Złapanie pociągu eksploracji danych

Wybrałeś ekscytujący moment, aby zostać eksploratorem danych. Według niektórych szacunków każdego roku powstaje obecnie ponad 15 eksabajtów nowych danych. Ile to kosztuje? Jest naprawdę, absurdalnie dużo! Dlaczego to jest ważne? Większość organizacji ma dostęp tylko do malutkiego, niewielkiego ułamka tych danych i nie czerpią zbytniej wartości z tego, co mają. Dane mogą być cennym zasobem dla firm, instytucji rządowych i organizacji non-profit, ale nie chodzi o ilość. Większa ilość danych nie gwarantuje lepszego zrozumienia ani przewagi konkurencyjnej. W rzeczywistości, dobrze wykorzystana, odrobina odpowiednich danych zapewnia większą wartość niż jakakolwiek źle używana olbrzymia baza danych. Twoim zadaniem jako eksploratora danych jest maksymalne wykorzystanie posiadanych danych.

Prawdziwe informacje o eksploracji danych

Być może słyszałeś wiadomości lub reklamy sugerujące, że wszystko, czego potrzebujesz, aby cenne informacje wyskakiwały jak magia, to duża baza danych i najnowsze oprogramowanie. To kompletna bzdura. Eksploratorzy danych muszą pracować i myśleć, aby dokonać cennych odkryć. Być może słyszałeś, że aby uzyskać wyniki z bazy danych, musisz najpierw zatrudnić specjalną rasę ludzi, którzy mają prawie ponadludzką wiedzę na temat danych, ludzi znanych jako istoty bardzo drogie, prawie niemożliwe do znalezienia i absolutnie niezbędne do Twojego sukcesu. To też jest nonsens. Poszukiwacze danych to zwyczajni, zmotywowani ludzie, którzy uzupełniają swoją wiedzę biznesową o podstawy analizy danych. Eksploracja danych to nie magia ani sztuka. To rzemiosło, którego zwykli śmiertelnicy uczą się każdego dnia. Ty też możesz się o tym dowiedzieć.

Nie statystyki twojego profesora

Być może dawno temu wziąłeś udział w zajęciach ze statystyki i czułeś się przytłoczony naleganiem profesora na rygorystyczne metody. Zrelaksuj się. Musisz znaleźć informacje, które pomogą w codziennych decyzjach biznesowych, a wiele codziennych problemów biznesowych można rozwiązać za pomocą mniej formalnych metod analizy niż te, których nauczyłeś się w szkole. Daj sobie trochę luzu. Jak dajesz sobie luz? Tak właśnie wygląda eksploracja danych. Eksploracja danych to sposób, w jaki zwykli biznesmeni używają szeregu technik analizy danych, aby odkryć użyteczne informacje z danych i wykorzystać je w praktyce. Eksploratorzy danych używają narzędzi zaprojektowanych w celu przyspieszenia pracy. Nie przejmują się teorią i założeniami. Potwierdzają swoje odkrycia, testując. I rozumieją, że rzeczy się zmieniają, więc kiedy odkrycie, które działało jak urok wczoraj, dziś nie wytrzymuje, dostosowują się.

Wartość eksploracji danych

Menedżerowie biznesowi już mają biurka wypełnione raportami. Niektórzy mają dostęp do pulpitów nawigacyjnych komputera, które pozwalają im przeglądać swoje dane w niezliczonych segmentach i podsumowaniach. Czy eksploracja danych może naprawdę zwiększyć wartość? To może. Typowe raporty biznesowe zawierają podsumowania tego, co wydarzyło się w przeszłości. Nie oferują zbyt wiele, jeśli w ogóle, aby pomóc ci zrozumieć, dlaczego te rzeczy się wydarzyły lub jak możesz wpłynąć na to, co będzie dalej. Eksploracja danych jest inna. Oto przykłady informacji, które zostały odkryte podczas eksploracji danych:

* Sprzedawca odkrył, że rejestracja w programie lojalnościowym może posłużyć do określenia, którzy klienci najprawdopodobniej wydadzą dużo, a którzy spędzą trochę czasu, na podstawie tylko informacji zebranych podczas pierwszej wizyty klienta. Informacje te pozwoliły sprzedawcy skupić się na

inwestycjach marketingowych na tych, którzy dużo wydają, w celu maksymalizacji przychodów i obniżenia kosztów marketingu.

