

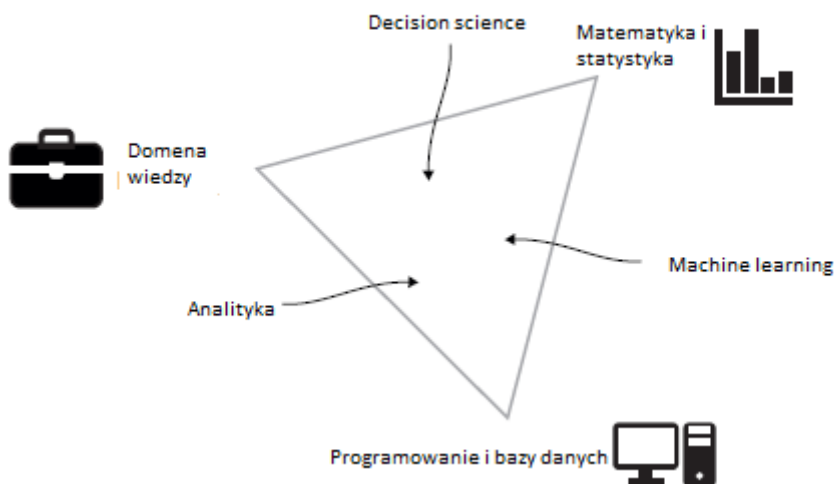
Przed Startem ...

„Najseksowniejsza praca XXI wieku”. „Najlepsza praca w Ameryce”. Naukowiec danych, tytuł, który nie istniał nawet przed 2008 r., jest obecnie stanowiskiem, na które pracodawcy nie mogą zatrudniać wystarczająco dużo, a osoby poszukujące pracy starają się zostać. Jest dobry powód do tego szumu: data science to niezwykle rozwijająca się dziedzina, z medianą pensji podstawowej w Stanach Zjednoczonych w 2019 r. przekraczającą 100 000 USD. W dobrej firmie analitycy danych cieszą się dużą autonomią i stale uczą się nowych rzeczy. Wykorzystują swoje umiejętności do rozwiązywania znaczących problemów, takich jak praca z lekarzami w celu analizy prób leków, pomoc drużynie sportowej w wyborze nowych kandydatów lub przeprojektowanie modelu cenowego dla biznesu budżetów. Wreszcie, jak omówimy, nie ma jednego sposobu, aby zostać naukowcem danych. Ludzie pochodzą z różnych środowisk, więc nie jesteś ograniczony na podstawie tego, co wybrałeś jako student. Ale nie wszystkie prace związane z nauką danych są doskonałe. Zarówno firmy, jak i osoby poszukujące pracy mogą mieć nierealistyczne oczekiwania. Firmy, które dopiero zaczynają naukę o danych, mogą pomyśleć, że jedna osoba może na przykład rozwiązać wszystkie problemy związane z danymi. Kiedy w końcu zatrudniony zostanie analityk danych, może stanąć przed niekończącą się listą zgłoszeń do zrobienia. Mogą otrzymać zadanie natychmiastowego wdrożenia systemu uczenia maszynowego, gdy nie wykonano żadnej pracy w celu przygotowania lub wyczyszczenia danych. Może nie być nikogo, kto byłby ich mentorem lub prowadził, a nawet wczuwał się w problemy, z którymi się borykają. Omówimy te kwestie bardziej szczegółowo później, gdzie pomożemy Ci uniknąć dołączania do firm, które mogą być nieodpowiednie dla nowego naukowca danych, oraz doradzimy co zrobić, jeśli znajdziesz się w negatywnej sytuacji. Z drugiej strony poszukujący pracy mogą pomyśleć, że w ich nowej karierze nigdy nie będzie nudnej chwili. Mogą oczekiwać, że interesariusze będą rutynowo postępować zgodnie z ich zaleceniami, że inżynierowie danych będą mogli natychmiast rozwiązać wszelkie problemy z jakością danych i że uzyskają najszybsze dostępne zasoby obliczeniowe do wdrożenia swoich modeli. W rzeczywistości naukowcy zajmujący się danymi spędzają dużo czasu na czyszczeniu i przygotowywaniu danych, a także na zarządzaniu oczekiwaniami i priorytetami innych zespołów. Projekty nie zawsze się udają. Kierownictwo wyższego szczebla może składać klientom nierealistyczne obietnice dotyczące tego, co mogą zapewnić modele analizy danych. Głównym zadaniem osoby może być praca z archaicznym systemem danych, którego nie da się zautomatyzować i wymaga godzin pracy otepiającej każdego tygodnia tylko po to, aby oczyścić dane. Naukowcy zajmujący się danymi mogą zauważyć wiele błędów statystycznych lub technicznych w starszych analizach, które mają realne konsekwencje, ale nikt nie jest tym zainteresowany, a oni są tak przeciążeni pracą, że nie mają czasu, aby spróbować je naprawić. Albo analityk danych może zostać poproszony o przygotowanie raportów potwierdzających to, co zdecydowało już kierownictwo wyższego szczebla, więc może martwić się, że zostanie zwolniony, jeśli udzieli niezależnej odpowiedzi. Jesteśmy tutaj, aby poprowadzić Cię przez proces zostania naukowcem danych i rozwoju kariery. Chcemy mieć pewność, że czytelnicy uzyskają wszystkie wspaniałe cechy bycia naukowcem danych i unikną większości pułapek. Być może pracujesz w sąsiedniej dziedzinie, takiej jak analityka marketingowa, i zastanawiasz się, jak dokonać zmiany. A może jesteś już analitykiem danych, ale szukasz nowej pracy i nie sądzisz, że podszedłeś dobrze do swojego pierwszego poszukiwania pracy. Lub chcesz rozwinąć swoją karierę, przemawiając na konferencjach, wspierając open source lub zostając niezależnym konsultantem. Niezależnie od Twojego poziomu jesteśmy przekonani, że okażemy się pomocni. Na początek omawiamy główne możliwości zdobycia umiejętności data science i budowania portfolio, aby obejść paradoks wymagający doświadczenia, aby zdobyć doświadczenie. Pokażemy, jak napisać list motywacyjny i życiorys, który zapewni Ci rozmowę kwalifikacyjną i jak zbudować swoją sieć kontaktów, aby uzyskać skierowanie. Omawiamy strategie negocjacyjne, które według badań pozwolą Ci uzyskać najlepszą możliwą ofertę. Kiedy pracujesz w nauce danych, będziesz pisać analizy, pracować z interesariuszami,

