Data Science to sztuka przekształcania danych w działania. Chodzi o rzemiosło. Tradecraft to proces, narzędzia i technologie umożliwiające ludziom i komputerom współpracę w celu przekształcenia danych w spostrzeżenia.

Data Science tradecraft, tworzy produkty danych. Produkty oparte na danych dostarczają użytecznych informacji bez narażania decydentów na podstawowe dane lub analizy (np. Strategie kupna / sprzedaży instrumentów finansowych, zestaw działań mających na celu poprawę wydajności produktu lub kroki w celu poprawy marketingu produktu)).

Data Science wspiera i zachęca do przechodzenia między rozumowaniem dedukcyjnym (opartym na hipotezach) a rozumowaniem indukcyjnym (opartym na wzorcach). Jest to fundamentalna zmiana w stosunku do tradycyjnych podejść analitycznych. Rozumowanie indukcyjne i eksploracyjna analiza danych zapewniają środki do formułowania lub udoskonalania hipotez i odkrywania nowych ścieżek analitycznych. Modele rzeczywistości nie muszą już być statyczne. Są stale testowane, aktualizowane i ulepszane, aż zostaną znalezione lepsze modele.

Data Science jest niezbędny, aby firmy mogły pozostać i konkurować w przyszłości. Organizacje nieustannie podejmują decyzje kierując się instynktem, najgłośniejszym głosem i najlepszą argumentacją - czasami są nawet informowane o prawdziwych informacjach. Zwycięzcy i przegrani w powstającej gospodarce opartej na danych zostaną ustaleni przez ich zespoły Data Science.

Funkcje nauki o danych można rozbudowywać w czasie. Organizacje dojrzewają poprzez szereg etapów - zbieranie, opisywanie, odkrywanie, przewidywanie, doradzanie - w miarę przechodzenia od zalewu danych do pełnej dojrzałości Data Science. Na każdym etapie mogą stawić czoła coraz bardziej złożonym celom analitycznym z szerszym zakresem możliwości analitycznych. Jednak organizacje nie muszą osiągnąć maksymalnej dojrzałości Data Science, aby osiągnąć sukces. Na każdym etapie można znaleźć znaczne korzyści.

Data Science to inny rodzaj sportu zespołowego. Zespoły Data Science potrzebują szerokiego spojrzenia na organizację. Liderzy muszą być kluczowymi orędownikami, którzy spotykają się z interesariuszami, aby wykryć najtrudniejsze wyzwania, zlokalizować dane, połączyć różne części firmy i uzyskać szerokie poparcie.

WPROWADZENIE DO NAUKI O DANYCH

Jeśli nie słyszałeś o nauce o danych, jesteś w tyle. Sama zmiana nazwy grupy Business Intelligence na Data Science nie jest rozwiązaniem.

Co rozumiemy przez naukę o danych?

Opisywanie nauki o danych przypomina próbę opisania zachodu słońca - powinno być łatwe, ale w jakiś sposób uchwycenie słów jest niemożliwe.

Definicja nauki o danych

Data Science to sztuka przekształcania danych w działania. Odbywa się to poprzez tworzenie produktów danych, które dostarczają użytecznych informacji bez narażania decydentów na podstawowe dane lub analizy (np. strategie kupna / sprzedaży instrumentów finansowych, zestaw działań mających na celu poprawę wydajności produktu lub kroki w celu poprawy marketingu produktu). Wykonywanie nauki o danych wymaga pozyskiwania aktualnych, przydatnych do działania informacji z różnych źródeł danych w celu napędzania produktów danych. Przykłady produktów opartych na

danych obejmują odpowiedzi na takie pytania, jak: „Które z moich produktów należy reklamować intensywniej, aby zwiększyć zyski? Jak mogę ulepszyć mój program zgodności, jednocześnie zmniejszając koszty? Jaka zmiana w procesie produkcyjnym pozwoli mi stworzyć lepszy produkt? ” Kluczem do odpowiedzi na te pytania jest: zrozumienie posiadanych danych i tego, co mówią one indukcyjnie.

Termin Data Science pojawił się w literaturze informatycznej w latach sześćdziesiątych i osiemdziesiątych XX wieku. Jednak dopiero pod koniec lat 90. dziedzina, którą tu opisujemy, zaczęła wyłaniać się ze społeczności zajmujących się statystykami i eksploracją danych. Data Science została po raz pierwszy wprowadzona jako niezależna dyscyplina w 2001 roku. Od tego czasu pojawiły się niezliczone artykuły rozwijające tę dyscyplinę, których kulminacją było uznanie Data Scientist za najseksowniejszą pracę XXI wieku.

Produkt danych

Produkt danych dostarcza użytecznych informacji bez narażania decydentów na podstawowe dane lub analizy. Przykłady obejmują:

• Zalecenia dotyczące filmów

• Prognoza pogody

• Prognozy giełdowe

• Udoskonalenia procesu produkcyjnego

• Diagnoza zdrowotna

• Prognozy dotyczące trendów grypy

• Reklama ukierunkowana

Co wyróżnia naukę o danych?

Data Science wspiera i zachęca do przechodzenia między rozumowaniem dedukcyjnym (opartym na hipotezach) i indukcyjnym (opartym na wzorcach). Jest to fundamentalna zmiana w stosunku do tradycyjnych podejść analitycznych. Rozumowanie indukcyjne i eksploracyjna analiza danych zapewniają środki do formułowania lub udoskonalania hipotez i odkrywania nowych ścieżek analitycznych. W rzeczywistości, aby dokonać odkrycia znaczących spostrzeżeń, które są cechą charakterystyczną nauki o danych, musisz mieć wiedzę fachową i wzajemne oddziaływanie między rozumowaniem indukcyjnym i dedukcyjnym. Aktywnie łącząc zdolność wnioskowania dedukcyjnego i indukcyjnego, Data Science tworzy środowisko, w którym modele rzeczywistości nie muszą już być statyczne i empiryczne. Zamiast tego są stale testowane, aktualizowane i ulepszane, aż zostaną znalezione lepsze modele. Pojęcia te podsumowano na rysunku,

Rodzaje rozumowania i ich rola w nauce o danych

RODZAJE ROZUMOWAŃ…

ROZUMOWANIE DEDUKCYJNE:

*Powszechnie kojarzony z„ logiką formalną ”.

*Obejmuje rozumowanie ze znanych przesłanek lub przesłanek, które uważa się za prawdziwe, do pewnego wniosku.

*Wyciągnięte wnioski są pewne, nieuniknione, nieuniknione

ROZUMOWANIE INDUKCYJNE

*Powszechnie znany jako„ logika nieformalna ”lub„ codzienna argumentacja ”.

*Wymaga wyciągania niepewnych wniosków w oparciu o rozumowanie probabilistyczne.

*Wyciągnięte wnioski są prawdopodobne, rozsądne, wiarygodne, wiarygodne.

… I ICH ROLA W RYNKU HANDLOWYM NAUK O DANYCH

ROZUMOWANIE DEDUKCYJNE:

*›Sformułuj hipotezy dotyczące relacji i leżących u ich podstaw modeli.

*Przeprowadź eksperymenty z danymi, aby przetestować hipotezy i modele.

ROZUMOWANIE INDUKCYJNE

*Eksploracyjna analiza danych w celu odkrycia lub udoskonalenia hipotez.

*Odkryj nowe relacje, spostrzeżenia i ścieżki analityczne na podstawie danych.

Różnice między nauką o danych a tradycyjnymi podejściami analitycznymi nie kończą się na płynnym przechodzeniu między rozumowaniem dedukcyjnym i indukcyjnym. Data Science oferuje wyraźnie inną perspektywę niż możliwości, takie jak Business Intelligence. Jednak nauka o danych nie powinna zastępować funkcji Business Intelligence w organizacji. Te dwie możliwości są addytywne i uzupełniające się, z których każda oferuje niezbędny wgląd w operacje biznesowe i środowisko operacyjne. Kluczowe kontrasty to:

* Pytania odkrywcze a pytania gotowe: nauka o danych w rzeczywistości pracuje nad znalezieniem pytania, które należy zadać, a nie tylko zadawaniem go.

* Siła wielu kontra umiejętność jednego: cały zespół zapewnia wspólne forum do łączenia wiedzy z zakresu informatyki, matematyki i dziedzin.

* Perspektywa kontra retrospektywa: nauka o danych koncentruje się na pozyskiwaniu praktycznych informacji z danych w przeciwieństwie do raportowania faktów historycznych.

Różnice między nimi

PATRZENIE W TYŁ I W PRZÓD

NAJPIERW BYŁA BUSINESS INTELLIGENCE

Rozumowanie dedukcyjne

Patrzenie wstecz

Dane dotyczące plasterków i kości

Przechowywane i przechowywane dane

Analizuj przeszłość, zgadnij przyszłość

Tworzy raporty

Wyjście analityczne

TERAZ DODALIŚMY DATA SCIENCE

Rozumowanie indukcyjne i dedukcyjne

Patrzenie w przyszłość

Interakcja z danymi

Rozproszone dane w czasie rzeczywistym

Przewiduj i doradzaj

Tworzy produkty danych

Odpowiadaj na pytania i twórz nowe

Odpowiednia odpowiedź

Jaki jest wpływ nauki o danych?

Kiedy przechodzimy do gospodarki opartej na danych, nauka o danych stanowi przewagę konkurencyjną dla organizacji zainteresowanych wygraną - niezależnie od tego, w jaki sposób wygrana jest zdefiniowana. Sposób definiowania korzyści polega na usprawnieniu procesu decyzyjnego. Były kolega lubił opisywać podejmowanie decyzji w oparciu o dane w następujący sposób: jeśli masz doskonałe informacji lub zero informacji, wtedy twoje zadanie jest łatwe - to pomiędzy tymi dwoma skrajnościami zaczyna się problem. Podkreślał surową rzeczywistość, że bez względu na to, czy informacje są dostępne, czy nie, decyzje muszą zostać podjęte. Sposób podejmowania decyzji przez organizacje ewoluował od pół wieku. Przed wprowadzeniem Business Intelligence jedynymi opcjami były instynkt, najgłośniejszy głos i najlepsza argumentacja. Niestety, ta metoda istnieje do dziś i u niektórych jest głównym środkiem działania organizacji. Skorzystaj z naszej rady i nigdy, przenigdy nie pracuj dla takiej firmy! Na szczęście dla naszej gospodarki większość organizacji zaczęła informować o swoich decyzjach za pomocą prawdziwych informacji, stosując proste statystyki. Ci, którzy zrobili to dobrze, zostali nagrodzeni; te, które nie , zawiodły. Jednakże przerastamy zdolność prostych statystyk do nadążania za wymaganiami rynku. Szybka ekspansja dostępnych danych oraz narzędzia dostępu do danych i korzystania z nich na dużą skalę umożliwiają fundamentalne zmiany w sposobie podejmowania decyzji przez organizacje. Nauka o danych jest niezbędna do utrzymania konkurencyjności w coraz bardziej bogatym w dane środowisku. Podobnie jak w przypadku stosowania prostych statystyk, organizacje, które stosują naukę o danych, zostaną nagrodzone, podczas gdy te, które tego nie zrobią, będą musiały dotrzymać kroku. Gdy dostępne staną się bardziej złożone, rozbieżne zbiory danych, przepaść między tymi grupami będzie się tylko poszerzać. Podkreślmy wartość czekającą organizacje, które wykorzystują Data Science.

Nauka danych jest konieczna ...

17-49% wzrost wydajności, gdy organizacje zwiększają użyteczność danych o 10%

11-42%, zwrot z aktywów (ROA) gdy organizacje zwiększają dostęp do danych o 10%

Wzrost zwrotu z inwestycji o 241%, gdy organizacje wykorzystują duże zbiory danych do poprawy konkurencyjności

1000% wzrost zwrotu z inwestycji w przypadku wdrażania analiz w większości organizacji, dostosowywania codziennych operacji do celów kierownictwa wyższego szczebla i uwzględniania dużych zbiorów danych

5-6% wzrost wydajności dla organizacji podejmujących decyzje oparte na danych.

... aby konkurować w przyszłości

Co jest teraz inne?

Przez 20 lat systemy informatyczne budowano w ten sam sposób. Oddzieliliśmy ludzi, którzy prowadzili biznes, od ludzi, którzy zarządzali infrastrukturą (i dlatego postrzegaliśmy dane jako po prostu kolejną rzecz, którą musieli zarządzać). Wraz z pojawieniem się nowych technologii i technik analitycznych, to sztuczne - i wysoce nieskuteczne - oddzielenie umiejętności krytycznych nie jest już konieczne. Po raz pierwszy organizacje mogą bezpośrednio łączyć decydentów biznesowych z danymi. Ten prosty krok przekształca dane z „czegoś, czym należy zarządzać", w „coś, co należy cenić". W następstwie transformacji organizacje stają przed trudnym wyborem: możesz dalej budować silosy danych i łączyć ze sobą różne informacje lub możesz skonsolidować Twoje dane i wyodrębnić odpowiedzi. Z punktu widzenia nauki o danych jest to fałszywy wybór: podejście wyciszone jest nie do utrzymania, jeśli weźmie się pod uwagę (a) koszt alternatywny niewykorzystania maksymalnego wykorzystania wszystkich dostępnych danych, aby pomóc organizacji odnieść sukces oraz koszty podążania tą samą ścieżką z przestarzałymi procesami. Wymierne korzyści z produktów danych obejmują:

* Koszty alternatywne: Ponieważ nauka o danych jest rozwijającą się dziedziną, koszty alternatywne pojawiają się, gdy konkurent wdraża i generuje wartość z danych, które są przed tobą. Brak wiedzy i brak uwzględnienia zmieniających się wymagań klientów nieuchronnie odciągnie klientów od Twojej obecnej oferty. Gdy konkurenci są w stanie skutecznie wykorzystać naukę danych w celu uzyskania wglądu, mogą przedstawiać zróżnicowane propozycje wartości dla klientów i dzięki temu przodować w swoich branżach.

* Ulepszone procesy: w wyniku coraz bardziej połączonego świata w każdej chwili generowane i przechowywane są ogromne ilości danych. Nauka o danych może służyć do przekształcania danych w spostrzeżenia, które pomagają ulepszyć istniejące procesy. Koszty operacyjne można radykalnie obniżyć poprzez efektywne uwzględnienie złożonych wzajemnych zależności w danych, jak nigdy dotąd. Skutkuje to lepszym zapewnieniem jakości, wyższą wydajnością produktu i bardziej efektywnymi operacjami.

Jak właściwie działa nauka o danych?

To nie jest fizyka jądrowa… to coś lepszego - nauka o danych Nie oszukujmy się - nauka o danych to złożona dziedzina. Jest to trudna, wymagająca intelektualnie praca, która wymaga wyrafinowanej integracji talentów, narzędzi i technik. Ale musimy przebić się przez złożoność i zapewnić jasny, ale skuteczny sposób zrozumienia tego nowego świata. Aby to zrobić, przekształcimy dziedzinę nauki o danych w zestaw uproszczonych działań. Mamy cztery kluczowe działania przedsięwzięcia naukowego o danych. Puryści Data Science prawdopodobnie nie zgodzą się z tym podejściem, ale z drugiej strony prawdopodobnie nie potrzebują przewodnika terenowego, siedząc w swoich wieżach z kości słoniowej! W prawdziwym świecie potrzebujemy jasnych i prostych modeli operacyjnych, które pomogą nam pchnąć nas do przodu.

Działania związane z nauką o danych

Zdobądź -> Przygotuj -> Analizuj -> Działaj

Działanie 1: Zdobądź

Ta aktywność koncentruje się na uzyskaniu potrzebnych danych. Biorąc pod uwagę charakter danych, szczegóły tego działania w dużym stopniu zależą od tego, kim jesteś i co robisz. Dzięki temu nie będziemy poświęcać zbyt wiele czasu na tę czynność, poza podkreśleniem jej wagi i zachęceniem do szerokiego spojrzenia na to, które dane można i należy wykorzystać.

Działanie 2: Przygotuj się

Wspaniałe wyniki nie pojawiają się same. Wiele zależy od przygotowania, a w Data Science oznacza to manipulowanie danymi w celu dopasowania ich do potrzeb analitycznych. Ten etap może pochłonąć dużo czasu, ale jest to doskonała inwestycja. Korzyści są natychmiastowe i długoterminowe.

Działanie 3: Analiza

Jest to czynność, która pochłania lwią część uwagi zespołu. Jest to również najtrudniejsze i najbardziej ekscytujące (zobaczysz wiele „chwil aha" w tej przestrzeni). Jako najtrudniejsze i najbardziej irytujące z czterech zadań, ten przewodnik terenowy koncentruje się na pomocy w wykonywaniu tego lepiej i szybciej.

Działanie 4: Działanie

Każdy skuteczny zespół Data Science analizuje swoje dane w celu - to jest, aby przekształcić dane w działania. Spostrzeżenia, które można podjąć i które mają wpływ, to Święty Graal Data Science. Jednak przekształcanie spostrzeżeń w działanie może być działaniem o charakterze politycznym. Ta aktywność zależy w dużej mierze od kultury i charakteru Twojej organizacji, więc zostawimy Ci samodzielne ustalenie tych szczegółów.