* Producent odkrył sekwencję zdarzeń poprzedzających przypadkowe uwolnienie materiałów toksycznych. Informacje te pozwoliły producentowi na utrzymanie obiektu w ruchu, jednocześnie zapobiegając niebezpiecznym wypadkom (chroniąc ludzi i środowisko) oraz unikając kar i innych kosztów.

* Firma ubezpieczeniowa odkryła, że jedno z jej biur było w stanie rozpatrywać niektóre typowe roszczenia szybciej niż inne o porównywalnej wielkości. Informacje te umożliwiły towarzystwu ubezpieczeniowemu określenie właściwego miejsca do poszukiwania najlepszych praktyk, które można by zastosować w całej organizacji w celu obniżenia kosztów i poprawy obsługi klienta.

Eksploracja danych pomaga zrozumieć, w jaki sposób elementy Twojej firmy są ze sobą powiązane. Zawiera wskazówki dotyczące działań, które możesz podjąć, aby Twoja firma działała sprawniej i generowała większe przychody. Może pomóc w określeniu, gdzie można obniżyć koszty bez szkody dla organizacji, a gdzie wydatki przynoszą największe zyski. Eksploracja danych zapewnia wartość, pomagając lepiej zrozumieć, jak działa Twoja firma.

Pracuję na to

Wiele osób ma nierealistyczne oczekiwania dotyczące eksploracji danych. To zrozumiałe, ponieważ większość ludzi uzyskuje informacje o eksploracji danych od osób, które nigdy tego nie robiły. Niektórzy ludzie oczekują, że eksploracja danych będzie tak łatwa, że będą musieli jedynie wprowadzić dane do odpowiedniego oprogramowania, a uporządkowane podsumowanie cennych informacji pojawi się automatycznie. Z drugiej strony, niektórzy spodziewają się, że eksploracja danych będzie tak trudna, że tylko ktoś z umiejętnościami programistycznymi na poziomie eksperckim i doktoratem w fizyce może sobie z tym poradzić. Niektórzy oczekują, że eksploracja danych przyniesie wspaniałe rezultaty, nawet jeśli eksplorator danych nie wie, co oznaczają dane. To wszystko są nierealistyczne oczekiwania, ale są zrozumiałe. Doniesienia prasowe, prezentacje sprzedażowe i źle poinformowani ludzie często rozpowszechniają poglądy na temat eksploracji danych, które są po prostu błędne. Jak ktoś ma wiedzieć, co jest rozsądne, a co jest szumem? Oto, co jest realistyczne: wielu początkujących eksploratorów danych uważa, że wystarczy kilka dni szkolenia i miesiąc ćwiczenia tego, czego się nauczyli (w niepełnym wymiarze godzin, nadal wykonując codzienne obowiązki), aby przygotować ich do uzyskiwania użytecznych, wartościowych wyników. Nie musisz mieć umysłu takiego jak Einstein, doktorat ani nawet umiejętności programowania. Musisz mieć podstawowe umiejętności obsługi komputera i wyczucie liczb. Trzeba też mieć cierpliwość i umiejętność metodycznej pracy. Eksploracja danych to ciężka praca. To nie jest trudne, jak wydobycie węgla lub operacja mózgu, ale jest trudne. Wymaga cierpliwości, organizacji i wysiłku.

Zaufaj danym lub swoim jelitom?

Czy intuicja może ci powiedzieć, co motywuje ludzi do kupowania, przekazywania darowizn lub podejmowania działań? Wiele osób uważa, że żadna analiza danych nie może prześcignąć ich intuicji przy podejmowaniu decyzji. Rzuciłem wyzwanie menedżerom biznesowym, aby przetestowali swoją intuicję. Pochodzili z różnych branż, małych i dużych firm i byli wśród nich zarówno młodzi, jak i doświadczeni menedżerowie. Każdy z nich obejrzał dziesięć par takich reklam:

* Dwie prawie identyczne reklamy, różniące się tylko tym, że jedna przedstawiała twarz kobiety, a druga mężczyznę. Która wygenerowała więcej potencjalnych klientów?

