompetencjami matematycznymi. Należy również wziąć pod uwagę inne ważne role, takie jak architekt danych, inżynier danych, ekspertyza domenowa i inżynier automatyki.

* Potrzebna infrastruktura: w przypadku nauki o danych koncentracja na infrastrukturze jest również inna, ponieważ wszystko jest skoncentrowane na danych. Wszystko sprowadza się do posiadania wydajnej infrastruktury, która może obsługiwać wydajne przechwytywanie, anonimizację, przesyłanie danych, zgodność z prawem, bezpieczeństwo i zarządzanie danymi, względy etyczne, przechowywanie i przetwarzanie, zanim jeszcze zaczniesz pracować nad analizą i opracowywaniem modeli. Jak można sobie wyobrazić, liczba i różnorodność potrzeb związanych z analizą danych stawia wysokie wymagania w zakresie przepustowości infrastruktury i jej zdolności do analizy i eksploracji ogromnych zbiorów danych przy jednoczesnym zachowaniu jakości i integralności danych. . Dla porównania, czysta infrastruktura programistyczna jest ogólnie bardziej samowystarczalna z mniejszą liczbą zależności zewnętrznych. Ma też mniejsze zapotrzebowanie na skalowalność pod względem mocy obliczeniowej i pamięci masowej. Innym aspektem jest to, że obszar tworzenia oprogramowania jest również znacznie bardziej dojrzały z perspektywy ekosystemu oprogramowania, co oznacza, że jest znacznie bardziej usprawniony i ustandaryzowany niż nauka o danych, która wciąż jest w trybie eksperymentalnym i eksploracyjnym.

* Inne ważne kwestie: w nauce o danych aspekty wykorzystania danych i wydajności modeli związane z etyką, prywatnością i bezpieczeństwem mają ogromne znaczenie, jeśli chodzi o zrozumienie zespołu zajmującego się analizą danych i zarządzanie nim. Jeśli nie zostanie przeprowadzona należyta staranność i zostaną naruszone jakiekolwiek wymogi etyczne lub regulacyjne, wszelkie działania związane z nauką danych musiałyby zostać wstrzymane. Oprócz kosztów przestoju możesz również spotkać się z ogromnymi grzywnami związanymi z łamaniem takich przepisów i regulacji. Jako lider ważne jest zatem, aby zasady firmy dotyczące takich kwestii były wbudowane w strategie, wytyczne, systemy i mechanizmy kontrolne podczas całego cyklu życia danych, a także modeli.

Zbliżamy się do pozycji lidera nauki o danych

Podstawowym aspektem zostania świetnym liderem w dziedzinie analityki danych jest przemyślenie, w jaki sposób wzbudzić zaufanie, autentyczność i lojalność w zespole i wobec lidera. Może to być prawdą w przypadku wszystkich relacji zespół/lider, ale jest to jednak szczególnie prawdziwe w przypadku nauki o danych, gdzie firmy nadal są bardzo zdezorientowane co do swojej roli w organizacji. Oznacza to, że lider data science jest odpowiedzialny za ochronę członków zespołu przed nieuzasadnionymi żądaniami i wyjaśnianie roli zespołu reszcie organizacji. Twój zespół musi ufać, że „będziesz miał ich plecy". Posiadanie pleców pracowników nie oznacza ślepej ich obrony za wszelką cenę. Oznacza to upewnienie się, że wiedzą, że cenisz ich wkład. Najlepszym sposobem na to jest upewnienie się, że członkowie zespołu mają ciekawe projekty do pracy i nie są przeciążeni projektami

o niejasnych wymaganiach, dziwnych przypadkach użycia lub nierealistycznych harmonogramach, którym towarzyszą niewystarczająca ilość danych i nieodpowiednie środowisko nauki o danych . Aby z czasem budować zaufanie, lider/menedżer ds. nauki danych powinien inwestować w otwartość i uczciwość. Analitycy danych to kompetentni ludzie przeszkoleni w zakresie analizowania i przetwarzania informacji. Zachowaj przejrzystość podczas całego procesu analizy danych, w tym rekrutacji, onboardingu, zapewniania codziennych operacji oraz omawiania wydajności i koncentracji zespołu – a także ogólnej strategii firmy. Bycie szczerym i otwartym we wszystkich aspektach może być bolesne, ale ma kluczowe znaczenie dla sukcesu zespołu. Spraw, aby informacje zwrotne były spójne i dwukierunkowe. Świetni naukowcy zajmujący się danymi doskonale sprawdzają się w wykrywaniu, czy masz na myśli to, co mówisz. Jeśli poprosisz o informację zwrotną, ale nie zamierzasz działać na podstawie opinii, wkrótce odkryjesz, że Twoi najlepsi pracownicy mogą chcieć odejść. Wartościowi pracownicy rzadko opuszczają firmę, ponieważ są niezadowoleni z samej firmy. Odchodzą z powodu złego przywództwa.

Znalezienie odpowiedniego lidera lub menedżera data science

Jak więc podejść do znalezienia odpowiedniego menedżera ds. nauki danych, aby zbudować odnoszący sukcesy zespół? Jednym z podejść jest użycie tego samego testu dla kandydatów na menedżera, co w przypadku analityków danych. Powinien zawierać te same wyzwania, ale powinieneś spodziewać się różnych wyników lub wyników testu. Dowiedz się, jak wygląda dobry wynik. Dokładnie zastanów się, jaki poziom biegłości wymagasz od lidera w dziedzinie nauki o danych. Na przykład, jeśli zatrudniasz menedżera zespołu uczenia maszynowego, niech menedżer rozwiąże określone zadanie z zakresu data science – takie, które dotyczy np. podobieństwa obrazów. Dobrym wynikiem tego ćwiczenia jest to, że menedżer potrafi wyjaśnić różne techniki potrzebne do rozwiązania zadania, zaczynając od głębokiego uczenia się i przechodząc do innych strategii uczenia maszynowego. Chodzi o to, aby ocenić zakres wiedzy kandydata, nawet jeśli nie prosisz go o stworzenie rzeczywistych modeli lub zakodowanie rozwiązania. Polecam również wdrożenie dwupoziomowego przywództwa w zespołach data science, gdzie jedna osoba będzie kierownikiem liniowym, wykazując zrozumienie ogólnych celów biznesowych i przyjmując odpowiedzialność za sukces zespołu, a druga osoba będzie ekspertem ds. data science, który raportuje do ogólnego kierownika liniowego. Ten rodzaj konfiguracji tworzy dynamiczne środowisko, które łączy w sobie głębokie przywództwo w zakresie technicznej analizy danych z dobrą orientacją biznesową. Te dwa poziomy przywództwa nie zawsze się zgadzają, ale taki jest cel konfiguracji. Menedżer zorientowany na biznes rzuci wyzwanie menedżerowi ds. technologii, a menedżer ds. technologii rzuci wyzwanie decyzjom dotyczącym współpracy. Niezwykle ważne jest, aby ci dwaj liderzy potrafili ściśle i otwarcie współpracować ze sobą, obdarzając się ogromnym zaufaniem. Należy to przetestować na jak najwcześniejszym etapie procesu rekrutacji.

Definiowanie warunków wstępnych dla odnoszącego sukcesy zespołu

Często można wymyślić dość specyficzne wymagania techniczne dla każdej roli w organizacji zajmującej się badaniem danych, ale musi istnieć wspólne zrozumienie tego, co jest wymagane, aby zespół ds. nauki danych odniósł sukces. Chociaż zestawy umiejętności technicznych są niezbędne dla skutecznego zespołu ds. analizy danych, istnieje znacznie bardziej krytyczny zestaw czynników sukcesu dla zespołu ds. analizy danych, o których musisz wiedzieć. Kilka następnych sekcji przeprowadzi Cię przez nie.

Rozwijanie struktury zespołu

Struktura zespołu ds. analizy danych ma kluczowe znaczenie dla skuteczności i wydajności zespołu. Nie należy lekceważyć znaczenia bliskiej współpracy między inżynierami danych a analitykami danych. Zazwyczaj te dwie różne role nie znajdują się w tym samym zespole, co oznacza, że jeszcze ważniejsze jest zapewnienie wydajnego środowiska współpracy między zespołami. Takie środowisko współpracy

powinno być w stanie obsłużyć wszystkie procesy analityczne od początku do końca w różnych systemach i organizacjach oraz zapewnić, że po drodze nie zostanie utracona produktywność. Jak można się domyślić, utrzymanie tych linii komunikacji nigdy nie jest łatwym zadaniem. Niemniej jednak powinieneś dążyć do zminimalizowania nieuniknionych przeniesień między inżynierami danych i analitykami danych, w procesie tworzenia płynnego przepływu pracy od przechwytywania danych do wdrażania modeli do produkcji. Zwiększa również wydajność, jeśli inżynierowie danych i analitycy danych są w stanie dzielić się spostrzeżeniami z użytkownikami biznesowymi z tego samego środowiska systemu danych.

Tworzenie infrastruktury

Infrastruktura data science musi być wysoce skalowalna – to znaczy powinna umożliwiać analitykowi danych skupienie się na analizie danych i budowaniu modeli, a nie zmaganie się z pozyskiwaniem danych lub ich obliczaniem. Wdrożenie bezpiecznej infrastruktury umożliwiającej działanie jest kluczem do sukcesu wszystkich inwestycji w sztuczną inteligencję — w tym inwestycji poczynionych w budowanie skutecznego zespołu zajmującego się analizą danych. Zbliżając się do prac nad infrastrukturą sztucznej inteligencji, wiele firm zapomina o zdefiniowaniu jasnej strategii analizy danych, która jest uzgadniana w całej firmie przed rozpoczęciem pracy. Pamiętaj, że źle przemyślana strategia infrastruktury nieuchronnie zwiększy złożoność i koszty rozwoju i operacji (DevOps). Jeśli nie postępujesz zgodnie z planem infrastruktury, podchodząc do niego w sposób bardziej budowany na bieżąco, złożoność może być trudniejsza do opanowania pod względem konfiguracji i utrzymania infrastruktury w czasie, zarządzania aktualizacjami i poprawkami, skalowania infrastruktury o rosnących ilościach danych, a także w dostarczaniu wysokowydajnej infrastruktury dużym zespołom analityków danych, czasami rozproszonym na dużym obszarze geograficznym. Jeśli chodzi o zwiększanie wydajności przetwarzania danych, przetwarzanie rozproszone jest często przedstawiane jako najlepsze rozwiązanie zapewniające wydajność na dużą skalę. . Należy jednak pamiętać, że złożoność zarządzania przetwarzaniem rozproszonym, od brzegów chmury do brzegów na poziomie urządzenia lub komponentu, wymaga specjalnych umiejętności, które mogą być trudne do zdobycia.

Zapewnienie dostępności danych

Umożliwienie dostępu do potrzebnych danych może wydawać się oczywistym zadaniem do ustalenia priorytetów, ale możesz być zdumiony, jak często jest to poważny problem dla naukowców zajmujących się danymi. Czasami problemy z dostępem są spowodowane czynnikami zewnętrznymi, którymi trudniej zarządzać, takimi jak nowe ograniczenia prawne dotyczące wykorzystania danych, ale wiele razy problemy pojawiają się, ponieważ firma nie ma wspólnej i uzgodnionej strategii umożliwiającej dostęp na spójnym poziomie. Zapewnienie, że dane są dostępne dla analityków danych do przeprowadzania analiz i budowania modeli, może wydawać się prostym zadaniem – na przykład o wiele prostsze niż poszukiwanie danych i naprawianie uszkodzonych środowisk danych. Zadanie nie jest wcale proste. W rzeczywistości zawodna dostępność danych jest powszechnym źródłem frustracji naukowców zajmujących się danymi, co powoduje, że odchodzą do innych firm, które lepiej zaspokajają potrzeby w zakresie danych. To naprawdę niepotrzebne ryzyko narażania Twojej firmy, więc upewnij się, że nie skończysz w takiej sytuacji. Jeśli uda Ci się zdobyć jednego z niewielu starszych analityków danych, którzy są dostępni na rynku, zrób wszystko, co jest potrzebne, aby zatrzymać tę osobę! Jednym ze sposobów na zminimalizowanie ryzyka jest obniżenie ambicji i rozpoczęcie od danych, które są już dla Ciebie dostępne. Następnie możesz pracować równolegle, aby upewnić się, że masz wszystkie niezbędne prawa do danych, że dane są bezpieczne i prawidłowo zarządzane oraz że masz środki do gromadzenia i przygotowywania danych dla analityków danych.

Obstawanie przy ciekawych projektach

Ważnym aspektem, który należy wziąć pod uwagę, starając się zapewnić sukces zespołowi zajmującemu się analizą danych, jest zaoferowanie analitykom danych interesujących wyzwań i trudnych problemów do rozwiązania. . Jeśli Twoje problemy są zbyt proste, starsi specjaliści ds. danych będą się nudzić. Jeśli problemy są niemożliwe do rozwiązania z powodu braku danych lub nieprawidłowo działającego środowiska, analitycy danych mogą zostać trochę dłużej i spróbować to naprawić, ale w końcu odejdą, jeśli nie będą wspierani. Analitycy danych mają obecnie najbardziej poszukiwane kompetencje na rynku. Znają swoją wartość i Ty też powinieneś. Nie zmarnują swojej wartości na firmę, która nie może zaoferować im odpowiednich możliwości bycia częścią odnoszącego sukcesy zespołu data science, pracującego nad ciekawymi i ambitnymi projektami. Tak więc, jeśli jesteś ekspertem musisz być gotowy, aby zaproponować im ciekawe pomysły i złożone problemy.

Promowanie ciągłego uczenia się

Ponieważ data science to obszar podlegający ciągłym zmianom, naukowcy zajmujący się danymi chcą (i powinni) być na bieżąco z nowymi technikami, metodami i trendami pojawiającymi się na rynku. Wszyscy pracownicy naukowi o danych muszą mieć możliwość kontynuowania nauki w ramach swojej pracy. Jednym ze sposobów na umożliwienie ciągłego uczenia się jest umożliwienie członkom zespołu ds. nauki danych uczestniczenia w różnych konferencjach technicznych dotyczących nauki o danych lub projektach lub miejscach open source. Formalne szkolenie niekoniecznie nadąża za wymaganym tempem, a nauka o danych nadal jest w dużej mierze napędzana z perspektywy open source, zarówno pod względem technologii, jak i metodologii.

Zachęcanie do badań naukowych

Członkowie zespołu zajmującego się analizą danych powinni mieć również możliwość spędzania czasu na badaniach, w tym pisaniu białych ksiąg i uczestniczeniu w sesjach czytania białych księgi w oparciu o pracę wykonaną przez kolegów. Ponieważ obszar nauki o danych szybko się rozwija, ważne jest, aby być na bieżąco z nowymi metodami, badaniami i rozwojem technologii, a także uczestniczyć w dyskusjach o standaryzacji i inicjatywach open source.

Budowanie zespołu

Tworząc zespół analityków danych, zawsze powinieneś skupiać się na zakresie i celu wokół pytań, które odzwierciedlają strategiczne cele biznesowe Twojej firmy. Ten nacisk może oznaczać na przykład przyciągnięcie nowych klientów, automatyzację procesów lub wprowadzenie do portfolio nowych, innowacyjnych produktów z zakresu danych. Chcesz mieć możliwość jak najwcześniejszego zaangażowania interesariuszy i decydentów w swoje ambicje w zakresie nauki o danych i mieć możliwość argumentowania za zwrotem z inwestycji (ROI). Powinieneś na przykład rozważyć następujące pytania:

* Co zapewni optymalny wynik i jakie są zachęty?

* Jak zespół ds. analityki danych będzie współpracował z interesariuszami?

* Jak podchodzi się do inwestycji w infrastrukturę w odniesieniu do priorytetów, procesu zatwierdzania, finansowania i zarządzania?

* Jak będą alokowane koszty?

* Jak będą działać zespoły biznesowe, prawne, IT i danych bez tworzenia niedopuszczalnego ryzyka?

Skuteczna strategia komunikacji, jasno określone priorytety i umiejętność zarządzania oczekiwaniami są niezbędne, niezależnie od tego, jakie podejście zdecydujesz się na strukturyzację i rozwój swoich możliwości w zakresie analizy danych.

Rozwój inteligentnych procesów rekrutacyjnych

Wielu profesjonalistów próbuje włamać się do „najseksowniejszego zawodu XXI wieku", więc jako menedżer ds. nauki danych na pewno otrzymasz wiele aplikacji i będziesz musiał być selektywny. Wykorzystaj to i bądź wybredny we właściwy sposób. Upewnij się, że dbasz o swój proces rekrutacji. Jednym wspólnym obszarem, w którym firmy ponoszą porażkę, jest kompromis między perspektywą krótko- i długoterminową. Na przykład, łatwo jest zacząć myśleć, że spóźniłeś się na grę naukową o danych i dlatego nie ma wystarczająco dużo czasu, aby zrekrutować wszystkich ludzi, których potrzebujesz. Takie podejście jest ogromnym błędem. Jeśli uważasz, że nie ma wystarczająco dużo czasu na znalezienie odpowiedniego talentu i przeanalizowanie procesu rozmowy kwalifikacyjnej i onboardingu, prawdopodobnie nie masz też czasu na zarządzanie nowym pracownikiem. Stworzenie świetnego procesu rekrutacyjnego opłaca się na dłuższą metę. Jak więc wygląda świetny proces rekrutacji? Po pierwsze, nie skupia się tylko na umiejętnościach technicznych. Umiejętności społeczne, takie jak empatia i komunikacja, są niedoceniane w nauce o danych i dyscyplinach, z których zwykle wychodzą naukowcy zajmujący się danymi, ale są one kluczowe dla zespołu. Niech to będzie częścią Twojego procesu rekrutacyjnego. Zamiast skupiać się na tym, czy potrafisz się dogadać z kandydatem, zadaj sobie pytanie, czy istnieje soczewka, przez którą ta osoba widzi świat poszerzający wiedzę i wartość zespołu. Ten wymiar powinien być ceniony tak wysoko, jak cenisz inne atrybuty, takie jak umiejętności techniczne i znajomość domeny. Dlatego ważne jest, aby priorytetowo traktować różnorodność. Obejmuje to różnorodność dyscypliny akademickiej i doświadczenia zawodowego, ale także przeżyte doświadczenie i perspektywę. Jeśli chodzi o różnorodność, kilka obszarów wyróżnia się jako szczególnie ważne w nauce o danych. Po pierwsze, nie powinieneś skupiać się na zatrudnianiu osób starszych. Są one nie tylko drogie i trudne do zdobycia, ale mniej doświadczeni pracownicy zwykle nie są tak podatni na historię i dlatego mogą łatwiej zadawać pytania o to, dlaczego sprawy są robione w określony sposób. Zadawane pytania są bardziej wolne od zwykłego założenia, o których bardziej doświadczeni specjaliści w pewnym momencie przestają być świadomi Nietrudno popaść w obsesję na punkcie określonego sposobu robienia rzeczy i zapomnieć o zastanowieniu się, czy preferowane podejście jest nadal najlepszym rozwiązaniem nowego zadania. Analitycy danych wywodzą się z różnych środowisk akademickich, w tym informatyki, matematyki, fizyki, statystyki i wielu innych. Najważniejszy jest kreatywny umysł połączony z pierwszorzędnymi umiejętnościami krytycznego myślenia. Inną ważną kwestią jest zatrudnianie osób, których mocne strony wzajemnie się uzupełniają, zamiast budowania zespołu, którego wszyscy członkowie wyróżniają się w tej samej dziedzinie. Posiadanie osoby, która zawsze widzi pełny obraz, kogoś, kto potrafi artykułować historie za pomocą danych, oraz współpracującego ze sobą kreatora wizualizacji, którzy mogą współpracować, aby uzyskać wyniki, których nikt nie byłby w stanie osiągnąć niezależnie. Chodzi o to, aby w jak największym stopniu wykorzystać te uzupełniające się umiejętności, aby w porządku tworzyć świetne rozwiązania, o których nikt wcześniej nie pomyślał.

Jednym ze sposobów upewnienia się, że zespół faktycznie działa jako zespół i współpracuje, jest poproszenie członków zespołu o regularne czytanie kodu i sprawdzanie swoich modeli. To świetny sposób na wspieranie współpracy zespołowej skupionej wokół dyskusji technicznych. Upewnienie się, że Twój zespół angażuje się w tego rodzaju zaplanowane działania oparte na współpracy, pomaga w maksymalnym wykorzystaniu tego rodzaju różnorodności w zespole. . Wreszcie, ważne jest, aby zbudować zespół, który będzie odzwierciedlał osoby, których dane analizujesz: na przykład, jeśli analizujesz dane z mediów społecznościowych z aplikacji używanej tylko przez kobiety, zespół data

science nie może składać się tylko z mężczyzn. Tylko w ten sposób zapewnisz sobie prężny zespół, który będzie zadawać lepsze pytania i mieć szerszą perspektywę, z której można je zadawać. W ten sposób martwe punkty każdej osoby są pokrywane przez wcześniejsze doświadczenia i zestaw umiejętności innego członka zespołu.

Pozwól swoim zespołom rozwijać się organicznie

Rozpoczynając tworzenie zespołów zajmujących się analizą danych, należy zacząć od małych, wszechstronnych zespołów, w których każdy zespół zachęca członków zespołu do „noszenia wielu kapeluszy" i wykonywania wielu różnych rodzajów analiz danych. W miarę dojrzewania zespołów i udowadniania swojej wartości na różne sposoby, role staną się bardziej zdefiniowane, a niektóre działania prawdopodobnie zostaną przeniesione do innych zespołów. Przykładem jest to, że gdy nauka o danych jest bardziej ugruntowana i zrozumiana w firmie, działania związane z infrastrukturą, operacjami, bezpieczeństwem itd. są zwykle obsługiwane oddzielnie, co pozwala zespołom analitycznym na większą specjalizację. A jednak zbyt wcześnie ostrzegałem przed wyspecjalizowanymi zespołami. Specjalizacja zespołu działa tylko wtedy, gdy obowiązki zespołu są jasno określone i wprowadzono efektywne sposoby pracy, aby zrównoważyć brak szybkości i dodatkowe koszty związane ze współpracą wielu zespołów. Ponieważ trudno jest znaleźć utalentowanych naukowców zajmujących się danymi, pozwól zespołom organicznie ewoluować w kierunku większej specjalizacji. (Moja mantra brzmi: „Specjalizacja nadejdzie; nie ma potrzeby się spieszyć.") I nie zapomnij połączyć doświadczonych analityków danych z mniej doświadczonymi. Zwykle łatwo jest znaleźć inteligentnych i zmotywowanych analityków danych, którzy (przy odrobinie dedykowanego coachingu) chcą dowiedzieć się więcej. Obejmuje to naukę, jak odpowiednio sformułować problem, zarządzać małym projektem, opracować i przeszkolić model, zintegrować z interfejsami API i wprowadzić projekt do produkcji. A jeśli w końcu uda ci się pozyskać kilku utalentowanych analityków danych, masz już strukturę umożliwiającą ich zintegrowanie z tymi zespołami uczącymi się, aby upewnić się, że możesz w pełni wykorzystać ich ogromne doświadczenie.

Łączenie zespołu z celem biznesowym

Analitycy danych są na ogół zorientowani na cel, więc aby jak najlepiej wykorzystać swój czas, muszą dobrze rozumieć zadanie i wierzyć w cel biznesowy stojący za projektami, do których zostali przydzieleni. . Musi być również miejsce na ustalenie priorytetów pomysłów pochodzących od samego zespołu zajmującego się badaniem danych. Zakotwiczenie pracy zespołu w kontekście strategii data science i ogólnych celów biznesowych to jedno z najważniejszych zadań, jakie musi wykonać lider zespołu data science. Niestety nie zawsze jest to łatwe zadanie. Projekt data science często zaczyna się od pytania kogoś spoza zespołu. Często jednak pytanie, które zadaje dana osoba, nie jest dokładnie tym, co zdaniem zespołu zajmującego się analizą danych powinno zostać zbadane. W związku z tym zarządzanie zespołem zajmującym się badaniem danych zwykle wymaga wielu dyskusji i dopracowywania pytań od interesariuszy, aby lepiej zrozumieć informacje, których faktycznie potrzebują, i sposób ich wykorzystania. Nie pozwól, aby pytania lub prośby stały się projektami Twojego zespołu, dopóki się nie dowiesz dokładnie to, co interesariusz chce zrozumieć i jak to zostanie wykorzystane. Posiadanie jasnych celów dotyczących pytań związanych z danymi, które pojawiają się na Twojej drodze, jest jedną z najważniejszych rzeczy, które możesz zapewnić swojemu zespołowi. Jednocześnie interesariusze nie zawsze będą w stanie odpowiedzieć na pytania zespołu zajmującego się analizą danych, nawet jeśli będą chcieli. Mogą nie znać pełnego kontekstu lub długoterminowego celu pytania, które zadają, jak wyglądałby gotowy produkt z zakresu nauki o danych, a nawet jak go zastosowali. Aby pokonać tę przeszkodę, wypróbuj niektóre czynności opisane na tej liście:

* Zrozum wartość biznesową. Upewnij się, że menedżerowie produktu rozumieją wartość biznesową nauki o danych. Muszą zrozumieć to pod kątem tego, w jaki sposób napędza innowacyjność i wartość biznesową, i być gotowym do nadawania im priorytetów, nawet jeśli nie rozumieją wszystkich jego aspektów. .

* Weź udział w spotkaniach strategicznych. Upewnij się, że członkowie zespołu data science są regularnie zapraszani na spotkania dotyczące produktów i strategii. W ten sposób mogą być częścią procesu twórczego, a nie tylko odpowiadać na prośby. (Nie zaszkodzi również zadawać wiele pytań). Integracja analityków danych z dialogiem biznesowym przyczynia się również do lepszego zrozumienia przez firmę podejścia opartego na badaniach danych do możliwości biznesowych.

* Współpracuj ponad granicami organizacyjnymi. Oczywiście zespoły data science nie są jedynymi, które poszukują współpracy z interesariuszami biznesowymi, niemniej jednak należy im nadać priorytet przy podejmowaniu decyzji dotyczących tego, kto współpracuje z jakimi zespołami w różnych działach biznesowych. Znacząco zwiększa motywację analityków danych, gdy:

wiedzą nie tylko, że przyczyniają się do czegoś, co rozumieją, ale także, że ich wkład jest ceniony i traktowany priorytetowo przez organizację jako całość.

* Udowodnij wartość. Zespoły zajmujące się analizą danych muszą być w stanie udowodnić swoją wartość. Jeśli zespół wpadnie na nowy pomysł, organizacja i kierownik ds. nauki danych muszą rzucić zespołowi pytania typu „Jak możemy udowodnić, że to przyczynia się do naszej działalności?” oraz „Skąd wiemy, że to najlepsze rozwiązanie?”

Zbliżanie się do konfiguracji organizacji zajmującej się badaniem danych

Siła rewolucji danych pozostaje silna, a firmy różnej wielkości aktywnie budują i rozszerzają swoje zespoły zajmujące się analizą danych na różne sposoby. Firmy zdają sobie sprawę, że muszą być w stanie wykorzystać dane, aby poprawić podejmowanie decyzji i efektywność operacyjną, ale widzą również, że muszą mieć możliwość tworzenia nowych produktów i procesów w oparciu o spostrzeżenia oparte na danych. Aby to osiągnąć, firmy muszą wprowadzić na poziomie korporacyjnym niezbędne zmiany organizacyjne i kulturowe, które będą potrzebne, aby odnieść sukces. Struktura organizacyjna potrzebna zespołowi zajmującemu się analizą danych różni się w zależności od wielkości firmy, liczby różnych funkcji biznesowych, rozmieszczenia geograficznego, kultury firmy i innych podobnych aspektów. Istnieje jednak kilka wspólnych czynników, które należy wziąć pod uwagę podczas integrowania analityków danych z większą organizacją opartą na danych. W tym rozdziale przyjrzymy się tym czynnikom.

Znalezienie odpowiedniego projektu organizacyjnego

Aby ustalić najlepszą konfigurację zespołu dla funkcji analizy danych z perspektywy organizacyjnej, rozważ pięć głównych zadań:

* Zdecyduj, w którym miejscu organizacji chcesz umieścić funkcję analizy danych, a następnie ustal optymalną konfigurację. Na przykład, czy chcesz model scentralizowany czy zdecentralizowany?

* Dopasuj projekt organizacyjny do jego funkcji biznesowych. W ramach tego procesu definiujesz i wdrażasz wydajną strukturę zarządzania opartą na otwartości i przejrzystości wobec biznesu, a także musisz budować struktury decyzyjne, które pozwalają firmie wpływać na problemy i możliwości, które zespoły ds. analityki danych powinny traktować priorytetowo .

* Upewnij się, że konfiguracja organizacyjna jest zgodna z ogólną strategią biznesową, w tym konfiguracją partnerską. Struktura powinna również umożliwiać zespołom zajmującym się nauką danych łatwe łączenie się z potrzebnymi ekosystemami nauki o danych związanymi z danymi, narzędziami i modelami. .

W nauce o danych bycie częścią ekosystemu jest niezbędne. Ewolucja postępuje zbyt szybko, jest zbyt skomplikowana i kosztowna, aby zająć się nią samemu. Szybki rozwój wymaga od firm udostępniania, a także ponownego wykorzystywania komponentów i możliwości w ramach ekosystemu nauki o danych. Wiąże się to z podejściem organizacyjnym, które umożliwia uczestnictwo i wkład w społeczności open source i inne otwarte struktury.

* Określ, jakich ról potrzebujesz w zespole i ilu członków zespołu przypada na każdą rolę. Zastanów się, ile ról można obsadzić w ramach rekrutacji wewnętrznej. Ten etap obejmuje określenie wymaganego poziomu szkolenia dla zespołu i zapewnienie zaangażowania kadry kierowniczej firmy w szkolenie.

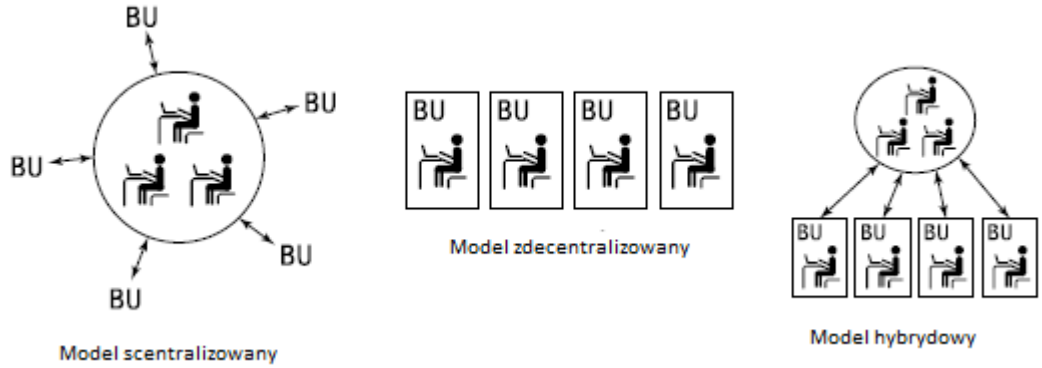
* Skaluj zespół w perspektywie długoterminowej. Zastanów się, jak do tego podejdziemy i jakie są terminy. Czy rekrutacja zewnętrzna będzie częścią scale-up, czy będzie opierać się na przykład na organicznym wzroście. Rozważ również standaryzację procesów i struktur zarządzania przed zwiększeniem skali, aby lepiej zarządzać wzrostem.

To jest ogólny obraz - nagi zarys tego, co musi się wydarzyć. Kolejne kilka sekcji zawiera szczegóły.

Projektowanie funkcji data science

Czas na bliższe przyjrzenie się, jak podejść do pierwszego z głównych zadań do rozważenia: Określ konfigurację funkcji analizy danych w Twojej organizacji. Opcje konfiguracji wahają się od modelu

scentralizowanego do modelu wysoce rozproszonego. Model scentralizowany jest czasami nazywany modelem współdzielonym lub centrum doskonałości, w którym wszyscy analitycy danych pracują razem w tej samej jednostce organizacyjnej. Ten model zachęca do współpracy i spójności zespołu ds. analizy danych, umożliwiając członkom zespołu wzajemne odbijanie pomysłów i uzyskiwanie szybkiej pomocy w przypadku pytań związanych z pracą w zakresie analizy danych. Taka konfiguracja pozwala również na szkolenie w miejscu pracy dla mniej doświadczonych analityków danych. W dużych firmach, w których jednostki biznesowe są silne i niezależne finansowo lub w których jednostki biznesowe prowadzą różne rodzaje działalności, istnieje tendencja do wyboru modelu rozproszonego lub zdecentralizowanego. W modelu rozproszonym jednostki biznesowe (BU) same zatrudniają własny zestaw analityków danych. Pozwala to analitykom danych na bliższą i bardziej ciągłą współpracę z menedżerami, inżynierami i interesariuszami. W tej konfiguracji naukowcy zajmujący się danymi mają możliwość zdobycia cennej wiedzy specjalistycznej w danej dziedzinie, a także wglądu w rzeczywiste problemy, z którymi borykają się ich koledzy każdego dnia. Ta konfiguracja zwykle zwiększa możliwości zespołów zajmujących się analizą danych, ale istnieje również ryzyko zduplikowania części pracy wykonywanej przez zespoły, ponieważ koordynacja i współpraca między zespołami w różnych jednostkach biznesowych są zwykle mniej wyraźne. Gdy małe zespoły analityków danych lub nawet pojedynczy analitycy danych są osadzani w różnych jednostkach biznesowych, może to mieć efekt uboczny polegający na nadmiernej izolacji analityków danych. Jeśli analitykom danych brakuje współpracowników zajmujących się analizą danych, z którymi można by dyskutować i rozwijać, wydajność spada, a motywacja spada. Istnieje również tendencja do kwestionowania odizolowanych naukowców zajmujących się danymi, ponieważ nauka o danych nie jest ogólnie rozumiana w tradycyjnych organizacjach zajmujących się oprogramowaniem. Najlepszym sposobem na złagodzenie tego problemu jest uniknięcie zbytniego rozpowszechniania analityków danych w organizacji. . Hybrydowy model organizacyjny, składający się z jednej jednostki centralnej z jednym lub kilkoma zespołami data science połączonych z wieloma wyspecjalizowanymi zespołami osadzonymi w jednostkach biznesowych w ramach kilku projektów, zaczyna wyłaniać się jako najlepiej sprawdzona strategia dla wielu firm. Z punktu widzenia analityka danych, główną korzyścią tego modelu jest możliwość pracy bliżej rzeczywistych funkcji biznesowych, co może oznaczać zdobycie większej wiedzy na temat tego, jak rzeczy działają w praktyce. . Ponieważ model hybrydowy przyjmuje aspekty zarówno podejścia scentralizowanego, jak i zdecentralizowanego, równoważy zalety i wady tych modeli. W modelu hybrydowym jednostka centralna służy jako hub promujący udostępnianie i ponowne wykorzystywanie najlepszych praktyk oraz propaguje je w każdym z rozproszonych zespołów analityki danych. Innym, rzadziej stosowanym sposobem konfigurowania modelu hybrydowego jest przypisanie wszystkich analityków danych do określonych jednostek biznesowych, ale raportowanie do wspólnej scentralizowanej jednostki nauki o danych. Rysunek przedstawia graficzną ilustrację trzech różnych sposobów konfigurowania funkcji analizy danych w firmie.



Model scentralizowany

Model zdecentralizowany

Model hybrydowy

To, na który z nich się zdecydujesz, zależy od takich aspektów, jak dojrzałość Twojej firmy, ale także od Twojego celu w zakresie nauki o danych i dostępności różnych kompetencji w zakresie nauki o danych. Jeśli tworzysz funkcję data science, która jest bardziej skoncentrowana na analityce niż na przykład uczenie maszynowe i sztuczna inteligencja, dostęp do wymaganych kompetencji jest znacznie mniejszym problemem niż w przestrzeni uczenia maszynowego. Kiedy możesz liczyć na dostęp do wszystkich potrzebnych kompetencji, możesz zaprojektować organizację według potrzeb, bezpośrednio od samego początku. Najlepszy model może następnie być wybierane w zależności od wielkości firmy, branży i ambicji związanych z analityką, a nie ograniczać się brakiem wymaganych kompetencji w zakresie analizy danych. W przestrzeni uczenia maszynowego i sztucznej inteligencji, gdzie dostęp do naukowców zajmujących się danymi jest ograniczony, a ogólne zrozumienie, w jaki sposób techniki ML/AI będą wykorzystywane w firmie jest znacznie mniej rozumiane, sytuacja jest inna. Możesz dojść do wniosku, że zdecentralizowany model jest tym, czego chcesz, ale zdajesz sobie sprawę, że przy tak niewielkiej liczbie naukowców zajmujących się danymi, których możesz zrekrutować, rozprzestrzenisz się w firmie za bardzo i nie będziesz w stanie osiągnąć swoich celów przy użyciu tego podejścia. Dlatego rozpoczęcie od scentralizowanego modelu, który oferuje odpowiednią liczbę zespołów zajmujących się analizą danych współpracujących ze sobą, aby dokonać zmian, może być jedynym sposobem na rozpoczęcie. Jednak z czasem model scentralizowany może przekształcić się w model hybrydowy, ponieważ pula naukowców zajmujących się danymi rośnie organicznie, ale także w wyniku rekrutacji. Ponadto, gdy obszar uczenia maszynowego stanie się bardziej towarem w branży, analitycy danych mogą stać się integralną częścią każdej jednostki rozwojowej i nie ma już potrzeby, aby scentralizowana jednostka trzymała to wszystko razem.

Ocena korzyści płynących z centrum doskonałości dla nauki o danych

Wybór scentralizowanego modelu dla funkcji nauki o danych jako preferowanej lub niezbędnego punktu wyjścia naprawdę oznacza utworzenie jednostki centralnej w firmie, znanej również jako centrum doskonałości (CoE) w zakresie nauki o danych. Chociaż istnieją różne opinie na temat skuteczności i wydajności centrów doskonałości w branży, nadal istnieje kilka udowodnionych korzyści wynikających z zastosowania scentralizowanego podejścia podczas budowania funkcji data science w firmie, jak wymieniono poniżej.

* Szybkość: CoE w zakresie analizy danych jest niezbędne do przyspieszenia podejścia opartego na danych w całej firmie na dużą skalę. Znacznie skraca czas wdrożenia, a tym samym czas wprowadzania produktów na rynek potrzebny do wdrożenia nowych produktów i usług opartych na danych.

* Ponowne wykorzystanie: CoE ułatwia dzielenie się najlepszymi praktykami i metodologiami między różnymi zespołami w organizacji.

* Ewolucja: CoE ułatwia przydzielanie czasu zespołowi, aby być na bieżąco z trendami rynkowymi i ewolucją technologii w nauce o danych.

* Zestawy umiejętności: CoE wyposaża firmę w niezbędny zestaw umiejętności w zakresie analizy danych, gdy jest to potrzebne.

* Terminologia: scentralizowana funkcja pomaga zapewnić, że cała organizacja używa wspólnej terminologii i „mówi tym samym językiem", opracowując wspólny zestaw standardów przy jednoczesnym wdrażaniu metod i technik analizy danych.

* Kultura: CoE może służyć jako siła napędowa zmian kulturowych, aby stać się organizacją opartą na danych i nastawioną na działanie, jeśli chodzi o korzystanie z technik analizy danych.

Nie jest konieczne ustawianie scentralizowanej funkcji CoE ściśle według modelu scentralizowanego. Możesz użyć lekkiej wersji modelu hybrydowego i nadal czerpać korzyści z centralnego CoE. Posiadanie konfiguracji ważonej 80-20 może nadal działać z dostępnością zasobów analizy danych - co oznacza, że masz 80% COE centralnie zlokalizowanej i 20% wbudowanej w jednostki biznesowe, a następnie możesz pozwolić jej rosnąć w czasie do 50-50 a nawet podejście 20-80

Identyfikacja czynników sukcesu centrum doskonałości data science science

Jeśli chcesz mieć pewność, że CoE przyniesie Twojej firmie rzeczywistą wartość biznesową, musisz upewnić się, że z powodzeniem osiągnie te trzy różne cele:

* CoE musi być postrzegane jako czynnik umożliwiający wartość biznesową: Innymi słowy, CoE należy uznać za zdolne do umożliwienia głębokiej zmiany kulturowej wokół wykorzystania automatyzacji, analityki, uczenia maszynowego i sztucznej inteligencji. Powinno to obejmować zdolność do przyciągania jak najlepszych talentów i ogólnie postrzegane jako czynnik wartości.

* CoE musi być postrzegane jako autonomiczny podmiot, ale w pełni wspierany przez kierownictwo. CoE musi funkcjonować jako niezależna jednostka, która jest właścicielem narzędzi, standardów i metodologii w nauce o danych. Od nikogo w organizacji nie może należeć optymalizacja infrastruktury i zestawów narzędzi pod kątem konkretnych korzyści związanych z produktami lub usługami w różnych obszarach, a nie pod kątem ogólnej wartości firmy. CoE powinno również stale angażować wspierających seniorów , przywództwo w dalszym wdrażaniu nauki o danych w organizacji w ramach jej codziennej działalności.

* RE musi być zorientowana na oddziaływanie. Jednostka powinna nadać priorytet pracom nad przypadkami użycia zgodnymi z priorytetami strategicznymi i wykorzystać opartą na wartościach mapę drogową przypadków użycia o wymiernym wpływie.

Upewnienie się, że Twoje CoE osiąga wszystkie trzy cele, znacznie zwiększa szanse na ustanowienie wydajnej i skutecznej funkcji centrum doskonałości data science.

Zastosowanie wspólnej funkcji nauki o danych

Zbliżając się do ustanowienia wspólnej dla całej firmy funkcji analizy danych, musisz podjąć kilka ważnych decyzji. W przypadku większych przedsiębiorstw lokalizacja geograficzna jest zwykle jedną z głównych decyzji, którymi należy się zająć. A jeśli zaczniesz od ważnego aspektu, gdzie powinna zostać umieszczona wspólna funkcja data science, może się wydawać, że dobrym pomysłem może być kolokacja jej z siedzibą główną firmy. Ponieważ jednak jest to strategicznie ważna decyzja, upewnij się, że opiera się ona na rzeczywistych danych, a także na przemyślanych kryteriach wyboru. Przyjrzyjmy się temu nieco dalej.

Wybór lokalizacji

W tradycyjnym centrum doskonałości skoncentrowanym na analityce, a nie na inteligencji maszynowej, faktyczna lokalizacja zwykle nie jest ważnym czynnikiem. A jednak, ponieważ wspólna funkcja nauki o danych zwykle wymaga nie tylko umiejętności niszowych, które są trudne do zdobycia, ale także dostępności talentów dla przyszłego potencjału, a także obecności ekosystemu branżowego w celu zapewnienia dostosowania i wpływu na standardy rynkowe w miarę ich ewolucji, umieszczając funkcja data science w tym samym miejscu, co centrala, może nie mieć większego sensu. Oczywiście zależy to jednak od tego, jak dobrze siedziba spełnia wybrane kryteria wyboru lokalizacji. Musisz także wziąć pod uwagę inne aspekty związane z lokalizacją. Poniższa lista opisuje cztery główne kryteria

wyboru lokalizacji, wymienione w bardziej uporządkowany sposób, w tym oszacowanie ich znaczenia w ogólnej decyzji:

* Talent: odnosi się do dostępności nowej i istniejącej puli talentów w rzeczywistych liczbach w danej lokalizacji lub w jej pobliżu. To kryterium jest bardzo ważne - niech niesie ze sobą około połowy całkowitej wagi decyzji. Obejmuje takie aspekty, jak absolwenci związani z technologiami, istniejąca pula talentów w potrzebnych rolach i obszarach kompetencji oraz jakość edukacji w tym obszarze. .

* Ekosystem: Odnosi się do zestawu czynników, które pozwalają na płynniejsze działanie zespołów data science, dzięki jakościowym cechom lokalizacji, takim jak zdolność do zatrzymywania i przyciągania talentów, dostępność analityków danych i innych kluczowych kompetencji, dostępność najnowszych technologii, zdolność do innowacji, liczba start-upów i potencjał finansowania kapitału podwyższonego ryzyka. To ważne kryterium powinno nieść około jednej trzeciej całkowitej wagi decyzji.

* Infrastruktura: chodzi o praktyczną realizację środowiska nauki o danych, w tym dobrą łączność, łącza danych, dostępność samych danych i kanały cyfrowe, z których można korzystać. (Ważyłbym to na około 10 procent). Chodzi o upewnienie się, że infrastruktura sprawia, że środowisko nauki o danych jest łatwe w obsłudze i obsłudze również dla pracowników, w tym aspekty takie jak dostępność nieruchomości, atrakcyjne środowisko biznesowe i mieszkalne oraz lotniska i inne środki transportu wspierające dostępność lokalizacji.

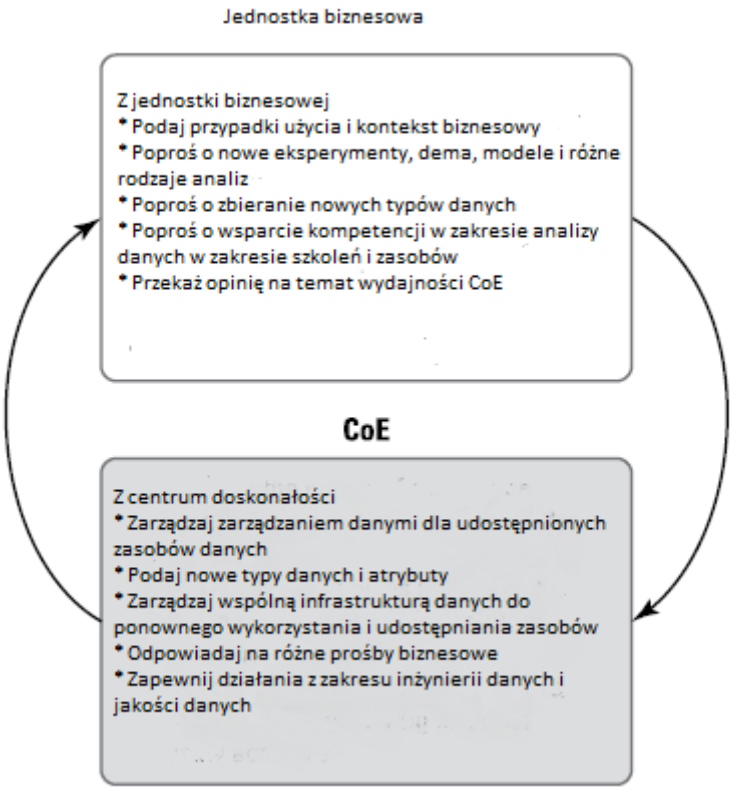
* Koszt: Odnosi się to do wskaźników kosztów głównie w trzech obszarach - koszty pracy i nieruchomości oraz inflacja wynagrodzeń. Jednak to kryterium jest znacznie mniej ważne niż dostępność odpowiedniego talentu i kompetencji i dlatego niesie tylko około 10 procent wagi decyzji.

Nawet jeśli planujesz scentralizowaną konfigurację funkcji analizy danych, nadal możesz rozdzielić tę funkcję na kilka lokalizacji geograficznych. Może to być nawet wskazane, zwłaszcza jeśli oznacza to dostęp do większej liczby lepszych talentów, a także poprawę zdolności do korzystania z istniejących ekosystemów nauki o danych. (Faktem jest, że istnieje niewielka szansa na spełnienie wszystkich kryteriów w jednej lokalizacji, więc i tak możesz być zmuszony do rozgałęzienia).

Zbliżające się sposoby pracy

Po uzgodnieniu lokalizacji funkcji analizy danych nadszedł czas, aby zacząć myśleć o tym, jak sprawić, by funkcja analizy danych działała dla Ciebie. Ważnymi aspektami do rozważenia są oczywiście to, jak upewnić się, że nowa funkcja jest zarówno skuteczna (robienie właściwych rzeczy), jak i wydajna (robienie rzeczy właściwie), a także utrzymanie pozostałej części organizacji w poczuciu, że udowadnia swoją wartość. Jeśli zostanie prawidłowo skonfigurowana, nowa funkcja analizy danych będzie odgrywać ważną rolę w organizacji, łącząc różne jednostki biznesowe i znajdując jasne i uzgodnione obowiązki związane z funkcją informatyczną.

Rysunek wyjaśnia, jak mogą wyglądać niektóre interakcje między jednostkami biznesowymi a wspólną funkcją analizy danych.



Rysunek ilustruje, jak może działać interakcja, ale pamiętaj, że podczas dzielenia ról i obowiązków należy wziąć pod uwagę wiele aspektów. Nie bądź zbyt surowy, jeśli chodzi o to, co robi która jednostka. Jeśli starsi analitycy danych są osadzeni w jednostkach biznesowych i są w stanie kierować pewnymi strategicznymi obszarami, pozwól im. Daj ludziom możliwość przejęcia kontroli i wprowadzania zmian dzięki analizie danych. (W każdym razie jest to cel długoterminowy). Po prostu upewnij się, że te rozproszone grupy nie są zbyt izolowane i odłączone od wspólnej funkcji nauki o danych. (Jednym ze sposobów zapewnienia niezbędnego stopnia integracji jest naleganie, aby takie grupy dzieliły się modelami i spostrzeżeniami przez wspólny zespół, w tym przestrzegając uzgodnionych standardów i zasad firmy).

Zachęcaj pracowników do rozwijania swoich umiejętności w zakresie analizy danych poprzez udział w zalecanych programach szkoleniowych oraz badanie danych i przypadków użycia w różnych obszarach. W miarę powiększania się puli pracowników zajmujących się analizą danych, wspólna praca nad realizacją funkcji analizy danych powinna naturalnie prowadzić do tego, że różne zespoły i grupy osób będą znajdować sposoby na nawiązywanie kontaktów i dzielenie się spostrzeżeniami i wiedzą w różnych segmentach biznesowych. Chociaż ważne jest, aby upewnić się, że dział IT ma do odegrania rolę w budowanej nowej organizacji opartej na danych, należy pamiętać, że zawsze istnieje równowaga, jeśli chodzi o zaangażowanie wewnętrznych funkcji IT. Różni się to oczywiście w zależności od firmy, ale z mojego osobistego doświadczenia wynika, że działa IT zazwyczaj chce zrobić więcej, niż jest w stanie, co często prowadzi do sytuacji, w których dział IT składa obietnice, których nigdy nie może spełnić – nie dlatego, że nie chcą, ale dlatego, że kierują się głównie kosztami i brakiem potrzebnego kontekstu biznesowego zrozumieć priorytety i podejmować niezbędne decyzje strategiczne.

W tej rzeczywistości należy wyraźnie oddzielić zadania, za które odpowiada dział IT w obszarze data science od tych, za które odpowiada zespół data science. Być może będziesz musiał skupić się na

infrastrukturze pamięci masowej lub niektórych aspektach zarządzania danymi. A może chcesz, aby skupili się na obsłudze modelu zarządzania danymi z perspektywy systemu. Jednak bez względu na to, jak podzielisz obowiązki, wyjaśnij działowi IT, że dbanie o ich obowiązki w kontekście nauki o danych to nie tylko drobny problem IT, ale raczej kluczowa część ogólnych operacji biznesowych w firmie. Działy IT zwykle lepiej radzą sobie z obsługą rozwiązań niż ich definiowaniem i rozwijaniem. Tak więc, bez względu na obszar, za który odpowiadasz dział IT, nie przypisuj im ogólnej odpowiedzialności za analizę danych. Po prostu brakuje im kompetencji biznesowych i data science, aby skutecznie nim zarządzać.

Zarządzanie oczekiwaniami

Innym ważnym aspektem do rozważenia jest jasne przedstawienie celu wspólnej funkcji nauki o danych. Każdy musi zrozumieć, że nie chodzi o to, aby cała praca związana z nauką o danych była wykonywana w ramach tej wspólnej funkcji. Daleko stąd. Aby firma mogła być oparta na danych i koncentrować się na nauce danych przez cały czas, ważne jest, aby była to działalność wszystkich w firmie. Jaka jest więc rola tej wspólnej funkcji nauki o danych? Przede wszystkim ma zapewnić współpracę w całej firmie, ułatwić udostępnianie danych i algorytmów, wspierać organizację wysoko wykwalifikowanymi analitykami danych oraz zapewniać inne odpowiednie kompetencje do realizacji celów firmy. Wszystko to musi się wydarzyć niezależnie od tego, czy działania są kierowane przez scentralizowaną funkcję, czy też kompetencje są wstrzykiwane do jednostek biznesowych, aby umożliwić i przyspieszyć pożądane rezultaty. Chcesz się upewnić, że organizacja nie postrzega tej nowej wspólnej funkcji nauki o danych jako „odpowiedzialnej" za tworzenie nauki o danych w firmie. Nie jest to pożądana sytuacja, ponieważ wprowadza pracowników w stan umysłu, w którym nie jest już ambicją firmy ani obowiązkiem wszystkich, aby urzeczywistnić naukę danych. Postrzega się wtedy, że rolą wspólnej funkcji nauki o danych jest to, aby tak się stało. Dlatego bardzo ważne jest jasne określenie roli wspólnej funkcji data science. Jest to organizacja wspierająca, mająca na celu umożliwienie firmie odniesienia sukcesu z inwestycją w naukę danych. Dla niektórych firm ta funkcja ma większe znaczenie przy starcie niż w dłuższej perspektywie, gdy data science jest bardziej znana i kompetencje są bardziej rozpowszechnione. Ale dla innych firm wspólna funkcja analizy danych staje się centrum podejścia opartego na danych, kluczowe dla przetrwania data science w firmie, nawet z długoterminowej perspektywy.

Wybór podejścia do realizacji

Po podjęciu wszystkich strategicznych decyzji dotyczących lokalizacji wspólnej funkcji analityki danych i uzgodnieniu, jak nowa funkcja analityki danych będzie działać w stosunku do reszty organizacji, nadszedł czas, aby zacząć. Ale gdzie Ty zaczynasz? Moim zdaniem masz tylko dwa logiczne sposoby, aby to zrobić: podejście Wielkiego Wybuchu lub podejście oparte na przypadkach użycia i skalowaniu. W następnych kilku sekcjach dokładnie opisano, z czym wiążą się te dwa podejścia i wymieniono ich zalety i wady.

Podejście typu „big bang"

Jednym ze sposobów ustanowienia nowej wspólnej funkcji analizy danych jest zastosowanie podejścia „big bang", w którym bezpośrednio definiujesz i wdrażasz długoterminową wizję i organizację docelową z pożądaną liczbą pracowników i pełną infrastrukturą analizy danych jednocześnie . Obejmuje to pełny zakres danych, architekturę danych, ramy zarządzania danymi, programy kompetencyjne i inne aspekty, wsparte jasnym planem wdrożenia zgodnym z priorytetami strategicznymi firmy. Zaletą takiego podejścia jest to, że komunikuje ono silne zaangażowanie firmy w naukę danych, zarówno wewnętrznie w stosunku do pracowników, jak i zewnętrznie w stosunku do rynku, klientów, dostawców, a nawet konkurentów. Takie podejście prawdopodobnie wygeneruje

motywację i zainteresowanie wśród pracowników, którzy czują się wzmocnieni i zainspirowani do poszukiwania nowych możliwości w tym obszarze. . Jest również prawdopodobne, że inwestycja faktycznie się wydarzy, ponieważ kierownictwo firmy zbytnio straciłoby wiarygodność, gdyby zdecydowało się nie dotrzymać jasno zakomunikowanego zobowiązania. Potencjalne wady podejścia „big bang" wiążą się z ideą, że struktura organizacyjna nauki o danych może być postrzegana jako narzucona organizacji odgórnie, bez wcześniejszego wypróbowania jej lub zakotwiczenia jej w organizacji. Jest to również ogromna inwestycja, którą należy nadrobić, zanim zostanie udowodniona jakakolwiek wartość dodana inwestycji w naukę danych. Ponadto podejście „big bang" utrudnia nowej wspólnej funkcji analityki danych znalezienie czasu na udowodnienie swojej wartości i uzyskanie wsparcia ze strony funkcji biznesowych, gdy utknie na zdefiniowaniu strategii badania danych, zatrudnianiu niezbędnych specjalistów ds. analizy danych, budowanie solidnej infrastruktury, zabezpieczanie danych oraz identyfikowanie i ustalanie priorytetów pierwszych interesujących przypadków.

Podejście do skalowania oparte na przypadkach użyciacase

Jeśli podejście typu „big bang" nie odpowiada potrzebom Twojej firmy, być może możesz wprowadzić nową wspólną funkcję analizy danych, korzystając z podejścia skalowania opartego na przypadkach użycia. To bardziej ostrożne podejście zaczyna się od szczegółowego zaprojektowania strategicznych kroków, które należy podjąć, w jakiej kolejności i skupienia się na rodzaju wartości. Takie podejście umożliwia strategiczny wybór przypadków, które potwierdzą wartość zarówno wewnętrznie (dla pracowników, zapewnienie zaangażowania i upodmiotowienia), jak i zewnętrznie (generowanie zaangażowania zewnętrznego poprzez udowodnienie, w jaki sposób można osiągnąć rzeczywistą wartość za pomocą nauki o danych w obszarach, które wcześniej okazały się być trudny). Korzyści z podejścia opartego na przypadkach użycia wynikają z faktu, że inwestycja z góry jest znacznie mniejsza, a także można udowodnić wartość użytkową funkcji analizy danych na podstawie każdego przypadku z osobna przed skalowaniem do następnego kroku. Możesz wybierać między korzyściami krótkoterminowymi a przypadkami użycia o wysokiej wartości. Na koniec zyskujesz przestrzeń do oddychania, ustalając wolniejsze tempo, co oznacza, że masz czas na zakotwiczenie konfiguracji i podejścia wśród pracowników, co oznacza, że mogą bardziej zaangażować się w definiowanie wymagań i priorytetów strategicznych. Możesz także zatrzymać się na mniej zaawansowanym poziomie konfiguracji bez marnowania czasu i pieniędzy, jeśli okaże się to lepsze dla firmy. W podejściu opartym na przypadkach użycia istnieje ryzyko, że nigdy nie udowodnisz wystarczająco wartości funkcji analizy danych dla interesariuszy, aby rzeczywiście uzyskać potrzebną wspólną funkcję analizy danych. Utkniesz w udowadnianiu każdego przypadku i nigdy nie uzyskasz ostatecznej akceptacji, co oznacza, że firma nigdy nie czerpie korzyści ze skali, w której jednostka może realizować wspólną strategię analizy danych, która promuje udostępnianie i ponowne wykorzystywanie danych, modeli i spostrzeżeń Firma.