Zdobycie

Cała analiza rozpoczyna się od dostępu do danych, a dla naukowca zajmującego się danymi ten aksjomat jest prawdziwy. Ale są pewne istotne różnice - szczególnie w kwestii tego, kto przechowuje, utrzymuje i jest właścicielem danych w organizacji. Ale zanim tam pójdziemy, spójrzmy, co się zmienia. Tradycyjnie sztywne silosy danych sztucznie definiują gromadzone dane. Innymi słowy, silosy tworzą filtr, który przepuszcza bardzo małą ilość danych i ignoruje resztę. Te filtrowane procesy dają nam sztuczny obraz świata oparty na „danych, które przetrwały", a nie taki, który pokazuje pełną rzeczywistość i znaczenie. Bez szerokiego i ekspansywnego zbioru danych nigdy nie możemy zanurzyć się w różnorodności danych. Zamiast tego podejmujemy decyzje w oparciu o ograniczone i ograniczone informacje. Eliminacja konieczności posiadania silosów daje nam dostęp do wszystkich danych naraz, w tym danych z wielu zewnętrznych źródeł. Obejmuje rzeczywistość, że różnorodność jest dobra, a złożoność jest w porządku. Ten sposób myślenia tworzy zupełnie inny sposób myślenia o danych w organizacji, nadając im nową i zróżnicowaną rolę. Dane stanowią dla organizacji znaczącą, nową szansę na zwiększenie zysków i misji. Ale jak wspomniano wcześniej, ta pierwsza czynność jest silnie zależna

od sytuacji i okoliczności. Nie możemy zostawić Ci nic poza ogólnymi wskazówkami, które pomogą Ci zapewnić maksymalną wartość:

*Zajrzyj najpierw do środka: do jakich danych masz obecnie dostęp, a których nie używasz? Jest to w dużej mierze dane pozostawione przez proces filtrowania i mogą być niezwykle cenne.

*Usuń ograniczenia formatu: przestań ograniczać swój sposób pozyskiwania danych do sfery ustrukturyzowanych baz danych. Zamiast tego potraktuj dane nieustrukturyzowane i częściowo ustrukturyzowane jako realne źródła.

*Dowiedz się, czego brakuje: zadaj sobie pytanie, jakie dane miałyby duży wpływ na Twoje procesy, gdybyś miał do nich dostęp. Następnie znajdź to!

*Przestrzegaj różnorodności: staraj się angażować i łączyć się z publicznie dostępnymi źródłami danych, które mogą mieć znaczenie dla obszaru Twojej domeny.

Nie wszystkie dane są równe

Rozpoczynając gromadzenie danych, pamiętaj, że nie wszystkie dane są tworzone jednakowo. Organizacje mają tendencję do gromadzenia wszelkich dostępnych danych. Dane, które są w pobliżu (łatwo dostępne i łatwo dostępne) mogą być tanie w gromadzeniu, ale nie ma gwarancji, że są to właściwe dane do zebrania. Skoncentruj się na danych zapewniających najwyższy zwrot z inwestycji dla Twojej organizacji. Twój zespół Data Science może pomóc zidentyfikować te dane. Pamiętaj również, że musisz zachować równowagę między danymi, których potrzebujesz, a danymi, które masz. Zbieranie ogromnych ilości danych jest bezużyteczne i kosztowne, jeśli nie są to dane, których potrzebujesz

Przygotowanie

Gdy masz już dane, musisz je przygotować do analizy. Organizacje często podejmują decyzje na podstawie niedokładnych danych. Przekazywanie hierarchicznie danych oznacza, że organizacje mogą mieć martwe punkty. Nie są w stanie zobaczyć całego obrazu i nie spojrzeć na swoje dane i wyzwania w sposób holistyczny. W rezultacie ważne informacje są ukrywane przed decydentami. Badania wykazały, że prawie 33% decyzji podejmowanych jest bez dobrych danych lub informacji. Kiedy naukowcy zajmujący się danymi są w stanie eksplorować i analizować wszystkie dane, pojawiają się nowe możliwości analizy i podejmowania decyzji w oparciu o dane. Spostrzeżenia uzyskane dzięki tym nowym możliwościom znacznie zmienią kierunek działań i decyzje w organizacji. Jednak uzyskanie dostępu do pełnego repozytorium danych organizacji wymaga przygotowania. Doświadczenie wielokrotnie pokazuje, że najlepszym narzędziem dla naukowców zajmujących się danymi do przygotowania analizy jest jezioro - a konkretnie jezioro danych. To nowe podejście do gromadzenia, przechowywania i integracji danych, które pomaga organizacjom maksymalizować użyteczność ich danych. Zamiast przechowywać informacje w dyskretnych strukturach danych, usługa Data Lake konsoliduje pełne repozytorium danych organizacji w jednym, dużym widoku. Eliminuje kosztowny i kłopotliwy proces przygotowania danych, znany jako wyodrębnianie / przekształcanie / ładowanie (ETL), niezbędny w przypadku silosów danych. Wszystkie informacje w Data Lake są dostępne dla każdego zapytania - i wszystkie naraz.

Analiza

Uzyskaliśmy dane… przygotowaliśmy je… teraz pora na ich analizę. Działanie Analiza wymaga największego wysiłku ze wszystkich działań w ramach analizy danych. Data Scientist faktycznie tworzy analizy, które tworzą wartość z danych. Analiza w tym kontekście to iteracyjne zastosowanie

wyspecjalizowanych i skalowalnych zasobów i narzędzi obliczeniowych, które zapewniają odpowiedni wgląd w wykładniczo rosnących danych. Ten rodzaj analizy umożliwia zrozumienie ryzyk i szans w czasie rzeczywistym poprzez ocenę danych sytuacyjnych, operacyjnych i behawioralnych. Dzięki pełnemu dostępowi do danych w Data Lake organizacje mogą korzystać z narzędzi analitycznych, aby znaleźć rodzaje połączeń i wzorców wskazujących na obiecujące możliwości. To szybkie połączenie analityczne jest realizowane w ramach usługi Data Lake, w przeciwieństwie do starszych metod próbkowania, które mogą wykorzystywać tylko wąski wycinek danych. Aby zrozumieć, co jest w jeziorze, trzeba było wydobyć dane i przestudiować je. Teraz możesz zanurzyć się w jeziorze, przenosząc swoje analizy do danych. Naukowcy zajmujący się danymi zajmują się różnymi celami analitycznymi - opisywanie, odkrywanie, przewidywanie i doradzaj. Dojrzałość zdolności analitycznej determinuje cele analityczne, które obejmuje. Wiele zmiennych odgrywa kluczową rolę w określaniu trudności i przydatności każdego celu dla organizacji. Niektóre z tych zmiennych to rozmiar i budżet organizacji i rodzaj produktów danych, których wymaga decyzja twórcy. Oprócz pochłaniania największego wysiłku, czynność Analiza jest zdecydowanie najbardziej złożona. Tradycja nauki o danych to sztuka. Chociaż nie możemy nauczyć Cię, jak być artystą, możemy podzielić się podstawowymi narzędziami i technikami, które pomogą Ci odnieść sukces.

Teraz, gdy przeanalizowaliśmy dane, czas podjąć działania. Umiejętność wykorzystania analizy jest krytyczna. Jest też bardzo sytuacyjny. Podobnie jak w przypadku działania Zdobądź, najlepsze, na co możemy liczyć, to przedstawienie pewnych zasad przewodnich, które pomogą Ci ułożyć wynik tak, aby uzyskać maksymalny efekt. Oto kilka kluczowych punktów, o których należy pamiętać podczas prezentowania wyników:

1. Ustalenie musi mieć sens przy stosunkowo niewielkim wstępnym przeszkoleniu lub przygotowaniu ze strony decydenta.

2. Odkrycie musi sprawić, że najbardziej znaczące wzorce, trendy i wyjątki będą łatwe do dostrzeżenia i zinterpretowania.

3. Należy dołożyć wszelkich starań, aby dokładnie zakodować dane ilościowe, tak aby osoba podejmująca decyzję mogła dokładnie zinterpretować i porównać dane.

4. Logika zastosowana do ustalenia wyniku musi być jasna i przekonująca, a także możliwa do prześledzenia wstecz poprzez dane.

5. Wyniki muszą odpowiadać na prawdziwe pytania biznesowe.

Dojrzałość nauki o danych w organizacji

Cztery omówione dotychczas działania zapewniają uproszczony obraz nauki o danych. Organizacje będą powtarzać te czynności przy każdym nowym przedsięwzięciu Data Science. Z czasem jednak poziom wysiłku niezbędnego do wykonania każdego działania będzie się zmieniał. Ponieważ na przykład więcej danych jest pozyskiwanych i przygotowywanych w Data Lake, działania te będą wymagały znacznie mniejszego wysiłku. Wskazuje to na dojrzewanie zdolności Data Science. Ocena dojrzałości Twoich możliwości Data Science wymaga nieco innego spojrzenia. Używamy Modelu dojrzałości nauki o danych jako wspólnej ramy do opisu progresji dojrzałości i komponentów, które składają się na zdolność nauki o danych. Ramy te można zastosować do zdolności organizacji do nauki o danych lub nawet do dojrzałości konkretnego rozwiązania, a mianowicie produktu danych. Na każdym etapie dojrzałości można uzyskać potężny wgląd. Kiedy organizacje zaczynają, mają silosy danych. Na tym etapie nie prowadzili żadnych ogólnych działań w ramach Agregatu. Mogą nie znać wszystkich posiadanych danych lub potrzebnych im danych. Decyzja o utworzeniu zdolności Data

Science sygnalizuje przejście do etapu zbierania. Cały twój początkowy wysiłek będzie skupiony na identyfikacji i agregacji danych. Z biegiem czasu będziesz mieć potrzebne dane, a mniejsza część Twojego wysiłku może skupić się na zbieraniu. Możesz teraz rozpocząć opisywanie swoich danych. Należy jednak pamiętać, że chociaż udział czasu spędzonego na zbieraniu drastycznie spada, nigdy nie znika całkowicie. Wskazuje to na cztery działania opisane wcześniej - będziesz kontynuować gromadzenie i przygotowywanie danych, gdy pojawią się nowe pytania analityczne, potrzebne są dodatkowe dane i pojawią się nowe źródła danych. Organizacje nadal osiągają dojrzałość, przechodząc przez etapy od Opisz do Doradzania. Na każdym etapie mogą zajmować się coraz bardziej złożonymi celami analitycznymi z szerszym zakresem możliwości analitycznych. Jak opisano w sekcji Zbieranie, każdy etap nigdy nie znika całkowicie. Zamiast tego zmniejsza się proporcja czasu poświęcanego na to i rozpoczynają się nowe, bardziej dojrzałe czynności. Krótki opis każdego etapu dojrzałości przedstawiono w tabeli Etapy dojrzałości nauki o danych.

Etap: Opis: Przykład

Gromadzenie: koncentruje się na zbieraniu wewnętrznych lub zewnętrznych zbiorów danych.: Gromadzenie rekordów sprzedaży i odpowiednich danych pogodowych.

Opis: dąży do ulepszenia lub udoskonalenia surowych danych, a także wykorzystania podstawowych funkcji analitycznych, takich jak liczenie. : W jaki sposób moi klienci są rozdzielani pod względem lokalizacji, czyli kodu pocztowego?

Odkryj: Identyfikuje ukryte relacje lub wzorce.: Czy wśród moich stałych klientów są grupy, które dokonują podobnych zakupów?

Przewidywanie: Wykorzystuje wcześniejsze obserwacje do przewidywania przyszłych obserwacji .: Czy możemy przewidzieć, które produkty są bardziej skłonne do zakupu określonych grup klientów?

Doradzaj: określa możliwe decyzje, optymalizuje je i radzi, aby zastosować decyzję, która daje najlepszy wynik. : Twoja rada jest taka, aby kierować reklamy określonych produktów do określonych grup, aby zmaksymalizować przychody.

Model dojrzałości zapewnia potężne narzędzie do zrozumienia i docenienia dojrzałości umiejętności Data Science. Organizacje nie muszą osiągać maksymalnej dojrzałości, aby osiągnąć sukces. Na każdym etapie można znaleźć znaczne korzyści. Jesteśmy głęboko przekonani, że nikt nie angażuje się w działania związane z nauką o danych, chyba że ma to na celu wytworzenie wyniku - to znaczy, że masz zamiar doradzić. Oznacza to po prostu, że każdy krok naprzód w dojrzałości prowadzi cię w prawo na diagramie modelu. Przejście w prawo wymaga odpowiednich procesów, ludzi, kultury i modelu operacyjnego - solidnych możliwości Data Science. Co jest potrzebne do stworzenia możliwości nauki o danych?. Zaobserwowaliśmy bardzo niewiele organizacji faktycznie działających na najwyższych poziomach dojrzałości, na etapach Przewiduj i Doradzaj. Tradycja Odkryj dopiero teraz dojrzewa do tego stopnia, że organizacje mogą skupić się na zaawansowanych działaniach typu Przewiduj i Doradzaj. To jest nowa granica nauki o danych. To jest przestrzeń, w której zaczniemy rozumieć, jak zlikwidować lukę poznawczą między ludźmi a komputerami. Organizacje, które sięgną po Doradzaj, uzyskają prawdziwy wgląd i rzeczywistą przewagę konkurencyjną.

Na jakim etapie dojrzałości analitycznej znajduje się Twoja organizacja?

Rozwiąż quiz!

1. Ile źródeł danych gromadzisz?

a. Po co nam dużo danych? - 0 punktów, koniec tutaj.

b. Nie znam dokładnej liczby. - 5 punktów

c. Zidentyfikowaliśmy wymagane dane i zbieramy je. - 10 punktów

2. Czy wiesz, na jakie pytania stara się odpowiedzieć Twój zespół Data Science?

a. Dlaczego potrzebujemy pytań? - 0 punktów

b. Nie, sami to sobie radzą. - 5 punktów

c. Tak, oceniliśmy pytania, które będą miały największy wpływ na biznes. - 10 punktów

3. Czy znasz ważne czynniki napędzające Twój biznes?

a. Nie mam pojęcia. - 0 punktów

b. Nasze kwanty pomagają mi to rozgryźć. - 5 punktów

c. Mamy do tego produkt danych. - 10 punktów

4. Czy rozumiesz przyszłe warunki?

a. Patrzę na aktualne warunki i czytam z liścii herbaty. - 0 punktów

b. Mamy do tego produkt danych. - 5 punktów

5. Czy znasz najlepszy sposób postępowania w przypadku kluczowych decyzji?

a. Patrzę na prognozy i planuję kurs. - 0 punktów

b. Mamy do tego produkt danych. - 5 punktów

Sprawdź swój wynik:

0 - Silosy danych, 5-10 - Gromadzenie,

10-20 - Opisz, 20-30 - Odkryj,

30-35 - Przewidywanie, 35-40 – Doradztwo

Co jest potrzebne do stworzenia możliwości analizy danych?

Data Science to przede wszystkim budowanie zespołów i kultury. Podobnie jak w przypadku każdego sportu zespołowego, nauka o danych zależy od różnorodnego zestawu umiejętności, aby osiągnąć swój cel - zdobywanie lepszych spostrzeżeń. Potrzebujesz trzech umiejętności, aby stworzyć zwycięski zespół w świecie Data Science.

EKSPERTYZA W DOMENIE: Zapewnia zrozumienie rzeczywistości, w której istnieje przestrzeń problemowa

MATEMATYKA: zapewnia strukturę teoretyczną, w której badane są problemy nauki o danych.

INFORMATYKA: Zapewnia środowisko, w którym tworzone są produkty danych.

Budowanie zespołów Data Science jest trudne. Wymaga zrozumienia typów osobowości, które umożliwiają naukę o danych, a także chęci stworzenia kultury innowacji i ciekawości w Twojej organizacji. Musisz także zastanowić się, jak wdrożyć zespół i uzyskać powszechne poparcie w całej organizacji.

Zrozumienie, co wyróżnia naukowca danych

Nauka o danych często wymaga znacznych nakładów czasu na wykonanie różnych zadań. Należy generować hipotezy, gromadzić, przygotowywać, analizować i wykorzystywać dane. Często stosuje się wiele technik, zanim jedna przyniesie interesujące rezultaty. Jeśli wydaje się to zniechęcające, to dlatego, że tak jest. Data Science to trudna, wymagająca intelektualnie praca, która wymaga dużego talentu: zarówno namacalnych umiejętności technicznych, jak i niematerialnych „czynników x". Najważniejszymi cechami naukowców zajmujących się danymi są zazwyczaj niematerialne aspekty ich osobowości. Naukowcy danych są z natury ciekawi, kreatywni, skupieni i zorientowani na szczegóły.

*Ciekawość jest niezbędna, aby oddzielić problem i zbadać współzależności między danymi, które mogą wydawać się pozornie niepowiązane.

