DATA SCIENCE. TWORZENIE ZESPOŁÓW, KTÓRE ZADAJĄ ODPOWIEDNIE PYTANIA I DOSTARCZAJĄ PRAWDZIWĄ WARTOŚĆ

Zrozumienie nauki o danych

Zacznę od zdefiniowania, kim jest analityk danych i czym się zajmuje. Następnie omówię różne typy oprogramowania i narzędzi używanych do zbierania, przeglądania i analizowania danych. Kiedy już poznasz różne typy oprogramowania i narzędzi używanych w nauce o danych, pokrótce omówię, jak ważne jest skupienie się na wiedzy organizacyjnej.

Definiowanie praktyki multidyscyplinarnej o wielu znaczeniach

Więc kim jest analityk danych? Naukowiec danych jest nieco trudniejszy do zdefiniowania niż inne typy naukowców. Jeśli jesteś politologiem lub klimatologiem, masz dyplom z uznanego programu. Termin „naukowiec danych" stał się szeroko stosowany, zanim „nauka o danych" stała się dobrze zdefiniowaną dyscypliną. Nawet teraz ludzie nazywający siebie naukowcami danych pochodzą z różnych dziedzin. Jako dyscyplina „nauka o danych" wciąż się rozwija. To trochę jak wczesna archeologia. Każdy mógł nazwać się archeologiem, pod warunkiem, że podniósł łopatę i zaczął kopać w poszukiwaniu artefaktów. W dzisiejszych czasach musisz przejść przez uniwersytet i spędzić lata na badaniach, aby zostać archeologiem. Podobnie jak wczesna archeologia, nauka o danych jest nadal bardziej praktyką niż dyscypliną. Jesteś naukowcem danych, jeśli pracujesz z danymi w sposób naukowy. To, czy zdecydujesz się nazywać siebie naukowcem danych, nadal zależy od Ciebie. Z pewnością są grupy ludzi, które lepiej pasują do tytułowego „naukowca danych" niż inne. Jeśli jesteś statystykiem lub analitykiem danych albo pracujesz w jednej z nauk biologicznych, prawdopodobnie możesz argumentować, że zawsze byłeś analitykiem danych. Niektórzy z pierwszych ludzi, którzy nazywali siebie naukowcami danych, byli w rzeczywistości matematykami; inni wywodzili się z systemów i inżynierii informacji, a niektórzy nawet z biznesu i finansów. Gdybyś pracował z liczbami i wiedział trochę o danych, mógłbyś łatwo nazwać siebie naukowcem od danych. Teraz, wraz ze wzrostem zapotrzebowania na analityków danych, będzie więcej ruchów w tworzeniu znormalizowanego zestawu umiejętności. Zaczynasz już to widzieć dzięki nowym programom na Berkeley, Syracuse University i Columbia University. Nowe programy studiów pozwolą firmom polegać na wspólnym zestawie umiejętności podczas zatrudniania. Na razie tak nie jest. W rzeczywistości nadal istnieje pewne niebezpieczeństwo, że naukowcy zajmujący się danymi będą postrzegani jako każdy, kto pracuje z danymi i może zaktualizować swój profil LinkedIn. Najlepszym sposobem myślenia o nauce o danych jest skupienie się na nauce, a nie na danych. W tym kontekście nauka posługuje się metodą naukową. Powinieneś przeprowadzić eksperymenty i zobaczyć wyniki, stosując podejście empiryczne. Empiryzm to jeden ze sposobów, w jaki naukowcy zdobywają wgląd i wiedzę, reagując na dane za pomocą eksperymentów i pytań. Analityk danych powinien używać tej umiejętności każdego dnia. Podejście empiryczne to połączenie wiedzy i praktyki. Prawdopodobnie używasz podejścia empirycznego i nie zdajesz sobie z tego sprawy. Jako coach i trener muszę sporo podróżować. Zwykle oznacza to, że znajduję się w różnych hotelach. Zawsze jestem zdumiony, ile różnych typów baterii i armatury jest na świecie. Jedyną rzeczą, z którą zawsze się borykam, jest radzenie sobie ze złożonością hotelowych pryszniców. Po chwili zdałem sobie sprawę, że najlepszym sposobem rozwiązania problemu jest podejście empiryczne. Najpierw muszę zgadnąć, jak włączyć prysznic. Zaczynam od zadania pytania empirycznego. Jak włączyć prysznic? Potem próbuję przeprowadzić eksperyment. Po naciśnięciu jednego przycisku woda napełnia wannę. Jeśli nacisnę inny, prysznic budzi się do życia. Po włączeniu wody muszę przekręcić różne pokrętła, aby sprawdzić, czy mogę kontrolować temperaturę. Jeśli przekręcę jedną gałkę, robi się za gorąco. Jeśli przekręcę inny, robi się za zimno. Więc zadaję pytania i ponownie oceniam, aż będę mógł sprawić, że woda będzie odpowiednia. Nie chciałbym stosować podejścia teoretycznego. Mogłem teoretyzować, jak sprawić, by woda była odpowiednia potem mogłem przekręcić pokrętło i wskoczyć pod prysznic.

