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).

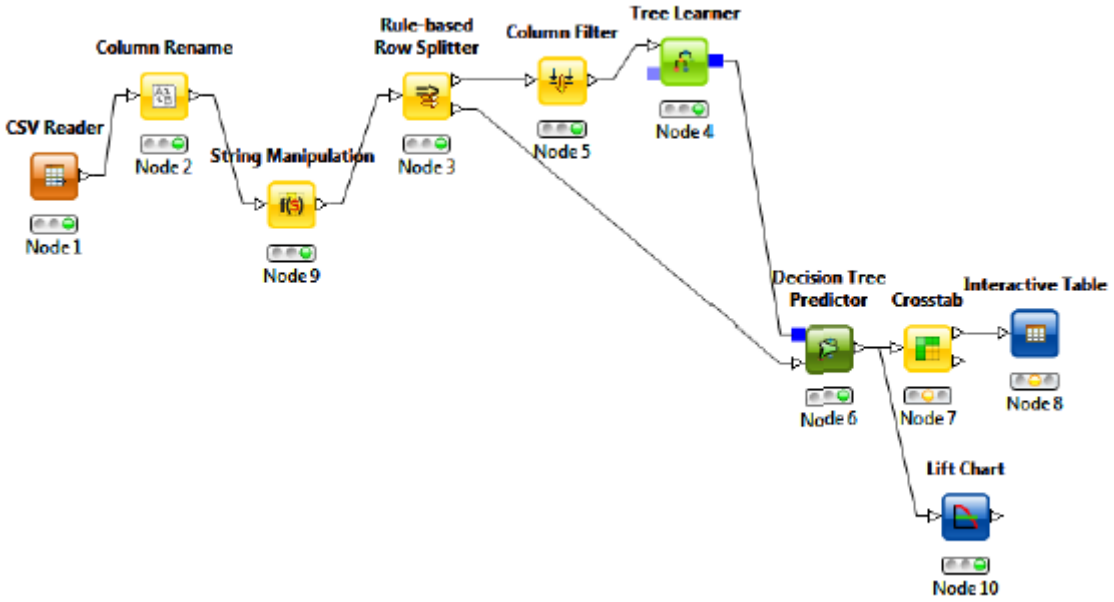
* Użyj modelu do prognozowania partii. (Na przykład możesz ocenić wewnętrzną listę klientów, aby zdecydować, którzy klienci powinni otrzymać określoną ofertę).

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!

Dzień z życia jako eksplorator danych

Dzień dobry! Witamy w zwykłym dniu w Twojej karierze w eksploracji danych. Dzisiaj spotkasz się z innymi członkami zespołu eksploracji danych, aby omówić trwający już projekt. Ekspert merytoryczny pomoże Ci zrozumieć cele biznesowe projektu i wyjaśnić, dlaczego są one ważne dla Twojej organizacji, aby upewnić się, że wszyscy dążą do tego samego celu. Inny członek zespołu rozpoczął już zbieranie danych i przygotowywanie ich do eksploracji i modelowania. (Masz szczęście, że masz silny zespół!) Po spotkaniu zaczniesz od praktycznej pracy z danymi. Poznasz dane. Chociaż część prac związanych z przygotowaniem danych została wykonana, nadal będziesz mieć więcej do zrobienia, zanim zaczniesz budować modele predykcyjne. Eksploratorzy danych spędzają dużo czasu na przygotowywaniu danych! Później dzisiaj zaczniesz przeglądać dane. Być może zaczniesz budować model, który będziesz udoskonalać i ulepszać w nadchodzących dniach. Oczywiście na bieżąco dokumentujesz całą swoją pracę. To tylko kolejny dzień z życia eksploratora danych. Ta część pokazuje, jak to się robi.

Właściwe rozpoczęcie dnia wolnego

Dobrze się wyspałeś, a teraz wstajesz wcześnie, żeby trochę poćwiczyć i zjeść dobre śniadanie. Ma to niewiele wspólnego z eksploracją danych, ale jest to dobry sposób na rozpoczęcie dnia. W drodze do pracy zastanów się nad tym: pomyślna eksploracja danych to praca zespołowa. Nikt nie posiada całej wiedzy, wszystkich zasobów ani wszystkich uprawnień wymaganych do przeprowadzenia typowego projektu eksploracji danych i wprowadzenia jego wyników w życie. Do załatwienia sprawy potrzebny jest cały zespół. Twoi współpracownicy mogą być czarującymi ludźmi z najlepszymi umiejętnościami i najczystszymi motywacjami lub mogą mieć trudne osobowości i ukryte plany, ale obiecujesz rozpocząć dzień eksploracji danych od razu, traktując każdą osobę z cierpliwością i słuchaniem do wszystkich z szacunkiem i wyjaśniania się w sposób zrozumiały dla innych członków zespołu.

