Dane: Paliwo dla AI

Pinterest to jeden z najgorętszych start-upów w Dolinie Krzemowej, umożliwiający użytkownikom przypinanie ulubionych elementów w celu tworzenia angażujących tablic. Witryna ma 250 milionów MAU (aktywnych użytkowników miesięcznie) i odnotowała przychody w wysokości 756 milionów dolarów w 2018 roku. Popularną aktywnością Pinteresta jest planowanie ślubów. Przyszła panna młoda będzie miała szpilki do sukni, lokale, miejsc na miesiąc miodowy, ciasta, zaproszenia i tak dalej. Oznacza to również, że Pinterest ma tę zaletę, że gromadzi ogromne ilości cennych danych. Część tego pomaga zapewnić ukierunkowane reklamy. Istnieją jednak również możliwości prowadzenia kampanii e-mailowych. W jednym przypadku Pinterest wysłał taki, który mówił:

Żenisz się! A ponieważ uwielbiamy planowanie ślubów - zwłaszcza wszystkich uroczych papeterii - zapraszamy do przeglądania naszych najlepszych plansz, których kuratorami są graficy, fotografowie i inne przyszłe panny młode, wszystkie Pinnerki z bystrym okiem i myślącym o małżeństwie.

Problem: wielu odbiorców wiadomości e-mail było już w związku małżeńskim lub nie spodziewało się, że wkrótce się pobierze.

Pinterest działał szybko i przeprosił:

Co tydzień wysyłamy e-mailem kolekcje pinów i tablic dla poszczególnych kategorii do pinnerów, którzy mamy nadzieję, że będą nimi zainteresowani. Niestety, jeden z tych ostatnich e-maili sugerował, że pinners faktycznie brali ślub, a nie tylko potencjalnie zainteresowani treściami związanymi ze ślubem. Przykro nam, że wypadliśmy jak apodyktyczna matka, która zawsze pyta, kiedy znajdziesz miłego chłopca lub dziewczynę.

To ważna lekcja. Nawet niektóre z najbardziej zaawansowanych technologicznie firm to rozwalają. Na przykład w niektórych przypadkach dane mogą być bezbłędne, ale rezultatem może być porażka. Rozważ sprawę z Target. Firma wykorzystała swoje ogromne dane, aby wysyłać spersonalizowane oferty do przyszłych matek. Zostało to oparte na tych klientach, którzy dokonali określonych rodzajów zakupów, takich jak bezzapachowe balsamy. System Target utworzyłby wynik ciąży, który dostarczyłby nawet szacunkowych terminów porodu. Cóż, ojciec jednego z klientów zobaczył e-mail i był wściekły, mówiąc, że jego córka nie jest w ciąży. Ale była – i tak, ukrywała ten fakt przed ojcem. Nie ma wątpliwości, że dane są niezwykle potężne i krytyczne dla sztucznej inteligencji. Ale musisz być rozważny i rozumieć ryzyko.

Podstawy danych

Dobrze jest rozumieć żargon danych. Przede wszystkim bit (skrót od „cyfra binarna”) to najmniejsza forma danych w komputerze. Pomyśl o tym jak o atomie. Bit może mieć wartość 0 lub 1, co jest binarne. Jest również powszechnie używany do pomiaru ilości przesyłanych danych (powiedzmy w sieci lub Internecie). Z drugiej strony bajt służy głównie do przechowywania. Oczywiście liczba bajtów może bardzo szybko rosnąć. Zobaczmy :

Jednostka: Wartość: Przypadek użycia

Megabajt: 1000 kilobajtów: Mała książka

Gigabajt: 1000 megabajtów: około 230 utworów

Terabajt: 1000 gigabajtów: 500 godzin filmów

Petabajt: 1000 terabajtów: pięć lat systemu obserwacji Ziemi (EOS)

Eksabajt: 1000 petabajtów: Cała Biblioteka Kongresu 3000 razy więcej

Zettabyte : 1000 eksabajtów : 36 000 lat wideo HD-TV

Yottabajty: 1000 zettabajtów: Wymagałoby to centrum danych wielkości Delaware i Rhode Island łącznie

Dane mogą również pochodzić z wielu różnych źródeł. Oto tylko próbka:

• Internetowe/społecznościowe (Facebook, Twitter, Instagram, YouTube)

• Dane biometryczne (monitorowanie kondycji, testy genetyczne)

• Systemy punktów sprzedaży (ze sklepów stacjonarnych i witryn e-commerce)

• Internet rzeczy lub IoT (tagi identyfikacyjne i urządzenia inteligentne)

• Systemy chmurowe (aplikacje biznesowe, takie jak Salesforce.com)

• Korporacyjne bazy danych i arkusze kalkulacyjne

Rodzaje danych

Istnieją cztery sposoby organizowania danych. Po pierwsze, istnieją dane strukturalne, które zwykle są przechowywane w relacyjnej bazie danych lub arkuszu kalkulacyjnym. Oto kilka przykładów:

• Informacje finansowe

• Numery ubezpieczenia społecznego

• Adresy

• Informacje o produkcie

• Dane punktów sprzedaży

• Numery telefoniczne

W większości przypadków łatwiej jest pracować z danymi strukturalnymi. Dane te często pochodzą z systemów CRM (zarządzanie relacjami z klientami) i ERP (planowanie zasobów przedsiębiorstwa) – i zwykle mają mniejsze ilości. Wydaje się również, że jest to prostsze, powiedzmy pod względem analizy. Istnieją różne programy BI (Business Intelligence), które mogą pomóc uzyskać wgląd w dane strukturalne. Jednak tego typu dane stanowią około 20% projektu AI. Większość będzie pochodzić z danych nieustrukturyzowanych, czyli informacji, które nie mają wstępnie zdefiniowanego formatowania. Musisz to zrobić sam, co może być żmudne i czasochłonne. Istnieją jednak narzędzia, takie jak bazy danych nowej generacji, takie jak te oparte na NoSQL, które mogą pomóc w tym procesie. Systemy AI są również skuteczne pod względem zarządzania danymi i ich strukturyzacji, ponieważ algorytmy potrafią rozpoznawać wzorce. Oto przykłady nieuporządkowanych danych:

• Obrazy

• Filmy

• Pliki audio

• Pliki tekstowe

• Informacje z sieci społecznościowych, takie jak tweety i posty

• Zdjęcia satelitarne

Teraz są pewne dane, które są hybrydą źródeł ustrukturyzowanych i nieustrukturyzowanych, zwanych danymi częściowo ustrukturyzowanymi. Informacje mają kilka wewnętrznych znaczników, które pomagają w kategoryzacji. Przykładami danych częściowo ustrukturyzowanych są XML (Extensible Markup Language), który opiera się na różnych regułach identyfikacji elementów dokumentu, oraz JSON (JavaScript Object Notation), który jest sposobem przesyłania informacji w sieci poprzez API (Programowanie aplikacji interfejsy). Ale dane częściowo ustrukturyzowane stanowią tylko około 5% do 10% wszystkich danych. Wreszcie istnieją dane szeregów czasowych, które mogą dotyczyć zarówno danych ustrukturyzowanych, nieustrukturyzowanych, jak i częściowo ustrukturyzowanych. Ten rodzaj informacji służy do interakcji, powiedzmy do śledzenia „podróży klienta". Będzie to zbieranie informacji, gdy użytkownik wchodzi na stronę internetową, korzysta z aplikacji, a nawet wchodzi do sklepu. Jednak tego rodzaju dane są często chaotyczne i trudne do zrozumienia. Częściowo wynika to ze zrozumienia intencji użytkowników, które mogą się znacznie różnić. Istnieją również ogromne ilości danych interaktywnych, które mogą obejmować biliony punktów danych. Aha, a mierniki sukcesu mogą nie być jasne. Dlaczego użytkownik robi coś w witrynie? Ale sztuczna inteligencja prawdopodobnie będzie miała kluczowe znaczenie dla takich problemów. Chociaż w większości analiza danych szeregów czasowych jest wciąż na wczesnym etapie.

Big Data

Wraz z wszechobecnym dostępem do Internetu, urządzeń mobilnych i urządzeń do noszenia, nastąpiło uwolnienie potoku danych. Co sekundę Google przetwarza ponad 40 000 wyszukiwań, czyli 3,5 miliarda dziennie. Z minuty na minutę użytkownicy Snapchata udostępniają 527 760 zdjęć, a użytkownicy YouTube oglądają ponad 4,1 miliona filmów. Są też staromodne systemy, takie jak e-maile, które nadal odnotowują znaczny wzrost. Co minutę wysyłanych jest 156 milionów wiadomości. Ale jest jeszcze coś do rozważenia: firmy i maszyny również generują ogromne ilości danych. Według badań Statista do 2020 r. liczba czujników osiągnie 12,86 miliarda. W świetle tego wydaje się, że warto założyć, że ilość danych będzie rosła w szybkim tempie. W raporcie International Data Corporation (IDC) zatytułowanym „Data Age 2025" oczekuje się, że do 2025 roku ilość utworzonych danych osiągnie oszałamiającą liczbę 163 zetabajtów. To około dziesięciokrotnie więcej niż w 2017 roku. Aby sobie z tym wszystkim poradzić, powstała kategoria technologii zwana Big Data. W ten sposób Oracle wyjaśnia znaczenie tego trendu: Dzisiaj duże zbiory danych stały się kapitałem. Pomyśl o największych światowych firmach technologicznych. Duża część oferowanej przez nich wartości pochodzi z ich danych, które stale analizują, aby zwiększyć wydajność i opracować nowe produkty. A więc tak, Big Data pozostanie krytyczną częścią wielu projektów AI. Czym właściwie jest Big Data? Jaka jest dobra definicja? Właściwie nie ma ani jednego, chociaż jest wiele firm, które skupiają się na tym rynku! Ale Big Data ma następujące cechy, zwane trzema V (analityk Gartnera, Doug Laney, wymyślił tę strukturę w 2001 r.): Volume [wolumen] , Variety [różnorodność] i Veloity [prędkość].

Wolumen

Taka jest skala danych, które często są nieustrukturyzowane. Nie ma twardej i szybkiej reguły na progu, ale zwykle jest to dziesiątki terabajtów. Wolumen to często duże wyzwanie, jeśli chodzi o Big Data. Jednak przetwarzanie w chmurze i bazy danych nowej generacji bardzo pomogły pod względem pojemności i niższych kosztów.

Różnorodność

Opisuje to różnorodność danych, powiedzmy kombinację danych ustrukturyzowanych, częściowo ustrukturyzowanych i nieustrukturyzowanych (wyjaśnionych powyżej). Pokazuje również różne źródła

danych i zastosowań. Bez wątpienia wysoki wzrost liczby nieustrukturyzowanych danych był kluczem do różnorodności Big Data. Zarządzanie tym może szybko stać się poważnym wyzwaniem. Jednak uczenie maszynowe jest często czymś, co może pomóc usprawnić ten proces.

Prędkość

Pokazuje szybkość, z jaką tworzone są dane. Jak widać wcześniej w tym rozdziale, serwisy takie jak YouTube i Snapchat charakteryzują się ekstremalną szybkością (często nazywa się to „remizą" danych). Wymaga to dużych inwestycji w technologie i centra danych nowej generacji. Dane są również często przetwarzane w pamięci, a nie w systemach dyskowych. Z powodu tych problemów prędkość jest często uważana za najtrudniejszą, jeśli chodzi o trzy V. Spójrzmy prawdzie w oczy, w dzisiejszym cyfrowym świecie ludzie chcą swoich danych tak szybko, jak to możliwe. Jeśli będzie zbyt wolny, ludzie będą sfrustrowani i odejdą gdzieś indziej. Jednak z biegiem lat, wraz z ewolucją Big Data, dodano więcej V. Obecnie jest ich kilkanaście. Ale oto niektóre z typowych:

• Veracity [Prawdziwość]: dotyczy danych, które są uważane za dokładne. W tym rozdziale przyjrzymy się niektórym technikom oceny prawdziwości.