*Kreatywność jest niezbędna do wymyślenia i wypróbowania nowych podejść do rozwiązania problemu, które często nigdy wcześniej nie były stosowane w takim kontekście.

*Skupienie jest wymagane do zaprojektowania i przetestowania techniki przez wiele dni i tygodni, stwierdzenia, że nie działa, wyciągnięcia wniosków z błędu i spróbuj ponownie.

*Zwracanie uwagi na szczegóły jest potrzebne, aby zachować rygor oraz wykryć i uniknąć nadmiernego polegania na intuicji podczas badania danych.

Sukces zespołu Data Science wymaga biegłości w trzech podstawowych umiejętnościach technicznych: informatyce, matematyce i wiedzy domenowej. Komputery zapewniają środowisko, w którym testowane są hipotezy oparte na danych, i jako taka informatyka jest niezbędna do manipulacji i przetwarzania danych. Matematyka zapewnia strukturę teoretyczną, w której bada się problemy nauki o danych. Bogate podstawy statystyki, geometrii, algebry liniowej i rachunku różniczkowego są ważne, aby zrozumieć podstawy wielu algorytmów i narzędzi. Wreszcie, wiedza specjalistyczna w dziedzinie przyczynia się do zrozumienia, jakie problemy faktycznie wymagają rozwiązania, jakiego rodzaju dane istnieją w domenie oraz w jaki sposób można instrumentować i mierzyć przestrzeń problemową.

Jednorożec potrójnego zagrożenia

Osoby, które świetnie radzą sobie ze wszystkimi trzema podstawowymi umiejętnościami technicznymi Data Science, są jak jednorożce - bardzo rzadkie i jeśli kiedykolwiek uda ci się znaleźć taką, należy traktować ją ostrożnie. Kiedy zarządzasz tymi osobami:

*Zachęcaj ich do kierowania Twoim zespołem, ale nie do zarządzania nim. Nie obciążaj ich obowiązkami kierownictwa, które mógłby wykonać inny personel.

*Włóż dodatkowy wysiłek w zarządzanie karierą i zainteresowaniami w Twojej organizacji. Twórz możliwości awansu w swojej organizacji, które pozwolą im skupić się na mentorowaniu innych naukowców zajmujących się danymi i rozwijaniu najnowocześniejszych technologii, jednocześnie rozwijając ich karierę.

*Zadbaj o to, by mieli możliwość prezentowania i rozpowszechniania swoich pomysłów na wielu różnych forach, ale też bądź wrażliwy na swój czas.

Znajdowanie sportowców dla swojego zespołu

Budowanie zespołu Data Science jest skomplikowane. Organizacje muszą jednocześnie zaangażować istniejący personel wewnętrzny do stworzenia „kotwicy", której można użyć do rekrutacji i rozwoju zespołu, jednocześnie przechodząc zmiany organizacyjne i transformacje, aby w znaczący sposób włączyć tę nową klasę pracowników. Budowanie zespołu zaczyna się od zidentyfikowania istniejącego personelu w organizacji, który ma duże predyspozycje do nauki o danych. Dobrzy kandydaci będą mieli

formalne doświadczenie w jednej z trzech podstawowych umiejętności technicznych, o których wspomnieliśmy, a co najważniejsze będą posiadać cechy osobowości niezbędne do nauki o danych. Często mogą mieć stopnie naukowe (magisterskie lub wyższe), ale nie zawsze. Pierwszy zidentyfikowany personel powinien również mieć dobre cechy przywódcze i poczucie celu dla organizacji, ponieważ będzie on kierował późniejszymi wysiłkami związanymi z zatrudnianiem i rekrutacją. Nie dyskontuj nikogo - naukowców danych znajdziesz w najdziwniejszych miejscach z najdziwniejszymi kombinacjami środowisk.

Kształtowanie kultury

Dobra nauka o danych wymaga wysoce akademickiej kultury recenzowania, w której żaden członek organizacji nie jest odporny na konstruktywną krytykę. Tworząc praktykę Data Science, powinieneś być przygotowany na poddanie wszystkich aspektów swojej działalności korporacyjnej ciekawemu charakterowi Twoich zespołów Data Science. Niezastosowanie się do tego stwarza negatywny obraz kultury, która nie „zjada własnej psiej karmy" i zachęci do negatywnej refleksji nad marką, zarówno wewnętrznie, jak i zewnętrznie. Powinieneś być świadomy wszelkich kulturowych spuścizny istniejących w organizacji, które są sprzeczne z nauką o danych. Naukowcy danych są zasadniczo ciekawi i pomysłowi. „Nie jesteśmy wścibscy, jesteśmy naukowcami danych". Cechy te mają fundamentalne znaczenie dla powodzenia projektu i nadania nowego wymiaru wyzwaniom i pytaniom. Często projekty Data Science są utrudnione przez brak możliwości wyobrażenia sobie czegoś nowego i innego. Zasadniczo organizacje muszą wspierać zaufanie i przejrzystą komunikację na wszystkich poziomach, zamiast szanować w celu stworzenia silnego zespołu Data Science. Menedżerowie powinni być przygotowani na częstsze zapraszanie do udziału i rzadziej przedstawianie wyjaśnień lub przeprosin.

W tym przypadku nie oceniaj książki po okładce ani naukowca zajmującego się danymi na podstawie stopnia naukowego. Naukowców danych można znaleźć wszędzie. Wystarczy spojrzeć na różnorodne i zaskakujące przykłady stopni naukowych posiadanych przez wielu ekspertów:

* Bioinformatyka

* Inżynieria biomedyczna

* Biofizyka

* Biznes

* Grafika komputerowa

* Informatyka

* Język angielski

* Zarządzanie lasem

* Historia

* Inżynieria przemysłowa

* Technologia informacyjna

*Matematyka

* Studia nad bezpieczeństwem narodowym

* Badania operacyjne

* Fizyka

* Zarządzanie dziką przyrodą i rybołówstwem

Wybór modelu operacyjnego

W zależności od rozmiaru, złożoności i czynników biznesowych, organizacje powinny rozważyć jeden z trzech modeli operacyjnych Data Science: scentralizowany, wdrożeniowy lub rozproszony.

Scentralizowane zespoły Data Science obsługują organizację we wszystkich jednostkach biznesowych. Zespół jest scentralizowany pod kierownictwem głównego badacza danych. Służą wszystkim analitycznym potrzebom organizacji i wszystkie znajdują się razem. Eksperci dziedzinowi przyjeżdżają do tej organizacji na krótkie okresy rotacyjne, aby rozwiązywać problemy związane z biznesem.

Wdrożeniowe zespoły Data Science trafiają do jednostki biznesowej lub grupy i przebywają tam na krótko lub długoterminowe zadania. Są oni własnym podmiotem i współpracują z ekspertami dziedzinowymi w grupie, aby rozwiązywać trudne problemy. Mogą pracować niezależnie nad określonymi wyzwaniami, ale zawsze powinny współpracować z innymi zespołami w celu wymiany narzędzi, technik i historii wojennych.

Rozproszony zespół Data Science to taki, który jest w pełni osadzony w każdej grupie i staje się częścią długoterminowej organizacji. Te zespoły działają najlepiej, gdy charakter domeny lub jednostki biznesowej jest już skoncentrowany na analityce. Jednak zbudowanie przekrojowego spojrzenia na zespół, który może współpracować z innymi zespołami Data Science, ma kluczowe znaczenie dla sukcesu.

RÓWNOWAŻENIE RÓWNANIA ZESPOŁU DS. DANYCH

Równoważenie składu zespołu Data Science jest podobne do równoważenia reagentów i produktów w reakcji chemicznej. Każda strona równania musi reprezentować tę samą ilość każdego konkretnego elementu. W przypadku nauki o danych elementy te to podstawowe umiejętności techniczne informatyka (CS), matematyka (M) i wiedza domenowa (DE). Każdy z reagentów, twoich naukowców zajmujących się danymi, ma swój własny, unikalny zestaw umiejętności. Musisz zrównoważyć skład personelu, aby spełnić wymagania umiejętności zespołu Data Science, produkt w reakcji. Jeśli nie zrównoważysz poprawnie równania, Twój zespół Data Science nie będzie miał pożądanego wpływu na organizację.

$2\ CS\ M_2 + 2\ CS + M\ DE \rightarrow CS_4\ M_5\ DE$

W powyższym przykładzie Twój projekt wymaga czterech części informatyki, pięciu części matematyki i jednej części wiedzy specjalistycznej. Biorąc pod uwagę różnorodność umiejętności personelu, potrzeba pięciu osób, aby zrównoważyć równanie. W trakcie Twojego projektu Data Science wymagania dotyczące umiejętności zespołu będą się zmieniać. Będziesz musiał ponownie zbilansować równanie, aby zapewnić równowagę reagentów z produktami.

Sukces zaczyna się na szczycie

Zespoły Data Science, bez względu na sposób ich rozmieszczenia, muszą mieć sponsora. Mogą one zacząć się od oddolnych wysiłków kilku osób, aby zacząć rozwiązywać trudne problemy, lub jako wysiłki kierowane przez dyrektora generalnego. W zależności od złożoności organizacji, kierunek od góry do

dołu dla dużych organizacji jest najlepszym sposobem na złagodzenie obaw i wątpliwości tych nowych grup. Zespoły Data Science często stają w obliczu trudniejszych politycznych przeszkód podczas rozwiązywania problemów niż przeszkody techniczne. Aby udowodnić wartość zespołu Data Science, zespół musi początkowo skupić się na najtrudniejszych problemach w organizacji, które przynoszą największy zwrot dla kluczowych interesariuszy i zmieni sposób, w jaki organizacja podchodzi do wyzwań w przyszłości. Dzięki temu zespół jest zmotywowany i zachęcony w obliczu trudnych wyzwań. Liderzy muszą być kluczowymi rzecznikami, którzy spotykają się z interesariuszami, aby wykryć najtrudniejsze problemy, zlokalizować dane, połączyć różne części firmy i uzyskać szerokie poparcie.

Zasady przewodnie

Niepowodzenie jest dobre; szybka awaria jest jeszcze lepsza. Zestaw zasad przewodnich, które regulują sposób, w jaki prowadzimy tradycję nauki o danych, jest luźno oparty na centralnych założeniach innowacji, ponieważ te dwa obszary są ze sobą ściśle powiązane. Zasady te nie są twardymi i szybkimi zasadami, których należy ściśle przestrzegać, ale raczej kluczowymi zasadami, które pojawiły się w naszej zbiorowej świadomości. Powinieneś ich używać do podejmowania decyzji, od dekompozycji problemu do implementacji.

* Bądź gotowy na porażkę. U podstaw nauki o danych leży idea eksperymentowania. Prawdziwie innowacyjne rozwiązania pojawiają się tylko wtedy, gdy eksperymentujesz z nowymi pomysłami i aplikacjami. Niepowodzenie jest akceptowalnym produktem ubocznym eksperymentów. Błędy lokalizują regiony, które nie muszą już być brane pod uwagę podczas wyszukiwania rozwiązania.

* Często zawodzą i szybko się uczą. Oprócz gotowości do porażki, bądź gotowy na wielokrotne niepowodzenie. Są chwile, kiedy trzeba zbadać kilkanaście podejść, aby znaleźć to, które działa. Chociaż nie powinieneś martwić się porażką, powinieneś starać się szybko uczyć na tej próbie. Jedynym sposobem na zbadanie dużej liczby rozwiązań jest zrobienie tego szybko.

* Pamiętaj o celu. Często można zgubić się w szczegółach i wyzwaniach związanych z wdrożeniem. Kiedy tak się dzieje, tracisz z oczu swój cel i zaczynasz zsuwać się ze ścieżki od danych do działań analitycznych. Od czasu do czasu cofaj się, kontempluj swój cel i oceniaj, czy Twoje obecne podejście naprawdę może doprowadzić Cię tam, gdzie chcesz.

* Poświęcenie i skupienie prowadzą do sukcesu. Często musisz zbadać wiele podejść, zanim znajdziesz to, które działa. Łatwo się zniechęcić. Musisz pozostać oddany swojemu celowi analitycznemu. Skoncentruj się na szczegółach i spostrzeżeniach ujawnionych przez dane. Czasami pozornie małe obserwacje prowadzą do dużych sukcesów.

* Skomplikowane nie znaczy lepsze. Jako praktycy techniczni mamy tendencję do odkrywania bardzo złożonych, zaawansowanych podejść. Chociaż są chwile, kiedy jest to konieczne, prostsze podejście może często zapewnić ten sam wgląd. Prostsze oznacza łatwiejsze i szybsze prototypowanie, wdrażanie i weryfikowanie.

Wskazówki od profesjonalistów

Łatwiej jest wykluczyć rozwiązanie niż potwierdzić jego poprawność. W rezultacie skup się na odkrywaniu oczywistych niedociągnięć, które mogą szybko zdyskwalifikować podejście. Pozwoli ci to skupić się na odkrywaniu naprawdę wykonalnych podejść, a nie ślepych uliczek.

Jeśli pierwszą rzeczą, którą spróbujesz zrobić, jest stworzenie ostatecznego rozwiązania, poniesiesz porażkę, ale dopiero po kilku tygodniach uderzenia głową o ścianę

Znaczenie rozumu

Uwaga: w świecie Data Science, jeśli chodzi jak kaczka i kwacze jak kaczka, może to być po prostu łoś. Data Science wspiera i zachęca do przechodzenia między rozumowaniem dedukcyjnym (opartym na hipotezach) i indukcyjnym (opartym na wzorcach). Rozumowanie indukcyjne i eksploracyjna analiza danych zapewniają środki do formułowania lub udoskonalania hipotez i odkrywania nowych ścieżek analitycznych. Modele rzeczywistości nie muszą już być statyczne. Są stale testowane, aktualizowane i ulepszane, aż zostaną znalezione lepsze modele. Analiza dużych zbiorów danych wysunęła na pierwszy plan rozumowanie indukcyjne. Analizowane są ogromne ilości danych w celu zidentyfikowania korelacji. Jednak częstą pułapką tego podejścia jest mylenie korelacji z przyczynowością. Korelacja implikuje, ale nie dowodzi związku przyczynowego. Nie można wyciągać wniosków z korelacji, dopóki nie zostaną zrozumiane podstawowe mechanizmy, które odnoszą się do elementów danych. Bez odpowiedniego modelu wiążącego dane korelacja może być po prostu zbiegiem okoliczności.

Korelacja bez przyczynowości

Typowym przykładem tego zjawiska jest wysoka korelacja między spożyciem lodów a wskaźnikiem morderstw w miesiącach letnich. Czy to oznacza, że spożycie lodów powoduje morderstwo, czy też odwrotnie, morderstwo powoduje ich spożycie? Najprawdopodobniej nie, ale można dostrzec niebezpieczeństwo w myleniu korelacji z przyczyną. Nasza praca jako Data Scientists polega na upewnieniu się, że rozumiemy różnicę.

Niebezpieczeństwa odrzucenia

W erze dużych zbiorów danych jednym z często pomijanych elementów analizy jest problem znajdowania wzorców, gdy w rzeczywistości nie ma widocznych wzorców. W statystykach określa się to jako błąd typu I. Jako naukowcy zawsze szukamy nowego lub interesującego przełomu, który mógłby wyjaśnić zjawisko. Mamy nadzieję, że w naszych danych zobaczymy wzór, który coś wyjaśnia lub może dać nam odpowiedź. Podstawowym celem testowania hipotez jest ograniczenie błędu typu I. Osiąga się to za pomocą małych wartości α. Na przykład wartość α równa 0,05 oznacza, że istnieje 1 na 20 szans, że test wykaże, że jest coś znaczącego, podczas gdy w rzeczywistości tak nie jest. To związki problematyczne podczas testowania wielu hipotez. Podczas przeprowadzania testów wielu hipotez prawdopodobnie napotkamy błąd typu I. Tak jak dostępnych jest więcej danych do analizy, należy kontrolować błąd typu I. Jeden z moich projektów wymagał sprawdzenia różnicy między średnią z dwóch próbek danych z mikromacierzy. Dane mikromacierzy zawierają tysiące pomiarów, ale liczba obserwacji jest ograniczona. Powszechnym podejściem do analizy jest pomiar tych samych genów w różnych warunkach. Jeśli jest wystarczająco znacząca różnica w ilości ekspresji genów między dwiema próbkami, możemy powiedzieć, że gen jest skorelowany z określonym fenotypem. Jednym ze sposobów jest wzięcie średniej każdego fenotypu dla określonego genu i sformułowanie hipotezy, aby sprawdzić, czy istnieje znacząca różnica między średnimi. Biorąc pod uwagę, że wykonaliśmy tysiące tych testów przy α = 0,05, znaleźliśmy kilka różnic, które były znaczące. Problem w tym, że niektóre z nich może być spowodowane przypadkowym przypadkiem. Istnieje wiele poprawek w celu kontroli fałszywych wskazań istotności. Poprawka Bonferroniego jest jedną z najbardziej konserwatywnych. To obliczenie obniża poziom, poniżej którego odrzucisz hipotezę zerową (twoją wartość p). Formuła to alfa / n, gdzie n oznacza liczbę przeprowadzanych testów hipotez. Tak więc, jeśli miałbyś przeprowadzić 1000 testów istotności przy α = 0,05, twoja wartość p powinna być mniejsza niż 0,00005 (0,05 / 1000), aby odrzucić hipotezę zerową. Jest to oczywiście znacznie bardziej rygorystyczna wartość. Duża liczba wcześniej istotnych wartości nie była już znacząca, ujawniając prawdziwe relacje w danych. Skorygowana istotność dała nam pewność, że obserwowane poziomy ekspresji wynikały raczej z różnic w ekspresji genów komórkowych niż z szumu. Byliśmy w stanie wykorzystać te informacje, aby rozpocząć badanie, jakie białka i szlaki były aktywne w genach wyrażających fenotyp będący przedmiotem zainteresowania. Umacniając naszą wiedzę na temat związków przyczynowych,

skupiliśmy się w naszych badaniach na obszarach, które mogą prowadzić do nowych odkryć dotyczących funkcji genów, a ostatecznie do ulepszonych metod leczenia.