Pozycjonowanie roli Chief Data Officera (CDO)

Zakładając, że istnieje strategiczne porozumienie w sprawie ustanowienia wspólnej funkcji zajmującej się badaniem danych w Twojej firmie, kto będzie strategicznym rzecznikiem wszystkich działań związanych z badaniem danych? Kto zapewni, że nauka o danych zostanie zrozumiana i uwzględniona w agendzie kierownictwa firmy? To coś więcej niż od czasu do czasu prezentacja w sali konferencyjnej, aby upewnić się, że najwyższa kadra ma pojęcie o tym, co się dzieje; chodzi o upewnienie się, że nauka o danych stanie się istotną częścią codziennego programu. Tutaj rola CDO staje się niezwykle ważna. Chief Data Officer (CDO) jest dyrektorem korporacyjnym odpowiedzialnym za nadzorowanie szeregu funkcji związanych z danymi, aby zapewnić Twojej organizacji maksymalne korzyści z tego, co może być jej najcenniejszym zasobem – jej danych. Zakres stanowiska obejmuje nadzór w całym przedsiębiorstwie oraz wykorzystanie danych i informacji jako zasobu, w tym wszystkie aspekty związane z architekturą danych, zarządzaniem danymi i zarządzaniem, wykorzystaniem danych i komercjalizacją danych realizowanych poprzez funkcję lub funkcje data science, niezależnie od tego, czy odbywa się to poprzez scentralizowaną, zdecentralizowaną lub hybrydową konfigurację. Jednak faktyczne umiejscowienie CDO w strukturze korporacyjnej nie jest podane i istnieje wiele przykładów. Chociaż znaczenie tej roli rośnie wraz ze wzrostem zrozumienia wartości danych, nadal rzadko zdarza się, aby CDO raportował bezpośrednio do dyrektora generalnego (CEO). Zazwyczaj CDO jest powiązany z funkcjami CIO (Chief Information Officer), CTO (Chief Technology Officer), a nawet CSO (Chief Strategy Officer) lub CMO (Chief Marketing Officer).

Zakres roli Chief Data Officer (CDO)

Aby opisać zakres CDO, musisz najpierw określić, w jaki sposób stanowisko ma się do stanowiska dyrektora ds. analityki (CAO). Chociaż CDO i CAO są dwiema odrębnymi rolami, te dwie pozycje są zwyczajowo pełnione przez tę samą osobę lub używana jest tylko jedna rola, rola CDO, ale gdy te role są połączone w jedną, czasami jest również określana jako rolę CDAO. Jednak w sytuacjach, w których te dwie role są oddzielone i pełnione przez dwie różne funkcje, główną różnicę można podsumować samym tytułem: dane kontra analityka. Główną różnicę w obszarze odpowiedzialności ujęto na rycinie

Jeśli CDO dotyczy włączania danych, rola CAO dotyczy sposobu uzyskiwania szczegółowych informacji na podstawie tych danych — innymi słowy, jak sprawić, by dane były przydatne. CAO jest znacznie bardziej prawdopodobne, że ma wykształcenie w zakresie analizy danych, a CDO – do inżynierii danych. Pozwolę sobie wyjaśnić, że zarówno stanowiska CDO, jak i CAO są zasadniczo wydzieleniami z tradycyjnej pracy CIO w dziedzinie IT. W przypadku roli CDO, CIO mógł z zadowoleniem przyjąć eliminację niektórych z tych obowiązków. Jednak jeśli chodzi o część roli CIO, która dotyczy kosztów IT dla nowych zasobów danych, CIO może być poważnie zakwestionowany przez nowe realia Big Data. Zarówno CDO, jak i CAO musiałyby argumentować za początkowym przechowywaniem ogromnych ilości danych, nawet jeśli ich wartość nie jest od razu widoczna. Aspekty te stanowią istotną, ale ważną zmianę w sposobie myślenia o roli CIO, która prawdopodobnie nie zostałaby rozpoznana w ten sam sposób bez wprowadzenia ról CDO i CAO. Jeśli chodzi o praktyczne wdrożenie tej roli, wszystko sprowadza się do zapewnienia wydajnej, kompleksowej konfiguracji i realizacji ogólnej strategii analizy danych w całej firmie. To, które rozwiązanie może działać jako najbardziej zoptymalizowana konfiguracja dla Twojej firmy, będzie zależeć od Twojej branży i sposobu organizacji. Pamiętaj tylko, aby te dwie role ściśle ze sobą współpracowały, w tym zespoły przypisane do ról. Separacja między zespołami inżynierii danych a zespołami analizy danych nie jest zalecana, zwłaszcza że istnieje potrzeba silnego wspólnego fundamentu opartego na tych dwóch częściach w nauce o danych. Zespoły mogą mieć inny cel, ale muszą ściśle ze sobą współpracować w ramach iteracji w sposób na osiągnięcie szybkości, elastyczności i wyników oczekiwanych przez interesariuszy biznesowych. W przypadkach, w których rola CDO jest jedyną rolą w firmie – gdzie obowiązki CDO i CAO zostały połączone, innymi słowy – mandat pełnienia funkcji CDO jest zwykle opisywany za pomocą rysunku 12-2. Obszar nawiązujący do prowadzonej działalności mandat odnosi się głównie do obszarów jazdy takich jak:

* Ustanowienie ogólnofirmowej strategii analizy danych

* Zapewnienie przyjęcia dominującej kultury danych w firmie

* Budowanie zaufania i legitymizacja wykorzystania danych.

* Wykorzystanie danych w celu uzyskania przewagi konkurencyjnej

* Udostępnianie możliwości biznesowych opartych na danych

* Zapewnienie przestrzegania zasad prawnych, bezpieczeństwa i etycznych

Jeśli chodzi o mandat technologiczny, zwykle uwzględnia się następujące aspekty:

* Ustanowienie architektury danych

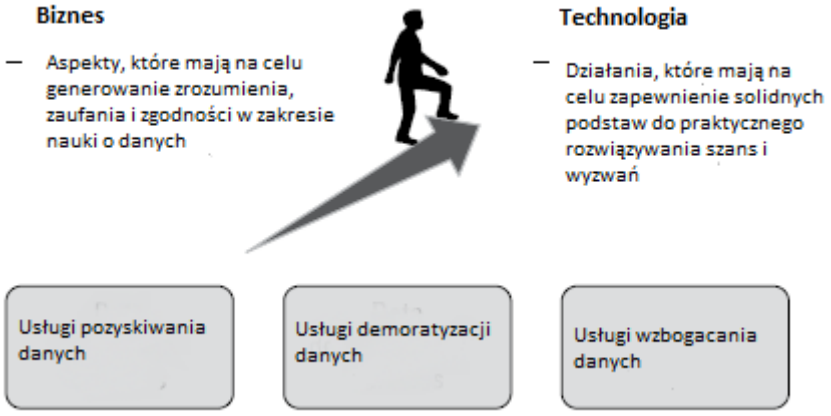
* Zapewnienie efektywnego zarządzania danymi

* Budowa infrastruktury umożliwiającej eksploracyjną i eksperymentalną analizę danych

* Promowanie ciągłej ewolucji metod i technik nauki o danych

* Projektowanie zasad pod kątem aspektów prawnych, bezpieczeństwa i etycznych

* Zapewnienie wydajnego zarządzania danymi i cyklem życia modelu



Zauważ, że na rysunku wyróżniam trzy różne usługi, którymi musi zarządzać połączony CDO/CAO. Ta lista daje wyobrażenie o tym, co pociąga za sobą każda usługa:

* Usługi fundacji danych obejmują takie obszary, jak zarządzanie pochodzeniem danych i zarządzanie danymi, definicja architektury danych, standardy danych i zarządzanie danymi, a także zarządzanie ryzykiem i różne rodzaje zgodności.

* Usługi demokratyzacji danych dotyczą takich obszarów, jak tworzenie kultury organizacyjnej opartej na danych poprzez walidację biznesową inicjatyw dotyczących danych, udostępnianie danych niewrażliwych wszystkim pracownikom (demokratyzacja danych) oraz właściwą ocenę dostępnych danych.

* Usługi wzbogacania danych obejmują takie obszary, jak pozyskiwanie i tworzenie wartości z danych poprzez stosowanie różnych metod i technik analitycznych i maszynowych, eksplorację i eksperymentowanie z danymi, a także zapewnienie inteligentnej i wydajnej konfiguracji jeziora danych / potoku danych wspierającego realizację wartości inwestycje w naukę danych.

Wyjaśnienie, dlaczego potrzebny jest dyrektor ds. danych

Oprócz badania możliwości uzyskania przychodów, opracowywania strategii przejęć i formułowania zasad dotyczących danych klientów, główny specjalista ds. danych jest odpowiedzialny za wyjaśnianie strategicznej wartości danych i ich ważnej roli jako aktywa biznesowego i siły napędowej przychodów kadrze kierowniczej, pracownikom i klientom. Dyrektorzy ds. danych odnoszą sukcesy, gdy ustanawiają autorytet, zabezpieczają budżet i zasoby oraz monetyzują zasoby informacyjne swojej organizacji. Rola CDO jest stosunkowo nowa i szybko ewoluuje, ale jednym z wygodnych sposobów patrzenia na tę rolę jest postrzeganie tej osoby jako głównego obrońcy i głównego stewarda zasobów danych organizacji. Organizacje mają coraz większy udział w agregowaniu danych i wykorzystywaniu ich do podejmowania lepszych decyzji. W związku z tym zadaniem CDO jest wykorzystywanie danych do automatyzacji procesów biznesowych, lepszego zrozumienia klientów, rozwijania lepszych relacji z partnerami, a docelowo szybszej sprzedaży większej liczby produktów i usług. Szereg ostatnich analiz trendów rynkowych wskazuje, że do 2020 r. 50 procent wiodących organizacji będzie miało CDO o podobnym poziomie wpływu na strategię i autorytecie, co dyrektorzy ds. informatyki. CDO mogą ustanowić rolę lidera, dopasowując swoje priorytety do priorytetów swoich organizacji. W dużej mierze rola dotyczy zarządzania zmianą. CDO muszą najpierw zdefiniować rolę i zarządzać oczekiwaniami, biorąc pod uwagę udostępnione im zasoby. Pomimo niedawnego szumu wokół koncepcji CDO, w praktyce okazało się, że trudno jest im zabezpieczyć cokolwiek innego niż umiarkowane budżety i ograniczone zasoby

podczas raportowania do istniejących jednostek biznesowych, takich jak IT. Co więcej, mając zwykle tylko garstkę personelu, grupa CDO musi działać wirtualnie, dołączając się do istniejących projektów i inicjatyw w całej organizacji i włączając się w nie. To oczywiście nie jest optymalna konfiguracja, jeśli chodzi o udowodnienie wartości funkcji CDO. Aby funkcja CDO naprawdę się opłaciła, musisz rozbić silosy i zoptymalizować struktury firmy wokół danych. Chodzi o rozdzielenie zakresu odpowiedzialności za dział IT, tak aby można było oddzielić zasoby danych od zasobów technologicznych i pozwolić CDO przejąć na własność część danych i informacji, a także pełny cykl nauki o danych, gdy jest nie wyznaczono żadnej roli CAO.

Ustanowienie roli CDO

Głównym zadaniem dyrektora ds. danych jest przekształcenie kultury firmy w taką, która obejmuje podejście oparte na wglądach i danych. Wartości tego nie należy lekceważyć. Ustanowienie mentalności skoncentrowania się na danych zmusza menedżerów na wszystkich poziomach do traktowania danych jako zasobu. Kiedy menedżerowie zaczną prosić o dane w nowy sposób i postrzegają kompetencje w zakresie analizy danych jako kluczowy zestaw umiejętności, nabiorą nowego znaczenia i priorytetów na wszystkich poziomach firmy. Zmiana kulturowego sposobu myślenia firmy nie jest łatwym zadaniem – potrzeba więcej niż tylko kilku warsztatów i serii poważnych dyrektyw z góry, aby wykonać zadanie. Pomysł, że należy traktować dane jako zasób, musi być mocno zakotwiczony w wyższej warstwie zarządzania – stąd tak ważna rola CDO. Niech ta lista wspólnych zleceń CDO w różnych branżach będzie inspiracją dla tego, co może zmieścić się w Twojej firmie. CDO może

* Stworzyć kulturę opartą na danych ze skutecznym zarządzaniem danymi. W ramach tego projektu ważne jest również zdobycie zaufania różnych jednostek biznesowych, aby można było stworzyć poczucie własności danych w całym przedsiębiorstwie. Chodzi o to, aby wspierać, a nie utrudniać efektywne wykorzystanie danych.

* Prowadzić zarządzanie danymi, wdrażając przydatne zasady i standardy zarządzania danymi zgodnie z ustaloną strategią dotyczącą danych. Ważne jest również uprzemysłowienie wydajnego zarządzania jakością danych, ponieważ zapewnienie jakości danych przez cały cykl życia danych wymaga znacznego wsparcia systemowego.

* Wpływać na podejmowanie decyzji w całej firmie, wspierany przez wysokiej jakości dane, które pozwalają na analitykę i spostrzeżenia, którym można zaufać.

* Wpływać na zwrot z inwestycji (ROI) poprzez wzbogacanie danych i lepsze zrozumienie potrzeb klientów. Pomysł polega na tym, aby pomóc firmie w dostarczaniu najwyższej jakości doświadczeń klientom poprzez wykorzystanie danych na wszystkie możliwe sposoby. .

* Zachęcać do ciągłych innowacji opartych na danych poprzez eksperymentowanie i eksplorację danych, w tym upewnianie się, że infrastruktura danych umożliwia to skutecznie i wydajnie.

Jak w większości ról, zawsze stajesz przed szeregiem wyzwań wpływających na poziom sukcesu, jaki można osiągnąć. Tylko będąc świadomym tych wyzwań, będziesz w stanie lepiej ich unikać lub przynajmniej mieć strategie radzenia sobie z nimi, jeśli lub kiedy się pojawią. Poniższa lista podsumowuje niektóre z najczęstszych wyzwań związanych z rolą CDO:

* Przypisywanie znaczenia biznesowego do danych: CDO musi upewnić się, że dane są traktowane priorytetowo, przetwarzane i analizowane we właściwym kontekście biznesowym, aby generować cenne i przydatne informacje. Jednym z aspektów tego może być czas generowania wglądu: jeśli czas

potrzebny na wygenerowanie wglądu jest zbyt wolny z perspektywy użyteczności biznesowej, wgląd, zamiast kierować biznesem, będzie jedynie potwierdzał to, co się właśnie wydarzyło.

* Ustanowienie i poprawa zarządzania danymi: obszar zarządzania danymi ma kluczowe znaczenie dla zachowania integralności danych w cyklu życia danych. Nie chodzi tylko o zarządzanie prawami dostępu do danych, ale w dużej mierze o zarządzanie jakością i wiarygodnością danych. Gdy tylko zadania wykonywane ręcznie staną się częścią czynności związanych z przetwarzaniem danych, istnieje ryzyko wprowadzenia błędów lub stronniczości do zbiorów danych, co sprawi, że analizy i spostrzeżenia wynikające z danych staną się mniej wiarygodne. Przetwarzanie danych oparte na automatyzacji jest zatem istotną częścią poprawy zarządzania danymi.

* Promowanie kultury udostępniania danych: W praktyce to wspólna funkcja nauki o danych będzie napędzać działania związane z nauką danych w całej firmie, ponieważ promuje codzienne udostępnianie danych. Jednak ważne jest również posiadanie silnego rzecznika w zarządzie, który wymusza zrozumienie i akceptację udostępniania danych. Głównym celem powinno być ustalenie, że nie będziesz w stanie czerpać wartości z danych, chyba że dane będą używane i udostępniane. Blokowanie danych poprzez ograniczanie dostępu i użytkowania to niewłaściwa droga – punktem wyjścia powinna być polityka otwartych danych w firmie. Mając to na uwadze, możesz ograniczyć dostęp do danych wrażliwych i nadal upewnić się, że takie ograniczenia są dobrze umotywowane i nie można ich rozwiązać za pomocą anonimizacji lub innych środków.

* Budowanie nowych strumieni przychodów, wzbogacanie i wykorzystywanie danych jako usługi: osoba pełniąca rolę CDO promuje również i wspiera innowacje związane z monetyzacją danych. To inspirujące zadanie, ale nie zawsze łatwe. Prowadzenie nowych rozwiązań biznesowych, które wymagają modeli biznesowych opartych na danych i potencjalnie zupełnie nowych modeli dostarczania, może wzbudzać strach i opór w firmie i ogólnie w zarządzie. Pamiętaj, że nowe pomysły monetyzacji danych mogą kwestionować istniejące modele biznesowe i być postrzegane jako zagrożenie, a nie jako nowy i obiecujący potencjał biznesowy. „Bądź uważny i poruszaj się powoli" to dobre podejście. Posługiwanie się przykładami z innych firm lub innych branż może również okazać się skuteczne w zdobyciu zaufania i wsparcia ze strony kierownictwa dla nowych pomysłów na monetyzację danych.

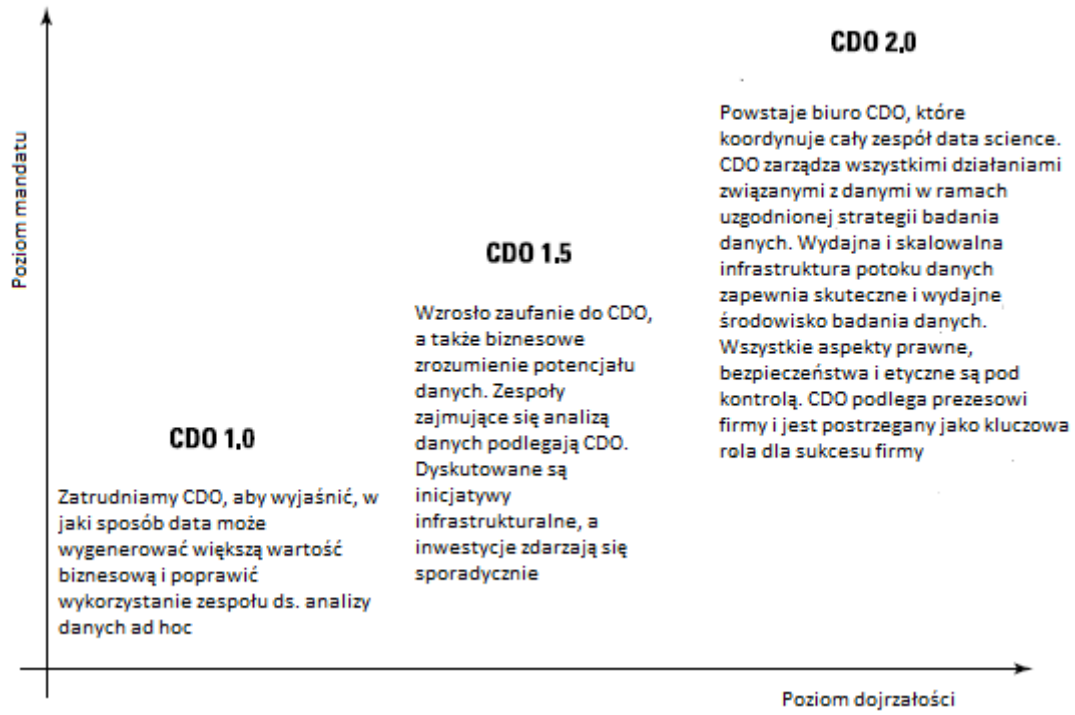
* Dostarczanie funkcji Poznaj swojego klienta (KYC) w sposób rzeczywisty i namacalny: wykorzystywanie danych w taki sposób, aby umożliwić sprzedaż opartą na danych, jest proaktywnym i skutecznym sposobem na wzmocnienie relacji z klientami i udowodnienie, jak kompetentna jest firma. Powinna jednak istnieć równowaga w sposobie pozyskiwania i wykorzystywania danych: ostatnią rzeczą, jakiej oczekujesz, jest to, aby Twoi klienci czuli się naruszeni. Chcesz, aby firma była postrzegana jako proaktywna i zorientowana na przyszłość z innowacyjnym dążeniem do dbania o swoich klientów, a nie jako firma, która wdziera się w sferę prywatną swoich klientów, wykorzystując ich dane do przekształcenia ich w przewagę w negocjacjach. CDO musi opanować tę równowagę i znaleźć sposób na strategiczne podporządkowanie się tej linii – linii, która może być zupełnie inna, w zależności od celów biznesowych i linii biznesowej.

* Naprawianie problemów ze starą infrastrukturą danych przy jednoczesnym inwestowaniu w przyszłość nauki o danych: To wyzwanie jest trudne do rozwiązania. Nie można po prostu przełączyć się ze starych, starszych infrastruktur (często skupiających się na transakcjach na danych i raportowaniu) na nową, często opartą na chmurze infrastrukturę, skoncentrowaną na obsłudze danych i zarabianiu na zupełnie nowe i różne sposoby. Musi być okres przejściowy, a w tym okresie musisz zajmować się utrzymaniem starszej infrastruktury, nawet jeśli jest to kosztowne i wydaje się niepotrzebnym obciążeniem. Jednocześnie kierownictwo oczekuje szybkich i wymiernych wyników,

opartych na potrzebnych głównych inwestycjach. Należy jednak pamiętać, że im dłużej dwie infrastruktury działają równolegle, tym trudniej jest naprawdę skłonić ludzi do zmiany sposobu myślenia i zachowania w kierunku nowego podejścia opartego na danych, realizowanego dzięki nowym inwestycjom w infrastrukturę. Nie zobaczysz też żadnych realnych oszczędności, ponieważ musisz zarządzać kosztami zarówno w starszych środowiskach, jak i w nowych. Nawet jeśli okaże się to trudne, spróbuj prowadzić tę zamianę z ambitnym harmonogramem, pamiętając, że nie ma odwrotu.

Przyszłość roli CDO

W ostatnich latach nasiliło się mianowanie dyrektorów ds. danych w dużych organizacjach, ponieważ firmy zdają sobie sprawę ze znaczenia danych jako podstawowego zasobu biznesowego. Oczekuje się, że dziewięć na dziesięć przedsiębiorstw ma pełnić tę rolę do 2020 r. Firmy coraz częściej umieszczają dyrektora ds. danych (CDO) w centrum ich działalności, z odpowiedzialnością powiązaną ze wszystkimi funkcjami, ponieważ wzrasta zależność od informacji i podejmowania decyzji w oparciu o dane. Jednocześnie rola CDO staje się szersza i mniej techniczna. W wielu firmach inteligentniejsze narzędzia analityczne eliminują niektóre złożone wcześniej wymagane analizy danych. Jeszcze kilka lat temu oczekiwano, że CDO ma wysoce techniczne zaplecze, ale teraz rola wyłania się z różnych części biznesowych CDO, teraz muszą wiedzieć znacznie więcej o kontekstach biznesowych, strategiach i ryzykach niż to, co było wymagane zaledwie dziesięć lat temu. Dane, na których dziś koncentrujemy się, to nie tylko „dane klientów" – to wszystko w firmie i dlatego ich rola się zmienia. Możliwość działania na danych w czasie rzeczywistym ma kluczowe znaczenie dla strategii biznesowych, dlatego podniesienie roli dyrektora ds. danych jako całościowej odpowiedzialności jest tak wielką sprawą. Na wyższym poziomie CDO musi być kimś, kto rozumie biznes, potrafi wyposażyć zespoły w odpowiednią infrastrukturę i zestaw narzędzi oraz wie, jak sprawić, by dostęp do danych był zarówno wydajny, jak i prosty. Tak więc rola CDO oczywiście ewoluuje, zarówno pod względem miejsca w hierarchii organizacyjnej, jak i rosnącego zakresu i mandatu. W miarę dojrzewania roli CDO wydaje się, że podąża pewną ścieżką, szczególnie w tych firmach, które ustanowiły tę rolę wcześnie. Możesz zobaczyć ogólny widok tych kroków na drodze do dojrzałości na rysunku



Kiedy mówisz o kontekście firmy CDO 1.0, prawdopodobnie zaczynasz od przetestowania kilku początkowych inicjatyw, co zwykle oznacza sporadyczne zatrudnianie analityków danych w całej firmie. Po ustaleniu, że fragmentaryczne podejście nie generuje dużej wartości, zwykle ustanawiasz kilka zespołów analityków danych jako wspierających role wokół naukowców zajmujących się danymi. Na koniec zatrudnia się dyrektora ds. danych, który próbuje uporządkować chaos, upewniając się, że zespoły analityków danych są efektywnie wykorzystywane. Osoba pełniąca rolę CDO zwykle podlega kierownictwu wyższego szczebla w zespole zarządzającym firmy lub o jeden szczebel zarządzania niżej. W kontekście CDO 1.5 zaczyna się pojawiać pewne uznanie i zaufanie wokół roli jako takiej, i dokonywane są wspólne inwestycje. Niemniej jednak nadal brakuje dostosowania strategicznego w skali całej firmy i nadal pojawia się wiele działań ad hoc. Jednak zazwyczaj trwają pewne uzgodnienia między różnymi zespołami zajmującymi się analizą danych, a kierownictwo zaczyna wyrażać swoje oczekiwania, jeśli chodzi o wyniki. Na tym poziomie dojrzałości CDO nie jest jeszcze częścią zespołu zarządzającego firmą. Wreszcie, na poziomie dojrzałości CDO 2.0, firma naprawdę dostrzega znaczenie roli CDO dla sukcesu firmy. Zazwyczaj tworzone jest biuro CDO, które podlega bezpośrednio dyrektorowi generalnemu, a wszystkie działania i inwestycje są prowadzone w oparciu o uzgodnioną i zatwierdzoną strategię analizy danych. Konfiguracja organizacji nauki o danych jest uzgadniana i koordynowana w całej firmie za pośrednictwem biura CDO, które zapewnia również dostosowanie standardów, zasad i infrastruktury. Gdzie więc CDO mogą zwrócić się o pomoc, aby zwiększyć rozpoznawalność i dojrzałość tej roli? Wiele problemów dla CDO wiąże się z udowodnieniem stabilności funkcji, gdy wszystko szybko się rozwija. Kiedy członkowie kierownictwa firmy mają problemy z nadążaniem za ciągłymi zmianami w obszarze data science, jest on postrzegany jako obszar, który należy traktować inaczej – bardziej jak start-up lub jednostka innowacyjna, a nie jak każda inna funkcja biznesowa. Ta marginalizacja, będąca poważnym problemem, utrudnia CDO stawanie się istotną i zintegrowaną funkcją w firmie. Na poziomie codziennym, prawdziwe problemy, przed którymi stoi CDO, wynikają z konieczności przebrnięcia przez ogromną liczbę usług oferowanych przez branżę danych: na przykład agencje transformacji, usługi w chmurze, narzędzia do czyszczenia danych i projektantów algorytmów. Jak CDO może znaleźć w tym wszystkim odpowiednie usługi? W tym przypadku trudno jest ocenić sukces, zwłaszcza w roli, która nie została jeszcze dobrze zdefiniowana przez przemysł, podczas gdy porażka może być dość oczywista. Jeśli Twoja firma jest na pierwszej stronie wiadomości z powodu poważnego naruszenia danych lub naruszenia prywatności, to zły dzień na CDO. Aby dowiedzieć się, co sprawia, że dzień jest dobry – cóż, wymaga to od większej liczby firm, które odważy się zaufać i zainwestować w CDO.

Pozyskiwanie zasobów i kompetencji

Prawie każda firma ma teraz możliwość zbierania danych, a ilość danych jest coraz większa. Doprowadziło to do większego zapotrzebowania na pracowników o specjalistycznych umiejętnościach, którzy potrafią skutecznie organizować i analizować te dane w celu uzyskania spostrzeżeń biznesowych. Niestety, nie tylko zapotrzebowanie na naukowców zajmujących się danymi przewyższa dostępną podaż, wielu początkujących naukowców zajmujących się danymi na rynku nie ma umiejętności ani doświadczenia potrzebnego do stanowiska. Wyspecjalizowany, złożony charakter pracy z data science stanowi istotny problem przy zatrudnianiu. W rzeczywistości na rynku pracy wciąż panuje zamieszanie co do tego, co właściwie oznacza termin data scientist. Często istnieją określone wymagania techniczne, których wymagają różne role w organizacji zajmującej się badaniem danych, ale musi istnieć wspólne zrozumienie tego, co jest wymagane, aby zespół ds. nauki danych odniósł sukces.

Identyfikacja ról w zespole Data Science

W ciągu ostatnich kilku lat rynek zdominowała lawina różnych ról związanych z analityką danych, a komuś, kto ma niewielkie doświadczenie w tej dziedzinie lub nie ma go wcale, trudno jest ogólnie zrozumieć, czym różnią się te role i jakie podstawowe umiejętności są w rzeczywistości wymagany. Faktem jest, że tym różnym rolom często przypisuje się różne tytuły, ale zwykle odnoszą się do tych samych lub podobnych zadań – co prawda czasami z nakładającymi się obowiązkami. Ta szalona seria tytułów zawodowych i obowiązków zawodowych to kolejny obszar nauki o danych, który wymaga większej standaryzacji. Zanim przystąpię do twardych i szybkich definicji ról, zacznę od naszkicowania różnych zestawów zadań, które zwykle można znaleźć w zespole analityków danych. Pomysł polega na tym, aby określić zakres obszarów kompetencji wysokiego poziomu, które muszą być uwzględnione w zespole analityków danych, niezależnie od tego, kto faktycznie wykonuje dane zadanie. Trzy główne obszary to matematyka/statystyka, informatyka i wiedza z dziedziny biznesu.

Naukowiec danych

Ogólnie rzecz biorąc, badacz danych tworzy modele matematyczne do celów predykcji. A ponieważ tworzenie i interpretacja modeli matematycznych wymaga głębokiej wiedzy technicznej, większość naukowców zajmujących się danymi ma wykształcenie na poziomie absolwenta w zakresie informatyki, matematyki lub statystyki. Naukowcy zajmujący się danymi potrzebują również silnych umiejętności programistycznych, aby skutecznie wykorzystać szereg dostępnych narzędzi programowych. Oprócz umiejętności technicznych, analitycy danych potrzebują umiejętności krytycznego myślenia, opartego na zdrowym rozsądku, a także na dogłębnym zrozumieniu celów biznesowych firmy, aby tworzyć modele wysokiej jakości. Czasami rola określana jako analityk danych jest oddzielona od roli naukowca danych. W takich przypadkach rola analityka danych jest podobna do Sherlocka Holmesa z zespołu data science, ponieważ skupia się na zbieraniu i interpretacji danych, a także analizowaniu wzorców i trendów w danych, z których wyciągają wnioski w kontekście biznesowym. Analityk danych musi opanować języki takie jak R, Python, SQL i C i, podobnie jak analityk danych, umiejętności i talenty potrzebne do tej roli są zróżnicowane i obejmują całe spektrum zadań w procesie data science. A do tego wszystkiego, analityk danych musi wykazać się zdrową postawą „ja-potrafię to wymyślić". Tak naprawdę to Ty decydujesz, czy chcesz, aby wszyscy analitycy danych w Twojej firmie zajęli się zadaniami związanymi z analitykiem danych, czy też chcesz skonfigurować analityka danych jako oddzielną rolę. W roli naukowca danych znajdziesz ukrytą inną, bardziej tradycyjną rolę: statystyka. W kategoriach historycznych statystyk był liderem, jeśli chodzi o dane i informacje, które mógł dostarczyć. Rola statystyka, chociaż często zapominana lub zastępowana przez bardziej wymyślnie brzmiące tytuły, reprezentuje to, co oznacza dziedzina nauki o danych: uzyskiwanie przydatnych informacji na

podstawie danych. Dzięki silnemu zapleczu w teoriach statystycznych i metodologiach oraz logicznemu nastawieniu, statystycy zbierają dane i przekształcają je w informacje i wiedzę. Mogą obsługiwać wszelkiego rodzaju dane. Co więcej, dzięki zapleczu ilościowemu, współcześni statystycy często są w stanie szybko opanować nowe technologie i wykorzystać je do zwiększenia swoich zdolności intelektualnych. Statystyk przedstawia magię matematyki dzięki spostrzeżeniom, które mogą radykalnie zmienić firmy.

Inżynier danych

Rola inżyniera danych ma fundamentalne znaczenie dla nauki o danych. Bez danych nie może istnieć nauka o danych, a praca naukowców zajmujących się danymi jest a) całkowicie niemożliwa, jeśli wymagane dane nie są dostępne i b) zdecydowanie zniechęcająca, jeśli dane są dostępne, ale tylko na niespójnych podstawach. Z problemem niespójności często borykają się naukowcy zajmujący się danymi, którzy często narzekają, że zbyt dużo czasu poświęcają na akwizycję i czyszczenie danych. W tym miejscu wkracza inżynier danych: rolą tej osoby jest tworzenie spójnych i łatwo dostępnych potoków danych do wykorzystania przez analityków danych. Innymi słowy, inżynierowie danych są odpowiedzialni za mechanikę pozyskiwania, przetwarzania i przechowywania danych, z których wszystkie powinny być niewidoczne dla analityków danych. Jeśli masz do czynienia z małymi zestawami danych, inżynieria danych zasadniczo polega na wprowadzaniu pewnych liczb do arkusza kalkulacyjnego. Kiedy działasz na bardziej imponującą skalę, inżynieria danych staje się samą w sobie wyrafinowaną dyscypliną. Ktoś z Twojego zespołu będzie musiał wziąć odpowiedzialność za radzenie sobie z trudnymi inżynieryjnymi aspektami dostarczania danych, z którymi może pracować reszta Twojego personelu. Inżynierowie danych nie muszą wiedzieć nic o uczeniu maszynowym (ML) ani statystykach, aby odnieść sukces. Nie muszą nawet należeć do głównego zespołu ds. analityki danych, ale mogą być częścią większego, oddzielnego zespołu ds. inżynierii danych, który dostarcza dane do wszystkich zespołów zajmujących się analizą danych. Z mojego doświadczenia wynika jednak, że nigdy nie należy umieszczać inżynierów danych i analityków danych zbyt daleko od siebie organizacyjnie. Jeśli te role są rozdzielone na różne organizacje, z potencjalnie różnymi priorytetami, może to mieć duży wpływ na produktywność zespołu ds. analizy danych. Metody data science mają dość eksperymentalny i iteracyjny charakter, co oznacza, że musi istnieć możliwość ciągłego modyfikowania zbiorów danych w miarę postępu analizy i rozwoju algorytmów. Aby tak się stało, analitycy danych muszą być w stanie polegać na szybkiej odpowiedzi inżynierów danych w przypadku wystąpienia problemów. Bez tej szybkiej reakcji istnieje ryzyko spowolnienia wydajności zespołu zajmującego się analizą danych.

Inżynier uczenia maszynowego

Analitycy danych budują modele matematyczne, a inżynierowie danych udostępniają dane analitykom danych jako „surowiec”, z którego powstają modele matematyczne. Aby uzupełnić obraz, modele te muszą najpierw zostać wdrożone (innymi słowy wprowadzone w życie), a po drugie muszą być w stanie działać na podstawie spostrzeżeń uzyskanych z analizy danych w celu wytworzenia wartości biznesowej. To zadanie należy do inżyniera uczenia maszynowego. Rola inżyniera uczenia maszynowego to rola inżyniera oprogramowania, z tą różnicą, że inżynier ML ma duże doświadczenie w nauce danych. Ta wiedza jest wymagana, ponieważ inżynierowie ML wypełniają lukę między naukowcami zajmującymi się danymi a szerszą organizacją inżynierii oprogramowania. Dzięki inżynierom ML zajmującym się wdrażaniem modeli naukowcy zajmujący się danymi mogą stale rozwijać i udoskonalać swoje modele. Warianty są zawsze możliwe podczas tworzenia zespołu ds. analizy danych. Na przykład obowiązki związane z wdrożeniem inżyniera ML są często obsługiwane przez rolę naukowca danych. W zależności od tego, jak ważne jest środowisko operacyjne dla konkretnej firmy, mniej lub bardziej sensowne może być oddzielenie tej roli od obowiązków analityka danych. Ponownie do Ciebie należy wdrożenie tej odpowiedzialności w zespole.

Architekt danych

Architektura danych to zestaw reguł, zasad, standardów i modeli, które regulują i definiują typ gromadzonych danych oraz sposób ich wykorzystania, przechowywania, zarządzania i integracji w ramach organizacji i jej systemów danych. Osoba odpowiedzialna za projektowanie, tworzenie, wdrażanie i zarządzanie architekturą danych organizacji nazywana jest architektem danych i zdecydowanie musi być rozliczana na dany zespół naukowy. Architekci danych określają, w jaki sposób dane będą przechowywane, wykorzystywane, chronione, integrowane i zarządzane przez różne jednostki danych i systemy informatyczne, a także przez wszelkie aplikacje wykorzystujące lub przetwarzające te dane w jakiś sposób. Architekt danych zwykle nie jest stałym członkiem jednego zespołu zajmującego się analizą danych, ale raczej obsługuje kilka zespołów zajmujących się analizą danych, ściśle współpracując z każdym zespołem w celu zapewnienia wydajności i wysokiej produktywności.

Analityk Biznesowy

Analityk biznesowy często pochodzi z innego środowiska niż reszta zespołu. Choć często mniej zorientowani technicznie, analitycy biznesowi nadrabiają to głęboką znajomością różnych procesów biznesowych zachodzących w firmie - procesów operacyjnych (proces sprzedaży), procesów zarządzania (proces budżetowania) i procesów wspierających (proces zatrudniania). Analityk biznesowy opanowuje umiejętność łączenia spostrzeżeń dotyczących danych z praktycznymi spostrzeżeniami biznesowymi i może wykorzystywać techniki opowiadania historii do rozpowszechniania wiadomości w całej organizacji. Ta osoba często działa jako pośrednik między „facetami z biznesu” i „technikami”.

Inżynier oprogramowania

Główną rolą inżyniera oprogramowania w zespole zajmującym się badaniem danych jest zapewnienie większej struktury w pracy z nauką danych, aby stała się ona bardziej stosowana i mniej eksperymentalna. Inżynier oprogramowania odgrywa ważną rolę we współpracy z analitykami danych, architektami danych i analitykami biznesowymi w celu zapewnienia zgodności między celami biznesowymi a rzeczywistym rozwiązaniem. Można powiedzieć, że inżynier oprogramowania jest odpowiedzialny za wprowadzenie kultury inżynierii oprogramowania do procesu nauki o danych. To ogromne przedsięwzięcie, które obejmuje takie zadania, jak automatyzacja infrastruktury zespołu zajmującego się badaniem danych, zapewnienie ciągłej integracji i kontroli wersji, automatyzacja testowania i opracowywanie interfejsów API, które pomagają integrować produkty danych z różnymi aplikacjami.

Ekspert domeny

Nauka danych wymaga wielu rozmów. Analitycy danych nie mogą tego zrobić samodzielnie. Sukces w nauce o danych wymaga wieloumiejętnego zespołu projektowego, w którym ściśle współpracują naukowcy zajmujący się danymi i eksperci dziedzinowi. Ekspert domeny wnosi techniczne zrozumienie swojego obszaru specjalizacji, czasami w połączeniu z dogłębnym zrozumieniem biznesowym również tego obszaru. Zwykle obejmuje znajomość podstaw analizy danych, co oznacza, że eksperci dziedzinowi mogą wspierać wiele ról w zespole data science. Jednak ekspert domeny zwykle nie jest stałym członkiem zespołu zajmującego się badaniem danych; częściej niż nie, ta osoba jest angażowana do określonych zadań, takich jak walidacja danych lub dostarczanie analizy lub wglądu z perspektywy eksperta. Czasami ekspert dziedzinowy jest przydzielany na dłuższe okresy do określonego zespołu, w zależności od zadania i skupienia. Czasami jeden lub kilku ekspertów domenowych jest przypisywanych do obsługi wielu zespołów jednocześnie.

Zobacz, co czyni wielkim naukowca danych

Z rolą naukowca danych wiąże się wiele obietnic. Problem polega nie tylko na tym, że nie istnieje doskonały naukowiec zajmujący się danymi, ale także na tym, że tych kilku naprawdę wykwalifikowanych specjalistów jest zbyt niewielu i zbyt trudno jest zdobyć je na obecnym rynku. Co więc powinieneś zrobić zamiast szukać idealnego naukowca danych? Należy skoncentrować się na znalezieniu kogoś, kto potrafi rozwiązać konkretne problemy, na których koncentruje się Twoja firma - lub, mówiąc jeszcze bardziej szczegółowo, na czym skupia się Twój własny zespół ds. analizy danych. Nie chodzi o zatrudnienie doskonałego analityka danych i nadzieję, że zrobi on wszystko, co musisz zrobić teraz i w przyszłości. Zamiast tego lepiej zatrudnić kogoś z konkretnymi umiejętnościami potrzebnymi do spełnienia jasno określonych celów organizacyjnych, które znasz dzisiaj. Na przykład zastanów się, czy Twoja potrzeba jest bardziej związana z analizą danych ad hoc lub rozwojem produktu. Firmy, które mają większe zapotrzebowanie na wgląd w dane ad hoc, powinny szukać analityków danych z elastycznym i eksperymentalnym podejściem oraz umiejętnością dobrej komunikacji ze stroną biznesową organizacji. Z drugiej strony, jeśli rozwój produktu jest ważniejszy w stosunku do problemów, które próbujesz rozwiązać, powinieneś poszukać silnych umiejętności w zakresie inżynierii oprogramowania, z solidną podstawą w procesie inżynieryjnym w połączeniu z ich umiejętnościami analitycznymi. Jeśli masz nadzieję znaleźć przydatną listę kontrolną wszystkich kluczowych umiejętności, które powinieneś szukać, zatrudniając analityka danych, będziesz bardzo rozczarowany. Faktem jest, że w branży nie uzgodniono nawet podstawowego opisu ważnych cech, jakie powinna posiadać rola. Jest wiele opinii i pomysłów na ten temat, ale znowu brak standaryzacji jest kłopotliwy. Co zatem sprawia, że wymyślenie prostej listy kontrolnej potrzebnych zestawów narzędzi, wymaganych kompetencji i umiejętności technicznych jest tak trudne? Po pierwsze, obszar ten wciąż szybko się rozwija, a narzędzia i techniki, które były ważne do opanowania w zeszłym roku, mogą być mniej ważne w tym roku. Dlatego pozostawanie w zgodzie z ewolucją pola i nieustanne poznawanie nowych metod, narzędzi i technik jest kluczem w tej przestrzeni. Innym powodem, dla którego trudno jest określić konkretną listę kontrolną umiejętności, jest to, że potrzebne zestawy umiejętności krytycznych znajdują się w rzeczywistości poza obszarem nauki o danych – kwalifikują się one bardziej jako umiejętności miękkie, takie jak komunikacja interpersonalna i projektowanie właściwej postawy. Wystarczy spojrzeć na diagram Venna przedstawiający dane naukowca przedstawiający potrzebne umiejętności, cechy i postawę. Różnorodność zestawów umiejętności i cech mentalności, które musi opanować doskonały analityk danych, jest niemal absurdalna.

Tak więc, mając na uwadze, że określenie kompetencji potrzebnych naukowcowi danych jest bardziej kwestią nastawienia i sposobu myślenia w połączeniu z pewnym zestawem umiejętności, wciąż sporządziłem tę listę:

* Zrozumienie biznesu: Umiejętność przełożenia problemu z języka biznesowego na hipotezę jest ważna i odnosi się do tego, w jaki sposób analityk danych powinien być w stanie zrozumieć, co opisuje osoba biznesowa, a następnie być w stanie przetłumaczyć to na terminy techniczne i przedstawić potencjalne rozwiązanie w tym kontekście.

* Imponujący kontra interesujący: naukowcy zajmujący się danymi muszą być w stanie oprzeć się pokusie, aby zawsze nadawać priorytet interesującym problemom, gdy mogą istnieć problemy, które są ważniejsze do rozwiązania ze względu na duży wpływ na biznes, jaki miałyby takie rozwiązania.

* Ciekawość: Posiadanie ciekawości intelektualnej i umiejętność wyszczególnienia problemu w jasny zestaw hipotez, które można przetestować, jest dużym plusem.

* Dbałość o szczegóły: Jako analityk danych zwracaj uwagę na szczegóły z technicznego punktu widzenia. Model nie może być prawie poprawny. Zbudowanie zaawansowanego algorytmu technicznego wymaga czasu i poświęcenia na szczegóły.

* Łatwy uczeń: Naukowiec zajmujący się danymi musi mieć zdolność do szybkiego uczenia się, ponieważ szybko zmieniający się charakter przestrzeni nauki o danych obejmuje technologie i metodologie, ale także nowe narzędzia i modele open-source, które są udostępniane i gotowe do budowania.

* Zwinny sposób myślenia: Bądź elastyczny i sprawny w zakresie tego, co jest możliwe, w jaki sposób podchodzi się do problemów, jak badane są rozwiązania i jak problemy są rozwiązywane.

* Nastawienie eksperymentalne: badacz danych nie może obawiać się porażki ani próbować założeń, które mogą być błędne, aby znaleźć najbardziej udaną drogę naprzód.

* Komunikacja: analityk danych musi być w stanie opowiedzieć historię i opisać problem w centrum uwagi lub okazję, do której dąży, a także opisać, jak wspaniałe są modele po ich ukończeniu i co faktycznie umożliwiają.

Oczywiście istnieją dodatkowe interesujące umiejętności, takie jak statystyka, uczenie maszynowe i programowanie, ale pamiętaj, że nie potrzebujesz tutaj jednej osoby, aby dopasować wszystkie kategorie. Przede wszystkim powinieneś szukać analityków danych, którzy posiadają najważniejsze umiejętności odpowiadające Twoim potrzebom. Jednak szukając tego najlepszego naukowca danych, pamiętaj, że powyższa lista może być również wykorzystana do zatrudnienia uzupełniającego zespołu analityków danych, którzy razem posiadają potrzebne umiejętności i sposób myślenia. Po utworzeniu zespołu analityków danych zachęcaj ich do rozwoju zawodowego i uczenia się przez całe życie. Wielu naukowców zajmujących się danymi ma akademicki sposób myślenia i chęć eksperymentowania, ale w dążeniu do idealnego rozwiązania czasami gubią się wśród wszystkich danych i problemów, które próbują rozwiązać. Dlatego ważne jest, aby pozostawali w kontakcie z zespołem, chociaż należy zapewnić im wystarczającą niezależność, aby mogli nadal publikować białe księgi, przyczyniać się do open source lub prowadzić inne znaczące działania w swojej dziedzinie.

Tworzenie zespołu Data Science

Tworząc zespół ds. analityki danych z odpowiednim typem zestawów umiejętności, chodzi o znalezienie zoptymalizowanego zestawu członków zespołu. Jakie są kluczowe czynniki napędzające różne rodzaje ról i jak należy je połączyć w jeden zespół? Bądźmy szczerzy: nie ma formuły, którą możesz zastosować, która rozwiąże to równanie za Ciebie. To trochę bardziej skomplikowane. To, jak struktura zespołu powinna wyglądać i być zrównoważona, jest w dużym stopniu związane z celami, wyglądem procesów, zdefiniowaniem zamierzonego środowiska docelowego i tak dalej. Jednak zawsze możesz zacząć od prostej standardowej konfiguracji opartej na opisach ról w poprzedniej sekcji i dalej pracować. Podczas równoważenia potrzebnych zasobów na rolę w zespole ds. analizy danych należy wziąć pod uwagę aspekty, takie jak te opisane na tej liście:

* Zakres i złożoność danych: Jaki jest zakres Twojego wyzwania w zakresie nauki o danych? Czy dążysz do wzrostu wydajności wewnętrznej firmy, czy zamierzasz prowadzić działalność w zakresie danych komercyjnych? Zakres potrzebnych danych i złożoność akwizycji danych wpłyną na przykład na potrzebę wsparcia architektury danych i doświadczonych inżynierów danych.

* Typ produktu lub usługi związanej z danymi: jeśli wybierasz komercyjny produkt związany z danymi, czy chcesz sprzedawać coś „z półki”, czy też zamierzasz zbudować model operacyjny z perspektywy usługi danych? Jak zamierzasz dostarczyć swój produkt lub usługę związaną z danymi? W zależności od

typu oferty i modelu dostarczania, z pewnością wpłynie to na liczbę inżynierów oprogramowania potrzebnych do opracowania kompleksowego rozwiązania i stworzenia platformy dostarczania.

* Poziom zastosowanych technik uczenia maszynowego/sztucznej inteligencji: jak złożone są Twoje przypadki użycia i docelowe rozwiązanie? Czy istnieje zapotrzebowanie na wysoce techniczne rozwiązanie z dużą ilością samouczących się algorytmów, czy wystarczy prostszy model? Wymagany poziom złożoności będzie napędzał zapotrzebowanie na różne zestawy umiejętności naukowców zajmujących się danymi, od analityki po zaawansowaną analitykę i od uczenia maszynowego po kompetencje w zakresie sztucznej inteligencji.

* Konfiguracja środowiska nauki o danych (opracowywanie i produkcja): jak będzie wyglądać infrastruktura nauki o danych? Czy działa w chmurze czy lokalnie? Czy jest dystrybuowany globalnie, czy istnieje pojedyncza instancja lokalna? A może masz globalną konfigurację ze scentralizowaną instancją i instancjami brzegowymi w chmurze w różnych krajach, a nawet instancjami brzegowymi na poziomie urządzenia? Konfiguracja infrastruktury może się znacznie różnić między firmami, w zależności od wielkości firmy, czy jest ona lokalna czy globalna, branży, na której się koncentruje, czy dane są własnością, czy potrzebujesz praw do ich używania i tak dalej. Jednak z punktu widzenia równoważenia zasobów wkrótce zdasz sobie sprawę, że konfiguracja środowiska nauki o danych wpływa na liczbę zespołów zajmujących się nauką danych, których będziesz potrzebować, a także na to, których kompetencji potrzebujesz więcej - a także na to, gdzie na świecie zespoły muszą być umieszczony.

* Model organizacyjny nauki o danych: odnosi się do konfiguracji organizacyjnej, na którą zdecydowałeś się w zakresie posiadania scentralizowanego zespołu, zdecentralizowanego zespołu lub hybrydy. (Więcej informacji na temat tych modeli można znaleźć w rozdziale 11.) Równoważenie zasobów zależy od roli, jaką scentralizowana funkcja będzie odgrywać w firmie. Na przykład, jeśli masz wspólną scentralizowaną funkcję analizy danych, czy oznacza to, że wszyscy analitycy danych powinni tam pracować, obsługując całą firmę? A może oznacza to, że scentralizowana funkcja działa tylko na wspólnych częściach, co jest istotne w różnych jednostkach biznesowych, co oznacza, że reszta organizacji może nabywać i budować własne kompetencje w zakresie analizy danych? Są to ważne pytania, które należy wyjaśnić, aby lepiej zrozumieć parametry równoważenia zasobów analizy danych.

Zatrudnianie i ocena potrzebnych talentów w zakresie analityki danych

Podejmując decyzje o zatrudnieniu w data science, Twoim celem jest posiadanie dobrze działającego zespołu, a nie tylko zestawu wykwalifikowanych osób. Równie ważna jest potrzeba stworzenia zróżnicowanego zespołu, w którym osoby z różnych środowisk i różnych doświadczeń życiowych mogą wygodnie ze sobą współpracować. Sztuczka polega na tym, aby rozpocząć wyszukiwanie od osób reprezentujących różne dyscypliny - na przykład naukowców zajmujących się danymi, inżynierów danych i inżynierów oprogramowania - ale potem zawsze podejmuj ostateczną decyzję w oparciu o zdolność kandydata do dobrego funkcjonowania w kontekście nauki o danych. Jak więc stwierdzić, czy ktoś będzie dobrze funkcjonował w takim kontekście? Zawsze przyglądam się trzem głównym obszarom, oceniając, czy kandydat ma odpowiedni zestaw umiejętności i cechy osobowości, aby odnieść sukces (nawiasem mówiąc, te kryteria zawsze stosuję do wszystkich osób w zespole, a nie tylko do wybranych):

* Dopasowanie kulturowe

* Umiejętności inżynierskie

* Kompetencje w zakresie nauki danych

W obszarze umiejętności inżynierskich poszukaj kompetencji w zakresie projektowania systemów, kodowania formalnego i opracowywania algorytmów. Jeśli chodzi o umiejętności z zakresu analizy danych, powinieneś nalegać na kompetencje w zakresie modelowania instancji i algorytmów, frameworku i narzędzi ML (takich jak TensorFlow) oraz przetwarzania danych. Jeśli chodzi o dopasowanie kulturowe, zacznij od przyjrzenia się osobom, które wyraźnie podzielają wartości firmy, a także pozostałym członkom zespołu. Następnie przejdź do oceny, jak kandydat pracuje w zespole, a następnie oceń jego osobistą motywację i motywację. Pamiętaj jednak o zasadzie niezatrudniania osób reprezentujących to samo pochodzenie, płeć, wykształcenie i wiek. Zróżnicowany zespół zapewnia różnorodne wyniki i aktywnie działa przeciwko stronniczości danych i spostrzeżeń. Aby ocenić kandydatów pod kątem wszystkich tych aspektów, ważne jest, aby mieć jasną i wspólnie uzgodnioną percepcję tego, jak wygląda dobro. Powinno to mieć formę ogólnych kryteriów oceny, co do których wszyscy uczestniczący w procesie rekrutacji są zgodni – i zdecydowanie powinieneś mieć je na piśmie. Wydaje się to łatwe, ale często jest to trudne w praktyce, ponieważ to, jak wygląda „dobre", jest w dużej mierze kwestią osobistej opinii. Ale choć jest to trudne, określenie kryteriów oceny jest naprawdę ważne. Wszyscy kandydaci muszą być oceniani w jak najbardziej bezstronny sposób według tych samych kryteriów. Aby uzyskać dane o kandydatach, których potrzebujesz do oceny różnych obszarów, powinieneś użyć kombinacji zadań i testów z rozmowami kwalifikacyjnymi. Wyniki testów, wyniki i wszelkie spostrzeżenia uzyskane z wywiadów osobistych należy następnie przyporządkować do tego, jak wygląda dobro. Na przykład możesz przeprowadzić rozmowę kwalifikacyjną, rozmowę o inżynierii i projektowaniu systemu, zadanie z zakresu nauki o danych, a następnie rozmowę o podstawach nauki o danych. Jeśli kandydat przejdzie przez wszystkie te etapy, lider data science musi ocenić dopasowanie kulturowe, aby upewnić się, że zostanie zatrudniona właściwa osoba (a nie tylko odpowiedni zestaw umiejętności).

Być może zastanawiasz się, dlaczego powinieneś oceniać inżyniera danych, patrząc na te same obszary i stosując te same kryteria, których używałbyś dla naukowca danych. Twierdzę, że opłaca się być w stanie ocenić zakres wiedzy kandydata, zwłaszcza w sąsiednich dziedzinach wiedzy. Niezależnie od wyników, warto wiedzieć, czy kandydat jest wykwalifikowanym naukowcem zajmującym się danymi, ale raczej złym inżynierem, nawet jeśli ta osoba może nigdy nie zostać poproszona o wykonanie jakiegokolwiek kodu. A kiedy przejdziesz do tego, naprawdę istnieją ważne umiejętności międzyfunkcyjne, które są niezbędne dla zespołu zajmującego się analizą danych. Na przykład najlepiej byłoby mieć inżyniera danych, który zna się na kodowaniu i wie o projektowaniu systemu i DevOps, ale wie wystarczająco dużo o podstawach nauki o danych, aby wiedzieć, w jakim stopniu i w jaki sposób można wykorzystać ich talenty w kontekście nauki o danych. Nie, ta osoba nie musi mieć takiego samego poziomu wiedzy o danych, jak inżynier uczenia maszynowego z wykształceniem, jak matematyk, ale musi być na takim poziomie, na którym wie, jak może przyczynić się. Aby zaspokoić te potrzebne wielofunkcyjne zestawy umiejętności, należy zmapować, które umiejętności i obowiązki są mniej ważne lub ważniejsze dla różnych ról w zespole. Tak, naukowiec zajmujący się danymi powinien przystąpić do testu projektu systemu, ale jeśli uzyska słaby wynik, nie powinien to być natychmiastowy czarny znak przeciwko niemu. Ale wtedy powinieneś wymagać od tego kandydata wysokiego wyniku w zakresie zrozumienia tematów, takich jak naiwne metody Bayesa i techniki regresji logistycznej jako kontrapunkt. Po zmapowaniu względnego znaczenia różnych zestawów umiejętności, upewnij się, że wyznaczyłeś obszary lub zestawy umiejętności, których oczekujesz od rozmów kwalifikacyjnych, zadań i testów, aby upewnić się, że obejmujesz wszystkie niezbędne punkty danych do prawidłowej oceny kandydata . Pamiętaj, że naprawdę trudno jest zbudować udane zespoły, a także kosztuje to sporo. Potraktuj tę pracę poważnie i przyjdź dobrze przygotowany do rozmów kwalifikacyjnych. Jest to ważne również dlatego, że dostęp do doświadczonych kompetencji w data science jest znikomy, a nie tylko Ty oceniasz kandydata - kandydat ocenia również Ciebie i Twój poziom kompetencji w danym obszarze,

a także dojrzałość firmy . Dobrym punktem wyjścia jest dobre przygotowanie i przemyślana struktura wywiadu.