Spotkanie z zespołem

Dzisiaj spotkasz się ze swoim zespołem: Virginia, źródło wiedzy biznesowej, oraz Matt, ekspert ds. Pozyskiwania danych i programowania. To czarujący ludzie z najlepszymi umiejętnościami i najczystszymi motywacjami. Virginia będzie łącznikiem z klientami i wyjaśni cele biznesowe Twojej organizacji. Wyjaśni problem biznesowy i jego wpływ na organizację. Potrafi wskazać czynniki, które mogą być ważne. Może też odpowiedzieć na większość Twoich pytań dotyczących funkcjonowania firmy lub pomóc Ci dotrzeć do kogoś, kto to potrafi. Matt bardzo dobrze zna dane, których będziesz używać. Przygotował dla Ciebie zbiory danych, pochodzące ze źródeł publicznych i dalej rozwijane za pomocą kilku własnych obliczeń. Ułatwia to pracę i oszczędza dużo czasu. Będzie osobą, na której możesz polegać, jeśli chodzi o informacje o źródłach danych, dokumentację oraz szczegółowe informacje o tym, jak i dlaczego dokonał restrukturyzacji danych. Virginia i Matt też na tobie polegają. Matt potrzebuje twoich danych wejściowych, aby zrozumieć, które dane są najbardziej przydatne do eksploracji danych i jak organizować dane do użytku. Chce, abyś wskazał wszelkie błędy (lub podejrzewane błędy) w danych, aby mógł zbadać i rozwiązać wszelkie problemy. Inni zależą od dostarczonych przez niego informacji - nie tylko od Ciebie - więc nie pozwól, aby błędy się przeciągały! Virginia potrzebuje twojego wkładu w zakresie rodzajów analiz, które możesz dostarczyć, jasnych informacji o twoich wynikach i dobrej dokumentacji twojej pracy.

Eksploracja z celem

Powiedzenie, że eksploratorzy danych eksplorują dane w poszukiwaniu cennych wzorców, może stworzyć mentalny obraz, który jest nieco magiczny lub tajemniczy. Zamierzasz zastąpić ten obraz takim, który jest znacznie bardziej praktyczny i przystępny. Eksploracja danych nie jest magiczna, a jej celem jest stopniowe eliminowanie tajemnic z Twojej firmy. Możesz zwiedzić centrum handlowe lub

urocze miasteczko tylko po to, by się rozejrzeć, ale eksplorując dane, odkrywasz je w określonym celu. Pierwszą rzeczą, którą zrobisz w każdym projekcie eksploracji danych, będzie jasne zrozumienie tego celu. Podczas pracy z danymi często powracasz do swoich celów i zastanawiasz się, czy i w jaki sposób informacje, które znajdziesz w danych, je wspierają. Od czasu do czasu będziesz miał do czynienia z pokusą, pokusą poświęcenia czasu na badanie pewnych wzorców w danych, które nie są bezpośrednio związane z wyznaczonymi celami. Podobnie jak w przypadku innych pokus, możesz sobie pozwolić na odrobinę czasu, jeśli masz trochę czasu i zasobów do stracenia, ale Twoim głównym priorytetem zawsze musi być osiągnięcie celów biznesowych ustalonych na początku projektu.

Przedstaw prawdziwych ludzi w swoim zespole projektowym

Projekt opisany tu jest prawdziwy pod każdym względem. Dotyczy rzeczywistego problemu biznesowego, który ma wpływ na ludzi i firmy w prawdziwej społeczności. Dane są prawdziwe. A ludzie w twoim zespole, Virginia i Matt, też są prawdziwi. Virginia Carlson jest strategiem danych. Jest głównym badaczem zajmującym się integracją danych w Impact Planning Council (www.impactinc. Org / impact-planowanie-council), Milwaukee w stanie Wisconsin, organizacja zajmująca się poprawą życia członków społeczności i profesor nadzwyczajny na Uniwersytecie Wisconsin w Milwaukee. Jest ekspertem w zbieraniu i wykorzystywaniu danych do wspierania inicjatyw sektora społecznego. Kierowała znaczącymi organizacjami i projektami zajmującymi się badaniami gospodarczymi, a także jest współautorką Podręcznika konkursów aplikacji obywatelskich, przewodnika po planowaniu, organizowaniu i rozwiązywaniu problemów. Matt Schumwinger jest niezależnym analitykiem danych. Jest właścicielem Big Lake Data (http://biglakedata.com), firmy usługowej, która pomaga swoim klientom wizualizować, analizować i prezentować informacje ilościowe. Matt studiował ekonomię pracy i stosunki pracy na Cornell University i poświęcił większość swojej kariery na poprawę dobrobytu Amerykanów poprzez organizowanie pracowników o niskich płacach w całych Stanach Zjednoczonych. Virginia i Matt mają wspólne zainteresowania poprawą życia obywateli publicznych i wykorzystaniem danych do wspierania społeczności. W tym kontekście pracowali razem jako zespół, łącząc uzupełniające się talenty i doświadczenia, aby pracować nad wspólnymi celami. Twój projekt jest przedłużeniem prawdziwej pracy Virginii i Matta. Przykład opiera się na projektach, które wykonali w przeszłości, aby stworzyć coś zupełnie nowego. Jako członkowie Twojego zespołu zapewniają oni wiedzę w zakresie rozwoju społeczności i zarządzania danymi. Każdy z nich jest zdolny do eksploracji danych, ale mają do wykonania swoje własne zadania! Poza tym wiesz rzeczy, których nie wiedzą, i masz umiejętności, których nie mają. Chcą, abyś wniósł do zespołu swoją własną mieszankę wiedzy i doświadczenia oraz wzbogacił wiedzę wszystkich. Razem z Virginią i Mattem możesz dokonywać odkryć, które pomogą budować silniejsze społeczności.