Problem w tym, że najprawdopodobniej byłbym zmarznięty lub poparzony. Naukowcy zajmujący się danymi zawsze używają tego samego empirycznego podejścia. Zadają pytania dotyczące danych i dokonują niewielkich korekt, aby sprawdzić, czy mogą uzyskać wgląd. Przekręcają gałki i zadają ciekawsze pytania. Dla naszych celów skupiam się na analityku danych jako kimś, kto stosuje podejście empiryczne do uzyskania wglądu w dane i koncentruje się na metodzie naukowej. Kładziemy nacisk na naukę w „nauce o danych", a nie na danych.

Korzystanie ze statystyk i oprogramowania

Ponieważ nauka o danych jest nadal definiowana w praktyce, kładzie się dodatkowy nacisk na używanie typowych narzędzi i oprogramowania. Pamiętaj, że naukowcy zajmujący się danymi są jak pierwsi archeolodzy. Pomyśl więc o oprogramowaniu jako o pędzlach i kilofach, których potrzebujesz, aby dokonywać odkryć. Staraj się jednak nie skupiać się zbytnio na nauce wszystkich narzędzi, ponieważ to nie wszystko, co musisz wiedzieć. To metoda naukowa sprawia, że ktoś jest analitykiem danych, a nie narzędzia. Narzędzia potrzebne naukowcom zajmującym się danymi można podzielić na trzy ogólne kategorie:

• Oprogramowanie do przechowywania danych: są to arkusze kalkulacyjne, bazy danych i magazyny kluczy / wartości. Niektóre popularne programy obejmują Hadoop, Cassandra i PostgreSQL.

• Narzędzia używane do czyszczenia danych: Czyszczenie danych, ułatwia pracę z danymi, modyfikując lub zmieniając dane lub usuwając zduplikowane, nieprawidłowo sformatowane, niepoprawne lub niekompletne dane. Typowe narzędzia używane do czyszczenia danych to edytory tekstu, narzędzia skryptowe i języki programowania, takie jak Python i Scala.

• Pakiety statystyczne ułatwiające analizę danych: Najpopularniejsze to środowisko oprogramowania typu open source R, oprogramowanie do analizy predykcyjnej IBM SPSS oraz język programowania Python. Większość z nich obejmuje możliwość wizualizacji danych. Będzie to potrzebne do tworzenia ładnych wykresów i wykresów.

Przechowywanie danych

Najpierw przyjrzyjmy się narzędziom, które musisz znać, aby przechowywać dane. Jednym z terminów, które często słyszysz, są duże zbiory danych. Big data brzmi jak tytuł horroru z lat 60. Wyobrażasz sobie krzyczącą kobietę w kocich okularach, która została pochłonięta przez sączącą górę danych. Big data to zbiory danych, które są tak duże, że nie pasują do większości systemów zarządzania danymi. Niektórzy ludzie mylą naukę o danych i duże zbiory danych, ponieważ w tym samym czasie zostały one zahipnotyzowane i często mieszane razem. Pamiętaj jednak, że nauka o danych stosuje metodę naukową do danych. Nie oznacza to, że Twoje dane muszą być duże. Niemniej jednak jednym z najbardziej aktywnych obszarów nauki o danych są duże zbiory danych, a istnieje oprogramowanie zaprojektowane specjalnie do obsługi dużych zbiorów danych. Pakiet oprogramowania open source Hadoop jest obecnie najpopularniejszy. Hadoop używa rozproszonego systemu plików do przechowywania danych na grupie serwerów, zwykle nazywanej klastrem Hadoop. Klaster dystrybuuje również zadania na serwerach, dzięki czemu można również uruchamiać na nich aplikacje. Oznacza to, że można przechowywać petabajty danych na setkach lub nawet tysiącach serwerów i uruchamiać procesy na danych w klastrze. Dwa najczęściej uruchamiane procesy klastru Hadoop to MapReduce i Apache Spark. MapReduce współpracuje z danymi w partiach, a Spark może przetwarzać dane w czasie rzeczywistym.