* Reklama z wieloma obrazami została zestawiona z reklamą, która miała tylko kilka. Która spowodowała więcej zakupów?

* Dwie reklamy miały tę samą kopię (tekst), ale różne układy. Która przyciągnęłaby więcej darowizn na cele charytatywne?

Niewielkie różnice w obrazach, układzie lub treści mogą znacząco wpłynąć na skuteczność reklamy. Testy próbek w tej grze w zgadywanie wykazały, że właściwy wybór może zwiększyć konwersje (działania ze strony klienta, takie jak kupowanie, przekazywanie darowizn lub proszenie o informacje) o 10%, 30%, a czasem więcej. W jednym przypadku lepsza reklama przyniosła 100 procent więcej konwersji niż alternatywa. Czy ktokolwiek mógłby stwierdzić, po prostu patrząc, które alternatywy byłyby najlepsze? Nie. Żaden z menedżerów nie był skuteczny w wyborze najlepszych reklam. Rzucanie monetą działało równie dobrze. Jeśli chcesz podejmować dobre decyzje biznesowe, potrzebujesz danych. Użyj mózgu, a nie jelit!

Robią to, co robią eksploratorzy danych

Jeśli myślisz o danych jako o surowcu, a informacje, które możesz uzyskać z danych, jako o czymś cennym i względnie wyrafinowanym, proces wydobywania informacji można porównać do wydobywania metalu z rudy lub klejnotów z ziemi. Tak powstał termin eksploracja danych. Czy słowa eksplorator danych wywołują w pamięci obraz szorstkiego pracownika w kombinezonie? To nie jest tak dalekie od celu. Oczywiście nic nie jest fizycznie brudne w eksploracji danych, ale kopacze danych robią problemy i brudzą się danymi. W eksploracji danych chodzi o władzę dla ludzi, dając możliwość analizy danych zwykłym biznesmenom.

Koncentrując się na biznesie

Eksploratorzy danych nie tylko rozważają dane bez celu, mając nadzieję na znalezienie czegoś interesującego, a projekt eksploracji danych zaczyna się od konkretnego problemu biznesowego i celu, któremu należy sprostać. Jako eksplorator danych prawdopodobnie nie będziesz mieć uprawnień do podejmowania ostatecznych decyzji biznesowych, dlatego ważne jest, aby dostosować swoją pracę do potrzeb decydentów. Musisz zrozumieć ich problemy, potrzeby i preferencje oraz skupić się na dostarczaniu informacji wspierających dobre decyzje biznesowe. Twoja własna wiedza biznesowa jest bardzo ważna. Kierownictwo nie będzie siedzieć obok Ciebie podczas pracy i przekazywać informacji zwrotnych na temat związku Twoich odkryć z ich obawami. Podczas pracy musisz korzystać z własnego doświadczenia i bystrości, aby ocenić to samodzielnie. Możesz nawet być zaznajomiony z aspektami działalności, którymi nie jest dyrektor, i być w stanie przedstawić nowe spojrzenie na problem biznesowy oraz możliwe przyczyny i środki zaradcze.

Zrozumienie, jak osoby poszukujące danych spędzają czas

Byłoby wspaniale, gdyby eksploratorzy danych mogli spędzić cały dzień na dokonywaniu odkryć zmieniających życie, tworzeniu wartościowych modeli i integrowaniu ich z codziennym biznesem. Ale to tak, jakby powiedzieć, że byłoby wspaniale, gdyby sportowcy mogli spędzić cały dzień na wygrywaniu turniejów. Przygotowanie do tych chwil triumfu wymaga wielu przygotowań. Tak więc, podobnie jak sportowcy, eksploratorzy danych spędzają dużo czasu na przygotowaniach. Największa część idzie na przygotowanie danych.