Utrzymanie kompetencji w zakresie nauki o danych

Co firmy mogą zrobić, aby jak najlepiej wykorzystać swoje zespoły zajmujące się analizą danych i zmotywować je do większego wkładu w biznes? Jedną z ważnych części jest zapewnienie analitykom danych czasu potrzebnego na wymyślenie. Pamiętaj, że masz do czynienia z ludźmi, którzy z jednej strony chcą przekraczać granice, a z drugiej łatwo się nudzą, gdy są proszeni o robienie tego samego w kółko. Ci naukowcy to rzadkie talenty, które chcą pracować nad najważniejszymi funkcjami firmy. Jeśli prosi się ich o spędzenie czasu na wykonywaniu powtarzalnych zadań, takich jak pozyskiwanie danych, zarządzanie danymi i szeroko zakrojone masowanie prognozowania wyników, często czują się niewykorzystani. Zlecanie analitykom danych przyszłościowych projektów daje im możliwość wymyślenia sposobu, w jaki firma może czerpać korzyści z dużych zbiorów danych. Upewnij się również, że kierownictwo firmy jest zaangażowane na odpowiednim etapie projektów danych. Bez dostępu do wyższego kierownictwa zespoły zajmujące się analizą danych mogą skoncentrować się na niewłaściwych problemach. Powinno to być najlepiej zarządzane w ramach roli CDO, jeśli taka rola istnieje w firmie. Ogólnie rzecz biorąc, dla zespołów zajmujących się analizą danych kluczowe jest zaangażowanie kierownictwa wyższego szczebla w każdym projekcie na trzech etapach: na wczesnym etapie, aby pomóc w zdefiniowaniu problemu, który firma chce rozwiązać; po pierwszych wynikach zaczną się toczyć; i kiedy nadejdzie czas, aby wynikające z tego spostrzeżenia zostały wdrożone lub podjęte w oparciu o nie. Prawidłowo obsłużeni specjaliści ds. danych mogą zyskać ogromną reputację w zakresie wiedzy w organizacji. Zapewnienie wczesnego dialogu między zespołami analityków danych a kierownictwem wyższego szczebla, co często zwiększa prawdopodobieństwo, że sugestie analityków danych zostaną faktycznie wdrożone. Ponownie, jest to kluczowa rola do odegrania dla CDO. Jeśli chodzi o motywowanie talentów zajmujących się strategicznymi analizami danych do pozostania w firmie, inną strategią jest wypuszczenie analityków danych z pudełka danych. Analitycy danych to naturalni uczniowie, którzy są w stanie postrzegać wszystkie aspekty biznesu jako oparte na danych, a nie przez tradycyjne tworzenie oprogramowania lub przez pryzmat marketingu. Ze względu na tę perspektywę mogą nawiązywać połączenia, których nie mogą inni, z szerszymi rozmowami i innowacyjnymi pomysłami poprzez obserwację całego biznesu. Polecam również rozważenie przeszkolenia analityków danych z Twojej firmy. To, czy naukowcy zajmujący się danymi mają najseksowniejszą pracę w XXI wieku, jak deklarował Harvard Business Review, jest dyskusyjne, ale nie podlega dyskusji, że trudno ich zidentyfikować, trudno zrekrutować i brakuje ich. Przeszkolenie naukowców zajmujących się danymi oznacza przeniesienie ludzi z organizacji zajmującej się badaniem danych do zarządzania operacjami, marketingu cyfrowego lub zarządzania relacjami z klientami, które są dyscyplinami ugruntowanymi analitycznie i mogą otworzyć nowe możliwości rozwoju osobistego i zawodowego, nie tylko dla naukowców zajmujących się danymi, ale także dla firmy jako całości, gdy ich kompetencje i wiedza są rozłożone w bardziej praktyczny sposób, napędzając myślenie oparte na danych i stosując modele statystyczne w praktyce poza tradycyjnymi dziedzinami. Takie szkolenie krzyżowe może również działać jako motywator dla większej liczby osób, aby chciały dowiedzieć się więcej o nauce danych i kontynuować karierę w tym obszarze poprzez szkolenia formalne i w miejscu pracy. Kiedy liderzy biznesowi mylą raportowanie danych z analizą, firma może mieć problemy z efektywnym rozwiązywaniem problemów. Z tego samego powodu naukowcy zajmujący się danymi muszą nauczyć się, jak zwracać się do wyższej kadry zarządzającej na warunkach wyższego kierownictwa. Analitycy danych zwykle chcą wyjaśnić wszystko, co zrobili, opisać, jak ciężko pracowali i podkreślić, jakie to było osiągnięcie. Z drugiej strony, kierownictwo wyższego szczebla ma trzy zasady: Bądź jasny, bądź szybki i odejdź. Rozwijanie świadomości biznesowej analityków danych pomaga im w bardziej holistyczny sposób wnosić wkład do rozmów w firmie, umożliwiając im inicjowanie analiz i

eksperymentów, a nie tylko reagowanie na prośby. Jest to długoterminowa korzyść, której wdrożenie kosztuje firmy niewiele, ale w dłuższej perspektywie jest to kluczowa kompetencja.

Zrozumienie, co powoduje odejście analityka danych

Niestety, bez względu na to, jakie masz ambicje w zakresie nowych analityków danych, których wprowadzasz do firmy, wielu analityków danych ma tendencję do kontynuowania pracy, często w ciągu pierwszego roku. Dlaczego? Aby uzupełnić wcześniejszą sekcję o tym, co należy zrobić, aby zachować cenne zasoby do nauki danych, ta sekcja ma na celu wskazanie czterech głównych problemów, które powodują niezadowolenie naukowców zajmujących się danymi. Oto moja lista:

* Oczekiwanie nie pasuje do rzeczywistości. Wiele firm zatrudnia naukowców zajmujących się danymi, nie rozumiejąc, na czym polega nauka o danych. Na przykład bez odpowiedniej infrastruktury, aby zacząć czerpać korzyści z inwestycji w naukę danych, w połączeniu z faktem, że firmy te nie zatrudniają starszych lub doświadczonych praktyków danych przed zatrudnieniem juniorów, masz teraz przepis na rozczarowanie i nieszczęśliwą relację dla obu stron. Analityk danych prawdopodobnie wchodzi do firmy z ambicją napisania inteligentnych algorytmów uczenia maszynowego w celu generowania spostrzeżeń, ale wkrótce odkrywa, że nie może tego zrobić, ponieważ ich pierwszym zadaniem jest uporządkowanie infrastruktury danych i/lub tworzenie raportów na żądanie. W przeciwieństwie do tego, wiele razy poziom dojrzałości data science w firmie jest tak niski, że jedyne, czego chcą, to wykres, który mogą prezentować na posiedzeniu zarządu każdego dnia. Liderzy w takich firmach są sfrustrowani, ponieważ nie widzą wystarczająco szybkiego generowania wartości, a wszystko to oczywiście prowadzi do tego, że data scientist nie jest zadowolony z tej roli i ostatecznie odchodzi.

* Polityka firmy jest ważniejsza niż umiejętności analizy danych. Data scientist często zakłada, że znajomość wielu algorytmów uczenia maszynowego sprawi, że będzie najbardziej wartościową osobą w firmie. Jednak badacz danych wkrótce odkrywa, że te oczekiwania nie pasują do rzeczywistości. Prawda jest taka, że najbardziej wpływowe osoby w firmie muszą zobaczyć wartość każdego pracownika, któremu mają zamiar powierzyć większe obowiązki, niezależnie od tego, czy są specjalistami ds. danych, czy nie. Z perspektywy analityka danych oznacza to najpierw udostępnienie siebie, a następnie pracę nad tym, aby stać się niezastąpionym. Aby tak się stało, musisz być gotowy do obsługi ciągłego przepływu pracy ad hoc, takiej jak pobieranie liczb z bazy danych, aby przekazać je właściwym osobom we właściwym czasie, wykonywanie prostych projektów tylko po to, aby właściwi ludzie otrzymali właściwe postrzeganie Ciebie, naukowca danych, jako kogoś godnego zaufania, rzetelnego i innowacyjnego. Jakkolwiek frustrujące może to zabrzmieć, postawienie się na tym, że jest niezbędną częścią pracy, którą każdy analityk danych musi zaakceptować, jeśli ma nadzieję dotrzeć do punktu, w którym może osiągnąć coś bardziej interesującego i wywierającego wpływ.

* Rola badacza danych nie jest zrozumiała. Kontynuując robienie czegokolwiek, aby zadowolić właściwych ludzi, ci sami ludzie z całą tą mocą często nie rozumieją, co oznacza termin naukowiec danych. Oznacza to, że od analityków danych oczekuje się, że będą ekspertami w dziedzinie analityki, jak również osobami, które powinny się zgłaszać, i nie zapominajmy również o ekspertach od baz danych. Nie tylko dyrektorzy nietechniczni robią zbyt wiele założeń dotyczących umiejętności analityka danych: inni koledzy z dziedziny technologii zakładają, że badacz danych wie wszystko, co jest związane z danymi. Powszechna mądrość głosi, że specjalista ds. danych powinien znać się na Spark, Hadoop, Hive, Pig, SQL, Neo4J, MySQL, Python, R, Scala, TensorFlow, testowaniu A/B, NLP, wszystkim, co jest związane z uczeniem maszynowym i wszystkim innym, o czym możesz pomyśleć to jest związane z danymi. Ale to nie koniec. Ponieważ specjalista ds. danych podobno wie to wszystko i oczywiście ma dostęp do wszystkich danych, oczekuje się, że specjalista ds. danych otrzyma odpowiedzi na wszystkie

pytania w ciągu kilku minut. Próba poinformowania wszystkich o tym, co faktycznie wiesz i nad czym masz kontrolę, może być zarówno trudna, jak i frustrująca.

* Praca w odizolowanym zespole ogranicza produktywność. Kiedy widzisz udane komercyjne produkty danych, często widzisz profesjonalnie zaprojektowane interfejsy użytkownika z inteligentnymi funkcjami i, co najważniejsze, użytecznymi danymi wyjściowymi, które przynajmniej są postrzegane przez użytkowników jako rozwiązanie istotnego problemu. Teraz, jeśli naukowiec zajmujący się danymi spędza czas tylko na uczeniu się pisania i wykonywania algorytmów uczenia maszynowego, może być tylko niewielką (choć niezbędną) częścią zespołu, który prowadzi do sukcesu długiego wysiłku, który kończy się wytworzeniem cennych danych produkt. To jeden scenariusz – tak, część większego zespołu, ale często oznacza to bycie małym trybikiem w znacznie większej maszynie. Mimo to jest to prawdopodobnie lepsze niż odsunięcie się na bok i poproszenie o pracę w izolacji nad czymś „naukowym o danych". Po odcięciu od tych procesów, które faktycznie tworzą produkty do sprzedaży, zespoły zajmujące się analizą danych mają problemy z dostarczaniem wartości. Mimo to wiele firm wciąż prosi zespoły zajmujące się analizą danych o opracowanie własnych projektów i napisanie kodu, który rozwiąże zdefiniowany przez siebie problem. W niektórych przypadkach może to wystarczyć. Na przykład, jeśli wszystko, co jest potrzebne, to statyczny arkusz kalkulacyjny, który jest tworzony raz na kwartał, zespół może zapewnić pewną wartość. Z drugiej strony, jeśli zamiast tego celem jest optymalizacja dostarczania inteligentnych sugestii w regulowanej witrynie będzie to wymagało wielu różnych umiejętności, których nie należy oczekiwać od ogromnej większości naukowców zajmujących się danymi. (Tylko prawdziwy jednorożec zajmujący się analizą danych może rozwiązać ten problem.) Tak więc, jeśli zlecisz ten projekt izolowanemu zespołowi zajmującemu się analizą danych, odciętym od wszystkich innych zasobów, najprawdopodobniej się nie powiedzie (lub zajmie bardzo dużo czasu, ponieważ organizowanie odizolowanych zespołów do pracy nad wspólnymi projektami w dużych przedsiębiorstwach nie jest łatwe).

Sprawdzona w czasie mądrość dotycząca zarządzania zespołami wymaga powtórzenia: Zespoły naukowców zajmujących się danymi, podobnie jak inne, najlepiej rozwijają się, gdy istnieje skuteczne przywództwo, silny mandat od zespołu zarządzającego firmy i jasne cele oparte na solidnej i uzgodnionej strategii. Pamiętaj, że utrzymanie cenionych naukowców zajmujących się danymi w Twojej firmie wymaga nie tylko ścieżki dla zespołów zajmujących się analizą danych, aby podejmować kluczowe inicjatywy w sposób oparty na współpracy i zwinnie, od projektu po wdrożenie umożliwione przez odpowiednią infrastrukturę danych, ale także bardzo zarządzanie oczekiwaniami. W obu kierunkach.

Opracowanie architektury danych

Budowanie architektury danych jest podobne do tego, co dzieje się, gdy tradycyjny architekt projektuje dom lub budynek: najpierw stwórz plan, który jest zgodny z krótko- i długoterminowymi celami organizacji, a następnie upewnij się, że plan stanie się rzeczywistością. Ogólny pogląd jest taki, że architektura danych definiuje standardowy zestaw produktów i narzędzi, których organizacja używa do zarządzania danymi. Ale to znacznie więcej. Każda naprawdę efektywna architektura danych musi uwzględniać wyjątkowe wymagania kulturowe i kontekstowe organizacji, takie jak wielkość firmy, konfiguracja i branża, a także potencjalne ograniczenia techniczne, prawne, bezpieczeństwa lub inne. Ponadto architektura danych musi określać procesy przechwytywania, przekształcania i dostarczania użytecznych danych użytkownikom biznesowym. Co najważniejsze, identyfikuje ludzi, którzy będą konsumować te dane i ich unikalne wymagania biznesowe. Omówię to wszystko (i wiele więcej).

Definiowanie, co składa się na architekturę danych

W obszarze technologii informatycznych architektura danych składa się z modeli, polityk, reguł i standardów, które regulują, jakie dane są gromadzone, a także w jaki sposób są przechowywane, porządkowane, integrowane i wykorzystywane w systemach danych i organizacjach. Dane są zwykle jedną z kilku domen architektonicznych, które tworzą filary architektury korporacyjnej lub architektury rozwiązań dla wewnętrznych operacji biznesowych lub dla komercyjnego portfolio produktów danych lub usług oferowanych na zewnątrz. Architektura danych powinna określać standardy danych dla wszystkich systemów danych jako wizję lub model interakcji między różnymi systemami danych organizacji. Na przykład integracja danych zależy od standardów i struktur architektury danych używanych przez różne jednostki biznesowe oraz wybrane aplikacje systemowe i definiuje sposób, w jaki musi działać interakcja danych. Te standardy i struktury dotyczą danych w pamięci i danych w ruchu oraz zawierają opisy rozwiązań do przechowywania danych, kategorii danych i typów danych, w tym mapowania tych jednostek danych do poziomów jakości danych, odpowiednich aplikacji, wykorzystania lub lokalizacji przechowywania i tak dalej. Jednym z kluczowych elementów realizacji celów biznesowych firmy przez architekturę danych jest sposób, w jaki architektura danych opisuje sposób przetwarzania, przechowywania i wykorzystywania danych w środowisku uprzemysłowionym lub w pracy systemu. Musi zapewniać kryteria operacji przetwarzania danych, aby umożliwić projektowanie przepływów danych, a także sterowanie przepływem danych w cyklu życia nauki o danych. Jeśli chodzi o zdefiniowanie ogólnej architektury danych, stroną odpowiedzialną jest tutaj oczywiście architekt danych. Jednak architekt danych jest również zazwyczaj kluczową osobą, której zadaniem jest upewnienie się, że projekt architektury danych jest przestrzegany i rozumiany jako część realizacji i rozbudowy rzeczywistej infrastruktury nauki o danych. Może to oczywiście obejmować również modyfikacje samej architektury danych, ze względu na rzeczywiste dostosowania, które muszą nastąpić w oparciu o potencjalne ograniczenia prawne, bezpieczeństwa, etyczne, geograficzne, kulturowe lub techniczne, które pojawiają się po wprowadzeniu planu architektury danych w praktyce.

Opisywanie tradycyjnych podejść architektonicznych

Architektura danych obejmuje pełną analizę relacji między funkcjami organizacji, dostępnymi technologiami i typami danych. Definiując architekturę danych dla swojej firmy, powinieneś podejść do swojego zadania, mając na uwadze te trzy perspektywy:

* Koncepcyjne: Koncepcyjna architektura danych, czasami nazywana również semantycznym modelem danych, reprezentuje wszystkie istotne jednostki biznesowe z perspektywy danych.

* Logiczne: Logiczna architektura danych, zwana także modelem danych systemowych, reprezentuje logikę tego, w jaki sposób zawarte jednostki danych są powiązane i połączone ze sobą z perspektywy przepływu danych.

* Fizyczne: Fizyczna architektura danych reprezentuje rzeczywistą realizację architektury w jej fizycznym środowisku - innymi słowy, w jaki sposób rzeczywista architektura danych jest implementowana jako część infrastruktury technologicznej.

Architekturę danych należy określić na etapie planowania nowej konfiguracji infrastruktury danych. W ramach tego procesu Twoja strategia dotycząca danych musi uchwycić – w sposób kompletny, spójny i zrozumiały – wszystkie główne kategorie i typy danych, a także źródła danych niezbędne do wspierania strategicznych ambicji przedsiębiorstwa. Podstawowym wymaganiem na tym wczesnym etapie planowania jest zdefiniowanie wszystkich odpowiednich kategorii danych i typów danych w odniesieniu do potrzeb i celów biznesowych Twojej organizacji, a nie określanie, które narzędzia lub aplikacje powinny być używane do radzenia sobie z nimi.

Elementy architektury danych

Jeśli chodzi o architekturę danych, ważne jest, aby pewne elementy zostały zdefiniowane już na etapie projektowania. Na przykład musisz zdefiniować strukturę administracyjną oraz powiązane metodologie i procesy wymagane do zarządzania danymi na różnych etapach ich cyklu życia. Niezwracanie wystarczającej uwagi na to, jak ważne jest administrowanie zarówno danymi, jak i architekturą danych, może spowodować chaos, uszkodzenie danych lub poważny cios w integralność danych – a każdy z nich może poważnie wpłynąć na wartość i użyteczność danych dla Twojej firmy. Ważną częścią architektury danych jest opis wyborów technologicznych. Czy Twoja architektura zostanie zrealizowana w środowisku zwirtualizowanym i opartym na chmurze, czy na przykład za pomocą rozwiązania lokalnego? A może realizacja będzie obejmować lokalną, pojedynczą lokację i instancję, czy też zostanie wdrożona w większej, wielolokacyjnej konfiguracji? Czy może w ogóle będzie dystrybuowany globalnie? Wszystkie te pytania muszą zostać zrozumiane i odpowiedzieć na nie już na wczesnym etapie, aby architektura danych została zaprojektowana w sposób wspierający cele biznesowe. Rozważ rodzaje interfejsów, których inne systemy będą potrzebowały, aby uzyskać dostęp do danych, a także rodzaj projektu infrastruktury niezbędnej do obsługi typowych operacji na danych (na przykład procedury awaryjne, import danych, tworzenie kopii zapasowych danych i zewnętrzne transfery danych). wskazówki dotyczące prawidłowo zaimplementowanego projektu architektury danych, możesz mieć wspólne operacje na danych zaimplementowane na bardzo różne sposoby, w zależności od tego, gdzie jesteś w organizacji. Tak szalone podejście do pikowania sprawia, że niezwykle trudno jest zrozumieć i kontrolować przepływ danych w Twojej organizacji. Ten rodzaj fragmentacji jest wysoce niepożądany, nie tylko ze względu na potencjalnie zwiększony koszt, ale także z powodu rozłączeń danych, które może się wiązać. Tego rodzaju trudności nie są rzadkością w szybko rozwijających się przedsiębiorstwach lub w przedsiębiorstwach, które mają szerokie portfolio produktów i usług obsługujących różne branże. Prawidłowo przeprowadzona faza projektowania architektury danych zmusza organizację do precyzyjnego określenia i opisania zarówno wewnętrznych, jak i zewnętrznych przepływów informacji. Są to wzorce, których organizacja mogła wcześniej nie poświęcić czasu na konceptualizację i właściwe przemyślenie. Na tym etapie możliwe jest zatem zidentyfikowanie kosztownych braków informacji, rozłączeń między działami oraz rozłączeń między systemami organizacyjnymi a danymi, które mogły nie być widoczne przed analizą architektury danych.

Poznawanie cech nowoczesnej architektury danych

Nadal czekasz na konkretną definicję tego, czym właściwie jest architektura danych? Zacznij od przyjrzenia się tym cechom, które architektura danych musi po prostu obejmować:

* Orientacja biznesowa: zamiast koncentrować się na danych lub technologii w fazie definiowania, nowoczesna architektura danych zaczyna się od użytkowników biznesowych i ogólnego celu biznesowego, a następnie płynie wstecz. Klienci mogą być wewnętrzni lub zewnętrzni w organizacji, a ich potrzeby mogą się różnić w zależności od roli, działu i w czasie. Dlatego dobra architektura danych stale ewoluuje, aby sprostać nowym i zmieniającym się potrzebom danych biznesowych i klientów.

* Możliwość dostosowania: W nowoczesnej architekturze danych dane łatwo przepływają z systemów źródłowych do użytkowników biznesowych. Celem architektury jest zarządzanie tym przepływem poprzez utworzenie serii połączonych i dwukierunkowych potoków danych, które służą różnym stale zmieniającym się potrzebom biznesowym.

* Automatyzacja: Aby stworzyć łatwo adaptowalną architekturę, w której dane przepływają w sposób ciągły, a integralność danych jest chroniona, architektura musi być maksymalnie zautomatyzowana. Architektura musi zapewniać profilowanie i tagowanie danych w momencie pozyskiwania danych oraz mapować je do istniejących zestawów danych i atrybutów – nawiasem mówiąc, jest to również kluczowa funkcja tworzenia katalogów danych. W ten sam sposób architektura danych musi również umożliwiać wykrywanie zmian w źródłach danych, a także kwantyfikację wpływu zmiany na dowolnym elemencie architektonicznym w dowolnym momencie. W środowisku produkcyjnym czasu rzeczywistego musi być w stanie wykrywać anomalie w locie i albo powiadamiać odpowiednie instancje (człowiek i/lub maszynę), albo w razie potrzeby wyzwalać alerty.

* Inteligencja: idealna architektura danych to coś więcej niż tylko automatyzacja; wykorzystuje uczenie maszynowe i sztuczną inteligencję do rzeczywistego budowania obiektów danych, tabel, widoków i modeli, które utrzymują przepływ danych. Innymi słowy, wykorzystuje inteligencję nie tylko do analizy danych, ale także w ramach zarządzania i przetwarzania danych. Uczenie maszynowe i sztuczną inteligencję można stosować do identyfikowania typów danych, znajdowania wspólnych kluczy i łączenia ścieżek, identyfikowania i naprawiania błędów jakości danych, mapowania tabel, identyfikowania relacji, rekomendowania powiązanych zestawów danych i analiz itd. Nowoczesna architektura danych wykorzystuje inteligencję do uczenia się, dostosowywania, ostrzegania i rekomendowania, dzięki czemu ludzie, którzy zarządzają środowiskiem i korzystają z niego, są wydajniejsi i wydajniejsi w swojej pracy.

* Elastyczność: Nowoczesna architektura danych musi być wystarczająco elastyczna, aby sprostać różnorodnym potrzebom biznesowym. Oznacza to, że musi obsługiwać wiele typów użytkowników biznesowych, operacje ładowania i częstotliwości odświeżania (wsadowe, miniwsadowe i strumieniowe), operacje zapytań (tworzenie, odczytywanie, aktualizowanie, usuwanie), wdrożenia (lokalne, w chmurze publicznej, prywatne chmura, hybryda), silniki przetwarzania danych (relacyjne, OLAP, MapReduce, SQL, graphing, mapping, programmatic) oraz potoki (hurtownia danych, data mart, kostki OLAP, visual discovery, aplikacje operacyjne czasu rzeczywistego). Nowoczesna architektura danych musi być wszystkim dla wszystkich osób w firmie w danym momencie.

* Duch współpracy: w przeciwieństwie do przeszłości, w której dział IT budował wszystko, nowoczesna architektura danych zwykle dzieli odpowiedzialność za pozyskiwanie i przekształcanie danych między dział IT a jednostki biznesowe. Dział IT może nadal wykonywać ciężkie zadania związane z pozyskiwaniem danych z wewnętrznych systemów operacyjnych i tworzyć ogólne bloki konstrukcyjne wielokrotnego użytku. Jednak dane z zewnętrznych źródeł danych, takich jak dane z mediów społecznościowych, dane klientów, dane dotyczące wydajności produktów ze środowisk na żywo itp., są zwykle gromadzone przez firmę. Powodem jest to, że jednostki biznesowe już posiadają ten interfejs, tak jak IT jest właścicielem interfejsu do systemów wewnętrznych. Pozwalanie działowi IT skoncentrować się na infrastrukturze szkieletowej konfiguracji i zarządzaniu przechowywaniem

danych, a także na transferze danych, jest zwykle dobrym podziałem — po pobraniu i przyjęciu danych inżynierowie danych w jednostkach biznesowych są gotowi do przygotowania danych i katalogu danych narzędzia do tworzenia niestandardowych zestawów danych w celu zasilania działań analitycznych i uczenia maszynowego jednostek biznesowych prowadzonych przez analityków danych i analityków biznesowych. Ta współpraca między inżynierami danych i analitykami danych oznacza, że dział IT nie musi być zaangażowany w szczegóły biznesowe związane z danymi

* Łatwość zarządzania: Nowoczesna architektura danych definiuje punkty dostępu dla każdego typu użytkownika w celu zaspokojenia ich potrzeb w zakresie danych. Z lotu ptaka zazwyczaj mamy cztery typy użytkowników biznesowych – konsumentów danych, eksploratorów danych, analityków danych i analityków danych – a każdy typ wymaga innego punktu dostępu do danych. Zapewnienie dostępu to podstawa zarządzania, co oznacza, że zarządzanie, co zaskakujące, jest tak naprawdę kluczem do dobrego środowiska samoobsługowego.

* Prostota: Twoim pierwszym założeniem powinno być zawsze, że najprostsza architektura jest zwykle najlepszą architekturą. Zapewnienie takiej prostoty może być jednak dość trudne, biorąc pod uwagę różnorodność potrzeb w zakresie danych i złożoność komponentów we współczesnej architekturze danych. Aby zastosować zasadę prostoty, organizacja z małymi zestawami danych powinna poważnie rozważyć gotowe narzędzie analityczne z wbudowanym środowiskiem zarządzania danymi. Aby zmniejszyć złożoność w kontekście dużych zbiorów danych i uniknąć tworzenia sztywnego środowiska, organizacje powinny dążyć do ograniczenia przepływu danych i powielania danych poprzez promowanie ujednoliconej struktury danych, ram integracji danych oraz zharmonizowanego środowiska analitycznego i uczenia maszynowego wspierającego innowacje i eksperymenty, bez zwiększania złożoności infrastruktury. Dokładnie, jak do tego podejść, opisano w dalszej części tego rozdziału.

* Skalowalność: w dobie dużych zbiorów danych i zmiennych obciążeń, organizacje potrzebują skalowalnej, elastycznej architektury, która na żądanie dostosowuje się do zmieniających się wymagań przetwarzania danych. Wiele firm gromadzi się obecnie wokół platform chmurowych (zarówno publicznych, jak i prywatnych), aby uzyskać skalowalność na żądanie w przystępnych cenach. Elastyczne architektury uwalniają administratorów od konieczności dokładnego kalibrowania pojemności, kontrolowania wykorzystania w razie potrzeby i ciągłego kupowania sprzętu. Skalowalność tworzy również wiele typów aplikacji i przypadków użycia, takich jak środowiska programistyczne i testowe na żądanie, analityczne piaskownice i place zabaw prototypów.

* Bezpieczeństwo: nowoczesna architektura danych musi być innowacyjną przestrzenią roboczą do współpracy, a jednocześnie bezpieczną, niezawodną i godną zaufania. Musi być w stanie zapewnić autoryzowanym użytkownikom łatwy dostęp do danych, jednocześnie trzymając na dystans hakerów i intruzów. Musi to wszystko robić, zachowując jednocześnie zgodność z przepisami dotyczącymi prywatności, w tym ustawami rządowymi, takimi jak ustawa o przenośności i odpowiedzialności w ubezpieczeniach zdrowotnych (HIPPA) w USA, a także przepisami, takimi jak ogólne rozporządzenie o ochronie danych UE (RODO). Architektura danych szyfruje dane po ich przyjęciu do magazynu danych, maskując informacje umożliwiające identyfikację osób i śledzi wszystkie elementy danych w katalogu danych, w tym ich pochodzenie, wykorzystanie i ścieżkę audytu. Zarządzanie cyklem życia zapewnia, że każdy obiekt danych ma właściciela, lokalizację i zdefiniowany okres przechowywania.

* Odporność: Architektura danych musi być również odporna, z wysoką dostępnością, solidnymi możliwościami odzyskiwania po awarii oraz stabilną infrastrukturą do tworzenia kopii zapasowych i przywracania danych. Jest to szczególnie widoczne w nowoczesnej architekturze danych, która często działa na ogromnych farmach serwerów w chmurze, gdzie awarie są powszechne. Dobrą wiadomością

jest to, że wielu dostawców usług w chmurze oferuje wbudowaną nadmiarowość i przełączanie awaryjne z dobrymi umowami o poziomie usług (SLA) i umożliwia firmom konfigurowanie obrazów lustrzanych na potrzeby odzyskiwania po awarii w geograficznie rozproszonych centrach danych przy niskich kosztach.

Wyjaśnienie warstw architektury danych

Decydując się na architekturę danych, musisz być świadomy różnych ograniczeń i wpływów, ponieważ może to mieć wpływ na projekt architektury. Obejmują one aspekty, których można się spodziewać, takie jak wymagania biznesowe, kluczowe wybory technologiczne, względy finansowe, różne rodzaje zasad biznesowych i potrzeby w zakresie przetwarzania danych. Ale ważne jest również, aby zrozumieć główne warstwy architektoniczne, które stanowią podstawę każdej architektury danych. W przeszłości organizacje budowały dość statyczne, oparte na IT architektury danych, w których projektowanie systemów było złożone, trudne i czasochłonne, a uszkodzenie bazy danych wpływało na praktycznie wszystkie aplikacje działające w środowiskach. Nazywano je hurtowniami danych. Ze względu na podstawową technologię i wzorce projektowe większość hurtowni danych zajmuje się tworzeniem i zarządzaniem armią ludzi, co zapewnia minimalny zwrot z inwestycji. Większość z nich to przewartościowane korporacyjne zrzuty danych, w których organizacja przechowuje wszystkie zebrane dane bez określonego celu i struktury w przekonaniu, że samo zbieranie i zrzucanie danych w hurtowni danych stanowi wartość dodaną samą w sobie. Istnieje jednak kilka przykładów dobrze zaprojektowanych i udanych wdrożeń, które zapewniają dobrze funkcjonujące środowisko do analizy danych. Nowoczesna architektura danych może nadal działać częściowo jak hurtownia danych, ale najlepiej byłoby, gdyby była to elastyczna, skalowalna i sprawna hurtownia danych. Pamiętaj tylko, że aspekty przechowywania w hurtowni danych stanowią tylko jeden potencjalny element nowoczesnej architektury danych. Do nowego środowiska danych należy podchodzić jak do żywego, oddychającego organizmu, który wykrywa zmiany i reaguje na nie, nieustannie uczy się i dostosowuje oraz zapewnia kontrolowany, dostosowany dostęp dla każdej osoby.

Każda warstwa ma określony cel w architekturze danych, w oparciu o specyficzny kontekst biznesowy Twojej firmy. Oznacza to, że konfiguracja realizacji architektury danych i komponenty, które zdecydujesz się użyć, mogą wyglądać zupełnie inaczej, w zależności od tego, czy koncentrujesz się na wewnętrznej analizie biznesowej, czy też rozwijasz architekturę do obsługi komercyjnego produktu lub usługi danych na skalę globalną. Wykorzystywanie metod nauki o danych do pracy nad projektowaniem architektury danych to doskonały sposób na uzyskanie architektury danych, która obsługuje sposób, w jaki Twoja firma musi działać. Kiedy już to zdefiniujesz, możesz zastosować potrzebne systemy i narzędzia, aby zrealizować swoją architekturę w kompleksowej infrastrukturze nauki o danych. Ale zacznijmy od bardziej szczegółowego przyjrzenia się każdej warstwie.

* Warstwa źródła i pozyskiwania danych: ta warstwa ma na celu upewnienie się, że rozumiesz swoje potrzeby w zakresie danych w oparciu o cel biznesowy oraz co to oznacza, jeśli chodzi o lokalizację danych, ich właściciel, ich wielkość, czy są one wrażliwe (i w związku z tym musi być zanonimizowany), w jaki sposób należy je zbierać, częstotliwość zbierania i tak dalej. Wskazane jest również wykonanie pierwszych czynności przetwarzania danych już w miejscu ich zbierania, ponieważ nie chcesz tracić czasu i pieniędzy na zbieranie i przechowywanie brudnych danych. Przykłady czynności przetwarzania na wczesnym etapie przed przechowywaniem obejmują sprawdzanie kompletności zebranych danych, identyfikację i usuwanie zduplikowanych rekordów danych, agregację danych w celu zminimalizowania wpływu na zdolność przesyłową, anonimizację danych osobowych, szyfrowanie danych i tak dalej.

* Warstwa przechowywania danych i obliczeń: W tej warstwie architektury danych należy wziąć pod uwagę takie aspekty, jak sposób przechowywania danych (na przykład okresy przechowywania dla

różnych typów danych). Należy również rozważyć następny poziom przetwarzania danych (czyszczenie danych, mapowanie, etykietowanie itd.) oraz sposób, w jaki można to zrobić przy jak najmniejszej interwencji ręcznej. (Automatyzacja zadań zarządzania danymi pomaga chronić integralność danych, nie dodając nieświadomie stronniczości do danych). W tej warstwie musisz również zdecydować, czy chcesz przechowywać i przetwarzać dane przy użyciu rozwiązania lokalnego, czy rozwiązania w chmurze. To, co wybierzesz, może zależeć od rozmiaru danych, z którymi pracujesz, ale także od tego, czy spodziewasz się szybkiego i nieoczekiwanego wzrostu danych, co sugerowałoby, że potrzebujesz rozwiązania opartego na chmurze do szybkiego i łatwego skalowania środowisko. Jednak rozwiązanie chmury publicznej – na przykład Amazon, Azure lub Google – pasuje idealnie (nawet w przypadku małych środowisk danych), ponieważ eliminuje początkowe koszty inwestycji we własną infrastrukturę, ponieważ płacisz tylko za wykorzystaną pojemność. Innym czynnikiem, który należy wziąć pod uwagę, jest to, że jeśli musisz zbierać i obliczać dane z wielu różnych krajów w swoim środowisku, możesz nie być w stanie przesłać danych ze wszystkich krajów ze względu na surowe przepisy i regulacje. Aby rozwiązać ten problem, należy rozważyć konfigurację chmury rozproszonej, w której dane mogą być przetwarzane w centrum danych w granicach kraju, ale wgląd i modele mogą być przesyłane do instancji centralnej w celu ponownego wykorzystania i udostępniania w konfiguracjach rozproszonych.

* Warstwa aplikacyjna: Warstwa aplikacyjna jest prosta. Jak sama nazwa wskazuje, jest to miejsce, w którym wdrażasz aplikacje i narzędzia, które chcesz uruchomić na swoich danych. Może to być mieszanka różnych aplikacji, takich jak narzędzia i frameworki open source, takie jak TensorFlow, Scikit-learn lub Keras, ale także gotowe aplikacje od uznanych dostawców analityki, takich jak SAS, IBM lub Tableau. Tutaj ważne jest, aby zastanowić się, jakiego typu użytkowników będziesz mieć w swoim środowisku, a także ich poziom kompetencji i zainteresowania. Być może wszystko, co musisz zrobić, to upewnić się, że spostrzeżenia są łatwo przekazywane za pomocą różnych wstępnie zaprojektowanych pulpitów nawigacyjnych i wizualizacji wspierających podejmowanie decyzji; jednak najlepszym podejściem jest zwykle upewnienie się, że zbudowałeś architekturę danych w taki sposób, aby można było wymieniać aplikacje i wymieniać je, w zależności od tego, co stanie się dostępne w branży, a także wspierać zmieniające się oczekiwania i potrzeby w organizacji. Użytkownicy mogą zacząć od chęci dostarczenia im rzeczy, ale wraz ze wzrostem dojrzałości użytkownicy mogą chcieć robić więcej sami. Dotyczy to zwłaszcza firm inwestujących w komercyjne produkty i usługi związane z danymi. Elastyczność w warstwie aplikacji nie jest głównym kosztem architektury. Jeśli dolne warstwy w architekturze są wspólne dla wszystkich, wiele można wygrać pod względem obniżenia kosztów oraz zwiększenia integralności i niezawodności danych. Elastyczność w warstwie aplikacji zapewnia zadowolenie użytkowników i spełnienie różnych potrzeb. Minimalizuje również ryzyko rozgałęziania się niezadowolonych użytkowników i budowania własnego środowiska, zwiększając w ten sposób całkowite koszty firmy oraz tworząc silosowe dane i insighty.

* Środowisko docelowe: odnosi się do środowiska, w którym zamierzasz wdrożyć swoje spostrzeżenia i modele. Ponownie, to, co tak naprawdę jest, różni się w zależności od firmy. Jeśli architektura danych jest budowana wyłącznie na potrzeby analityki wewnętrznej, środowiskiem docelowym mogą być różne systemy wewnętrzne, ale może również odnosić się do tego, w jaki sposób wgląd w różne struktury organizacyjne i fora decyzyjne są wykorzystywane w różnych strukturach i forach decyzyjnych. Na żywo operacyjny lub w przypadku komercyjnego produktu lub usługi danych, środowisko docelowe może odnosić się do rzeczywistego środowiska produkcyjnego. Dane wyjściowe z tych środowisk docelowych są następnie przesyłane z powrotem do źródła danych i warstwy akwizycji, dostarczając dane zwrotne poprzez zmianę wdrożoną w ramach działającego środowiska produkcyjnego, a cykl analizy danych rozpoczyna się od nowa.

Lista najważniejszych technologii dla nowoczesnej architektury danych

Obecnie głównym celem jest refaktoryzacja platformy technologicznej dla przedsiębiorstw, aby umożliwić szybszy, łatwiejszy i bardziej elastyczny dostęp do dużych ilości cennych danych. Ta refaktoryzacja jest niemałym przedsięwzięciem i zwykle jest inicjowana przez zmieniający się zestaw kluczowych czynników biznesowych. Mówiąc najprościej, platformy, które dominowały w IT dla przedsiębiorstw od prawie 30 lat, nie są już w stanie obsłużyć obciążeń potrzebnych do napędzania opartego na danych firmy do przodu. Organizacje od dawna są ograniczane w korzystaniu z danych przez niekompatybilne formaty, ograniczenia tradycyjnych baz danych oraz niemożność elastycznego łączenia danych z wielu źródeł. Nowe technologie zaczynają teraz spełniać obietnicę zmiany tego wszystkiego. Ulepszenie modelu wdrażania oprogramowania to jeden z głównych kroków w usuwaniu barier w wykorzystaniu danych. Większa elastyczność danych wymaga również bardziej elastycznych baz danych i bardziej skalowalnych platform przesyłania strumieniowego w czasie rzeczywistym. W rzeczywistości potrzeba co najmniej siedmiu podstawowych technologii, aby zapewnić przedsiębiorstwu elastyczną, nowoczesną architekturę danych w czasie rzeczywistym. Te siedem kluczowych technologii opisano w poniższych sekcjach.

Bazy danych NoSQL

System zarządzania relacyjnymi bazami danych (RDBMS) dominuje na rynku baz danych od prawie 30 lat, jednak tradycyjna relacyjna baza danych okazała się mniej niż wystarczająca do obsługi stale rosnących ilości danych i przyspieszonego tempa, w jakim dane muszą być obsługiwane. Bazy danych NoSQL – „bez SQL”, ponieważ jest zdecydowanie nierelacyjne – przejęły kontrolę ze względu na szybkość i możliwość skalowania. Zapewniają mechanizm przechowywania i wyszukiwania danych modelowanych w sposób inny niż relacje tabelaryczne stosowane w relacyjnych bazach danych. Ze względu na swoją szybkość bazy danych NoSQL są coraz częściej wykorzystywane w aplikacjach internetowych typu big data i czasu rzeczywistego. Bazy danych NoSQL oferują prostotę projektowania, prostsze skalowanie poziome do klastrów maszyn (prawdziwy problem w przypadku relacyjnych baz danych) i lepszą kontrolę nad dostępnością. Struktury danych używane przez bazy danych NoSQL (na przykład klucz-wartość, szeroka kolumna, wykres lub dokument) różnią się od struktur używanych domyślnie w relacyjnych bazach danych, dzięki czemu niektóre operacje są szybsze w NoSQL. Konkretna przydatność danej bazy danych NoSQL zależy od problemu, który musi rozwiązać. Czasami struktury danych używane przez bazy danych NoSQL są również postrzegane jako bardziej elastyczne niż tabele relacyjnych baz danych.

Platformy strumieniowe w czasie rzeczywistym

Odpowiadanie klientom w czasie rzeczywistym ma kluczowe znaczenie dla obsługi klienta. Nie jest tajemnicą, dlaczego branże mające kontakt z konsumentami, czyli konfiguracje B2C (Business-to-Consumer), innymi słowy, doświadczyły ogromnych zakłóceń w ciągu ostatnich dziesięciu lat. Ma to wszystko, co wiąże się ze zdolnością firm do reagowania na użytkownika w czasie rzeczywistym. Poinformowanie klienta, że w ciągu 24 godzin będzie gotowa oferta, nie jest dobre, ponieważ wykonał już decyzję, którą podjął 23 godziny temu. Przejście na model czasu rzeczywistego wymaga przesyłania strumieniowego zdarzeń. Aplikacje oparte na wiadomościach istnieją od lat, ale dzisiejsze platformy strumieniowe skalują się znacznie lepiej i przy znacznie niższych kosztach niż ich poprzednicy. Niedawny postęp w technologiach przesyłania strumieniowego otwiera drzwi na wiele nowych sposobów optymalizacji biznesu. Jedną z korzyści jest reagowanie na klienta w czasie rzeczywistym. Kolejnym aspektem do rozważenia są korzyści dla rozwoju. Zapewniając zespołom programistów pętlę informacji zwrotnych w czasie rzeczywistym, strumienie zdarzeń mogą również pomóc firmom poprawić jakość produktów i szybciej udostępniać nowe oprogramowanie.

Docker i kontenery

Docker to program komputerowy, który wykonuje wirtualizację na poziomie systemu operacyjnego, znaną również jako konteneryzacja. Wydany po raz pierwszy w 2013 roku przez firmę Docker, Inc., Docker służy do uruchamiania pakietów oprogramowania zwanych kontenerami, metody wirtualizacji, która pakuje kod aplikacji, konfiguracje i zależności w bloki konstrukcyjne w celu zapewnienia spójności, wydajności, produktywności i kontroli wersji. Kontenery są odizolowane od siebie i zawierają własne aplikacje, narzędzia, biblioteki i pliki konfiguracyjne oraz mogą komunikować się ze sobą za pomocą dobrze zdefiniowanych kanałów. Wszystkie kontenery są obsługiwane przez jedno jądro systemu operacyjnego, dzięki czemu są lżejsze niż maszyny wirtualne. Kontenery są tworzone z obrazów, które określają ich dokładną zawartość. Obraz kontenera to samodzielny program, który zawiera wszystko, czego potrzebuje do uruchomienia, na przykład kod, narzędzia i zasoby. Kontenery przynoszą znaczne korzyści zarówno deweloperom i operatorom, jak i samej organizacji. Tradycyjne podejście do izolacji infrastruktury polegało na partycjonowaniu statycznym, czyli przydzielaniu do każdego obciążenia oddzielnego, stałego wycinka zasobów, takiego jak serwer fizyczny lub maszyna wirtualna. Partycje statyczne ułatwiały rozwiązywanie problemów, ale kosztem dostarczenia znacznie niewykorzystanego sprzętu. Na przykład serwery internetowe zużywałyby średnio tylko około 10 procent całkowitej dostępnej mocy obliczeniowej. Ogromną zaletą technologii kontenerowej jest jej zdolność do tworzenia nowego rodzaju izolacji. Ci, którzy najmniej rozumieją kontenery, mogą wierzyć, że mogą osiągnąć te same korzyści, korzystając z narzędzi do automatyzacji, takich jak Ansible, Puppet lub Chef, ale w rzeczywistości technologie te nie mają istotnych możliwości. Bez względu na to, jak bardzo się starasz, te narzędzia do automatyzacji nie mogą zapewnić izolacji wymaganej do swobodnego przenoszenia obciążeń między różnymi konfiguracjami infrastruktury i sprzętu. Ten sam kontener może działać na sprzęcie bare-metal w lokalnym centrum danych lub na maszynie wirtualnej w chmurze publicznej. Żadne zmiany nie są konieczne. Na tym polega prawdziwa mobilność przy obciążeniach.

Repozytoria kontenerów

Repozytorium obrazów kontenerów to zbiór powiązanych obrazów kontenerów, zwykle udostępniających różne wersje tej samej aplikacji lub usługi. Ma to kluczowe znaczenie dla utrzymania sprawności w Twojej infrastrukturze. Bez procesu DevOps z ciągłymi dostawami do tworzenia obrazów kontenerów i repozytorium do ich przechowywania, każdy kontener musiałby zostać zbudowany na każdej maszynie, na której ten kontener mógłby działać. Dzięki repozytorium obrazy kontenerów można uruchamiać na dowolnym komputerze skonfigurowanym do odczytu z tego repozytorium. Sytuacja staje się jeszcze bardziej skomplikowana, gdy mamy do czynienia z wieloma centrami danych. Jeśli obraz kontenera jest zbudowany w jednym centrum danych, jak przenieść obraz do innego centrum danych? W idealnym przypadku, wykorzystując konwergentną platformę danych, będziesz mieć możliwość dublowania repozytorium między centrami danych. Kluczowym szczegółem jest tutaj to, że możliwości dublowania między lokalnymi a chmurą mogą znacznie różnić się od lokalnych centrów danych. Konwergentna platforma danych rozwiąże ten problem, oferując te możliwości niezależnie od infrastruktury fizycznej lub chmury, z której korzystasz w swojej organizacji.

Orkiestracja kontenerów

Zamiast statycznych partycji sprzętowych, każdy kontener wydaje się być całkowicie własnym prywatnym systemem operacyjnym. W przeciwieństwie do maszyn wirtualnych kontenery nie wymagają statycznej partycji obliczeniowej danych i pamięci. Umożliwia to administratorom uruchamianie dużej liczby kontenerów na serwerach bez konieczności martwienia się o dokładną ilość pamięci. Dzięki narzędziom do orkiestracji kontenerów, takim jak Kubernetes, można łatwo

uruchamiać kontenery, zabijać je, przenosić i ponownie uruchamiać w innym miejscu środowiska. Zakładając, że masz na miejscu nowe komponenty infrastruktury (na przykład bazę danych dokumentów, taką jak MapR-DB lub MongoDB) oraz platformę do strumieniowego przesyłania zdarzeń (może MapR-ES lub Apache Kafka) z narzędziem do orkiestracji (na przykład Kubernetes), jaki jest następny krok? Z pewnością będziesz musiał zaimplementować proces DevOps, aby wymyślać ciągłe kompilacje oprogramowania, które można następnie wdrożyć jako kontenery Dockera. Większym pytaniem jest jednak to, co właściwie powinieneś wdrożyć w utworzonych przez siebie kontenerach. To prowadzi nas do mikrousług.

Mikrousługi

Mikrousługi to technika tworzenia oprogramowania, która tworzy aplikację jako zbiór usług, które:

* Są łatwe w utrzymaniu i testowaniu

* Są luźno połączone

* Są zorganizowane wokół możliwości biznesowych

* Może być wdrażany niezależnie

W związku z tym mikrousługi łączą się, tworząc architekturę mikrousług, która umożliwia ciągłe dostarczanie/wdrażanie dużych, złożonych aplikacji, a także umożliwia organizacji rozwijanie stosu technologicznego – zestawu oprogramowania zapewniającego infrastrukturę dla komputera lub serwera . Zaletą podziału aplikacji na różne, mniejsze usługi jest to, że poprawia modułowość, co z kolei sprawia, że aplikacja jest łatwiejsza do zrozumienia, rozwijania i testowania, a także staje się bardziej odporna na erozję architektury – naruszenia architektury systemu, które prowadzą do znaczących problemy w systemie i przyczyniają się do jego wzrastającej kruchości. Dzięki architekturze mikrousług małe autonomiczne zespoły mogą działać równolegle, aby niezależnie opracowywać, wdrażać i skalować swoje usługi. Pozwala także na wyłonienie się architektury indywidualnej usługi poprzez ciągłą refaktoryzację - zdyscyplinowaną technikę restrukturyzacji istniejącego kodu, zmieniającą jego wewnętrzną strukturę bez zmiany jej zewnętrznego zachowania (a tym samym zapewniając, że nadal pasuje do otoczenia architektonicznego). Pojęcie mikroserwisów nie jest niczym nowym. Dzisiejsza różnica polega na tym, że technologie wspomagające, takie jak bazy danych NoSQL, przesyłanie strumieniowe zdarzeń i orkiestracja kontenerów, można skalować wraz z tworzeniem tysięcy mikrousług. Bez tych nowych podejść do przechowywania danych, przesyłania strumieniowego zdarzeń i aranżacji infrastruktury wdrażanie mikrousług na dużą skalę nie byłoby możliwe. Infrastruktura potrzebna do zarządzania ogromnymi ilościami danych, zdarzeń i instancji kontenerów nie byłaby w stanie skalować się do wymaganych poziomów. Mikroserwisy mają na celu zapewnienie elastyczności. Usługa, która ma charakter mikro, zazwyczaj składa się z pojedynczej funkcji lub niewielkiej grupy powiązanych funkcji. Im mniejsza i bardziej skoncentrowana jednostka funkcjonalna pracy, tym łatwiej będzie tworzyć, testować i wdrażać usługę. Usługi te muszą być oddzielone, co oznacza, że możesz wprowadzać zmiany w dowolnej usłudze bez wpływu na jakąkolwiek inną usługę. Jeśli tak nie jest, tracisz elastyczność obiecaną przez koncepcję mikrousług. Wprawdzie rozdzielenie nie może być bezwzględne — mikrousługi mogą oczywiście polegać na innych usługach — ale zależność powinna opierać się na zrównoważonych interfejsach API REST lub strumieniach zdarzeń. (Korzystanie ze strumieni zdarzeń pozwala na wykorzystanie tematów żądań i odpowiedzi, dzięki czemu można łatwo śledzić historię zdarzeń; takie podejście jest dużym plusem, jeśli chodzi o rozwiązywanie problemów, ponieważ cały przepływ żądań i wszystkie dane w żądania mogą być odtwarzane w dowolnym momencie.)

Funkcjonować jako usługa

Podobnie jak idea mikrousług przyciągnęła duże zainteresowanie w branży oprogramowania, tak samo pojawił się rozwój przetwarzania bezserwerowego – być może dokładniej określanego jako funkcja jako usługa (FaaS). Amazon Lambda to przykład frameworka FaaS, w którym pozwala na uruchamianie kodu bez udostępniania serwerów lub zarządzania nimi, oraz płacisz tylko za wykorzystany czas obliczeniowy. FaaS umożliwia tworzenie mikrousług w taki sposób, aby kod mógł zostać opakowany w lekki framework wbudowany w kontener, wykonywany na żądanie w oparciu o jakiś wyzwalacz, a następnie automatycznie równoważony obciążeniem, dzięki wspomnianemu wcześniej light frameworkowi. Główną zaletą FaaS jest to, że pozwala deweloperowi skupić się prawie wyłącznie na samej funkcji, co sprawia, że FaaS jest logicznym wnioskiem z podejścia mikrousługowego. Zdarzenie wyzwalające jest krytycznym elementem FaaS. Bez tego nie ma możliwości wywoływania funkcji (i zużywania zasobów) na żądanie. Ta możliwość automatycznego żądania funkcji w razie potrzeby sprawia, że FaaS jest naprawdę cenny. Wyobraź sobie przez chwilę, że ktoś czytający profil użytkownika uruchamia zdarzenie audytu, funkcję, która musi zostać uruchomiona, aby powiadomić zespół ds. bezpieczeństwa. Mówiąc dokładniej, może odfiltrowuje tylko niektóre typy rekordów, które mają być oznaczone jako monitujące o wyzwalacz. Innymi słowy, może być selektywna, co podkreśla fakt, że jako funkcja biznesowa można ją w pełni dostosować. (Zauważę, że wprowadzenie takiego przepływu pracy jest niezwykle proste w przypadku modelu wdrażania, takiego jak FaaS). Magia stojąca za usługą wyzwalania to tak naprawdę nic innego jak praca ze zdarzeniami w strumieniu zdarzeń. Niektóre typy zdarzeń są używane jako wyzwalacze częściej niż inne, ale dowolne zdarzenie można przekształcić w wyzwalacz. Zdarzeniem może być aktualizacja dokumentu lub uruchomienie procesu OCR na nowym dokumencie, a następnie dodanie tekstu z procesu OCR do bazy danych NoSQL. Możliwości tutaj są nieograniczone. FaaS jest również doskonałym obszarem do kreatywnego wykorzystania uczenia maszynowego – być może uczenia maszynowego jako usługi, a dokładniej „funkcji uczenia maszynowego aaS". Weź pod uwagę, że za każdym razem, gdy obraz jest przesyłany, może być uruchamiany przez platformę uczenia maszynowego w celu identyfikacji obrazu i oceniania. Nie ma tutaj podstawowych ograniczeń. Zdarzenie wyzwalające jest zdefiniowane, coś się dzieje, zdarzenie wyzwala funkcję, a funkcja wykonuje swoje zadanie. FaaS jest już ważną częścią przyjęcia mikrousług, ale musisz wziąć pod uwagę jeden główny czynnik, gdy zbliżasz się do FaaS: uzależnienie od dostawcy. Ideą FaaS jest to, że został zaprojektowany w celu ukrycia określonych mechanizmów przechowywania, określonej infrastruktury sprzętowej i orkiestracji komponentów oprogramowania – wszystkie wspaniałe funkcje, jeśli jesteś programistą. Ale z powodu tej abstrakcji oferta hostowanego FaaS jest jedną z największych możliwości zablokowania dostawców, jakie kiedykolwiek widziała branża oprogramowania. Ponieważ interfejsy API nie są ustandaryzowane, migracja z jednej oferty FaaS w chmurze publicznej do innej jest trudna bez odrzucenia znacznej części wykonanej pracy. Jeśli do FaaS podchodzimy w sposób bardziej metodyczny – na przykład wykorzystując zdarzenia z platformy konwergentnych danych – łatwiej jest przemieszczać się między dostawcami chmury.

Tworzenie nowoczesnej architektury danych

W wielu większych firmach funkcja IT zwykle zajmuje się definiowaniem i budowaniem systemów danych, zwłaszcza danych generowanych przez wewnętrzne systemy informatyczne. Często jednak zdarza się, że dane pochodzące ze źródeł zewnętrznych – klientów, produktów czy dostawców – są przechowywane i zarządzane oddzielnie przez odpowiedzialne jednostki biznesowe. W takim przypadku stajesz przed wyzwaniem upewnienia się, że wszystkie mają wspólne podejście do architektury danych, takie, które umożliwia połączenie wszystkich tych różnych typów danych i potrzeb użytkowników za pomocą wydajnego i umożliwiającego potok danych. Ten potok danych ma na celu zapewnienie pełnego przepływu danych, w którym stosowane zasady zarządzania danymi i nadzoru

koncentrują się na równowadze między wydajnością użytkownika a zapewnieniem zgodności z odpowiednimi przepisami i regulacjami. W mniejszych firmach lub nowoczesnych przedsiębiorstwach opartych na danych funkcja IT jest zwykle wysoce zintegrowana z różnymi funkcjami biznesowymi, co obejmuje ścisłą współpracę z inżynierami danych w jednostkach biznesowych w celu zminimalizowania luki między IT a funkcjami biznesowymi. Takie podejście okazało się bardzo skuteczne. Tak więc, gdy zdecydujesz, która funkcja zostanie skonfigurowana i steruje częścią architektury danych, nadszedł czas, aby zacząć. Korzystając z przewodnika krok po kroku znajdującego się na tej liście, błyskawicznie będziesz w drodze:

1. Zidentyfikuj swoje przypadki użycia, a także niezbędne dane dla tych przypadków użycia. Pierwszym krokiem, jaki należy wykonać, rozpoczynając tworzenie architektury danych, jest współpraca z użytkownikami biznesowymi w celu zidentyfikowania przypadków użycia i typu danych, które są w danym momencie najodpowiedniejsze lub po prostu mają najwyższy priorytet. Pamiętaj, że celem dobrej architektury danych jest połączenie biznesowej i technologicznej strony firmy, aby upewnić się, że pracują one na rzecz wspólnego celu. Aby znaleźć najcenniejsze dane dla Twojej firmy, powinieneś poszukać danych, które mogą generować spostrzeżenia o dużym wpływie biznesowym. Dane te mogą znajdować się w środowiskach danych korporacyjnych i mogły tam być już od jakiegoś czasu, ale być może środki i technologie umożliwiające odkrywanie takich danych i wyciąganie z nich wniosków były zbyt drogie lub niewystarczające. Dostępność dzisiejszych technologii open source i ofert w chmurze umożliwia przedsiębiorstwom wyciąganie takich danych i pracę z nimi w bardzo krótkim czasie bardziej opłacalny i uproszczony sposób.

2. Skonfiguruj zarządzanie danymi.

Niezwykle ważne jest, aby działania związane z zarządzaniem danymi stały się priorytetem. Proces identyfikacji i pozyskiwania danych oraz budowania modeli dla Twoich danych musi zapewnić jakość i trafność z perspektywy biznesowej, jest ważny i powinien również obejmować wydajne mechanizmy kontrolne w ramach wsparcia systemu. Należy również ustalić odpowiedzialność za dane, niezależnie od tego, czy dotyczy to indywidualnych właścicieli danych, czy różnych funkcji związanych z nauką o danych.