Czyszczenie danych

Po zebraniu danych najprawdopodobniej zechcesz użyć niektórych narzędzi do wyczyszczenia danych, aby były bardziej użyteczne. Czyszczenie danych ułatwia pracę, modyfikując lub zmieniając dane lub usuwając zduplikowane, nieprawidłowo sformatowane, niepoprawne lub niekompletne dane. Wyobraź sobie, że zbierasz miliony tweetów swoich klientów, które mogą zawierać tekst, zdjęcia, a nawet filmy. Gromadząc te dane, możesz utworzyć skrypt, który podzieli wszystkie przychodzące tweety na różne typy (tekst, obrazy, filmy i inne). Umożliwiłoby to analizę każdej z tych grup oddzielnie i przy użyciu różnych parametrów. Jeśli często przeprowadzasz tę analizę, może lepiej jest utworzyć małą aplikację w języku Python do wykonywania operacji na klastrze zamiast skryptu, który robi to w miarę pojawiania się tweetów. Naukowcy zajmujący się danymi mogą poświęcić do 90% czasu na dostosowywanie i czyści swoje dane, aby były bardziej użyteczne, więc automatyzacja tego procesu ma kluczowe znaczenie na tym etapie.

Analiza danych

Ostatnia grupa narzędzi to te służące do analizy danych. Dwa najpopularniejsze to R i Python. R to statystyczny język programowania i środowisko oprogramowania, które umożliwia tworzenie połączeń i korelacji w danych, a następnie przedstawianie ich za pomocą wbudowanej wizualizacji danych języka R. Dzięki temu możesz mieć ładny diagram do swojego raportu. Na przykład wyobraź sobie, że Twoja firma potrzebuje raportu, aby sprawdzić, czy istnieje związek między jej pozytywną opinią a tym, czy ta opinia pojawia się w dzień czy w nocy. Jednym ze sposobów zebrania tych informacji jest przechwycenie danych Twittera w klastrze Hadoop, a następnie użycie czyszczenia danych w celu sklasyfikowania tweetów jako pozytywnych lub negatywnych. Następnie możesz użyć pakietu statystycznego, takiego jak R, aby stworzyć korelację między pozytywnymi i negatywnymi tweetami a czasem ich opublikowania i wydrukować raport przedstawiający wyniki w ładnym diagramie. Pamiętaj, że są to najpopularniejsze narzędzia. Jeśli jesteś częścią zespołu analityków danych, prawie na pewno słyszysz, że przynajmniej jeden z nich pojawia się w rozmowie. Istnieje wiele innych narzędzi, które automatyzują zbieranie, czyszczenie i analizę danych. Jest wiele organizacji, które wydają dużo pieniędzy, próbując wkupić się w tę przestrzeń. Pamiętaj, aby skupić się na analizie. Dane i narzędzia są po prostu środkiem do uzyskania lepszego wglądu. Wydawaj pieniądze ostrożnie w tej rozwijającej się dziedzinie.

Odkrywanie spostrzeżeń i tworzenie wiedzy

W ciągu ostatnich 20 lat większość organizacji skupiła się na zwiększeniu swojej wydajności operacyjnej, aby być szczuplejszą i bardziej elastyczną poprzez usprawnienie procesów biznesowych. Zadawali pytania operacyjne, takie jak „Jak możemy lepiej współpracować?" Nauka o danych jest inna; nie jest nastawiony na cel. Ma charakter eksploracyjny i wykorzystuje metodę naukową. Nie chodzi o to, jak dobrze działa organizacja; chodzi o zdobycie przydatnej wiedzy biznesowej. W przypadku nauki o danych zadajesz różnego rodzaju pytania, takie jak:

• Co wiemy o naszym kliencie?