Strukturyzacja czasu z odpowiednim procesem

Wielu potencjalnych poszukiwaczy danych pobrało i zainstalowało oprogramowanie, uruchomiło je i zastanawiało się: „Co teraz?" To ci się dzisiaj nie przydarzy. Dowiesz się, jak wykorzystać swój czas, ponieważ wykorzystasz podstawy, które górnicy danych z setek organizacji wykonali dla Ciebie, opracowując i publikując modelowy proces eksploracji danych. Międzybranżowy standardowy proces eksploracji danych (CRISP-DM), otwarty standard, zawiera wytyczne dotyczące organizacji i dokumentowania pracy. Jest to sześciofazowy proces, który zaczyna się od zdefiniowania celów biznesowych, a kończy na zintegrowaniu wyników z rutynową działalnością biznesową i przejrzeniu swojej pracy pod kątem kolejnych kroków i możliwości poprawy. Tam zobaczysz, że każda z sześciu faz wymaga kilku zdefiniowanych zadań i że każde zadanie ma jeden lub więcej elementów dostarczanych, którymi mogą być raporty, prezentacje, dane lub modele. W tym rozdziale nie zobaczysz wszystkich tych szczegółów, ale dotkniesz każdej z sześciu głównych faz procesu CRISP-DM.

Zrozumienie celów biznesowych

Virginia wyjaśnia najnowszy projekt zespołu zajmującego się eksploracją danych: pomoc lokalnej radzie ds. planowania. Jej misją jest promowanie dobrobytu gospodarczego poprzez zachęcanie do użytkowania gruntów, które czyni społeczność atrakcyjną dla przedsiębiorstw i mieszkańców. Kluczową częścią jej pracy jest zatrzymywanie i przyciąganie firm, które zatrudniają lokalnych mieszkańców i oferują dobre wynagrodzenie. Zadaniem Twojego zespołu jest dostarczanie nowych i istotnych informacji, opartych na danych i analizach, na podstawie których rada planistyczna może zdecydować, gdzie skoncentrować wysiłki, aby jak najlepiej wykorzystać swoje zasoby. Virginia i Matt byli już zaangażowani w projekty wspierające te cele. We wcześniejszych projektach opracowali analizy czynników, które mają wpływ na użytkowanie gruntów, oraz udostępnili informacje poprzez konsultacje i prezentacje, pisemne raporty i interaktywne mapy. Rada rozumie, że najlepsza okazja do wpłynięcia na użytkowanie określonej działki ma miejsce, gdy ziemia ma zamiar zmienić właściciela. Ale właściciele gruntów nie zamierzają po prostu wpaść i ogłosić swoich zamiarów sprzedaży. Wiele znaczących transakcji dotyczących nieruchomości jest zawieranych po cichu, więc rada może nie wiedzieć nic o tej okazji, dopóki nieruchomość nie zostanie sprzedana. Tak więc celem biznesowym rady jest zidentyfikowanie działek, które mają zmienić właściciela, i zrobienie tego na tyle wcześnie, aby wpłynąć na sposób użytkowania gruntu. W jaki sposób rada zdecyduje, czy uda jej się osiągnąć ten cel? Na tym etapie rada ma tylko nieformalne (i nie do końca spójne) sposoby przewidywania, które działki mają wkrótce zmienić właściciela. Podane kryteria sukcesu wymagają po prostu ustanowienia procesu przewidywania zmiany własności w spójny sposób. (Przyszłe projekty będą opierać się na tym celu i będą miały ilościowe kryteria sukcesu). Kiedy przedstawiany jest cel, zawsze omawiaj i dokumentuj kryteria sukcesu od samego początku. Chociaż możesz być odpowiedzialny tylko za wąską część pracy potrzebnej do osiągnięcia celu biznesowego, zrozumienie, w jaki sposób zostaną ocenione ostateczne wyniki, pomoże ci zrozumieć najlepsze sposoby przyczynienia się do sukcesu projektu. Te kryteria sukcesu mogą wydawać się proste, ale masz wątpliwości. Zadajesz takie pytania:

* Czy rada spodziewa się, że tylko jeden model będzie działał dla wszystkich rodzajów nieruchomości? Przemysłowe, komercyjne, jednorodzinne, wielorodzinne itd. - nierealistyczne jest myślenie, że znajdziesz jedno wielkie równanie, które rozwiąże je wszystkie.