a może nawet wprowadzać model do produkcji . Pomożemy zrozumieć, jak wyglądają wszystkie te procesy i jak przygotować się na sukces. Znajdziesz strategię, jak odzyskać siły, gdy projekt nieuchronnie się nie powiedzie, a kiedy będziesz gotowy, jesteśmy tutaj, aby poprowadzić Cię przez decyzję, gdzie rozpocząć karierę: przejście do zarządzania, kontynuowanie być indywidualnym współpracownikiem lub nawet występować jako niezależny konsultant. Zanim jednak rozpoczniesz tę podróż, musisz wiedzieć, kim są naukowcy zajmujący się danymi i jaką pracę wykonują. Nauka o danych to szeroka dziedzina, która obejmuje wiele rodzajów pracy i im lepiej rozumiesz różnice między tymi obszarami, tym lepiej możesz się w nich rozwijać.

Co to jest nauka o danych?

Nauka o danych to praktyka wykorzystywania danych do próby zrozumienia i rozwiązania rzeczywistych problemów. Ta koncepcja nie jest zupełnie nowa; ludzie analizowali dane dotyczące sprzedaży i trendy od czasu wynalezienia zera. Jednak w ostatnim dziesięcioleciu uzyskaliśmy dostęp do wykładniczo większej liczby danych niż wcześniej. Pojawienie się komputerów pomogło w generowaniu wszystkich tych danych, ale komputery to także nasz jedyny sposób przetwarzania stosów informacji. Za pomocą kodu komputerowego analityk danych może przekształcać lub agregować dane, przeprowadzać analizy statystyczne lub trenować modele uczenia maszynowego. Wynikiem tego kodu może być raport lub pulpitanawigacyjny do spożycia przez ludzi lub może to być model uczenia maszynowego, który zostanie wdrożony w celu ciągłego działania. Jeśli firma zajmująca się handlem detalicznym ma problem z podjęciem decyzji, na przykład, gdzie postawić nowy sklep, może wezwać analityka danych do przeprowadzenia analizy. Analityk danych mógłby przyjrzeć się historycznym danym o lokalizacjach, do których wysyłane są zamówienia online, aby zrozumieć, gdzie jest popyt klientów. Mogą również łączyć te dane dotyczące lokalizacji klientów z informacjami demograficznymi i informacjami o dochodach dla tych miejscowości ze spisów powszechnych. Dzięki tym zbiorom danych mogli znaleźć optymalne miejsce na nowy sklep i stworzyć prezentację Microsoft PowerPoint, aby przedstawić swoje rekomendacje wiceprezesowi firmy ds. handlu detalicznego. W innej sytuacji ta sama firma detaliczna może chcieć zwiększyć rozmiary zamówień online, polecając produkty klientom podczas zakupów. Analityk danych mógłby załadować historyczne dane dotyczące zamówień internetowych i stworzyć model uczenia maszynowego, który - biorąc pod uwagę zestaw produktów znajdujących się obecnie w koszyku - przewiduje najlepszy produkt do polecenia kupującemu. Po utworzeniu tego modelu analityk danych współpracowałby z zespołem inżynierów firmy, aby za każdym razem, gdy klient robił zakupy, nowy model uczenia maszynowego wyświetlał zalecane pozycje. Kiedy wiele osób zaczyna zajmować się nauką o danych, jednym z wyzwań, przed którymi stają, jest przytłoczenie ilością rzeczy, których muszą się nauczyć, takich jak kodowanie (ale jakiego języka?), Statystyki (ale które metody są najważniejsze w praktyce, a które głównie akademickie?), uczenie maszynowe (ale czym różni się uczenie maszynowe od statystyki lub sztucznej inteligencji?) oraz znajomość domeny w dowolnej branży, w której chcą pracować (ale co, jeśli nie wiesz, gdzie chcesz pracować?). Ponadto muszą nauczyć się umiejętności biznesowych, takich jak skuteczne przekazywanie wyników odbiorcom, od innych analityków danych po dyrektora generalnego. Ten niepokój może być spotęgowany przez ogłoszenia o pracę, w których pojawia się prośba o tytuł doktora, wieloletnie doświadczenie w nauce o danych oraz znajomość listy metod statystycznych i programistycznych. Jak możesz nauczyć się wszystkich tych umiejętności? Od których powinieneś zacząć? Jakie są podstawy? Jeśli przyjrzałeś się różnym obszarom nauki o danych, być może znasz popularny diagram Venna Drew Conwaya. Zdaniem Conwaya (w czasie tworzenia diagramu) nauka o danych znalazła się na przecięciu wiedzy matematycznej i statystycznej, wiedzy specjalistycznej w danej dziedzinie i umiejętności hakerskich (czyli kodowania). Ten obraz jest często używany jako podstawa do zdefiniowania, kim jest analityk danych. Z naszej perspektywy elementy nauki o danych różnią się nieco od tego, co zaproponował .