Rozsądek i zdrowy rozsądek są podstawą nauki o danych. Bez tego dane są po prostu zbiorem bitów. Kontekst, wnioski i modele są tworzone przez ludzi i niosą ze sobą uprzedzenia i założenia. Ślepe zaufanie do analiz jest niebezpieczną rzeczą, która może prowadzić do błędnych wniosków. Podchodząc do wyzwania analitycznego, zawsze powinieneś zatrzymać się i zadać sobie następujące pytania:

* Jaki problem próbujemy rozwiązać? Sformułuj odpowiedź w formie zdania, szczególnie podczas komunikacji z użytkownikiem końcowym. Upewnij się, że brzmi to jak odpowiedź. Na przykład: „Biorąc pod uwagę ustaloną ilość kapitału ludzkiego, rozmieszczenie ludzi z tymi priorytetami zapewni najlepszy zwrot z ich czasu".

*Czy podejście ma sens? Napisz swój plan analityczny. Przyjmij dyscyplinę pisania, która nadaje strukturę twojemu myśleniu. Obliczenia z tyłu koperty są dowodem na istnienie Twojego podejścia. Bez takiego przygotowania komputery są elektronarzędzia, które bardzo szybko mogą dać wiele złych odpowiedzi.

*Czy odpowiedź ma sens? Czy możesz wyjaśnić odpowiedź? Komputery, w przeciwieństwie do dzieci, robią to, co im każą. Upewnij się, że rozmawiałeś z nim wyraźnie, potwierdzając, że podane przez Ciebie instrukcje są zgodne z zamierzeniami. Dokumentuj swoje założenia i upewnij się, że nie wprowadziły one uprzedzeń do Twojej pracy.

*Czy to ustalenie czy błąd? Bądź sceptyczny wobec niespodziewanych odkryć. Doświadczenie mówi, że jeśli wydaje się to złe, to prawdopodobnie jest złe. Zanim jednak zaakceptujesz ten wniosek, upewnij się, że rozumiesz i możesz jasno wyjaśnić, dlaczego jest źle.

*Czy analiza odnosi się do pierwotnych zamiarów? Upewnij się, że nie dopasowujesz odpowiedzi do oczekiwań klienta. Zawsze mów prawdę, ale pamiętaj, że odpowiedzi „Twoje dziecko jest brzydkie" wymagają więcej, a nie mniej analiz.

*Czy historia jest kompletna? Celem twojej analizy jest opowiedzenie historii, która da się wykorzystać. Nie możesz liczyć na to, że publiczność zszyje kawałki razem. Zidentyfikuj potencjalne dziury w swojej historii i wypełnij je, aby uniknąć niespodzianki. Gramatyka, pisownia i grafika mają znaczenie; Twoi odbiorcy stracą zaufanie do Twojej analizy, jeśli wyniki będą wyglądać niechlujnie.

*Dokąd zmierzalibyśmy dalej? Żadna analiza nie jest zakończona, po prostu zabrakło Ci zasobów. Zrozum i wyjaśnij, jakie dodatkowe środki można by podjąć, gdyby znalazło się więcej zasobów.

Wskazówki od profesjonalistów

Lepszy krótki ołówek niż długa pamięć. Kończ każdy dzień, dokumentując, gdzie jesteś; po drodze możesz się czegoś nauczyć. Dokumentuj, czego się nauczyłeś i dlaczego zmieniłeś swój plan.

Przetestuj swoje odpowiedzi z przyjazną publicznością, aby upewnić się, że wyniki są wiarygodne. Niezależne grupy ratują kariery.

Części składowe nauki o danych

Istnieje sieć komponentów, które współdziałają ze sobą, tworząc przestrzeń rozwiązań. Zrozumienie, w jaki sposób są one połączone, ma kluczowe znaczenie dla umiejętności projektowania rozwiązań problemów związanych z nauką o danych. Komponenty zaangażowane w każdy projekt Data Science należą do wielu różnych kategorii, w tym analizowanych typów danych, używanych klas analitycznych,

stosowanych modeli uczenia się i modeli wykonywania używanych do uruchamiania analiz. Wzajemne połączenia między tymi komponentami, pokazane na rysunku, Wzajemne połączenia między częściami składowymi nauki o danych, mówią o złożoności rozwiązań inżynieryjnych w dziedzinie nauki o danych. Wybór jednego elementu wpływa na wybory dokonane dla innych kategorii. Na przykład typy danych prowadzą do wyborów w klasach analitycznych i modelach uczenia się, podczas gdy opóźnienia, terminowość i algorytmiczna strategia zrównoleglania wpływają na model wykonania. Gdy zagłębiamy się w techniczne aspekty Data Science, zaczniemy od eksploracji tych komponentów i omówimy przykłady każdego z nich.

Przeczytaj to, aby nie niedbale:

Podczas projektowania rozwiązania do nauki o danych pracuj nad zrozumieniem komponentów, które definiują przestrzeń rozwiązania. Niezależnie od celu analitycznego, musisz wziąć pod uwagę typy danych, z którymi będziesz pracować, klasy analiz, których użyjesz do wygenerowania produktu danych, sposób działania i ewolucji zawartych w nim modeli uczenia się oraz modele wykonawcze, które będą regulować sposób analiza zostanie uruchomiona. Dopiero po rozważeniu każdego z tych aspektów będziesz w stanie sformułować kompletne rozwiązanie Data Science.

Typy danych

Typy danych i cele analityczne idą w parze, podobnie jak kura i jajko; nie zawsze jest jasne, co jest pierwsze. Cele analityczne są wyprowadzane z celów biznesowych, ale typ danych również wpływa na cele. Na przykład cel biznesowy polegający na zrozumieniu postrzegania produktu konsumenckiego kieruje analitycznym celem analizy nastrojów. Podobnie, cel analizy nastrojów kieruje wyborem typu danych typu tekstowego, takiego jak zawartość mediów społecznościowych. Typ danych wpływa również na wiele innych wyborów podczas projektowania rozwiązań. Dane można klasyfikować na wiele sposobów. Często określa się dane jako ustrukturyzowane lub nieustrukturyzowane. Dane strukturalne istnieją, gdy informacje są wyraźnie podzielone na pola z rozszerzeniem wyraźne znaczenie i są wysoce kategorialne, porządkowe lub numeryczne. Powiązana kategoria, częściowo ustrukturyzowana, jest czasami używana do opisu danych ustrukturyzowanych, które nie są zgodne z formalną strukturą modeli danych związanych z relacyjnymi bazami danych lub innymi formami tabel danych, ale mimo to zawierają znaczniki lub inne znaczniki. Dane nieustrukturyzowane, takie jak tekst w języku naturalnym, mają mniej jasno określone znaczenie. Obrazy nieruchome, wideo i audio często należą do kategorii danych nieustrukturyzowanych. Dane w tej formie wymagają wstępnego przetwarzania w celu zidentyfikowania i wyodrębnienia odpowiednich „cech”. Funkcje to ustrukturyzowane informacje używane do indeksowania i pobierania lub uczenia klasyfikacji lub modeli klastrowych. Dane mogą być również klasyfikowane według tempa ich generowania, gromadzenia lub przetwarzania. Rozróżnia się dane przesyłane strumieniowo, które nieustannie napływają jak strumień wody z węża strażackiego, a dane wsadowe, które docierają w wiadrach. Chociaż rzadko istnieje połączenie między typem danych a szybkością transmisji danych, szybkość transmisji ma znaczący wpływ na model wykonania wybrany do realizacji analitycznej, a także może wpływać na decyzję klasy analitycznej lub model uczenia się.

Klasy technik analitycznych

Aby pomóc w konceptualizacji wszechświata możliwych technik analitycznych, pogrupowaliśmy je w dziewięć podstawowych klas. Zwróć uwagę, że techniki z danej klasy można stosować na wiele sposobów, aby osiągnąć różne cele analityczne. Członkostwo w klasie po prostu wskazuje na podobną funkcję analityczną.

* Transformacja Analityki

-Agregacja: Techniki podsumowania danych. Obejmują one podstawowe statystyki (np. Średnią, odchylenie standardowe), dopasowywanie rozkładu i wykresy graficzne.

-Wzbogacanie: Techniki dodawania dodatkowych informacji do danych, takich jak informacje o źródle lub inne etykiety.

-Przetwarzanie: Techniki dotyczące czyszczenia, przygotowania i separacji danych. Ta grupa obejmuje również typowe czynności przetwarzania wstępnego algorytmów, takie jak transformacje i wyodrębnianie cech.

*Uczenie się Analityki

-Regresja: Techniki szacowania relacji między zmiennymi, w tym zrozumienie, które zmienne są ważne w przewidywaniu przyszłych wartości.

-Klastrowanie: Techniki segmentacji danych na naturalnie podobne grupy.

-Klasyfikacja: Techniki identyfikacji członkostwa w grupach elementów danych.

-Zalecenie: Techniki przewidywania ratingu lub preferencji dla nowego podmiotu na podstawie historycznych preferencji lub zachowania.

*Analizy predykcyjne

-Symulacja: Techniki naśladowania działania procesu lub systemu w świecie rzeczywistym. Są one przydatne do przewidywania zachowania w nowych warunkach.

-Optymalizacja: Techniki badań operacyjnych skupiały się na wyborze najlepszego elementu z zestawu dostępnych alternatyw w celu maksymalizacji funkcji użyteczności.

Modele uczenia się

Klasy analityczne, które wykonują przewidywania, takie jak regresja, grupowanie, klasyfikacja i rekomendacje, wykorzystują modele uczenia się. Modele te charakteryzują sposób, w jaki analityk jest szkolony do dokonywania ocen na nowych danych w oparciu o obserwacje historyczne. Aspekty modeli uczenia się opisują zarówno typy dokonywanych osądów, jak i ewolucję modeli w czasie. Modele uczenia się są zwykle opisywane jako wykorzystujące uczenie się nienadzorowane lub nadzorowane. Uczenie nadzorowane ma miejsce, gdy model jest szkolony przy użyciu oznaczonego zestawu danych, który ma znaną klasę lub kategorię skojarzoną z każdym elementem danych. Model łączy funkcje znalezione w instancjach uczących z etykietami, dzięki czemu można tworzyć prognozy dla instancji bez etykiet. Modele uczenia się nienadzorowanego nie mają wiedzy a-priori o klasach, w których można umieścić dane. Używają funkcji w zestawie danych do tworzenia grup na podstawie podobieństwa funkcji. Użytecznym rozróżnieniem między modelami uczenia się są modele szkolone w jednym przebiegu, które są znane jako modele offline, oraz modele trenowane stopniowo w czasie, zwane modelami online. Wiele metod uczenia się ma warianty online lub offline. Decyzja o zastosowaniu jednego lub drugiego jest podejmowana na podstawie wybranych celów analitycznych i modeli wykonania. Generowanie modelu offline wymaga przejścia przez cały zestaw danych szkoleniowych. Udoskonalenie modelu wymaga wykonania oddzielnych przejść przez dane. Modele te są statyczne, ponieważ po przeszkoleniu ich przewidywania nie zmienią się, dopóki nowy model nie zostanie utworzony w kolejnym etapie szkolenia. Wydajność modelu offline jest łatwiejsza do oszacowania ze względu na to deterministyczne zachowanie. Wdrożenie modelu w środowisku produkcyjnym obejmuje wymianę starego modelu na nowy. Modele online mają zarówno zalety, jak i wady. Dynamicznie ewoluują w czasie, co oznacza, że wymagają tylko jednego wdrożenia w środowisku

produkcyjnym. Fakt, że te modele nie mają dostępnego całego zestawu danych podczas szkolenia, stanowi jednak wyzwanie. Muszą przyjąć założenia dotyczące danych na podstawie zaobserwowanych przykładów; te założenia mogą być nieoptymalne. Można to nieco zrównoważyć w przypadkach, gdy dostępna jest informacja zwrotna na temat prognoz modelu, ponieważ modele online mogą szybko uwzględniać informacje zwrotne w celu poprawy wydajności.

Modele wykonania

Modele wykonawcze opisują, w jaki sposób manipuluje się danymi w celu wykonania funkcji analitycznej. Można je podzielić na różne kategorie. Modele wykonania są zawarte w strukturze wykonywania, która organizuje sekwencjonowanie obliczeń analitycznych. W tym sensie struktura może być tak prosta, jak środowisko uruchomieniowe języka programowania, takie jak interpreter języka Python lub struktura przetwarzania rozproszonego, która zapewnia określony interfejs API dla jednego lub większej liczby języków programowania, takich jak Hadoop, MapReduce lub Spark. Grupowanie modeli wykonywania na podstawie sposobu, w jaki obsługują one dane, jest powszechne, klasyfikując je jako modele wykonywania wsadowego lub strumieniowego. Model wykonywania wsadowego oznacza, że dane są analizowane w dużych segmentach, że narzędzie analityczne ma stan, w którym działa i stan, w którym nie działa, a ten niewielki stan jest przechowywany w pamięci między wykonaniami. Wykonanie wsadowe może również oznaczać, że analiza generuje wyniki z częstotliwością rzędu kilku minut lub więcej. Obciążenia wsadowe są dość łatwe do konceptualizacji, ponieważ reprezentują dyskretne jednostki pracy. W związku z tym łatwo jest zidentyfikować określoną serię kroków wykonania, jak również odpowiednią częstotliwość wykonywania i ograniczenia czasowe w oparciu o szybkość, z jaką docierają dane. W zależności od wyboru algorytmu modele wykonywania wsadowego są łatwo skalowalne dzięki równoległości. Istnieje wiele struktur obsługujących równoległe wykonywanie analizy wsadowej. Najbardziej znany jest model Hadoop, który udostępnia rozproszony model wykonywania wsadowego w swojej strukturze MapReduce. I odwrotnie, model przesyłania strumieniowego analizuje dane w momencie ich nadejścia. Modele wykonywania przesyłania strumieniowego sugerują, że w normalnym działaniu narzędzie analityczne jest zawsze wykonywane. Narzędzie analityczne może przechowywać stan w pamięci i stale dostarczać wyniki w miarę napływu nowych danych, rzędu kilku sekund lub mniej. Wiele koncepcji przesyłania strumieniowego jest nieodłącznie związanych z filozofią projektowania Unixpipeline; procesy są połączone ze sobą poprzez połączenie danych wyjściowych jednego procesu z danymi wejściowymi następnego. W rezultacie wielu programistów zna już podstawowe koncepcje przesyłania strumieniowego. Dostępnych jest wiele platform obsługujących równoległe wykonywanie analiz strumieniowych, takich jak Storm, S4 i Samza. Wybór między modelami wykonywania wsadowego i strumieniowego często zależy od wymagań dotyczących opóźnienia analitycznego i terminowości. Opóźnienie odnosi się do ilości czasu wymaganego do przeanalizowania danych po ich dotarciu do systemu, podczas gdy terminowość odnosi się do średniego wieku odpowiedzi lub wyniku wygenerowanego przez system analityczny. W przypadku wielu celów analitycznych opóźnienie wynoszące godziny i terminowość dni jest akceptowalne, a zatem można je wdrożyć dzięki podejściu wsadowemu. Niektóre cele analityczne mają wymagania co do sekundy, a wynik sprzed kilku minut ma niewielką wartość. Model wykonywania przesyłania strumieniowego lepiej obsługuje takie cele. Modele wykonywania wsadowego i strumieniowego nie są jedynymi wymiarami, w ramach których można kategoryzować metody wykonywania analitycznego. Myśląc o skalowalności, dokonuje się innego rozróżnienia. W wielu przypadkach skalę można osiągnąć poprzez rozproszenie obliczeń na wielu komputerach. W tym kontekście niektóre algorytmy wymagają dużego stanu pamięci współużytkowanej, podczas gdy inne można łatwo zrównoleglać w kontekście, w którym nie istnieje stan współdzielony między maszynami. To rozróżnienie ma znaczący wpływ zarówno na wybór oprogramowania, jak i sprzętu podczas tworzenia równoległego środowiska analitycznego.