Poznanie procesu eksploracji danych

Dobry proces pracy pomaga maksymalnie wykorzystać czas, dane i wszystkie inne zasoby. Poznasz najpopularniejszy proces przetwarzania danych, CRISP-DM. Jest to sześciofazowy cykl odkrywania i

działania stworzony przez konsorcjum eksploratorów danych z wielu branż i otwarty standard, z którego każdy może skorzystać. Fazy procesu CRISP-DM to

1. Zrozumienie biznesu

2. Zrozumienie danych

3. Przygotowanie danych

4. Modelowanie

5. Ocena

6. Wdrożenie (używanie modeli w codziennym biznesie)

Każda faza ma równe znaczenie dla jakości wyników i wartości dla firmy. Ale pod względem wymaganego czasu dominuje przygotowywanie danych. Przygotowanie danych rutynowo zajmuje więcej czasu niż wszystkie inne fazy procesu eksploracji danych łącznie.

Tworzenie modeli

Kiedy cele są zrozumiałe, a dane oczyszczone i gotowe do użycia, możesz skupić się na budowaniu modeli predykcyjnych. Modele robią to, czego nie potrafią raporty; dostarczają informacji, które wspierają działanie. Raport może powiedzieć, że sprzedaż spadła. Może rozbić sprzedaż według regionu, produktu i kanału, dzięki czemu wiesz, gdzie spadła sprzedaż i czy spadki te były powszechne lub dotyczyły tylko niektórych obszarów. Ale nie dają żadnych wskazówek, dlaczego sprzedaż spadła ani jakie działania mogą pomóc ożywić firmę.

Modele pomagają zrozumieć czynniki wpływające na sprzedaż, działania, które mają tendencję do zwiększania lub zmniejszania sprzedaży, oraz strategie i taktyki, które zapewniają płynne działanie Twojej firmy. To ekscytujące, prawda? Może dlatego większość eksploratorów danych uważa modelowanie za fajną część pracy.

Zrozumienie modeli matematycznych

Modele matematyczne mają kluczowe znaczenie dla eksploracji danych, ale czym one są? Co robią, jak działają i jak powstają?

Model matematyczny jest prostym i prostym równaniem lub zbiorem równań, które opisują związek między dwiema lub więcej rzeczami. Takie równania są skrótem dla teorii o funkcjonowaniu przyrody i społeczeństwa. Teoria może być poparta pokaźną ilością dowodów lub może być tylko szalonym przypuszczeniem. Język matematyki jest taki sam w obu przypadkach. Terminy takie jak model predykcyjny, model statystyczny lub model liniowy odnoszą się do określonych typów modeli matematycznych, nazw odzwierciedlających zamierzone zastosowanie, formę lub metodę wyprowadzenia określonego modelu. Te trzy przykłady to tylko kilka z wielu takich terminów. Kiedy model jest wymieniany w otoczeniu biznesowym, najprawdopodobniej jest to model używany do prognozowania. Modele są używane między innymi do przewidywania cen akcji, sprzedaży produktów i stóp bezrobocia. Prognozy te mogą być dokładne lub nie, ale dla dowolnego zestawu wartości (znane czynniki, takie jak te nazywane są zmiennymi niezależnymi lub wejściami), uwzględniono w modelu, znajdziesz dobrze zdefiniowaną prognozę (nazywaną również zmienną zależną, wyjściem lub wynikiem). Modele matematyczne są wykorzystywane również do innych celów w biznesie, takich jak opis mechanizmów roboczych, które kierują określonym procesem. W eksploracji danych tworzymy modele, znajdując wzorce w danych za pomocą uczenia maszynowego lub metod statystycznych. Osoby zajmujące się eksploracją danych nie przestrzegają tego samego rygorystycznego podejścia,

które stosują klasyczni statystycy, ale wszystkie nasze modele pochodzą z rzeczywistych danych i spójnych matematycznych technik modelowania. Wszystkie modele przetwarzania danych są poparte materiałami dowodowymi. Po co używać modeli matematycznych? Nie można opisać tych samych relacji używając słów? Jest to możliwe, ale stosowanie równań ma pewne zalety. Obejmują one

* Wygodę: w porównaniu z równoważnymi opisami zawartymi w zdaniach, równania są krótkie. Symbolika matematyczna rozwinęła się specjalnie w celu przedstawienia związków matematycznych; języki takie jak angielski nie.