3. Buduj dla elastyczności.

Zasadą jest to, że powinieneś budować systemy danych zaprojektowane z myślą o zmianach, a nie takie, które mają trwać. Kluczową zasadą dla każdej architektury danych w dzisiejszych czasach jest niezależność od konkretnej technologii lub rozwiązania. Jeśli na rynku pojawi się nowe kluczowe rozwiązanie lub technologia, architektura powinna być w stanie je dostosować. Rodzaje danych napływających do przedsiębiorstw mogą się zmieniać, podobnie jak narzędzia i platformy, które są wykorzystywane do ich obsługi. Kluczem jest zatem zaprojektowanie środowiska danych, które może pomieścić taką zmianę.

4. Zdecyduj się na techniki przechwytywania danych.

Musisz wziąć pod uwagę swoje techniki pozyskiwania danych, a zwłaszcza upewnić się, że Twoja architektura danych może w pewnym momencie obsłużyć przesyłanie danych w czasie rzeczywistym, nawet jeśli nie jest to bezwzględne wymaganie od samego początku. Należy zbudować nowoczesną architekturę danych, aby wspierać przesyłanie i analizę danych do decydentów, kiedy i gdzie jest to potrzebne. Skoncentruj się na przesyłaniu danych w czasie rzeczywistym z dwóch perspektyw: potrzeby ułatwienia dostępu do danych w czasie rzeczywistym (dane, które mogą być historyczne) oraz wymogu obsługi danych ze zdarzeń w miarę ich występowania. W przypadku pierwszej kategorii istniejąca infrastruktura, taka jak hurtownie danych, ma do odegrania kluczową rolę. Po drugie,

kluczowe znaczenie mają nowe podejścia, takie jak analiza strumieniowa i uczenie maszynowe. Dane mogą pochodzić z dowolnego miejsca – z aplikacji transakcyjnych, urządzeń i czujników na różnych podłączonych urządzeniach, urządzeniach mobilnych i sprzęcie telekomunikacyjnym oraz „kto wie gdzie jeszcze". Nowoczesna architektura danych musi obsługiwać przesyłanie danych przy wszystkich prędkościach, niezależnie od tego, czy są to prędkości poniżej sekundy, czy z 24-godzinnym opóźnieniem.

5. Stosuj odpowiednie środki bezpieczeństwa danych.

Nie zapomnij wbudować bezpieczeństwa w architekturę. Nowoczesna architektura danych rozpoznaje, że zagrożenia dla bezpieczeństwa danych pojawiają się nieustannie, zarówno zewnętrznie, jak i wewnętrznie. Zagrożenia te nieustannie ewoluują i mogą pojawiać się za pośrednictwem poczty e-mail w jednym miesiącu, a dysków flash w następnym. Menedżerowie danych i architekci danych mają zwykle największą wiedzę, jeśli chodzi o zrozumienie, co jest wymagane w celu zapewnienia bezpieczeństwa danych w dzisiejszych środowiskach, dlatego warto wykorzystać ich wiedzę specjalistyczną.

6. Zintegruj zarządzanie danymi podstawowymi.

Upewnij się, że zajmujesz się zarządzaniem danymi podstawowymi, metodą stosowaną do definiowania krytycznych danych organizacji i zarządzania nimi, aby za pomocą integracji danych zapewnić jeden punkt odniesienia. Dzięki uzgodnionej i wbudowanej strategii zarządzania danymi podstawowymi (MDM) Twoje przedsiębiorstwo może mieć jedną wersję prawdy, która synchronizuje dane z aplikacjami uzyskującymi do nich dostęp. Potrzeba architektury opartej na MDM ma kluczowe znaczenie, ponieważ organizacje nieustannie przechodzą zmiany, w tym wzrost, dostosowania, fuzje i przejęcia. Często przedsiębiorstwa kończą z systemami danych działającymi równolegle, a często krytyczne rekordy i informacje mogą być duplikowane i nakładać się na siebie w tych silosach. MDM zapewnia, że aplikacje i systemy w całym przedsiębiorstwie mają taki sam widok ważnych danych.

7. Oferuj dane jako usługę (aaS).

Ten konkretny krok jest stosunkowo nowym podejściem, ale okazał się całkiem udanym komponentem - upewnij się, że Twoja architektura danych jest w stanie pozycjonować dane jako usługę (aaS). Wiele przedsiębiorstw posiada szereg baz danych i starszych środowisk, co utrudnia pobieranie informacji z różnych źródeł. W podejściu aaS dostęp jest możliwy poprzez warstwę zwirtualizowanych usług danych, która standaryzuje wszystkie źródła danych, niezależnie od urządzenia, aplikatora czy systemu. Dane jako usługa są z definicji formą wewnętrznej usługi w chmurze firmy, w której dane – wraz z różnymi platformami, narzędziami i aplikacjami do zarządzania danymi – są udostępniane przedsiębiorstwu jako standardowe usługi wielokrotnego użytku. Potencjalną zaletą danych jako usługi jest to, że procesy i zasoby mogą być wstępnie pakowane w oparciu o standardy korporacyjne lub zgodności i łatwo dostępne w chmurze korporacyjnej.

8. Włącz funkcje samoobsługi.

Jako ostatni krok w budowaniu architektury danych, zdecydowanie powinieneś zainwestować w środowiska samoobsługowe. Dzięki samoobsłudze użytkownicy biznesowi mogą konfigurować własne zapytania i uzyskiwać żądane dane lub analizy, lub mogą przeprowadzać własne wyszukiwanie danych bez konieczności oczekiwania na dostarczenie danych przez działy IT lub działy zarządzania danymi. Droga do samoobsługi polega na zapewnieniu interfejsów front-end, które są proste i łatwe w użyciu dla celu

publiczność. W trakcie tego procesu można opracować logiczną warstwę usług, którą można ponownie wykorzystać w różnych projektach, działach i jednostkach biznesowych. IT może nadal odgrywać ważną rolę w architekturze samoobsługowej, w tym w takich aspektach, jak operacje potoku danych (sprzęt, oprogramowanie i chmura) oraz mechanizmy kontroli zarządzania danymi, ale musiałoby wydawać coraz mniej swoich czas i zasoby na spełnienie żądań użytkowników, które mogą być lepiej sformułowane i zaadresowane przez samego użytkownika.

Skupienie zarządzania danymi na właściwych aspektach

Teraz bardziej niż kiedykolwiek umiejętność obsługi ogromnych ilości danych w sposób, który pozwala zrównoważyć ryzyko i możliwości biznesowe, ma kluczowe znaczenie dla sukcesu firmy. Jednak nawet po pojawieniu się funkcji zarządzania danymi i dyrektorów ds. danych (CDO) większość firm nie radzi sobie najlepiej, jeśli chodzi o zarządzanie danymi i zarabianie na nich. Badania międzybranżowe pokazują, że średnio mniej niż połowa ustrukturyzowanych danych organizacji jest aktywnie wykorzystywana do podejmowania decyzji, a mniej niż 1 procent nieustrukturyzowanych danych jest analizowanych lub w ogóle używanych. Wiele razy firmy blokują dane, aby być po bezpiecznej stronie, a pracownicy są zmuszeni spędzać dużo czasu na wyjaśnianiu, dlaczego potrzebują określonego zestawu danych. Jednocześnie coraz częściej dochodzi do naruszeń danych, ponieważ zbiory danych są rozproszone w silosach, a ich wartość często nie jest odpowiednio rozumiana lub traktowana. Ponadto częstym problemem jest to, że infrastruktura danych i aplikacje firmy nie spełniają pokładanych w nich oczekiwań. W tym rozdziale przeprowadzę Cię przez kluczowe elementy zarządzania danymi i pokażę Ci najlepszy sposób podejścia do tematu.

Porządkowanie zarządzania danymi

Pojęcie zarządzania danymi odnosi się do ludzi, procesów i wsparcia systemowego wymaganego do ustanowienia spójnego i prawidłowego postępowania z danymi organizacji w całej firmie. Wspiera zarządzanie danymi z niezbędną podstawą, strategią i strukturą niezbędną do zapewnienia, że dane są zarządzane jako zasób i przekształcane w sensowne i praktyczne spostrzeżenia. Kluczowe obszary zarządzania danymi obejmują zapewnienie dostępności, użyteczności, spójności, integralności i bezpieczeństwa danych przez cały czas, a także ustanowienie procesów zapewniających skuteczne zarządzanie danymi w całej organizacji. Różne obszary zarządzane przez zarządzanie danymi można wyjaśnić zgodnie z poniższą listą:

* Dostępność danych: ten termin jest używany do opisania, w jakim stopniu element danych może być łatwo dostępny na dowolnym poziomie wydajności. Poziom dostępności danych można mierzyć za pomocą takich czynników, jak łatwość zarządzania danymi i ich utrzymywania, możliwość przywrócenia lub odzyskania dowolnych usług lub danych w przypadku wystąpienia błędu lub awarii, możliwość dostarczenia usługi oraz zdolność do zrozumieć problemy z danymi, zdiagnozować ich podstawową przyczynę i jak najszybciej je naprawić.

* Użyteczność danych: Odnosi się to do stanu danych, które obecnie posiadasz (w ich surowym formacie) i tego, jak spełniają one swoje przeznaczenie. Skąd wiesz, czy Twoje dane nadają się do użytku? Oto kilka pytań, które możesz zadać: Czy wartość surowych danych jest poprawna czy niepoprawna? Jak szczegółowe i precyzyjne są atrybuty danych? Jak zintegrowane są dane z innymi źródłami danych i obiektami danych?

* Spójność danych: ten termin jest używany do opisania, jak przydatne i wiarygodne są dane z punktu widzenia wiarygodności. Spójność jest zwykle sprawdzana z trzech głównych perspektyw: spójność punktu w czasie (że dane pozostają spójne w czasie), spójność transakcji (że dane pozostają spójne podczas transakcji) oraz spójność aplikacji (że dane pozostają spójne między różnymi aplikacjami).

* Integralność danych: Odnosi się to do utrzymania i zapewnienia dokładności i spójności danych w całym cyklu ich życia. Jest to krytyczny aspekt projektowania, wdrażania i użytkowania dowolnego systemu, który przechowuje, przetwarza lub pobiera dane. (Integralność danych jest przeciwieństwem uszkodzenia danych).

* Bezpieczeństwo danych: Odnosi się to do ochrony danych cyfrowych, takich jak te znajdujące się w bazie danych, przed siłami destrukcyjnymi i niepożądanymi działaniami nieautoryzowanych użytkowników – użytkowników, którzy mogą np. zainicjować cyberatak lub naruszenie danych.

Na początek zrozum i poproś firmę o porozumienie w sprawie tego, czego oczekuje się od działań i ram zarządzania danymi. Czy jest to coś, w co angażujesz się tylko po to, aby upewnić się, że trzymasz się linii, jeśli chodzi o przepisy i regulacje, czy też ambicją jest wykorzystanie zarządzania danymi jako czynnika umożliwiającego większy, znacznie bardziej niezawodny biznes? Punkt wyjścia jest ważny, ponieważ pomoże Ci określić priorytety i podejście podczas ustanawiania i wdrażania zarządzania danymi.

Zarządzanie danymi na potrzeby obrony lub przestępstwa

Jeśli chodzi o zarządzanie danymi, możesz grać w obronie lub ofensywie. Ochrona danych polega na minimalizowaniu ryzyka i obejmuje takie aspekty, jak zapewnienie zgodności z przepisami, wykorzystywanie analiz lub uczenia maszynowego do wykrywania i ograniczania oszustw oraz budowanie systemów zapobiegających kradzieży. Wysiłki obronne obejmują również środki zaprojektowane w celu zapewnienia integralności danych przepływających przez wewnętrzne systemy firmy. Odbywa się to poprzez identyfikację, standaryzację i zarządzanie głównymi źródłami danych (na przykład dane dotyczące klientów, produktów lub sprzedaży), aby firma mogła polegać na jednym źródle prawdy w przypadku najważniejszych danych. Z drugiej strony, przestępstwo danych koncentruje się na wspieraniu celów biznesowych, takich jak zwiększenie przychodów, rentowności i zadowolenia klientów. Zazwyczaj obejmuje działania, takie jak analiza i modelowanie danych, które mają na celu generowanie informacji o klientach w celu wsparcia procesu podejmowania decyzji zarządczych. Przestępstwo dotyczące danych obejmuje jednak również działania związane z realizacją możliwości biznesowych z produktami lub usługami związanymi z danymi opartymi na badaniach lub pracach rozwojowych. Każda firma potrzebuje zarówno ataku, jak i obrony, aby odnieść sukces, ale uzyskanie właściwej równowagi jest skomplikowane i może się znacznie różnić między różnymi typami firm. Te dwa podejścia zwykle konkurują o ograniczone zasoby, fundusze i ludzi. Należy jednak pamiętać, że choć położenie równego nacisku na te dwa jest optymalne dla niektórych firm, dla innych może być dużo rozsądniej faworyzować jedną lub drugą. Duży wpływ na to ma kontekst firmy związany z branżą, której jest częścią i jak konkurencyjne jest otoczenie, a także ograniczenia regulacyjne. Na przykład szpitale istnieją na mało konkurencyjnym rynku o silnie uregulowanym kontekście, gdzie wymagania dotyczące jakości danych i ochrony prywatności są niezwykle wysokie. Muszą zatem nadać priorytet obronie danych przed przestępstwami. Z drugiej strony firmy z branży detalicznej są znacznie mniej regulowane i dlatego mogą pracować z wrażliwymi danymi osobowymi w ramach strategii pokonania konkurencji i szybkiego reagowania na zmiany rynkowe. Tego typu firmy zazwyczaj przedkładają atak nad obronę.

Cele zarządzania danymi

Ty (a może Twoi szefowie) możesz zapytać, dlaczego zarządzanie danymi jest tak ważne. Od razu mogę powiedzieć, że nieefektywne zarządzanie danymi w firmie nieuchronnie prowadzi do jednego: kiepskich danych. Te słabe dane są widoczne w niespójnych definicjach, duplikatach, brakujących polach i innych klasycznych problemach z danymi. Są to jasne kwestie, których należy unikać. Cele dotyczące zarządzania danymi można zdefiniować na wszystkich poziomach firmy, ale zachęcając interesariuszy do udziału w ustalaniu celów, możesz upewnić się, że uznają oni znaczenie zarządzania danymi. Na poniższej liście znajdziesz kilka przydatnych przykładów celów zarządzania danymi, podzielonych na obozy defensywne i ofensywne:

<u>Obronny:</u>

* Zwiększenie spójności i pewności w podejmowaniu decyzji

* Zmniejszenie ryzyka grzywien regulacyjnych

* Poprawa bezpieczeństwa danych oraz zdefiniowanie i weryfikacja wymagań dla polityk dystrybucji danych

* Wyznaczanie odpowiedzialności za jakość danych

* Umożliwienie lepszego planowania przez personel nadzoru

* Zmniejszenie tarcia operacyjnego

* Ochrona potrzeb interesariuszy danych

* Szkolenie kierownictwa i personelu w celu przyjęcia wspólnego podejścia do kwestii danych

Ofensywy:

* Maksymalizacja potencjału generowania dochodu przez komercyjne produkty danych

* Zachęcanie do wysokiego stopnia udostępniania i ponownego wykorzystywania danych i spostrzeżeń

* Wykorzystywanie spostrzeżeń uzyskanych na podstawie danych do podejmowania decyzji dotyczących inwestycji biznesowych

* Widząc, że działania badawczo-rozwojowe są napędzane danymi, analizami i uczeniem maszynowym/sztuczną inteligencją i koncentrują się na odkrywaniu nowych możliwości biznesowych

* Optymalizacja efektywności pracowników

* Minimalizacja lub eliminacja „przeróbek”

* Ustalenie podstaw wydajności procesu w celu umożliwienia działań doskonalących

Każdy z tych celów można zrealizować, wdrażając programy zarządzania danymi lub uruchamiając inicjatywy wykorzystujące techniki zarządzania zmianą.

Wyjaśnienie, dlaczego potrzebne jest zarządzanie danymi

Wyobraź sobie, że przesyłasz budżet mający na celu usprawnienie praktyk zarządzania danymi. Słyszysz, że nie zostanie przyznany ani grosz, chyba że możesz przekonać wyższe kierownictwo, że to dobra inwestycja. Jakie są najlepsze argumenty, aby ocalić swoje inicjatywy i budżet? Czytaj dalej, aby się dowiedzieć.

Zarządzanie danymi oszczędza pieniądze

Przede wszystkim zarządzanie danymi zwiększa efektywność w organizacji. Zduplikowane konta prowadzą do zduplikowanych wysiłków lub przynajmniej do marnowania czasu na śledzenie duplikatów kont w działaniach marketingowych, sprzedażowych, finansowych, programistycznych lub analitycznych. Zarządzanie danymi redukuje błędy w danych źródłowych, dając Twojej firmie solidną podstawę do pracy i oszczędza cenny czas, który w przeciwnym razie zostałby wykorzystany na poprawianie istniejących danych. Zaoszczędzony czas to zaoszczędzone pieniądze. Zarządzanie danymi zmusza Twoją firmę do zdefiniowania swoich danych podstawowych, a także zasad rządzących tymi danymi podstawowymi. Rozpoczęcie projektu zarządzania danymi może być doskonałą okazją, aby wszyscy byli na tej samej stronie na temat podstawowych definicji danych. Egzekwowanie tych definicji zapewnia z czasem większą wydajność operacyjną.

Złe zarządzanie danymi jest niebezpieczne

Brak skutecznego zarządzania danymi jest problemem bezpieczeństwa z dwóch powodów:

* Z brudnymi, nieustrukturyzowanymi danymi wiążą się zewnętrzne zagrożenia bezpieczeństwa.

* Złe zarządzanie danymi może spowodować problemy ze zgodnością z przepisami.

Źle ustrukturyzowane dane stanowią zagrożenie bezpieczeństwa z prostego powodu, że jeśli masz brudne, nieustrukturyzowane dane, które zatykają potok danych, nie możesz szybko stwierdzić, kiedy coś ma pójść nie tak, i nie możesz skutecznie monitorować, które dane są zagrożone. Zgodność z przepisami i zarządzanie danymi stają się z każdym dniem coraz gorętszym tematem. Ponieważ ludzie stają się coraz bardziej świadomi znaczenia ich danych osobowych, rządy zaczynają bardzo poważnie traktować sposób, w jaki firmy przechowują, chronią i wykorzystują dane swoich klientów. W przypadku bałaganu, niekontrolowanego bagna danych może okazać się niemożliwe, aby firma mogła zagwarantować, że wszystkie dane dotyczące konkretnej osoby zostaną usunięte na żądanie. To naraża Twoją firmę na duże ryzyko i potencjalnie ogromne kary.

Dobre zarządzanie danymi zapewnia przejrzystość

Skuteczne zarządzanie danymi zapewnia spokój ducha, że dane Twojej firmy są ogólnie czyste, ustandaryzowane i dokładne. Efekty tego zapewnienia odbijają się echem w całej firmie i zapewniają ważne korzyści. Oczywistą, ale ważną korzyścią jest zapewnienie, że integralność danych jest zachowana przez cały cykl ich życia, co oznacza, że dane są godne zaufania i wykorzystywane jako podstawa ważnych decyzji biznesowych, a także do badań i rozwoju nowych spostrzeżeń lub produktów i usług związanych z danymi.

Ustanowienie zarządzania danymi w celu egzekwowania zasad zarządzania danymi

W organizacji administrator danych jest odpowiedzialny za wykorzystanie procesów zarządzania danymi organizacji w celu zapewnienia przydatności różnych elementów danych. W związku z tym administratorzy danych pełnią specjalistyczną rolę, która obejmuje procesy, zasady, wytyczne i obowiązki związane z administrowaniem całym zakresem danych firmy zgodnie z wymogami polityki i/lub przepisów. Zarządzanie danymi dotyczy dbania o zasoby danych, które nie należą do samych zarządców i mogą reprezentować potrzeby całej organizacji. Inni mogą mieć za zadanie reprezentowanie mniejszego zakresu danych związanego z konkretną jednostką biznesową lub działem, a nawet z określonym typem danych. W niektórych organizacjach administratorzy danych to starsi przedstawiciele wyznaczonych grup interesariuszy – struktury zaprojektowanej w celu zapewnienia wystarczającego zaangażowania i podejmowania decyzji dotyczących traktowania pewnych zasobów danych. Jednak w innych organizacjach administratorzy danych działają niezależnie, zapewniając, że ogólne zasady i kontrole są odpowiednio stosowane do danych w całej organizacji. Ogólnym celem administratora danych jest zapewnienie jakości danych dla głównych elementów danych, co do których zdecydowano. Obejmuje to przechwytywanie metadanych dla każdego elementu danych, takich jak definicje, powiązane reguły/zarządzanie, manifestacje fizyczne i powiązane modele danych. Zarządcy danych rozpoczynają proces zarządzania od identyfikacji elementów, którymi będą zarządzać, z ostatecznym wynikiem standardów, kontroli i wprowadzania danych. Administrator danych musi ściśle współpracować z interesariuszami zaangażowanymi w standaryzację danych w celu dostosowania do standardów danych; z architektami danych w celu zrozumienia i zapewnienia przestrzegania zależności danych; i z ekspertów wspierających system w celu zabezpieczenia zautomatyzowanych i wbudowanych mechanizmów kontrolnych dla kontroli

jakości danych. Kontrole danych mogą mieć charakter prewencyjny, detektywistyczny i korygujący i mogą być wykonywane ręcznie, wspomagane technologią lub całkowicie zautomatyzowane.

Wdrażanie ustrukturyzowanego podejścia do zarządzania danymi

Wszystkie organizacje muszą być w stanie podejmować decyzje dotyczące zarządzania danymi, czerpać z nich wartość, minimalizować koszty i złożoność, zarządzać ryzykiem i zapewniać zgodność z ciągle rosnącymi wymaganiami prawnymi, regulacyjnymi i innymi. Firma musi stworzyć reguły, zapewnić ich przestrzeganie i być w stanie poradzić sobie z niezgodnością, niejasnościami i wszelkimi problemami związanymi z danymi, które mogą się pojawić. Krótko mówiąc, firma musi robić więcej niż tylko zarządzać danymi. Potrzebny jest system zarządzania, który ustala zasady zaangażowania w działania zarządcze w całej organizacji. Małe organizacje lub te z prostymi środowiskami danych mogą odnieść sukces w realizacji tych celów dzięki nieformalnemu systemowi zarządzania. Mogą nawet nie zdawać sobie sprawy, kiedy przechodzą między podejmowaniem decyzji zarządczych a szerszymi decyzjami zarządczymi. Z drugiej strony, większe organizacje lub te z bardziej złożonymi środowiskami danych lub zgodności na ogół stwierdzają, że muszą wycofać się i uzgodnić bardziej formalny system zarządzania. Definiowanie i ustanawianie ram lub ustrukturyzowanego podejścia do zarządzania danymi obejmuje działania takie jak:

* Określ swoje cele. Zadaj sobie pytanie, czy dla Twojej branży, sytuacji rynkowej i regulacyjnej najważniejsze jest podejście defensywne lub ofensywne.

* Wybierz obszar ostrości. Od czegoś trzeba zacząć, więc zadaj sobie pytanie, od czego zaczyna się Twój projekt data governance - pełny zakres w całej firmie czy tylko dla wybranej jednostki lub działu?

* Ustaw definicje danych i zasady. Czym dokładnie musisz rządzić? Unikaj zbyt szerokiego zarzucania sieci, ponieważ może to utrudnić innowacyjność i wydajność.

* Określ uprawnienia decyzyjne. Kto będzie mógł decydować o zasadach zarządzania? Jakie ramy zarządzania danymi są potrzebne Twojej firmie?

* Zdefiniuj i zaimplementuj mechanizmy kontrolne. Jak zapewnisz przestrzeganie zasad? Unikaj ręcznej (ludzkiej) kontroli tak bardzo, jak to możliwe, wbudowując mechanizmy kontroli do swoich systemów danych o jak najwyższym poziomie automatyzacji.

* Zidentyfikuj interesariuszy danych. Określ, kto i w jaki sposób będzie korzystać z danych. Poświęć trochę czasu, aby naprawdę zrozumieć potrzeby biznesowe i interesariuszy wybranego obszaru tematycznego.

* Skonfiguruj kartę zarządzania danymi (DGB). Zarządcy danych tworzą DGB. Są oni ostatecznie odpowiedzialni za wykorzystanie danych biznesowych, jakość danych i ustalanie priorytetów w kwestiach związanych z danymi. Podejmują decyzje, które mają wpływ na dane na podstawie zaleceń od zarządców danych. Rada jest uprawniona do decydowania o sposobie wydatkowania budżetu na usprawnienia zarządzania danymi związane z zarządzaniem danymi. Ten krok ma zastosowanie głównie w przypadku większych firm lub większych wdrożeń, gdy złożoność i zależności są znaczne.

* Przypisywanie i szkolenie stewardów danych. Kto będzie na co dzień zajmował się zarządzaniem danymi? W jaki sposób uchwycisz potrzebę w całej organizacji i upewnisz się, że struktura pozostaje odpowiednia w czasie? Ważnym krokiem jest ustanowienie roli administratorów danych w zarządzaniu danymi zgodnie z Twoją branżą.

* Projektowanie i wdrażanie potrzebnych procesów. Ostatnim krokiem jest upewnienie się, że masz procesy, które są wystarczające, a jednocześnie jasne i proste, aby Twoja firma mogła realizować zadowalające zarządzanie danymi.

Jednym z najważniejszych czynników, jeśli chodzi o zarządzanie danymi, jest zapewnienie wspólnego, uzgodnionego zestawu zasad i najlepszych praktyk wspólnych dla wszystkich zespołów i osób odpowiedzialnych za gromadzenie, zarządzanie i wykorzystywanie danych. Upewnij się, że wszyscy są na pokładzie i że istnieją jasne cele, jasno zdefiniowane procesy i jasne poziomy uprawnień, aby wszystko przebiegało sprawnie. Kluczem do zarządzania danymi jest efektywna współpraca. Właściwe narzędzia zarządzania danymi powinny iść w parze z tymi zasadami. Upewnij się, że narzędzia, które chcesz wdrożyć, są łatwe w użyciu zarówno dla użytkowników biznesowych, jak i IT, umożliwiają bezproblemową współpracę między zespołami i są wystarczająco elastyczne, aby ewoluować wraz ze zmieniającymi się potrzebami biznesowymi.

Zarządzanie modelami podczas rozwoju i produkcji

Chociaż zarządzanie danymi jest niezbędne, aby odnieść sukces z inwestycją w naukę danych, równie ważne jest zrozumienie, dlaczego zarządzanie modelami jest kluczową częścią. W tym rozdziale pokrótce zbadam, na czym polega zarządzanie modelami, a także wymienię niektóre z ważnych aspektów, które należy wziąć pod uwagę przy opracowywaniu i wdrażaniu modeli.

Odkrywanie podstaw zarządzania modelami

Algorytm to metoda rozwiązywania problemu krok po kroku, powszechnie stosowana do przetwarzania danych, obliczeń i innych powiązanych operacji komputerowych i matematycznych. Algorytm jest również używany do manipulowania danymi na różne sposoby, np. wstawianie nowego elementu danych, wyszukiwanie określonego elementu lub sortowanie elementu. Technicznie rzecz biorąc, komputery używają algorytmów, aby wyświetlić szczegółowe instrukcje dotyczące wykonania zadania. Na przykład, aby obliczyć pensję pracownika, komputer używa algorytmu. Aby wykonać to zadanie, należy wprowadzić do systemu odpowiednie dane. Pod względem wydajności różne algorytmy są w stanie łatwo i szybko wykonywać operacje lub rozwiązywać problemy. Tak więc algorytm jest ogólnym podejściem, które przyjmiesz. Model jest tym, co otrzymujesz, gdy uruchomisz algorytm na swoich danych treningowych, a następnie użyjesz go do wykonania prognozy dotyczące nowych danych. Możesz wygenerować nowy model z tym samym algorytmem, ale z innymi danymi, lub możesz uzyskać nowy model z tych samych danych, ale z innym algorytmem.

Praca z wieloma modelami

Powszechnym nieporozumieniem dotyczącym modeli uczenia maszynowego w porównaniu z przestrzenią tworzenia oprogramowania jest przekonanie, że celem sesji modelowania uczenia maszynowego jest zbudowanie jednego udanego modelu, wdrożenie go, a następnie odejście, poklepywanie się po plecach w celu uzyskania dobrej pracy zrobione. To fantazja. W rzeczywistości praca z modelami uczenia maszynowego obejmuje pracę z wieloma modelami przez dłuższy czas, nawet po wdrożeniu modelu w środowisku produkcyjnym. Często zdarza się, że kilka modeli jest w produkcji w tym samym czasie, a nowe modele są gotowe do zastąpienia starszych modeli w produkcji, gdy zmienią się warunki. Ważne jest również, aby móc zarządzać tymi wymianami modeli w płynny sposób, bez zakłócania bieżącej usługi. Podczas opracowywania modelu pracujesz również z więcej niż jednym modelem, eksperymentując z wieloma narzędziami i porównując wydajność modelu, aby znaleźć model o najlepszej wydajności. Takie jest ogólne ukształtowanie terenu, ale nie do końca wyjaśnia, na czym tak naprawdę polega zarządzanie modelami. Aby lepiej sobie z tym poradzić, rozważ sytuację, w której organizacja ma setki modeli osadzonych w różnych systemach produkcyjnych, aby wspierać podejmowanie decyzji w zakresie marketingu, wyceny, ryzyka kredytowego, ryzyka operacyjnego, oszustw i finansów. W tym przykładzie, który jest powszechny w branży oprogramowania, analitycy danych w różnych jednostkach biznesowych mogą swobodnie opracowywać swoje modele bez sformalizowanych lub ustandaryzowanych procesów przechowywania, wdrażania i zarządzania nimi. Niektóre modele nie posiadają niezbędnej dokumentacji opisującej właściciela modelu, cel biznesowy, wytyczne użytkowania lub inne informacje niezbędne do zarządzania modelem lub wyjaśnienia go regulatorom, ponieważ jednostkom powiedziano, że mają priorytet nad szybkością nad właściwą dokumentacją. Co więcej, po osiągnięciu wyników modelu w tym wyimaginowanym scenariuszu, podlegają one ograniczonym kontrolom i wymaganiom, gdy trafiają do decydentów. Nic dziwnego, ponieważ do tworzenia modeli wykorzystano różne zbiory danych i zmienne, wyniki okazują się niespójne. Niewiele jest walidacji lub testów wstecznych pod kątem dokładności, a decyzje są podejmowane na podstawie wyników modelu w takim stanie, w jakim są – a wtedy wszyscy po prostu mają nadzieję na najlepsze. Opisany właśnie

scenariusz, z całkowitym zamieszaniem w modelowaniu, może wydawać się zbyt znajomy wielu organizacjom. W zróżnicowanym i luźno zarządzanym środowisku modelowania może być dość trudno odpowiedzieć na krytyczne pytania dotyczące predykcyjnych modeli analitycznych lub modeli uczenia maszynowego, na których opiera się Twoja organizacja nie tylko w codziennych operacjach biznesowych, ale także w podejmowaniu strategicznych decyzji. Po prostu ważne jest, aby Twoja organizacja zbudowała solidną podstawę do zarządzania modelami w sposób niezawodny, przejrzysty, skalowalny i wielokrotnego użytku. Ale jak to zrobić? Dobrym punktem wyjścia przy przeglądaniu bieżącej sytuacji zarządzania modelami lub próbie ustalenia, jak ją rozumie zespół architektoniczny, jest zadanie następujących pytań:

* Czy śledzimy, kto stworzył modele i dlaczego?

* Czy wiemy, które zmienne wejściowe są używane do prognozowania lub, w przypadku uczenia maszynowego, które dane szkoleniowe zostały użyte do trenowania modelu?

* Czy śledzimy, w jaki sposób wykorzystywane są modele?

* Czy mierzymy wydajność modeli i czy wiemy, kiedy te modele były ostatnio aktualizowane?

* Czy istnieje wystarczająca ilość dokumentacji pomocniczej, aby umożliwić ponowne wykorzystanie modelu przez inne zespoły zajmujące się analizą danych?

* Czy wprowadzenie do produkcji nowych lub zaktualizowanych modeli zajmuje dużo czasu?

W obliczu tej listy pytań firmy zwykle odpowiadają na dwa sposoby: albo są w stanie odpowiedzieć pozytywnie na te pytania, ale dochodzą do wniosku, że można zrobić więcej, aby zwiększyć wydajność i wartość z modeli, albo: nie jestem w stanie odpowiedzieć twierdząco na żadne z tych pytań. Dlaczego to drugie? Ponieważ nie zdali sobie sprawy, jak ważne jest dobre zarządzanie modelami w firmie opartej na danych i uczeniu maszynowym/sztucznej inteligencji. W przedsiębiorstwie opartym na danych i modelach modele są podstawą kluczowych decyzji biznesowych. Modele mogą identyfikować nowe możliwości, pomagać w nawiązywaniu nowych lub lepszych relacji z klientami oraz umożliwiać zarządzanie niepewnością i ryzykiem. Z tych i wielu innych powodów należy je tworzyć i traktować jako aktywa organizacyjne o wysokiej wartości. Zarządzanie modelami to nie tylko stosowanie nowych wytycznych lub nowej struktury zarządzania; musisz mieć pod ręką oprogramowanie, które może uporządkować dane w kształt i szybko stworzyć wiele dokładnych modeli, na których możesz polegać. Co więcej, zarządzanie modelami i śledzenie ich w celu uzyskania optymalnej wydajności przez cały cykl życia wymaga wydajnych i powtarzalnych procesów, które są w pełni wspierane przez niezawodną infrastrukturę i różne elementy architektoniczne.

Przekonanie do efektywnego zarządzania modelami

Pogląd, że efektywne zarządzanie modelami ma kluczowe znaczenie dla sukcesu organizacji, zyskuje coraz większe znaczenie, ponieważ znaczenie nauki o danych jest dostrzegane w szerszym znaczeniu. Niektórzy twierdzą nawet, że w przyszłości silniejszą przewagą konkurencyjną będzie kierowanie się modelem, w przeciwieństwie do kierowania się tylko danymi. Wygląda również na to, że piętrzą się dowody na to, że firmy, którym udaje się uzyskać wartość z inwestycji w naukę danych, to te, które traktują modele jako nowy rodzaj aktywów biznesowych. Firmy, które odnoszą dziś największe sukcesy w nauce o danych, traktują modele zupełnie inaczej niż traktują dane i oprogramowanie. Sposób, w jaki budują modele, opracowują modele, wdrażają modele, zarządzają modelami i sprawują nad nimi nadzór, a także jak tworzą systemy infrastruktury technologicznej w celu obsługi modeli, różnią się od tego, co robili w przeszłości podczas konfigurowania systemów dla dane lub oprogramowanie. Dlaczego? Po pierwsze, surowe dane, których naukowcy używają do tworzenia modeli, różnią się od

innych aktywów biznesowych, ponieważ modele wymagają algorytmów wymagających dużej mocy obliczeniowej. To wymaganie oczywiście napędza rosnące zapotrzebowanie na elastyczną, skalowalną moc obliczeniową, a także na wyspecjalizowany sprzęt, taki jak procesory graficzne (GPU). Są to komponenty architektoniczne, których zespoły inżynierów oprogramowania zwykle nie potrzebują. Innym krytycznym surowcem do tworzenia modeli jest ekosystem open source. Każdego dnia pojawiają się nowe narzędzia, nowe pakiety lub zaktualizowane pakiety, zwłaszcza dotyczące Pythona i R. Jeśli firma próbuje konkurować i mieć najlepsze, najbardziej innowacyjne modele, potrzebuje sposobów, aby zapewnić analitykom danych szybki dostęp do bardzo szybko ewoluujący ekosystem bez ograniczania ich elastyczności. Trzecią właściwością, która odróżnia modele od innych rodzajów tworzenia oprogramowania, jest proces. Modele ze swej natury wyłaniają się z procesu badawczego, a procesy te są z natury eksperymentalne, wyłaniające się i eksploracyjne. To zupełnie różni się od tego, jak działa tworzenie oprogramowania i jest zupełnie inne od sposobu, w jaki systemy pozyskują dane. Zespół zajmujący się analizą danych opracowujący modele może wypróbować setki pomysłów, zanim znajdzie taki, który działa; to jest w porządku, ale stwarza inne wymagania dotyczące podstawowej infrastruktury. Zespoły opracowujące modele potrzebują różnych możliwości, aby ułatwić szybkie eksperymentowanie i szybką eksplorację, aby mogły dokonywać przełomów. W oprogramowaniu chodzi o zmniejszenie ryzyka i dążenie do jasności wymagań. W tworzeniu modeli chodzi o szybkie eksperymentowanie, w którym wypróbowujesz jak najwięcej pomysłów tak szybko, jak to możliwe. Czwartą kluczową właściwością modeli, o której należy pamiętać, jest ich zachowanie. W inżynierii oprogramowania zazwyczaj istnieje specyfikacja, do której dążą programiści, oraz testy, które mogą potwierdzić, czy specyfikacja została spełniona. Nie ma nic takiego podczas budowania modeli predykcyjnych. Zamiast tego modele nauki o danych są probabilistyczne. Nie mają prawidłowej odpowiedzi. Po prostu mają lepsze lub gorsze odpowiedzi, kiedy żyją w prawdziwym świecie. Oznacza to, że organizacje potrzebują nowych sposobów kontroli jakości, monitorowania, zarządzania i przeglądu modeli, aby zapewnić niezawodność modelu i przewidywane zachowanie algorytmu, co oznacza, że działa on zgodnie z oczekiwaniami. Wyjątkowe wymagania, aby odnieść sukces w zarządzaniu modelami w nauce o danych, sugerują, że zarządzanie modelami należy traktować jako dyscyplinę dedykowaną. Infrastruktury danych nie powinny ograniczać zarządzania modelami do aplikacji, ale powinny obejmować znacznie więcej podejścia opartego na modelu, nie tylko opartego na danych.

Wdrażanie zarządzania modelami

Następnym przełomem w nauce danych prawdopodobnie nie będą nowe rewolucyjne algorytmy (te będą się pojawiać bez względu na wszystko), ale raczej zdolność do szybkiego łączenia, wdrażania i utrzymywania istniejących algorytmów w szybko zmieniających się środowiskach na żywo. Wiele korporacji uświadomiło sobie teraz potrzebę stworzenia scentralizowanego repozytorium do przechowywania modeli predykcyjnych wraz ze szczegółowymi metadanymi w celu wydajnej współpracy w grupie roboczej i kontroli wersji różnych modeli. Skuteczne zarządzanie modelami obejmuje współpracujący zespół modelarzy, architektów, menedżerów oceniających modele, audytorów modeli i testerów walidacyjnych. Jednak wiele firm zmaga się z procesem podpisywania etapów rozwoju, walidacji, wdrażania i zarządzania cyklem życia swoich modeli. Musisz od razu dokładnie wiedzieć, gdzie znajduje się każdy model w cyklu życia, ile ma lat, kto opracował model i kto używa modelu do jakich zastosowań. Możliwość kontrolowania wersji modelu w czasie to kolejna krytyczna potrzeba biznesowa, która obejmuje rejestrowanie zdarzeń i śledzenie zmian w celu zrozumienia, w jaki sposób forma i wykorzystanie modelu ewoluują w czasie. Degradacja modelu polegająca na tym, że model nie działa już zgodnie z oczekiwaniami, a dokładność modelu spada, to kolejne poważne wyzwanie, z którym boryka się wiele organizacji. Pilnie potrzebna jest standaryzacja metryk wykorzystywanych do pomiaru wydajności modelu. Obecnie każda firma – a nawet każdy

analityk danych – musi zdefiniować i określić, kiedy model należy przeszkolić lub wymienić. Ponadto zawsze istnieje potrzeba zarządzania wycofanymi modelami, ponieważ należy je archiwizować, a nie tylko wyrzucać. Wreszcie, bardziej niezawodny proces zarządzania punktacją modelu jest koniecznością, ponieważ jest to kluczowy wymóg, aby zapewnić możliwość oceny wydajności i rentowności modelu w czasie. Organizacje odnoszące sukcesy zdają sobie sprawę, że modele są podstawowymi zasobami korporacyjnymi, które wytwarzają i dostarczają odpowiedzi dla systemów produkcyjnych w celu poprawy relacji z klientami, usprawnienia operacji, zwiększenia przychodów i zmniejszenia ryzyka. Jednak niewiele firm jest w stanie w pełni zarządzać wszystkimi zawiłościami całego cyklu życia modelu, tylko dlatego, że jest to tak wieloaspektowe zadanie. Więc jeśli nie możesz zrobić tego wszystkiego, na czym powinieneś się skupić? Innymi słowy, czego potrzeba, z perspektywy zarządzania modelami, aby móc podejmować wiele dobrych, szybkich decyzji operacyjnych, które konsekwentnie odzwierciedlają ogólną strategię organizacji, a jednocześnie sprawiają, że Twoja organizacja jest szybsza i lepsza niż ktokolwiek inny? Cóż, jak w większości przypadków, nie ma srebrnej kuli, ale jest kilka głównych aspektów, na których należy się skupić:

* Systemy oparte na danych: Cała konfiguracja operacyjna w firmie musi wykorzystywać dane do generowania odpowiedzi dla osób lub systemów, w zależności od poziomu automatyzacji, aby rozpocząć właściwe działania.

* Niezawodne i zaktualizowane modele: kluczowe znaczenie ma posiadanie odpowiednich i aktualnych modeli, na których firma może polegać w celu podejmowania optymalnych decyzji i działań we właściwym czasie. Decyzje te mogą być podejmowane przez systemy oparte na inteligencji maszynowej, które wykorzystują Twoje modele do automatycznego podejmowania decyzji w warunkach operacyjnych.

* Zintegrowane reguły biznesowe: Integracja reguł biznesowych i predyktywne podejście analityczne do przepływów decyzji operacyjnych musi nastąpić, jeśli chcesz zapewnić wgląd instruktażowy potrzebny do zweryfikowanych, zaufanych decyzji.

* Monitorowanie modeli: nic nie zadziała na dłuższą metę, jeśli nie znajdziesz sposobu na zarządzanie i monitorowanie modeli analitycznych, aby upewnić się, że działają dobrze i nadal dostarczają właściwych odpowiedzi.

* Nowoczesna architektura danych: upewnij się, że masz nowoczesną architekturę danych, która odpowiada Twoim potrzebom i jest obsługiwana przez wydajne i odpowiednie procesy, które mogą rosnąć w celu zaspokojenia nowych potrzeb, takie jak przesyłanie strumieniowe danych i budowanie bardziej szczegółowych modeli predykcyjnych szybciej niż kiedykolwiek.

Wskazanie wyzwań wdrożeniowych

Chociaż przejście od dobrze przemyślanej strategii analizy danych do fazy wdrożenia może wydawać się łatwe, nie zawsze tak jest. Nadal istnieje wiele problemów, które mogą się pojawić podczas wdrażania. Świadomość tych typowych wyzwań z perspektywy zarządzania modelami pomaga poruszać się po nich lub przynajmniej daje kilka wskazówek, jak sobie z nimi radzić, gdy już się pojawią:

* Zbyt powolne wprowadzanie modeli do produkcji: ponieważ procesy są często ręczne i doraźne, uzyskanie modelu może zająć miesiące do wdrożenia do środowiska produkcyjnego. A ponieważ przechodzenie modeli przez fazy programowania i testowania może trwać tak długo, mogą one stać się przestarzałe, zanim dotrą do produkcji - lub nigdy nie zostaną wdrożone. Wewnętrzne i zewnętrzne kwestie zgodności mogą sprawić, że proces będzie jeszcze trudniejszy, zwłaszcza że sytuacja regulacyjna w nauce o danych ewoluuje w tak szybkim tempie.

* Trudności w interpretacji zaleceń modelu: etap przekładania wyników modelu na działania biznesowe w celu podejmowania decyzji operacyjnych wymaga jasnych, uzgodnionych reguł biznesowych, które muszą stać się częścią zarządzanego środowiska, ponieważ są to reguły definiujące sposób wykorzystania wyników modelu . Na przykład model wykrywania oszustw może zwracać ocenę ryzyka oszustwa jako liczbę od 100 do 1000 (podobnie jak ocena kredytowa FICO). To zależy od firmy, aby zdecydować, jaki poziom ryzyka wymaga działania. Jeśli wyzwalacz alertu o oszustwie jest ustawiony zbyt wysoko, oszustwo może pozostać niezauważone. Jeśli wyzwalacz oszustwa jest ustawiony zbyt nisko, alerty powodują zbyt wiele fałszywych alarmów. Oba wyniki zmniejszą wartość, jaką te modele tworzą, a także zmniejszą zaufanie do wyników.

* Modele, które nie działają zgodnie z oczekiwaniami: zbyt często modele o słabej skuteczności pozostają w produkcji, mimo że generują niedokładne wyniki, które prowadzą do złych decyzji biznesowych. Wyniki modelu zmieniają się wraz ze zmianami danych, co jest odzwierciedleniem nowych warunków i zachowań, do których model może nie być w stanie dostosować, nawet jeśli jest to model uczenia maszynowego. Nie stanowiłoby to problemu, gdyby nieścisłości zostały wykryte wystarczająco szybko, ale często tak nie jest. Głównymi przyczynami tej sytuacji są brak centralnego repozytorium modeli, brak spójnych metryk do monitorowania wydajności modelu oraz niewystarczające wytyczne lub mechanizmy kontrolne pozwalające określić, kiedy model należy przeszkolić lub wymienić.

* Procesy zarządzania modelami nie działają w praktyce: Organizacje często znajdują się w trybie reaktywnym, gdy znajdują się pod presją i reagują w pośpiechu, aby dotrzymać terminów. (Dotyczy to zwłaszcza zespołów zajmujących się analizą danych na ich wczesnych etapach, kiedy ich procesy są po raz pierwszy ustanawiane i mają poczucie, że mają wiele do udowodnienia kierownictwu). Może to powodować sytuacje, w których każda grupa ma inne podejście do obsługi i walidacji modelu , co może skutkować szeroką gamą raportów o różnym poziomie szczegółowości do przeglądu lub niespójnie opisanych modeli, co utrudnia interpretację. Nikt nie jest pewien, w jaki sposób wybrano model o najwyższej punktacji (model mistrz), jak obliczono punktację konkretnego modelu ani co rządzi regułami biznesowymi, które uruchamiają model.

* Brak przejrzystości: jeśli nie zajmiesz się aktywnie przejrzystością w zarządzaniu modelami, nie uzyskasz dużego wglądu w różne etapy opracowywania modelu ani dużej wiedzy na temat tego, kto dotyka modelu w trakcie jego cyklu życia. W małej firmie może to być w porządku, ale w większym przedsiębiorstwie szybko przekonasz się, że taka sytuacja może być dość uciążliwa. Mogą pojawić się sprzeczne założenia i spowodować dodatkowe zamieszanie, a kiedy, w ostateczności, bezstronni recenzenci są wzywani do weryfikacji modeli przechodzących przez każdą grupę, stajesz w obliczu dużego drenażu zasobów i dodatkowego uderzenia w lidera rozwoju- czas.

* Utrata ważnej wiedzy o modelach: przy nieodpowiedniej dokumentacji modeli ważna własność intelektualna pozostanie w pamięci twórcy modelu, co poważnie wpłynie na możliwość ponownego wykorzystania modelu. Podejście, które w dużej mierze zależy od kluczowych osób, zwiększa również ryzyko całkowitej utraty ważnych informacji - gdy ta osoba odchodzi z firmy, wiedza znika.

* Niewystarczające zestawy umiejętności: nawet przy rosnącej liczbie analityków danych wchodzących na rynek, brak umiejętności analitycznych potrzebnych do tworzenia i wdrażania modeli jest nadal dużym wyzwaniem dla wielu organizacji. Bez odpowiednich umiejętności w firmie postęp może być powolny, a wyniki słabe.

Zarządzanie ryzykiem modeli

Istotną częścią Twojego podejścia do wdrażania zarządzania modelami powinno być skoncentrowanie się na zrozumieniu i pomiarze ryzyka związanego z wykorzystaniem i zaufaniem do modelu sztucznej inteligencji do podejmowania strategicznych decyzji i konfiguracji operacyjnych. Ryzyko jest związane z terytorium, ponieważ modele uczenia maszynowego/sztucznej inteligencji są probabilistyczne: - dają najlepszą możliwą odpowiedź, ale ta odpowiedź może nadal być błędna. To nie jest prawda absolutna. Kolejna ważna kwestia dotycząca ryzyka wynika z faktu, że modele uczenia maszynowego/sztucznej inteligencji są zaprojektowane do dynamicznego uczenia się, co oznacza, że jeśli zostaną wdrożone w sposób dynamiczny, będą ewoluować w środowisku produkcyjnym na żywo. Oznacza to również, że ramy decyzyjne modelu mogą zmieniać się w czasie, odchodząc od zasad, na których był pierwotnie szkolony w środowisku laboratoryjnym, gdy model zostanie wystawiony na nowe dane, które uruchamiają model do uczenia się i reagowania na nowe zachowania. Dlatego ważne jest, aby wdrożyć wystarczające ograniczenia polityczne, aby upewnić się, że uczenie się modeli pozostaje pod kontrolą. Bardziej oczywiste ryzyko, którym musisz zarządzać, jeśli chodzi o modele uczenia maszynowego/sztucznej inteligencji, polega na tym, że słabe szkolenie, stronniczość oraz złe lub uszkodzone dane mogą wpływać na wyniki modelu. (Śmieci wchodzą, śmieci wychodzą.) Zapewnienie różnorodnych i reprezentatywnych danych treningowych o odpowiedniej jakości to dobry początek, jeśli chodzi o obniżenie tego ryzyka.

Pomiar poziomu ryzyka

Cel pomiaru ryzyka modelu uczenia maszynowego/sztucznej inteligencji jest bardzo związany ze zrozumieniem i zdefiniowaniem profilu ryzyka dla określonego modelu. Mówiąc najprościej, jeśli rozważasz wykorzystanie ustaleń i zaleceń wynikających z konkretnego modelu do bardzo ważnej (i kosztownej) decyzji inwestycyjnej lub ważnej rekomendacji klienta, czy nie chciałbyś wiedzieć, jakie ryzyko wiąże się z zaufaniem do tego modelu że? Jak więc wyglądałaby taka ocena ryzyka? Musisz zacząć od pełnego zrozumienia ryzyka technicznego związanego z zaufaniem modelowi w odniesieniu do wpływu, jaki miałoby takie niesłuszne zaufanie, gdyby model zawiódł. Ryzyko techniczne związane z modelem obejmuje takie aspekty, jak:

* Cele modelu

* Jego możliwości funkcjonalne

* Podejście do uczenia się modelu

* Warunki środowiskowe

* Poziom nadzoru ludzkiego

Wpływ mierzy się przede wszystkim na podstawie potencjalnego finansowego, emocjonalnego i fizycznego wpływu, jaki awaria modelu może mieć na użytkowników zewnętrznych i wewnętrznych, ale obejmuje również szacowanie wpływu z perspektywy reputacyjnej, regulacyjnej i prawnej. Kiedy już zrozumiesz zagrożenia techniczne i skutki awarii, musisz zastanowić się, jak ustanowić właściwy rodzaj mechanizmów kontrolnych. Kluczem jest tutaj znalezienie odpowiedniej równowagi w infrastrukturze analizy danych między przestrzenią roboczą, która jest innowacyjna i wydajna, a jednocześnie pozostaje dokładna i niezawodna.

Identyfikacja odpowiednich mechanizmów kontrolnych

Zrozumienie, które mechanizmy kontrolne należy wprowadzić dla jakich zagrożeń, nie jest łatwym zadaniem, ale Tabela 16-1 pokazuje kilka przykładów, jak można temu zaradzić.

Ryzyko modelu: mechanizm kontroli

Niewystarczający nadzór nad modelem: istnieją procedury monitorowania wydajności modelu i reagowania na odchylenia od oczekiwanych wyników. Stosowane są ścisłe środki bezpieczeństwa i kontroli, aby zapobiec niekontrolowanej ewolucji modelu.

Brak możliwości wyjaśnienia modelu : Przeprowadzono ocenę wpływu na ochronę danych, a wyniki przekazano odpowiednim zainteresowanym stronom.

Stronnicze wyniki: Zbieranie zróżnicowanego zestawu danych treningowych we wszystkich odpowiednich klasach w celu uniknięcia utajonej stronniczości. Stosowanie technik regularyzacyjnych w celu karania za brak równowagi w selekcji w grupie docelowej typu danych . Zróżnicowany zespół jest używany do testowania wyników modelu pod kątem utajonego uprzedzenia.

Niska wydajność modelu: istnieją procedury testowania modelu w różnych warunkach na żywo, aby zapewnić wymaganą wydajność podczas wdrażania

Stosowane są techniki walidacji krzyżowej i składania (łączenie prognoz z kilku różnych modeli), aby zapobiec nadmiernemu dopasowaniu modelu, które występuje, gdy wydajność modelu jest zbyt ściśle powiązana z określonym zestawem danych

Chociaż świadomość i kontrola ryzyka są zdecydowanie dobrymi rzeczami, nauka o danych z samej swojej natury polega na wykorzystaniu podejścia opartego na eksperymentach i maszynach, aby poprawić ludzkie zachowanie, aby wyjść poza to, co jest obecnie możliwe. Nadmierne kontrolowanie i ograniczanie tego podejścia z perspektywy infrastruktury i procesów tylko utrudni innowacyjność i produktywność modeli nauki o danych. Kluczem jest znalezienie zrównoważonego podejścia do zarządzania ryzykiem i kontrolą modeli w porównaniu z innowacjami modeli i kreatywnością biznesową

Odkrywanie znaczenia otwartego oprogramowania

Największe nazwiska w dziedzinie nauki o danych to open source, a wiele z nich jest nawet częścią tej samej (open source) rodziny Apache: Spark, Hadoop, Kafka i Cassandra. Chociaż zamknięte bazy danych są nadal niezwykle popularne, alternatywy open source rozwijają się w szybkim tempie. Oczywiste jest, że jeśli będą się rozwijać, te zamknięte bazy danych nie będą już tak popularne na długo. Ten rozdział koncentruje się na wyjaśnieniu, dlaczego open source jest ważne w nauce o danych, a także przegląd popularnych narzędzi i frameworków.

Odkrywanie roli otwartego oprogramowania

Popularność systemów open source w nauce o danych rośnie z wielu powodów. Po pierwsze, zasady open source opierają się na dzieleniu się zasobami, podejściu, które pozwala różnym ludziom w różnych obszarach efektywnie współpracować. Kiedy firmy dzielą się swoją pracą i pozwalają innym wnieść swój wkład, więcej osób może znaleźć zarówno nowe problemy, jak i nowe możliwości. Na przykład techniki takie jak głębokie uczenie się wiele zawdzięczają dużym graczom, takim jak Google i Facebook, które aktywnie oddają swoje dane i zasoby społeczności. Zawsze wygląda na to, że technologia rozwija się bardzo szybko, ale sam proces nie jest szybki. Gdyby firmy same próbowały poradzić sobie z oprogramowaniem Big Data, bez wkładu lub pomocy różnych programów open source, byłby to boleśnie powolny proces. Istnieje poważna potrzeba nadążania za czasem i nauką o danych , to szybko rozwijająca się dziedzina, w której stale brakuje odpowiednich kompetencji i umiejętności. Wpływa to nie tylko na małe firmy, które chcą nadążać za nimi, ale także na dużych inwestorów, którzy mogą zmienić ogólny bieg biznesu. Firmy chcą szybko rozwinąć swoje działy i zastosowania analityki danych, ale pula talentów i technologia jeszcze się nie pojawiły. Otwarte zaopatrzenie, w którym dane i technologia przynajmniej zmniejszają obciążenie i pozwalają firmom iść naprzód w równym tempie. Podejście społecznościowe oznacza również, że użytkownicy mają możliwość zadawania pytań i uzyskiwania pomocnych odpowiedzi. Zamiast wpadać w korkociąg za każdym razem, gdy pojawia się problem, użytkownik prawdopodobnie znajdzie w społeczności kilku innych, którzy znają odpowiedź lub, co bardziej prawdopodobne, wiedzą, jak ją znaleźć. Kreatywni użytkownicy open source również szukają sposobów na ekonomiczną pracę i oszczędzanie pieniędzy. Prawdopodobnie znajdą lub ulepszą niedrogi sprzęt, podczas gdy duża firma programistyczna z monopolem może skłonić użytkowników do zakupu bardzo specyficznego i drogiego sprzętu.

Zrozumienie znaczenia open source w mniejszych firmach

Gdy firmy zdecydują się wykorzystać swoje dane, często są pochłonięte działaniami skoncentrowanymi na wdrożeniu jeziora danych. Nadrzędnym celem nagle staje się gromadzenie i przechowywanie danych, ale bez odpowiednich zasobów lub, w przypadku mniejszych firm, środków na ich wykorzystanie, dane są całkowicie bezużyteczne. Gdyby mała firma miała płacić za każdy kawałek oprogramowania i szkolenia wymagane do korzystania z danych, zachęta do integracji dużych zbiorów danych w miejscu pracy byłaby znacznie mniejsza. Open source ma jednak mentalność „próbuj, zanim kupisz”. W przypadku firm oferujących produkty oparte na oprogramowaniu open source potencjalni klienci często są zaznajomieni z aspektami open source produktów. Nowi użytkownicy mogą zaryzykować dane przy niewielkim ryzyku, a eksperci mogą stosunkowo łatwo przechodzić między różnymi rozwiązaniami.