* Ile istnieje typów nieruchomości? Możesz mieć dziesiątki.

* Czy rada jest w równym stopniu zainteresowana wszystkimi nieruchomościami? Można by pomyśleć, że najważniejsze byłyby duże, przemysłowe paczki.

* Jakie rodzaje nieruchomości są najważniejsze dla rady? Możesz chcieć naciskać na modelowanie tylko jednej lub dwóch ważnych kategorii w pierwszej rundzie.

Zawsze pytaj o ostatnie wpadki. Niewypowiedziane cele często obejmują unikanie powtarzania czegoś, co po prostu poszło nie tak. Zadawanie pytań pomaga oczywiście uzyskać więcej informacji, ale pytania mają większe znaczenie. Pomagają innym członkom zespołu (w tym kierownictwu, jeśli masz okazję się z nimi spotkać) uświadomić sobie, czego brakuje, co będzie wyzwaniem i co jest o wiele bardziej skomplikowane, niż myśleli! Zadając dociekliwe pytania w fazie zrozumienia biznesu, pomagasz każdemu wyjaśnić myślenie, zdefiniować rozsądne cele i ustalić realistyczne oczekiwania. Po krótkiej dyskusji uzgodniono (i udokumentowano!), że celem biznesowym tego projektu będzie wykazanie wykonalności modelowania w celu przewidywania zmiany własności gruntów - węższy i mniej ambitny cel niż pierwotnie sugerowany. Nie oczekuje się od Ciebie stworzenia megamodelu (nie, to nie jest termin techniczny) obejmującego wszystkie rodzaje nieruchomości. Jeśli rada uzna, że choćby jeden czynnik ma wartość predykcyjną dla transferów własności, będzie to zadowalające dla pierwszej rundy. W pierwszym badaniu nie zostaną określone kryteria ilościowe dotyczące wydajności modelu. Celem

jest po prostu wykazanie, że istnieje potencjał do opracowania użytecznego modelu do przewidywania zmian własności nieruchomości przy użyciu dostępnych danych. Cele biznesowe są określane przez klienta (zewnętrznego lub wewnętrznego), a nie eksploratora danych. Jeśli Ty i Twój zespół macie wątpliwości co do konkretnego celu, nie zmieniajcie go samodzielnie. Klienci tego nie zaakceptują! Zamiast tego rozpocznij dyskusję z klientem, wyjaśnij swoje obawy i uzgodnij rozsądne cele biznesowe projektu. Na podstawie celów biznesowych definiujesz cele eksploracji danych. Ponieważ celem biznesowym jest wykazanie wykonalności modelowania w celu przewidywania zmian własności gruntów, należy wyznaczyć cel eksploracji danych polegający na stworzeniu podstawowego modelu predykcyjnego zmiany własności nieruchomości. Ponieważ nie masz konkretnych liczb dotyczących wydajności obecnego, nieformalnego podejścia. Aby przewidzieć zmiany własności, będziesz po prostu dążyć do wykazania, że co najmniej jedna zmienna ma wymierną wartość do prognozowania. (Podobnie jak w przypadku celów biznesowych, przyszłe projekty będą opierać się na tym, a na tym etapie ustalisz bardziej szczegółowe ilościowe kryteria sukcesu). Ukończysz tę fazę procesu eksploracji danych, opisując swoje działania krok po kroku plan zakończenia pracy (w tym harmonogram i szczegóły zasobów wymaganych na każdym etapie) oraz wstępną ocenę odpowiednich narzędzi i technik dla projektu.

Zrozumienie Twoich danych

Na etapie zrozumienia danych najpierw zbierzesz i szeroko opiszesz swoje dane. Nie musisz zaczynać od zera, aby zbierać dane, ponieważ Matt zebrał już kilka zbiorów danych, których możesz używać. Zostały zaczerpnięte z danych używanych we wcześniejszych projektach i wyprowadził dodatkowe pola, których będziesz potrzebować. Następnie przeanalizujesz dane bardziej szczegółowo, eksplorując dane po jednej zmiennej (polu) na raz, sprawdzając zgodność z oczekiwaniami i wszelkie oczywiste oznaki problemów z jakością danych. Zaczynasz przeglądać dane, robiąc notatki do raportu w trakcie pracy.