* Jasność: Równania zwięźle przekazują pomysły i są jednoznaczne. Nie podlegają różnym interpretacjom ze względu na kulturę, a symbolika matematyki jest rodzajem powszechnego języka używanego na całym świecie.

* Spójność: ponieważ reprezentacje matematyczne są jednoznaczne, implikacje każdej konkretnej sytuacji są jasno określone przez model matematyczny.

Wprowadzanie informacji w czyn

Model zapewnia wartość tylko wtedy, gdy jest używany w biznesie. Prognozy modelu mogą wspierać podejmowanie decyzji na różne sposoby.

* Włącz prognozy do raportu lub prezentacji do wykorzystania przy podjęciu konkretnej decyzji.

* Zintegruj model z systemem operacyjnym (takim jak system obsługi klienta), aby zapewnić prognozy w czasie rzeczywistym do codziennego użytku. (Na przykład możesz oznaczyć roszczenia ubezpieczeniowe do natychmiastowej płatności, natychmiastowej odmowy lub dalszego dochodzenia).

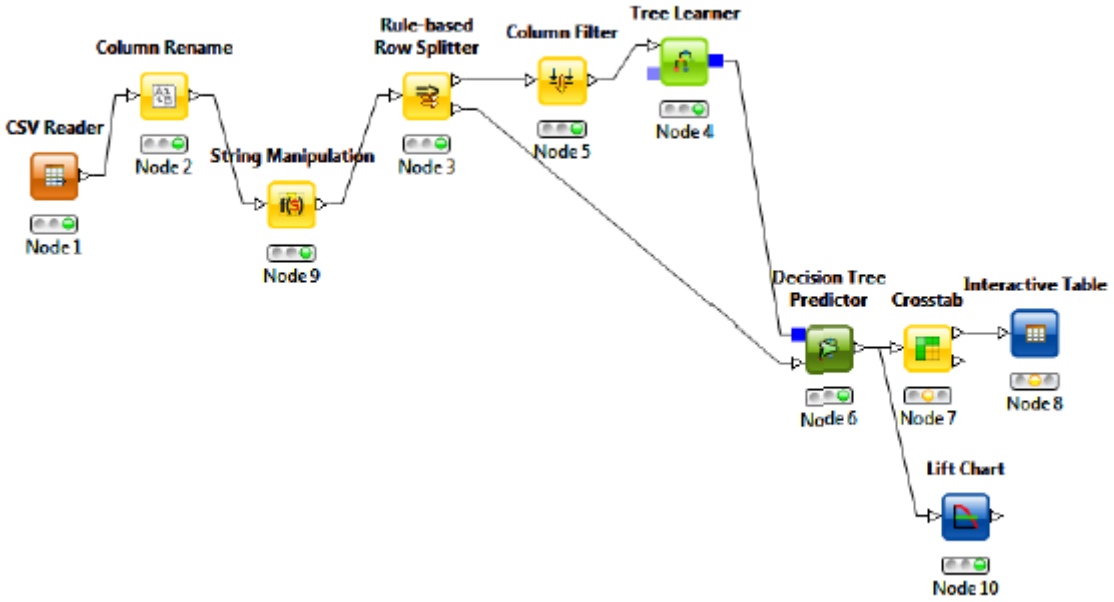
* Użyj modelu do prognozowania partii. (Na przykład możesz ocenić wewnętrzną listę klientów, aby zdecydować, którzy klienci powinni otrzymać określoną ofertę).

Narzędzia i metody wykrywania

Kopacze danych pracują szybko. Aby uzyskać prędkość, musisz użyć odpowiednich narzędzi i odkryć sztuczki związane z handlem.