• Value [Wartość]: pokazuje przydatność danych. Często chodzi o posiadanie zaufanego źródła.

• Vaaraibility [Zmienność]: Oznacza to, że dane zwykle zmieniają się w czasie. Tak jest na przykład w przypadku treści w mediach społecznościowych, które mogą się zmieniać w oparciu o ogólne nastroje dotyczące nowych wydarzeń i najświeższych wiadomości.

• Visualisation [Wizualizacja]: wykorzystuje wizualizacje podobne do wykresów, aby lepiej zrozumieć dane.

Jak widać, zarządzanie Big Data składa się z wielu ruchomych części, co prowadzi do złożoności. To pomaga wyjaśnić, dlaczego wiele firm nadal wykorzystuje tylko niewielki ułamek swoich danych.

Bazy danych i inne narzędzia

Istnieje mnóstwo narzędzi, które pomagają w przetwarzaniu danych. Sednem tego jest baza danych. Nie powinno dziwić, że na przestrzeni dziesięcioleci nastąpiła ewolucja tej kluczowej technologii. Ale nawet starsze technologie, takie jak relacyjne bazy danych, są nadal bardzo często używane. Jeśli chodzi o dane o znaczeniu krytycznym, firmy niechętnie wprowadzają zmiany, nawet jeśli niosą ze sobą wyraźne korzyści. Aby zrozumieć ten rynek, cofnijmy się do 1970 r., kiedy informatyk IBM Edgar Codd opublikował „A Relational Model of Data for Large Shared Data Banks". To było przełomowe, ponieważ wprowadziło strukturę relacyjnych baz danych. Do tego momentu bazy danych były dość złożone i miały sztywną strukturę jako hierarchie. To sprawiło, że wyszukiwanie i znajdowanie relacji w danych było czasochłonne. Jeśli chodzi o podejście do relacyjnych baz danych Codda, zostało ono zbudowane dla bardziej nowoczesnych maszyn. Język skryptowy SQL był łatwy w użyciu, pozwalając na operacje CRUD (Create, Read, Update i Delete). Tabele miały również połączenia z kluczami podstawowymi i obcymi, które tworzyły ważne połączenia, takie jak:

• Jeden do jednego: Jeden wiersz w tabeli jest połączony tylko z jednym wierszem w innej tabeli. Przykład: Numer prawa jazdy, który jest unikalny, jest powiązany z jednym pracownikiem.

• Jeden-do-wielu: W tym miejscu jeden wiersz w tabeli jest połączony z innymi tabelami. Przykład: Klient ma wiele zamówień zakupu.

• Wiele do wielu: Wiersze z jednej tabeli są skojarzone z wierszami innej. Przykład: Różne raporty mają różnych autorów.

Przy tego typu strukturach relacyjna baza danych może usprawnić proces tworzenia zaawansowanych raportów. To było naprawdę rewolucyjne. Jednak pomimo zalet, IBM nie był zainteresowany technologią i nadal skupiał się na własnych systemach. Firma uznała, że relacyjne bazy danych są zbyt wolne i kruche dla klientów korporacyjnych. Ale był ktoś, kto miał inne zdanie w tej sprawie: Larry Ellison. Przeczytał artykuł Codda i wiedział, że to zmieni zasady gry. Aby to udowodnić, w 1977 r. współzałożył firmę Oracle, skupiając się na budowaniu relacyjnych baz danych, które szybko stałyby się ogromnym rynkiem. Dopiero w 1993 roku IBM wyszedł z własną relacyjną bazą danych DB2. Ale było za późno. W tym czasie Oracle był liderem na rynku baz danych. W latach 80. i 90. relacyjna baza danych była standardem dla systemów mainframe i klient-serwer. Ale kiedy Big Data stało się czynnikiem, technologia miała poważne wady, takie jak:

• Rozciąganie się danych: Z biegiem czasu różne bazy danych rozprzestrzeniłyby się w całej organizacji. W rezultacie centralizacja danych stała się trudniejsza.

• Nowe środowiska: Technologia relacyjnych baz danych nie została stworzona do przetwarzania w chmurze, danych o dużej prędkości lub danych nieustrukturyzowanych.

• Wysokie koszty: relacyjne bazy danych mogą być drogie. Oznacza to, że korzystanie z technologii w projektach AI może być niedopuszczalne.

• Wyzwania programistyczne: Współczesne tworzenie oprogramowania w dużej mierze opiera się na iteracji. Ale relacyjne bazy danych okazały się trudne w tym procesie.