Zmieniliśmy oryginalny diagram Venna Conwaya na trójkąt, ponieważ nie oznacza to, że albo masz umiejętności, albo nie; chodzi o to, że możesz go posiadać w różnym stopniu niż inni w tej dziedzinie. Chociaż prawdą jest, że wszystkie trzy umiejętności są fundamentalne i każda z nich jest do pewnego stopnia potrzebna, nie musisz być ekspertem w każdej z nich. W trójkącie umieściliśmy różne typy specjalizacji w dziedzinie nauki o danych. Te specjalizacje nie zawsze są przypisane jeden do jednego z nazwami stanowisk, a nawet jeśli to robią, różne firmy czasami nazywają je różnymi rzeczami. Więc co oznacza każdy z tych składników?

Matematyka / statystyka

Na poziomie podstawowym znajomość matematyki i statystyki to umiejętność korzystania z danych. Rozbijamy tę umiejętność czytania na trzy poziomy wiedzy:

- * Te techniki istnieją - jeśli nie wiesz, że coś jest możliwe, nie możesz tego użyć. Gdyby analityk danych próbował pogrupować podobnych klientów, wiedząc, że mogą to zrobić metody statystyczne (zwane grupowaniem), byłby to pierwszy krok.
- * Jak stosować techniki - chociaż naukowiec zajmujący się danymi może znać wiele technik, musi również umieć zrozumieć zawiłości ich stosowania - nie tylko jak napisać kod, aby zastosować metody, ale także jak je skonfigurować. Jeśli analityk danych chce użyć metody, takiej jak grupowanie k-średnich, do grupowania klientów, musiałby zrozumieć, jak tworzyć klastry k-średnich w języku programowania, takim jak R lub Python. Musieliby również wiedzieć, jak dostosować parametry metody, na przykład wybierając liczbę utworzonych grup.
- * Jak wybrać techniki do wypróbowania - Ponieważ w nauce o danych można zastosować tak wiele możliwych technik, ważne jest, aby analityk danych mógł szybko ocenić, czy dana technika będzie dobrze działać. W naszym przykładzie grupowania klientów nawet po tym, jak naukowiec zajmujący się danymi koncentruje się na klastrach, muszą wziąć pod uwagę dziesiątki różnych metod i algorytmów. Zamiast wypróbowywać każdą metodę, muszą być w stanie szybko wykluczyć metody i skupić się tylko na kilku.

Tego rodzaju umiejętności są stale używane w ramach roli nauk o danych. Aby rozważyć inny przykład, załóżmy, że pracujesz w firmie zajmującej się handlem elektronicznym. Twój partner biznesowy może być zainteresowany tym, które kraje mają najwyższą średnią wartość zamówienia. Jeśli masz dostępne dane, odpowiedź na to pytanie jest łatwa. Ale zamiast po prostu przedstawiać te informacje i pozwalać

klientowi wyciągnąć własne wnioski, możesz sięgnąć głębiej. Jeśli masz jedno zamówienie z kraju A za 100 USD i tysiąc zamówień z kraju B o średniej wartości 75 USD, to prawdą jest, że kraj A ma wyższą średnią wartość zamówienia. Ale czy byłbyś przekonany, mówiąc, że oznacza to, że Twój partner biznesowy zdecydowanie powinien zainwestować w reklamę w kraju A, aby zwiększyć liczbę zamówień? Prawdopodobnie nie. Masz tylko jeden punkt danych dla kraju A i być może jest to wartość odstająca. Gdyby zamiast tego kraj A miał 500 zamówień, możesz użyć testu statystycznego, aby sprawdzić, czy wartość zamówienia znacznie się różniła, co oznacza, że jeśli naprawdę nie było różnicy między A i B w tej mierze, prawdopodobnie nie zobaczysz różnicy zrobili. W tym jednym paragrafowym przykładzie dokonano wielu różnych ocen, które podejścia są rozsądne, co należy wziąć pod uwagę, a jakie wyniki uznano za nieistotne.