Programowanie wizualne

Twoim najlepszym narzędziem do eksploracji danych jest mózg z odrobiną wiedzy. Drugim najlepszym narzędziem jest aplikacja do eksploracji danych z wizualnym interfejsem programowania. W przypadku programowania wizualnego etapy procesu pracy są reprezentowane przez małe obrazy, które organizujesz na ekranie, aby stworzyć obraz przepływu i logiki Twojej pracy. Programowanie wizualne ułatwia zobaczenie, co robisz w kilku krokach, niż w przypadku poleceń (programowanie) lub konwencjonalnych menu. W tym przykładzie możesz zobaczyć proces pracy w głównym obszarze aplikacji do przetwarzania danych. Wokół niego znajdują się menu ostatnich projektów, narzędzia do funkcji przetwarzania danych, przeglądarka ułatwiająca nawigację po złożonych procesach oraz dziennik. Te szczegóły różnią się nieco w zależności od produktu. Przyjrzyj się dokładniej procesowi



Chociaż dopiero zaczynasz swoją misję, aby zostać eksploratorem danych, prawdopodobnie możesz zrozumieć wiele z tego, co się dzieje, po prostu patrząc na ten diagram, w tym:

* Możesz zobaczyć CSV Reader. Jeśli wiesz, że .csv (wartości rozdzielane przecinkami), prawdopodobnie już wiesz, że jest to import danych. (I to jest pierwszy krok; do zrobienia czegokolwiek innego potrzebujesz danych).

* Następnie zobaczysz narzędzia wyraźnie oznaczone funkcjami, takimi jak Zmiana nazwy kolumny i Manipulacja ciągami. To są kroki przygotowania danych.

* Tree Learner może być tajemniczy, jeśli dopiero zaczynasz modelować, ale to narzędzie tworzy model drzewa decyzyjnego z podzbioru danych.

* Na koniec zastosuj model do danych, które były przechowywane oddzielnie na potrzeby testów, i wykonaj kilka technik oceny

Praca szybka i brudna

Programowanie wizualne pomaga eksploratorom danych w szybkiej pracy. O wiele łatwiej i szybciej zaplanować proces pracy przy użyciu tych małych obrazów, niż programując od podstaw. Łatwo jest zobaczyć, co robisz, gdy widzisz coś w rodzaju mapy wielu kroków naraz, więc programowanie wizualne jest również szybsze niż przy użyciu konwencjonalnego oprogramowania sterowanego menu. Kopacze danych mają inny ważny sposób na szybką pracę. Eksploratorzy danych nie zawsze przejmują się każdym szczegółem teorii i założeń matematycznych. Dobra wiadomość jest taka, że brak zamieszania pozwala szybciej budować modele. Zła wiadomość jest taka, że jeśli nie będziesz się przejmować teorią i założeniami, Twój model może nie być dobry. Eksploratorzy danych łamią reguły statystyki, ponieważ eksploratorzy danych wybierają modele na podstawie eksperymentu, a nie na podstawie teorii i założeń statystycznych. Ale górnicy danych również łamią własne zasady, ponieważ niektórzy eksploratorzy danych mają statystyki wiedzy i starają się rozważać założenia. (Niewiele wiadomo, że standardowy proces eksploracji danych CRISP-DM obejmuje etap raportowania założeń).

Testowanie, testowanie i jeszcze raz testowanie

Jako eksplorator danych nie będziesz w stanie obronić modeli, które tworzysz w oparciu o teorię statystyczną, ponieważ Twoje metody pracy nie uwzględniają teorii Korzystasz z danych, które możesz uzyskać, i na pewno masz pewne problemy, które nie są To nie jest zgodne z teorią stojącą za modelem, którego używasz: * Możesz nie mieć wystarczającej wiedzy statystycznej, aby formułować teoretyczne argumenty. Ale to w porządku. Eksploratorzy danych oceniają swoje modele głównie poprzez testowanie, testowanie i jeszcze raz testowanie. Wiele narzędzi do modelowania przeprowadza wewnętrzne testy podczas tworzenia modeli. Odkładasz dane na bok, aby przetestować model po jego utworzeniu. Będziesz testować w terenie, gdy tylko będzie to możliwe. Po wdrożeniu będziesz monitorować wydajność modelu. Kiedy jesteś eksploratorem danych, testy nigdy się nie kończą!