Pod koniec lat 90. opracowano projekty open source, które miały pomóc w tworzeniu systemów baz danych nowej generacji. Być może najbardziej krytyczny pochodził od Douga Cuttinga, który opracował Lucene, który służył do wyszukiwania tekstu. Technologia została oparta na wyrafinowanym systemie indeksów, który umożliwiał działanie z niskimi opóźnieniami. Lucene stał się natychmiastowym hitem i zaczął ewoluować, tak jak Apache Nutch, który sprawnie przeszukiwał sieć i przechowywał dane w indeksie. Ale był duży problem: aby przeszukiwać sieć, potrzebna była infrastruktura zdolna do hiperskalowania. Tak więc pod koniec 2003 roku Cutting rozpoczął prace nad nowym rodzajem platformy infrastrukturalnej, która może rozwiązać ten problem. Pomysł wpadł na pomysł z artykułu opublikowanego przez Google, który opisywał jego ogromny system plików. Rok później Cutting zbudował swoją nową platformę, która pozwoliła na wyrafinowane przechowywanie bez komplikacji. U podstaw tego był MapReduce, który umożliwiał przetwarzanie na wielu serwerach. Wyniki byłyby następnie scalane, umożliwiając tworzenie znaczących raportów. Ostatecznie system Cutting przekształcił się w platformę o nazwie Hadoop - i byłby niezbędny do zarządzania Big Data, na przykład do tworzenia wyrafinowanych hurtowni danych. Początkowo Yahoo! używał go, a następnie szybko się rozprzestrzenił, ponieważ firmy takie jak Facebook i Twitter przyjęły tę technologię. Firmy te mogły teraz uzyskać pełny widok swoich danych, a nie tylko ich podzbiorów. Oznaczało to, że mogą być bardziej efektywne eksperymenty z danymi. Jednak jako projekt open source Hadoop wciąż nie dysponował zaawansowanymi systemami dla klientów korporacyjnych. Aby sobie z tym poradzić, startup Hortonworks zbudował nowe technologie, takie jak YARN, na platformie Hadoop. Posiadał funkcje, takie jak przetwarzanie analityczne w pamięci, przetwarzanie danych online i interaktywne przetwarzanie SQL. Takie możliwości wspierały przyjęcie Hadoopa w wielu korporacjach. Ale oczywiście pojawiły się inne projekty hurtowni danych typu open source. Te dobrze znane, jak Storm i Spark, skupiały się na strumieniowaniu danych. Z drugiej strony Hadoop został zoptymalizowany pod kątem przetwarzania wsadowego. Oprócz hurtowni danych pojawiła się również innowacja tradycyjnego biznesu bazodanowego. Często były one znane jako systemy NoSQL. Weź MongoDB. Rozpoczęło się jako projekt open source i przekształciło się w odnoszącą sukcesy firmę, która weszła na giełdę w październiku 2017 r. Baza danych MongoDB, która ma ponad 40 milionów pobrań, została

stworzona do obsługi środowisk chmurowych, lokalnych i hybrydowych. Istnieje również duża elastyczność strukturyzowania danych, która opiera się na modelu dokumentu. MongoDB może nawet zarządzać ustrukturyzowanymi i nieustrukturyzowanymi danymi na dużą skalę petabajtów. Chociaż startupy były źródłem innowacji w systemach baz danych i pamięci masowej, ważne jest, aby pamiętać, że operatorzy mega tech również byli krytyczni. Z drugiej strony firmy takie jak Amazon.com i Google musiały znaleźć sposoby radzenia sobie z ogromną ilością danych ze względu na potrzebę zarządzania swoimi ogromnymi platformami. Jedną z innowacji jest jezioro danych, które umożliwia bezproblemowe przechowywanie ustrukturyzowanych i nieustrukturyzowanych danych. Pamiętaj, że nie ma potrzeby ponownego formatowania danych. Jezioro danych poradzi sobie z tym i umożliwi szybkie wykonywanie funkcji AI. Według badania przeprowadzonego przez Aberdeen firmy korzystające z tej technologii osiągają średnio 9% wzrost organiczny w porównaniu z tymi, które tego nie robią. Teraz nie oznacza to, że musisz pozbyć się hurtowni danych. Oba służą raczej określonym funkcjom i przypadkom użycia. Hurtownia danych jest ogólnie dobra w przypadku danych strukturalnych, podczas gdy jezioro danych jest lepsze w przypadku zróżnicowanych środowisk. Co więcej, prawdopodobnie duża część danych nigdy nie zostanie wykorzystana. W większości dostępnych jest mnóstwo narzędzi. I oczekuj, że w miarę jak środowiska danych stają się coraz bardziej złożone, rozwinie się więcej. Ale to nie znaczy, że powinieneś wybrać najnowszą technologię. Ponownie, nawet starsze relacyjne bazy danych mogą być całkiem skuteczne w przypadku projektów AI. Kluczem jest zrozumienie zalet i wad każdego z nich, a następnie opracowanie jasnej strategii.

Przetwarzanie danych

Ilość pieniędzy wydanych na dane jest ogromna. Według IDC, przewiduje się, że wydatki na Big Data i rozwiązania analityczne wzrosną z 166 miliardów dolarów w 2018 r. do 260 miliardów dolarów do 2022 r.11 Stanowi to 11,9% łączną stopę wzrostu rocznego. Najwięksi wydający to banki, dyskretni producenci, producenci procesów, profesjonalne firmy usługowe i rząd federalny. Stanowią one blisko połowę ogólnej kwoty. Oto, co powiedziała Jessica Goepfert z IDC — wiceprezes programu (VP) Customer Insights and Analysis:

Na wysokim poziomie organizacje zwracają się do Big Data i rozwiązań analitycznych, aby poruszać się po konwergencji swojego świata fizycznego i cyfrowego. Ta transformacja przybiera różny kształt w zależności od branży. Na przykład w bankowości i handlu detalicznym – w dwóch najszybciej rozwijających się obszarach Big Data i inwestycji w analitykę – chodzi o zarządzanie i ożywianie doświadczeń klientów. . Podczas gdy w branży produkcyjnej, firmy wymyślają się na nowo, aby zasadniczo stać się firmami high-tech, wykorzystując swoje produkty jako platformę umożliwiającą i dostarczającą usługi cyfrowe.

Ale wysoki poziom wydatków niekoniecznie przekłada się na dobre wyniki. Badanie Gartnera szacuje, że około 85% projektów Big Data jest porzucanych, zanim dotrą do etapu pilotażowego.13 Niektóre z przyczyn są następujące:

- Brak wyraźnego skupienia
- Brudne dane
- Inwestycja w niewłaściwe narzędzia informatyczne
- Problemy z gromadzeniem danych
- Brak akceptacji ze strony kluczowych interesariuszy i czempionów w organizacji.