Opisywanie danych

Dane znajdują się w kilku plikach tekstowych, każdy w formacie wartości rozdzielanych przecinkami (.csv). Pliki są dość duże, od 50 do 100 MB, ale nie są zbyt duże, aby można je było obsługiwać za pomocą dostępnego komputera i oprogramowania. Zapisujesz nazwę i rozmiar każdego pliku. Twoim pierwszym problemem jest zidentyfikowanie zmiennych w każdym pliku i potwierdzenie, że masz odpowiednią dokumentację dla każdego z nich. Kilka plików zawiera historyczne rejestry własności publicznej; obszerny dokument definiuje te zmienne. Otrzymałeś również uwagi wyjaśniające, w jaki sposób powstały zmienne pochodne. Przeglądasz każdą zmienną w danych, porównując nazwy zmiennych z informacjami w dokumentacji. Odnotowujesz ustalenia dotyczące danych i dokumentacji, w tym następujące:

* Większość pól jest zgodna z posiadaną dokumentacją.

* Niektóre pola w plikach danych rekordów właściwości nie są wyjaśnione w dokumentacji.

* Niektóre pola opisane w dokumentacji rekordów właściwości nie pojawiają się w danych.

* Jeden z plików danych rekordów właściwości zawiera o wiele więcej pól niż inne, a te pola nie są wyjaśnione w dokumentacji.

Piszesz szczegółowe notatki o każdym pliku i każdej zmiennej. Wykorzystując swoje notatki jako punkt odniesienia, szukasz informacji, które pozwolą rozwiązać te rozbieżności. Znajdziesz to

* Kilka pól w danych ze źródeł publicznych po prostu nie pasuje do dostarczonej dokumentacji (dane publiczne nie zawsze są idealnymi danymi).

* Dostępne są dodatkowe uwagi wyjaśniające, w jaki sposób utworzono niektóre z pól pochodnych.

* Niektóre z nieudokumentowanych danych uzyskano poprzez skrobanie stron internetowych (przy użyciu specjalistycznego oprogramowania do automatycznego wyodrębniania informacji ze stron internetowych) i nie można znaleźć dla nich żadnej niezawodnej dokumentacji.

Aktualizujesz swoje notatki o danych, poprawiając je o dodatkową dokumentację. Zwracasz uwagę, które zmienne są nadal nieudokumentowane. Chociaż wydaje się, że niektóre z tych zmiennych mogą mieć wartość predykcyjną dla modelowania zmian własności nieruchomości (takich jak przejęcia nieruchomości), istnieje szereg wad stosowania ich do modelowania predykcyjnego, w tym:

* Niektóre dane zostały zebrane przez skrobanie stron internetowych. Nie masz pewności, że będziesz w stanie uzyskać te dane w przyszłości.

* Nie masz szczegółowych informacji na temat procesu skrobania, więc nie możesz mieć pewności, że pobierane dane zostały zdefiniowane spójnie.

* Będziesz miał mnóstwo czasu na wyjaśnienie znaczenia danych bez dokumentacji.

Decydujesz więc, że przy pierwszej próbie opracowania modelu predykcyjnego zmiany własności nieruchomości użyjesz tylko tych pól, które zostały odpowiednio udokumentowane. W przyszłym projekcie możesz poszukać alternatywnych źródeł dla niektórych innych dziedzin.

Eksploracja danych

Nadszedł czas, aby krótko przeanalizować dane dla każdej zmiennej w każdym pliku. Musisz sprawdzić podstawowe informacje, takie jak czy dane są ciągami czy liczbami, czy zakres wartości jest odpowiedni, a rozkład wartości wygląda na rozsądny. Zanotujesz wszelkie rozbieżności w dokumentacji i własne uzasadnione oczekiwania. Procedury, których będziesz używać do generowania informacji diagnostycznych o danych, różnią się w zależności od rodzaju posiadanych danych, dostępnych narzędzi i sposobu, w jaki lubisz pracować. Możesz używać wysoce zautomatyzowanych funkcji lub możesz pracować ze zmiennymi w małych grupach lub pojedynczo. Prawie zawsze będziesz mieć wybór sposobów, aby to zrobić. Dla każdego pola przygotowujesz krótkie podsumowanie, podając nazwę i opis, liczbę brakujących obserwacji oraz zakres wartości (niski i wysoki). Możesz również dołączyć dodatkowe informacje, takie jak wykres rozkładu, średnia (średnia) i najczęściej występująca (mod) wartość zmiennej. W tym momencie nie będziesz próbował powiązać jednej zmiennej z inną. Zaczynasz od korzystania z oprogramowania, które generuje podstawowy raport dla każdej zmiennej w danych, w tym informacje, takie jak zakres wartości, średnia dla zmiennych ciągłych, najczęstsza wartość dla zmiennych kategorialnych itd.