Bazy danych / programowanie

Programowanie i bazy danych odnoszą się do możliwości pobierania danych z firmowych baz danych i pisania czystego, wydajnego, łatwego w utrzymaniu kodu. Umiejętności te są pod wieloma względami podobne do tego, co musi wiedzieć programista, z tym wyjątkiem, że naukowcy zajmujący się danymi muszą pisać kod, który przeprowadza otwartą analizę, zamiast generować predefiniowane dane wyjściowe. Stos danych każdej firmy jest unikalny, więc od analityka danych nie jest wymagany jeden zestaw umiejętności technicznych. Ogólnie rzecz biorąc, musisz wiedzieć, jak pobierać dane z bazy danych i jak czyścić, manipulować, podsumowywać, wizualizować i udostępniać dane. W większości prac związanych z nauką danych głównym językiem jest R lub Python. R to język programowania, który ma swoje korzenie w statystykach, więc ogólnie jest najsilniejszy do analizy i modelowania statystycznego, wizualizacji i generowania raportów z wynikami. Python to język programowania, który powstał jako ogólny język programowania i stał się niezwykle popularny w nauce o danych. Python jest znany z tego, że jest lepszy niż R w pracy z dużymi zbiorami danych, uczeniu maszynowym i obsłudze algorytmów czasu rzeczywistego (takich jak silniki rekomendacji Amazona). Jednak dzięki pracy wielu współpracowników możliwości tych dwóch języków są obecnie prawie równe. Naukowcy zajmujący się danymi z powodzeniem używają języka R do tworzenia modeli uczenia maszynowego, które są uruchamiane miliony razy w tygodniu, a także wykonują przejrzyste, reprezentatywne analizy statystyczne w Pythonie. R i Python to najpopularniejsze języki do nauki o danych z kilku powodów:

* Są darmowe i typu open source, co oznacza, że wiele osób, a nie tylko jedna firma czy grupa, udostępnia kod, którego możesz używać. Mają wiele pakietów lub bibliotek (zestawów kodu) do zbierania danych, manipulacji, wizualizacji, analizy statystycznej i uczenia maszynowego.

* Co ważne, ponieważ każdy język ma tak wielu zwolenników, naukowcom zajmującym się danymi łatwo jest znaleźć pomoc, gdy napotkają problemy. Chociaż niektóre firmy nadal używają SAS, SPSS, STATA, MATLAB lub innych płatnych programów, wiele z nich zaczyna zamiast tego przechodzić na R lub Python.

Chociaż większość analiz nauki o danych jest wykonywana w języku R lub Python, często trzeba pracować z bazą danych, aby uzyskać dane. Tutaj pojawia się język SQL. SQL jest językiem programowania używanym przez większość baz danych do manipulowania zawartymi w nich danymi lub ich wyodrębniania. Weźmy pod uwagę naukowca zajmującego się danymi, który chce przeanalizować setki milionów rekordów zamówień klientów w firmie, aby przewidzieć, jak dzienne zamówienia będą się zmieniać w czasie. Po pierwsze, prawdopodobnie napiszą zapytanie SQL, aby uzyskać liczbę zamówień każdego dnia. Następnie obliczali dzienne liczby zamówień i generowali prognozę statystyczną w języku R lub Python. Z tego powodu język SQL jest niezwykle popularny w społeczności analityków danych i trudno jest zająć zbyt daleko, nie wiedząc o tym. Inną podstawową umiejętnością jest używanie kontroli wersji - metody śledzenia zmian kodu w czasie. Kontrola wersji

umożliwia przechowywanie plików; przywrócić je do poprzedniego czasu; i zobacz, kto zmienił jaki plik, jak i kiedy. Ta umiejętność jest niezwykle ważna w nauce o danych i inżynierii oprogramowania, ponieważ jeśli ktoś przypadkowo zmieni plik, który zepsuje twój kod, chcesz mieć możliwość przywrócenia lub zobaczenia, co się zmieniło. Git jest zdecydowanie najczęściej używanym systemem do kontroli wersji i jest często używany w połączeniu z GitHub, usługą hostingową opartą na sieci Web dla Git. Git umożliwia zapisywanie (zatwierdzanie) zmian, a także przeglądanie całej historii projektu i tego, jak zmieniał się on z każdym zatwierdzeniem. Jeśli dwie osoby pracują osobno nad tym samym plikiem, Git zapewnia, że żadna praca nie zostanie przypadkowo usunięta ani nadpisana. W wielu firmach, szczególnie tych z silnymi zespołami inżynierów, musisz używać Git, jeśli chcesz udostępnić swój kod lub wprowadzić coś do produkcji.

Czy możesz zostać naukowcem danych bez programowania?