W związku z tym bardzo ważne jest posiadanie procesu danych. Pomimo tego, że istnieje wiele podejść -często wychwalanych przez dostawców oprogramowania — jest jedno, które cieszy się powszechną akceptacją. Grupa ekspertów, programistów, konsultantów i naukowców stworzyła proces CRISP-DM pod koniec lat 90-tych. W tym rozdziale przyjrzymy się krokom od 1 do 3. Następnie w dalszej części książki omówimy pozostałe (czyli przyjrzymy się modelowaniu i ocenie w rozdziale 3 i wdrażaniu w rozdziale 8). Należy zauważyć, że kroki 1–3 mogą stanowić 80% czasu przetwarzania danych, co opiera się na doświadczeniu Atifa Kureishy, który jest globalnym wiceprezesem ds. pojawiających się praktyk w firmie Teradata.14 Wynika to z takich czynników, jak: Dane nie są dobrze zorganizowane i pochodzą z różnych źródeł (od różnych dostawców lub silosów w organizacji), nie ma wystarczającego skupienia na narzędziach automatyzacji, a wstępne planowanie było niewystarczające dla zakresu projektu. Warto również pamiętać, że proces CRISP-DM nie jest procesem ściśle liniowym. Kiedy mamy do czynienia z danymi, może być wiele iteracji. Na przykład może być wiele prób wymyślenia właściwych danych i przetestowania ich.

Krok #1-Zrozumienie biznesowe

Powinieneś mieć jasny obraz problemu biznesowego do rozwiązania. Kilka przykładów:

• W jaki sposób korekta ceny może wpłynąć na sprzedaż?

• Czy zmiana tekstu doprowadzi do poprawy konwersji reklam cyfrowych?

• Czy spadek zaangażowania oznacza wzrost rezygnacji?

Następnie musisz ustalić, jak będziesz mierzyć sukces. Czy to możliwe, że sprzedaż powinna wzrosnąć o co najmniej 1% lub że konwersje powinny wzrosnąć o 5%? Oto przypadek Prasada Vuyyuru, który jest partnerem w Enterprise Insights Practice firmy Infosys Consulting:

Identyfikacja problemu biznesowego do rozwiązania za pomocą sztucznej inteligencji i ocena, jaka wartość zostanie stworzona, ma kluczowe znaczenie dla powodzenia wszystkich projektów AI. Bez tak starannego skupienia się na wartości biznesowej projekty AI mogą nie zostać przyjęte w organizacji. Doświadczenie AB Inbev w wykorzystywaniu sztucznej inteligencji do identyfikacji silników linii pakujących, które mogą ulec awarii, jest doskonałym przykładem tego, jak sztuczna inteligencja tworzy praktyczną wartość. ABInbev zainstalował 20 bezprzewodowych czujników do pomiaru drgań na silnikach linii pakujących. Porównali dźwięki z normalnie działającymi silnikami, aby zidentyfikować anomalie, które przewidywały ostateczną awarię silników.

Niezależnie od celu ważne jest, aby proces był wolny od wszelkich uprzedzeń lub uprzedzeń. Celem jest znalezienie najlepszych wyników. Bez wątpienia w niektórych przypadkach nie będzie satysfakcjonującego wyniku. Lub w innych sytuacjach mogą być duże niespodzianki. Słynny przykład tego pochodzi z książki „Moneyball" Michaela Lewisa, która również została nakręcona w 2011 roku na film z Bradem Pittem w roli głównej. To prawdziwa historia o tym, jak Oakland A wykorzystał techniki analizy danych do rekrutacji graczy. Tradycją w baseballu było poleganie na wskaźnikach, takich jak średnie mrugnięcia. Jednak przy użyciu wyrafinowanych technik analizy danych uzyskano zaskakujące wyniki. Oakland A zdał sobie sprawę, że należy skupić się na obijaniu i procentach bazowych. Dzięki tym informacjom zespół był w stanie zrekrutować najlepszych graczy na niższych poziomach wynagrodzenia. W rezultacie musisz być otwarty i chętny do eksperymentowania. W kroku #1 powinieneś również zebrać odpowiedni zespół do projektu. Teraz, o ile nie pracujesz w firmie takiej jak Facebook czy Google, nie będziesz miał luksusu wyboru grupy doktorów z zakresu uczenia maszynowego i data science. Taki talent jest dość rzadki i drogi. Ale nie potrzebujesz też armii najwyższej klasy inżynierów do projektu AI. W rzeczywistości stosowanie modeli uczenia maszynowego

i głębokiego uczenia się staje się coraz łatwiejsze dzięki systemom open source, takim jak TensorFlow i platformom w chmurze Google, Amazon.com i Microsoft. Innymi słowy, możesz potrzebować tylko kilku osób z doświadczeniem w nauce danych. Następnie powinieneś znaleźć osoby – prawdopodobnie z Twojej organizacji – które mają odpowiednią wiedzę specjalistyczną w zakresie domeny dla projektu AI. Będą musieli przemyśleć przepływy pracy, modele i dane szkoleniowe – ze szczególnym zrozumieniem branży i wymagań klientów. Na koniec musisz ocenić potrzeby techniczne. Jaka infrastruktura i narzędzia programowe zostaną wykorzystane? Czy będzie potrzeba zwiększenia wydajności lub zakupu nowych rozwiązań?

Krok #2 - Zrozumienie danych

W tym kroku przyjrzysz się źródłom danych projektu. Weź pod uwagę, że są trzy główne, do których należą:

• Dane wewnętrzne: Dane te mogą pochodzić ze strony internetowej, sygnałów nawigacyjnych w lokalizacji sklepu, czujników IoT, aplikacji mobilnych i tak dalej. Główną zaletą tych danych jest to, że są bezpłatne i dostosowane do Twojej firmy. Ale z drugiej strony istnieje pewne ryzyko. Mogą wystąpić problemy, jeśli nie poświęcono wystarczającej uwagi formatowaniu danych lub jakie dane należy wybrać.