Wskazówki od profesjonalistów

Aby zrozumieć pojemność systemu w kontekście wykonywania analizy strumieniowej, zbierz metryki, w tym: ilość zużytych danych, wyemitowane dane i opóźnienia. Pomoże Ci to zrozumieć, kiedy zostaną osiągnięte limity skali.

Fraktalny model analityczny

Analityka Data Science jest bardzo podobna do brokułów. Fraktale to zbiory matematyczne, które wyświetlają samopodobne wzory. Gdy powiększasz fraktal, ponownie pojawiają się te same wzory. Wyobraź sobie łodygę brokułów. Oderwij kawałek brokuła, a kawałek wygląda podobnie do oryginalnej łodygi. Stopniowo mniejsze kawałki brokułów nadal wyglądają jak oryginalna łodyga. Analityka Data Science jest bardzo podobna do brokułów - z natury fraktale zarówno pod względem czasu, jak i konstrukcji. Wczesne wersje narzędzia analitycznego podlegają temu samemu procesowi programowania, co nowsze wersje. W dowolnej iteracji sama analiza jest zbiorem mniejszych analiz, które często rozkładają się na jeszcze mniejsze analizy.

Iteracyjne z natury

Dobra nauka o danych to fraktal w czasie - proces iteracyjny. Szybkie uzyskanie niedoskonałego rozwiązania za drzwiami wzbudzi większe zainteresowanie interesariuszy niż idealne rozwiązanie, które nigdy nie zostanie ukończone. Skonfiguruj infrastrukturę, zagreguj i przygotuj dane oraz wykorzystaj wiedzę ekspercką w tej dziedzinie. Wypróbuj różne techniki analityczne i modele na podzbiorach danych. Oceń modele, dopracuj, oceń ponownie i wybierz model. Zrób coś ze swoimi modelami i wynikami - zastosuj modele, aby informować, inspirować do działania i działać. Oceń wyniki biznesowe, aby ulepszyć cały produkt.

Mniejsze kawałki brokułów: produkt nauki o danych

Komponenty wewnątrz i na zewnątrz produktu Data Science będą się zmieniać z każdą iteracją. Przyjrzyjmy się produktowi Data Science i zbadajmy jego komponenty podczas jednej takiej iteracji. Aby osiągnąć wyższy cel analityczny, musisz najpierw rozłożyć problem na podkomponenty, aby podzielić i pokonać.

Cel: Najpierw musisz mieć pojęcie o swoim celu analitycznym i końcowym stanie analizy. Czy ma to na celu odkrywanie, opisywanie, przewidywanie czy doradzanie? To jest prawdopodobnie połączenie kilku z nich. Upewnij się, że zanim zaczniesz, zdefiniujesz wartość biznesową danych i sposób, w jaki planujesz wykorzystać spostrzeżenia do podejmowania decyzji, lub ryzykujesz, że skończysz na interesujących, ale niepodlegających działaniu ciekawostkach.

Dane: dane dyktują potencjalne spostrzeżenia, które mogą dostarczyć analizy. Data Science polega na znajdowaniu wzorców w zmiennych danych i porównywaniu tych wzorców. Jeśli dane nie są reprezentatywne dla całego świata zdarzeń, które chcesz analizować, będziesz chciał zebrać te dane poprzez starannie zaplanowane zmiany zdarzeń lub procesów poprzez testy A / B lub projektowanie eksperymentów. Zestawy danych nigdy nie są idealne, więc nie czekaj na idealne dane, aby rozpocząć. Dobry analityk danych jest biegły w obsłudze niechlujnych danych z brakującymi lub błędnymi wartościami. Po prostu pamiętaj, aby poświęcić czas na wyczyszczenie danych lub zaryzykować generowanie śmieci.

Obliczenia: Obliczenia dostosowują dane do celów poprzez proces tworzenia spostrzeżeń. Poprzez dzielenie i zdobywanie, obliczenia rozkładają się na kilka mniejszych możliwości analitycznych z własnymi celami, danymi, obliczeniami i wynikającymi z nich działaniami, tak jak mniejszy kawałek

brokułów zachowuje strukturę oryginalnej łodygi. W ten sposób same obliczenia są fraktalami. Budowanie zdolności bloku mogą wykorzystywać różne typy modeli wykonywania, takie jak obliczenia wsadowe lub przesyłanie strumieniowe, które indywidualnie realizują małe zadania. Właściwie połączone razem małe zadania dają złożone, wykonalne wyniki.

Działanie: Jak inżynierowie powinni zmienić proces produkcyjny, aby generować wyższą wydajność produktu? W jaki sposób firma ubezpieczeniowa powinna wybrać, które polisy komu zaoferować i za jaką cenę? Wynik obliczeń powinien umożliwiać działania, które są zgodne z celami produktu danych. Wyniki, które nie wspierają ani nie inspirują do działania, to nic innego jak ciekawe ciekawostki. Biorąc pod uwagę fraktalny charakter analityki Data Science w czasie i konstrukcji, istnieje wiele możliwości wyboru fantastycznych lub tandetnych analitycznych bloków konstrukcyjnych. Proces selekcji analitycznej dostarcza pewnych wskazówek.

Proces selekcji analitycznej

Jeśli skupisz się tylko na naukowym aspekcie Data Science, nigdy nie zostaniesz artystą danych. Krytycznym krokiem w nauce o danych jest zidentyfikowanie techniki analitycznej, która przyniesie pożądane działanie. Czasami jest to jasne; charakterystyka problemu (np. typ danych) wskazuje na technikę, którą należy zaimplementować. Jednak w innych przypadkach może być trudno wiedzieć, od czego zacząć. Wszechświat możliwych technik analitycznych jest duży. Znalezienie drogi w tym wszechświecie to sztuka, którą trzeba ćwiczyć. Poprowadzimy Cię przez kolejną część Twojej podróży - zostanie artystą danych.

Rozkładanie problemu

Rozłożenie problemu na możliwe do zarządzania części jest pierwszym krokiem w procesie analitycznej selekcji. Osiągnięcie pożądanego działania analitycznego często wymaga połączenia wielu technik analitycznych w całościowe, kompleksowe rozwiązanie. Inżynieria kompletnego rozwiązania wymaga rozłożenia problemu na coraz mniejsze podproblemy. Model analityczny fraktali ucieleśnia to podejście. Na każdym etapie sama analiza jest zbiorem mniejszych obliczeń, które rozkładają się na jeszcze mniejsze obliczenia. Gdy problem jest wystarczająco rozłożony, do osiągnięcia celu analitycznego potrzebna jest tylko jedna technika analityczna. Rozkład problemów tworzy wiele podproblemów, z których każdy ma własne cele, dane, obliczenia i działania. Na pozór problematyczny rozkład wydaje się być mechanicznym, powtarzalnym procesem. Chociaż koncepcyjnie może to być prawdą, tak naprawdę jest to wykonanie sztuki, a nie rozwiązanie problemu inżynieryjnego. Problem może być rozwiązany na wiele sposobów, a każdy z nich prowadzi do innego rozwiązania. Mogą istnieć ukryte zależności lub ograniczenia, które pojawiają się dopiero po rozpoczęciu opracowywania rozwiązania. Tu sztuka spotyka się z nauką. Chociaż sztuki stojącej za rozkładem problemów nie można się nauczyć, wyodrębniliśmy kilka pomocnych wskazówek, które mogą Cię poprowadzić. Kiedy zaczynasz myśleć o rozłożeniu problemu, poszukaj:

*Złożone cele analityczne, które tworzą naturalną segmentację. Na przykład wiele problemów związanych z przewidywaniem przyszłych warunków obejmuje zarówno cele Odkrywaj, jak i Przewiduj.

*Naturalne uporządkowanie celów analitycznych. Na przykład podczas wyodrębniania cech należy najpierw zidentyfikować cechy kandydujące, a następnie wybrać zestaw cech o najwyższej wartości informacyjnej. Te dwie czynności stanowią odrębne cele analityczne.

*Typy danych, które dyktują czynności przetwarzania. Na przykład tekst lub obrazy wymagają wyodrębnienia cech.

*Wymagania dotyczące sprzężenia zwrotnego człowieka w pętli. Na przykład podczas opracowywania progów ostrzegawczych może być konieczne zwrócenie się do analityków o opinie i zaktualizowanie progu na podstawie ich oceny.

*Konieczność łączenia wielu źródeł danych. Na przykład może być konieczne skorelowanie dwóch zestawów danych, aby osiągnąć szerszy cel. Często oznacza to obecność celu Odkryj

Oprócz dekompozycji problemu, zapewniającej wykonalne podejście do selekcji analitycznej, ma dodatkową zaletę w postaci uproszczenia bardzo złożonego problemu. Zamiast stawiać czoła zrozumieniu całego kompleksowego rozwiązania, obliczenia są dyskretnymi segmentami, które można zbadać. Należy jednak pamiętać, że chociaż ta technika pomaga naukowcom zajmującym się danymi w podejściu do problemu, należy ocenić kompleksowe rozwiązanie.

Identyfikacja sfałszowanych domen

Identyfikacja sfałszowanych domen jest ważna dla organizacji, aby zachować wizerunek marki i uniknąć utraty zaufania klientów. Sfałszowane domeny występują, gdy złośliwy autor tworzy witrynę internetową, adres URL lub adres e-mail, który zdaniem użytkowników jest powiązany z prawidłową organizacją. Kiedy użytkownicy klikają łącze, odwiedzają witrynę internetową lub otrzymują e-maile, są poddawani jakiejś niegodziwej aktywności. Zespół stanął przed problemem identyfikacji sfałszowanych domen dla firmy handlowej. Z pozoru problem wydawał się łatwy; wziąć niedawno zarejestrowaną domenę, sprawdzić, czy jest podobna do domeny firmy, i powiadomić, gdy podobieństwo jest wystarczająco duże. Jednak po rozłożeniu problemu główne obliczenia szybko się skomplikowały. Potrzebowali obliczenia, które określi podobieństwo między dwiema domenami. Gdy rozkładali obliczenia podobieństwa, problemem stała się złożoność i szybkość. Podobnie jak w przypadku wielu problemów związanych z bezpieczeństwem, bardzo ważne są szybkie szybkości alertów. Szybkość wyników stworzyła ograniczenie implementacyjne, które zmusiło do ponownej oceny sposobu, w jaki rozwiązali problem. Ponowna analiza procesu rozkładu doprowadziła do zupełnie nowego podejścia. Na koniec sporządzili listę domen podobnych do tych, które są zarejestrowane przez firmę. Następnie porównali tę listę z listą ostatnio zarejestrowanych domen.

Ograniczenia wdrożeniowe

W studium przypadku sfałszowanych domen pojawienie się ograniczenia implementacyjnego spowodowało, że zespół zrewidował swoje podejście. To pokazuje, że selekcja analityczna nie oznacza po prostu wyboru techniki analitycznej w celu osiągnięcia pożądanego wyniku. Oznacza to również zapewnienie wykonalności rozwiązania. Data Scientist może napotkać wiele różnych ograniczeń implementacyjnych. Można je jednak konceptualizować w kontekście pięciu wymiarów, które konkurują o twoją uwagę: złożoność analityczna, szybkość, dokładność i precyzja, rozmiar danych oraz złożoność danych. Równoważenie tych wymiarów jest grą o sumie zerowej - rozwiązanie analityczne nie może jednocześnie pokazywać wszystkich pięciu wymiarów, ale zamiast tego musi dokonywać transakcji między nimi.

SZYBKOŚĆ: Szybkość, z jaką musi zostać wygenerowany wynik analityczny (np. Prawie w czasie rzeczywistym, co godzinę, codziennie) lub czas potrzebny na opracowanie i wdrożenie rozwiązania analitycznego

ZŁOŻONOŚĆ ANALITYCZNA: Złożoność algorytmiczna (np. Klasa złożoności i zasoby wykonawcze)

DOKŁADNOŚĆ I PRECYZJA: Możliwość tworzenia dokładnych w porównaniu z przybliżonymi rozwiązaniami, a także zdolność do zapewnienia miary pewności

ROZMIAR DANYCH: rozmiar zbioru danych (np. Liczba wierszy)

ZŁOŻONOŚĆ DANYCH: typ danych, formalne miary złożoności, w tym miary nakładania się i liniowej rozdzielności, liczba wymiarów / kolumn i powiązania między zestawami danych.

Ograniczenia implementacyjne pojawiają się, gdy aspekt problemu dyktuje wartość jednego lub więcej z tych wymiarów. Gdy tylko jeden wymiar zostanie ustalony, Data Scientist jest zmuszony dokonać transakcji między innymi. Na przykład, jeśli problem analityczny wymaga podjęcia działań w czasie zbliżonym do rzeczywistego, wymiar prędkości jest stały, a transakcje muszą być dokonywane między innymi czterema wymiarami. Zrozumienie, które zawody osiągną właściwą równowagę wśród pięciu wymiarów, jest sztuką, której należy się nauczyć z czasem.

Niektóre typowe przykłady ograniczeń implementacji obejmują:

• Częstotliwość obliczeń. Rozwiązanie może wymagać regularnego uruchamiania (np. co godzinę), co wymaga wykonania obliczeń w określonym przedziale czasu. Najlepsze narzędzie analityczne jest bezużyteczne, jeśli nie może rozwiązać problemu w wymaganym czasie.

• Terminowość rozwiązań. Niektóre aplikacje wymagają wyników niemal w czasie rzeczywistym, wskazując na podejście do przesyłania strumieniowego. Podczas gdy niektóre algorytmy można zaimplementować w strukturach przesyłania strumieniowego, wiele innych nie.

• Szybkość wdrażania. Projekt może wymagać szybkiego opracowania i wdrożenia rozwiązania, aby szybko uzyskać informacje analityczne. W takich przypadkach może być konieczne skupienie się na mniej złożonych technikach, które można szybko wdrożyć i zweryfikować.

• Ograniczenia zasobów obliczeniowych. Chociaż możesz być w stanie przechowywać i analizować swoje dane, rozmiar danych może być na tyle duży, że algorytmy wymagające wielu obliczeń w całym zestawie danych wymagają zbyt dużej ilości zasobów. Może to wskazywać na potrzebę podejść, które wymagają tylko jednego przejścia danych (np. Klaster baldachimu w przeciwieństwie do grupowania k-średnich).

• Ograniczenia przechowywania danych. Są chwile, kiedy duże zbiory danych stają się tak duże, że nie można ich przechowywać lub można przechowywać tylko krótki horyzont czasowy. Metody analityczne, które wymagają długich horyzontów czasowych, mogą nie być możliwe.

Zasady organizacyjne i wymagania regulacyjne są głównym źródłem ukrytych ograniczeń, które zasługują na krótkie omówienie. Zasady są często ustalane wokół określonych klas danych, takich jak dane osobowe (PII) lub dane osobowe (PHI). Chociaż obecnie dostępne technologie mogą bezpiecznie przechowywać informacje z różnymi kontrolami bezpieczeństwa w jednym systemie, zasady te wymuszają specjalne rozważania dotyczące obsługi danych, w tym ograniczone okresy przechowywania i dostęp do danych. Ograniczenia danych wpływają na rozmiar danych i wymiary złożoności opisane wcześniej, tworząc kolejną warstwę ograniczeń, które należy wziąć pod uwagę.

Wskazówki od profesjonalistów

Jeśli to możliwe, rozważ podejścia, które wykorzystują wcześniej obliczone wyniki. Twój algorytm będzie działał znacznie szybciej, jeśli możesz tego uniknąć ponowne obliczenie wartości w całym horyzoncie czasowym danych.

Nasz cykl życia produktu do nauki o danych ewoluował, aby szybko uzyskiwać wyniki, a następnie stopniowo ulepszać rozwiązanie.

Podejścia do przesyłania strumieniowego mogą być przydatne do pokonywania ograniczeń magazynu.

Szczegółowa tabela analiz

Nie wystarczy dotrzeć do właściwego punktu wyjścia. Zapewniamy również tłumacza, abyś zrozumiał, co Ci powiedziano. Zidentyfikowanie kilku technik analitycznych, które można zastosować do twojego problemu, jest przydatne, ale sama ich nazwa nie będzie zbyt pomocna. Szczegółowa tabela danych analitycznych przekłada nazwy na coś bardziej znaczącego.

Algorytmy lub nazwa metody: Opis: Wskazówki od profesjonalistów

Symulacja oparta na agentach: symuluje działania i interakcje autonomicznych agentów. : W wielu systemach złożone zachowanie wynika z zaskakująco prostych reguł. : Zachowaj prostą logikę swoich agentów i stopniowo budować wyrafinowanie.

Filtrowanie oparte na współpracy: nazywane również „zaleceniem" lub eliminowanie pozycji ze zbioru porównując historię działań z pozycjami wykonanymi przez użytkowników. Znajduje podobne elementy na podstawie tego, kto ich używał lub podobnych użytkowników na podstawie używanych przez nich przedmiotów. : Użyj rekomendacji opartej na rozkładzie wartości osobliwych w przypadkach, gdy w Twojej domenie występują czynniki ukryte, np. gatunki w filmach.