Ten raport jest punktem wyjścia do zrozumienia danych. Używasz go do określenia, jakie dane posiadasz i czy są one zgodne z tym, czego oczekujesz od dokumentacji i twoich współpracowników. Dodajesz do tego za pomocą wykresów lub innych prostych metod dodawania szczegółów do zrozumienia każdej zmiennej. Przeglądając każdą zmienną, opisujesz ją i odnotowujesz wszelkie obawy oraz co należy zrobić, aby je rozwiązać. W swoich podsumowaniach stwierdzasz, czy zmienna wydaje się gotowa do użycia w modelowaniu, wymaga dalszego przygotowania, czy jest w tak złym stanie, że nie można jej użyć. Twoje indywidualne podsumowania zmiennych wyglądają jak przykłady pokazane w tabeli

Nazwa zmiennej : Opis

BI_VIOL:

Opis: nieznany (brak dokumentacji dla tej zmiennej)

Typ zmiennej: ciąg

Zakres: XXXX do XXXX

Liczba brakujących spraw: 0

Ocena: Niedopuszczalne do modelowania. Wszystkie sprawy mają tą samą wartość. Przyczyna nieznana.

Kolejne kroki: nie będą używane w tym projekcie.

TAXKEY:

Opis: Dziesięciocyfrowy numer kodu identyfikacyjnego nieruchomości

Typ zmiennej: Identyfikator (ciąg)

Zakres: 9999000–7369999110

Liczba brakujących spraw: 30

Ocena: brakuje niewielkiej liczby przypadków. W niektórych przypadkach mają mniej niż dziesięć cyfr, prawdopodobnie z powodu obcięcia wiodących zer ponieważ format zmiennej był odczytywany jako liczba całkowita, a nie jako łańcuch.

Kolejne kroki: Należy jak najlepiej wyczyścić tę zmienną, ponieważ jest to unikalny identyfikator dla każdej usługi. Zmień typ zmiennej z liczby całkowitej na ciąg znaków. Ponownie oceń.

C_A_CLASS:

Opis: Kod zajęć egzaminacyjnych - określa wykorzystanie właściwości.

Typ zmiennej: Nominalna

Zakres: 1-9

Liczba brakujących spraw: 0

Ocena: Dystrybucja wygląda odpowiednio z klasą 1 (mieszkalna), która jest najczęściej występującą kategorią. Brak widocznych oznak problemów z jakością.

Kolejne kroki: To pole jest gotowe do użycia w modelowaniu.

DIV_ORG:

Opis: numer kontrolny używany w biurze osoby oceniającej

Typ zmiennej: ciąg

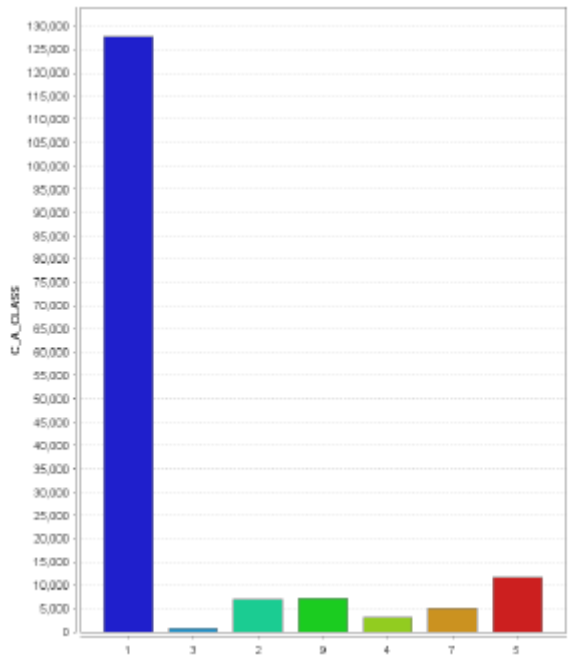
Zakres: 0-999

Liczba brakujących spraw: 0

Ocena: służy do administrowania w ramach oceniającego i nie wydaje się mieć żadnej wartości do celów modelowania.

Kolejne kroki: brak.

Niektóre z tych zmiennych nie będą przydatne w modelowaniu. Na przykład BI_VIOL brzmi tak, jakby mógł reprezentować liczbę lub rodzaj naruszeń inspekcji budynku zgłoszonych dla nieruchomości. Być może był kiedyś używany do tego celu, ale w tym zbiorze danych każdy przypadek ma tę samą wartość „XXXX”. Pole nie jest wymienione w żadnej posiadanej dokumentacji. Naruszenia dotyczące budynków mogą być cennymi informacjami przy prognozowaniu przeniesień własności, ale być może będziesz musiał poczekać na przyszły projekt, gdy będziesz mieć czas na wyśledzenie innego źródła tych informacji. Na szczęście niektóre pola są w znacznie lepszym stanie. Na przykład C_A_CLASS, kod klasy oceny, identyfikuje wykorzystanie nieruchomości w głównych klasach, takich jak mieszkalne, produkcyjne i komercyjne. Może to być bardzo ważne przy modelowaniu, ponieważ oczekujesz różnych wzorców zachowań dla różnych zastosowań właściwości. Nie brakuje przypadków dla C_A_CLASS, zakres wartości jest zgodny z dokumentacją, a wykres słupkowy