• Jak możemy dostarczyć lepszy produkt?

• Dlaczego jesteśmy lepsi od naszych konkurentów?

Wszystkie te pytania wymagają wyższego poziomu organizacyjnego myślenia, a większość organizacji nie jest gotowych do zadawania tego typu pytań. Są skłonni do wyznaczania kamieni milowych i tworzenia budżetów. Nie zostali nagrodzeni za sceptycyzm lub dociekliwość. Wyobraź sobie, że jesteś na spotkaniu biznesowym i ktoś zadaje takie pytania. Dlaczego robimy to w ten sposób? Dlaczego myślisz, że to zadziała? Dlaczego to dobry pomysł? Są szanse, że osoba, która o to pyta, będzie

postrzegana jako irytująca. Zwykle ktoś odpowie w stylu „Czy nie przeczytałeś notatki?” Są to jednak umiejętności potrzebne do budowania wiedzy organizacyjnej. Oto pytania, których oczekujesz od swojego zespołu analityków danych. Mimo to większość ludzi w organizacjach koncentruje się na załatwianiu spraw. Pytania takie jak te, o których mowa, są postrzegane jako przeszkoda w posuwaniu się naprzód. Jednak jako organizacja zdobywasz wiedzę zadając ciekawe pytania. Kiedyś pracowałem dla strony internetowej, która łączyła potencjalnych nabywców samochodów z dealerami. Na stronie internetowej znajdowały się setki tagów informacyjnych, które wskazywały, czy klient znajduje się na nim, czy też klika w jego linki. Wszystkie te dane wpływały do klastra Hadoop i co tydzień były ich terabajty. Firma posiadała dane historyczne z lat wstecz. Wydali dużo pieniędzy, a nawet skonfigurowali działy, które skupią się na gromadzeniu i utrzymywaniu tych danych. Zbieranie danych było łatwe. Oprogramowanie, którego używali, było proste i łatwe do stworzenia. Najtrudniejsze było ustalenie, co zrobić z danymi. Wydaje się, że jest to częste wyzwanie dla wielu organizacji rozpoczynających działalność w dziedzinie nauki o danych. Organizacje te postrzegają to głównie jako wyzwanie operacyjne. Skupiają się na technicznej stronie danych. Chodzi o zbieranie danych, ponieważ są one stosunkowo tanie i łatwe do zrozumienia. Jest przyjazny dla spotkań i każdy może się z tym pogodzić. Utworzą nawet wiele klastrów lub jezior danych, aby zebrać dane z całej organizacji. To łatwa część. To, z czym zmagają się organizacje, to nauka. Nie są przyzwyczajeni do zadawania interesujących pytań i odpowiadania na nie. Pomyśl o eksperymentach i pytaniach, które mógłbyś zadać, gdybyś był naukowcem zajmującym się badaniem danych w tej witrynie samochodowej. Możesz przeprowadzić eksperyment, który zmienił kolory obrazów, aby sprawdzić, czy klienci byli bardziej skłonni do klikania obrazu, gdyby był czerwony, niebieski lub żółty. Jeśli z raportów wynika, że klienci są o 2% bardziej skłonni do kliknięcia samochodu, jeśli jest czerwony, organizacja mogłaby podzielić się tym z salonami samochodowymi, aby wygenerować nowe przychody. Możesz przeprowadzić eksperyment, aby sprawdzić, czy witryna zawiera zbyt wiele samochodów na stronie, zmniejszając liczbę wyświetlanych samochodów. Następnie możesz wygenerować raport, aby sprawdzić prawdopodobieństwo faktu, że klient kliknął łącze, zwiększył się wraz z mniejszą liczbą wyświetlanych samochodów. To jest rodzaj badań empirycznych, o których powinien myśleć naukowiec. Powinni przekręcać dane, zadawać interesujące pytania, przeprowadzać szybkie eksperymenty i tworzyć dobrze zaprojektowane raporty.