Transformacja współrzędnych: zapewnia inne spojrzenie na dane. : Zmiana układu współrzędnych dla danych, na przykład przy użyciu współrzędnych biegunowych lub cylindrycznych, może łatwiej uwydatnić kluczową strukturę w dane. Kluczowym krokiem w transformacji współrzędnych jest docenienie wielowymiarowości i systematyczne analizowanie podprzestrzeni danych.

Projektowanie eksperymentów: stosuje kontrolowane eksperymenty w celu ilościowego określenia wpływu na wydajność systemu spowodowanego zmianami danych wejściowych. : Ułamkowe plany czynnikowe mogą znacznie zmniejszyć liczbę różnych typów eksperymentów, które musisz przeprowadzić.

Równania różniczkowe: używane do wyrażania relacji między funkcjami i ich pochodnymi, na przykład zmieniającymi się w czasie. : Równania różniczkowe mogą służyć do formalizowania modeli i prognozowania. Same równania można rozwiązać numerycznie i przetestować w różnych warunkach początkowych w celu zbadania trajektorii systemu.

Symulacja zdarzeń dyskretnych: Symuluje dyskretną sekwencję zdarzeń, w których każde zdarzenie występuje w określonym momencie. : Model aktualizuje swój stan tylko w momencie wystąpienia zdarzeń.: Symulacja zdarzeń dyskretnych jest przydatna podczas analizowania procesów opartych na zdarzeniach, takich jak linie produkcyjne i centra serwisowe, w celu określenia, jak zmienia się zachowanie na poziomie systemu wraz ze zmianą różnych parametrów procesu. Optymalizację można zintegrować z symulacją, aby zwiększyć wydajność procesu.

Dyskretna transformata falkowa: przekształca dane szeregów czasowych w dziedzinę częstotliwości z zachowaniem informacji o lokalizacji. : Oferuje bardzo dobrą lokalizację w czasie i częstotliwości. Zaletą w stosunku do transformacji Fouriera jest to, że zachowuje zarówno częstotliwość, jak i lokalność.

Wygładzanie wykładnicze: służy do usuwania artefaktów oczekiwanych od błędu kolekcji lub wartości odstających. : W porównaniu do średniej ruchomej, w której poprzednie obserwacje są ważone jednakowo, wygładzanie wykładnicze przypisuje wagi malejące wykładniczo w czasie.

Analiza czynnikowa: opisuje zmienność między skorelowanymi zmiennymi w celu zmniejszenia liczby nieobserwowanych zmiennych, czyli czynników. : Jeśli podejrzewasz, że na Twoje dane mają niewymierny wpływ, możesz spróbować analizy czynnikowej.

Szybka transformata Fouriera: wydajnie przekształca szeregi czasowe od czasu do dziedziny częstotliwości. Może być również używany do ulepszania obrazu poprzez transformacje przestrzenne. : Filtrowanie zmieniającego się w czasie sygnału można przeprowadzić skuteczniej w dziedzinie częstotliwości. Szum można również często zidentyfikować w takich sygnałach, obserwując moc przy nieprawidłowych częstotliwościach.

Konwersja formatu: tworzy standardową reprezentację danych niezależnie od formatu źródłowego. Na przykład wyodrębnianie surowego tekstu zakodowanego w formacie UTF-8 z plików binarnych, takich jak Microsoft Word lub PDF. : Istnieje wiele pakietów oprogramowania typu open source, które obsługują konwersję formatów i mogą interpretować wiele różnych formatów. Jednym z godnych uwagi pakietów jest Apache Tikia.

Filtrowanie Gaussa: Działa w celu usunięcia szumu lub rozmycia danych. : Może być używany do usuwania szumu plamkowego z obrazów.

Uogólnione modele liniowe: rozszerza zwykłą regresję liniową w celu dopuszczenia lub rozkładu błędu, który nie jest normalny. : Użyj, jeśli zaobserwowany błąd w systemie nie ma normalnego rozkładu.

Algorytmy genetyczne: Ewoluuje modele kandydatów na przestrzeni pokoleń przez inspirowane ewolucją operatory mutacji i krzyżowania się parametrów. : Zwiększenie rozmiaru generacji zwiększa różnorodność w rozważaniu kombinacji parametrów, ale wymaga bardziej obiektywnej oceny funkcji. Obliczanie liczby osób w ramach pokolenia jest bardzo równoległe. Przedstawienie proponowanych rozwiązań może wpłynąć na wydajność.

Grid Search: Systematyczne przeszukiwanie dyskretnych wartości parametrów pod kątem problemów z eksploracją parametrów. : Siatka obejmująca parametry służy do wizualizacji krajobrazu parametrów i oceny, czy istnieje wiele minimów.

Ukryte modele Markowa: Modeluje dane sekwencyjne poprzez określenie dyskretnych zmiennych utajonych, ale obserwable mogą być ciągłe lub dyskretne. : Jedną z najpotężniejszych właściwości ukrytych modeli Markowa jest ich zdolność do wykazywania pewnego stopnia niezmienności lokalnego wypaczenia (ściskania i rozciągania) osi czasu. Jednak istotną słabością Ukrytego Modelu Markowa jest sposób, w jaki przedstawia on rozkład czasów, przez które system pozostaje w danym stanie.

Klastrowanie hierarchiczne: podejście oparte na łączności, które sekwencyjnie tworzy większe (aglomeracyjne) lub mniejsze (dzielące) klastry w danych. : Zapewnia widoki klastrów w wielu rozdzielczościach bliskości. Algorytmy zaczynają zwalniać w przypadku większych zestawów danych, ponieważ większość implementacji wykazuje złożoność O (N3) lub O (N2).

K-średnich i X-średnich Clustering: Algorytmy klastrowania oparte na centroidach, gdzie K oznacza liczbę klastrów, a X oznacza liczbę klastrów jest nieznana. : Stosując techniki grupowania, pamiętaj o zrozumieniu kształtu danych. Techniki grupowania dadzą słabe wyniki, jeśli dane nie są okrągłe lub elipsoidalne.

Programowanie liniowe, nieliniowe i całkowite: Zestaw technik minimalizacji lub maksymalizacji funkcji w ograniczonym zestawie parametrów wejściowych .: Rozpocznij od programów liniowych, ponieważ algorytmy dla zmiennych całkowitych i nieliniowych mogą trwać znacznie dłużej.

Markov Chain Monte Carlo (MCMC): Metoda próbkowania stosowana zazwyczaj w modelach bayesowskich do oszacowania łącznego rozkładu parametrów podanych danych. : Problemy, które są trudne do rozwiązania przy użyciu metod analitycznych, mogą stać się możliwe do rozwiązania za pomocą MCMC, nawet biorąc pod uwagę problemy wielowymiarowe. Wykonalność jest wynikiem

wykorzystania statystyk dotyczących leżących u podstaw rozkładów zainteresowania, a mianowicie próbkowania metodą Monte Carlo i rozważenia stochastycznego sekwencyjnego procesu łańcuchów Markowa.

Metody Monte Carlo: zbiór technik obliczeniowych do generowania liczb losowych. : Szczególnie przydatny do całkowania numerycznego, rozwiązywania równań różniczkowych, obliczeń bocznych bayesowskich i wielowymiarowego próbkowania wielowymiarowego.

Naiwny Bayes: przewiduje klasy zgodnie z twierdzeniem Bayesa, które określa prawdopodobieństwo wyniku przy danym zestawie cech na podstawie prawdopodobieństwa cech danego wyniku. : Zakłada, że wszystkie zmienne są niezależne, więc może mieć problemy z uczeniem się w kontekście wysoce współzależnych zmiennych. Model można nauczyć się podczas pojedynczego przebiegu danych przy użyciu prostych liczników, a zatem jest przydatny w określaniu, czy możliwe do wykorzystania wzorce istnieją w dużych zestawach danych przy minimalnym czasie tworzenia.

Sieci neuronowe: poznaje najważniejsze cechy danych, dostosowując wagi między węzłami za pomocą reguły uczenia się. : Uczenie sieci neuronowej trwa znacznie dłużej niż ocena nowych danych przy użyciu już wyuczonej sieci. Rzadsza łączność sieciowa może pomóc w segmentowaniu przestrzeni wejściowej i poprawie wydajności zadań klasyfikacyjnych.

Usuwanie wartości odstających: metoda identyfikacji i usuwania szumu lub artefaktów z danych. : Zachowaj ostrożność podczas usuwania wartości odstających. Czasami najbardziej interesującym zachowaniem systemu są momenty, w których występują nieprawidłowe punkty danych.

Analiza głównych komponentów: umożliwia redukcję wymiarowości poprzez identyfikację wysoce skorelowanych wymiarów. : Wiele dużych zbiorów danych zawiera korelacje między wymiarami; dlatego część zbioru danych jest zbędna. Analizując otrzymane główne składniki, uszereguj je według wariancji, ponieważ jest to najwyższy widok informacyjny danych. Użyj wykresów skree, aby wywnioskować optymalną liczbę komponentów.

Regresja ze skurczem (Lasso): Metoda selekcji zmiennych i przewidywania połączona w potencjalnie obciążony model liniowy. : Istnieją różne metody wyboru parametru lambda. Typowym wyborem jest walidacja krzyżowa z MSE jako metryką.

Analiza wrażliwości: obejmuje testowanie poszczególnych parametrów w analityce lub modelu i obserwowanie wielkości efektu. : Niewrażliwe parametry modelu podczas optymalizacji są kandydatami do ustawienia na stałe. Zmniejsza to wymiarowość problemów optymalizacji i daje możliwość przyspieszenia.

Symulowane wyżarzanie: Nazwane po kontrolowanym procesie chłodzenia w metalurgii i przez analogię przy użyciu zmiany temperatury lub harmonogramu wyżarzania w celu zmiany zbieżności algorytmicznej. : Standardowa funkcja wyżarzania pozwala na początkową szerokie badanie przestrzeni parametrów, po której następuje węższe wyszukiwanie. W zależności od priorytetu wyszukiwania można zmodyfikować funkcję wyżarzania, aby umożliwić dłuższe poszukiwania eksploracyjne w wysokiej temperaturze.

Regresja krokowa: metoda wyboru i przewidywania zmiennych. Kryterium informacyjne Akaike AIC jest używane jako metryka do wyboru. Powstały model predykcyjny oparty jest na zwykłych najmniejszych kwadratach lub ogólnym modelu liniowym z estymacją parametrów poprzez maksymalne prawdopodobieństwo. : Należy zachować ostrożność rozważając regresję krokową, ponieważ często występuje nadmierne dopasowanie. Aby złagodzić nadmierne dopasowanie, spróbuj ograniczyć liczbę używanych wolnych zmiennych.

Stochastic Gradient Descent: Optymalizacja ogólnego przeznaczenia do uczenia się sieci neuronowych, maszyn wektorów nośnych i modeli regresji logistycznej. : Stosowane w przypadkach, gdy funkcja celu nie jest całkowicie różniczkowalna przy użyciu gradientów podrzędnych.

Maszyny wektorów pomocniczych: Projekcja wektorów cech za pomocą funkcji jądra do przestrzeni, w której klasy są bardziej rozdzielne. : Wypróbuj wiele jąder i użyj k-krotnej weryfikacji krzyżowej, aby potwierdzić wybór najlepszego.

Odwrotna częstotliwość terminów Częstotliwość dokumentu: Statystyka, która mierzy względne znaczenie terminu z korpusu. : Zwykle używane w eksploracji tekstu. Zakładając zbiór artykułów z wiadomościami, termin, który jest bardzo częsty, taki jak „the", będzie prawdopodobnie pojawiał się wiele razy w wielu dokumentach i będzie miał niską wartość. Termin, który jest rzadki, taki jak nazwisko osoby, które pojawia się w jednym artykule, będzie miał wyższy wynik TDF.

Modelowanie tematów (ukryta alokacja Dirichleta): identyfikuje ukryte tematy w tekście poprzez badanie współwystępowania słów. : Zastosuj tagowanie części mowy, aby wyeliminować słowa inne niż rzeczowniki i czasowniki. Używaj surowych liczników terminów zamiast terminów ważonych TF / IDF.

Metody oparte na drzewie: modele mają strukturę drzew graficznych, których gałęzie wskazują decyzje. : Może służyć do usystematyzowania procesu lub działać jako klasyfikator.

Metody opakowania: metoda redukcji zestawu funkcji, która wykorzystuje wydajność zestawu funkcji modelu jako miarę wydajności zestawu funkcji. Może pomóc zidentyfikować kombinacje funkcji w modelach, które osiągają wysoką wydajność. : Wykorzystaj walidację krzyżową k-fold, aby kontrolować dopasowanie.

Inżynieria funkcji

Inżynieria cech jest bardzo podobna do tlenu. Nie możesz się bez tego obejść, ale rzadko poświęcasz temu dużo uwagi.

Inżynieria cech to proces, w którym ustala się reprezentację danych w kontekście podejścia analitycznego. To podstawowa umiejętność w Data Science. Bez inżynierii cech nie byłoby możliwe zrozumienie i przedstawienie świata za pomocą modelu matematycznego. Inżynieria funkcji to trudna sztuka. Podobnie jak inne sztuki, jest to proces twórczy, który przejawia się w wyjątkowy sposób w każdym Data Scientist. Istotny wpływ na to będą miały doświadczenia, gusta i zrozumienie danej dziedziny naukowca. Jak sama nazwa wskazuje, inżynieria funkcji może być złożonym zadaniem, które może obejmować tworzenie łańcuchów i testowanie różnych podejść. Funkcje mogą być proste, takie jak „worek słów", popularna technika w dziedzinie przetwarzania tekstu, lub mogą być oparte na bardziej złożonych reprezentacjach uzyskanych w wyniku działań, takich jak uczenie maszynowe. Korzystasz z wyniku jednej techniki analitycznej, aby utworzyć reprezentację, która jest używana przez inną. Najczęściej znajdziesz się w świecie bardzo złożonych działań.

Wyszukiwanie chemoinformatyczne

Podczas jednego zadania zespół stanął przed wyzwaniem opracowania wyszukiwarki związków chemicznych. Cel poszukiwań chemoinformatycznych polega na przewidywaniu właściwości, które cząsteczka będzie wykazywać, a także zapewniać wskaźniki dotyczące przewidywanych właściwości, aby ułatwić odkrywanie danych w badaniach opartych na chemii. Te właściwości mogą być dyskretne (np. „Cząsteczka dobrze leczy chorobę") lub ciągłe (np. „Cząsteczka może rozpuszczać się do 100,21 g / ml "). Cząsteczki są złożonymi strukturami 3D, które są zwykle przedstawiane jako lista atomów

połączonych wiązaniami chemicznymi o różnej długości, o różnych domenach elektronowych i geometriach molekularnych. Struktury są określone przez współrzędne w 3 przestrzeni i potencjał elektrostatyczny powierzchni atomów w cząsteczce. Przeszukiwanie tych danych jest trudnym zadaniem, jeśli weźmie się pod uwagę, że naiwne podejście do problemu ma istotne podobieństwo do problemu izomorfizmu grafów. Opracowano rozwiązanie oparte na wcześniejszych pracach nad molekularnymi odciskami palców (czasami nazywanymi również mieszaniem lub mieszaniem z uwzględnieniem lokalizacji). Odciski palców to technika redukcji wymiarowości, która radykalnie ogranicza problematykę poprzez podsumowanie wielu cech, często z niewielkim uwzględnieniem znaczenia cechy. Gdy dokładne rozwiązanie jest prawdopodobnie niewykonalne, często zwracamy się do podejść heurystycznych, takich jak pobieranie odcisków palców. W naszym podejściu wykorzystaliśmy zestaw treningowy, w którym dostępne były wszystkie zmierzone właściwości cząsteczek. Stworzyliśmy model tego, jak podobieństwa struktur molekularnych mogą wpływać na ich właściwości. Zaczęliśmy od znalezienia wszystkich pod-grafów każdej cząsteczki o długości n, co dało reprezentację podobną do podejścia worka słów z przetwarzania języka naturalnego. Podsumowaliśmy każdy fragment cząsteczki w rodzaju odcisków palców zwanych „Counting Bloom Filter". Następnie wykorzystaliśmy kilka przykładów z zestawu do stworzenia nowych funkcji. Znaleźliśmy odległość od każdego członka pełnego zestawu treningowego do każdego z przykładów. Wprowadziliśmy te funkcje do algorytmu regresji nieliniowej, aby uzyskać model, którego można by użyć na danych, których nie było w oryginalnym zestawie uczącym. Podejście to można konceptualizować jako „ukrytą rozmaitość", w której ukryta powierzchnia lub kształt definiuje sposób, w jaki cząsteczka będzie wykazywać właściwość. Przybliżamy ten kształt za pomocą regresji nieliniowej i zestawu danych o znanych właściwościach. Kiedy już mamy przybliżony kształt, możemy go użyć do przewidywania właściwości nowych cząsteczek. Nasze podejście było wieloetapowe i złożone - wygenerowaliśmy pod-wykresy, stworzyliśmy filtry poświaty, obliczyliśmy metryki odległości i dopasowaliśmy model regresji liniowej. Ten przykład ilustruje, ile etapów może być zaangażowanych w tworzenie złożonej reprezentacji funkcji. Dzięki kreatywnemu łączeniu i budowaniu „funkcji na funkcjach" byliśmy w stanie tworzyć nowe reprezentacje danych, które były bogatsze i bardziej opisowe, a jednocześnie działały szybciej i dawały lepsze wyniki.