pokazuje, że rozkład użytkowania nieruchomości wydaje się rozsądny, a klasa mieszkalna występuje znacznie częściej niż jakiekolwiek inne wykorzystanie . Zauważasz, że wiele pól, które początkowo mogą wydawać się wartościowe dla modelowania, nie jest w dobrym stanie. Wiele z nich jest nieudokumentowanych, niektóre nie są utrzymywane przez źródło publiczne (a dokumentacja tak mówi), a inne nie różnią się lub nie wydają się zgodne z Twoimi oczekiwaniami. Masz wątpliwości, czy pozostałe dane wystarczą do zbudowania użytecznego modelu

Dane dotyczące czyszczenia

Po zbadaniu danych okazało się, że niektóre zmienne, które wydają się mieć wartość dla modelowania, zawierają drobne błędy lub inne problemy, które chcesz najpierw naprawić. Dobrym przykładem jest pole TAXKEY. Jest to numer kodu, który identyfikuje każdą indywidualną działkę nieruchomości. Ściśle mówiąc, identyfikator nie jest zmienną modelującą, ale model nie będzie miał żadnej wartości, chyba że dopasujesz swoje przewidywania do określonych właściwości. Zauważyłeś dwa problemy w danych:

* W kilku przypadkach (ułamek procenta całości) brakuje kodów identyfikacyjnych.

* Wiele spraw ma mniej niż dziesięć cyfr, które zgodnie z dokumentacją powinny.

Poświęć chwilę, aby zastanowić się nad brakującymi przypadkami (30 z ponad 160 000). Teoretycznie instytucja publiczna, która udostępniła dane, może wypełnić te luki. Ale wyobrażasz sobie dzwonienie do biura rzeczoznawcy majątkowego i wyjaśnianie problemu, być może wielokrotnie, szukanie kogoś, kto go rozumie i jest chętny do pomocy. Kiedy docierasz do tej osoby, nie masz pewności, że chęć pomocy przełoży się na sukces w skorygowaniu błędów w danych. Myślisz, że w tym czasie możesz zrobić bardziej produktywne rzeczy i zdecydować się żyć bez tych 30 przypadków. Następnie niektóre przypadki mają mniej niż dziesięć cyfr w swoich kodach właściwości. Ten problem występuje często, ale podejrzewasz, że można go łatwo naprawić. Ponieważ kod jest numeryczny, oprogramowanie zinterpretowało go jako liczbę całkowitą, ale ciąg byłby bardziej odpowiedni. Zmiana typu pola na ciąg uniemożliwiłaby programowi obcinanie jakichkolwiek wiodących zer w kodach właściwości. Więc ponownie importujesz dane do oprogramowania, tym razem upewniając się, że pole jest nominalne (jak nazwa). Mimo to można znaleźć wiele przypadków, w których wartości mają mniej niż dziesięć

cyfr. Twoja łatwa poprawka niczego nie naprawiła. Zaglądasz do danych w edytorze tekstu (ponieważ są to dane w formacie tekstowym, możesz użyć edytora tekstu lub arkusza kalkulacyjnego, aby je wyświetlić) i potwierdzić, że problem nie ma nic wspólnego z przycinaniem zer wiodących. Niektóre wartości są po prostu krótsze niż dziesięć cyfr, których oczekiwałeś w dokumentacji. Zapisujesz to w swoim raporcie i decydujesz (na dziś), że zaufasz danym, a nie dokumentacji. Podobny proces przechodzisz dla każdego pola, które wydaje Ci się potencjalnie przydatne, ale nie jest w idealnym stanie. Podczas pracy dokumentujesz swoje obserwacje i wszelkie wprowadzane zmiany. Dla każdej dziedziny oceniasz, czy jest wystarczająco dobra do wykorzystania w modelowaniu. (Nie decydujesz, czy zmienna znajdzie się w ostatecznym modelu, czy też dobrze sprawdzi się jako predyktor, tylko czy jest wystarczającej jakości do przetestowania). Na koniec łączysz swoje notatki z tych obserwacji i działań w raporcie dotyczącym jakości danych.