• Dane typu Open Source: są one zazwyczaj dostępne bezpłatnie, co z pewnością jest miłą korzyścią. Niektóre przykłady danych o otwartym kodzie źródłowym obejmują informacje rządowe i naukowe. Dostęp do danych jest często uzyskiwany za pośrednictwem interfejsu API, co sprawia, że proces jest dość prosty. Dane open source są również zwykle dobrze sformatowane. Jednak niektóre zmienne mogą być niejasne i mogą występować uprzedzenia, takie jak przeskalowanie do określonej grupy demograficznej.

• Dane stron trzecich: Są to dane od komercyjnego dostawcy. Ale opłaty mogą być wysokie. W rzeczywistości w niektórych przypadkach może brakować jakości danych.

Według Teradata, na podstawie własnych działań firmy w zakresie sztucznej inteligencji, około 70% źródeł danych to źródła wewnętrzne, 20% z open source, a reszta od dostawców komercyjnych. Ale niezależnie od źródła wszystkie dane muszą być zaufane. Jeśli nie, prawdopodobnie pojawi się problem „śmieci wchodzą, śmieci wychodzą". Aby ocenić dane, musisz odpowiedzieć na następujące pytania:

• Czy dane są kompletne? Czego może brakować?

• Skąd pochodzą dane?

• Jakie były punkty zbiórki?

• Kto dotykał danych i je przetwarzał?

• Jakie były zmiany w danych?

• Jakie są problemy z jakością?

Jeśli pracujesz z danymi strukturalnymi, ten etap powinien być łatwiejszy. Jednak jeśli chodzi o dane nieustrukturyzowane i częściowo ustrukturyzowane, będziesz musiał oznaczyć dane, co może być długotrwałym procesem. Na rynku pojawia się jednak kilka narzędzi, które mogą pomóc zautomatyzować ten proces.

Krok #3 – Przygotowanie danych

Pierwszym krokiem w procesie przygotowania danych jest podjęcie decyzji, jakich zestawów danych użyć. Przyjrzyjmy się scenariuszowi: Załóżmy, że pracujesz dla firmy wydawniczej i chcesz opracować strategię poprawy utrzymania klientów. Niektóre dane, które powinny pomóc, obejmują informacje demograficzne dotyczące bazy klientów, takie jak wiek, płeć, dochód i wykształcenie. Aby zapewnić więcej kolorów, możesz również spojrzeć na informacje o przeglądarce. Jaki rodzaj treści interesuje klientów? Jaka jest częstotliwość i czas trwania? Jakieś inne ciekawe wzorce - powiedzmy, że dostęp do informacji w weekendy? Łącząc źródła informacji, możesz stworzyć potężny model. Na przykład, jeśli w niektórych obszarach nastąpi rezygnacja z aktywności, może to stwarzać ryzyko anulowania. To ostrzegałoby sprzedawców, aby skontaktowali się z klientami. Chociaż jest to sprytny proces, nadal istnieją miny lądowe. Uwzględnienie lub wykluczenie nawet jednej zmiennej może mieć istotny negatywny wpływ na model AI. W celu zobacz dlaczego, spójrz wstecz na kryzys finansowy. Modele gwarantowania kredytów hipotecznych były wyrafinowane i oparte na ogromnych ilościach danych. W normalnych czasach ekonomicznych działały całkiem dobrze, ponieważ duże instytucje finansowe, takie jak Goldman Sachs, JP Morgan i AIG, w dużym stopniu na nich polegały. Ale był problem: modele nie uwzględniały spadających cen mieszkań! Głównym powodem było to, że przez dziesięciolecia nigdy nie było krajowego spadku. Założono, że mieszkalnictwo jest głównie zjawiskiem lokalnym. Oczywiście ceny mieszkań nie tylko spadły, ale spadły. Modele okazały się wtedy dalekie od celu, a miliardy dolarów strat niemal zniszczyły amerykański system finansowy. Rząd federalny nie miał innego wyjścia, jak pożyczyć 700 miliardów dolarów na ratowanie Wall Street. To prawda, to skrajny przypadek. Podkreśla jednak znaczenie selekcji danych. W tym przypadku niezbędne może być posiadanie solidnego zespołu ekspertów dziedzinowych i analityków danych. Następnie, na etapie przygotowania danych, konieczne będzie czyszczenie danych. Faktem jest, że wszystkie dane mają problemy. Nawet firmy takie jak Facebook mają luki, niejasności i wartości odstające w swoich zbiorach danych. To nieuniknione. Oto kilka działań, które możesz podjąć, aby wyczyścić dane:

• Deduplikacja: Ustaw testy, aby zidentyfikować wszelkie duplikaty i usunąć zbędne dane.

• Wartości odstające: są to dane, które znacznie wykraczają poza zakres większości pozostałych danych. Może to wskazywać, że informacje nie są przydatne. Ale oczywiście zdarzają się sytuacje, w których jest odwrotnie. To byłoby do odliczenia oszustwa.

• Spójność: Upewnij się, że masz jasne definicje zmiennych. Nawet terminy takie jak „przychody” czy „klient” mogą mieć wiele znaczeń.

• Reguły walidacji: Patrząc na dane, spróbuj znaleźć nieodłączne ograniczenia. Na przykład możesz mieć flagę dla kolumny wiek. Jeśli w wielu przypadkach jest ponad 120, to dane mają poważne problemy.

• Binning: Niektóre dane nie muszą być szczegółowe. Czy to naprawdę ma znaczenie, czy ktoś ma 35 czy 37 lat? Prawdopodobnie nie. Ale porównanie tych od 30-40 do 41-50 prawdopodobnie tak.

• Nieaktualność: czy dane są aktualne i istotne?

• Scalanie: W niektórych przypadkach kolumny danych mogą zawierać bardzo podobne informacje. Być może jeden ma wzrost w calach, a drugi w stopach. Jeśli Twój model nie wymaga bardziej szczegółowego numeru, możesz po prostu użyć tego dla stóp.