Wybór funkcji

Wybór cech to proces określania zbioru cech o najwyższej wartości informacyjnej dla modelu. Dwa główne podejścia to metody filtrowania i opakowywania. Metody filtrowania analizują cechy za pomocą statystyki testowej i eliminują funkcje nadmiarowe lub nieinformacyjne. Na przykład metoda filtrowania może wyeliminować cechy, które mają niewielką korelację z etykietami klas. Metody opakowujące wykorzystują model klasyfikacji jako część wyboru cech. Model jest uczony na zbiorze cech, a dokładność klasyfikacji jest używana do pomiaru wartości informacyjnej zbioru cech. Jednym z przykładów jest uczenie sieci neuronowej za pomocą zestawu funkcji i ocena dokładności modelu. Jeśli model uzyska wysoką ocenę w zbiorze testowym, wówczas cechy mają wysoką wartość informacyjną. Wszystkie możliwe kombinacje funkcji są testowane w celu znalezienia najlepszego zestawu funkcji. Istnieją kompromisy między tymi technikami. Metody filtrowania są szybsze do obliczenia, ponieważ każda funkcja musi być porównana tylko z etykietą klasy. Z drugiej strony metody opakowujące oceniają zestawy funkcji, konstruując modele i mierząc wydajność. Wymaga to przeszkolenia i oceny dużej liczby modeli (ilość, która rośnie wykładniczo wraz z liczbą cech). Dlaczego ktoś miałby używać metody opakowującej? Zestawy funkcji mogą działać lepiej niż pojedyncze funkcje. Dzięki metodom filtrującym eliminowana jest funkcja o słabej korelacji z etykietami klas. Niektóre z tych wyeliminowanych funkcji mogły jednak działać dobrze w połączeniu z innymi funkcjami.

Klasyfikacja komórek rakowych

W jednym z projektów zespół miał za zadanie sklasyfikować profile komórek rakowych. Nadrzędnym celem była klasyfikacja różnych typów białaczki na podstawie profili mikromacierzy z 72 próbek przy użyciu niewielkiego zestawu cech. Wykorzystali hybrydową sztuczną sieć neuronową (ANN) i algorytm genetyczny do identyfikacji podzbiorów 10 cech wybranych z tysięcy. Przeszkolili SSN i przetestowali wydajność przy użyciu weryfikacji krzyżowej. Miara wydajności została wykorzystana jako sprzężenie zwrotne do algorytmu genetycznego. Gdy zestaw funkcji nie zawierał użytecznych informacji, model działał słabo i badany byłby inny zestaw funkcji. Z biegiem czasu ta metoda wybrała zestaw funkcji, które działają z dużą dokładnością. Wybrana funkcja zapewniła większą szybkość i wydajność, a także umożliwiła lepszy wgląd w czynniki, które mogą rządzić systemem. Pozwoliło to naszemu zespołowi zaprojektować test diagnostyczny tylko dla kilku markerów genetycznych zamiast tysięcy, znacznie zmniejszając złożoność i koszt testów diagnostycznych.

Wiarygodność danych

Jesteśmy naukowcami danych, a nie alchemikami danych. Nie możemy stworzyć analitycznego złota na podstawie danych. Podczas gdy większość ludzi kojarzy objętość, szybkość i różnorodność danych z dużymi zbiorami danych, istnieje równie ważny, ale często pomijany wymiar - prawdziwość danych. Prawdziwość danych odnosi się do ogólnej jakości i poprawności danych. Musisz ocenić prawdziwość i dokładność danych, a także zidentyfikować brakujące lub niekompletne informacje. Jak to się mówi, „śmieci wchodzą, śmieci wychodzą". Jeśli Twoje dane są niedokładne lub brakuje w nich informacji, nie możesz mieć nadziei na uzyskanie złota analitycznego. Ocena prawdziwości danych jest często subiektywna. Musisz polegać na swoim doświadczeniu i zrozumieniu pochodzenia danych i kontekstu. Znajomość domeny jest szczególnie ważna w przypadku tych ostatnich. Chociaż ocena dokładności danych może być również subiektywna, czasami można zastosować metody ilościowe. Możesz ponownie pobrać próbkę z populacji i przeprowadzić porównanie statystyczne z zapisanymi wartościami, zapewniając w ten sposób miary dokładności. Najczęściej napotykane problemy to brakujące lub niepełne informacje. Istnieją dwie podstawowe strategie radzenia sobie z brakującymi wartościami - usuwanie i imputacja. W pierwszym przypadku całe obserwacje są wyłączone z analizy, zmniejszając wielkość próby i potencjalnie wprowadzając błąd. Imputacja, czyli zastępowanie brakujących lub błędnych wartości, wykorzystuje różne techniki, takie jak losowe pobieranie próbek (imputacja gorącej płyty) lub zastępowanie za pomocą średniej, rozkładów statystycznych lub modeli.

Wskazówki od profesjonalistów

Znajdź podejście, które działa, zastosuj je i idź dalej. Możesz martwić się optymalizacją i dostrajaniem swoich metod później podczas stopniowego ulepszania

Modelowanie szeregów czasowych

Na jednym z projektów zespół stanął przed skorelowaniem szeregów czasowych dla różnych parametrów. Wstępna analiza wykazała, że korelacje prawie nie istniały. Przeanalizowali dane i szybko odkryli problemy z ich prawdziwością. Brakowało wartości zerowych a także obserwacje o wartości ujemnej, co było niemożliwe z uwagi na kontekst pomiarów (patrz rysunek, Dane szeregów czasowych przed oczyszczeniem). Dane o śmieciach oznaczały śmieciowe wyniki. Ponieważ wielkość próby była już niewielka, usuwanie obserwacji było niepożądane. Niestabilny charakter szeregów czasowych oznaczał, że nie można było ufać, że imputacja poprzez próbkowanie da wartości, których zespół byłby pewien. W rezultacie szybko zdali sobie sprawę, że najlepszą strategią jest podejście umożliwiające filtrowanie i korygowanie szumów w danych. Początkowo wypróbowali uproszczone podejście, w którym każdą obserwację zastępowaliśmy średnią ruchomą. Chociaż poprawiło to niektóre szumy, w

tym wartości odstające w naszych obliczeniach średniej ruchomej, przesunęło szeregi czasowe. To spowodowało niepożądane zniekształcenie sygnału bazowego i szybko porzucili to podejście. Jeden z członków zespołu, który miał doświadczenie w przetwarzaniu sygnałów, zasugerował zastosowanie filtra medianowego. Filtr mediany to technika okienkowa, która porusza się po danych punkt po punkcie i zastępuje ją medianą obliczoną dla bieżącego okna. Eksperymentowano z różnymi rozmiarami okien, aby uzyskać akceptowalny kompromis między wygładzaniem szumu a wygładzaniem sygnału. Zastosowanie metody filtra mediany było niezwykle skuteczne. Wizualna inspekcja wykresów szeregów czasowych ujawnia wygładzanie wartości odstających bez tłumienia naturalnie występujących szczytów i dołków (brak utraty sygnału). Przed wygładzeniem nie widzieliśmy żadnej korelacji w naszych danych, ale później wartość Rho Spearmana wynosiła ~ 0,5 dla prawie wszystkich parametrów. Rozwiązując nasze problemy z prawdziwością danych, byliśmy w stanie stworzyć analityczne złoto. Chociaż inne podejścia również mogły być skuteczne, ograniczenia szybkości implementacji uniemożliwiły nam dalszą analizę. Odnieśliśmy sukces, o który nam chodzi, i zajęliśmy się innymi aspektami problemu.

Zastosowanie wiedzy domenowej

Wszyscy jesteśmy wyjątkowi na swój sposób. Nie lekceważ tego, co wiesz. Znajomość dziedziny, w której tkwi problem, jest niezwykle cenna i niezastąpiona. Zapewnia dogłębne zrozumienie danych i czynników wpływających na cel analityczny. Często wiedza domenowa jest kluczowym wyróżnikiem sukcesu zespołu Data Science. Wiedza domenowa wpływa na to, jak projektujemy i wybieramy funkcje, wprowadzamy dane, wybieramy algorytm i określamy sukces. Jedna osoba nie może być jednak ekspertem w każdej dziedzinie. Polegamy na naszym zespole, innych analitykach i ekspertach dziedzinowych, a także konsultujemy artykuły badawcze i publikacje w celu zrozumienia domeny.

Kradzież pojazdu silnikowego

W ramach jednego projektu zespół zbadał, w jaki sposób można zastosować naukę o danych w celu poprawy bezpieczeństwa publicznego. Według FBI około 8 miliardów dolarów jest traconych rocznie z powodu kradzieży samochodów. Odzyskanie miliona pojazdów skradzionych każdego roku w USA to mniej niż 60%. Radzenie sobie z tymi przestępstwami stanowi znaczną inwestycję w środki ochrony porządku publicznego. Chciano sprawdzić, czy możemy określić, jak ograniczyć kradzież samochodów, efektywnie korzystając z zasobów organów ścigania. Zespół rozpoczął od przeanalizowania i zweryfikowania danych o przestępstwach w San Francisco. Wzbogacili raportowanie kradzieży samochodu o ogólne dane miasta. Po przeprowadzeniu kilku eksperymentów z danymi w przestrzeni i czasie, wyłoniły się trzy geoprzestrzenne i jeden czasowy hotspot. Ekspert dziedzinowy z zespołu był w stanie stwierdzić, że główny hotspot geoprzestrzenny odpowiadał obszarowi otoczonemu parkami. Parki stworzyły miejską górę z wieloma punktami dojścia pieszego, które sprzyjały kradzieży samochodów. Nasz zespół wykorzystał informacje o czasowych hotspotach w połączeniu ze spostrzeżeniami eksperta dziedzinowego, aby opracować model Monte Carlo do przewidywania prawdopodobieństwa kradzieży pojazdu silnikowego na poszczególnych skrzyżowaniach miasta. Ustalając priorytety skrzyżowań zidentyfikowanych przez model, samorządy będą miały informacje niezbędne do efektywnego rozmieszczenia patroli. Można by ograniczyć kradzieże pojazdów silnikowych i skuteczniej rozmieścić środki ścigania. Analiza, oparta na wiedzy specjalistycznej, dostarczyła praktycznych informacji, które mogą uczynić ulice bezpieczniejszymi.

Klątwa wymiarowości

Nie ma magicznej mikstury, która wyleczyłaby klątwę, ale jest PCA. „Przekleństwo wymiarowości” jest jednym z najważniejszych rezultatów uczenia maszynowego. Większość tekstów poświęconych uczeniu maszynowemu wspomina o tym zjawisku w pierwszym lub drugim rozdziale, ale często

potrzeba wielu lat praktyki, aby zrozumieć jego prawdziwe konsekwencje. Metody klasyfikacji, podobnie jak większość metod uczenia maszynowego, podlegają konsekwencjom przekleństwa wymiarowości. Podstawowa intuicja w tym przypadku jest taka, że wraz ze wzrostem liczby wymiarów danych, tworzenie modeli klasyfikacji dających się uogólnić (modele, które dobrze sprawdzają się w stosunku do zjawisk nieobserwowanych w zbiorze uczącym) staje się trudniejsze. Ta trudność jest zwykle niemożliwa do pokonania w warunkach rzeczywistych. Istnieją pewne wyjątki w domenach, w których sprawy się układają, ale zazwyczaj musisz pracować nad zminimalizowaniem liczby wymiarów. Wymaga to połączenia sprytnej inżynierii cech i wykorzystania technik redukcji wymiarowości. Z naszego praktycznego doświadczenia wynika, że maksymalna liczba wymiarów wydaje się wynosić ~ 10 dla podejść opartych na modelu liniowym. Wydaje się, że limit wynosi dziesiątki tysięcy dla bardziej wyrafinowanych metod, takich jak maszyny wektorów nośnych, ale nadal istnieje. Przeciw intuicyjną konsekwencją przekleństwa wymiarowości jest to, że ogranicza ilość danych potrzebnych do wytrenowania modelu klasyfikacyjnego. Zjawisko to ma mniej więcej dwie przyczyny. W jednym przypadku wymiarowość jest na tyle mała, że model można trenować na jednej maszynie. W drugim przypadku wykładniczo rozszerzająca się złożoność problemu o dużej wymiarowości sprawia, że trenowanie modelu jest (praktycznie) obliczeniowo niemożliwe. Z doświadczenia wynika, że rzadko zdarza się, aby problem znajdował się w „słodkim miejscu" między tymi dwoma skrajnościami. To spostrzeżenie nie oznacza, że taki stan nigdy nie występuje. Uważamy jednak, że rzadko zdarza się, aby praktycy nie musieli zajmować się tym, jak rozwiązać ten przypadek. Zamiast próbować tworzyć superskalowalne implementacje algorytmów, skup się na rozwiązywaniu bezpośrednich problemów za pomocą podstawowych metod. Poczekaj, aż napotkasz problem polegający na tym, że algorytm nie jest zbieżny lub zapewnia słabe wyniki weryfikacji krzyżowej, a następnie poszukaj nowych podejść. Dopiero gdy okaże się, że alternatywne podejścia jeszcze nie istnieją, należy rozpocząć tworzenie nowych implementacji. Oczekiwany koszt tego schematu pracy jest niższy niż nadmierna inżynieria zaraz po wyjściu z bramy. Inaczej mówiąc: „Niech to będzie proste, głupie".

Pieczenie ciasta

Kiedyś otrzymałem szereg czasowy zawierający około 1600 zmiennych predykcyjnych i 16 zmiennych docelowych i poproszono mnie o zaimplementowanie szeregu modeli techniki przewidywania wartości zmiennych docelowych. Klient został wezwany do radzenia sobie ze złożonością związaną z dużą liczbą zmiennych i potrzebował pomocy. Nie tylko miałem przypadek klątwy, ale także zmienne predykcyjne były dość zróżnicowane. Na pierwszy rzut oka wyglądało to jak próba upieczenia ciasta ze wszystkim, co jest w kredensie. To nie jest dobry sposób na pieczenie ani przewidywanie! Zróżnicowanie danych można częściowo wyjaśnić faktem, że nie wszystkie predyktory szeregów czasowych miały taką samą okresowość. Docelowe szeregi czasowe obejmowały wszystkie wartości dzienne, podczas gdy predyktory były dziennymi, tygodniowymi, kwartalnymi i miesięcznymi. Było to trudne do rozwiązania, biorąc pod uwagę, że przypisywanie zer prawdopodobnie nie przyniesie dobrych wyników. Z tego konkretnego powodu zdecydowałem się użyć sieci neuronowych do oceny tygodniowych udziałów zmiennych. Stosując to podejście, mogłem kondycjonować na tydzień, bez znacznego zwiększania wymiarowości. W przypadku innych predyktorów użyłem różnych technik, w tym projekcji i korelacji, aby utworzyć orły lub ogony predyktorów. Moje podejście z powodzeniem ograniczyło liczbę zmiennych, realizując cel klienta, jakim jest uczynienie problemu możliwym do rozwiązania. W rezultacie ciasto wyszło dobrze.

Walidacja modelu

Powtarzanie tego, co właśnie usłyszałeś, nie oznacza, że się czegoś nauczyłeś. Walidacja modelu ma zasadnicze znaczenie dla konstrukcji każdego modelu. To odpowiada na pytanie „Jak dobrze moja hipoteza pasuje do obserwowanych danych?" Jeśli nie mamy wystarczającej ilości danych, nasze

modele nie mogą połączyć kropek. Z drugiej strony, biorąc pod uwagę zbyt wiele danych, model nie może myśleć nieszablonowo. Model uczy się szczegółowych informacji o danych uczących, które nie są generalizowane na populację. To jest problem nadmiernego dopasowania modelu. Istnieje wiele technik zwalczania nadmiernego dopasowania modelu. Najprostszą metodą jest podzielenie zbioru danych na zbiory uczące, testujące i walidacyjne. Dane szkoleniowe są używane do konstruowania modelu. Model skonstruowany na podstawie danych uczących jest następnie oceniany na podstawie danych testowych. Wydajność modelu względem zestawu testowego jest wykorzystywana do dalszego zmniejszania błędu modelu. To pośrednio obejmuje dane testowe w konstrukcji modelu, pomagając zredukować nadmierne dopasowanie modelu. Na koniec model jest oceniany na podstawie danych walidacyjnych, aby ocenić, jak dobrze model uogólnia. Kilka metod, w których dane są podzielone na zestawy uczące i testowe, obejmują: k-krotną weryfikację krzyżową, walidację krzyżową typu Leave-One-Out, metody ładowania początkowego i metody ponownego próbkowania. Walidacja krzyżowa typu Leave-One-Out może być użyta do uzyskania poczucia idealnej wydajności modelu w zbiorze uczącym. Z danych wybierana jest próbka, która ma działać jako próbka testowa, a model jest szkolony na podstawie pozostałych danych. Błąd próbki testowej jest obliczany i zapisywany, a próbka jest zwracana do zestawu danych. Następnie wybiera się inną próbkę i proces jest powtarzany. Trwa to do momentu wykorzystania wszystkich próbek w zestawie testowym. Średni błąd w przykładach testowych jest miarą błędu modelu. Istnieją inne podejścia do testowania, jak dobrze Twoja hipoteza odzwierciedla dane. Metody statystyczne, takie jak obliczanie współczynnika determinacji, powszechnie nazywanego wartością R-kwadrat, są używane do określenia, jak dużą zmienność danych wyjaśnia twój model. Zauważ, że wraz ze wzrostem wymiarowości przestrzeni cech rośnie również wartość R-kwadrat. Skorygowana wartość R-kwadrat kompensuje to zjawisko poprzez uwzględnienie kary za złożoność modelu. Podczas testowania istotności regresji jako całości, test F porównuje wyjaśnioną wariancję z niewyjaśnioną wariancją. Wynik regresji z wysoką statystyką F i skorygowanym R-kwadrat powyżej 0,7 jest prawie na pewno istotny.