Zrozumienie trendu

Jeśli obecne trendy utrzymają się, cała platforma danych nowej generacji będzie open source, co oznacza korzyści dla firm open source i tych, którzy na nich bazują. W nauce o danych open source jest normą. Nawet szkolenia w tej dziedzinie wspierają społeczność open source, często pozostając

bezpłatne. Chociaż dyplomy uniwersyteckie z pewnością okażą się przydatne w przyszłości, wiele firm i programistów po prostu szuka dalszych szkoleń na tematy związane z big data, które można by dodać do swojego arsenału. Darmowe kursy online z zakresu danych są obfite, a programy z Udacity (www.udacity.com), IBM Cognitive Class (https://cognitiveclass.ai, dawniej Big Data University) i inne starają się wypełnić lukę między zainteresowanymi nauką o danych i użytkowników. Nawet Google zorganizował bezpłatne kursy na temat korzystania z danych. Niesamowity wzrost, jakiego doświadczają programy i społeczności open source, jest prawdziwym dowodem na to, że jest to przyszłość danych. Firmy, które podejmują wielkie kroki w zakresie danych, również promują open source. Świadczy to nie tylko o skuteczności open source, ale także pokazuje, dokąd zmierzają finanse w danych i właśnie tam, gdzie firmy stawiają swoje zakłady. SAS Institute, amerykański gigant w dziedzinie danych i analityki, inwestuje wiele w kompatybilność open source, rozumiejąc, że pomysł nie ma na celu konkurowania z open source, ale raczej wymyślenie, jak go uzupełnić.

Na przykład w nowym rozwiązaniu SAS opartym na chmurze analityk danych może kontynuować pracę w środowisku open source, korzystając z ulubionych narzędzi open source lub języków programowania podczas opracowywania, ale gdy nadejdzie czas wdrożenia algorytmu do produkcji, osoba ta może wdrożyć do zwirtualizowanego środowiska SAS, aby zapewnić większą niezawodność, wydajność i monitorowanie. Przeszłość, teraźniejszość i przyszłość big data są silnie zakorzenione w technologii open source i to będzie jeden z jej największych atutów. W obliczu niedoboru naukowców zajmujących się danymi i wykwalifikowanych pracowników najważniejsze będzie, aby firmy i osoby prywatne miały łatwy dostęp do zaawansowanych i aktualnych rozwiązań bez obawy, że zapłacą każdy grosz, aby pozostać w grze. Zwłaszcza, gdy firmy takie jak Google i Facebook dzielą się swoją wiedzą, przyszłość danych będzie tylko lepsza i potężniejsza.

Opisywanie kontekstu języków programowania nauki o danych

Krajobraz nauki o danych szybko ewoluuje, a liczba narzędzi wykorzystywanych do wydobywania wartości z nauki o danych również wzrosła. Aby w pełni wykorzystać potencjał narzędzi i struktur open source, ważne jest, aby strategicznie upewnić się, że Twoja firma buduje i nabywa umiejętności w zakresie języków programowania open source w przestrzeni nauki o danych. Chociaż nie ma określonej kolejności na tej liście popularnych języków do nauki danych, Python i R walczą o pierwsze miejsce. Jednak posiadanie analityków danych z więcej niż jedną umiejętnością językową zapewnia organizacji większą elastyczność. Ile języków programowania open source jest w użyciu? Policzmy sposoby:

* Python: Python jest niezwykle popularnym, dynamicznym i powszechnie używanym językiem ogólnego przeznaczenia w społeczności naukowców zajmujących się danymi. Jest powszechnie określany jako najłatwiejszy do czytania i uczenia się język programowania. Ponieważ łączy szybkie doskonalenie z możliwością współpracy z wysokowydajnymi algorytmami napisanymi w Fortran lub C, stał się wiodącym językiem programowania dla nauki o danych typu open source. Wraz z rozwojem technologii, takich jak sztuczna inteligencja, uczenie maszynowe i analityka predykcyjna, zapotrzebowanie na ekspertów z umiejętnościami Pythona znacząco rośnie. Słabością Pythona jest to, że wykonuje się za pomocą interpretera zamiast kompilatora, co czyni go nieco wolniejszym niż na przykład C lub C++. Python ma również dość duże zużycie pamięci ze względu na elastyczność w zarządzaniu różnymi typami danych.

* R: R to język i środowisko oprogramowania typu open source do obliczeń statystycznych i grafiki, wspierane przez R Foundation for Statistical Computing. Ten zestaw umiejętności cieszy się dużym zainteresowaniem wśród rekruterów zajmujących się uczeniem maszynowym i nauką o danych. R dostarcza wiele modeli statystycznych, a liczni analitycy skomponowali swoje aplikacje w języku R. Jest to ulubiony język otwartej analizy statystycznej i wyraźnie koncentruje się na modelach statystycznych

utworzonych przy użyciu języka R. Publiczne archiwum pakietów R zawiera ponad 8000 wniesionych pakiety. W R jednostką udostępnianego kodu jest pakiet. Microsoft, R Studio i wiele innych organizacji zapewnia wsparcie biznesowe dla przetwarzania opartego na R. Jedną z wad R jest to, że trudniej jest utrzymać, gdy kod się rozrasta. Inną kwestią jest to, że ponieważ R jest niezwykle elastyczny, możesz znaleźć się w wielu sytuacjach, w których możesz zrobić coś dobrze na setki sposobów. Ze względu na łatwość konserwacji i pracę w zespołach może to nie być to, czego chcesz. Aby wesprzeć użycie Pythona i R, powinieneś również rozważyć użycie Anacondy (dotyczy zarówno Pythona, jak i R) lub RStudio (tylko dla R). Anaconda oferuje łatwy sposób wykonywania uczenia maszynowego Python i R w systemach Linux, Windows i Mac OS X i ma ponad 11 milionów użytkowników na całym świecie. do pobierania pakietów do nauki danych Python i R, zarządzania bibliotekami, zależnościami i środowiskami. Oferuje wsparcie w zakresie opracowywania i trenowania modeli uczenia maszynowego i głębokiego uczenia za pomocą Scikit-learn, TensorFlow i Theano, a także analizowania i wizualizacji danych przy użyciu innego specjalistycznego oprogramowania typu open source.

* Java: Java to popularny język ogólnego przeznaczenia, który działa na wirtualnej maszynie Java (JVM). Wiele organizacji, zwłaszcza międzynarodowych korporacji, używa tego języka do tworzenia systemów zaplecza i aplikacji desktopowych/mobilnych/webowych. Jest to system komputerowy obsługiwany przez Oracle, który umożliwia przenoszenie między platformami. Jedną z wielkich zalet Javy jest ogromna baza użytkowników oprogramowania dla przedsiębiorstw, co oznacza, że istnieje dość duża społeczność, z wieloma wykwalifikowanych programistów dostępnych dla Ciebie. Istnieje od dłuższego czasu, a większość inżynierów oprogramowania pakuje umiejętności Java. Jednak nawet Java ma swoje wady. Na przykład Java jest stosunkowo wolniejsza i zajmuje więcej miejsca w pamięci niż inne natywne języki programowania, takie jak C i C++.

* SQL: SQL (Structured Query Language) to kolejny popularny język programowania w dziedzinie nauki o danych, który istnieje już od jakiegoś czasu. Świetnie nadaje się do odpytywania i edytowania informacji przechowywanych w relacyjnych bazach danych i jest używany od dziesięcioleci do przechowywania i pobierania danych. Okazał się szczególnie przydatny do zarządzania szczególnie dużymi bazami danych, skracając czas realizacji żądań online dzięki szybkiemu czasowi przetwarzania. Posiadanie umiejętności SQL mogą być ważnym zasobem dla specjalistów zajmujących się uczeniem maszynowym i analityką danych, ponieważ SQL jest preferowanym zestawem umiejętności w wielu organizacjach.

* Julia: Julia to dynamiczny język programowania wysokiego poziomu zaprojektowany w celu zaspokojenia potrzeb wysokowydajnej analizy numerycznej i obliczeń naukowych. W związku z tym szybko zyskuje popularność wśród naukowców zajmujących się danymi. Ze względu na szybszą realizację Julia stała się idealnym wyborem do radzenia sobie ze złożonymi projektami zawierającymi duże zbiory danych. W przypadku wielu podstawowych testów porównawczych działa 30 razy szybciej niż Python i regularnie działa nieco szybciej niż kod C. Jeśli podoba Ci się składnia Pythona, ale masz do czynienia z ogromną ilością danych, Julia jest kolejnym językiem programowania do nauki.

* Scala: Scala (skrót od języka skalowalnego), jest teraz językiem do programowania funkcjonalnego. Ten uniwersalny, otwarty język programowania działa na JVM. Jest to idealny wybór dla osób pracujących z dużymi zestawami danych i posiada pełne wsparcie dla programowania funkcjonalnego. Ponieważ został opracowany do działania na JVM, umożliwia interoperacyjność z samą Javą, dzięki czemu Scala jest świetnym językiem ogólnego przeznaczenia, a jednocześnie jest idealną opcją dla nauki o danych. (Tak się składa, że framework do przetwarzania klastrów Apache Spark jest napisany w Scali, więc jeśli chcesz żonglować danymi w klastrze z tysiącami procesorów i mieć stos starszego kodu Java, Scala jest dobrym rozwiązaniem typu open source). języki programowania mają wady, a Scala nie jest wyjątkiem. Scala jest zdecydowanie trudna do nauczenia i dlatego trudna do

zaadoptowania. Co więcej, nie ma zbyt dużej obecności w społeczności i jest utrudniony przez ograniczoną kompatybilność wsteczną. Jeśli te minusy przeważają nad plusami w twoich oczach, Scala nie jest dla ciebie.

Rozwijanie frameworków Open Source dla modeli AI/ML

Platforma uczenia maszynowego to interfejs, biblioteka lub narzędzie, które umożliwia programistom łatwiejsze i szybsze tworzenie modeli uczenia maszynowego bez wchodzenia w sedno podstawowych algorytmów. Zapewnia przejrzysty, zwięzły sposób definiowania modeli uczenia maszynowego przy użyciu zbioru gotowych, zoptymalizowanych komponentów. Ogólnie rzecz biorąc, wydajna struktura uczenia maszynowego zmniejsza złożoność uczenia maszynowego, dzięki czemu jest dostępna dla większej liczby programistów. Niektóre z kluczowych cech dobrej struktury uczenia maszynowego to:

* Jest zoptymalizowany pod kątem wydajności w czasie wykonywania

* Jest przyjazny programistom i wykorzystuje tradycyjne sposoby budowania modeli

* Jest łatwy do zrozumienia i kodowania

* Zapewnia równoległość w celu rozproszenia procesu obliczeniowego, aby przyspieszyć

Dramatyczny rozwój sztucznej inteligencji w ostatniej dekadzie wywołał ogromne zapotrzebowanie na sztuczną inteligencję i umiejętności uczenia maszynowego na dzisiejszym rynku pracy. Technologia oparta na uczeniu maszynowym jest obecnie stosowana w prawie każdej branży, od finansów po opiekę zdrowotną. W kolejnych podrozdziałach opisuję wybór popularnych frameworków i bibliotek uczenia maszynowego, wskazując ich mocne i słabe strony, jeśli chodzi o budowanie modeli uczenia maszynowego. Będziesz podejmował kilka ważnych decyzji, jeśli chodzi o wybór frameworka, ale nie zapominaj, że open source pozwala najpierw je wypróbować. Pamiętaj tylko, że nie wszystkie frameworki uczenia maszynowego są zoptymalizowane pod kątem wszystkich rodzajów technik uczenia maszynowego. Chociaż niektóre są dobre do przetwarzania języka naturalnego (NLP), inne zostały zbudowane, aby skupić się na uczeniu głębokim (DL), i chociaż niektóre są bardziej odpowiednie dla różnych typów sprzętu, inne są dostosowane do chmury. Ważne jest, aby zastanowić się, na czym się koncentrujesz, a ponieważ te frameworki stale ewoluują, a także uzupełniają się nawzajem, pozwalają na pewną swobodę wyboru dla użytkowników w warstwie aplikacji.

TensoraFlow

Opracowana przez Google, TensorFlow to biblioteka oprogramowania typu open source stworzona z myślą o głębokim uczeniu lub sztucznych sieciach neuronowych. Dzięki TensorFlow możesz tworzyć sieci neuronowe i modele obliczeniowe za pomocą wykresów przepływu. Jest to jedna z najlepiej utrzymanych i popularnych bibliotek open source dostępnych do głębokiego uczenia się. Framework TensorFlow jest dostępny zarówno w formatach C++, jak i Python. Inne podobne struktury uczenia głębokiego, które są oparte na Pythonie, obejmują Theano, Torch, Lasagne, Blocks, MXNet, PyTorch i Caffe. Możesz użyć TensorBoard do łatwej wizualizacji, dzięki czemu możesz zobaczyć potok obliczeń. Jego elastyczna architektura umożliwia łatwe wdrażanie na różnych rodzajach urządzeń. Z drugiej strony TensorFlow nie ma pętli symbolicznych (sterowanych symbolami i dynamicznych programów do wyszukiwania błędów) i nie obsługuje rozproszonego uczenia się, w którym algorytmy uczenia maszynowego są uruchamiane w konfiguracji przetwarzania rozproszonego rozłożonej na kilka witryn lub środowisk docelowych. Co więcej, nie obsługuje systemu Windows.

Theano

Theano to biblioteka Pythona przeznaczona do głębokiego uczenia się. Za pomocą tego narzędzia można definiować i oceniać wyrażenia matematyczne, w tym tablice wielowymiarowe. Zoptymalizowane pod kątem GPU narzędzie zawiera wiele przydatnych funkcji, w tym integrację z NumPy, dynamiczne generowanie kodu C i różnicowanie symboliczne. Aby jednak uzyskać wyższy poziom, bardziej intuicyjny widok, który to ułatwi aby opracować modele uczenia głębokiego niezależnie od używanego zaplecza obliczeniowego, narzędzie będzie musiało być używane z innymi bibliotekami, takimi jak Keras, Lasagne i Blocks. Narzędzie doskonale nadaje się do pracy na wielu platformach, ponieważ jest kompatybilne z systemami operacyjnymi Linux, Mac OS X i Windows.

Torch

Torch to łatwa w użyciu platforma obliczeniowa typu open source dla algorytmów uczenia maszynowego. Narzędzie oferuje wydajną obsługę GPU, tablice N-wymiarowe, procedury optymalizacji numerycznej, procedury algebry liniowej oraz procedury indeksowania, dzielenia i transpozycji. Oparte na języku skryptowym o nazwie Lua, narzędzie zawiera dużą liczbę wstępnie wytrenowanych modeli. To elastyczne i wydajne narzędzie do badania uczenia maszynowego obsługuje szeroką gamę głównych platform, w tym Linux, Android, Mac OS X, iOS i Windows.

Caffe i Caffe2

Caffe to popularne narzędzie do głębokiego uczenia się przeznaczone do tworzenia aplikacji i tak się składa, że ma dobry interfejs Matlab/C++/Python. Narzędzie pozwala szybko zastosować sieci neuronowe do problemu za pomocą tekstu bez pisania kodu. Narzędzie obsługuje różne systemy operacyjne, takie jak Ubuntu, Mac OS X i Windows. W miarę pojawiania się nowych wzorców obliczeniowych – obliczeń rozproszonych, obliczeń mobilnych, obliczeń o zmniejszonej precyzji i większej liczby przypadków użycia niewizji, co oznacza, że przypadek użycia nie ma reprezentacji obrazu – projekt Caffe wykazał pewne ograniczenia. Wprowadzenie Caffe2 poprawia Caffe 1.0 na wiele sposobów, w tym pierwszorzędną obsługę rozproszonych szkoleń na dużą skalę, wdrażanie mobilne, obsługę nowego sprzętu (oprócz CPU i CUDA) oraz elastyczność w przyszłych kierunkach, takich jak obliczenia kwantowe.

Microsoft Cognitive Toolkit (wcześniej znany jako Microsoft CNTK)

Microsoft Cognitive Toolkit umożliwia programistom wykorzystanie inteligencji w ogromnych zestawach danych poprzez głębokie uczenie oraz zapewnianie skalowania, szybkości i dokładności z jakością klasy komercyjnej i zgodnością z wieloma różnymi językami programowania i algorytmami. Jest to jedna z najszybszych platform uczenia głębokiego z obsługą interfejsu C#/C++/Python. Platforma open source jest dostarczana z potężnym API C++ i jest szybsza i dokładniejsza niż TensorFlow. Narzędzie obsługuje również rozproszone uczenie się dzięki wbudowanym czytnikom danych, zapewniającym bardzo wydajny sposób dostępu do danych. Obsługuje algorytmy takie jak feedforward, CNN, RNN, LSTM i sekwencja do sekwencji. Obsługa platformy narzędzia jest odrobinę ograniczona, ponieważ działa tylko z systemami Windows i Linux.

Keras

Keras to napisana w języku Python biblioteka typu open source, zaprojektowana w celu ułatwienia tworzenia nowych modeli uczenia głębokiego. Ten wysokopoziomowy i intuicyjny interfejs API sieci neuronowej ułatwia tworzenie modeli uczenia głębokiego niezależnie od zaplecza obliczeniowego i może być uruchamiany na platformach uczenia głębokiego, takich jak TensorFlow, Microsoft CNTK i tak dalej. Keras jest znany z łatwości obsługi i modułowości, co czyni go idealnym narzędziem do szybkiego prototypowania. Narzędzie jest zoptymalizowane zarówno pod kątem CPU, jak i GPU.

Scikit-learn

Scikit-learn (dawniej scikits.learn) to bezpłatna biblioteka do uczenia maszynowego dla języka programowania Python. Zawiera różne algorytmy klasyfikacji, regresji i klastrowania, w tym maszyny wektorów pomocniczych, losowe lasy, zwiększanie gradientu, k-średnie i DBSCAN, i jest zaprojektowany do współpracy z numerycznymi i naukowymi bibliotekami Pythona NumPy i SciPy. Narzędzie obsługuje systemy operacyjne, takie jak Windows i Linux. Z drugiej strony nie jest zbyt wydajny w przypadku GPU.

Spark MLlib

Apache Spark MLlib to skalowalna biblioteka uczenia maszynowego, która obejmuje klastrowanie, wymiarowość, regresję, filtrowanie zespołowe, drzewa decyzyjne i interfejsy API potoku wyższego poziomu. Jest to rozproszona struktura uczenia maszynowego, która może być używana w językach Java, Scala, Python i R. Zaprojektowana do przetwarzania danych na dużą skalę, została opracowana na bazie Apache Spark Core i jest szeroko stosowana i koncentruje się na ułatwieniu uczenia maszynowego . Narzędzie współpracuje z NumPy w bibliotekach Python i R.

Azure ML Studio

Azure ML Studio to nowoczesna platforma w chmurze, której naukowcy danych mogą używać do opracowywania modeli uczenia maszynowego w chmurze. Dzięki szerokiej gamie opcji i algorytmów modelowania platforma Azure nadaje się do tworzenia większych modeli uczenia maszynowego. Usługa zapewnia różne miejsca do przechowywania na konto i stosuje model „płatności zgodnie z rzeczywistym użyciem” i może być używana z programami R i Python.

Uczenie maszynowe Amazon

Amazon Machine Learning (Amazon ML) to solidna usługa oparta na chmurze, która ułatwia programistom na wszystkich poziomach umiejętności korzystanie z technologii uczenia maszynowego. Amazon ML zapewnia narzędzia do wizualizacji i kreatory, które poprowadzą Cię przez proces tworzenia modeli uczenia maszynowego bez konieczności poznawania złożonych algorytmów i technologii uczenia maszynowego, w tym innych usług związanych z nauką o danych, takich jak frameworki i usługi bezpieczeństwa danych.

Wybierasz Open Source czy nie?

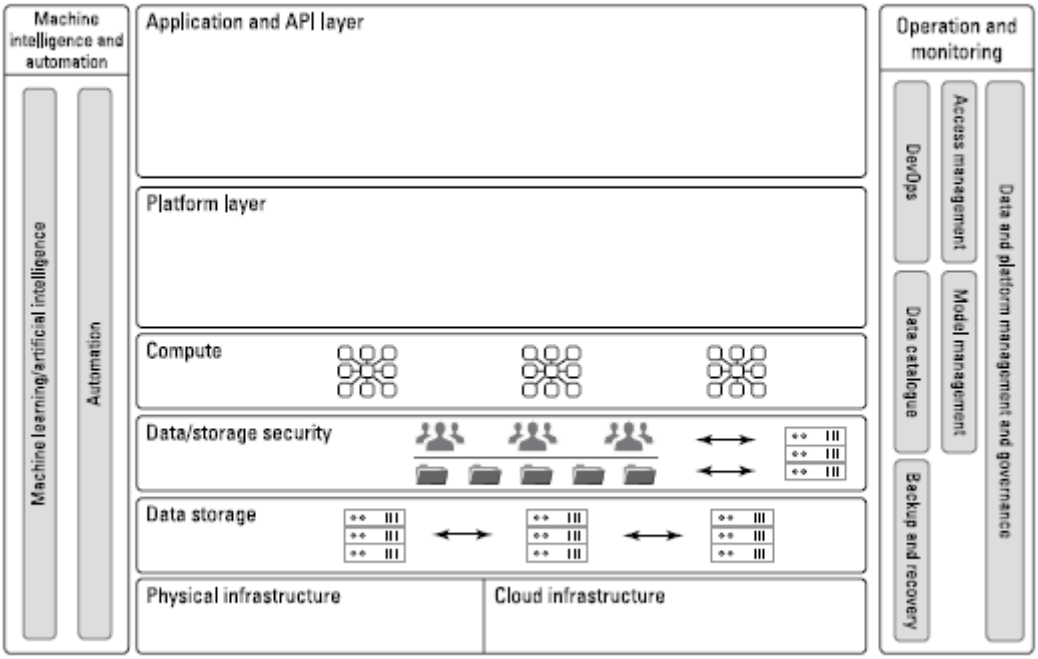
Oczywiście masz nieskończony wybór narzędzi i struktur w przestrzeni open source w nauce o danych, a ponieważ jest to przestrzeń open source, która napędza ewolucję nauki o danych, czy oznacza to, że open source jest jedynym sposobem na udaną naukę danych? Nie, oczywiście nie. Należy jednak zrozumieć, że open source jest potężnym czynnikiem w przestrzeni data science. To nie jest coś, co należy lekceważyć lub marginalizować. Chociaż Twoja inwestycja w naukę danych może wydawać się małą rybką w oceanie, nadal jest częścią oceanu, co oznacza, że musi funkcjonować w rozległym ekosystemie analizy danych. Niewiele poczynionych dzisiaj inwestycji w naukę o danych można postrzegać jako odizolowane środowiska, które nie są zależne od niczego zewnętrznego, takiego jak dane, wymagania prawne, dostawcy, klienci i tak dalej. Biorąc to pod uwagę oraz fakt, że większość dzisiejszej standaryzacji jest również napędzana de facto podejściem standaryzacyjnym w społecznościach open source, musisz zrozumieć i uważnie monitorować to, co dzieje się w przestrzeni open source, nawet jeśli zdecydujesz się zainwestować w całkowicie komercyjny rozwiązanie gotowe do użycia.

Realizacja Infrastruktury

Dane odgrywają kluczową rolę w każdym przypadku użycia nauki o danych, chociaż rodzaj wykorzystywanych danych może się różnić. Na przykład innowacje mogą być napędzane dzięki modelom uczenia maszynowego, które znajdują wgląd w duże ilości danych generowanych przez firmy. W rzeczywistości firma może kultywować całkowicie nowy sposób myślenia wewnątrz organizacji, oparty wyłącznie na nauce o danych, jeśli kierownictwo pchnie w tym kierunku. Kluczem jest zrozumienie roli, jaką dane odgrywają na każdym etapie przepływu pracy w nauce o danych oraz tego, jak należy projektować i obsługiwać infrastrukturę, aby zmaksymalizować wykorzystanie danych, a także zapewnić wysoką produktywność w nauce o danych. Pomogę Ci skoncentrować się na tym, jak wszystkie elementy muszą zostać połączone, aby stworzyć produktywną infrastrukturę danych wspierającą konfigurację analizy danych.

Zbliżanie się do realizacji infrastruktury

Infrastruktura danych to zdecydowanie infrastruktura cyfrowa promująca udostępnianie i konsumpcję danych. Podobnie jak w przypadku innych infrastruktur, jest to struktura potrzebna do funkcjonowania systemu. Sposób realizacji i konfiguracji infrastruktury danych zależy w dużej mierze od celów firmy, wielkości firmy, jej branży i innych czynników. Jednak korzystając z modelu referencyjnego dla swojej infrastruktury danych, będziesz w stanie uzyskać przegląd obszarów, które należy uwzględnić i co należy wziąć pod uwagę, aby dokonać strategicznych wyborów w każdym obszarze. Model referencyjny dla infrastruktury danych to abstrakcyjna struktura do zrozumienia istotnych jednostek i relacji między nimi. Celem jest ułatwienie zrozumienia istniejących infrastruktur danych podczas porównywania ich pod względem funkcjonalności, usług i warunków brzegowych związanych z zakresem tego, co jest uwzględnione lub nie. Rysunek przedstawia widok z lotu ptaka struktury infrastruktury danych bez dołączonych komponentów dla każdej warstwy i obszaru.



Każdy obszar objęty tymi ramami należy dokładnie przeanalizować, a komponenty starannie dobrać w odniesieniu do ogólnej strategii i ambicji firmy, a także kontekstu branżowego, linii biznesowej i dotychczasowej konfiguracji infrastruktury. Różne obszary oznaczone na rysunku - na przykład infrastruktura fizyczna i bezpieczeństwo danych – reprezentują znaczące jednostki w infrastrukturze

danych. Prace nad realizacją infrastruktury danych należy rozpocząć od tych ważnych jednostek, w których należy określić, w jaki sposób będą służyć ogólnemu celowi nauki o danych w zakresie optymalizacji wykorzystania danych i produktywności w zakresie nauki o danych.

Najlepiej znasz swoje dane, klienta i rynek, a stosując tę wiedzę i zrozumienie, w jaki sposób infrastruktura musi wspierać Twoje sposoby osiągania celów, będziesz w stanie dowiedzieć się, jak zrealizować tę infrastrukturę pod kątem jakich komponentów należy umieścić w której warstwie. Wracając do rysunku 18-1, oto krótkie spojrzenie (zaczynając od dołu) na encje i warstwy, z którymi będziesz musiał pracować przy rozwijaniu infrastruktury danych:

* Infrastruktura fizyczna: zasoby sprzętowe niezbędne do wdrożenia, w tym sieci komunikacyjne (kable) i centra danych.

* Infrastruktura chmury: komponenty sprzętowe i programowe — serwery, pamięć masowa, sieć i oprogramowanie do wirtualizacji — potrzebne do obsługi wymagań obliczeniowych modelu przetwarzania w chmurze.

* Przechowywanie danych: dotyczy archiwizacji danych w różnych formach. Jednak różne rodzaje przechowywania danych odgrywają różne role w środowisku komputerowym. Oprócz form twardego przechowywania danych istnieją zdalne przechowywanie danych, takie jak przetwarzanie w chmurze, które znacznie poprawiają sposoby uzyskiwania przez użytkowników dostępu do danych.

* Bezpieczeństwo przechowywania danych: Specjalny obszar bezpieczeństwa związany z zabezpieczaniem systemów przechowywania danych i ekosystemów oraz danych znajdujących się w tych systemach.

* Obliczenia: zasoby wykorzystywane do przetwarzania danych są nazywane zasobami obliczeniowymi, aw przypadku przetwarzania w chmurze są one zwykle dostarczane przez jednostki centralne (CPU) współpracujące ze sobą w klastrach. Aby umożliwić szybkie i wydajne obliczenia dużych zbiorów danych, istnieją również inne zasoby obliczeniowe, takie jak jednostki przyspieszonego przetwarzania (APU) i jednostki przetwarzania grafiki (GPU).

* Platforma: środowisko, w którym wykonywane jest oprogramowanie. Może to być sprzęt lub system operacyjny (OS), a nawet przeglądarka internetowa i powiązane interfejsy programowania aplikacji lub inne podstawowe oprogramowanie, o ile kod programu jest z nim wykonywany. Warstwa platformy to etap, na którym mogą działać programy komputerowe.

* Aplikacje i interfejsy API: Warstwa określająca współużytkowane protokoły komunikacyjne i metody interfejsów używane przez hosty w sieci komunikacyjnej. Składa się z protokołów, które koncentrują się na komunikacji między procesami w sieci IP i zapewniają solidny interfejs komunikacyjny i usługi dla użytkownika końcowego. Warstwa aplikacji jest wykorzystywana w obu standardowych modelach sieci komputerowych: Internet Protocol Suite (TCP/IP) oraz Open Systems Interconnection (OSI).

Realizując infrastrukturę danych, należy również rozważyć, w jaki sposób techniki automatyzacji i uczenia maszynowego będą wykorzystywane w różnych warstwach infrastruktury, ponieważ stosowanie tych technik w ramach działań związanych z przetwarzaniem infrastruktury jest niezbędne ze względu na szybkość, koszt i dane perspektywa integralności. Kolejnym aspektem jest oczywiście to, jak zamierzasz w praktyce obsługiwać infrastrukturę i jej systemy danych. Obejmuje to takie aspekty, jak określenie zasad zarządzania danymi i zarządzania nimi, aspekty regulacyjne, utrzymanie i monitorowanie kompleksowej infrastruktury. Definiowanie i konfigurowanie infrastruktury to jedno, ale aby osiągnąć sukces, należy odpowiednio przemyśleć zarówno operacje, jak i zarządzanie cyklem życia środowiska end-to-end.

Lista kluczowych kwestii dotyczących infrastruktury dla wsparcia AI i ML

W przeszłości infrastruktura dla projektów sztucznej inteligencji i uczenia maszynowego była konfigurowana i prowadzona przez same zespoły zajmujące się analizą danych, ale w większych firmach zadania te są teraz powoli przekazywane specjalistom ds. infrastruktury IT, ponieważ technologie te zaczynają wchodzić do infrastruktury głównego nurtu. Wraz z tym przejściem i inicjatywy sztucznej inteligencji stają się bardziej rozpowszechnione, organizacje IT muszą zacząć uważnie zastanawiać się, jaki rodzaj infrastruktury najlepiej umożliwia wydajność sztucznej inteligencji w tej stale zmieniającej się przestrzeni nauki o danych. Zamiast kupować serwery, infrastrukturę sieciową i inne komponenty do konkretnych projektów, celem powinno być szersze spojrzenie na potrzeby biznesowe zarówno dziś, jak i jutro, podobnie jak w dzisiejszych centrach danych. W jaki sposób można zbudować infrastrukturę w taki sposób, aby składała się zarówno ze stabilnego fundamentu w dolnych warstwach, aby zapewnić szybkość, efektywność kosztową i ponowne wykorzystanie, ale jednocześnie z elastycznością użytkownika i podejściem samoobsługowym?

Lokalizacja

Inicjatywy związane ze sztuczną inteligencją i uczeniem maszynowym nie są prowadzone wyłącznie w chmurze ani nie są obsługiwane wyłącznie lokalnie. Inicjatywy te powinny być realizowane w miejscu, które ma największy sens, biorąc pod uwagę oczekiwany efekt. Na przykład system rozpoznawania twarzy na lotnisku może być zmuszony do przeprowadzenia analizy lokalnie ze względu na ochronę danych i bezpieczeństwo, a w niektórych przypadkach być może w zależności od wymagań dotyczących opóźnień (czas odpowiedzi do i z chmury). Niezbędne jest zapewnienie, że infrastruktura może być wdrażana na różne sposoby; w chmurze, w lokalnym centrum danych lub na krawędzi (na urządzeniu), aby zoptymalizować wydajność inicjatyw AI w zależności od wymagań i kontekstu. Lokalizacja to także coś więcej niż lokalizacja infrastruktury; może również odnosić się do położenia geograficznego. To, czy Twoja firma jest globalna, czy lokalna, wpłynie na zasięg geograficzny potrzebny do konfiguracji Twojej infrastruktury. Ostatni aspekt lokalizacji, który należy wziąć pod uwagę, dotyczy tego, czy dążysz do stworzenia środowiska nauki o danych w celu zwiększenia wewnętrznej wydajności biznesowej, czy też infrastruktura powinna również (lub tylko) wspierać komercyjną ofertę biznesową produktów i usług związanych z danymi. Lokalizacja staje się tutaj kluczowym elementem, jeśli istnieje regulacja danych, opóźnienie systemu lub inna potrzeba, która zmusza Cię do bycia jak najbliżej klientów.

Pojemność

Wydajność sztucznej inteligencji w dużym stopniu zależy od infrastruktury bazowej. Na przykład procesory graficzne (GPU) mogą przyspieszyć głębokie uczenie się 100 razy w porównaniu z tradycyjnymi jednostkami centralnymi (CPU). Słaby serwer spowoduje opóźnienia w procesie, podczas gdy serwer o zbyt dużej mocy marnuje pieniądze. Niezależnie od tego, czy strategia jest kompleksowa, czy najlepsza w swojej klasie, upewnij się, że sprzęt obliczeniowy ma odpowiednią kombinację możliwości przetwarzania i szybkiej pamięci masowej. Wymaga to wyboru dostawcy, który ma szerokie portfolio, które może zająć się dowolnym etapem cyklu życia sztucznej inteligencji.

Konfiguracja centrum danych

Jeśli chodzi o centra danych, infrastruktura danych nie żyje w izolacji - zawsze jest uważana za rozszerzenie obecnej konfiguracji, poprzez zwiększenie liczby dostępnych centrów danych lub przekształcenie części obecnej infrastruktury w taką, która opiera się na nauce o danych. Najlepiej byłoby, gdyby firmy szukały rozwiązania, którym można zarządzać za pomocą istniejących narzędzi (lub przynajmniej w ramach konfiguracji, która uzupełnia to, co już masz), zamiast rezygnować z wszystkiego i zaczynać od nowa. W niektórych przypadkach może to jednak nie być możliwe, nawet

jako środek tymczasowy. Jeśli Twoja obecna infrastruktura jest beznadziejnie przestarzała, z kosztownymi i mało wydajnymi starszymi narzędziami, których nie można zwirtualizować ani skonteneryzować, dalsze korzystanie z niej nie jest warte wysiłku. Jeśli ambicją Twojej firmy jest przekształcenie się w firmę w pełni napędzaną danymi i sztuczną inteligencją, która zamierza zbudować zwirtualizowany biznes w oparciu o infrastrukturę w pełni opartą na chmurze, obecna konfiguracja musi zniknąć.

Kompleksowe zarządzanie

Nie ma jednej „AI w pudełku", którą można wrzucić i włączyć, aby rozpocząć proces AI. Składa się z kilku ruchomych części, w tym serwerów, pamięci masowej, sieci i oprogramowania, z wieloma możliwościami wyboru w każdej pozycji. Najlepszym rozwiązaniem jest rozwiązanie całościowe, które obejmuje wszystkie (lub przynajmniej większość) komponentów, którymi można zarządzać za pomocą jednego interfejsu. Chociaż jest to skomplikowane, spróbuj myśleć o „prostoty" w kategoriach tego, jak należy nią zarządzać. Wykorzystaj automatyzację w jak największej liczbie aspektów w aspektach zarządzania i operacyjnych środowiska danych.

Infrastruktura sieci

Wdrażając rozwiązania z zakresu sztucznej inteligencji, połóż nacisk na serwery obsługujące procesory GPU, pamięć flash i inne elementy infrastruktury obliczeniowej. Ma to sens, ponieważ sztuczna inteligencja w dużym stopniu obciąża procesor i pamięć. Nie zapominaj jednak, że Twoje dane muszą w jakiś sposób dostać się do systemów pamięci masowej i serwerów, co oznacza, że musisz zwracać uwagę na możliwości swojej sieci. Pomyśl o infrastrukturze sztucznej inteligencji jak o trójnożnym stołku, w którym jedna noga składa się z serwerów; inny system przechowywania; a trzecią sieć. Każdy musi być równie szybki, aby nadążyć za sobą i nie powodować nierównowagi w infrastrukturze, i nigdy nie może być silniejszy ani szybszy niż najsłabsza lub najwolniejsza część. Opóźnienie w którymkolwiek z tych komponentów może pogorszyć wydajność. Taki sam poziom należytej staranności jak serwery i pamięć masowa należy do sieci – sprawdzając przepustowość łącza do przesyłania danych między punktami na przykład odbiór do punktu obliczeniowego. Pamiętaj, że konfiguracja infrastruktury sieciowej może obejmować więcej niż jeden kraj, przynajmniej w przypadku firm globalnych lub firm zależnych od dużych ilości danych z innych krajów.

Bezpieczeństwo i etyka

Sztuczna inteligencja często obejmuje niezwykle wrażliwe dane, takie jak akta pacjentów, informacje finansowe i dane osobowe. Naruszenie tych danych może być katastrofalne dla organizacji. Ponadto infuzja złych lub stronniczych danych może spowodować, że system AI będzie dokonywał błędnych wniosków, prowadząc do błędnych decyzji. Infrastruktura sztucznej inteligencji musi być od początku do końca zabezpieczona najnowocześniejszą technologią, obejmującą zarówno aspekty bezpieczeństwa, jak i mechanizmy kontroli etyki sztucznej inteligencji. I chociaż rzadko zapomina się o aspektach bezpieczeństwa, o tyle etyczne tak. Więcej szczegółów na temat mechanizmów kontrolnych dotyczących zarządzania aspektami etycznymi w modelach sztucznej inteligencji można znaleźć w rozdziale 16. Ponieważ ograniczenia regulacyjne stają się coraz bardziej rygorystyczne, zaczynają uwzględniać zasady infrastruktury etycznej – na przykład wykorzystywanie reprezentatywnych i bezstronnych danych lub tworzenie zróżnicowanych i bezstronnych danych zespoły naukowe – które muszą być spełnione, aby uniknąć sytuacji, w których źle przetwarzane dane mogą spowodować, że systemy sztucznej inteligencji staną się dyskryminujące, inwazyjne, a nawet niebezpieczne.

Usługi doradcze i wspierające

Chociaż usługi takie jak szkolenia w zakresie nauki o danych i różne zadania doradcze nie są technicznie uważane za część infrastruktury, muszą być częścią decyzji infrastrukturalnej. Większość organizacji, zwłaszcza niedoświadczonych, nie posiada wewnętrznych umiejętności niezbędnych do tego, by sztuczna inteligencja była skuteczna i wydajna. A partner usługowy może świadczyć niezbędne usługi szkoleniowe, doradcze, wdrożeniowe i optymalizacyjne w całym cyklu życia nauki o danych i powinien być uważany za kluczowy element wdrożenia.

Dopasowanie ekosystemu

Ekosystemy danych można opisać jako składające się z wielu podmiotów, które współdziałają ze sobą w celu wymiany, wytwarzania i konsumowania danych. Takie ekosystemy dostarczają różnych istotnych komponentów do tworzenia, zarządzania i utrzymywania danych w ramach infrastruktury danych. Niektórzy twierdzą, że data science, szczególnie w odniesieniu do uczenia maszynowego i sztucznej inteligencji, jest na poziomie dojrzałości porównywalnym z tym, w którym biznes oprogramowania znajdował się w latach 70. lub 80. Dlatego wyprzedzanie konkurencji w zakresie zarządzania wydajnymi ekosystemami inteligencji maszynowej może stać się w przyszłości główną przewagą konkurencyjną. Żaden pojedynczy dostawca sztucznej inteligencji nie jest w stanie zapewnić całej technologii wszędzie. Musisz wybrać dostawców, którzy zapewniają szerokie wsparcie ekosystemu i mogą połączyć wszystkie lub wiele komponentów sztucznej inteligencji, aby dostarczyć w pełni wydajne, kompleksowe rozwiązanie. Konieczność samodzielnego składania komponentów zwykle prowadzi do niepotrzebnych opóźnień, a nawet awarii. Wybór dostawcy z silnym ekosystemem może zamiast tego zapewnić szybką ścieżkę do sukcesu.

Automatyzacja przepływów pracy w Twojej infrastrukturze danych

W organizacji opartej na danych dane napędzają każdy proces biznesowy, ale nie można w pełni przyspieszyć i zoptymalizować jego procesów, jeśli nie zautomatyzujesz obciążeń związanych z zarządzaniem danymi na każdym etapie. Ręczne punkty styku i przepływy pracy w całym cyklu życia zarządzania danymi utrudniają wielu firmom przyjmowanie, agregowanie, przechowywanie, przetwarzanie, analizowanie, wykorzystywanie i inne maksymalne wykorzystanie ich zasobów danych. Automatyzacja większej liczby procesów potoku danych może pomóc Twojej firmie w przeprowadzaniu transakcji, podejmowaniu decyzji, przemyśleniu strategii oraz lepszym i szybszym wykorzystywaniu konkurencyjnych okazji. Coraz więcej specjalistów od danych przyjmuje podejście skoncentrowane na tworzeniu architektury opartej na automatyzacji, w której powtarzalne zadania, takie jak skrypty integracji danych i modele uczenia maszynowego, można szybko wdrożyć do środowiska produkcji. Jednak automatyzacja dużych części potoku danych wymaga integracji praktyk DevOps ze wszystkimi funkcjami i rolami zależnymi od potoku. Obejmuje to role, takie jak analitycy danych, inżynierowie danych, analitycy biznesowi i administratorzy danych. Może to mieć wpływ nawet na operacje IT i innych interesariuszy, w zależności od konfiguracji firmy i tego, czy potok jest skonfigurowany i obsługiwany przez dział IT, czy nie. Automatyzacja wymaga również, aby praktyki DevOps obejmowały całą infrastrukturę, w tym zróżnicowane centra danych, komputery mainframe i chmury prywatne, a także wszelkie zewnętrzne oferty „jako usługa”, z których korzysta firma. Idealnie, z perspektywy operacyjnej, powinieneś mieć jeden wizualny interfejs, za pomocą którego można tworzyć powtarzalne skrypty, uruchamiać zaplanowane zadania, opracowywać dopracowane reguły i orkiestracje oraz w inny sposób w pełni automatyzować planowanie, zużycie i administrowanie zasobami w całym środowisku danych rozproszonych. wykorzystać oszczędność czasu i kosztów dzięki zautomatyzowanym przepływom pracy.

Zapewnienie wydajnej przestrzeni roboczej dla inżynierów danych i naukowców zajmujących się danymi

W ciągu ostatnich pięciu lat słyszałem wiele historii zespołów zajmujących się analizą danych o ich sukcesach i wyzwaniach związanych z tworzeniem, wdrażaniem i monitorowaniem modeli. Niestety słyszałem też o błędnym przekonaniu, że data science należy traktować tak samo, jak tworzenie oprogramowania. To nieporozumienie jest całkowicie zrozumiałe. Tak, nauka o danych obejmuje kod i dane, ale ludzie wykorzystują naukę o danych, aby znaleźć odpowiedzi na wcześniejsze pytania nierozwiązywalne. W rezultacie praca z nauką o danych jest bardziej eksperymentalna, iteracyjna i odkrywcza niż tworzenie oprogramowania. Praca z nauką o danych obejmuje intensywne obliczeniowo algorytmy, które korzystają ze skalowalnych zasobów obliczeniowych, a czasami wymagają specjalistycznego sprzętu, takiego jak procesory graficzne. Praca w zakresie analizy danych również wymaga danych — o wiele więcej danych niż wymaga to typowe oprogramowanie. Wszystkie te potrzeby i nie tylko pokazują, czym różni się nauka o danych od tworzenia oprogramowania. Potrzeby te podkreślają również kluczowe znaczenie współpracy między nauką o danych i inżynierią dla innowacyjnych, opartych na modelach firm, które dążą do utrzymania lub zwiększenia swojej przewagi konkurencyjnej. Wraz ze wzrostem ilości danych w organizacji rośnie liczba inżynierów, analitycy i analitycy danych potrzebowali do przeanalizowania tych danych. Obecnie zespoły IT nieustannie zmagają się ze znalezieniem sposobu na alokację budżetów infrastruktury Big Data między różnych użytkowników w celu optymalizacji wydajności. Użytkownicy danych, tacy jak analitycy i analitycy danych, również spędzają ogromną ilość czasu na dostrajaniu swojej infrastruktury Big Data, co może nie być ich podstawową wiedzą, a przynajmniej nie tym, nad czym są przypisani do pracy. Nie należy lekceważyć znaczenia wydajnej wspólnej przestrzeni roboczej między zespołami. Taka wspólna przestrzeń robocza powinna być w stanie obsłużyć wszystkie procesy analityczne od początku do końca, w różnych systemach i organizacjach, i zapewnić, że po drodze nie zostanie utracona produktywność. Jak można sobie wyobrazić, nie jest to łatwe zadanie, ale dzięki wspólnemu podejściu do obszaru roboczego dla całego zespołu analityków danych minimalizujesz nieuniknione wymiany między inżynierami danych i analitykami danych, tworząc płynny przepływ pracy od przechwytywania danych do wdrażania modele do produkcji. Wspólna przestrzeń robocza do współpracy powinna również mieć wystarczające wsparcie ekosystemu dla najpopularniejszych języków i narzędzi, co pozwoli praktykom korzystać z preferowanego zestawu narzędzi. Dobra przestrzeń robocza do współpracy między zespołami powinna również wspierać pracę zespołową między inżynierami danych i analitykami danych za pomocą interaktywnych notatników (które działają jako interaktywne środowisko kodowania), interfejsów API lub ich ulubionych zintegrowanych środowisk programistycznych (IDE), a wszystko to jest wspierane kontrolą wersji i zarządzaniem zmianami Pomoc. Praktycy muszą również mieć dostęp do wszystkich potrzebnych danych w jednym miejscu i zautomatyzować najbardziej złożone potoki danych z planowaniem zadań, monitorowaniem we wstępnie zdefiniowanych przepływach pracy. Ten dostęp zapewnia zespołom ds. analizy danych pełną elastyczność w uruchamianiu i utrzymywaniu potoków danych na dużą skalę i na każdym etapie cyklu życia analizy danych.

Inwestowanie w dane jako firma

Pomysł „dane to atut" nie jest nowy. Jednak pomimo dużej liczby osób, które złożyły oświadczenie, nadal istnieje ogromna różnica między mówieniem o uczynieniu danych zasobem, a faktycznym zrobieniem tego. A jeśli chodzi o przeforsowanie nieco bardziej rozbudowanego przekazu „dane jako biznes", niewielu jest skłonnych podjąć ten ambitny krok, mimo że (jak bym argumentował) możliwości są nieograniczone dla tych, którzy mają odwagę iść dalej i robić to. Więc tak, od jakiegoś czasu było trochę szumu wokół danych, ale obecnie interesującym pytaniem jest, dlaczego coraz więcej uwagi poświęca się danym jako zasobom i dlaczego wydaje się, że niektóre firmy są odkrywasz to po raz pierwszy? I dlaczego wiele firm wciąż nie docenia danych i informacji – lub nie jest w stanie ich wykorzystać – chociaż analitycy, dostawcy i inni wciąż powtarzają przesłanie o znaczeniu danych? Ten rozdział stara się odpowiedzieć na te pytania.

Odkrywanie, jak zarabiać na danych

Do tej pory prawie każdy zdaje sobie sprawę, że na danych można zarobić, ale nie każdy dobrze rozumie, jak te pieniądze się zarabia. Poniższa lista punktowana ujawnia tajemnice, pokazując przykłady niektórych różnych rodzajów możliwości monetyzacji danych.

* Reklama cyfrowa: odpowiednia treść, odpowiednia publiczność, właściwy czas

* Usługi finansowe: krzyżowanie i sprzedaż dodatkowa oraz wykrywanie oszustw

* Zarządzanie ruchem: zmniejsz zatory i zoptymalizuj trasy dostaw

* Zoptymalizowane reklamy billboardowe: zrozum ruch i dostosuj przekaz

* Transport publiczny: zadowolenie pasażerów, wydajność operacyjna, możliwości uzyskania przychodów

* Handel detaliczny: zoptymalizuj rozmieszczenie sklepów i personel, monitoruj konkurencję

* Rozrywka i wydarzenia: zarządzaj ruchem, ukierunkowuj promocje

* IoT (Internet of Things): Dodawaj wartość dzięki danym o lokalizacji i nie tylko Dane są nie tylko głównym czynnikiem łączącym ludzi – są podstawą wielu perspektyw firmy, takich jak zrozumienie i poprawa jakości obsługi klienta i obsługi klienta. I nie chodzi tylko o dane – chodzi również o uzyskanie do nich dostępu, niezależnie od tego, czy powodem jest budowanie nowych modeli biznesowych, prowadzenie marketingu opartego na danych, czy po prostu uzyskanie dostępu do właściwych danych, które umożliwiają lepsze podejmowanie decyzji. Nie chodzi o to, że firmy nie rozumieją znaczenia danych, informacji i inteligencji umożliwiającej działanie. (Cóż, niektórzy tak.) Chodzi głównie o to, że wiele firm często nie rozumie w pełni, jaka część danych dotyczących aktywów biznesowych naprawdę jest lub jak dane są odróżniane od technologii, przez którą przepływają. Jednak pojawienie się dyrektora ds. danych (CDO) w wielu organizacjach i w różnych branżach wskazuje na rosnące uznanie danych jako strategicznego zasobu biznesowego, który sam w sobie jest. W większości organizacji posiadających CDO rola ta będzie albo uczestniczyć, albo skutecznie kierować działaniami w zakresie monetyzacji danych, jako sposób na zademonstrowanie własnej wartości CDO. Ta przywódcza rola może wiele zrobić, aby wzmocnić rosnący wpływ CDO, ale aby naprawdę osiągnąć prawdziwe korzyści ekonomiczne płynące z danych, firmy muszą wspierać

Inicjatywy CDO i zacznij traktować dane jako prawdziwy zasób biznesowy.

Podejście do zarabiania na danych polega na traktowaniu danych jako zasobu.

Wykonanie siedmiu kroków opisanych w tej sekcji pomoże Twojej organizacji lub firmie przyjąć ustrukturyzowane podejście do zarabiania na danych, dzięki czemu możesz zacząć traktować dane jako zasób:

1. Przypisz menedżera produktu danych. Daj swoim danym taką samą uwagę w zarządzaniu produktami, jaką poświęcasz innym cennym zasobom i kompetencjom. Organizacje zazwyczaj mają zdefiniowane podejście do zarządzania i marketingu produktów. Podobnie, jeśli rozważasz licencjonowanie danych w dowolnej formie, potrzebujesz kogoś, kto zajmie się definiowaniem i rozwijaniem rynku zasobu danych i przekształceniem go w prawdziwy produkt danych.

2. Poznaj swoje dane i zrób inwentaryzację dostępnych zasobów danych. Bardzo ważne jest, abyś ocenił, do jakich danych masz dostęp i czego potrzebujesz w przyszłości. A jeśli nie nauczysz się w pełni posługiwać się danymi i naprawdę nie zrozumiesz szczegółów swoich danych, nie będziesz w stanie wykorzystać ich jako zasobu. Upewnij się, że identyfikujesz wszystkie rodzaje danych – w tym operacyjne, komercyjne, publiczne, media społecznościowe i treści internetowe – które możesz wydobyć w poszukiwaniu nowych form wartości. Następnie pomóż liderom biznesowym zrozumieć zakres dostępnych danych i korzystać z różnych technik zarządzania danymi i eksploracji danych, aby udoskonalić surowe dane w bardziej przystępne i komunikowalne formy.

3. Oceń bezpośrednie i pośrednie metody monetyzacji danych. Zarabianie pośrednie wymaga wewnętrznego wykorzystania danych w celu usprawnienia procesu lub produkt w sposób, który przyniesie wymierne rezultaty, takie jak wzrost dochodów lub oszczędności kosztów. Zarabianie bezpośrednie obejmuje pewnego rodzaju transakcję lub włączenie danych do nowego produktu lub usługi związanej z danymi. Może to również oznaczać faktyczną sprzedaż samych danych (zwykle poprzez licencjonowanie ich) w takiej czy innej formie.

4. Obserwuj innych. Pożyczaj pomysły z innych branż. Sprawdź, co robią inne branże, aby rozpocząć własne działania monetyzacyjne. Coraz ważniejsze staje się spojrzenie poza określoną branżę, nie tylko po to, aby znaleźć dobre pomysły, ale także jako wczesny sygnał ostrzegawczy o tym, w jaki sposób inne firmy rozwijają swoje inicjatywy monetyzacji informacji, które mogą w nieoczekiwany sposób zakłócić Twój rynek.

5. Przetestuj pomysły pod kątem wykonalności. Przetestuj pomysły, zadając serię pytań dotyczących wykonalności dotyczących tego, czy Twoje pomysły są praktyczne, rynkowe, skalowalne, prawne, etyczne, ekonomiczne i tak dalej.

6. Przygotuj dane. Zastanów się, w jaki sposób zamierzasz zbierać dane i z jakich źródeł danych. Następnie musisz pracować z danymi, aby zwiększyć ich wartość analityczną i potencjalną ekonomiczną. Ponownie zastanów się, w jaki sposób fizyczne procesy produkcyjne wykorzystują surowce do ostatecznego wytworzenia gotowych produktów.

7. Zdecyduj się na strategię marketingową. Wreszcie, w przypadku produktów z danymi, które chcesz wprowadzić na rynek komercyjny, powinieneś skupić się na aspekcie marketingowym. Ważnym aspektem, od którego należy zacząć, jest spakowanie produktu danych i określenie, w jaki sposób będzie on pozycjonowany, wyceniony i sprzedany. Należy również zastanowić się, jakie warunki będą miały zastosowanie do konkretnego zastosowania produktu danych. Wyznaczony menedżer produktu danych będzie odgrywał ważną rolę w tym procesie.

Nie traktuj monetyzacji danych jako nadzwyczajnego zadania lub zadania przeznaczonego tylko dla najnowocześniejszych zastosowań cyfrowego biznesu. Można ją raczej postrzegać jako kluczową kompetencję dla każdej organizacji w dzisiejszych czasach – innymi słowy, taką, która ma potencjał

generowania znaczącej wartości ekonomicznej z różnych zasobów danych, którymi dysponuje każda firma. ROI z produktów danych zostanie osiągnięty, gdy produkty danych będą traktowane jako rzeczywiste produkty i nic więcej. Ważnym krokiem na tej drodze jest zastosowanie do danych tradycyjnych metod zarządzania produktami. Musisz wyjaśnić pogląd, że komercyjny produkt danych to coś innego, co należy traktować z określonymi zasadami i metodami.

Monetyzacja danych w gospodarce opartej na danych

Traktując swoje produkty z danymi jak każdy inny produkt z portfolio produktów swojej firmy, musisz stale przypominać sobie, że produkty z danymi są częścią globalnego zjawiska rynkowego zwanego gospodarką danych. Zrozumienie ekosystemu tej globalnej gospodarki opartej na danych ma kluczowe znaczenie dla odniesienia sukcesu przy wprowadzaniu do tej gospodarki nowego produktu lub usługi związanej z danymi. Mówiąc najprościej, gospodarka oparta na danych to gospodarka oparta na danych, technologiach danych oraz produktach i usługach związanych z danymi. Wywodzi się z nowej gospodarki światowej – opartej na przejściu od gospodarki opartej na produkcji do gospodarki opartej na usługach i informacji. Gospodarka danymi to cyfrowy ekosystem i sieć różnych graczy, takich jak dostawcy danych i użytkownicy danych. Termin gospodarka oparta na danych odnosi się do zdolności organizacji i ludzi do wykorzystywania danych jako zasobu. Dane są wykorzystywane do podejmowania strategicznych decyzji, poprawy wydajności operacyjnej i zrównoważonego rozwoju , wzrost, dobrobyt i innowacje. Wartość i wpływ danych zwiększają czynniki sytuacyjne, kontekstowe, historyczne i czasowe. Integrowanie, udoskonalanie i udostępnianie danych zwiększa ich wartość i wpływ w gospodarce opartej na danych. Efektywne wykorzystanie danych może prowadzić do rozwoju firmy, poprawy jakości życia i tworzenia sprawnych społeczeństw. Jednak skuteczne wykorzystanie danych może być utrudnione przez krajowe lub regionalne przepisy ustawowe i wykonawcze ograniczające wykorzystanie danych i wydajną wymianę danych, co wyjaśniono bardziej szczegółowo w rozdziale 24. Firmy zwiększają swoją obecność i potencjalną wartość w gospodarce opartej na danych na wiele sposobów i nie jest konieczne, aby firma pozostawała w jednej z tych warstw. Wiodące firmy zwykle rozwijają się w ramach jednej warstwy, wielu warstw lub całego stosu technologicznego.

GDZIE MYDATA CHCE NAS ZABRAĆ

MyData to skoncentrowane na człowieku podejście do zarządzania danymi osobowymi, które łączy zapotrzebowanie branży na dane z cyfrowymi prawami człowieka. Misją z misją ruchu MyData jest wzmocnienie pozycji osób fizycznych poprzez poprawę ich prawa do samostanowienia w odniesieniu do ich danych osobowych. Paradygmat skoncentrowany na człowieku dąży do sprawiedliwego, zrównoważonego i dostatniego społeczeństwa cyfrowego, w którym dzielenie się danymi osobowymi opiera się na zaufaniu. Celem jest budowanie zrównoważonych relacji między osobami i organizacjami. „Ruch MyData" wystartował we wrześniu 2016 roku w ramach konferencji MyData2016 w Helsinkach w Finlandii.

Gospodarkę danymi można również postrzegać na różne sposoby pod kątem tego, jak dane są postrzegane z perspektywy gospodarki światowej.

* Gospodarka dużych zbiorów danych: można ją zdefiniować jako opartą na algorytmach analizę wielkoskalowych danych cyfrowych w celu przewidywania, pomiaru i zarządzania zasobami danych.

* Gospodarka oparta na danych napędzana przez człowieka: jest to sprawiedliwa i funkcjonująca gospodarka oparta na danych, w której dane są kontrolowane i wykorzystywane w sposób uczciwy i etyczny w sposób zorientowany na człowieka. Oparta na ludziach gospodarka danych jest powiązana z ruchem MyData (patrz pasek boczny) i podejściem do zarządzania danymi osobowymi, które koncentruje się na człowieku.

* Gospodarka danymi osobowymi: jest to możliwe dzięki osobom skoncentrowanym na korzystaniu z danych osobowych, które wszystkie osoby generują i dostarczają bezpośrednio lub pośrednio. Konsumenci danych osobowych stają się dostawcami i administratorami, tak jak wtedy, gdy Facebook wykorzystuje nasze dane osobowe do dostarczania podobnych interesujących nas tematów w aplikacji. Podobnie Uber zadeklarował niedawno, że algorytmy będą analizować dane osobowe w czasie rzeczywistym i będą pobierać od klientów to, co przewiduje algorytm, a nie ryczałtową stawkę.