• One-Hot Encoding: Jest to sposób na zastąpienie danych kategorycznych jako liczb. Przykład: Załóżmy, że mamy bazę danych z kolumną, która ma trzy możliwe wartości: Apple, Ananas i Orange. Możesz reprezentować Apple jako 1, Ananas jako 2, a Orange jako 3. Brzmi rozsądnie, prawda? Może nie. Problem w tym, że algorytm AI może uznać, że Orange jest większy niż Apple. Ale dzięki kodowaniu

na gorąco możesz uniknąć tego problemu. Utworzysz trzy nowe kolumny: is_Apple, is_Pineapple i is_Orange. Dla każdego wiersza w danych wpiszesz 1 dla miejsca występowania owocu i 0 dla reszty.

• Tabele przeliczeniowe: Można tego użyć podczas tłumaczenia danych z jednego standardu na inny. Byłoby tak w przypadku, gdy masz dane w systemie dziesiętnym i chcesz przejść do systemu metrycznego.

Te kroki znacznie poprawią jakość danych. Istnieją również narzędzia do automatyzacji, które mogą pomóc, takie jak firmy takie jak SAS, Oracle, IBM, Lavastorm Analytics i Talend. Są też projekty typu open source, takie jak OpenRefine, plyr i reshape2. Niezależnie od tego dane nie będą idealne. Brak źródła danych. Prawdopodobnie nadal będą luki i nieścisłości. Dlatego musisz być kreatywny. Zobacz, co zrobił Eyal Lifshitz, który jest dyrektorem generalnym BlueVine. Jego firma wykorzystuje sztuczną inteligencję do finansowania małych firm. „Jednym z naszych źródeł danych są informacje kredytowe naszych klientów" – powiedział. „Odkryliśmy jednak, że właściciele małych firm błędnie identyfikują swój rodzaj działalności. Może to oznaczać złe wyniki dla naszego ubezpieczenia. Aby sobie z tym poradzić, zbieramy dane ze strony klienta za pomocą algorytmów AI, które pomagają zidentyfikować branżę". Podejścia do czyszczenia danych będą również zależeć od przypadków użycia w projekcie AI. Na przykład, jeśli budujesz system do konserwacji predykcyjnej w produkcji, wyzwaniem będzie radzenie sobie z dużą różnorodnością różnych czujników. W rezultacie duża ilość danych może mieć niewielką wartość i być w większości szumem.

Etyka i zarządzanie

Musisz pamiętać o wszelkich ograniczeniach dotyczących danych. Czy sprzedawca może zabronić Ci wykorzystywania informacji do określonych celów? Być może Twoja firma będzie na haku, jeśli coś pójdzie nie tak? Aby poradzić sobie z tymi kwestiami, wskazane jest, aby sprowadzić dział prawny. W większości przypadków dane należy traktować z ostrożnością. W końcu istnieje wiele głośnych przypadków, w których firmy naruszyły prywatność. Wybitnym tego przykładem jest Facebook. Jeden z partnerów firmy, Cambridge Analytica, uzyskał dostęp do milionów punktów danych z profili bez zgody użytkowników. Kiedy demaskator odkrył to, akcje Facebooka spadły, tracąc ponad 100 miliardów dolarów wartości. Firma znalazła się również pod presją rządów USA i Europy. Należy uważać na zbieranie danych ze źródeł publicznych. To prawda, że często jest to skuteczny sposób tworzenia dużych zestawów danych. Istnieje również wiele narzędzi, które mogą zautomatyzować ten proces. Jednak skrobanie może narazić Twoją firmę na odpowiedzialność prawną, ponieważ dane mogą podlegać prawom autorskim lub prawom prywatności. Istnieją również pewne środki ostrożności, które, jak na ironię, mogą mieć nieodłączne wady. Na przykład ostatnie badanie przeprowadzone przez MIT pokazuje, że dane zanonimizowane mogą nie być bardzo zanonimizowane. Naukowcy odkryli, że w rzeczywistości dość łatwo jest zrekonstruować tego typu dane i zidentyfikować osoby- na przykład przez połączenie dwóch zestawów danych. Dokonano tego przy użyciu danych w Singapurze z sieci komórkowej (śledzenie GPS) i lokalnego systemu transportowego. Po około 11 tygodniach analizy naukowcy byli w stanie zidentyfikować 95% osobników. Na koniec upewnij się, że podejmujesz kroki w celu zabezpieczenia danych. Liczba cyberataków i zagrożeń wciąż rośnie w zastraszającym tempie. Według Verizon w 2018 r. miało miejsce ponad 53 000 incydentów i około 2200 naruszeń. W raporcie odnotowano również następujące kwestie:

• 76% naruszeń było motywowanych finansami.

• 73% pochodziło z osób spoza firmy.

• Około połowa pochodziła ze zorganizowanych grup przestępczych, a 12% z państw narodowych lub podmiotów powiązanych z państwem.

Rosnące wykorzystanie danych w chmurze i danych lokalnych może również narazić firmę na luki w zabezpieczeniach. Do tego dochodzi mobilna siła robocza, co może oznaczać dostęp do danych, który może narazić ją na naruszenia. Ataki również stają się znacznie bardziej szkodliwe. W rezultacie firma może łatwo ponieść kary, procesy sądowe i uszczerbek na reputacji. Zasadniczo, przygotowując projekt AI, upewnij się, że istnieje plan bezpieczeństwa i że jest on przestrzegany.

Ile danych potrzebujesz do sztucznej inteligencji?

Im więcej danych, tym lepiej, prawda? Tak jest zwykle. Spójrz na coś, co nazywa się Hughes Phenomenon. Oznacza to, że wraz z dodawaniem funkcji do modelu wydajność ogólnie wzrasta. Ale ilość to nie koniec, wszystko. Może nadejść moment, w którym dane zaczną się degradować. Pamiętaj, że możesz natknąć się na coś, co nazywa się przekleństwem wymiarowości. Według Charlesa Isbella, profesora i starszego zastępcy dziekana School of Interactive Computing w Georgia Tech: „Wraz ze wzrostem liczby funkcji lub wymiarów ilość danych, które potrzebujemy do dokładnego uogólnienia, rośnie wykładniczo". Jaki jest praktyczny wpływ? Może to uniemożliwić posiadanie dobrego modelu, ponieważ może nie być wystarczającej ilości danych. To dlatego, jeśli chodzi o aplikacje takie jak rozpoznawanie wzroku, przekleństwo wymiarowości może być dość problematyczne. Nawet analizując obrazy RGB, liczba wymiarów wynosi około 7500. Wyobraź sobie, jak intensywny byłby ten proces przy użyciu wideo w wysokiej rozdzielczości w czasie rzeczywistym.