Analiza zachowań konsumentów na podstawie wieloterabajtowego zestawu danych

Wyzwanie analityczne

Po przechowywaniu ponad 10 lat transakcji detalicznych w naturalnej przestrzeni zdrowia, klient detaliczny był zainteresowany wykorzystaniem zaawansowanych technik uczenia maszynowego do wydobywania danych w celu uzyskania cennych informacji. Klient chciał opracować strukturę bazy danych do długoterminowego wdrażania analiz łańcucha dostaw w handlu detalicznym i wybrać odpowiednie algorytmy potrzebne do uzyskania wglądu w interakcje między dostawcami, detalistami i konsumentami. Szczególnie interesujące było również określenie rzeczywistej wartości zastosowania analizy dużych zbiorów danych w kompleksowym łańcuchu dostaw detalicznych.

Podejście

Dane klienta obejmowały 3 TB opisów produktów, informacje o lojalności klientów oraz transakcje B2B i B2C dla tysięcy sklepów ze zdrową żywnością w Ameryce Północnej. Ponieważ dane były przechowywane ad hoc, pierwszym krokiem było stworzenie praktycznej struktury bazy danych, która umożliwi analizę. Wybrano środowisko w chmurze, aby szybko wdrożyć analizy w różnych, a czasem zbędnych zbiorach danych klienta. Kiedy stworzyli odpowiednią architekturę analityczną, przeszli do identyfikacji odpowiednich technik uczenia maszynowego, które dodałyby wartość w trzech kluczowych obszarach tematycznych: chłonności produktu, analizie lojalnych programów i analizie koszyka rynkowego. Zespół użył metody bayesowskiej do przyjmowania produktów Belief Networks (BBN) w celu opracowania modeli probabilistycznych do przewidywania sukcesu, niepowodzenia i długowieczności nowego lub obecnego produktu. Połączyli dane transakcji z danymi atrybutów

zarówno produktów, które odniosły sukces, jak i tych, które zakończyły się niepowodzeniem, aby przekształcić dane w użyteczną formę. Po utworzeniu tego pliku danych użyli go do szkolenia BBN i stworzenia modelu predykcyjnego dla przyszłych produktów. W celu analizy programów lojalnościowych połączyli dane transakcyjne i dane o atrybutach klientów, które obejmowały informacje o lokalizacji i trendy zakupowe. Użyli grupowania k-średnich do segmentacji klientów na podstawie ich zachowania w czasie. To pozwoliło skupić i scharakteryzować grupy klientów, które wystawiały podobne wzorce lojalnościowe. Do analizy koszyka rynkowego zastosowali Latent Dirichlet Allocation (LDA), technikę przetwarzania języka naturalnego, aby stworzyć spójną kategoryzację produktów. Kategoryzacja produktów klienta była ad hoc, wprowadzana przez poszczególnych dostawców i detalistów. W rezultacie był niespójny i często zawierał błędy typograficzne lub braki wartości. LDA umożliwiło zespołowi wykorzystanie istniejącego tekstu do pozyskania nowego, spójnego klienta , kategorie do analizy koszyka rynkowego. Po połączeniu nowych danych kategoryzacji produktów z danymi transakcji, skorzystali z uczenia się reguł asocjacyjnych, aby zidentyfikować zestawy kategorii produktów, które klienci zwykle kupowali razem w poszczególnych punktach sprzedaży detalicznej.

Wpływ

Zespół przedstawił kluczowe ustalenia i zalecenia, aby opisać, w jaki sposób można zoperacjonalizować techniki uczenia maszynowego, aby zapewnić sprzedawcom detalicznym raportowanie w czasie rzeczywistym. Klient otrzymał sugestie dotyczące ulepszenia nomenklatury produktów, promocji produktów oraz pełnej widoczności produktu i cyklu życia procesu. Jako przykład wykorzystali analizę koszyka rynkowego do stworzenia rekomendacji produktów dla poszczególnych punktów sprzedaży. Nasze rekomendacje mogą potencjalnie zwiększyć sprzedaż w niektórych kategoriach produktów nawet o 50% w całej sieci detalicznej. Wraz z oszczędnością czasu uzyskaną dzięki automatycznemu przetwarzaniu danych (np. 300-krotny wzrost szybkości kategoryzacji produktów) te spostrzeżenia pokazały wyraźną wartość analizy dużych zbiorów danych dla organizacji klienta.

Wgląd strategiczny w terabajtach danych pasażerów

Wyzwanie analityczne

Klient komercyjnych linii lotniczych stanął w obliczu rosnącej konkurencji rynkowej i wyzwań związanych z rentownością. Chcieli stawić czoła tym wyzwaniom, szybko wdrażając zaawansowane, zglobalizowane narzędzia analityczne w swojej prywatnej elektronicznej hurtowni danych. W przeszłości klient samodzielnie analizował mniejsze zbiory danych. Ponieważ mniejsze zbiory danych są filtrowanymi lub rozcieńczonymi podzbiorami pełnych danych, linia lotnicza nie była w stanie wydobyć całościowego zrozumienia, którego szukała. Booz Allen był zaangażowany w stworzenie możliwości analizowania setek gigabajtów danych klientów. Ostatecznym celem było uzyskanie wglądu w działania linii lotniczych, decyzje inwestycyjne i preferencje konsumentów, które mogły nie wynikać z analizy podzbiorów danych. W szczególności linia lotnicza chciała być w stanie zrozumieć takie kwestie, jak: jak radzą sobie na różnych rynkach par miast w porównaniu z konkurentami; jak zmieniają się zachowania rezerwacyjne w zależności od charakterystyki pasażera i lotu; i jak czasy połączeń wpływają na popyt.

Rozwiązanie

Ze względu na problemy związane z prywatnością danych zespół utworzył środowisko chmurowe w elektronicznej hurtowni danych klienta. Wykorzystując to środowisko analityczne, analiza przebiegała zgodnie z podejściem, które koncentrowało się na trzech priorytetach klienta: wynikach rynkowych, zachowaniach rezerwacyjnych i wyborze pasażerów. Przeprowadziliśmy analizę probabilistyczną przy

użyciu technik uczenia maszynowego, w szczególności Bayesian Belief Networks (BBN). Połączyli rezerwacje pasażerów i inne dane, aby utworzyć plik szkoleniowy BBN. Zespół opracował i zweryfikował kompleksowe modele BBN, które reprezentują istotne zachowania klientów i czynniki rynkowe, które wpływają na preferencje pasażerów w odniesieniu do wyboru lotów według czasu przesiadek. Wreszcie, nasz zespół opracował niestandardowe wizualizacje Big Data, aby przekazać wyniki zarówno odbiorcom technicznym, jak i nietechnicznym.

Wpływ

Wykazano zdolność do szybkiego wdrażania narzędzi analitycznych do dużych zbiorów danych i uczenia maszynowego na ogromnych zbiorach danych znajdujących się w środowisku chmury prywatnej komercyjnych linii lotniczych. Wyniki obejmowały spostrzeżenia, które wydawały się sprzeczne z intuicją, ale mimo to mogły poprawić wyniki finansowe. Przykładem takiego ustalenia był fakt, że w pewnych okolicznościach pasażerowie są skłonni zapłacić dodatkową opłatę za rezerwację tras ze zmienionymi czasami przesiadek. Przekłada się to na potencjalny wzrost przychodów o wiele milionów dolarów. Te spostrzeżenia, często na poziomie klientów imiennych, można wykorzystać natychmiast, aby poprawić wyniki finansowe.

Oszczędności dzięki lepszej produkcji

Wyzwanie analityczne

Firma produkcyjna zaangażowała Booz Allena do zbadania danych związanych z produkcją związków chemicznych. Te procesy są dość złożone. Obejmują długi łańcuch wzajemnie powiązanych zdarzeń, co ostatecznie prowadzi do dużej zmienności w produkcji produktu. To sprawia, że produkcja jest bardzo droga. Zrozumienie procesu produkcyjnego nie jest łatwe - czujniki zbierają tysiące zmiennych szeregów czasowych i tysiące pomiarów punktowo-czasowych, uzyskując terabajty danych. Istniała ogromna szansa, gdyby klient mógł zrozumieć te dane. Zmniejszenie rozbieżności i poprawa wydajności produktu nawet o niewielką ilość może spowodować znaczne oszczędności kosztów.

Rozwiązanie

Ze względu na rozmiar i złożoność danych procesowych wcześniejsze analizy, które koncentrowały się na danych z tylko jednego podprocesu, zakończyły się ograniczonym sukcesem. Zespół Data Science przyjął inne podejście: przeanalizował wszystkie dane ze wszystkich podprocesów w celu zidentyfikowania czynników wpływających na zmienność. Gdy zrozumiemy te czynniki, będziemy mogli opracować zalecenia, jak je kontrolować, aby zwiększyć plony. Inżynierowie procesu klienta zawsze chcieli stosować to podejście, ale brakowało im narzędzi do przeprowadzenia analizy. Podzieliono problem na serię mniejszych problemów. Po pierwsze, konieczne było określenie, które parametry szeregów czasowych prawdopodobnie wpływały na wydajność produktu. Zaangażowano ekspertów domenowych klienta, aby zidentyfikować ich hipotezy dotyczące tego procesu. Po rozpoznaniu zestawu hipotez zidentyfikowano czujniki, które gromadziły odpowiednie dane. Rozpoczęto wstępne przetwarzanie danych, które obejmowało filtrowanie złych wartości i identyfikację wzorców w szeregach czasowych. Następnie musiano podzielić strumienie danych na poszczególne serie produkcyjne. Uszkodzono czujnik, który przechowuje informacje wysokiego poziomu wskazujące, kiedy rozpoczął się proces produkcji. Ten czujnik zapewnił dokładnie to, czego potrzebowano, ale szybko zauważono, że brakuje połowy oczekiwanych danych. Po dokładniejszej analizie danych zdano sobie sprawę, że czujnik był używany dopiero w ostatnich latach. Musisno cofnąć się o krok i ponownie ocenić nasz plan. Po rozmowach z ekspertami w dziedzinie zidentyfikowaliśmy inny czujnik, który podał nam surowe wartości bezpośrednio z procesu produkcyjnego. Surowe wartości obejmowały znacznik wskazujący początek cyklu produkcyjnego. Czujnik był aktywny dla każdej serii produkcyjnej i mógł być

niezawodnie używany do segmentowania strumieni danych na serie produkcyjne. Następnie musieliśmy określić, które parametry szeregów czasowych wpływają na wydajność produktu. Korzystając z oczyszczonych i przetworzonych danych oraz nieparametrycznej techniki korelacji, porównaliśmy każdy szereg czasowy w przebiegu produkcyjnym z wszystkimi innymi szeregami czasowymi w tym samym przebiegu. Biorąc pod uwagę podobieństwa parami, oszacowaliśmy korelację podobieństw z wydajnością produktu końcowego. Następnie użyliśmy korelacji jako danych wejściowych do algorytmu grupowania, aby znaleźć skupiska parametrów szeregów czasowych, które były ze sobą skorelowane pod względem wydajności produktu, a nie samych szeregów czasowych. Ta analiza danych była na skalę wcześniej niemożliwą - miliony porównań dla całych serii produkcyjnych. Inżynierowie mogli po raz pierwszy przejrzeć wszystkie dane i zobaczyć wpływ określonych parametrów na różne partie i czujniki. Oprócz określenia kluczowych parametrów inżynierowie musieli wiedzieć, jak kontrolować parametry, aby zwiększyć wydajność produktu. Dyskusje z ekspertami dziedzinowymi pozwoliły dowiedzieć się, które parametry szeregów czasowych można łatwo kontrolować. Ograniczało to potencjalne parametry tylko do tych, na które inżynierowie procesu mogli mieć wpływ. Wyodrębniono cechy z pozostałych sygnałów szeregów czasowych i wprowadziliśmy je do naszych modeli, aby przewidzieć wydajność. Modele określały ilościowo korelację między wzorcem wartości parametrów a wydajnością, dostarczając informacji na temat zwiększania wydajności produktu.

Wpływ

Po zidentyfikowaniu kontroli i kwantyfikacji pożądanych wzorców, zapewniliśmy inżynierom zestaw działań kontrolnych procesu w celu poprawy wydajności produktu. Surowe dane z czujników, które pochodziły bezpośrednio z procesu produkcyjnego, stanowiły podstawę naszej analizy i zaleceń, dając klientowi pewność co do zastosowanego podejścia. Zmniejszenie zmienności wydajności produktu umożliwi klientowi wytworzenie lepszego produktu przy niższym ryzyku przy niższych kosztach.

Wyższe zyski dzięki analizie predykcyjnej

Wyzwanie analityczne

Duży dom inwestycyjny chciał zbadać, czy zastosowanie technik Data Science może przynieść większe zwroty z inwestycji. W szczególności firma chciała przewidzieć przyszłe zmiany wartości towarów na podstawie wskaźników akcji na koniec dnia i z poprzedniego dnia. Klient miał nadzieję, że prognozy mogą zostać wykorzystane do optymalizacji jego działań handlowych. Przekładając podejście na cały portfel, mogli radykalnie poprawić krzywą dochodowości swoich inwestorów. Kilka wyzwań było od razu widocznych. Objętość danych była bardzo duża i obejmowała informacje z dziesiątek tysięcy akcji, towarów i opcji z większości głównych rynków światowych w wielu przedziałach czasowych. Jeszcze większym wyzwaniem była potrzeba rekomendowania akcji predykcyjnej (idź krótko, idź długo, zostań, zwiększ rozmiar pozycji lub zaangażuj się w grę konkretną opcją) z bardzo małym opóźnieniem. Zespół musiałby opracować podejście, które uwzględni oba te ukryte ograniczenia.

Rozwiązanie

Klient wezwał Booz Allena do wykorzystania 3500 zmiennych niezależnych do przewidywania dziennych zmian cen 16 instrumentów finansowych. Klient ukrył jednak znaczenie i kontekst zmiennych niezależnych, zmuszając zespół do przeprowadzenia analizy bez informacji jakościowych. Zespół natychmiast rozpoczął wyszukiwanie lub uzupełnianie źródeł danych. Zidentyfikowali nieustrukturyzowane dane z innych firm, instytucji finansowych, rządów i mediów społecznościowych, które mogą być wykorzystane w naszej analizie. Zespół poświęcił wiele uwagi wydajności i bezpieczeństwu dostępu do baz danych, a także szybkość obliczeń. Nasz zespół wdrożył

wielopłaszczyznowe podejście, obejmujące połączenie optymalizacji sieci neuronowej i różnych głównych komponentów, technik regresji i uczenia się bez nadzoru. Udało się wywnioskować wgląd w zdarzenia egzogeniczne na małą skalę, które dostarczyły bogatszej podstawy do przewidywania lokalnych fluktuacji cen akcji. Nasz zespół był w stanie wykorzystać te prognozy, aby określić optymalną kombinację działań, które wygenerowałyby najlepszy łączny zwrot w ciągu 12 miesięcy handlu. Dokładne rozważenie reszt i umiejętne modelowanie wariancji dodało dodatkową wartość do wyniku dla tego klienta.

ROZDZIELANIE MYŚLI

Możliwości Data Science tworzą analizy danych, które poprawiają każdy aspekt naszego życia, od ratujących życie leczenia chorób, przez bezpieczeństwo narodowe, po stabilność ekonomiczną, a nawet wygodę wyboru restauracji. Mamy nadzieję, że pomogliśmy Ci naprawdę zrozumieć potencjał Twoich danych i dowiedzieć się, jak stać się niezwykłymi myślicielami, zadając właściwe pytania dotyczące Twoich danych. Mamy nadzieję, że pomogliśmy w rozwoju nauki i sztuki Data Science. Co najważniejsze, mamy nadzieję, że odchodzisz z nowo odkrytą pasją i podekscytowaniem dla nauki o danych