* Ekonomia algorytmów: tutaj firmy i osoby fizyczne mogą kupować, sprzedawać, wymieniać lub przekazywać poszczególne algorytmy lub komponenty aplikacji.

Patrząc w przyszłość gospodarki opartej na danych

Gospodarka oparta na danych zwiększa konkurencyjność, innowacyjność i możliwości biznesowe na skalę światową. Według najnowszych szacunków rosnące globalne przepływy danych zwiększyły światowy PKB o ponad 10%. Można to porównać tylko z danymi liczbowymi dla Europy, gdzie oczekuje się, że nowe regulacje polityczne, warunki prawne i inwestycje w ICT zwiększą wartość europejskiej gospodarki opartej na danych do 739 mld euro do 2020 r., co stanowi 4% całkowitego PKB UE. Kluczowe sektory gospodarki opartej na danych albo są już oparte na danych, albo są na dobrej drodze, aby tak się stało, w takich obszarach jak produkcja, rolnictwo, motoryzacja, telekomunikacja i inteligentne środowiska życia. Opieka zdrowotna i farmacja również stanowią podstawę gospodarki opartej na danych. Świat zmierza również w kierunku bardziej sprawiedliwej gospodarki opartej na danych, która przynosi korzyści wszystkim. Odpowiedzialne zarządzanie danymi osobowymi ułatwia codzienne życie i przyczynia się do dobrego samopoczucia wielu osób. Ujednolicona procedura otwiera możliwości dla innowacji i działań biznesowych zorientowanych na użytkownika. Osoby fizyczne zaczynają mieć teraz większą kontrolę nad (i przejrzystością) dotyczących ich danych. Osoby fizyczne mogą aktywnie określać warunki, w jakich wykorzystywane są ich dane osobowe. Usługodawcy godni zaufania klientów mogą również uzyskać dostęp do znacznie bardziej rozbudowanych i zróżnicowanych e-usług danych. Wszystko nie jest słodyczą i światłem na horyzoncie gospodarki opartej na danych – niektóre prawdziwe wyzwania zdecydowanie spadają. Nie będzie łatwo stale wymyślać coraz to nowe podejścia do radzenia sobie z naruszeniami danych, ponieważ techniki hakerskie dostosowują się do nowych środków bezpieczeństwa. Wartości, które można uzyskać dzięki hakowaniu, również rosną wraz ze wzrostem gospodarki opartej na danych. Inne trudne kwestie obejmują ustalenie odszkodowania dla ofiar awarii produktów danych (na przykład wypadki z samojezdnymi samochodami) oraz opracowanie wystarczających zachęt dla przedsiębiorstw, aby podjęły niezbędne kroki w celu zainwestowania w bezpieczeństwo danych. Jeśli dodamy do tego niepewność firm dotyczącą obciążeń regulacyjnych związanych z danymi oraz ryzyko sporów sądowych, widać, że w miarę postępów firmy i społeczeństwa czeka wiele pracy. Regulacja gospodarki opartej na danych jest ściśle związana z prywatnością danych. Obecne podejście to elastyczność, znalezienie równowagi między ochroną prywatności a umożliwieniem obywatelom samodzielnego decydowania. Jednym z filarów tych nowych ram regulacyjnych jest unijne rozporządzenie RODO. Potrzebny jest nowy paradygmat zarządzania danymi, w którym etyka danych będzie centralnym elementem wszystkich reform regulacyjnych.

Wykorzystywanie danych do wglądu lub okazji handlowych

Jeśli planujesz wykorzystywać analitykę danych głównie do umożliwienia operacji opartych na danych i podejmowania decyzji opartych na faktach oraz do stymulowania wewnętrznej wydajności, musisz zrozumieć swoje główne strategiczne obszary zainteresowania z perspektywy optymalizatora biznesowego. Jeśli z drugiej strony twoją główną ambicją jest wspieranie swojej oferty komercyjnej za pomocą analizy danych, istnieją inne strategiczne kwestie, o których należy pamiętać jako innowator lub osoba zakłócająca rynek. W tym rozdziale wyjaśniono strategiczne aspekty, które należy wziąć pod uwagę w zależności od celów biznesowych.

Koncentracja inwestycji w badania danych

Jeśli Ty i Twoja firma dopiero niedawno rozpoczęliście przygodę z nauką o danych i inteligencją maszynową, strategiczna decyzja o tym, jak chcesz skoncentrować swoje inwestycje, nie jest łatwym zadaniem. Jednak wybór właściwego punktu wyjścia jest kluczowy, ponieważ przez cały czas wiemy, że taka decyzja ma ogromne znaczenie! Co więcej, obszary nauki o danych i inteligencji maszynowej są dość złożone, niezwykle transformacyjne i nieustannie ewoluują, a nowe techniki i metodologie (w tym nowe rozwiązania techniczne do szybszych i bardziej wydajnych obliczeń i analiz) pojawiają się, jak się wydaje, każdego dnia. Biorąc pod uwagę poziom inwestycji w postaci pieniędzy i wysiłek wymagany do wprowadzenia zmian, skąd możesz mieć pewność, że wybór, którego dokonasz dzisiaj, będzie nadal ważny za kilka lat? Prosta odpowiedź: nie możesz. Nie oznacza to jednak, że powinieneś poczekać, aż dziedzina nauki o danych ustabilizuje się i dojrzeje. Jeśli to zrobisz, możesz być pewien, że Twoja konkurencja Cię ominęła. Zamiast tego skoncentruj swoją inwestycję na najbardziej elastycznej możliwej konfiguracji architektury danych — takiej, która w razie potrzeby pozwoli zmienić kierunek w zakresie ukierunkowania działalności i zakresu danych. Pomysł polega na tym, że każda wybrana konfiguracja powinna umożliwiać wymianę starych aplikacji i narzędzi uczenia maszynowego/sztucznej inteligencji, aby można było włączyć nowe. Na wczesnym etapie należy podjąć jedną podstawową decyzję — czy skoncentrować swoje wysiłki w zakresie analizy danych wewnętrznie (na skuteczności biznesowej i wydajności) czy zewnętrznie (na ofertach komercyjnych). Pamiętaj tylko, że głównym celem tej (kluczowej) decyzji jest pokierowanie i skoncentrowanie wysiłków, a nie decydowanie o przyszłym kierunku Twojej firmy raz na zawsze. Jeśli Twoją strategią jest „rozwój" i od samego początku rozłożenie inwestycji w analitykę danych zarówno pod kątem wydajności wewnętrznej, jak i możliwości komercyjnych, jest to oczywiście również opcja. Pamiętaj jednak, że skupienie się w równym stopniu na obu aspektach jako pierwszej rzeczy, którą robisz, nie jest łatwym zadaniem – nawet jeśli Twoja firma ma już podstawową wiedzę na temat nauki o danych. Jeśli Twoja firma jest nowa w obszarze nauki o danych, zdecydowanie zalecam porzucenie tego podejścia: zamiast tego zacznij od wewnętrznego skupienia, a następnie przejdź na zewnątrz. Umożliwi to Twojej firmie wykorzystanie stabilnej podstawy opartej na danych, która będzie stanowić podstawę Twojej oferty handlowej.

Określanie czynników wpływających na wewnętrzne analizy biznesowe

Nauka o danych i inteligencja maszynowa mają fundamentalny wpływ na przedsiębiorstwa i szybko stają się zasobami krytycznymi dla zróżnicowania rynku, a czasem nawet dla przetrwania firmy. Chodzi o to, aby Twoja firma była skoncentrowana na robieniu właściwych rzeczy (efektywność) i robieniu ich we właściwy sposób (wydajność).

Rozpoznawanie kategorii data science do praktycznego wdrożenia

Chociaż możesz mieć najlepsze intencje, jeśli chodzi o inwestycję w naukę danych, jeśli chodzi o ustanowienie w firmie podejścia opartego na danych, nie jest to łatwe zadanie na co dzień zarządzać biznesem opartym na danych . Nawet jeśli zbudowałeś najlepszą możliwą obsługę infrastruktury i masz

świetny zespół zajmujący się analizą danych, który kieruje Twoimi wysiłkami, nadal ciężko jest uzyskać wszystkie aspekty z praktycznego punktu widzenia. Jest tak wiele aspektów do rozważenia i tak wiele części firmy, które wymagają fundamentalnej zmiany. Jednocześnie nękają Cię pytania nie tylko o to, od czego zacząć, ale także o to, jak uniknąć zagubienia się we wszystkich zmianach. Cóż, możliwość szybkiego zaklasyfikowania każdego potencjalnego wpływu do jednej z pięciu kategorii, a następnie komunikowania potencjału każdej z nich, jest skutecznym sposobem pomagania liderom w osiąganiu lepszych wyników dzięki analizie danych. Na szczęście dla ciebie wykonałem tę robotę i mogę teraz opisać dla ciebie te pięć kategorii:

* Innowacje: ta kategoria dotyczy wspierania nowego myślenia i identyfikowania potencjalnych zakłóceń biznesowych i rynkowych w oparciu o analizę danych. Naukowcy zajmujący się danymi posiadają zdolność do przedstawiania złożonych problemów biznesowych jako problemów z uczeniem maszynowym lub badaniami operacyjnymi, co jest kluczem do znalezienia lepszych, bardziej zoptymalizowanych rozwiązań starych problemów. Mogą nawet ujawnić nowe problemy i podejścia, które były wcześniej nieznane.

* Eksploracja: ta kategoria odnosi się do sposobu eksplorowania nieznanych wzorców transformacyjnych w danych, a tym samym identyfikowania nieznanego potencjału biznesowego poprzez nieszablonowe myślenie. Analitycy danych powinni być zachęcani do podejmowania ekspedycji odkrywania danych, w przypadku których nie ma jasnych celów innych niż eksploracja danych pod kątem wcześniej nieodkrytej wartości. Kierowanie się danymi polega na kwestionowaniu starych sposobów rozwiązywania problemów poprzez pozbycie się świadomych lub nieświadomych uprzedzeń na temat tego, jak rzeczy działają i dlatego należy się nimi zająć. Eksploracja danych pozwala utorować drogę do nowych, bardziej zoptymalizowanych rozwiązań opartych na faktach.

* Eksperymentowanie: pozwalaj na swobodne eksperymentowanie i prototypowanie, kwestionując status quo za pomocą radykalnie nowych pomysłów i rozwiązań, a nie tylko wglądu w dane. Eksperymenty zwykle odbywają się na żywo, a nie w laboratorium. Wraz z dostępnością coraz większej ilości danych oraz ciągle zmieniającymi się potrzebami i oczekiwaniami klientów, podejmowanie decyzji przez ludzi staje się coraz bardziej nieadekwatne. Nauka o danych, a zwłaszcza uczenie maszynowe, doskonale sprawdza się w rozwiązywaniu bardzo złożonych, bogatych w dane problemów, które przytłaczają nawet najmądrzejszą osobę. Lista wyzwań biznesowych lub rządowych, z którymi nauka o danych może się zmierzyć, jest potencjalnie nieskończona. Biorąc tylko jeden przykład, wyobraź sobie znalezienie najbardziej zoptymalizowanego silnika do polecania produktów, którymi klient może być zainteresowany na podstawie wcześniejszych zachowań, zakupów, preferencji i profilu. Jak działałby taki silnik rekomendacji? Eksperymentując, będziesz używać różnych algorytmów uczenia maszynowego równolegle w ustawieniach na żywo, aby generować rekomendacje. Wszystkim algorytmom przypisano ten sam cel i cel, którym jest maksymalizacja konwersji – innymi słowy maksymalizacja prawdopodobieństwa, że klienci kupią coś w serwisie. Różni potencjalni nabywcy są narażeni na działanie różnych algorytmów, a po upływie określonego czasu lub osiągnięciu określonej liczby wyników w celu zapewnienia niezbędnej istotności statystycznej wynik jest analizowany i jeden z algorytmów konkurujących w eksperymencie zostaje wybrany zwycięzcą .

* Ulepszenie: Ta kategoria dotyczy ciągłego ulepszania istniejących procesów biznesowych i bieżącej oferty portfela. Ulepszanie jest prawdopodobnie najczęstszym zastosowaniem nauki o danych, ponieważ naukowcy zajmujący się danymi wielokrotnie tworzyli modele do udoskonalania wewnętrznych procesów i metodologii związanych z danymi gromadzonymi przez ich organizację. Typowymi przykładami są firmy marketingowe wykorzystujące segmentację klientów w kampaniach marketingowych, detaliści dopracowujący dynamiczne modele cenowe oraz banki dostosowujące swoje modele ryzyka finansowego. W wymiarze rozwoju produktu ulepszenie może obejmować

zwiększenie wydajności rozwoju i dystrybucji pod względem czasu i kosztów, ale może również oznaczać ulepszenie oferty usług przy wsparciu funkcji inteligencji maszyn i automatyzacji zadań maszynowych.

* Pożarnictwo: Pożarnictwo dotyczy tego, jak zidentyfikować czynniki powodujące zachowanie reaktywne - złe rzeczy, które się zdarzają. (Oczywiście wolisz skoncentrować swoje wysiłki na zachowaniu predykcyjnym, proaktywnym i zapobiegawczym, ale jeśli nie gasisz pożarów, Twoja firma może spłonąć.) Spójrzmy prawdzie w oczy: gaszenie pożarów jest czasem koniecznością. Kiedy coś poszło nie tak w Twoim systemie – na przykład spada rentowność biznesu lub klient ma pilną reklamację – musisz zareagować i odpowiedzieć tak szybko i skutecznie, jak to możliwe. Naukowcy zajmujący się danymi mogą nie tylko szybko znaleźć najlepsze rozwiązanie problemu, ale także pomóc określić, dlaczego problem wystąpił w pierwszej kolejności i spróbować zapobiec jego ponownemu wystąpieniu, wdrażając algorytmy do przewidywania i zapobiegania jego wystąpieniu.

Stosowanie wewnętrznych analiz biznesowych opartych na nauce danych

Kategoryzacja wyzwań to jedno, ale jak określić, które działania związane z nauką danych są najważniejsze? Jak następnie zastosować te działania, aby uzyskać rzeczywistą wartość biznesową z inwestycji w naukę danych? I wreszcie, jaką rolę w tym wszystkim odgrywa badacz danych? Najpierw odpowiem na ostatnie pytanie, pokazując, w jaki sposób naukowcy danych mogą zastosować praktyczne wewnętrzne wartości biznesowe w Twojej firmie. Oto wartości, które chcesz promować, a także porady dotyczące tego, jak naukowcy danych mogą je promować:

* Wzmocnij kierownictwo, aby podejmować lepsze decyzje. Przy odpowiednim podejściu doświadczony analityk danych może być postrzegany jako zaufany doradca i strategiczny partner wyższego kierownictwa organizacji. Analityk danych może komunikować się i demonstrować wartość danych firmy, aby usprawnić procesy podejmowania decyzji w całej organizacji, nie tylko jako samodzielne działania lub predefiniowane pulpity menedżerskie, ale także poprzez zintegrowanie zapotrzebowania na dane i wgląd w operacje operacyjne. model w firmie. Analityk danych ma możliwość skonfigurowania modelu w taki sposób, aby dane naprawdę stały się paliwem dla całej działalności firmy, stanowiąc podstawę każdej decyzji, działania i oceny.

* Poznaj możliwości i dowiedz się, jak stosować statystyki. Rolą data scientist jest również badanie i eksploracja danych organizacji, po czym można sformułować rekomendacje przepisywania pewnych działań, które pomogą poprawić wyniki firmy, lepsze zaangażowanie klientów, a ostatecznie zwiększyć rentowność. * Zapoznanie pracowników z użytecznością środowiska data science. Innym obowiązkiem analityka danych jest upewnienie się, że pracownicy są zaznajomieni (i poinformowani o) środowisku rozwoju i produkcji nauki o danych w organizacji w celu analizowania i identyfikowania wartości. Analitycy danych przygotowują organizację na sukces, wykazując wykorzystanie systemu do pozyskiwania spostrzeżeń i motywowania do działania. Gdy pracownicy zrozumieją możliwości środowiska nauki o danych, mogą skupić się na rozwiązywaniu kluczowych wyzwań biznesowych.

* Zidentyfikuj nowe możliwości. Istotną częścią tej roli jest kwestionowanie istniejących procesów i założeń w celu opracowania dodatkowych metod oraz modeli analitycznych i algorytmów, aby stale i stale poprawiać wartość wywodzącą się z danych organizacji.

* Promuj podejmowanie decyzji na podstawie wymiernych dowodów opartych na danych. Wraz z pojawieniem się naukowców zajmujących się danymi, możliwość zbierania i analizowania danych z różnych kanałów wykluczyła konieczność podejmowania wysokiego ryzyka. Analitycy danych mogą teraz tworzyć modele przy użyciu istniejących danych, które symulują różne potencjalne działania; w ten sposób organizacja może określić, która ścieżka zapewnia najlepsze możliwe wyniki biznesowe.

* Decyzje testowe. W ostatecznym rozrachunku wszystko sprowadza się do podjęcia pewnych decyzji (a nie innych), a następnie wprowadzenia zmian. Ale to nie koniec. Kluczowe jest poznanie wpływu podjętych i wdrożonych decyzji pod kątem ich rzeczywistego wpływu na organizację. Analitycy danych mogą pomóc organizacji zidentyfikować kluczowe wskaźniki związane z ważnymi zmianami i określić ilościowo ich sukces.

* Zidentyfikuj i doprecyzuj widok klienta. Większość firm ma co najmniej jedno źródło danych klientów, z którym można pracować, ale jeśli nie jest ono dobrze wykorzystywane, dane są prawie bezwartościowe. Jednym z ważnych aspektów nauki o danych jest możliwość łączenia istniejących danych, które niekoniecznie są przydatne same w sobie, z innymi punktami danych w celu wygenerowania wglądu, którego organizacja może użyć, aby dowiedzieć się więcej o swoich klientach i innych odbiorcach docelowych. Analityk danych może pomóc w precyzyjnej identyfikacji kluczowych grup poprzez analizę różnych źródeł danych. Dzięki tej dogłębnej wiedzy organizacje mogą dostosowywać usługi i produkty do grup klientów i pomagać rosnąć marżom.

* Rekrutuj odpowiedni talent. Czytanie CV przez cały dzień to codzienne, powtarzalne zadanie dla rekrutera, ale teraz zaczyna się to zmieniać ze względu na możliwość wykorzystania analityki danych również do tego typu zadań. Wydobywanie ogromnej ilości danych, które są już dostępne — na przykład życiorysy i aplikacje wewnętrzne, a nawet zaawansowane testy umiejętności i gry oparte na danych — może pomóc zespołowi rekrutacyjnemu w podejmowaniu szybszych i dokładniejszych decyzji o zatrudnieniu.

Wykorzystywanie danych do okazji handlowych

Kiedy chcesz wykonać strategiczny ruch biznesowy, musisz mieć odpowiednie uzasadnienie i motywację do swoich działań. Ponadto, jeśli naprawdę chcesz wykorzystać okazje, nie możesz sobie pozwolić na czekanie miesiącami na regularne oceny biznesowe. Nauka o danych daje właścicielom firm możliwość szybkiego i skutecznego podejmowania decyzji przy jednoczesnym unikaniu ryzyka. Żeby było jasne, wykorzystywanie danych do identyfikowania i wykorzystywania nowych możliwości handlowych ma niewiele wspólnego z podejmowaniem lepszych decyzji biznesowych dotyczących bieżących przedsięwzięć Twojej firmy. Zamiast tego odnosi się do tego, w jaki sposób można wykorzystać dane do identyfikowania, określania zakresu i inwestowania w nowe komercyjne inicjatywy biznesowe oparte na danych i produktach związanych z danymi. Innymi słowy, oznacza to inwestowanie w zupełnie nowe produkty i usługi, które mogą albo uzupełnić obecną linię biznesową, albo całkowicie zakłócić istniejący model biznesowy. To naprawdę zależy od zakresu Twoich ambicji oraz od tego, jakie możliwości znajdziesz i w które chcesz zainwestować.

Definiowanie produktu danych

Oto proste pytanie: Czym jest produkt oparty na danych? Innymi słowy, tak zwany produkt danych? Okazuje się, że odpowiedź na to pytanie nie jest łatwa. Podobnie jak w wielu dziedzinach nauki o danych, nie ma jasnej definicji produktu danych, chociaż – jeśli zmusi się do wymyślenia działającej definicji – powiedziałbym, że jest to produkt, który ułatwia osiągnięcie celu końcowego poprzez wykorzystanie danych. Być może ta definicja nie pomaga w zrozumieniu tego pojęcia, ponieważ na pierwszy rzut oka może wydawać się dość szeroka – odnosząc się do prawie wszystkiego. Chociaż z pewnością prawdą jest, że dostępnych jest wiele różnych rodzajów produktów danych, można je jednak podzielić tylko na dwie główne kategorie:

* Dane włączone: produkt zorientowany funkcjonalnie, który potrzebuje danych, aby spełnić swój cel - innymi słowy, wykorzystuje dane wejściowe, aby spełnić swój cel funkcjonalny. Jednym z przykładów jest sytuacja, w której dane są wykorzystywane do uzyskania predykcyjnego wglądu w

zautomatyzowane działanie systemu, które jest potrzebne, aby zapobiec wystąpieniu problemu (zautomatyzowane podejmowanie decyzji). Bez zasilania systemu danymi, system nie może analizować i działać z wyprzedzeniem, aby zapobiec problemowi.

* Czyste dane: Ten typ produktu składa się z danych i ma cel skoncentrowany na danych (nie jest zorientowany na funkcjonalność). Innymi słowy, generuje wgląd jako wynik końcowy, a nie jakąkolwiek zdolność do wykonania funkcjonalnego zadania. Produkt czystych danych może również odnosić się do sytuacji, w której sprzedajesz albo surowe dane, albo same przetworzone dane, albo inne powiązane spostrzeżenia pochodzące z danych zebranych w raporcie lub zestawie zaleceń. Warto trochę zastanowić się nad różnicami w definicji między produktami danych a innymi produktami technologicznymi. Różne typy są ogólnie definiowane przez różne cechy i dlatego należy traktować je w różny sposób z perspektywy strategicznej. Chociaż wiele standardowych zasad opracowywania produktów ma zastosowanie zarówno do produktów technologicznych, jak i produktów danych – na przykład zaspokajanie potrzeb klientów, uczenie się na podstawie informacji zwrotnych lub ustalanie priorytetów wymagań – nadal istnieje wiele obszarów, w których te dwa rodzaje produktów znacznie się różnią. Ta lista wykorzystuje popularne produkty, aby określić niektóre z tych różnic:

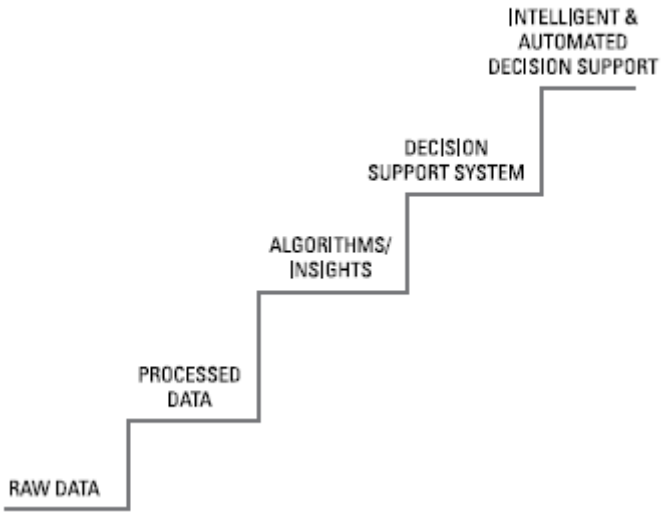
* Gmail: czy Gmail jest usługą transmisji danych? Nie, Gmail to usługa poczty e-mail, której głównym celem jest umożliwienie pisemnej, cyfrowej komunikacji między osobami. Jednak sortowanie naszych wiadomości e-mail przez Gmaila jako ważne lub nieważne jest usługą związaną z danymi, ponieważ głównym celem jest sortowanie wiadomości e-mail na podstawie zawartości i trafności danych, a nie ich funkcjonalności.

*Google Analytics: czy Google Analytics to usługa związana z danymi? Tak to jest. Głównym celem jest wyjaśnienie użytkownikowi ilościowego zrozumienia zachowań online. W tym przypadku dane mają kluczowe znaczenie dla interakcji z użytkownikiem i, w przeciwieństwie do innych produktów technologicznych skoncentrowanych na wyniku funkcjonalnym, ich głównym celem jest uzyskanie wglądu w te dane.

* Instagram: czy Instagram jest produktem danych? Nie, ale jeśli podzielisz Instagram na różne produkty, niektóre części (na przykład tagowanie danych, wyszukiwanie i odkrywanie) można uznać za produkty z danymi.

Rozróżnianie kategorii produktów danych

Wcześniej w tym rozdziale omówię rozróżnienie między produktami obsługującymi dane a produktami opartymi na danych, ale istnieją inne, bardziej szczegółowe sposoby podziału tortu produktów danych. Jednym ze sposobów jest posortowanie produktów na pięć głównych kategorii: dane surowe, dane przetworzone, algorytmy/wglądy, wspomaganie decyzji i automatyczne decyzje-zrobienie. Rysunek przedstawia graficzną reprezentację asortymentu, a poniższa lista opisuje niektóre szczegóły dla każdego typu produktu:



* Surowe dane: termin ten odnosi się do zbierania danych i udostępniania ich w takiej postaci, w jakiej są (lub może z kilkoma małymi etapami przetwarzania lub oczyszczania). Użytkownik może następnie wybrać odpowiednie wykorzystanie danych, chociaż większość pracy zostanie wykonana po stronie klienta lub użytkownika.

* Przetworzone dane: Przetworzone dane są o jeden krok w stosunku do nieprzetworzonych danych, co oznacza, że podczas konwersji nieprzetworzonych danych do formatu, który można następnie analizować i wizualizować w celu zapewnienia wglądu użytkownikowi produktu danych, nastąpiło pewne czyszczenie i transformacja danych lub zamierzonego klienta. W przypadku danych o klientach można dodać dodatkowe atrybuty dla dodatkowej wartości - np. przypisanie segmentu klienta do każdego klienta lub obliczenie prawdopodobieństwa kliknięcia przez klienta w reklamę lub zakupu produktu z określonej kategorii.

* Algorytmy/informacje: produkty danych związane z modelami i algorytmami lub algorytmy jako usługa to najnowsze rodzaje ofert produktów danych cyfrowych. Tutaj algorytm działa na niektórych danych – czasami w kontekście uczenia maszynowego, czasami nie – a wynikiem są nowe informacje lub spostrzeżenia. Przykładem jest algorytm używany w Google Images. Gdy użytkownik przesyła zdjęcie, otrzymuje zestaw obrazów, które są takie same lub podobne do przesłanego. Za kulisami produkt wyodrębnia najważniejsze cechy obrazu, klasyfikuje je i dopasowuje do zapisanych obrazów, zwracając te, które są najbardziej podobne. Statystyki są dodawane do tej samej kategorii, ponieważ produktem danych może być czasami sam wgląd, a nie algorytm, który wygenerował wgląd. Typowy nabywca wglądu (za pośrednictwem raportu lub modelu wglądu jako usługi) jest osobą nietechniczną. Typowy wgląd związany z Grafika Google można uzyskać, używając uczenia maszynowego do porównywania setek zdjęć Twojego produktu, aby na przykład wykryć pewne wzorce preferencji klientów podczas korzystania z Twoich produktów. Wgląd w to, jak klienci wolą korzystać z Twojego produktu, może być następnie wykorzystany w przyszłych kampaniach marketingowych lub jako wkład do modelowania przyszłych ulepszeń produktu.

* System wspomagania decyzji: Ta kategoria dostarcza użytkownikowi informacji w celu wsparcia procesu decyzyjnego, chociaż ostateczne decyzje są nadal podejmowane przez użytkownika. Produkty analityczne, takie jak Google Analytics, Flurry i SAS Visual Analytics to przykłady, które należą do tej kategorii. Stworzenie systemu wspomagania decyzji wymaga wiele wysiłku i oczekuje się, że wykona on większość pracy z zamiarem dostarczenia użytkownikowi istotnych informacji w łatwym do

przyswojenia formacie - pulpity nawigacyjne, aby umożliwić użytkownikom podejmowanie lepszych decyzji , na przykład. Podczas korzystania z tych narzędzi analitycznych zdobyte spostrzeżenia mogą prowadzić do zmian w strategii redakcyjnej, planów usuwania wycieków w lejku konwersji lub podwojenia strategii danego produktu. Ważną rzeczą do zapamiętania w przypadku tego typu produktu danych jest to, że chociaż produkt zebrał dane, skompilował dane i wyświetlił dane, nadal oczekuje się, że użytkownik zinterpretuje dane. Użytkownicy mają kontrolę nad decyzją o działaniu (lub niepodejmowaniu działania) na tych danych.

* Inteligentne i zautomatyzowane wspomaganie decyzji: w tego typu produkcie danych uwzględniana jest cała inteligencja w danej domenie, co oznacza, że produkt może działać samodzielnie, a także zapewniać i działać na podstawie informacji w ramach konkretnego produktu danych. Jednym z przykładów są rekomendacje produktów Netflix, gdzie dane z poprzednich preferencji użytkowników dotyczące seriali i filmów są wykorzystywane przez algorytm do rekomendowania nowych selekcji. Ponieważ wynik decyzji o tym, co wyświetlić, jest przechwytywany w tym samym środowisku (aplikacja Netflix) model może uchwycić każdy wybór i nauczyć się dawać jeszcze bardziej precyzyjne i inteligentne zalecenia w przyszłości. Inne przykłady bardziej złożonych produktów danych w tej kategorii obejmują automatyzację w pętli zamkniętej z przykładami, takimi jak samochody autonomiczne i zautomatyzowane drony. Ten rodzaj produktu danych umożliwia algorytmowi wykonanie zadania, uczenie się na podstawie danych i działań, a następnie przedstawienie użytkownikowi końcowego wyniku. Wynik czasami zawiera wyjaśnienie, dlaczego sztuczna inteligencja wybrała tę opcję; innym razem proces podejmowania decyzji jest całkowicie ukryty.

Główną różnicą między tymi kategoriami jest wbudowany poziom złożoności. Dokładniej, kategorie na rysunku są klasyfikowane pod względem ich rosnącej wewnętrznej złożoności i (powinny mieć) mniejszą złożoność po stronie użytkownika. Na przykład, chociaż surowe dane mają niewielką wbudowaną złożoność na początku, wymagają złożonych technik i umiejętności, aby opracować produkt, który generuje wartość z nieprzetworzonych danych. Z drugiej strony, dzięki produktowi opartemu na złożonym algorytmie uczenia maszynowego otrzymujesz prosty interfejs użytkownika dla klienta, który wymaga mniej myślenia. Produkt danych zarządza złożonością wewnętrznie poprzez algorytm uczenia maszynowego. Zazwyczaj (ale nie tylko) surowe dane, przetworzone dane i algorytmy koncentrują się na użytkownikach technicznych. Wglądy, wspomaganie decyzji i produkty zautomatyzowanego podejmowania decyzji mają zwykle bardziej zrównoważoną mieszankę użytkowników technicznych i nietechnicznych.

Równoważenie celów strategicznych

Załóżmy, że jesteś właścicielem firmy obuwniczej. Czy nie byłoby przydatne, gdybyś dowiedział się, że segment Twojego rynku docelowego woli kupować buty do biegania w czerwcu? Czy nie byłoby bardziej przydatne, gdybyś dowiedział się, że segment należy do grupy wiekowej 16-21, że preferują buty do biegania po drogach od butów do biegania w terenie, że mogą sobie pozwolić na wydanie 100 USD na parę butów i że kochasz kolory niebieski i czerwony? Nie zawsze musisz sprzedawać dane, aby na nich zarabiać. Dzięki stale rosnącej popularności Internetu dołączenie małego chipa do każdego produktu (na przykład butów) oraz śledzenie użytkowania i innych szczegółów jest teraz prostsze niż kiedykolwiek. Nie oznacza to, że Twoja firma musi słuchać prywatnych rozmów klientów lub śledzić wszystko, co robią Twoi klienci. Musisz tylko śledzić te dane, które Twoim zdaniem są korzystne dla Twojej firmy. Dane te mogą jednak obejmować zarówno dane niewrażliwe, jak i wrażliwe (na przykład korzystanie z produktów, zainteresowania i preferencje lub aktywność w Internecie), a nawet zainteresowania znajomych klientów oraz dzienniki SMS-ów i połączeń. Dlatego musisz mieć dobrą strategię postępowania z danymi w sposób etyczny i zgodny z prawem. W związku z tym zastosowanie danych jako danych wejściowych lub wyjściowych w firmie jest ważną decyzją strategiczną, która

wymaga pewnego rozważenia. Poświęcenie czasu na zrozumienie aktualnych trendów na rynku dla Twojej branży jest ważne, ale tak samo jest upewnienie się, że Twoje ambicje są wykonalne. Jeśli Twoja branża jest bardziej tradycyjna i nie jest jeszcze zdigitalizowana i nadal nie jest oparta na danych, być może najlepszym sposobem na rozpoczęcie jest wewnętrzne skupienie się na odwróceniu sytuacji, zanim spróbujesz zakłócić rynek nowymi produktami danych.

Inne podejście do klientów

Ponieważ ludzie na całym świecie żyją teraz w wieku klienta, nadszedł czas, aby wyjaśnić, co naprawdę oznacza pojęcie zarządzania doświadczeniami klientów (CEM). Może jednak pomóc najpierw zobaczyć, czym nie jest. CEM nie polega na zbieraniu informacji zwrotnych, odpowiadaniu na opinie lub śledzeniu sprawdzonego wskaźnika lojalności klientów, jakim jest Twój wynik Net Promoter Score. Żadne z tych działań nie reprezentuje indywidualnie CEM. Zamiast tego można powiedzieć, że CEM odnosi się do pełnej filozofii i metodologii, które sprawiają, że praca z Twoją firmą jest przyjemna dla klientów. W tym rozdziale chcę pokazać, jak skuteczna strategia danych może prowadzić do lepszego i bardziej wnikliwego podejścia do klientów.

Zrozumieć swoich klientów

Optymalizacja obsługi klienta to świetny sposób na przyciągnięcie nowych klientów, ale jest to również jeden z najlepszych sposobów na budowanie lojalności klientów, aby zachować tych, których już masz. Pomimo tej korzyści marketerzy i inni liderzy organizacyjni często zaniedbują klienta przed i po sprzedaży. Największą przeszkodą, aby nawet zacząć odwracać się od tej szkodliwej praktyki, jest zwykle brak głębokiego zrozumienia klienta.

Pełne zrozumienie klientów jest kluczem do osiągnięcia podstawowych celów biznesowych, niezależnie od tego, czy próbujesz budować (lub optymalizować) doświadczenie klienta, tworzyć bardziej angażujące treści, czy zwiększać sprzedaż. Aby zobaczyć, w jaki sposób możesz lepiej zrozumieć swoich klientów, radzę przyjrzeć się kilku kluczowym czynnościom, które musisz wykonać, aby naprawdę poznać swoich klientów. W następnych kilku sekcjach znajdziesz drogę.

Krok 1: Zaangażuj swoich klientów

Zoptymalizowane doświadczenie klienta jest oczywiście cenne dla przychodów i utrzymania, ale jeśli zrobisz to dobrze, może być również doskonałym źródłem wiedzy o klientach. Kontaktowanie się z klientami w czasie rzeczywistym stało się łatwiej dostępne dzięki wielu nowym narzędziom. Messenger staje się coraz bardziej popularnym kanałem obsługi klienta, a narzędzia takie jak Drift pozwalają rozmawiać z klientami podczas przeglądania Twojej witryny. Drift jest szczególnie ekscytujący, ponieważ działa jako zupełnie nowy sposób podejścia do obsługi klienta w czasie rzeczywistym w formacie konwersacyjnym w porównaniu z tradycyjnym zaangażowaniem klienta. Tak więc prędkość tutaj jest zdecydowanie plusem, ale jest wiele innych plusów. Chociaż kanały takie jak Messenger i Drift są zdecydowanie świetnymi sposobami gromadzenia informacji o klientach, nie są one często wykorzystywane w najbardziej efektywny sposób. Jeśli twoje zaangażowanie jest doraźne lub fragmentaryczne, nie wykorzystujesz prawdziwej mocy tych kanałów; takie zaangażowanie musi być częścią większego planu. Oznacza to, że firmy i organizacje muszą przewidywać, aby zainwestować czas i pieniądze potrzebne do zrozumienia całej podróży klienta. Nie wystarczy zająć się jednym punktem, aby przeprowadzić ankietę i zrozumieć klienta; bez szerszego kontekstu te wyrywkowe kontrole mogą być gorsze niż bezużyteczne. Nie będziesz w stanie odpowiedzieć na podstawowe pytania, takie jak jak klient dotarł do tego punktu, czego szukał lub dokąd zmierza w całej podróży, ponieważ nie masz informacji, których potrzebujesz, aby znaleźć świadomą odpowiedź. Jeśli zainwestujesz czas potrzebny na zrozumienie całej podróży klienta, będziesz w stanie określić, w jaki sposób Twoi klienci odbierają Twoją markę w trakcie ich relacji z tą marką. Ten kontekst pozwoli Ci zadawać klientom właściwe pytania we właściwym czasie, budując w ten sposób zaangażowanie marki, a także rodzaj zaufania klienta potrzebnego do poprowadzenia podróży do punktu zakupu. Pracując nad utrzymaniem zaangażowania klientów na pierwszych etapach podróży klienta, pomyśl o swojej relacji jako o ulicy dwukierunkowej. Zachęcaj klientów do dzielenia się swoimi przemyśleniami i opiniami, regularnie

umieszczając ankietę zadowolenia klientów w zwykłych wiadomościach e-mail. Podczas projektowania ankiety przestrzegaj tych trzech zasad:

* Usuń stronniczość. Zapytaj klientów o ich opinie bez projektowania własnej. Uzyskaj ich niezakłócone, bezstronne opinie. Chcesz prawdziwych spostrzeżeń, nawet jeśli są one negatywne. Przykładem może być coś tak prostego jak: „Jak myślisz, co moglibyśmy zrobić lepiej?"

* Bądź konkretny. Używaj prostego języka, który prosi o opinie na określony temat. Na przykład pytanie „Jak poprawiłeś swoją skuteczność marketingową za pomocą naszego algorytmu uczenia maszynowego?" pomoże określić wartość, jaką otrzymują od Ciebie Twoi klienci.

* Skupiać. Twoje ankiety powinny dotyczyć tylko jednego obszaru doświadczenia klienta. Celem jest uzyskanie spostrzeżeń, na podstawie których możesz działać.

Pamiętaj o tych zasadach, personalizując ankietę klienta za pomocą pytań dotyczących Twojej marki i produktu lub usługi.

Krok 2: Zidentyfikuj, co kieruje Twoimi klientami

Wielu marketerów popełnia błąd polegający na używaniu ogólnych danych demograficznych – takich jak wiek, zawód i lokalizacja – aby wyrobić sobie poczucie zakresu punktów danych ich bazy klientów, po prostu nie dostarczają wystarczającej ilości informacji do tworzenia wiadomości, które przemawiają do klientów na poziomie emocjonalnym . Jednym ze sposobów zagłębienia się w preferencje klientów jest skorzystanie z karty Pozyskiwanie w Google Analytics, aby sprawdzić, z jakich mediów społecznościowych, blogów branżowych i forów zawodowych pochodzi ruch w Twojej witrynie. Następnie zastosuj te informacje do swoich tożsamości, abyś mógł dowiedzieć się, gdzie i kiedy skuteczniej do nich dotrzeć. Pozyskiwanie danych o słowach kluczowych to kolejny pomocny sposób na poznanie terminów i opisów używanych przez niektórych tożsamości kupujących do opisywania Twoich usług. Aby na przykład podzielić klientów na podstawie wyszukiwań słów kluczowych, najpierw użyj Narzędzi Google dla webmasterów, aby utworzyć listę popularnych słów kluczowych, które przyciągają użytkowników do Twojej witryny, pogrupuj je w nadrzędne tematy, a następnie przypisz je do różnych kategorii klientów na podstawie danych, które mieć dostępne. Aby włożyć ten wysiłek w czyn, uwzględnij te słowa kluczowe w swojej witrynie, a następnie mapuj działania content marketingowe i inne interakcje online na nowe kategorie klientów w oparciu o to, co przyciąga różne typy kupujących. Uważność na preferencje klientów i mówienie tym samym językiem, co Twoi klienci, to subtelny sposób na sprawienie, aby Twoi obecni odbiorcy poczuli się bardziej mile widziani.

Krok 3: Zastosuj analizy i uczenie maszynowe do działań klientów

Od kliknięcia linku po przeczytanie strony internetowej, każde działanie klienta zapewnia cenny wgląd w zachowanie klienta. Aby określić, w jaki sposób klienci wchodzą w interakcję z Twoją witryną, możesz wypróbować narzędzia do śledzenia zachowań użytkowników, takie jak Google Analytics i Inspectlet. Są świetnymi narzędziami do zbierania informacji, takich jak czas spędzony na stronie i współczynnik odrzuceń. Inspectlet może nawet wyświetlać krótkie filmy wideo użytkowników na Twojej stronie w czasie rzeczywistym. Zebrane dane behawioralne powinny prowadzić do wniosków na temat tego, czego Twoi odbiorcy nie rozumieją, co im się podoba, a czego nie, oraz jak możesz stworzyć silniejszą witrynę. Jeśli na przykład ludzie mieli problemy z nawigacją do określonej strony sprzedaży, możesz dostosować interfejs, aby zapewnić bardziej przyjazne dla użytkownika wrażenia. Jeśli ludzie spędzają więcej czasu na jednej stronie niż na innych, przeanalizuj zawartość tej strony, aby zobaczyć, co może wymagać dodatkowej uwagi. Na przykład, jeśli ludzie spędzają zbyt dużo czasu na stronie kasy, być może nadszedł czas, aby poprawić obsługę płatności klientów w Twojej witrynie. Co jednak

najważniejsze, jeśli masz stronę o wysokim współczynniku odrzuceń, spróbuj zobaczyć, co sprawia, że ludzie odchodzą. Systemy rekomendacji po raz pierwszy stały się popularne w branży detalicznej, głównie w handlu internetowym lub e-commerce w celu spersonalizowanych rekomendacji produktów. Jednym z najczęstszych zastosowań jest sekcja Amazona „Klient, który kupił ten przedmiot, kupił również . . .". Systemy rekomendujące mogą być postrzegane jako inteligentny i wyrafinowany sprzedawca, który zna gust i styl klienta, dzięki czemu może podejmować bardziej inteligentne decyzje dotyczące tego, jakie rekomendacje przyniosą klientowi korzyści. Choć zaczynał w e-commerce, teraz zyskuje popularność w innych sektorach, zwłaszcza w mediach. Niektóre przykłady to „polecane filmy" z YouTube lub „inne filmy, które mogą Ci się spodobać" z serwisu Netflix. Inne branże również dostrzegają wartość stosowania systemów rekomendujących. (Branża transportowa jest jednym z przykładów.)

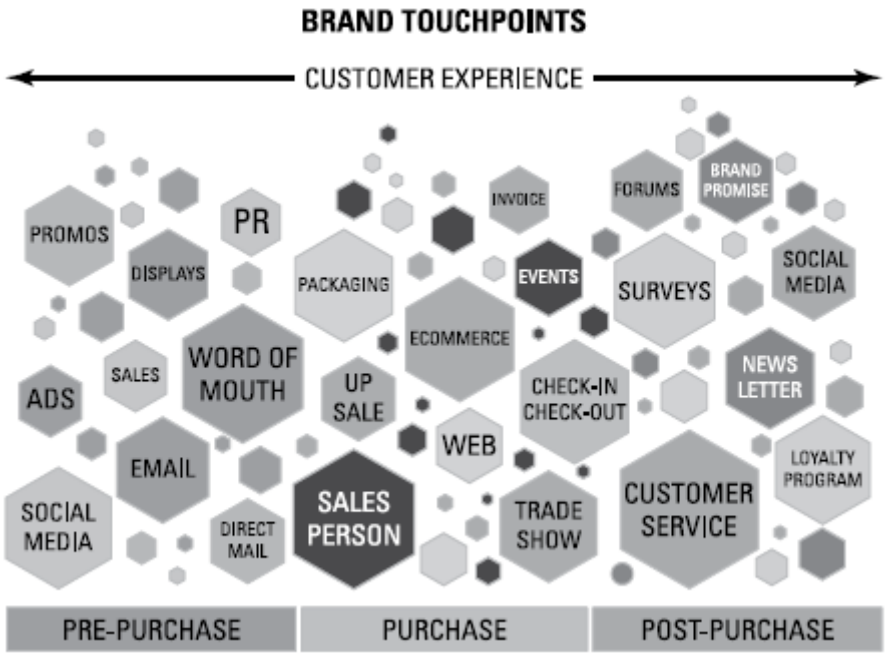
Krok 4: Przewiduj i przygotuj się na następny krok

Tworzenie planu przyszłego zaangażowania klienta jest tak samo ważne, jak tworzenie planu na teraźniejszość. Dzięki temu zespoły zajmujące się obsługą klienta są w odpowiednim nastroju, aby odpowiadać klientom w stresujących lub trudnych sytuacjach. Oprogramowanie do modelowania predykcyjnego przeszukuje istniejące dane klientów, aby zidentyfikować cykliczne wzorce i trendy, które mogą wpływać na proces podejmowania decyzji. Dwoma świetnymi narzędziami do tych zadań są niestandardowe programy analityczne od RapidMiner i Angoss, które tworzą realistyczne modele przyszłości. Aby zobaczyć, jak modelowanie predykcyjne wpływa na strategię klienta, wyobraź sobie, że pracujesz dla firmy SaaS, która chce dostosować swoją mapę drogową produktu do przewidywania potrzeb klientów. Patrząc na historyczne dane behawioralne, możesz zobaczyć, które funkcje klienci uznali z czasem za najbardziej wartościowe, a których nie używali. Zrozumienie najpopularniejszych i najczęściej odwiedzanych stron może również wpłynąć na strategię dotyczącą treści, skupiając się na tematach i formatach, które najlepiej rozwiążą problemy Twoich odbiorców. Zidentyfikuj podobieństwa w najczęściej używanych funkcjach, aby określić, dlaczego klienci je polubili. Dodatkowo, przyglądanie się trendom i analizom rynku daje dobre wyobrażenie o tym, co inne firmy w Twojej przestrzeni już osiągnęły, dzięki czemu możesz opracować nowe funkcje, które eksplorują te obszary. Wiele firm zwraca się do firm zajmujących się badaniem rynku tylko jako forma ubezpieczenia – innymi słowy sposób na zmniejszenie ryzyka biznesowego związanego z inwestowaniem w niewłaściwy produkt lub usługę. Jednak badania rynkowe mogą być wykorzystywane w rozwoju produktów nie tylko jako ubezpieczenie, ale także jako narzędzie do ustalania potrzeb rynku i lepszego zrozumienia potencjału rynku. Ciągłe badania rynku na mapie drogowej produktu naturalnie prowadzą do większej sprzedaży. Im lepiej rozumiesz swój rynek, tym lepiej pasujesz do produktu/rynku.

Krok 5: Wyobraź sobie przyszłość swojego klienta

Jedynym sposobem na zrozumienie wyjątkowej i dynamicznej ścieżki zakupowej klienta jest postawienie się na jego miejscu. Jest to możliwe dzięki zaawansowanej technice zwanej mapowaniem podróży klienta, metodzie, w której firmy tworzą szczegółową, graficzną reprezentację podróży klienta w oparciu o krytyczne punkty kontaktu. Te punkty styku to interakcje między klientem a Twoją marką przed, w trakcie lub po zakupie. Współczesne interakcje z klientami obejmują szeroki zakres punktów styku: mobilne, internetowe, społecznościowe, interaktywne odpowiedzi głosowe (IVR), sklepowe, chatboty i inne. To prowadzi mnie do koncepcji omnichannel, wielokanałowej strategii treści, której organizacje używają do poprawy jakości obsługi klienta. Obecnie klienci często przełączają się między wieloma różnymi kanałami – powiedzmy w trakcie zakupu lub nawet podczas odkrywania. Analizy trendów zachowań klientów w różnych branżach pokazują, że omnichannel będzie się powiększał wraz ze wzrostem i różnorodnością kanałów; dlatego mapowanie ścieżki klienta musi obejmować każdy punkt kontaktu i kanał, w którym obecni są Twoi klienci. Rysunek ilustruje szeroką gamę możliwych

punktów kontaktu, których możesz użyć, aby dotrzeć do klientów, a także ich różnice w zależności od tego, czy jest to przed zakupem, w trakcie zakupu czy po zakupie. Zbieranie danych ze wszystkich trzech faz pozwala lepiej zrozumieć, co skłania klienta do zakupu – i, miejmy nadzieję, do ponownego zakupu.



Pierwszym warunkiem bycia omnichannel jest wielokanałowość: musisz być dostępny wszędzie tam, gdzie są Twoi klienci. Jednak posiadanie i uruchamianie tych kanałów to jedno; Sprawienie, że będą ze sobą płynnie współpracować w ramach całej podróży, to kolejna rzecz. Sprawdź Ubera jako przykład, jak zdefiniować punkty styku i jak je zastosować do mapowania ścieżki klienta. Drobne punkty kontaktu obejmują czynności, takie jak pobieranie aplikacji lub po prostu śledzenie aplikacji w mediach społecznościowych. Główne punkty styku obejmują takie rzeczy, jak faktyczne zamówienie przejazdu lub ukończenie szkolenia kierowcy. Kiedy już zdefiniujesz punkty kontaktu, musisz zbadać okoliczności wpływające na każdy z nich. Na przykład marketer w Uber może zapytać: „Co skłoniło pasażera do pobrania aplikacji po raz pierwszy? Czy było to związane z programem poleceń klientów Ubera?" Zespół wewnętrzny powinien być zaangażowany w te kwestie, aby uzyskać wszechstronną perspektywę i promować wspólne rozwiązywanie problemów. Po zidentyfikowaniu nieudanych punktów kontaktu (na przykład, gdy klient nie korzysta z pobranej aplikacji Uber), musisz ustalić plan kontaktu z tymi klientami. Tworzenie kamieni milowych może być dobrym pomysłem, na przykład gdy użytkownik aplikacji nie zalogował się na konto od trzech miesięcy lub gdy zapalony klient nagle przestaje korzystać z produktu. Najlepiej, jeśli zespół obsługi klienta może zadzwonić, napisać lub spotkać się bezpośrednio z klientami, aby zrozumieć, dlaczego są niezaangażowani. Jeśli te zasoby nie są dostępne, możesz utworzyć wiadomość e-mail marketingową skoncentrowaną na ponownym zaangażowaniu klientów w oparciu o określone kamienie milowe.

Uszczęśliwianie klientów

Pierwszą rzeczą, którą musisz się zająć, gdy dążysz do zadowolenia klientów poprzez poprawę ich zadowolenia, jest lepsze zrozumienie tego, czym tak naprawdę jest obecny wskaźnik odpływu klientów. W ten sposób możesz skoncentrować swoje początkowe wysiłki na grupie klientów, którzy z największym prawdopodobieństwem odejdą lub są już w drodze do odejścia. Aby zmniejszyć utratę klientów (określaną również jako churn), należy używać historycznych danych klientów do mapowania migawek klientów wykonanych w danym momencie. Takie migawki rejestrowałyby na przykład, kim

są, co kupili i jak wchodzili w interakcję ze sprzedawanymi im produktami i/lub usługami. Powinieneś zmapować te informacje na to, czy później odeszli (innymi słowy, przestali być klientem). Następnie możesz przeanalizować każdego obecnego klienta, który według Ciebie prawdopodobnie odejdzie, a następnie ocenić, jak cenny jest ten klient. Na koniec możesz określić, jakie działania należy podjąć, aby zapobiec odpływowi najcenniejszych klientów. Oto kilka działań, które możesz podjąć, aby uniemożliwić klientom odejście:

* Ulepsz kampanie marketingowe poprzez sprzedaż krzyżową produktów. Upewnij się, że Twoje kampanie marketingowe rzeczywiście wykorzystują właściwy przekaz, aby dotrzeć do odpowiedniej grupy klientów. Nie ma nic bardziej irytującego niż bycie przedmiotem kampanii marketingowej, która oferuje coś, czym absolutnie nie jesteś zainteresowany. To naprawdę sprawia, że zastanawiasz się, czy firma w ogóle zna swoich klientów. Unikanie tego zakłopotania najlepiej rozwiązać poprzez sprzedaż krzyżową produktów. Zacznij od zmapowania par klient/produkt do wskaźników zakupu zapisanych w danych historycznych. Dzięki temu będziesz wiedział, do kogo kierować reklamy, wprowadzając nowy produkt lub promując już istniejący.

* Zoptymalizuj produkty i ceny. Nawet jeśli jesteś w stanie zaoferować odpowiedni produkt lub usługę właściwemu klientowi, ważne jest również, aby oferować go w odpowiedniej cenie. Aby dowiedzieć się, jaka cena jest odpowiednia dla danego produktu, musisz zmapować charakterystykę produktu i cenę do liczby sprzedaży. Następnie możesz zmienić cenę i inne cechy, aby zobaczyć, jak wpływają one na przychody (cena × liczba obrotów). Daje to dobre zrozumienie najbardziej zoptymalizowanego poziomu cen.

* Zwiększ zaangażowanie klientów. Na koniec ważne jest, aby dowiedzieć się więcej o tym, jak zwiększyć zaangażowanie w kontakt z klientem. Czym tak naprawdę interesuje się klient? Możesz dowiedzieć się więcej na ten temat, obserwując zachowanie klientów, gdy przedstawia się im różne produkty lub usługi. Jest to potrzebne w celu odwzorowania par klient/produkt na wskaźniki zainteresowania klienta. Umożliwia to przewidywanie potrzeb i zainteresowań oraz uwzględnienie ich przy ocenie usługi świadczonej klientowi.

Skąd więc wiesz, że klient jest dobrze obsługiwany? Mówiąc najprościej, dobra obsługa klientów oznacza, że musisz albo oferować swoim klientom produkty, którymi są zainteresowani i na które mogą sobie pozwolić, albo świadczyć usługi, z którymi się angażują. Aby Twoi klienci byli zadowoleni i w pełni usatysfakcjonowani, ważne jest, aby każda część Twojej firmy współpracowała ze sobą. Zadowolenie klientów zależy nie tylko od jakości obsługi klienta, ale także od innych działów (na przykład tych odpowiedzialnych za produkcję). Tylko wtedy, gdy wszystkie koła zębate w Twojej firmie są dobrze naoliwione i ściśle połączone, możesz oczekiwać najlepszych wyników. Nawet jeśli wdrożysz wszystkie środki, jakie możesz wymyślić, aby Twoja firma zorientowana na klienta, zawsze możesz zrobić lepiej. Ta sama zasada dotyczy obsługi klienta. Ustalenie niektórych celów biznesowych, związanych z obsługą klienta, a także kluczowych wskaźników wydajności (KPI), pomoże Ci śledzić wszystkie wysiłki, aby zwiększyć satysfakcję klientów. Jeśli Twój klient zgłasza problem w mniej preferowanym kanale komunikacji – na przykład Facebooku – nie zmuszaj go do skorzystania z wybranego przez Ciebie kanału w jego miejsce. Jeśli klient skontaktował się z Tobą w mediach społecznościowych, to dlatego, że był to dla niej najwygodniejszy sposób komunikacji. Zamiast tego powinieneś zaoferować klientowi kilka różnych wyborów, a nie tylko jedno rozwiązanie. Pamiętaj również, aby informować klienta o tym, kiedy problem może zostać rozwiązany, zamiast czekać. Chodzi o to, by oferować swoim klientom te same usługi, których oczekiwałbyś od innych.

Bardziej wydajna obsługa klientów

Bardziej wydajna obsługa klientów odnosi się głównie do usprawnienia operacji w celu obniżenia kosztów. Ale oczywiście oznacza to również dobrą obsługę klientów - przewidywanie problemów, zanim się pojawią, lub poprawę obsługi zgłoszeń do obsługi klienta, gdy się pojawią. Nauka o danych może zwiększyć wydajność dzięki wykorzystaniu nadzorowanych technik uczenia maszynowego. Pomysł polega na mapowaniu sytuacji na wyniki, aby można było przewidzieć wyniki w nowych sytuacjach. Jednym z przykładów jest sytuacja, w której klient miał kontakt z nowym produktem: wynik tutaj jest definiowany przez to, czy klient wykaże zainteresowanie nowym produktem. W takich sytuacjach techniki uczenia maszynowego wymagają przykładów do pracy i szkolenia. Oznacza to, że potrzebujesz danych o sytuacjach (jak najdokładniej scharakteryzowanych wraz z wszelkimi informacjami kontekstowymi) oraz wyników zaobserwowanych w tych sytuacjach. Analiza przykładowych danych pozwala najpierw znaleźć wzorce, a następnie relacje między sytuacjami a wynikami. Prognozy dotyczące wyników są tworzone automatycznie przy użyciu tych relacji. Poniższe działania (przedstawione w ich kontekstach biznesowych) mogą pomóc w poprawie efektywności w zakresie zarządzania klientami poprzez podejście predykcyjne i prewencyjne.

Przewidywanie popytu

Przewidywanie popytu jest ważne dla firm, które obserwują dużą zmienność popytu na swoje usługi i/lub produkty – na przykład firmy, które sprzedają świeże towary i muszą unikać posiadania zbyt dużej lub małej ilości zapasów. Daje to korzyści w postaci możliwości pomiaru popytu i kontekstu, w którym to się stało, dzięki czemu można mapować kontekst do popytu. Możesz także wykorzystać zdobyte spostrzeżenia, aby określić, ile pracowników należy zatrudnić, przewidując, jak zajęty będzie biznes.