Więcej terminów i pojęć dotyczących danych

Angażując się w analizę danych powinieneś znać podstawowe pojęcia. Oto kilka, które często słyszysz:

Dane kategoryczne: są to dane, które nie mają znaczenia liczbowego. Ma raczej znaczenie tekstowe, jak opis grupy (rasa i płeć). Chociaż możesz przypisać numery do każdego z elementów.

Typ danych: Jest to rodzaj informacji, które reprezentuje zmienna, na przykład Boolean, liczba całkowita, ciąg znaków lub liczba zmiennoprzecinkowa.

Analityka opisowa: to analiza danych w celu lepszego zrozumienia aktualnego stanu firmy. Niektóre przykłady obejmują mierzenie, które produkty sprzedają się lepiej lub określanie ryzyka w obsłudze klienta. Istnieje wiele tradycyjnych narzędzi programowych do analizy opisowej, takich jak aplikacja BI.

Analiza diagnostyczna: to zapytanie o dane, aby zobaczyć, dlaczego coś się stało. Ten rodzaj analizy wykorzystuje techniki takie jak eksploracja danych, drzewa decyzyjne i korelacje.

ETL (Ekstrakcja, Transformacja i Ładowanie): Jest to forma integracji danych i jest zwykle używana w hurtowni danych.

Funkcja: To jest kolumna danych.

Instancja: To jest rząd danych.

Metadane: Są to dane dotyczące danych, czyli opisów. Na przykład plik muzyczny może zawierać metadane, takie jak rozmiar, długość, data przesłania, komentarze, gatunek, wykonawca itd. Ten rodzaj danych może okazać się bardzo przydatny w projekcie AI.

Dane liczbowe: są to dowolne dane, które mogą być reprezentowane przez liczbę. Ale dane liczbowe mogą mieć dwie formy. Istnieją dane dyskretne, które są liczbą całkowitą, czyli liczbą bez kropki dziesiętnej. Następnie są dane ciągłe, które mają przepływ, powiedzmy temperaturę lub czas.

OLAP (Online Analytical Processing): Jest to technologia, która umożliwia analizowanie informacji z różnych baz danych.

Dane porządkowe: Jest to połączenie danych liczbowych i kategorycznych. Typowym tego przykładem jest pięciogwiazdkowa ocena na Amazon.com. Ma zarówno gwiazdkę, jak i powiązany z nią numer.

Analityka predykcyjna: obejmuje wykorzystanie danych do tworzenia prognoz. Modele do tego są zwykle wyrafinowane i opierają się na podejściach AI, takich jak uczenie maszynowe. Aby być skutecznym, ważne jest aktualizowanie bazowego modelu o nowe dane. Niektóre z narzędzi do analizy predykcyjnej obejmują podejścia do uczenia maszynowego, takie jak regresje.

Analiza nakazowa: chodzi o wykorzystanie Big Data do podejmowania lepszych decyzji. Koncentruje się to nie tylko na przewidywaniu wyników, ale także na zrozumieniu przesłanek. I tutaj AI odgrywa dużą rolę.

Zmienne skalarne: są to zmienne, które przechowują pojedyncze wartości, takie jak imię i nazwisko lub numer karty kredytowej.

Dane transakcyjne: Są to dane rejestrowane na temat działań finansowych, biznesowych i logistycznych. Przykłady obejmują płatności, faktury i roszczenia ubezpieczeniowe.

Wniosek

Odniesienie sukcesu dzięki sztucznej inteligencji oznacza posiadanie kultury opartej na danych. To jest kluczowe dla firm takich jak Amazon.com, Google i Facebook. Podejmując decyzje, najpierw patrzą na dane. Powinna również istnieć szeroka dostępność danych w całej organizacji. Bez tego podejścia sukces z AI będzie ulotny, niezależnie od tego, jak planujesz. Być może pomaga to wyjaśnić, że = według badania przeprowadzonego przez NewVantage Partners - około 77% respondentów twierdzi, że „przyjęcie biznesowe” Big Data i sztucznej inteligencji pozostaje wyzwaniem.

Kluczowe dania na wynos

* Dane strukturalne są oznaczone i sformatowane - i często są przechowywane w relacyjnej bazie danych lub arkuszu kalkulacyjnym.

* Dane nieustrukturyzowane to informacje, które nie mają wstępnie zdefiniowanego formatowania.

* Częściowo ustrukturyzowane dane mają kilka wewnętrznych znaczników, które pomagają w kategoryzacji.

* Big Data opisuje sposób obsługi ogromnych ilości informacji.

* Relacyjna baza danych jest oparta na relacjach danych. Ale ta struktura może okazać się trudna dla współczesnych aplikacji, takich jak sztuczna inteligencja.

* Baza danych NoSQL jest bardziej swobodna, oparta na modelu dokumentu. Dzięki temu lepiej radzi sobie z danymi nieustrukturyzowanymi i częściowo ustrukturyzowanymi.

* Proces CRISP-DM zapewnia sposób zarządzania danymi projektu, obejmujący etapy, które obejmują zrozumienie biznesowe, zrozumienie danych, przygotowanie danych, modelowanie, ocenę i wdrażanie.

* Ilość danych jest z pewnością ważna, ale trzeba też dużo pracy nad jakością. Nawet małe błędy mogą mieć ogromny wpływ na wyniki modelu AI.