Możliwe jest wykonanie dużej ilości danych przy użyciu tylko programu Excel, Tableau lub innych narzędzi analizy biznesowej, które mają interfejsy graficzne. Chociaż nie piszesz kodu, narzędzia te mają podobną funkcjonalność, co języki takie jak R czy Python, a wielu analityków danych czasami ich używa. Ale czy mogą być kompletnym zestawem narzędzi do nauki o danych? Mówimy nie. Praktycznie bardzo niewiele firm ma zespół analityki danych, w którym nie trzeba by było programować. Ale nawet gdyby tak nie było, programowanie ma przewagę nad używaniem tych narzędzi. Pierwszą zaletą programowania jest powtarzalność. Pisząc kod zamiast korzystać z oprogramowania typu „wskaż i kliknij”, możesz uruchomić go ponownie za każdym razem, gdy zmieniają się dane, czy to codziennie, czy co sześć miesięcy. Ta zaleta wiąże się również z kontrolą wersji: zamiast zmieniać nazwę pliku za każdym razem, gdy zmienia się kod, możesz zachować jeden plik, ale zobaczyć całą jego historię. Drugą zaletą jest elastyczność. Jeśli na przykład Tableau nie ma dostępnego typu wykresu, nie będzie można go utworzyć. Ale dzięki programowaniu możesz napisać własny kod, aby stworzyć coś, o czym twórcy i opiekunowie narzędzia nigdy nie pomyśleli. Trzecią i ostatnią zaletą języków open source, takich jak Python i R, jest wkład społeczności. Tysiące ludzi tworzy pakiety i publikuje je otwarcie na GitHub i / lub CRAN (dla R) i pip (dla Pythona). Możesz pobrać ten kod i użyć go do własnych problemów. Nie musisz polegać na jednej firmie lub grupie osób, które dodają funkcje

Zrozumienie biznesu

Firmy muszą, delikatnie mówiąc, różnie rozumieć, jak działa nauka o danych. Często kierownictwo chce po prostu coś zrobić i zwraca się do swoich jednoróżców z analityki danych, aby to się stało. Podstawową umiejętnością w nauce o danych jest umiejętność przełożenia sytuacji biznesowej na pytanie dotyczące danych, znalezienia odpowiedzi na dane i wreszcie udzielenia odpowiedzi biznesowej. Przedsiębiorca może zapytać na przykład „Dlaczego nasi klienci odchodzą?” Ale nie ma pakietu Pythona „dlaczego-klienci-odchodzą”, który można by zaimportować - to do Ciebie należy wywnioskowanie, jak odpowiedzieć na to pytanie za pomocą danych. Zrozumienie biznesu to miejsce, w którym ideały data science spotykają się z praktycznymi aspektami prawdziwego świata. Nie wystarczy chcieć określonej informacji, nie wiedząc, jak dane są przechowywane i aktualizowane w Twojej firmie. Jeśli Twoja firma jest usługą abonamentową, gdzie znajdują się dane? Co się stanie, jeśli ktoś zmieni subskrypcję? Czy wiersz tego subskrybenta jest aktualizowany, czy też jest dodawany inny wiersz do tabeli? Czy musisz obejść jakieś błędy lub niespójności w danych? Jeśli nie znasz odpowiedzi na te pytania, nie będziesz w stanie udzielić dokładnej odpowiedzi na podstawowe pytanie, takie jak „Ilu subskrybentów mieliśmy 2 marca 2019 roku?” Zrozumienie biznesu pomaga również dowiedzieć się, jakie pytania zadać. Pytanie „Co powinniśmy zrobić dalej?” ze strony interesariusza to trochę tak, jakby ktoś zapytał go: „Dlaczego tego nie robimy aby mieć więcej pieniędzy?” Tego typu pytanie wymaga więcej pytań. Rozwinięcie zrozumienia podstawowej działalności (a także zaangażowanych osobowości) może pomóc w lepszym przeanalizowaniu sytuacji. Możesz odpowiedzieć na pytanie „W

sprawie której linii produktów szukasz wskazówek?” lub „Czy chciałbyś, aby był większy udział określonego sektora naszej publiczności?”. Inną częścią zrozumienia biznesu jest rozwijanie ogólnych umiejętności biznesowych, takich jak możliwość dostosowania prezentacji i raportów do różnych odbiorców. Czasami będziesz omawiać lepszą metodologię w sali pełnej doktoratów ze statystyki, a czasami będziesz przed wiceprezesem, który nie brał udziału w zajęciach z matematyki od 20 lat. Musisz poinformować swoją publiczność, nie rozmawiając z nią ani nie komplikując jej zbytnio. Wreszcie, gdy osiągasz wyższy wiek, częścią twojej pracy jest określenie, gdzie firma może skorzystać na analizie danych. Jeśli chciałeś zbudować system prognozowania dla swojej firmy, ale nigdy nie miałeś wsparcia kierownictwa, dołączenie do zespołu zarządzającego może pomóc rozwiązać ten problem. Starszy analityk danych będzie szukał miejsc do wdrożenia uczenia maszynowego, ponieważ zna jego ograniczenia i możliwości, a także jakie rodzaje zadań skorzystałyby na automatyzacji.

Czy nauka o danych zniknie?