Automatyzacja zadań

Możesz zaoszczędzić czas, zlecając maszynom automatyczne wykonywanie pewnych powtarzalnych lub inteligentnych zadań. Czasami możesz już wykonywać te zadania za pomocą ręcznie opracowanych reguł, ale gdy zamiast tego wprowadzisz umiejętności uczenia maszynowego, aby uruchomić działanie, możesz wykorzystać nowy potencjał optymalizacji sposobu wykonywania działania. Wykorzystanie uczenia maszynowego do automatyzacji wykonywania zadania oznacza, że maszyna automatycznie uczy się reguł na podstawie przykładowych danych i z czasem może zoptymalizować sposób wykonywania zadania, w zależności od używanych technik. W tym kontekście oczywistym przykładem jest punktacja wniosków kredytowych lub roszczeń ubezpieczeniowych, w przypadku których oczekuje się, że coś zatwierdzisz lub odrzucisz. Innym przykładem jest automatyczne wykonywanie analiz ryzyka na podstawie danych historycznych z wykorzystaniem uczenia maszynowego. Ten daje lepszą podstawę do podejmowania decyzji przed zainwestowaniem czasu i pieniędzy w nowe projekty

Predykcja aplikacji firmowych

Możesz wiele zyskać, tworząc aplikacje używane przez pracowników związane z zarządzaniem relacjami z klientami, planowaniem zasobów przedsiębiorstwa, predykcją zasobów ludzkich. Dodając do tych aplikacji funkcje przewidywania, ludzie mogą wykonywać swoją pracę wydajniej. Oto niektóre z korzyści płynących z używania aplikacji predykcyjnych:

* Możesz priorytetyzować rzeczy. Predykcyjne aplikacje firmowe umożliwiają skierowanie uwagi użytkownika na to, co najważniejsze. Mogą to być na przykład e-maile (takie jak Google Priority Inbox), prośby o obsługę klienta (abyś mógł szybciej odpowiadać na najważniejsze) lub inne zewnętrzne prośby kierowane do Twojej firmy, na które musisz odpowiedzieć w trybie pilnym.

* Możesz lepiej dostosować przepływy pracy. Proaktywne podejście umożliwia korzystanie z adaptacyjnych przepływów pracy opartych na przewidywaniach, a nie na predefiniowanych ręcznych

regułach. Możesz na przykład kierować zgłoszenia do obsługi klienta do osób najlepiej przygotowanych do ich obsługi, w którym to przypadku wynikiem jest zespół obsługi klienta lub osoba.

* Możesz dostosować interfejs użytkownika. Możesz łatwo zwiększyć wydajność użytkownika, dostosowując interfejs, aby pokazać, czego potrzebują użytkownicy w momencie korzystania z aplikacji. Wszystko, co musisz zrobić, to zmapować kontekst do akcji, która będzie musiała zostać wykonana, aby wyzwolić korektę.

* Możesz zautomatyzować ustawienia użytkownika. Aplikacje predykcyjne umożliwiają automatyczne ustawianie konfiguracji i preferencji poprzez analizę danych o użytkowaniu aplikacji, a tym samym przyspieszają wydajność użytkownika.

Przedstawiamy modele biznesowe oparte na danych

Rosnące wykorzystanie danych zmienia sposób, w jaki firmy prowadzą działalność. Dzięki zaawansowanej analityce, uczeniu maszynowemu i dostępowi do nowych źródeł danych firmy z jednego sektora mogą odgrywać rolę w produktach i usługach innych — nawet tych, które są daleko od swojej tradycyjnej branży. To zaciera granice między branżami i zmienia dynamikę konkurencji. Firmy, które wykorzystują pełen zakres możliwości i równolegle z tymi zmianami przekształcają swoje modele biznesowe, znajdą nowe możliwości dla strumieni przychodów, klientów, produktów i usług. W tym rozdziale opiszę, jak można podejść do obszaru modeli biznesowych opartych na danych.

Definiowanie modeli biznesowych

Najpierw potrzebujesz działającej definicji modelu biznesowego. Jest ogólnie opisywany jako podstawa tego, jak organizacja tworzy, dostarcza i wychwytuje wartość w kontekście gospodarczym, społecznym, kulturowym lub innym. Proces budowy i modyfikacji modelu biznesowego, zwany także innowacją modelu biznesowego, jest częścią zwykłego opracowywania strategii biznesowej. Faktem jest jednak, że termin model biznesowy jest używany do szerokiego zakresu nieformalnych i formalnych opisów w celu wyjaśnienia podstawowych aspektów działalności, w tym celu, procesu biznesowego, klientów docelowych, ofert, strategii, infrastruktury, struktur organizacyjnych, zaopatrzenia, praktyki handlowe oraz procesy i polityki operacyjne, w tym kulturę firmy. Biorąc pod uwagę szerokie zastosowanie terminu „model biznesowy", zalecam jak najszersze zdefiniowanie go. Dla mnie oznacza to definiowanie modeli biznesowych po prostu jako projektowanie struktur organizacyjnych w celu wspierania możliwości komercyjnych. Modele biznesowe są wykorzystywane do opisywania i klasyfikowania przedsiębiorstw, zwłaszcza w środowisku przedsiębiorczym, ale są również wykorzystywane przez menedżerów w firmach do badania możliwości przyszłego rozwoju. Obecnie rodzaj modelu biznesowego, który jest potrzebny określonej firmie, może w rzeczywistości zależeć od tego, w jaki sposób wykorzystywana jest podstawowa technologia. Na przykład przedsiębiorcy w Internecie stworzyli również zupełnie nowe modele, które całkowicie zależą od istniejącej lub powstającej technologii. Korzystając z technologii, firmy mogą dotrzeć do dużej liczby klientów przy minimalnych kosztach. Ponadto rozwój outsourcingu i globalizacji oznacza, że modele biznesowe muszą również uwzględniać zaopatrzenie strategiczne, złożone łańcuchy dostaw i przechodzenie do opartych na współpracy, relacyjnych struktur kontraktowych. Jak można się spodziewać, projektowanie modelu biznesowego ogólnie odnosi się do działań, które koncentrują się na projektowaniu modelu biznesowego firmy. Jest to część procesu rozwoju biznesu i strategii biznesowej i obejmuje metody projektowania. Istnieje jednak duża różnica między zdefiniowaniem całkowicie nowego modelu biznesowego, którego nie ma, a zmianą istniejącego modelu biznesowego. W przypadku projektowania nowego modelu biznesowego, częstym wyzwaniem jest zwykle zrozumienie i przydzielenie potrzebnych zasobów w czasie. Jednak przy zmianie istniejącego modelu na nowy model biznesowy wyzwaniem jest raczej zarządzanie oporem lub brakiem zainteresowania ze strony pracowników oraz dostosowanie struktur organizacyjnych i produktowych do nowych sposobów rozwoju i sprzedaży. W zależności od wielkości i rozmieszczenia pracowników może to być trudne zadanie. Społeczności skoncentrowane na technologii czasami mają określone ramy modelowania biznesowego, które próbują zdefiniować to, co często może być trudnym podejściem do definiowania strumieni wartości biznesowej. (W pewnym momencie start-upy technologiczne muszą zacząć zarabiać, prawda?) Ramy modelu biznesowego stanowią kluczowy aspekt każdej firmy, dążąc do przedstawienia pełnego obrazu tego, jak firma wybiera swoich klientów, ale obejmują również sposób, w jaki firma definiuje i różnicuje swoją ofertę, definiuje zadania, które będzie wykonywać sam i te, które zleci, konfiguruje swoje zasoby, trafia na rynek i tworzy użyteczność dla klientów oraz przechwytuje zyski. Patrząc na ramy modelowania biznesowego, należy wziąć pod uwagę jeden

ostateczny pryzmat: czy skupienie się na czynnikach wewnętrznych, takich jak analiza rynku, promocja produktów/usług, rozwój zaufania, wpływ społeczny i dzielenie się wiedzą, czy też koncentracja bardziej na czynnikach zewnętrznych , jak konkurencja i aspekty technologiczne? Wydawałoby się, że wymagamy od ram modelowania biznesowego zbyt wiele, i prawdą jest, że zakres może być bardzo szeroki – czasami zbyt szeroki. A jednak, jeśli są używane prawidłowo, frameworki do modelowania biznesowego mogą być niezwykle użytecznymi narzędziami. Jednak w kontekście nauki o danych pojawiły się nowe ramy modelowania biznesowego - ramy modeli biznesowych opartych na danych. Zostaną one wyjaśnione i zilustrowane bardziej szczegółowo w dalszej części tego rozdziału, ale najpierw chcę wyjaśnić, czym właściwie jest model biznesowy oparty na danych.

Eksploracja modeli biznesowych opartych na danych

Zwiększone wykorzystanie danych w każdej dzisiejszej nowoczesnej firmie stanowi wyzwanie dla tradycyjnych sposobów dodawania wartości biznesowej i stanowi poważne ryzyko dla firm, które nie reagują odpowiednio. I oczywiście oferuje możliwości tym, którzy to robią. Firmy, które równolegle z tymi zmianami przekształcą swoje modele biznesowe, znajdą dla siebie nowe drzwi. Na przykład na rynku termostatów domowych, który jest tradycyjnie stosunkowo stabilnym sektorem z niewielką, ustaloną listą konkurentów, start-up o nazwie Nest był w stanie rzucić wyzwanie uznanym firmom, wprowadzając termostat, który wykorzystuje analitykę do uczenia się klientów preferencje poprzez analizę wzorców danych - wzorców, które są następnie wykorzystywane do budowania modelu, w jaki sposób termostat powinien się odpowiednio dostosować. Przykład tego, jak nowatorski, oparty na danych model biznesowy firmy Nest umożliwił jej wejście na rynek od dawna zamknięty dla osób z zewnątrz, jest dobrym przykładem tego, jak modele biznesowe oparte na danych mogą całkowicie zakłócić każdy tradycyjny rynek. Jednak wypłata dotyczy nie tylko nowych graczy. W przypadku firm o ugruntowanej pozycji nowe modele biznesowe oparte na danych mogą pomóc utrzymać i zwiększyć ich udział w istniejącym rynku. Jednym z ostatnich przykładów w sektorze ubezpieczeń komunikacyjnych jest aplikacja Snapshot oferowana przez dużego gracza Progressive. Dzięki funkcji Snapshot dane są zbierane z małego urządzenia, które klienci podłączają do portu diagnostycznego samochodu, aby pomóc obliczyć składki na podstawie rzeczywistych nawyków jazdy. Wśród analizowanych danych jest to, kiedy i jak daleko jedzie klient oraz ile hamuje. Dobrzy kierowcy są nagradzani niższymi składkami. Średnio może to oznaczać oszczędności od 10 do 15 procent, co dla wielu kierowców może być atrakcyjną propozycją wartości.

Tworzenie firm zorientowanych na dane

Duża ilość danych generowanych przez firmy i spostrzeżenia, które generują, mogą mieć wartość dla innych firm i organizacji, zarówno w branży, jak i poza nią. Na przykład serwisy społecznościowe często przechwytują dane związane z preferencjami i opiniami użytkowników, które mogą być informacjami interesującymi dla producentów, którzy chcą lepiej skoncentrować swoje działania na rzecz rozwoju produktów i kampanie marketingowe. Operatorzy sieci komórkowych rutynowo zbierają dane o lokalizacji abonentów, które mogą być wartościowe dla detalistów, którzy chcą wiedzieć, gdzie konsumenci robią zakupy. Udostępniając te informacje (oczywiście za opłatą), firmy mogą rozwijać nowe strumienie przychodów poprzez monetyzację danych. Chociaż sprzedaż danych osobowych identyfikowalnych do konkretnych osób może budzić obawy dotyczące prywatności, firmy mogą znacznie zmniejszyć wrażliwość poprzez agregację i zapewnienie anonimowości danych, na przykład poprzez segmentację. Oznacza to, że osoby są najpierw umieszczane w grupie lub segmencie na podstawie ich nawyków konsumpcyjnych, sąsiedztwa, wieku i tak dalej. Po zakończeniu grupowania wszystkie dane osobowe (na przykład imię i nazwisko, adres i numer telefonu) są następnie usuwane, dzięki czemu stają się anonimowe. Identyfikacja odpowiednich aplikacji to dopiero pierwszy krok do czerpania wartości z dużych zbiorów danych. Potrzebne będą również nowe możliwości, nowe

struktury organizacyjne (i sposób myślenia) oraz znaczące zmiany wewnętrzne. Nie należy jednak lekceważyć znaczenia powiększania właściwych okazji. Musisz myśleć nieszablonowo, przyjmować nowe modele, a nawet na nowo wyobrazić sobie, jak i gdzie chcesz prowadzić biznes. Kultura, która zachęca do innowacji i eksperymentów, a nawet do radykalnego myślenia, dobrze przysłuży się temu przedsięwzięciu, ale w razie potrzeby będzie także wzywać pomoc z zewnątrz, aby ocenić, ustalić priorytety i opracować różne ścieżki dochodzenia do wartości. Dane i inteligencja maszyn to nie tylko zmiana konkurencyjnego środowiska; to fundamentalnie je przekształca. A wraz z tym Twoja firma musi się zmieniać. Zobaczenie, gdzie leżą możliwości i stworzenie strategii ich wykorzystania, pomoże Twojej firmie urzeczywistnić obietnicę dotyczącą danych. A ta nowa rzeczywistość pozwoli Tobie i Twojej firmie pozyskać nowych klientów, nowe przychody, a nawet nowe rynki.

Badanie różnych typów modeli biznesowych opartych na danych

Ważnym pierwszym krokiem w uświadomieniu sobie potencjalnych korzyści płynących z danych w Twojej firmie jest podjęcie decyzji o tym, jaki będzie model biznesowy. Gospodarka oparta na danych wspiera cały ekosystem firm i innych organizacji. Często są one zależne od swoich produktów i usług, dlatego siła sektora jako całości ma kluczowe znaczenie. Na przykład firmy i organizacje mogą udostępniać lub sprzedawać dane, modele, algorytmy i spostrzeżenia, które są włączane do nowych lub ulepszonych rozwiązań innych firm. Warto również wziąć pod uwagę, że produkty danych wymagają modelu biznesowego, aby określić, w jaki sposób użytkownicy odniosą korzyści ze świadczonej usługi i jak będzie generowana wartość z produktów i usług danych. Dostępnych jest wiele modeli kapitalizacji wartości danych i usług opartych na danych. Wybór, na który Ty i Twoja firma powinniście się zdecydować, naprawdę zależy od takich czynników, jak rodzaj świadczonej usługi, czy jest ona powiązana z platformą lub produktem oraz jakie korzyści odniesie z tego klient. (Jednym z typowych przykładów zarabiania jest model freemium, w którym użytkownikom oferuje się część usługi za darmo, ale pobierają opłatę za uaktualnienie do pełnej usługi lub pobierają opłatę za dodatkowe usługi danych w ramach istniejącego produktu.)

Różnicowanie przez dane

Kategoria zróżnicowania przez dane modeli biznesowych opartych na danych odnosi się głównie do tego, w jaki sposób wykorzystujesz dane w celu zróżnicowania obecnej firmy - podejmowanie kroków w celu jej wzmocnienia i uczynienia bardziej konkurencyjną. Można to zrobić, wykorzystując dane, aby lepiej zrozumieć swój rynek i klientów, wykorzystując dane do podejmowania decyzji w całej firmie lub stając się bardziej predykcyjnym, proaktywnym i zapobiegawczym w operacjach biznesowych i wobec klientów. Ta kategoria może również obejmować takie obszary, jak rozszerzenie obecnej działalności poprzez rozwój nowych rodzajów usług w oparciu o dane związane z obecną działalnością. W tym sensie różnicowanie tworzy również nowe doświadczenia. Od mniej więcej dekady świat widzi, jak technologia i dane dodają nowe poziomy personalizacji i znaczenia dla reklam i usług opartych na lokalizacji, jako dwa przykłady. Google AdSense dostarcza reklamy, które są faktycznie związane z tematami, których szukają użytkownicy. Sprzedawcy internetowi mogą oferować — za pośrednictwem FedEx, UPS, a nawet US Postal Service - śledzenie z dokładnością do minuty, gdzie znajdują się Twoje paczki. Usługi map Google, Microsoft, Yahoo!, a teraz także Apple dostarczają informacji związanych z miejscem, w którym się znajdujesz.

Pośrednictwo danych i informacji

Kolejna kategoria modeli biznesowych dotyczy tego, w jaki sposób możesz zostać brokerem danych i spostrzeżeń. Obejmuje to sprzedaż surowych, zagregowanych lub przetworzonych danych (na przykład oczyszczonych, oznaczonych, a nawet skorelowanych danych), których jesteś właścicielem. Może to również obejmować sprzedaż danych, których pierwotnym właścicielem nie jesteś, ale wtedy musisz

upewnić się, że masz prawa do sprzedaży danych stronie trzeciej. Inny model biznesowy w tej kategorii obejmuje sprzedaż określonych modeli analitycznych do celów samodzielnych lub do integracji z innym rozwiązaniem lub innymi odpowiednimi zastosowaniami. Modele biznesowe oparte na danych oferują możliwości dla wielu innych ofert usług, które zwiększą satysfakcję klientów i zapewnią znaczenie kontekstowe. Wyobraź sobie usługę opartą na mapach, która łączy dostawę paliwa z dostępnością stacji paliw. Jeśli miałeś mało paliwa, a Twój samochód rozmawiał z aplikacją map, może nie tylko zapewnić Ci trasy do najbliższych otwartych stacji benzynowych w promieniu 10 mil, ale także otrzymać cenę za galon. Kto nie zapłaciłby kilku dolarów miesięcznie za usługę kontekstową, która zapewnia spokój ducha i nigdy nie zabraknie paliwa w drodze? Oto kolejny przykład. W tym scenariuszu, w ramach działalności biznesowej, zarządzasz posiadaniem milionów zdjęć przedmiotów, w tym opisów tego, co przedstawiają te zdjęcia. Oprócz wykorzystania tych danych do rozwijania własnej firmy, możesz sprzedawać dostęp do tego zestawu danych do trenowania modeli uczenia maszynowego, na przykład modeli Deep Learning, jako dodatkowe źródło przychodów. Pośrednictwo danych i informacji może również obejmować sprzedaż ogólnych lub konkretnych modeli opartych na uczeniu maszynowym, zaprojektowanych do działania jako samodzielne produkty lub do zintegrowania z istniejącym oprogramowaniem w celu zwiększenia jego wydajności. Przykładem tej ostatniej jest aplikacja internetowa, której celem jest zapewnienie miejsca spotkań lub internetowego rynku dla osób, które chcą sprzedawać i kupować używane towary - na przykład szwedzka firma Blocket lub amerykański odpowiednik eBay. Żadna z tych firm nie miała na początku modeli uczenia maszynowego ulepszających aplikacje; tego typu funkcjonalności – wspierające wyszukiwanie, rekomendacje i automatyczną klasyfikację zdjęć podczas publikowania nowego ogłoszenia – zostały dodane później. Wreszcie, takie pośrednictwo może również obejmować usługi porównawcze, które w tym kontekście odnoszą się do wykorzystywania danych z kilku firm z określonego rynku lub segmentu biznesowego do porównywania aspektów, takich jak penetracja rynku, ocena klientów lub przestrzeganie określonego obowiązującego standardu.

Pośrednictwo w infrastrukturze

Pośrednictwo w infrastrukturze ma nieco inną strategię niż poprzednie dwie kategorie. Ma na celu sprzedaż produktów potrzebnych do umożliwienia pierwszych dwóch kategorii. Mogłaby na przykład oferować rozwiązania infrastrukturalne do pozyskiwania i gromadzenia danych lub przechowywania i przetwarzania danych oferowanych za pośrednictwem usług w chmurze. Może również odnosić się do różnego rodzaju narzędzi raportowych lub analitycznych, oferowanych lokalnie lub w chmurze. Rozwiązania mogą być wykorzystywane do wielu celów, w tym do eksploracji danych, wizualizacji danych, uzyskiwania spostrzeżeń do użytku wewnętrznego, a nawet komercjalizacji wyników poprzez sprzedaż uzyskanych spostrzeżeń. Wreszcie kategoria ta może również obejmować usługi doradcze związane z konfiguracją i użytkowaniem infrastruktury.

Sieci dostarczania danych

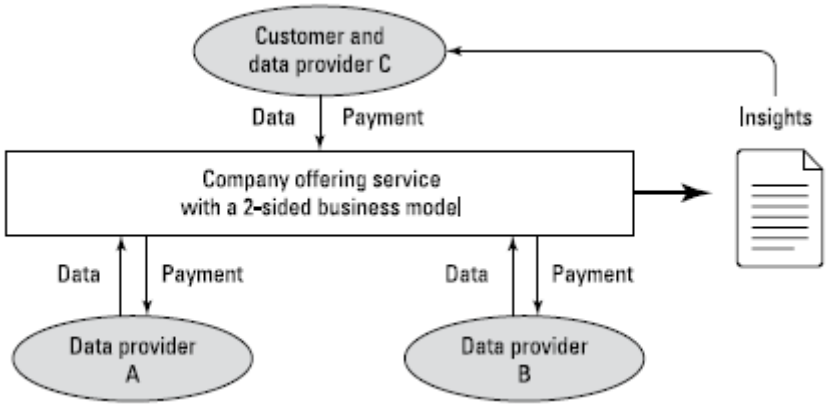
Kategoria modeli biznesowych sieci dostarczania danych odnosi się do tych obszarów, w których zysk pochodzi z łączenia różnych firm na różnych rynkach w celu udostępniania i sprzedaży produktów danych - innymi słowy, dogodne miejsce do spotkań i udostępniania, nawet dla konkurentów . W takim scenariuszu detaliści tacy jak Amazon mogliby sprzedawać surowe informacje na temat najgorętszych kategorii zakupów, a dodatkowe dane dotyczące wzorców pogodowych i wielkości płatności od innych partnerów mogą pomóc dostawcom jeszcze dokładniej określić sygnały popytu. Te nowe strumienie analiz i wglądów mogą być tworzone i utrzymywane przez brokerów informacji, którzy mogliby sortować według wieku, lokalizacji, zainteresowań i innych kategorii. Dzięki nieskończonym wariantom modele biznesowe brokerów byłyby dostosowane do branży, lokalizacji i roli użytkownika. Licencjonowanie krzyżowe danych to jeden z rodzajów ofert, z których korzystają czasami konkurenci.

W tym przypadku obie strony zgadzają się udzielić drugiej stronie licencji na zbieranie i wykorzystywanie danych należących do drugiej strony i przez nią zablokowanych. Korzystając z modelu cross-licensing, każda ze stron zyskuje (lub traci) równy wgląd w dane konkurencji. Często spotyka się takie licencje krzyżowe stosowane w branży dostawców sprzętu telekomunikacyjnego. Dzieje się tak, ponieważ wiele sieci telekomunikacyjnych jest obsługiwanych przez wielu dostawców, którzy dostarczyli sprzęt telekomunikacyjny do sieci operatora. Biorąc pod uwagę to środowisko telekomunikacyjne wielu dostawców, wzajemne licencjonowanie danych przydaje się, gdy trzeba dowiedzieć się więcej o zainstalowanej bazie sprzętu konkurencji – na przykład, jaki sprzęt, oprogramowanie lub zestaw konfiguracji jest używany – lub gdzie może być potrzeba uzyskania dostępu do danych o wydajności ze sprzętu konkurencji w celu uzyskania pełnego wglądu w wydajność całej sieci. Uzyskane dane i spostrzeżenia można następnie wykorzystać do sprzedaży spostrzeżeń lub innych rodzajów usług związanych z danymi. Sieci dostarczania danych umożliwiają monetyzację danych na większą skalę. Aby były naprawdę wartościowe, wszystkie te dane muszą być dostarczone do rąk tych, którzy mogą z nich korzystać, kiedy mogą z nich korzystać, za pośrednictwem różnego rodzaju rynków. Sieci dostarczania danych zbierają dane i agregują je, wymieniają i odtwarzają w nowsze i czystsze strumienie wglądu - podobnie jak telewizja kablowa w zakresie dostarczania treści. Te sieci dostarczania danych będą podstawowym ścieżką, przez którą oferty oparte na informacjach znajdują swoje rynki i zarabiają. Chociaż ich podstawową funkcją jest bycie rynkiem do prowadzenia biznesu, funkcjonują również jako hybryda między nowym rodzajem oferty a modelem dostawy. Niewiele organizacji dysponuje kapitałem na tworzenie kompleksowych sieci dostarczania treści, które mogą przechodzić z chmury na urządzenie. Dziś tylko nieliczni giganci – tacy jak Amazon, Apple, Bloomberg, Google i Microsoft – wykazują taki potencjał, ponieważ są właścicielami łańcucha dystrybucji od chmury do urządzenia. Licencjonowanie krzyżowe opiera się na otwartym rynku, który działa jako platforma, na której dostawcy danych i modeli mogą spotykać się z użytkownikami. Dwustronny model biznesowy jest podobny, ponieważ opiera się na koncepcji łączenia danych i różnych stron, ale jest jedna istotna różnica – dwustronny model biznesowy nie jest dostępny dla wszystkich. Jest to ograniczona konfiguracja, stworzona tylko w określonym celu i wyposażona w aktywnego pośrednika, który łączy różne strony, zapewniając zarówno dostawę modeli, jak i monetyzację. Podstawowa koncepcja dwustronnego modelu biznesowego polega na tym, że obejmuje (co najmniej) trzy rodzaje zaangażowanych stron, chociaż może obejmować znacznie więcej. Główna strona działa jako pośrednik – ten, który oferuje usługę klientowi, który potrzebuje wglądu w dane, ale nie jest w stanie samodzielnie przeprowadzić tej analizy. Istnieje wiele powodów, dla których klient może nie być w stanie tego zrobić - być może firma nie ma praw dostępu do danych lub nie ma odpowiedniej infrastruktury do zarządzania nimi lub nie ma wiedzy domenowej do analizy i wyciągania wniosków z danych - ale to nie ukrywa, że potrzeba nadal istnieje. Tutaj z pomocą przychodzi pośrednik, który kupuje potrzebne dane od innych dostawców danych, a następnie wykonuje analizę w imieniu klienta, który z kolei płaci za spostrzeżenia.

Oto przykład: Wyobraź sobie, że globalny sprzedawca kawy chce zrozumieć, jak dobrze działają jego kampanie marketingowe dotyczące kawy na określonym obszarze w Stanach Zjednoczonych. Dostawcą tej konkretnej usługi sieciowej dostarczania danych jest globalny dostawca telekomunikacyjny, który uruchomił zupełnie nową usługę zbudowaną na przełomowym modelu biznesowym. Działa to tak, że dostawca usługi kupuje dane od dwóch operatorów w USA i wykorzystuje segmentowane dane lokalizacyjne dla grupy osób mieszkających na określonym obszarze. (Nie można zidentyfikować żadnych osób, ponieważ są one anonimowe jako część grupy mieszkającej na określonym obszarze geograficznym.) Następnie badane są wzorce przemieszczania się do i z najbliższych kawiarni przy użyciu danych od operatorów zebranych z telefonów komórkowych

użytkowników telefony. Dzięki tym danym można by określić, jak skutecznie kampania marketingowa okazała się skuteczna dla osób mieszkających na danym obszarze. I można to analizować bez

naruszanie prywatności osoby. Rysunek przedstawia ekstrapolację z mojego przykładu sprzedawcy kawy, pokazujący dwustronny model biznesowy w działaniu.



Należy zauważyć, że Rysunek pokazuje, że klient również dostarcza dane (jako dostawca danych C), w tym dane takie jak obszar geograficzny, zakres czasowy i lokalizacja kawiarni, ale brakuje mu potrzebnych danych od dostawców danych A i B (dane operatora z informacjami o lokalizacji klientów kawy). Bez tych danych globalny sprzedawca kawiarni nie może przeprowadzić analizy. Dostawcy danych A i B są konkurentami na tym samym rynku i dlatego odmówiliby po prostu przekazania swoich danych konkurentowi, nawet jeśli mieliby otrzymać wynagrodzenie, ponieważ mogłoby to ujawnić poufne informacje na temat ich działalności. Otwiera to możliwość wykorzystania dwustronnego modelu biznesowego, w którym strona trzecia oferująca zarówno dane, jak i biznesowe zrozumienie biznesu telekomunikacyjnego, mogłaby działać jako neutralny gracz lub pośrednik łączący działalność kawiarni z operatorami telekomunikacyjnymi.

Włączanie funkcji uczenia maszynowego/sztucznej inteligencji

Wszystkie modele biznesowe oparte na uczeniu maszynowym i sztucznej inteligencji potrzebują danych, aby istnieć i realizować swój cel (funkcjonalny lub inny), a zatem z definicji są modelami biznesowymi opartymi na danych. Możesz również użyć technologii uczenia maszynowego/sztucznej inteligencji, aby uzyskać wgląd w dane, a te spostrzeżenia mogą być sprzedawane tak samo, jak inne spostrzeżenia. Jednak pomimo podobieństw z innymi modelami, modele biznesowe oparte na uczeniu maszynowym i sztucznej inteligencji są nieco inne. Głównym powodem inwestowania w modele biznesowe oparte na uczeniu maszynowym/sztucznej inteligencji jest zwykle rozszerzenie obecnej działalności i technologii o nowe i zaawansowane techniki i funkcje. Ta ulepszona funkcjonalność może być na przykład wykorzystana do ewolucji automatyzacji na nowy poziom dzięki inteligentnej automatyzacji, skupiającej się głównie na optymalizacji sposobu wykonywania określonego zautomatyzowanego zadania. Na przykład, jeśli kroki automatyzacji wykonywane dzisiaj przez jakąkolwiek maszynę są tymi samymi krokami, które wcześniej wykonał człowiek w celu wykonania określonej usługi, dzięki uczeniu maszynowemu maszyna może zidentyfikować najlepszy sposób rozwiązania zadania, niezależnie od tego, które kroki były wcześniej wykonywane. Maszyna nie jest związana z góry wyobrażeniem „właściwego sposobu zrobienia czegoś" (zakładając, że dane, zespół i algorytm są bezstronne), ale raczej koncentruje się na rozwiązaniu zadania w najbardziej zoptymalizowany sposób. Możesz również użyć modelu biznesowego skoncentrowanego na uczeniu maszynowym/sztucznej inteligencji, aby rozwinąć funkcjonalność już istniejącego rozwiązania za

pomocą dynamicznych i regulowanych technik, które oferuje taki model. Aby zobaczyć, co mam na myśli, oto przykład sieci telekomunikacyjnej, w której wcześniej w oprogramowaniu używano tylko modeli uczenia maszynowego dla tysięcy stacji bazowych rozsianych po całym kraju. Tak było do 2017 roku. Teraz jednak kilka modeli języka maszynowego zostało wprowadzonych online do stacji bazowych, dzięki czemu można dynamicznie dostosowywać się i lepiej obsługiwać klientów w czasie rzeczywistym, ponieważ potrzeba zmiany przepustowości z dnia na dzień -tydzień, pora dnia, preferencje dla niektórych aplikacji, lokalizacja geograficzna itd. Modele uczenia maszynowego w stacjach bazowych są oczywiście szkolone na rzeczywistych danych przed ich wdrożeniem, ale mogą następnie nadal uczyć się wzorców dla różnych obszarów geograficznych, które obejmują, co oznacza, że mogą następnie przewidywać i przygotowywać się do obsługi klientów w miarę pojawiania się ich potrzeb. Zamiast jednego modelu służy wszystkim (lub nikomu), sieć dynamicznie dostosowuje się proaktywnie i w czasie rzeczywistym do stale zmieniających się potrzeb w połączonym społeczeństwie. Kolejny sposób, w jaki uczenie maszynowe/sztuczna inteligencja może wzmocnić twoje różne modele biznesowe polegają na wykorzystaniu go do zupełnie nowych i destrukcyjnych modeli biznesowych, takich jak robotyka. Jest to rozszerzający się obszar, który obecnie wykracza poza powtarzalną automatyzację, którą można znaleźć w fabrykach, gdzie roboty pracują same, do scenariuszy, w których stają się dynamicznymi i inteligentnymi asystentami ludzi w środowiskach laboratoryjnych (co-boty), w naszych samochodach (samojazda). samochodów), w naszych ogrodach (roboty koszące), a nawet w naszych domach (roboty odkurzające). Wiele możliwych ścieżek stoi przed Tobą otworem, jeśli chodzi o monetyzację rewolucji danych. Najważniejsze jest, aby mieć pomysł, którym chcesz się kierować w swojej firmie. Tylko dzięki zrozumieniu, który model biznesowy (lub modele) najlepiej pasuje do Twojej organizacji, możesz podejmować mądre decyzje dotyczące budowania, partnerstwa lub zdobywania drogi do następnej fali ewolucyjnej.

Korzystanie z modelu biznesowego opartego na danych

Choć wydaje się, że argumentowanie przeciwko biznesowej wartości danych wydaje się trudne, wykorzystanie potencjału danych nie jest tak łatwe, jak się wydaje na pierwszy rzut oka. Przyczyny są wielowarstwowe: wielu firmom brakuje specjalistycznej wiedzy w obszarach czyszczenia i przechowywania danych, a dane często stają się wartościowe dopiero wtedy, gdy są agregowane z danymi od konkurentów lub graczy z innej branży, co może być trudne lub nawet niemożliwe do osiągnięcia. Doprowadziło to do powstania kilku inicjatyw mających na celu zdefiniowanie ram wspierających firmy i organizacje oraz oferowanie bardziej ustrukturyzowanego podejścia do wprowadzania modeli biznesowych opartych na danych. Ważnym aspektem, od którego należy zacząć, jest aktywne kwestionowanie gotowości i chęci firmy do zmiany i inwestowania w dane jako podstawową część działalności, a nie tylko coś, co firma robi na boku. Modele biznesowe oparte na danych wymagają pełnego zaangażowania w całej firmie i dotyczą bezpośrednio klientów, dlatego ważne jest, abyś zrobił to właściwie, gdy już zdecydujesz się to zrobić. Ta krótka lista pytań może posłużyć jako przykład podejścia do samooceny własnej firmy pod kątem gotowości do wprowadzenia modeli biznesowych opartych na danych:

* Czy moja firma jest gotowa do omówienia modeli biznesowych opartych na danych?

* Jakie strategiczne cele biznesowe realizuje moja firma? Czy są one strategicznie zgodne z ambicją modelu biznesowego opartego na danych – czy też są z nimi sprzeczne?

*Czy ambicją jest zarabianie między przedsiębiorstwami (B2B) czy między przedsiębiorstwami (B2C)?

*Czy istnieje wsparcie organizacyjne (procesy, zarządzanie, ramy) dla pomysłu, który ma być oceniany w ustrukturyzowany sposób?

* Czy potencjał rynkowy można poprawić dzięki inicjatywom międzybranżowym, takim jak projekt open source lub wysiłek normalizacyjny – i jak byś podszedł do tego w takim przypadku?

Po przeprowadzeniu oceny firmy i (miejmy nadzieję) podjęciu decyzji o dalszym rozwoju i wprowadzeniu modelu biznesowego opartego na danych, masz kilka kluczowych obszarów do rozważenia:

* Dowiedz się, co oznacza model biznesowy oparty na danych w Twojej branży.

* Wykorzystaj potencjał modelu biznesowego opartego na danych i spróbuj określić i określić ilościowo jego znaczenie dla przyszłości Twojej firmy.

* Uznaj, że modele biznesowe oparte na danych nie są tajemnicą. Są już używane w różnych firmach w różnych segmentach przemysłu, nawet jeśli większość z nich jest wciąż stosunkowo nowa.

* Użyj prostej struktury, aby kierować myśleniem i podejściem do modeli biznesowych opartych na danych.

* Odkryj istniejące i odpowiednie wzorce w firmie i uzyskaj przegląd tego, od czego możesz zacząć i co wykorzystać na wczesnym etapie.

Tworzenie modelu biznesowego opartego na danych przy użyciu frameworka

Nie ma nic prostego w tworzeniu modelu biznesowego opartego na danych dla Twojej firmy, więc aby zaoferować Ci praktyczne podejście do tego zadania, opiszę w tej sekcji ramy modelu biznesowego opartego na danych (DDBM). DDBM można wykorzystać jako plan innowacji wysokiego poziomu do identyfikacji korzyści i wyzwań związanych z wykorzystaniem myślenia opartego na danych do konstruowania modeli biznesowych opartych na danych. Model biznesowy oparty na danych (DDBM) składa się z sześciu wymiarów: kluczowych zasobów, kluczowych działań, propozycji wartości, segmentu klientów, modelu przychodów i struktury kosztów.

Kluczowe zasoby

Firmy potrzebują zasobów, aby rozwijać swoje produkty lub usługi, a także tworzyć wartość. Z definicji DDBM ma dane jako kluczowy zasób, ale nie oznacza to, że dane są jedynym kluczowym zasobem danego modelu biznesowego. Twoja firma może potrzebować innych kluczowych zasobów, aby umożliwić Twój model biznesowy – na przykład kluczowych kompetencji i infrastruktury. Jednak głównym celem kluczowego zasobu DDBM jest zbadanie i zdefiniowanie, jakiego rodzaju źródła i typy danych są potrzebne do realizacji celu modelu biznesowego opartego na danych. Identyfikując potrzebne źródła danych, należy rozróżnić zewnętrzne i wewnętrzne:

* Zewnętrzne źródła danych mogą odnosić się do typów danych, takich jak preferencje klientów lub rynku, otwarte rekordy danych statystycznych, dane porównawcze lub różne dane z mediów społecznościowych, takich jak blogi.

*Wewnętrzne źródła danych mogą odnosić się do różnych baz danych z danymi historycznymi, danymi z wewnętrznych systemów transakcyjnych, danymi produktów i usług związanych z wydajnością, jakością, oceną klientów, a także danymi finansowymi firmy i tak dalej.

Kluczowe działania

Jak każda inna firma, Twoja firma musi wykonywać różne działania, aby rozwijać, produkować i dostarczać swoją ofertę. W tradycyjnych modelach biznesowych zorientowanych na produkt, kluczowe działania tworzące wartość można opisać za pomocą tradycyjnego łańcucha wartości - modelu

wysokiego poziomu używanego do opisania procesu, w którym firmy otrzymują surowce, dodają wartość do surowców poprzez różne procesy, aby stworzyć gotowy produkt, a następnie sprzedać gotowy produkt klientom. Ponieważ jednak tradycyjna koncepcja łańcucha wartości koncentruje się przede wszystkim na świecie fizycznym i traktuje dane jako element wspierający, a nie jako samo źródło wartości, ma ograniczone zastosowanie w kontekście modeli biznesowych opartych na danych. Zamiast tego musisz zidentyfikować wszystkie kluczowe działania związane z Twoim konkretnym pomysłem biznesowym opartym na danych, w tym wszystkie działania związane z cyklem życia danych, które wykonujesz na danych – ich przechwytywanie, utrzymywanie, przetwarzanie, analizowanie, komunikowanie, i uruchamiając go. Działania związane z nauką o danych będą miały różne znaczenie, w zależności od tego, jak wygląda Twój model biznesowy, ale warto poświęcić trochę czasu na przemyślenie niezbędnych kroków i tego, jak zabezpieczyć stabilną podstawę nauki o danych w swoim DDBM.

Oferta/propozycja wartości

Propozycję wartości można zdefiniować jako wyraz doświadczenia, które klient otrzyma od dostawcy i które można zmierzyć poprzez tworzenie wartości. Oznacza to, że propozycja wartości jest wartością stworzoną dla klientów poprzez ofertę. Ponieważ jednak trudno jest sformalizować i skategoryzować postrzeganą wartość przez klienta w jakiejkolwiek branży, ramy DDBM skupiają się na ofercie. To, co Twoja firma będzie oferować jako rdzeń modelu biznesowego opartego na danych, sprowadza się do tego, co opisano wcześniej w tym rozdziale jako różne typy modeli biznesowych: zróżnicowanie poprzez dane, pośrednictwo w zakresie danych i informacji, pośrednictwo w infrastrukturze, sieci dostarczania danych i maszyny włączanie funkcji uczenia się/sztucznej inteligencji. Ważne jest, aby mieć solidny pomysł na biznes, który następnie poświęcisz na odpowiednie określenie i zdefiniowanie kategorii.

Segment klientów

Wymiar segmentu klienta dotyczy grupy docelowej oferty. Chociaż istnieje kilka sposobów segmentacji klientów, najbardziej ogólna klasyfikacja dzieli klientów docelowych na firmy lub klientów biznesowych (B2B) oraz konsumentów indywidualnych lub klientów biznesowych (B2C). B2B to sytuacja, w której jeden biznes dokonuje transakcji handlowej z innym biznesem. Zwykle dzieje się tak, gdy firma pozyskuje materiały do swojego procesu produkcyjnego do produkcji (na przykład producent żywności kupuje sól jako surowiec w celu zwiększenia produkcji). Z drugiej strony B2C odnosi się do firmy, która bezpośrednio sprzedaje produkty lub świadczy usługi konsumentom końcowym. W wielu przypadkach firmy mogą kierować reklamy zarówno do przedsiębiorstw, jak i indywidualnych konsumentów. W handlu B2B często zdarza się, że strony relacji mają porównywalną siłę negocjacyjną, a nawet jeśli nie, każda ze stron zazwyczaj angażuje profesjonalny personel i radcę prawnego w negocjowanie warunków. B2C jest natomiast kształtowany w znacznie większym stopniu przez nierówną równowagę między stronami: firmą oferującą produkt lub usługę a użytkownikiem końcowym lub konsumentem. W tej relacji firma ma nadrzędną pozycję, jeśli chodzi o użytkownika końcowego pod względem konsekwencji ekonomicznych i dostępu do odpowiednich informacji.

Model przychodów

Aby przetrwać długoterminowo, każda firma musi mieć co najmniej jeden strumień przychodów. Można wyróżnić kilka różnych modeli przychodów za pomocą klasyfikacji, takich jak:

* Sprzedaż aktywów: zrzeczenie się praw własności do produktu (produktu danych, takiego jak zbiór danych, spostrzeżenia lub modele/algorytmy) lub usługi w zamian za pieniądze

* Użyczanie/wynajem/leasing: Tymczasowe przyznanie komuś wyłącznych praw do korzystania z aktywów (na przykład zbioru danych) na określony czas

* Licencjonowanie: udzielanie pozwolenia na korzystanie z chronionej własności intelektualnej, takiej jak patent (na przykład model/algorytm) lub prawo autorskie, w zamian za opłatę licencyjną

* Opłata za użytkowanie: Pobieranie opłaty za korzystanie z określonej usługi (np. zdefiniowany zakres usługi Insights as a Service)

* Subskrypcja: Pobieranie opłat za korzystanie z usługi lub oprogramowania (na przykład oprogramowania rozszerzonego do uczenia maszynowego) w ograniczonym i uzgodnionym okresie

* Pośrednictwo: Pobieranie opłat za usługę pośrednią, w której model biznesowy działa jako pośrednik, łącząc dane i spostrzeżenia z innymi podmiotami (czasami poprzez tworzenie nowych rynków, na których te podmioty mogą się spotykać i prowadzić interesy)

* Reklama: dostarczanie reklam w witrynie lub w związku z usługą. Może to zapewnić dodatkowe źródło dochodu lub być głównym źródłem dochodu. Aby była skuteczna dla reklamodawców, potrzebujesz dobrego zrozumienia grupy docelowej, aby kierować odpowiednie reklamy do właściwej grupy docelowej, wykorzystując dostępne dane do badania i analizy potencjału rynkowego i cenowego.

Poświęć trochę czasu, aby odpowiednio przemyśleć rodzaj modelu lub modeli przychodów, do których dąży Twoja firma, zarówno krótkoterminowych, jak i długoterminowych. Pomoże Ci to zidentyfikować i kierować innymi aspektami modelu biznesowego opartego na danych podczas korzystania z tej struktury.

Struktura kosztów

Aby móc tworzyć i dostarczać klientom wartość, firma generuje koszty pracy, technologii, zakupionych produktów i tak dalej. Czy zatem w ramach tego procesu wykorzystanie danych umożliwi konkretną przewagę kosztową? Cóż, zazwyczaj firma miałaby określoną przewagę kosztową, gdyby dane wykorzystane w jej produkcie lub usłudze zostały utworzone niezależnie od konkretnej oferty. Przykładem tego jest producent samochodów wykorzystujący dane, które są automatycznie tworzone i przechowywane przez elektronikę w samochodzie. Są też inne firmy, takie jak Automatic, start-up dostarczający analizy dla właścicieli samochodów, takie jak śledzenie parkowania, przypomnienia o konserwacji, diagnostyka silnika, historia jazdy i wgląd. Aby przechwycić te same dane, Automatic musi zainstalować określone urządzenie podłączone do samochodu, co najprawdopodobniej wymagałoby również konkretnej zgody każdego właściciela samochodu. Innym przykładem jest Twitter, który mógłby bez dodatkowych kosztów wykorzystywać własne dane do świadczenia różnych usług analitycznych, takich jak trendy w opiniach na dany temat poruszany w tweetach, podczas gdy firmy takie jak Gnip, start-up zajmujący się analizą mediów społecznościowych, musiałby zapłacić Twitter dla tych samych danych, bezpośrednio wpływ na jego strukturę kosztów.

Łącząc wszystko razem

Po odpowiednim zrozumieniu swojego pomysłu biznesowego i sposobu, w jaki chcesz go zrealizować za pomocą modelu biznesowego opartego na danych, po prostu śmiało i korzystając ze struktury DDBM opisanej w tym rozdziale, połącz sześć wymiarów i odpowiednie funkcje dla każdego wymiaru . Usiąść i wykonać tę pracę pomoże Ci zdefiniować i opracować w pełni oparty na danych model biznesowy. Dla każdego wymiaru należy wybrać co najmniej jedną cechę; jednak firma może mieć więcej niż jedną funkcję dla dowolnego wymiaru.

Obsługa nowych modeli dostaw

Być może masz już całkiem niezły pomysł na to, co oznacza data science w Twojej branży, zarówno pod względem wyzwań, jak i potencjału. Być może nawet zacząłeś pracować nad różnymi aspektami własnej strategii analizy danych w oparciu o swój pomysł na biznes. A może już budujesz pełnoskalowy model biznesowy oparty na danych i zbliżasz się do realizacji. Bez względu na to, jak daleko zaszedłeś lub nawet jeszcze nie zacząłeś, jeśli nie zastanowiłeś się, jak zamierzasz dostarczać nowe produkty i usługi związane z danymi, pominąłeś w swoich planach ważny strategiczny aspekt. Modele dostarczania mogą wydawać się czymś, co możesz później rozwiązać, myśląc, że kiedy zaczniesz, zrozumiesz to. Ale nie popełnij błędu, ważne jest, aby jak najwcześniej przemyśleć ten aspekt swojego modelu biznesowego. Sposób, w jaki zamierzasz dostarczać – lub możesz być zmuszony do dostarczenia, w zależności od wymagań klientów lub oczekiwań użytkowników – może narzucić Twojej firmie lub organizacji ogromne transformacyjne zmiany.

Definiowanie modeli dostarczania produktów i usług związanych z danymi

Model dostawy opisuje sposób, w jaki zamierzasz dostarczyć produkt lub usługę, którą planujesz sprzedać klientowi. W przypadku produktów fizycznych oznacza to ustalenie, w jaki sposób planujesz wysłać produkt z fabryki do sklepu, w którym zostanie on udostępniony do zakupu przez klienta. W zależności od modelu biznesowego może być również wysyłany bezpośrednio z fabryki do klienta końcowego – na przykład w przypadku sprzedaży za pośrednictwem sklepów internetowych. Rzeczy, które musisz wziąć pod uwagę w tym kontekście, są głównie związane z takimi aspektami, jak liczba fabryk i lokalizacja fabryk (wybór krajów dla globalnych przedsiębiorstw), twoja potrzeba, w zależności od zapotrzebowania klientów, oczekiwany czas dostawy oraz częstotliwość i pożądane sposoby do konsumpcji Twoich produktów. Jeśli chodzi o produkty i usługi związane z danymi, należy wziąć pod uwagę inne kwestie. W przypadku produktów danych jest to głównie przypadek produktów i usług cyfrowych i zwirtualizowanych, które wymagają innych typów modeli i platform dostarczania, takich jak usługi oparte na chmurze z różnymi typami modeli jako usługi (aaS). Lub może to być nielicencjonowane oprogramowanie open source lub różnego rodzaju rynki danych i uczenia maszynowego/sztucznej inteligencji, w których niektóre części są otwarte dla wszystkich, a inne są zablokowane, chyba że uzyskasz licencję. W przypadku produktów i usług związanych z danymi należy również wziąć pod uwagę aspekty prawne, etyczne i związane z bezpieczeństwem, które różnią się od wymagań dotyczących tradycyjnych modeli dostarczania sprzętu i oprogramowania. W zależności od ograniczeń prawnych w różnych krajach, może być konieczne rozważenie modelu dostarczania, w którym niektóre kraje, w których prowadzisz działalność, mogą mieć bardziej rygorystyczne przepisy dotyczące wykorzystania danych, zwłaszcza w odniesieniu do korzystania z danych zawierających dane osobowe. Jeśli dane nie mogą legalnie opuścić kraju, nie możesz wykonywać przetwarzania danych, opracowywać modeli i spostrzeżeń ani dostarczać wyników z innego kraju niż ten, z którego pochodzą dane.

Zrozumienie i dostosowanie się do nowych modeli dostarczania

W IT termin alternatywne modele dostarczania odnosi się do zastąpienia tradycyjnych modeli dostarczania produktów i usług oprogramowania nowymi rodzajami strategii i procesów, które mają na celu usprawnienie sposobu wykorzystania technologii. Dość szeroki model dostarczania terminów jest często ostrożnie stosowany do nowych modeli usług, które stały się możliwe dzięki postępowi technologicznemu, np. wspierających usługi dostarczane przez Internet. Niektóre z alternatywnych modeli dostarczania, o których najczęściej mówią eksperci, obejmują usługi w chmurze i modele oprogramowania jako usługi (SaaS). Tutaj zamiast sprzedawać oprogramowanie w pudełku na fizycznym dysku CD lub innym nośniku danych, oprogramowanie jest dostarczane przez Internet lub

inny rodzaj połączenia sieciowego. Dzięki tym nowym typom alternatywnych modeli dostawy użytkownicy mogą zdecydować się na zakup usług z opłatami za subskrypcję lub kupić cały pakiet, jednocześnie uzyskując jego wdrożenie przez Internet. Tym samym alternatywne modele dostaw stały się w rzeczywistości ważny termin do omówienia, reprezentujący szybką zmianę w świecie biznesu oraz w sposobach, w jakie ludzie kupują i korzystają z aplikacji. Dostarczanie jako usługa jest również odpowiednim modelem dostarczania produktów danych. Produkty danych jako usługa oznaczają dostarczanie ich na żądanie – skalowalne i bezpieczne. Interfejs użytkownika jest często implementowany za pośrednictwem aplikacji lub interfejsu internetowego, a cała usługa jest często udostępniana za pośrednictwem infrastruktury opartej na chmurze, obejmującej różne platformy i aplikacje. Organizacje, które zazwyczaj nie działają w branży oprogramowania, w rzeczywistości muszą zacząć zachowywać się jak firmy programistyczne, dostarczając produkty z danymi. Z modeli dostarczania oprogramowania można się wiele nauczyć, ale pamiętaj, że te tradycyjne również ulegają ciągłym zmianom. Nowe postępy technologiczne, wymagania dotyczące efektywności kosztowej i wymagania użytkowników to niektóre czynniki napędzające ciągłą potrzebę znalezienia lepszych i bardziej atrakcyjnych sposobów dostarczania oprogramowania oraz produktów i usług związanych z danymi. Z punktu widzenia strategii analizy danych łatwo zapomnieć lub nie docenić znaczenia, jakie dla Twojej firmy ma wybór odpowiedniego modelu dostarczania. To zadanie jest bardzo ważne, aby dotrzeć do klientów w sposób, jakiego oczekują i potrzebują oraz który odpowiada Twojej branży. A ponieważ oczekiwania klientów będą się zmieniać w czasie, Twój model dostarczania musi być elastyczny, skalowalny i zbudowany w sposób umożliwiający reagowanie na zmieniające się wymagania w czasie. Gdy zrozumiesz, w jaki sposób musisz dostarczać swoje oferty, odkryjesz, że wybrany model dostawy będzie miał znacznie większy wpływ niż myślisz. Zwykle wpływa na takie obszary, jak cykl życia produktów i usług, geograficzna obecność ośrodków rozwoju, strategie kompetencyjne, struktury organizacyjne i wspierające, a nawet faktyczną operacjonalizację produktów i usług związanych z danymi.

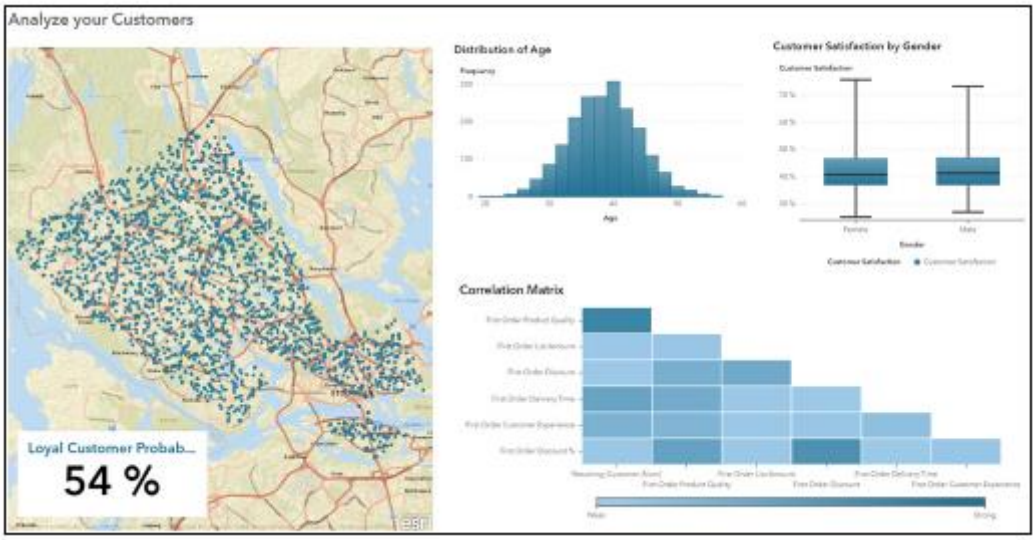
Przedstawiamy nowe sposoby dostarczania produktów z danymi

Obszar rekomendowania wydajnych modeli dostarczania dla różnych rodzajów produktów i usług związanych z danymi jest nadal badany i odkrywany w trakcie pisania tej książki. Jednak w tej sekcji znajdziesz ogólny przegląd kilku przykładów różnych modeli używanych w różnych kategoriach produktów i usług związanych z danymi. Istnieje wiele różnych typów modeli dostarczania, a czasami ten sam model dostarczania może być używany dla różnych ofert i różnych modeli biznesowych opartych na danych. Poniższe sekcje opisują te różne modele dostarczania bardziej szczegółowo i zawierają kilka przykładów kontekstowych.

Samoobsługowe środowiska analityczne jako model dostarczania

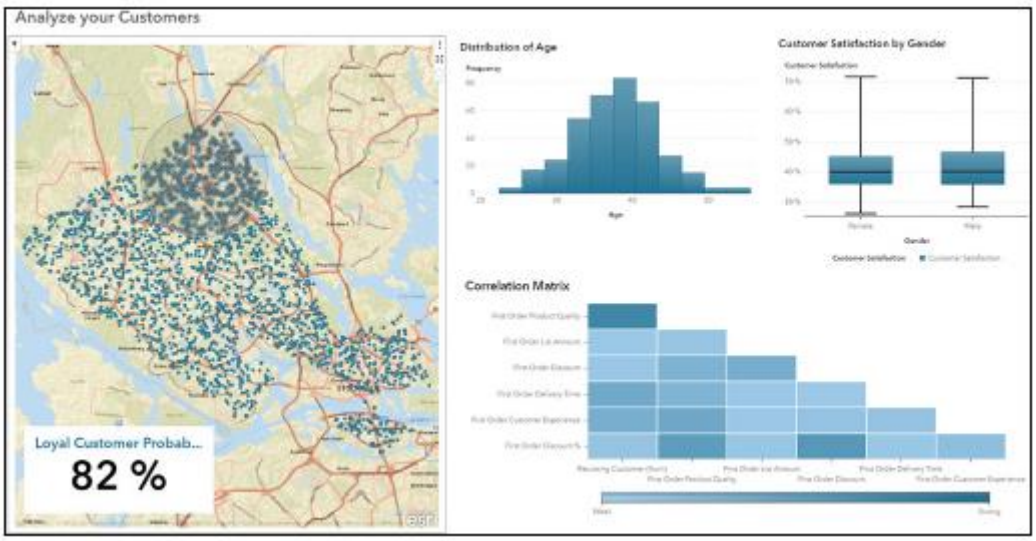
Inwestując w model biznesowy oparty na danych, którego celem jest różnicowanie na podstawie danych, jednym z przykładów modelu dostarczania jest wykorzystanie samoobsługowego środowiska analitycznego. Większość z tych gotowych do użycia narzędzi analitycznych jest łatwa w użyciu i zwykle jest dostępna zarówno jako instalacje lokalne, jak i rozwiązania oparte na chmurze. Korzystając z gotowego produktu do eksploracji i generowania spostrzeżeń biznesowych na podstawie danych w celu podejmowania lepszych decyzji, koncentrujesz swoje wysiłki na przygotowaniu danych i analizie danych, a nie na inwestowaniu w tworzenie od podstaw własnego narzędzia lub platformy. Jest to szczególnie przydatne dla firm, które dopiero zaczynają przygodę z nauką o danych (a zatem z niewielkimi lub zerowymi kompetencjami w zakresie analityki lub uczenia maszynowego/sztucznej inteligencji), ale także do wewnętrznych celów analizy biznesowej w dowolnym typie firmy, niezależnie od ich dojrzałości analitycznej poziom. Możliwe jest również wykorzystanie danych wyjściowych z gotowego narzędzia do eksploracji lub analizy w celu wygenerowania odpowiedniego pulpitu

nawigacyjnego lub innej wizualizacji, której można użyć również zewnętrznie – na przykład w stosunku do klientów – oszczędzając w ten sposób czas potrzebny na zaprojektuj swój własny. Kolejną zaletą korzystania z gotowych do użycia narzędzi analitycznych jest to, że są one wyposażone w interaktywne widoki wizualizacji, co rzadko zdarza się w przypadku tworzenia wizualizacji w Pythonie lub R, dwóch najpopularniejszych językach programowania dla naukowców zajmujących się danymi. Interaktywne wizualizacje oznaczają, że możesz łatwo klikać różne części wizualizacji, aby powiększyć lub pomniejszyć, wybrać obszar do dalszej analizy, a nawet zmienić zakres tego, czym się zajmujesz. Zacznij od jednej wizualizacji dla swojego zestawu danych, a następnie łatwo zmień na inny, gdy chcesz rozszerzyć analizę. Lub po prostu połącz różne wizualizacje w jeden widok (jak w poniższych przykładach) i połącz wykresy, aby po zmianie zakresu jednego widoku inne wykresy również dopasowywały się do tego zakresu. Poniższy przypadek pokazuje, jak możesz zwiększyć swoje zrozumienie, łatwo dodając dane o lokalizacji do tradycyjnego zestawu danych, który zwykle przeglądasz. Dodając kontekst geograficzny do analizy i wizualizacji poprzez połączenie tradycyjnych danych z danymi lokalizacji, analiza lokalizacji wysuwa na pierwszy plan wymiar „gdzie", dzięki czemu można analizować dane na nowe sposoby, aby uzyskać pełny obraz przed podjęciem decyzji, jednocześnie identyfikując lokalizację- konkretne możliwości. Rysunek 23-2 poniżej pokazuje, w jaki sposób można uzyskać przegląd danych klientów za pomocą zestawu zmiennych, takich jak miejsce zamieszkania, wiek i poziom zadowolenia klienta. Te różne zmienne danych można następnie wykorzystać do dalszej analizy i wyszukiwania możliwych zależności. Na przykład automatycznie wygenerowana macierz korelacji pokazuje silną korelację między „jakością produktu pierwszego zamówienia" a „klientem powracającym". Dane o lokalizacji \ pokazane na mapie wskazują, że średnie prawdopodobieństwo lojalności klientów we wszystkich klientów to 54 proc.