U podstaw pytania, czy nauka o danych pojawi się za dekadę lub dwie, leżą dwa główne obawy: że praca zostanie zautomatyzowana, że nauka o danych jest przereklamowana, a bańka na rynku pracy pęknie. Prawdą jest, że niektóre części procesu analizy danych można zautomatyzować. Zautomatyzowane uczenie maszynowe (AutoML) może porównywać wydajność różnych modeli i wykonywać określone części przygotowania danych (takie jak zmienne skalowania). Ale te zadania to tylko niewielka część procesu nauki o danych. Na przykład często będziesz musiał tworzyć dane samodzielnie; bardzo rzadko czekają na Ciebie idealnie czyste dane. Ponadto tworzenie danych zwykle wiąże się z rozmową z innymi osobami, takimi jak badacze doświadczeń użytkowników lub inżynierowie, którzy przeprowadzą ankietę lub zarejestrują działania użytkownika, które mogą wpłynąć na analizę. Jeśli chodzi o możliwość pojawienia się bańki na rynku pracy, dobrym porównaniem jest inżynieria oprogramowania w latach 80. W miarę jak komputery stawały się tańsze, szybsze i bardziej powszechne, pojawiły się obawy, że wkrótce komputer będzie w stanie zrobić wszystko i że nie będą potrzebni programiści. Ale stało się odwrotnie i obecnie w Stanach Zjednoczonych jest ponad 1,2 miliona inżynierów oprogramowania. Chociaż tytuły takie jak webmaster zniknęły, więcej ludzi niż kiedykolwiek pracuje nad rozwojem, utrzymaniem i ulepszaniem stron internetowych. Wierzmy, że będzie więcej specjalizacji w nauce o danych, co może doprowadzić do zniknięcia ogólnego tytułu naukowca ds. Danych, ale wiele firm wciąż jest na wczesnym etapie uczenia się, jak wykorzystać naukę o danych i pozostało jeszcze wiele pracy.

Różne rodzaje prac związanych z nauką danych

Możesz mieszać i łączyć trzy podstawowe umiejętności nauki o danych w kilku zawodach, z których wszystkie mają uzasadnienie dla posiadania tytułu naukowca danych. Z naszej perspektywy te umiejętności łączą się ze sobą na trzy główne sposoby: analityka, uczenie maszynowe i nauka o decyzjach. Każdy z tych obszarów służy firmie do innego celu i zasadniczo zapewnia inną rzecz. Szukając ofert pracy w dziedzinie nauki o danych, należy zwracać mniejszą uwagę na nazwy stanowisk, a bardziej na opisy stanowisk i to, o co są pytani podczas rozmów kwalifikacyjnych. Przyjrzyj się przeszłości osób pełniących role w dziedzinie nauki o danych, na przykład jakie zajmowali poprzednie stanowiska i jakie są ich stopnie naukowe. Może się okazać, że osoby pracujące na podobnych stanowiskach mają zupełnie inne tytuły lub osoby, które mają ten sam tytuł naukowca zajmującego się danymi, robią zupełnie inne rzeczy. Mówiąc w tej książce o różnych rodzajach zawodów związanych z nauką danych, pamiętaj, że faktyczne tytuły używane w firmach mogą się różnić.

Analityka

Analityk pobiera dane i przedstawia je odpowiednim osobom. Gdy firma określi swoje roczne cele, możesz umieścić te cele na pulpicie nawigacyjnym, aby kierownictwo mogło co tydzień śledzić postępy.

Możesz także wbudować funkcje, które pozwolą menedżerom łatwo rozbić liczby według kraju lub typu produktu. Ta praca wymaga dużo czyszczenia i przygotowania danych, ale generalnie mniej pracy przy interpretacji danych. Chociaż powinieneś być w stanie wykryć i naprawić problemy z jakością danych, główną osobą, która podejmuje decyzje na podstawie tych danych, jest partner biznesowy. Zatem zadaniem analityka jest pobieranie danych z firmy, efektywne ich formatowanie i porządkowanie oraz dostarczanie innym osobom. Ponieważ rola analityka nie obejmuje wielu statystyk i uczenia maszynowego, niektórzy ludzie i firmy uznałyby tę rolę za wykraczającą poza dziedzinę nauki o danych. Jednak duża część pracy, na przykład tworzenie znaczących wizualizacji i podejmowanie decyzji dotyczących konkretnych przekształceń danych, wymaga tych samych umiejętności, które są wykorzystywane w innych rodzajach ról związanych z nauką o danych. Analitykowi może zostać przydzielone zadanie, na przykład „Utwórz automatyczny pulpit nawigacyjny, który pokazuje, jak liczba naszych subskrybentów zmienia się w czasie i pozwala nam filtrować dane tylko do subskrybentów określonych produktów lub w określonych regionach geograficznych”. Analityk musiałby znaleźć odpowiednie dane w firmie, dowiedzieć się, jak odpowiednio przekształcić dane (na przykład zmieniając je z codziennych na cotygodniowe nowe subskrypcje), a następnie utworzyć znaczący zestaw pulpitu nawigacyjnego, które są atrakcyjne wizualnie i automatycznie aktualizują się każdego dnia bez błędów.

Krótką zasadą: analityk tworzy dashboardy i raporty, które dostarczają dane.

Nauczanie maszynowe

Inżynier uczenia maszynowego opracowuje modele uczenia maszynowego i wprowadza je do produkcji, gdzie działają one w sposób ciągły. Mogą zoptymalizować algorytm rankingu wyników wyszukiwania w witrynie handlu elektronicznego, stworzyć system rekomendacji lub monitorować model w środowisku produkcyjnym, aby upewnić się, że jego wydajność nie pogorszyła się od czasu jego wdrożenia. Inżynier uczenia maszynowego poświęca mniej czasu na takie rzeczy, jak tworzenie wizualizacji, które przekonają ludzi do czegoś, a więcej czasu na programowanie w naukach o danych. Duża różnica między tą rolą a innymi rodzajami stanowisk związanych z nauką o danych polega na tym, że wynik pracy jest przeznaczony głównie do wykorzystania przez maszynę. Możesz na przykład tworzyć modele uczenia maszynowego, które zostaną przekształcone w interfejsy programowania aplikacji (API) dla innych maszyn. Pod wieloma względami będziesz bliżej programisty niż innych osób na stanowiskach analityki danych. Chociaż każdy analityk danych powinien przestrzegać najlepszych praktyk kodowania, jako inżynier systemów uczących się musisz to zrobić. Twój kod musi być wydajny, przetestowany i napisany, aby inni ludzie mogli z nim pracować. Z tego powodu wielu inżynierów zajmujących się uczeniem maszynowym wywodzi się z informatyki. W roli inżyniera systemów uczących się osoba może zostać poproszona o utworzenie modelu uczenia maszynowego, który w czasie rzeczywistym może przewidzieć prawdopodobieństwo, że klient w witrynie faktycznie dokończy zamówienie. Inżynier uczenia maszynowego musiałby znaleźć dane historyczne w firmie, wytrenować na nich model uczenia maszynowego, przekształcić ten model w interfejs API, a następnie wdrożyć interfejs API, aby witryna internetowa mogła go uruchomić. Jeśli ten model z jakiegoś powodu przestanie działać, inżynier uczenia maszynowego zostanie wezwany do jego naprawy.

Krótką zasadą: inżynier uczenia maszynowego tworzy modele, które są uruchamiane w sposób ciągły.

Nauka decyzyjna

Specjalista ds. Decyzji przekształca surowe dane firmy w informacje, które pomagają firmie w podejmowaniu decyzji. Ta praca opiera się na głębokim zrozumieniu różnych metod matematycznych i statystycznych oraz znajomości procesu podejmowania decyzji biznesowych. Ponadto naukowcy

zajmujący się decyzjami muszą być w stanie tworzyć przekonujące wizualizacje i tabele, aby osoby nietechniczne, z którymi rozmawiają, zrozumiały ich analizę. Chociaż naukowiec zajmujący się decyzjami dużo programuje, jego praca jest generalnie uruchamiana tylko raz, aby przeprowadzić konkretną analizę, więc mogą uciec od posiadania kodu, który jest nieefektywny lub trudny w utrzymaniu. Naukowiec decyzyjny musi zrozumieć potrzeby innych osób w firmie i dowiedzieć się, jak generować konstruktywne informacje. Na przykład dyrektor ds. marketingu może poprosić naukowca zajmującego się podejmowaniem decyzji, aby pomógł mu zdecydować, na jakie rodzaje produktów należy zwrócić uwagę w firmowym przewodniku po prezentach świątecznych. Decydent może zbadać, które produkty dobrze się sprzedały, bez umieszczania ich w przewodniku po prezentach, porozmawiać z zespołem badawczym użytkowników o przeprowadzeniu ankiety i skorzystać z zasad nauk behawioralnych, aby przeprowadzić analizę, aby znaleźć optymalne pozycje do zasugerowania. Rezultatem będzie prawdopodobnie prezentacja PowerPoint lub raport, który zostanie udostępniony menedżerom produktu, wiceprezesom i innym biznesmenom. Naukowiec podejmujący decyzje często wykorzystuje swoją wiedzę o statystykach, aby pomóc firmie w podejmowaniu decyzji w warunkach niepewności. Na przykład naukowiec decyzyjny może być odpowiedzialny za prowadzenie firmowego systemu analizy eksperymentów. Wiele firm przeprowadza eksperymenty online lub testy A / B, aby sprawdzić, czy zmiana jest skuteczna. Ta zmiana może być tak prosta, jak dodanie nowego przycisku lub tak skomplikowana, jak zmiana systemu rankingowego wyników wyszukiwania lub całkowite przeprojektowanie strony. Podczas testu A / B odwiedzający są losowo przypisywani do jednego z dwóch lub więcej warunków, np. Połowa do starej wersji strony głównej, która jest kontrolką, a połowa do nowej wersji, czyli leczenia. Następnie porównuje się działania odwiedzających po wejściu do eksperymentu, aby sprawdzić, czy osoby w trakcie leczenia mają wyższy wskaźnik wykonywania pożądanych czynności, takich jak kupowanie produktów. Ze względu na losowość rzadko zdarza się, aby wskaźniki w grupie kontrolnej i leczeniu były dokładnie takie same. Załóżmy, że wyrzuciłeś dwie monety, a jedna z nich odwróciła reszkę 52 razy na 100, a jedna 49 razy na 100. Czy doszedłbyś do wniosku, że pierwsza moneta ma większe szanse na wypadnięcie orła? Oczywiście nie! Ale partner biznesowy może przyjrzeć się eksperymentowi, zobaczyć, że współczynnik konwersji wynosi 5,4% w grupie kontrolnej i 5,6% w leczeniu i zadeklarować, że leczenie zakończyło się sukcesem. Decydent pomaga w interpretacji danych, egzekwowaniu najlepszych praktyk w zakresie projektowania eksperymentów i nie tylko.