Wykres na rysunku przybliża jeden konkretny obszar geograficzny bazy klientów. Możesz zobaczyć okrągłe kółko w górnym rogu. Wykres następnie automatycznie dostosowuje inne połączone wykresy (wykres słupkowy przedstawiający rozkład wieku, wykres skrzynkowy pokazujący rozkład płci oraz macierz korelacji). Jak widać prawdopodobieństwo lojalności klientów w tym obszarze jest znacznie wyższe niż dla przeciętnego klienta - 82 proc. Decyzja o korzystaniu z łatwych w użyciu interaktywnych narzędzi analitycznych nie jest jednak właściwa dla wszystkich firm. Dojrzałe firmy zajmujące się analizą i inteligencją maszynową skłaniają się ku robieniu wszystkiego samemu, zwłaszcza jeśli chodzi o analizę w ramach komercyjnych produktów i usług związanych z danymi, które łączą klientów. Na przykład istnieje powszechne przekonanie, że bez unikalnego projektu interfejsu użytkownika wizualizującego dane i spostrzeżenia, nie będzie on wyróżniał się na tle konkurencji. Pamiętaj jednak, aby przemyśleć,

na czym skoncentrować swoją pracę programistyczną; chcesz opracować insighty lub narzędzie do wizualizacji Twoich insightów? Różnicowanie jest ważne, ale skoncentruj wysiłki firmy na właściwych zadaniach

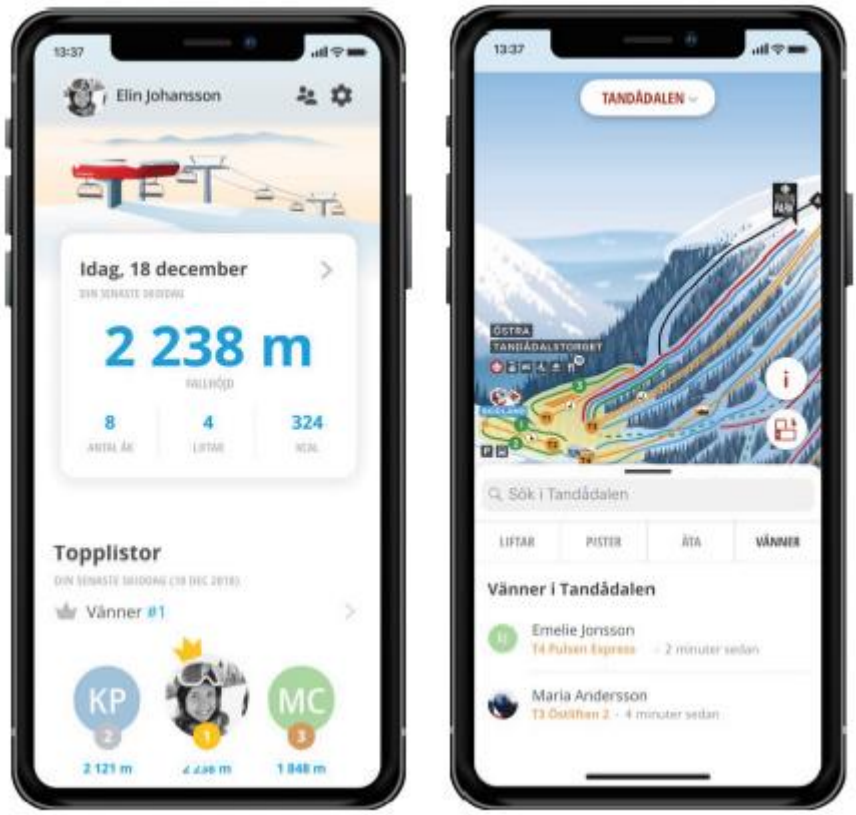


Firmy zajmujące się narzędziami analitycznymi inwestują dużo pieniędzy, kompetencji i czasu, aby te narzędzia były przyjazne dla użytkownika i najnowocześniejsze w większości wymiarów. W tym właśnie się specjalizują. Większość narzędzi można również dostosować do różnych potrzeb, więc przed rozpoczęciem tworzenia całkowicie własnego rozwiązania należy obliczyć alternatywny koszt i czas. Do celów czysto wewnętrznych gotowe narzędzia analityczne są zwykle bardziej ekonomiczne i szybsze od pomysłu do wglądu; oferować bardziej stabilne środowisko produkcyjne; i mieć możliwość udostępnienia analityki większej liczbie i różnym typom pracowników, wspierając wdrażanie organizacji opartej na danych w różnych segmentach firmy.

Aplikacje, strony internetowe i interfejsy produktów/usług jako modele dostarczania

Kiedy Twoją ambicją jest wykorzystanie aplikacji, witryn internetowych lub istniejących interfejsów produktów i usług jako modeli dostarczania do różnicowania danych, wszystko sprowadza się do udostępniania danych i wyników klientom. Można to zrobić, udostępniając im własne dane użytkowników za pośrednictwem kanału komunikacji, z którego korzystają, lub kanału komunikacji oferowanego przez Twoją firmę. Na przykład operator komórkowy może udostępniać swoim abonentom dane dotyczące ich własnych kosztów i wykorzystania, najlepszej oferty subskrypcji opartej na wzorcu użytkowania, usług opartych na lokalizacji i nie tylko. To wzmacnia subskrybentów i daje im kontekstowe zrozumienie, w jaki sposób faktycznie używają telefonów komórkowych, co oznacza, że mogą mieć większą kontrolę nad swoim obecnym i przyszłym użytkowaniem, w tym kosztami. Jednocześnie wzmacnia pozycję firmy udostępniającej dane, ponieważ wysyła sygnał przejrzystości do swoich klientów i generuje zaufanie, które może wzmocnić postrzeganie marki. Innym przykładem tego, jak firma udostępnia własne dane użytkowników, aby wzmocnić swoją markę, jest nordycka firma Skistar, która posiada zaplecze do narciarstwa zjazdowego w krajach skandynawskich. Firma posiada aplikację, w której można założyć konto i wgrać numer identyfikacyjny z cyfrowego karnetu narciarskiego. Karnet narciarski nadaje się do wielokrotnego użytku, dopóki się nie zepsuje, a po prostu reaktywujesz go, uiszczając odpowiednią opłatę, gdy jest potrzebny na nowy okres. Karnet narciarski automatycznie łączy się z systemem narciarskim w ciągu dnia na stoku za każdym razem, gdy korzystasz z wyciągu narciarskiego. Aplikacja dostarcza danych o liczbie przejechanych jazd, przebytym dystansie, osiągniętych metrach wysokości, spalonych kaloriach i tak dalej. Wszystko to jest ładnie agregowane

na dzień, tydzień, miesiąc lub rok. Pozwala także łączyć się ze znajomymi, dzięki czemu możesz porównać swoje wyniki. Rysunek przedstawia dwa widoki dostępne dla narciarzy korzystających z ośrodków narciarskich Skistar.



Pozorny przykład tego, jak Skistar zwraca dane narciarzom, łącząc cyfrowy karnet narciarski z ich aplikacją w celu generowania prostych, ale zabawnych informacji. Innym przykładem jest firma internetowa, która wykorzystuje dane wygenerowane na swojej stronie głównej, aby oferować swoim klientom inne polecane produkty lub usługi w oparciu o poprzednie wzorce zakupów w witrynie lub oferując obniżone nagrody użytkownikom, którzy wykazują szczególne zainteresowanie określonymi przedmiotami. Witryny mogą być również przydatne jako modele dostarczania produktów danych, takich jak informacje, które sprzedajesz. Wynikiem wglądu może być kokpit lub podobna wizualizacja. Po opublikowaniu pulpitu nawigacyjnego możesz udostępnić klientowi bezpieczne łącze do jego własnej witryny internetowej, w której będzie mógł korzystać ze spostrzeżeń za pomocą interaktywnego widoku, a także pobrać plik .pdf z widokiem statycznym. W przypadku istniejącego produktu lub usługi, które oferujesz, możesz użyć nowych lub istniejących danych, które nie były wcześniej udostępniane klientom. Dane mogą być następnie dodawane do tego samego interfejsu systemu, co poprzednio – na przykład systemu finansowego. Celem byłoby tutaj zwiększenie doświadczenia poprzez dodanie nowych danych, które mogłyby poprawić postrzeganie tego samego produktu lub usługi przez użytkownika bez jego funkcjonalnej poprawy.

Istniejące produkty i usługi

Integrując dane jako dane wejściowe lub kluczowy zasób z istniejącym produktem lub usługą i kierując swoją ofertę na dane, będziesz w stanie rozróżniać dane. W tym przypadku oryginalny produkt lub usługa nie byłyby produktem danych, ale mogłyby wykorzystywać różne formy danych do celów kontekstowych, a nie jako główną siłę napędową. Byłoby tak bez względu na to, czy produkty byłyby

sprzętowe, programowe, czy inne, oraz niezależnie od tego, czy produkty były lokalnie, czy też zostały zwirtualizowane i wdrożone za pośrednictwem usługi w chmurze. Model dostarczania obejmuje ulepszanie istniejących produktów lub usług lub identyfikowanie nowych możliwości związanych z produktami danych poprzez wykorzystanie danych. Jednym z przykładów takiego podejścia jest rozpoczęcie korzystania z opartego na danych i predykcyjnego podejścia do istniejącej oferty usług. Jest to szczególnie interesujące w przypadku ofert usług działających w czasie rzeczywistym, takich jak rozległe i złożone operacje sieci telekomunikacyjnej. Bez podejścia opartego na danych i bez pomocy technik, takich jak analityka predykcyjna i uczenie maszynowe/sztuczna inteligencja, usługi mają tendencję do reagowania na awarie lub różnego rodzaju alarmy. Ale gdy dane i modele są wykorzystywane proaktywnie do identyfikowania wzorców w danych, pomaga to zrozumieć, co powoduje pewne problemy, umożliwiając przewidywanie i zapobieganie wystąpieniu problemów w przyszłości. To z kolei poprawi wydajność usług i jakość sieci, a także satysfakcję z obsługi klienta, operatora sieci, a nawet użytkowników końcowych sieci komórkowej, takich jak Ty i ja.

Pliki do pobrania

W przypadku ofert związanych z maklerowaniem danych i informacji wybranym modelem dostarczania są często pliki do pobrania. Jeśli oferujesz zestaw danych wystarczająco mały, aby umieścić go w pliku i pobrać ze strony internetowej, jest to doskonały model dostarczania. Może tak być na przykład w przypadku pliku z danymi testowymi do uczenia modelu. Pliki do pobrania działają również, gdy sprzedajesz samodzielne rozwiązania analityczne lub uczenie maszynowe modele. Innym przykładem, kiedy ma to zastosowanie, jest sytuacja, w której oferowana jest usługa Insightas-a-Service lub jakiś rodzaj raportu, który ma zostać dostarczony. Zwykle rozmiar skompilowanego raportu ze spostrzeżeniami, zaleceniami i różnymi wizualizacjami statystycznymi całkiem dobrze pasuje do formatu do pobrania. Warto jednak zastanowić się, skąd klient pobiera pliki. Dobrym pomysłem jest zbudowanie łatwej w użyciu, ale bezpiecznej strony internetowej, na której będą mogli uzyskać dostęp do plików. W tej witrynie możesz również skorzystać z okazji, aby poinformować klienta o innych obecnych i przyszłych produktach, które oferujesz, a nawet otworzyć witrynę dla innych firm, aby kupić miejsca reklamowe dla innych powiązanych produktów danych dla klientów. Po prostu upewnij się, że wybierasz firmy, których nie postrzegasz jako obecnych lub potencjalnych przyszłych konkurentów, i utrzymuj witrynę prostą i przejrzystą, skupiając się głównie na dostarczanych plikach. Ułatw im dostęp do tego, czego szukają.

API

API to interfejs programowania aplikacji, zbiór jasno zdefiniowanych metod komunikacji pomiędzy różnymi komponentami, takimi jak systemy internetowe, systemy operacyjne, systemy bazodanowe, sprzęt komputerowy czy biblioteki oprogramowania. Korzystanie z interfejsu API jako metody dostarczania jest przydatne, gdy klienci chcą mieć bezpośredni dostęp do produktu lub usługi danych (danych, modelu lub wglądu) w celu zintegrowania produktu lub wyników bezpośrednio z ich środowiskiem systemowym. Interfejsy API mogą być również przydatne, gdy sprzedajesz określoną funkcję uczenia maszynowego jako usługę i uwzględniasz infrastrukturę niezbędną do uruchomienia modelu. Oznacza to, że nie sprzedajesz samego modelu, a jedynie możliwość korzystania z modelu uczenia maszynowego. Podejście oparte na modelu dostarczania polega na tym, że klient przesyła dane do Ciebie za pomocą interfejsu API, a następnie uruchamia model w Twoim środowisku. Jest to uczenie maszynowe jako usługa, a także forma pośrednictwa w infrastrukturze. Konkretnym tego przykładem jest Amazon, który oferuje tego typu usługi, korzystając ze swojego środowiska chmury i algorytmu uczenia maszynowego do rozpoznawania obrazów do różnych celów, takich jak rozpoznawanie i analiza twarzy, wykrywanie obiektów i aktywności, wykrywanie niebezpiecznych treści, wykrywanie celebrytów i nawet analiza tekstu na obrazach. Oferta rozpoznawania obrazów Amazon umożliwia

przeszukiwanie kolekcji obrazów pod kątem podobnych twarzy poprzez przechowywanie metadanych twarzy za pomocą funkcji API IndexFaces. Następnie możesz użyć funkcji SearchFaces, aby zwrócić dopasowania o wysokim stopniu ufności.

Usługi w chmurze

Usługa w chmurze to dowolna usługa udostępniana użytkownikom na żądanie za pośrednictwem Internetu z serwerów dostawcy chmury, w przeciwieństwie do usług świadczonych z własnych serwerów lokalnych firmy. Usługi w chmurze mają na celu zapewnienie łatwego, skalowalnego dostępu do aplikacji, zasobów i usług i są w pełni zarządzane przez dostawcę usług w chmurze. Przykładami niektórych znanych dostawców usług w chmurze są Amazon, Microsoft i Google. Mówiąc, że korzystasz z usług w chmurze jako platformy dostarczania, możesz oferować usługę w chmurze jako infrastrukturę lub usługę platformy dla różnych usług, takich jak przechowywanie danych, obliczanie danych lub aplikacje. Ale może również odnosić się do wykorzystania usługi w chmurze jako platformy dostarczania danych, wglądu i pośrednictwa modeli lub sieci dostarczania danych dla rynku. Ponieważ usługa w chmurze może dynamicznie skalować się w celu zaspokojenia potrzeb użytkowników, a dostawca usług dostarcza sprzęt i oprogramowanie niezbędne do świadczenia usługi, firma nie musi dostarczać ani wdrażać własnych zasobów ani przydzielać personelu IT do zarządzania usługi. To sprawia, że usługa w chmurze jest interesującym modelem dostarczania wielu różnych rodzajów produktów i usług związanych z danymi.

Rynki internetowe

Internetowe platformy handlowe są czasami określane również jako internetowe rynki handlu elektronicznego. Platforma handlowa to rodzaj witryny handlu elektronicznego, w której produkty lub usługi są dostarczane przez wiele stron trzecich, a transakcje są przetwarzane przez operatora platformy handlowej. Rynki internetowe są jak platformy dla wielu graczy i dobrze nadają się do wspierania rynków produktów danych i ofert usług danych. Już teraz są ważnymi mediami i mają potencjał, aby w przyszłości stać się głównym modelem dostarczania danych, napędzającym zawieranie transakcji w zakresie danych, analityki i sztucznej inteligencji. Daje również możliwości reklamowe w wielu kanałach i przemyśle. Rynek danych lub rynek danych to sklep internetowy, w którym ludzie mogą kupować dane. Rynki danych zazwyczaj oferują różne rodzaje danych dla różnych rynków i z różnych źródeł. Typowe rodzaje sprzedawanych danych obejmują analizy biznesowe, reklamy, dane demograficzne, dane osobowe, badania i rynek. Dane typy mogą być mieszane i strukturyzowane na różne sposoby. Dostawcy danych mogą oferować dane w określonych formatach dla klientów indywidualnych. Dane sprzedawane na tych platformach są wykorzystywane przez wszelkiego rodzaju firmy, rządy, agencje wywiadu gospodarczego i rynkowego oraz wiele rodzajów analityków. Rynki danych mnożyły się wraz z rozwojem big data, ponieważ ilość danych gromadzonych przez rządy, firmy, strony internetowe i usługi wzrosła, a wszystkie te dane są coraz częściej uznawane za aktywa. Rynki danych są często zintegrowane z usługami w chmurze.

Licencje do pobrania

Licencja na oprogramowanie to rodzaj licencji, która służy do określania zasad dotyczących tego, w jaki sposób można lub nie można używać danego oprogramowania. Po pobraniu lub zakupie oprogramowania musisz zgodzić się z licencją, aby z niego korzystać. W przypadku modelu biznesowego, takiego jak pośrednictwo w infrastrukturze, licencje do pobrania na różne typy oprogramowania analitycznego są powszechne i przydatne. Licencje na oprogramowanie do pobrania są zwykle wyposażone w ograniczenia czasowe, które uniemożliwiają korzystanie z nich po upływie daty wygaśnięcia, chyba że licencja zostanie odnowiona.

Usługi online

Usługa online to ogólny termin odnoszący się do wszelkich informacji i usług świadczonych przez Internet. Usługi te nie tylko umożliwiają abonentom komunikowanie się ze sobą, ale także zapewniają nieograniczony dostęp do informacji. Usługi online mogą być proste lub złożone. Podstawowa usługa online może pomóc abonentom w zdobyciu potrzebnych danych przez wyszukiwarkę, a złożoną może być aplikacja o kredyt hipoteczny online z banku. Usługi online mogą być bezpłatne lub płatne. Usługa online jest odpowiednia dla ofert takich jak usługi doradcze w zakresie danych i analiz, które nie muszą być świadczone na miejscu. Wielu analityków danych znalazło lukratywny biznes, sprzedając swoją wiedzę specjalistyczną online, dostarczając analitykę i wiedzę na temat uczenia maszynowego/sztucznej inteligencji w zakresie zaleceń i strategii opartych na danych, a także modeli i rozwiązań uczenia maszynowego/sztucznej inteligencji.

Usługi na miejscu

Usługi na miejscu to model dostawy, który odnosi się do usług, które odbywają się w tym samym lokalu lub w tej samej lokalizacji, co klient. Ten rodzaj usługi jest zwykle potrzebny, gdy nie można świadczyć usług poza siedzibą lub klient jest kompletnym nowicjuszem. W porównaniu z usługami zewnętrznymi lub usługami online, konfiguracja pomocy technicznej na miejscu zajmuje więcej czasu i jest zwykle droższa.

Dziesięć powodów, dla których warto opracować strategię analizy danych

W tej książce opisano wiele wyzwań, z jakimi przyjdzie Ci się zmierzyć, rozpoczynając przygodę z nauką o danych w Twojej firmie. Podkreśla to, co jest fundamentalne, ao czym nie należy zapominać, ale także wskazuje obszary szczególnego zainteresowania i wybory o szczególnym znaczeniu. Jedną z rzeczy, których nie zrobił (jeszcze), jest przedstawienie argumentu, dlaczego ważne jest, abyś opracował i udokumentował wszystkie swoje strategiczne ambicje w strategii analizy danych. O tym jest ten rozdział. Ciesz się!

Rozszerzenie swojego poglądu na naukę o danych

Poświęcenie czasu na opracowanie strategii analizy danych ma kluczowe znaczenie. Zmusza Cię do dowiedzenia się więcej o tym, czym naprawdę jest data science, zanim zaczniesz inwestować i dokonywać ważnych wyborów. Posiadanie strategii zmniejsza ryzyko pominięcia ważnych kroków i rozważań po drodze. Chociaż nauka o danych jest mieszanką różnych dyscyplin - takich jak matematyka, statystyka i informatyka - nie myl się: jest to dyscyplina sama w sobie. Zrozumienie kluczowych pojęć i rozważań kierujących dziedziną nauki o danych jest niezbędne, ale często nawet tego nie robi. Uważam, że kiedy naprawdę zrozumiesz, na czym polega data science, spojrzysz na swoją firmę w innym świetle, z innej perspektywy. Będzie dla Ciebie oczywiste, co należy zrobić inaczej, i będziesz w stanie wyjaśnić, dlaczego tak jest. Następnie możesz zmotywować osoby wokół ciebie do wprowadzenia niezbędnych zmian, ponieważ w nauce o danych wszystko zaczyna się i kończy na danych. Być może wiele firm, które istnieją już od jakiegoś czasu, nie myśli o sobie, że są ustrukturyzowane w ten sposób w oparciu o dane, ale muszą nimi być, jeśli mają odnieść sukces w nowej erze danych i sztucznej inteligencji. Google używa danych jako punktu wyjścia do wszystkiego. Wykorzystując sztuczną inteligencję i techniki uczenia maszynowego do wykrywania wzorców i odchyleń w danych, Google może decydować w oparciu o dane, którą działalność biznesową wybrać i które obszary mają priorytetyzować i podejmować działania. W Google dane napędzają zmiany organizacyjne, nowe innowacje i priorytety biznesowe. A jego głównym hasłem przewodnim jest, jak można sobie wyobrazić, AI na pierwszym miejscu.

Dopasowanie poglądu firmy

Jeśli odpowiednio pokierujesz swoją strategią analizy danych, będziesz mieć możliwość zgromadzenia ludzi wokół możliwości biznesowych, które z pewnością wynikną z inwestycji w naukę danych. Ważne jest, aby sformułować tę wizję i misję oraz uchwycić je w strategii analizy danych uzgodnionej przez wszystkich interesariuszy. W ten sposób zapewniasz, że wszyscy są zaangażowani w określone cele i na wczesnym etapie są zakotwiczeni w strukturze organizacyjnej. Daje to mocne i solidne podstawy do ogromnej i pełnej wyzwań pracy, którą czeka Cię przyszłość. Jednak łatwiej powiedzieć niż zrobić, aby dostosować organizację do nauki o danych. Dlaczego? Cóż, na początek opinie ludzi na temat tego, czym jest nauka o danych i jak będzie ona przekształcać różne firmy, są dość zróżnicowane. Oznacza to, że nie zaczniesz od tego samego poziomu zrozumienia, co to znaczy wprowadzić do firmy naukę o danych. Jeśli niektórzy przystąpią do przedsięwzięcia, zakładając, że analityka danych może zostać dodana do zakątka firmy jako swego rodzaju dodatek i oczekuje się, że będzie generować wartość, będziesz mieć problemy w dalszej kolejności. Aby w pełni wykorzystać potencjał inwestycji w naukę danych, należy traktować ją jako dyscyplinę dominującą. Jeśli jesteś w stanie dostosować swoją firmę do takiego postrzegania i uchwycić szczegóły tego, jak będzie to traktowane w ramach strategii analizy danych, zapewniłeś swojej firmie sukces.

Tworzenie solidnej podstawy do wykonania

Faktycznie spisując swoje podejście i priorytety związane z nauką o danych, tworzysz podstawę planów niezbędnych do realizacji strategii. Pomaga kierować firmą we właściwym kierunku i zapewnia punkt odniesienia, na którym można polegać, gdy pojawiają się wyzwania i pojawiają się nowe możliwości. Istotnym elementem tak solidnej podstawy jest rysunek architektoniczny, na którym można zrealizować i wdrożyć swoją infrastrukturę. Wymaga to dość dużo szczegółowego myślenia w konfiguracji zespołu międzydomenowego, nie tylko po to, aby szczegółowo opisać podejście do konfiguracji i wykonania w różnych domenach, ale także przemyśleć, w jaki sposób zostanie to wykonane w konfiguracji opartej na danych i maszynie w całej w sposób płynny. Oczywiście strategia może z czasem ulec zmianie ze względu na zmieniające się potrzeby lub priorytety w firmie, a nawet ewoluującą technologię analizy danych. Ale niezależnie od tego daje solidne podstawy, na których możesz stanąć i zacząć od rozważania nowego kierunku lub modyfikowania planów dotyczących realizacji.

Wczesne realizowanie priorytetów

Szczerze mówiąc, dla średniej i dużej firmy inwestycja typu end-to-end w naukę danych jest kosztowna. Ale ponieważ zwiększony potencjał biznesowy jest znacznie większy, firmy zdają sobie sprawę z konieczności inwestowania w przyszłość napędzaną nauką o danych. Aby nie zgubić się w całej nieuniknionej złożoności na drodze do tej przyszłości, ważne jest, aby wcześnie zrozumieć i jasno określić swoje priorytety, a następnie spróbować się ich trzymać, gdy sprawy staną się trudniejsze. Pomoże ci to poprowadzić cię w trudnych okresach. Jakie zatem są typowe priorytety, które należy wziąć pod uwagę na początku? Cóż, ważną częścią twojej pracy nad opracowaniem strategii powinno być przyjrzenie się obecnej konfiguracji biznesowej. W ten sposób wymieniono niektóre z ważniejszych pytań, które należy rozważyć na początku:

* Co tak naprawdę chcesz zmienić?

* Jaki jest potencjał data science w Twojej branży?

* Jakie są Twoje prawdziwe oczekiwania?

* Jak w praktyce zrealizujesz swoje oczekiwania?

Stawianie celu w perspektywie

Stworzenie kompleksowej strategii analizy danych zmusza Cię nie tylko do wyznaczenia jasnych celów, ale także do rozważenia ich z wielu perspektyw. Nie chodzi tylko o potencjał biznesowy; chodzi również o prawa, obawy dotyczące prywatności lub inne względy etyczne związane z danymi, z których korzystasz. Rozważając kontekst fundamentalnej zmiany, która musi nastąpić w Twoim środowisku biznesowym podczas wprowadzania nauki o danych, być może będziesz musiał zastanowić się, w jakim stopniu Twoja firma zależy od następujących czynników:

* Dane, których możesz nie posiadać, co może oznaczać ograniczenia w użytkowaniu

* Konieczność zdigitalizowania wszystkich części Twojej firmy, aby stać się danymi napędzanymi od początku do końca

* Konieczność nowych ról, kompetencji i zestawów umiejętności w data science wśród pracowników i menedżerów

* Nowe przepisy i regulacje, które wcześniej nie obowiązywały

* Współpraca z nowymi dostawcami i partnerami związanymi z obsługą danych i infrastruktury

*Potencjalnie adresowanie do zupełnie nowej bazy klientów

Tworzenie doskonałej bazy do komunikacji

Tak, wdrożenie strategii analizy danych, dopasowanej i zakotwiczonej z głównymi interesariuszami, to dużo pracy, ale gdy już zostanie wdrożona, zapewni doskonałą bazę do komunikacji. Możesz łatwo wykorzystać napisaną strategię i przełożyć jej zdefiniowane cele i wyzwania na materiały prezentacyjne i cele firmy. Strategię można wykorzystać do zbudowania planu komunikacji dla różnych zidentyfikowanych grup docelowych, w tym różnych uzgodnionych priorytetów i rozważań dotyczących nowego sposobu myślenia i kultury, które chcesz egzekwować. Możesz przekształcić treść ze swojej strategii analizy danych w bardziej użyteczne formaty, takie jak często zadawane pytania, a następnie zamienić wybrane części w materiały zewnętrzne do komunikacji z klientami, partnerami i dostawcami.

Zrozumienie, dlaczego wybory są ważne

Strategia polega na dokonywaniu wyborów – wyborów dotyczących tego, do czego należy dążyć, a do czego nie warto. Jeśli twoją strategią jest próba zrobienia wszystkiego, jesteś zgubiony. To gorsze niż zła strategia; to w ogóle żadna strategia. Ponieważ wybory, których dokonasz w strategii, będą gwiazdą przewodnią dla nadchodzących wyborów i priorytetów podczas wyzwania polegającego na pełnym wprowadzeniu nauki o danych, absolutnie musisz dokonać właściwych wyborów na początku. Dokonywanie niewłaściwych wyborów w tym momencie z pewnością będzie miało poważny wpływ na ogólny sukces Twojej inwestycji w naukę danych. Co więc możesz zrobić, aby upewnić się, że dokonujesz właściwych wyborów?

* Poświęć trochę czasu, aby opracować właściwą strategię i pamiętaj o iteracji. Nie spiesz się!

* Zapewnij swoim głównym interesariuszom (i Tobie) podstawowy poziom zrozumienia w nauce o danych.

* Zaangażuj wewnętrznych i zewnętrznych ekspertów data science w tej dziedzinie, aby mieć pewność, że zyskasz szerokie i zróżnicowane spojrzenie na sytuację rynkową.

* Skorzystaj z wewnętrznych porad swoich analityków danych (jeśli istnieją) i pozwól im aktywnie przyczynić się do realizacji strategii.

* Angażuj głównych interesariuszy w decyzje dotyczące wyzwań i priorytetów międzydziedzinowych, nawet jeśli jest to trudne i uciążliwe.

* Dokonuj trudnych wyborów, ale bądź gotowy na dostosowanie się po drodze w zależności od lepszego zrozumienia lub zmieniających się warunków.

Wczesna identyfikacja ryzyka

Poświęcając czas na rozważenie zagrożeń w ramach strategii analizy danych, możesz nie tylko wcześnie wykryć zagrożenia, ale także potencjalnie zapobiec ich urzeczywistnieniu. Kluczem jest znalezienie dobrej struktury do wykorzystania przy identyfikowaniu potencjalnych zagrożeń, która pomoże ci przejść przez to niezbyt zabawne ćwiczenie. Wiem, że o wiele bardziej atrakcyjne jest myślenie o wszystkich nowych możliwościach, które może przynieść przyszłość, niż o tym, co może potencjalnie pójść nie tak z twoją inwestycją. Jednak jest to czas dobrze zainwestowany, aby to zrobić. Niektóre główne obszary ryzyka do rozważenia obejmują te opisane na tej liście:

* Dane: Czy kontrolujesz własność wszystkich danych, których będziesz potrzebować, aby zrealizować swoje wewnętrzne ambicje dotyczące efektywności lub zrealizować zewnętrzne możliwości

biznesowe? Jeśli nie, czy zapewniłeś sobie niezbędne prawa do danych do tego, co chcesz robić dzisiaj i potencjalnie w przyszłości?

* Kompetencje: Czy posiadasz umiejętności niezbędne do realizacji swojej strategii? Jeśli nie, czy określiłeś odpowiednie ambicje biznesowe w odniesieniu do dostępności takich kompetencji – biorąc pod uwagę czas potrzebny na wewnętrzne budowanie doświadczenia i/lub na przykład przyciągnięcie i zatrzymanie analityków danych wśród niewielkiej liczby dostępnych na rynku?

* Infrastruktura: Czy dokładnie zbadałeś ryzyko związane z Twoimi ambicjami architektonicznymi? (Przykłady obejmują przejście na open source lub nie, środowisko zwirtualizowane i oparte na chmurze lub nie, oraz rozproszone i lokalne konfiguracje lub scentralizowaną konfigurację.) Istnieje wiele zagrożeń związanych z wyborem architektury infrastruktury, a także wyzwaniami związanymi z implementacją (konfiguracje dla na przykład dane i dystrybucja obliczeń i automatyzacji ponad granicami.)

Dokładne rozważenie Twoich potrzeb w zakresie danych

Przeprowadzenie dokładnej inwentaryzacji potrzeb w zakresie danych ma kluczowe znaczenie, jeśli chodzi o pracę nad strategią danych. Zapewnia praktyczne zrozumienie priorytetów biznesowych, potrzeb infrastrukturalnych, aspektów prawnych i etycznych, aspektów zarządzania danymi, a także potencjału biznesowego tej strategii. Wszystko zaczyna się od danych. To takie proste. Inwentaryzacja danych powinna obejmować takie aspekty, jak:

* Klasyfikacja danych pod względem rodzaju, formatu, stopnia wrażliwości, punktu(-ów) zbierania i własności

* Grupowanie typów danych w kategorie danych o podobnych atrybutach (zmniejszenie liczby poszczególnych typów do obsługi)

* Potrzeba użycia pod względem wymaganego poziomu szczegółowości danych, częstotliwości zbierania i okresów przechowywania danych.

Po utworzeniu spisu możesz użyć go do stworzenia modelu danych, który pomoże Ci zrozumieć (i przygotować się na) interoperacyjność danych (które dane należy połączyć z jakimi i jak można je razem analizować?); w jaki sposób zapotrzebowanie na dane wpływa na konfigurację infrastruktury (jakie wymagania można wyprowadzić ze zbiorowego zapotrzebowania na dane?); i co należy chronić z punktu widzenia prawa i bezpieczeństwa (jakie przepisy ustawowe i wykonawcze mają zastosowanie do jakich typów danych i co to oznacza?).

Zrozumienie wpływu zmian

Strategia dotycząca danych to dobry sposób, aby dobrze zrozumieć całkowity zakres zmian, które są potrzebne do osiągnięcia swoich celów. Oznacza to, że możesz wcześnie rozpocząć planowanie niezbędnej zmiany kulturowej w sposobie myślenia i zachowaniach, która jest potrzebna. Umożliwia to, aby przejście było dobrze zaplanowane i proaktywne, tak aby nie przebiegało w pośpiechu, ale było wprowadzane krok po kroku. Kolejnym krokiem jest umożliwienie pracownikom dojrzałości w postrzeganiu, czym jest data science i co to umożliwi. Jednym z ważnych aspektów nauki o danych, który trzeba przyznać, jest obawa ludzi przed tym, jak wprowadzenie automatyzacji wpłynie na zwykłe miejsca pracy. (Jest to ściśle powiązane z dalszym strachem, że algorytmy uczenia maszynowego zastąpią potrzebę ludzi w miejscu pracy). zadania i role Twoich pracowników, jakie przyniesie korzyści i jakie możliwości otwierają się w wyniku zmiany. Kierowanie się danymi i inwestowanie w naukę danych oznacza również, że powinieneś kierować się tym samym sposobem myślenia podczas

zarządzania zmianami. Podejdź do programu transformacji z perspektywy opartej na danych i użyj technik analitycznych i uczenia maszynowego, aby zmierzyć i zrozumieć efektywność i wpływ zmian. Na przykład stosowanie metody takiej jak analiza nastrojów pozwala zrozumieć, jak zmiana jest postrzegana przez interesariuszy. Inne aspekty, które chcesz omówić, obejmują stopień, w jakim zmiana faktycznie zachodzi i czy istnieją konkretne role związane ze zmianą, które są bardziej wydajne niż inne. Co robią, czego nie robią inni? Definiując strategię nauki o danych, zyskujesz możliwość szerszego zrozumienia przez kierownictwo tego, czym naprawdę jest nauka o danych, jakie możliwości umożliwia i fundamentalny wpływ zmian, które nakłada nauka o danych.

Dziesięć błędów, których należy unikać podczas inwestowania w analitykę danych

Chociaż aby odnieść sukces, musisz skoncentrować się na celach strategii analizy danych, nie zaszkodzi również uczyć się na błędach innych. Ten rozdział zawiera listę dziesięciu wyzwań, z którymi wiele firm radzi sobie w niewłaściwy sposób. Każda sekcja nie tylko opisuje, czego należy unikać, ale także wskazuje właściwe podejście do sytuacji.

Nie toleruj ignorancji przez najwyższe kierownictwo w zakresie nauki o danych

W obszarze data science pojawia się fundamentalne nieporozumienie dotyczące grupy docelowej szkolenia z data science. Powszechnie uważa się, że tak długo, jak poprawi się zestaw umiejętności samych naukowców zajmujących się danymi lub inżynierów oprogramowania, którzy kształcą się na analityków danych, jesteś na miejscu. Jednak przyjmując takie podejście, firma ponosi znaczne ryzyko wyobcowania zespołu data science od reszty organizacji. Często zapomina się o menedżerach i liderach. Jeśli menedżerowie nie rozumieją lub nie ufają pracy wykonanej przez analityków danych, wynik nie zostanie wykorzystany w organizacji, a spostrzeżenia nie zostaną zastosowane. Tak więc głównym pytaniem, jakie należy zadać, jest to, jak zapewnić pełne wykorzystanie inwestycji w badania danych, jeśli wyniki nie mogą być zinterpretowane przez kierownictwo. Jest to jeden z najczęstszych błędów popełnianych obecnie przez firmy, a faktem jest, że istnieje również niewiele szkoleń i coachingu dostępnego dla menedżerów liniowych i liderów. Ale bez pewnego poziomu zrozumienia nauki o danych na poziomie zarządzania, jak można wdrożyć właściwą strategię i jak można oczekiwać, że kierownictwo odważy się wykorzystać wyniki statystyczne do podejmowania istotnych decyzji? Bez zrozumienia przez kierownictwo data science nie tylko trudno jest uchwycić pełną szansę biznesową dla firmy, ale może to również prowadzić do dalszej alienacji zespołu data science lub całkowitego zakończenia pracy zespołu.

Nie wierz, że sztuczna inteligencja to magia

Nauka o danych to przede wszystkim dane, statystyki i algorytmy. Nie ma w tym nic magicznego – maszyna robi to, co jej każe. Jednak pogląd, że maszyna może się uczyć, sprawia, że niektórzy sądzą, iż ma ona pełną zdolność do samodzielnego uczenia się. W pewnym stopniu to prawda – maszyna może się uczyć – ale jest to poprawne tylko w granicach, które dla niej ustawiłeś. (Innymi słowy, bez magii!) Maszyna nie może sama rozwiązać problemów, chyba że pozwoli się jej na opracowanie takiego projektu. Ale to zaawansowana technologia, a nie dzisiejsza rzeczywistość. Przecenianie tego, co sztuczna inteligencja może zrobić dla Twojej firmy, może naprawdę skierować Cię na złą ścieżkę, budując oczekiwania, których nigdy nie można spełnić. Może to prowadzić do poważnych konsekwencji zarówno wewnątrz firmy, jak i na zewnątrz, z wpływem nie tylko na zaufanie i wiarygodność, ale także na wyniki finansowe. O ile nie należy lekceważyć potencjału sztucznej inteligencji, należy również unikać przeciwnej skrajności, gdzie jej potencjał jest przeceniany. Powtarzam: Sztuczna inteligencja to nie magia. Tak, nazywa się to sztuczną inteligencją, ale bardziej poprawną definicją jest inteligencja algorytmiczna. Czemu? Ponieważ w ostatecznym rozrachunku bardzo zaawansowana matematyka jest stosowana do ogromnych ilości danych, z możliwością dynamicznej interakcji ze zdefiniowanym środowiskiem w czasie rzeczywistym.

Nie podchodź do analizy danych jako wyścigu na śmierć między człowiekiem a maszyną

Niektórzy ludzie wierzą, że automatyzacja zadań, oparta na przewidywaniach uczenia maszynowego, naprawdę oznacza koniec ludzi w miejscu pracy. Ta prognoza nie jest taką, w którą wierzę. Oznacza to jednak znaczącą zmianę w zakresie kompetencji i zestawów umiejętności, a także zmianę, które role zawodowe będą istotne i jakie rodzaje obowiązków będą się koncentrować w miejscu pracy. Podobnie jak wprowadzenie Internetu w miejscu pracy, wprowadzenie sztucznej inteligencji w bardziej

powszechnym formacie zmieni rodzaje pracy i sposób ich wykonywania. Będzie o wiele mniej „praktycznej" pracy, nawet w branży oprogramowania. I tak, maszyny najprawdopodobniej wykonają wiele podstawowego rozwoju oprogramowania, co oznacza, że ludzie z branży związanej ze sprzętem nie zostaną zastąpieni jedynymi. Ostatecznie, w zasadzie wszyscy ludzie będą pod wpływem, ponieważ uczenie maszynowe/sztuczna inteligencja i możliwości automatyzacji będą się rozszerzać i ewoluować poza to, co jest możliwe dzisiaj. Oznacza to jednak również, że ludzie mogą przejść do wykonywania innych zadań, które różnią się od tych, które wykonujemy obecnie – na przykład zarządzania i monitorowania modeli i algorytmów oraz ich wydajności lub ustalania priorytetów i działania jako rozwiązanie awaryjne dla ludzi we współpracy z maszyna. Innymi typowymi zadaniami człowieka może być zarządzanie kwestiami prawnymi związanymi z danymi, ocena etycznych aspektów podejmowania decyzji w oparciu o algorytmy lub dążenie do standaryzacji w nauce o danych. Można powiedzieć, że nowe ludzkie zadania będą koncentrować się na zarządzaniu maszynami, które zarządzają pierwotnymi zadaniami - zadaniami, które wcześniej postrzegano jako nudne i powtarzalne lub zbyt skomplikowane do wykonania. Ten biznes „przeciwstawiający człowieka maszynie" nie jest sposobem na podejście do wdrożenia nauki o danych. Pozwolenie na takie sformułowanie narracji może przestraszyć Twoich pracowników, a nawet skłonić ich do odejścia z firmy, co nie jest tym, czego chcesz. Twoi pracownicy są cennymi zasobami, których będziesz potrzebować również na kolejnych etapach, ale być może w nowych rolach i z nowymi nabytymi zestawami umiejętności. Dowiedz się, co technologia uczenia maszynowego/sztucznej inteligencji może zrobić dla określonej branży. Liderzy firm, którzy rozumieją, jak wykorzystać te techniki w zrównoważonym podejściu między człowiekiem a maszyną, aby zwiększyć ogólną wydajność i pozwolić firmie rozwinąć się poza jej bieżącą działalność, są liderami, których firmy odniosą sukces.

Nie lekceważ potencjału sztucznej inteligencji

Choć może się to wydawać dziwne, niektóre firmy po prostu nie rozumieją, jak naprawdę transformująca jest sztuczna inteligencja. Nie widzą fundamentalnej zmiany, która już zaczyna przekształcać społeczeństwo, i nie mogą postrzegać sztucznej inteligencji jako czegoś innego niż tylko kolejną technikę oprogramowania lub zestaw nowych języków programowania. Kluczem jest tutaj a) poświęcenie czasu, aby naprawdę zrozumieć, na czym tak naprawdę polega data science i b) nie bać się skorzystać z pomocy ekspertów w celu zidentyfikowania i wyjaśnienia strategicznego potencjału Twojej konkretnej firmy. Ponieważ obszar data science jest złożony, wymaga wiedzy dziedzinowej i doświadczenia zarówno w zakresie opracowania strategii, jak i jej realizacji. Wymaga również umiejętności odczytywania i interpretowania kierunku, w którym porusza się rynek w tym obszarze. Nie doceniając wpływu, jaki sztuczna inteligencja może mieć na Twój biznes, ryzykujesz znaczne ograniczenie przyszłej ekspansji Twojej firmy. Później, kiedy naprawdę zrozumiesz prawdziwy potencjał, zauważysz, że wchodzisz do gry zbyt późno i jesteś wyposażony w niewłaściwy zestaw umiejętności. Możesz w końcu zostać wyrzucony z biznesu przez konkurencję, która znacznie wcześniej dostrzegła potencjał i dlatego wcześniej i mądrzej zainwestowała w sztuczną inteligencję.

Nie lekceważ potrzebnego zestawu umiejętności w zakresie analizy danych

Typową oznaką niedoinwestowania przez firmy w analitykę danych jest znalezienie małych, odizolowanych wysp kompetencji w zakresie analityki danych, rozproszonych w różnych częściach dużej firmy. W mniejszych firmach podobny objaw widać, gdy mały, ale kompetentny zespół data science pracuje nad najważniejszym projektem w firmie, ale jedynym poza zespołem, który zdaje sobie sprawę z jego znaczenia, jest osoba z zewnątrz, taka jak Ty. Oba te przykłady świadczą o tym, że najwyższe kierownictwo w firmie nie zrozumiało potencjału data science. Po prostu zdali sobie sprawę, że coś się dzieje w tym obszarze na rynku i po prostu podążają za trendem, aby upewnić się, że data science ich nie ominie. Jeśli poziom świadomości i kompetencji kierownictwa nie ulegnie poprawie,

obszar będzie nadal niedoinwestowany, rozproszony w taki sposób, że nie będzie mógł osiągnąć masy krytycznej, a tym samym nie będzie można go powiększyć na późniejszym etapie.

Nie myśl, że pulpit jest celem końcowym

Dla kogoś znającego się na danych może zabrzmieć dziwnie stwierdzenie, że każdy może myśleć, że głównym rezultatem nauki o danych jest tablica rozdzielcza. Zapewniam jednak, że jest to powszechne nieporozumienie. To nie tylko błąd — jest to również jeden z głównych powodów, dla których wiele firm ponosi porażkę z inwestycjami w naukę danych. W wielu firmach kierownictwo ma tendencję do myślenia, że głównym celem analityki i sztucznej inteligencji jest wykorzystanie wszystkich dużych zbiorów danych, które zostały wpompowane do drogiego jeziora danych, do automatyzacji zadań i raportowania postępów. Biorąc pod uwagę takie nastawienie, nie powinno dziwić, że głównym celem kierownictwa byłoby wykorzystanie tych technik do odpowiadania na pytania statystycznie sprawdzonymi metodami, które mogłyby dawać wyniki, które można zwizualizować na ładnie wyglądającym pulpicie nawigacyjnym. Dla kogoś nowego w dziedzinie nauki o danych może to wydawać się dobrym podejściem. Niestety byliby w błędzie. Aby było absolutnie jasne, głównym celem analityki i uczenia maszynowego/sztucznej inteligencji nie jest po prostu robienie tego, co zawsze robiłeś, ale korzystanie z większej liczby maszyn. Chodzi o to, aby móc wyjść poza to, co jesteś w stanie zrobić dzisiaj i pokonać nowe granice. Gdyby jedynym celem końcowym było stworzenie kokpitu, aby odpowiedzieć na niektóre pytania zadane przez menedżera, nie byłoby potrzeby tworzenia organizacji opartej na danych. Pomysł polega na tym, że w organizacji opartej na danych wszystko zaczyna się od danych, a nie od menedżera i pulpitu nawigacyjnego. Punktem wyjścia jest to, co wskazują dane, na co należy spojrzeć, przeanalizować, zrozumieć i podjąć działania. Analiza powinna mieć charakter predykcyjny, aby organizacja była proaktywna, a jej działania prewencyjne. Rolą pulpitu nawigacyjnego powinno być zaskakiwanie Cię nowymi spostrzeżeniami i odkrywanie nowych pytań, które powinieneś zadawać, a nie odpowiadanie na pytania, które już wymyśliłeś. Powinno to umożliwić zespołom monitorowanie i uczenie się na podstawie bieżących działań zapobiegawczych. Kokpit powinien również wspierać odkrywanie przez ludzi lub maszyny potencjalnych trendów i prognoz w celu podejmowania długofalowych decyzji strategicznych. W prawdziwym świecie kroki potrzebne do zaprojektowania pulpitu nawigacyjnego stają się najważniejszymi zadaniami do omówienia i skupienia się. Często kokpity menedżerskie kończą się wszystkim, co jest robione w programie wdrażania nauki o danych, całkowicie tracąc sedno o utrzymywaniu otwartego i eksploracyjnego podejścia do danych. Dzieje się tak, ponieważ deska rozdzielcza jest najprostszym i najbardziej konkretnym rezultatem do zrozumienia i utrzymania się w tym nowym, złożonym i stale zmieniającym się środowisku. W tym sensie działa jak kula dla tych, którzy nie chcą lub nie są w stanie uchwycić pełnego potencjału biznesu opartego na danych. Narażasz się na ogromne ryzyko, że przegapisz cały sens kierowania się danymi, gdy punktem wyjścia jest zaprojektowanie pulpitu nawigacyjnego i zadawanie wszystkich pytań od samego początku. W ten sposób zakładasz, że już wiesz, które pytania są ważne. Ale jak możesz być tego pewien? W społeczeństwie i na rynku przechodzącym obecnie ogromne przemiany, jeśli nie spojrzysz najpierw na dane i nie pozwolisz algorytmom na pracę polegającą na znalezieniu ukrytych tam wzorców i odchyleń, możesz w końcu spojrzeć na całkowicie niewłaściwy problem dla Twojej firmy .

Nie zapomnij o etycznych aspektach sztucznej inteligencji

Do czego właściwie odnosi się etyka sztucznej inteligencji i dlaczego uważasz, że ma to ogromne znaczenie? Cóż, jest wiele aspektów związanych z ideą etyki w AI, z których wiele może mieć poważny wpływ na wyniki sztucznej inteligencji. Jednym z oczywistych, ale ważnych rozważań etycznych jest potrzeba unikania stronniczości maszyn w algorytmach – stronniczości, w której ludzkie uprzedzenia dotyczące rasy, płci, klasy lub innych dyskryminujących aspektów są nieświadomie wbudowane w

modele i algorytmy. Zwykle ludzie wierzą, że nie mają stronniczych opinii, ale prawda jest taka, że wszyscy je mamy, mniej więcej. Ludzie skłaniają się w jednym kierunku, podświadomie lub nie. Modelowanie tej tendencji do samouczących się algorytmów może mieć poważne konsekwencje dla wydajności algorytmów firmy. Jednym z przykładów, który przychodzi mi do głowy, jest innowacyjny, internetowy i oparty na sztucznej inteligencji konkurs piękności. Algorytm nauczył się wyszukiwać dziesięć najpiękniejszych kobiet w USA, korzystając wyłącznie z cyfrowych zdjęć kobiet. Jednak przyglądając się wynikom konkursu, stało się jasne, że coś musiało pójść nie tak: wszystkie z dziesięciu najpiękniejszych kobiet wybranych przez algorytm były białe, blond i niebieskookie. Tak więc, ponownie studiując algorytm, okazało się, że zestaw treningowy użyty do algorytmu zawierał większość białych, blond i niebieskookich kobiet, co nauczyło maszynę, że jest to pożądany wygląd.

Inne aspekty oprócz stronniczości maszyn obejmują takie obszary, jak wykorzystanie danych osobowych, odtwarzalność wyników poza środowiskiem laboratoryjnym oraz wyjaśnialność spostrzeżeń lub decyzji AI. Warto również zauważyć, że ten ostatni aspekt jest obecnie prawem w ramach RODO (ogólnego rozporządzenia o ochronie danych) w UE.

Względy etyczne służą naszej własnej ochronie, ponieważ inteligencja maszyn ewoluuje w czasie. O takich aspektach trzeba myśleć wcześnie. Jest to nie tylko podstawowy aspekt, który należy wziąć pod uwagę w ramach inwestycji w naukę danych, ale w rzeczywistości niezwykle ważne jest, aby wziąć pod uwagę już na samym początku, projektując modele biznesowe, architekturę, infrastrukturę, sposoby pracy i same zespoły. Niechęć do łamania prawa jest oczywiście ważna, ale zapewnienie trwałej i godnej zaufania ewolucji sztucznej inteligencji w Twojej firmie jest znacznie ważniejsze.

Nie zapomnij wziąć pod uwagę praw do danych

Jednym z najczęstszych błędów, gdy stajemy się napędzani danymi, jest zapomnienie o dokonaniu właściwej analizy, które dane są potrzebne. Nawet jeśli Twoja główna ambicja związana z inwestycją w naukę danych koncentruje się na wewnętrznej wydajności i operacjach opartych na danych, nadal jest to podstawowy obszar, którym należy się zająć. Po przeanalizowaniu zapotrzebowania na dane nie jest niczym niezwykłym odkrycie, że potrzebujesz innych rodzajów danych, niż początkowo sądziłeś. Mogą to być dane inne niż tylko dane generowane wewnętrznie, należące do Ciebie. Przykładem mogą być błędy znalezione w Twoich produktach lub usługach, a może dane związane z wydajnością. Może to być nawet bardziej wrażliwy rodzaj danych, który należy do kategorii danych dotyczących prywatności, związanych z tym, jak Twoje produkty lub usługi są wykorzystywane przez Twoich klientów. Prywatność danych to obszar, który zyskuje coraz większą uwagę w społeczeństwie, w którym konsumenci mają większą świadomość tego, w jaki sposób wykorzystywane są ich dane, a także w zakresie nowych przepisów i regulacji dotyczących danych. Jednym z konkretnych przykładów jest ogólne prawo o ochronie i regulacji danych (RODO), wprowadzone w 2018 r. w UE, które nakłada wysokie kary na osoby naruszające. Chociaż możesz nie mieć żadnych planów dotyczących zarabiania na danych lub tworzenia nowych produktów w oparciu o te dane, cała kwestia praw jest nadal kluczowa – nawet jeśli chcesz tylko przeanalizować dane, aby lepiej zrozumieć swoją działalność, ulepszyć i unowocześnić obecne portfolio lub po prostu poprawić wydajność swojej operacji. Bez względu na powody, dla których korzystasz z danych, nadal potrzebujesz praw, aby z nich korzystać! Absolutnie konieczne jest zajęcie się tym na wczesnym etapie w ramach opracowywania strategii danych. Jeśli tego nie zrobisz, możesz skończyć z naruszeniem prawa regulującego wykorzystanie i własność danych lub utkniesz w sytuacji, gdy nie będziesz w stanie sprzedać swojego nowego fantastycznego produktu lub usługi, ponieważ wykorzystuje dane, do których nie masz prawa.

Nie ignoruj skali potrzebnych zmian

Jeśli nie poświęcisz czasu, aby odpowiednio naszkicować różne scenariusze zmian dla swojej firmy podczas wprowadzania strategii analizy danych, najprawdopodobniej poniesiesz porażkę. Fundamentalna zmiana potrzebna firmie, aby stać się naprawdę napędzaną danymi, analizą i maszynami, jest znacząca i nie należy jej lekceważyć. Oto najczęstsze błędy w data science związane z zarządzaniem zmianą:

* Niedocenianie zakresu zmiany i nietraktowanie wystarczająco poważnie tego, co ma się wydarzyć

* Brak uznania, że wprowadzenie nauki o danych z pewnością wpłynie na modele biznesowe

* Podchodzenie do klientów z argumentacją wartości opartą na wprowadzaniu technik analizy danych bez wyraźnego wyjaśnienia, jaka jest wartość dla klienta

* Modele cenowe pozostają takie same lub nie odzwierciedlają zwiększonej wartości, tylko obniżony koszt

* Jednomyślne skupienie się na efektywności kosztowej, jeśli chodzi o zmiany operacyjne firmy

* Ani mierzenie, ani rozumienie ulepszeń operacyjnych

* Przeprowadzanie zmian organizacyjnych na tak małą skalę, że w praktyce wszystko pozostaje takie samo, zapewniając, że rzeczywista zmiana nigdy nie nastąpi

* Budowanie modelu kosztów i wymiarowania na starych i nieaktualnych kryteriach, dzięki czemu model nie uchwyci nowych wartości

* Niedostrzeżenie zmiany, jaką data science nakłada na firmę i niezrozumienie tej zmiany z perspektywy ekosystemu

* Niedocenianie potrzeby komunikacji związanej ze zmianą

Nie zapomnij o pomiarach potrzebnych do udowodnienia wartości

Częstym błędem jest zapominanie o wprowadzeniu pomiarów bazowych przed dokonaniem i wdrożeniem inwestycji w naukę danych. W takich przypadkach większość uwagi skupia się na przyszłych pomiarach i wynikach inwestycji. Dzieje się tak zwykle z powodu oporu przed inwestowaniem w nowe pomiary w obecnej sytuacji, ponieważ jest to porzucane dla nowej strategii. Niestety oznacza to, że firmie zabraknie możliwości statystycznego udowodnienia wartości inwestycji w kolejnym kroku. Nie wpadnij w tę pułapkę! Może to naprawdę obrócić się przeciwko całej ambicji strategicznej, gdy najwyższe kierownictwo lub nawet rada dyrektorów zapyta, jaka była wartość tej poważnej inwestycji. Finansowo mógłbyś oczywiście być w stanie zmotywować inwestycję na wysokim poziomie; jednak trudno byłoby udowodnić poszczególne części. Przyrosty wydajności, takie jak szybkość, zwinność, poziom automatyzacji i reaktywność procesów w porównaniu z proaktywnością, to wartości, które są trudniejsze do udowodnienia i podania liczb, jeśli nie zabezpieczyłeś punktu odniesienia pomiaru przed realizacją strategii analizy danych.