Krótką zasadą: naukowiec podejmujący decyzje tworzy analizy, które dają rekomendacje

Powiązane prace

Chociaż trzy obszary omówione w poprzednich sekcjach to główne typy stanowisk związanych z nauką o danych, możesz zauważyć kilka innych odrębnych ról, które wykraczają poza te kategorie. Podajemy tutaj te oferty pracy, ponieważ dobrze jest zrozumieć stanowiska, które są dostępne, i ponieważ może być konieczna współpraca z kolegami na tych stanowiskach. To powiedziawszy, jeśli interesuje Cię jedna z tych ról, materiały zawarte w tej książce mogą być dla Ciebie mniej istotne.

ANALITYCZNA BUSINESS INTELLIGENCE

Analitik Business Intelligence działa podobnie do analityka, ale zazwyczaj wykorzystuje mniej wiedzy statystycznej i programistycznej. Ich narzędziem z wyboru może być Excel zamiast Python i mogą nigdy nie tworzyć modeli statystycznych. Chociaż ich funkcja zawodowa jest podobna do funkcji analityka, tworzą mniej wyrafinowane wyniki ze względu na ograniczenia ich narzędzi i technik. Jeśli chcesz zająć się uczeniem maszynowym lub programowaniem albo zastosować metody statystyczne, stanowisko analityka Business Intelligence może być bardzo frustrującą rolą, ponieważ nie pomoże Ci zdobyć nowych umiejętności. Ponadto te prace są zwykle mniej prestiżowe niż prace związane z nauką danych

i są uważane za mniej prestiżowe. Jednak praca analityka biznesowego może być dobrym początkiem do zostania naukowcem zajmującym się danymi, zwłaszcza jeśli wcześniej nie pracowałeś z danymi w środowisku biznesowym. Jeśli chcesz zacząć jako analityk Business Intelligence i rozwinąć się w naukowca danych, poszukaj stanowisk, na których możesz nauczyć się niektórych umiejętności, których możesz nie mieć, takich jak programowanie w języku R lub Python.

INŻYNIER DANYCH

Inżynier danych koncentruje się na utrzymywaniu danych w bazach danych i zapewnieniu ludziom dostępu do danych, których potrzebują. Nie generują raportów, nie analizują ani nie opracowują modeli; zamiast tego przechowują dane starannie przechowywane i sformatowane w dobrze ustrukturyzowanych bazach danych, aby inne osoby mogły robić te rzeczy. Inżynierowi danych można powierzyć zadanie utrzymania wszystkich rekordów klientów w wielkoskalowej bazie danych w chmurze i dodawania nowych tabel do tej bazy danych zgodnie z żądaniem. Inżynierowie danych znacznie różnią się od naukowców zajmujących się danymi, są jeszcze bardziej rzadcy i poszukiwani. Inżynier danych może pomóc w tworzeniu komponentów zaplecza danych wewnętrznego systemu eksperymentalnego firmy i aktualizować przepływ przetwarzania danych, kiedy prace trwają zbyt długo. Inni inżynierowie danych opracowują i monitorują środowiska wsadowe i strumieniowe, zarządzając danymi od gromadzenia, przez przetwarzanie, po przechowywanie danych. Jeśli interesujesz się inżynierią danych, będziesz potrzebować dużych umiejętności informatycznych; wielu inżynierów danych to byli inżynierowie oprogramowania.

BADACZ

Naukowiec opracowuje i wdraża nowe narzędzia, algorytmy i metodologie, które są często używane przez innych naukowców zajmujących się danymi w firmie. Tego typu stanowiska prawie zawsze wymagają doktoratu, zwykle z informatyki, statystyki, ilościowych nauk społecznych lub pokrewnej dziedziny. Naukowcy mogą spędzić tygodnie na badaniu i wypróbowywaniu metod zwiększania mocy eksperymentów online, uzyskiwania o 1% większej dokładności rozpoznawania obrazu w samochodach samojezdnych lub tworzenia nowego algorytmu głębokiego uczenia. Mogą nawet spędzać czas na pisaniu artykułów naukowych, które rzadko są wykorzystywane w firmie, ale które pomagają podnieść prestiż firmy i (najlepiej) posunąć się naprzód w tej dziedzinie.

Wybór swojej ścieżki

Wkrótce omówimy niektóre opcje zdobywania umiejętności w zakresie analizy danych, zalety i wady każdej z opcji oraz kilka sugestii dotyczących wyboru spośród nich. . W tym momencie dobrze jest zacząć zastanawiać się nad obszarem data science, w którym chcesz się specjalizować. Gdzie masz już doświadczenie? Widzieliśmy naukowców zajmujących się danymi, którzy są byłymi inżynierami, profesorami psychologii, menedżerami marketingu, studentami statystyki i pracownikami socjalnymi. Wiele razy wiedza, którą zdobyłeś w innych zawodach i dziedzinach akademickich, może pomóc ci być lepszym naukowcem zajmującym się danymi. Jeśli zajmujesz się już nauką o danych, warto zastanowić się teraz, w której części trójkąta się znajdujesz. Czy jesteś z tego zadowolony? Czy chcesz przełączyć się na inny typ pracy związanej z analizą danych? Przejście jest często dostępne.