Jak eksploratorzy danych spędzają czas

Kucharze serwujący pyszne obiady spędzają dużo czasu na siekaniu warzyw. Biegacze, którzy wygrywają wyścigi, spędzają dużo czasu na rozciąganiu i treningu. Eksploratorzy danych, którzy opracowują cenne modele predykcyjne, spędzają dużo czasu na przygotowywaniu danych. Ludzie, którzy nie próbowali jeszcze eksploracji danych, czasami myślą, że odkrywanie wspaniałych spostrzeżeń i opracowywanie potężnych modeli to niekończąca się ekscytująca przejażdżka. Tak nie jest. Większość twojego czasu idzie na robienie wszystkich rzeczy, które trzeba zrobić, zanim zaczniesz budować modele. Przygotowanie danych nie jest najbardziej efektownym aspektem pracy. To żmudna praca i masz dużo do zrobienia, do tego stopnia, że eksploratorzy danych spędzają więcej czasu na przygotowywaniu danych niż na czymkolwiek innym. Jednak przygotowanie danych jest warte wysiłku, ponieważ umożliwia znaczące ich odkrycie.

Przygotowywanie danych

Po zebraniu danych i przejrzeniu pól jeden po drugim, aby zapoznać się z danymi i sprawdzić, czy nie występują problemy z jakością, przechodzisz dalej i przygotowujesz dane do modelowania. Na tym etapie pracy wykonujesz niezbędne zadania, aby przekształcić dane z ich pierwotnej postaci do formy wymaganej do modelowania, np.

* Łączenie zbiorów danych

* Określenie roli pól

* Pobieranie próbek danych

* Dzielenie próbki na podzbiory w celu budowania i testowania modeli

Wiele projektów wymaga wyprowadzenia nowych pól na podstawie tych, które są już w danych. Na przykład zmienna wskaźnikowa, która będzie potrzebna do zidentyfikowania właściwości, które zmieniły własność, nie istnieje w danych publicznych. Należy to obliczyć na podstawie innych pól. Na szczęście dla ciebie, twój kolega Matt już utworzył tę zmienną i zapisał ci krok w tym projekcie. Ale będziesz musiał wyprowadzić dla siebie inne nowe pola.

Pierwsze kroki z danymi nieruchomości

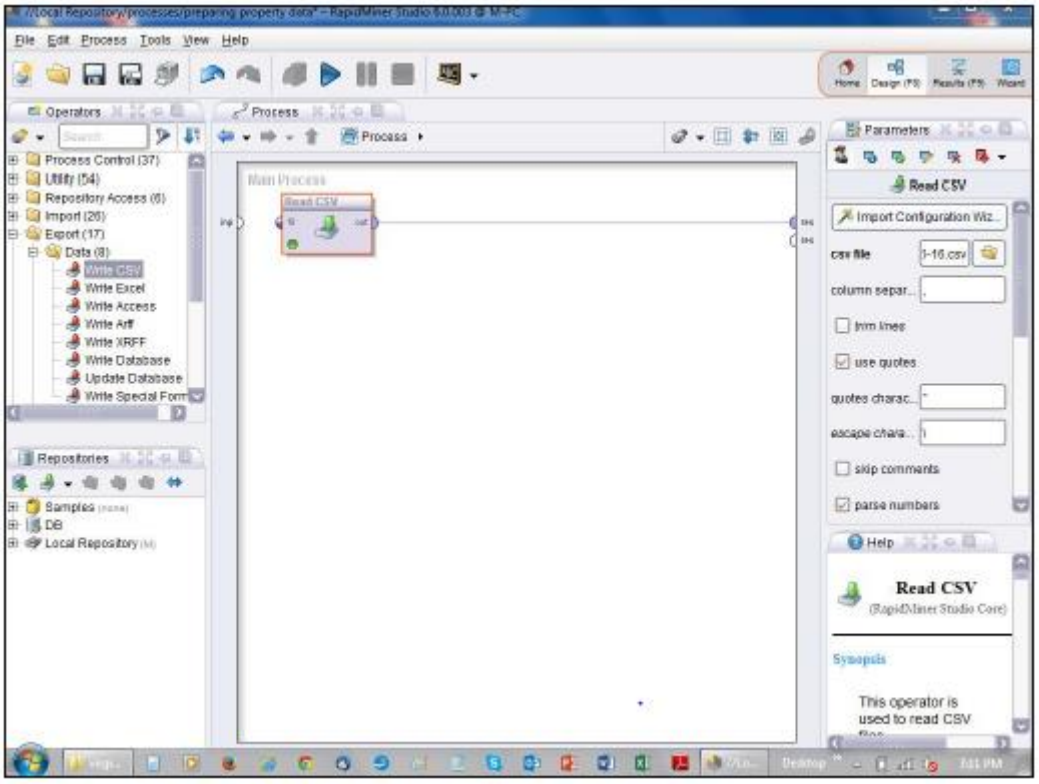
Na etapie zrozumienia danych zidentyfikowałeś szereg zmiennych, których nie będziesz używać do modelowania. Wykluczyłeś każdy z nich z jednego z następujących powodów:

* Nie ma sensu jako predyktor: obejmuje unikalne pola, takie jak adres lub nazwy, lub coś, co Twoim zdaniem nie ma związku ze zmianą własności

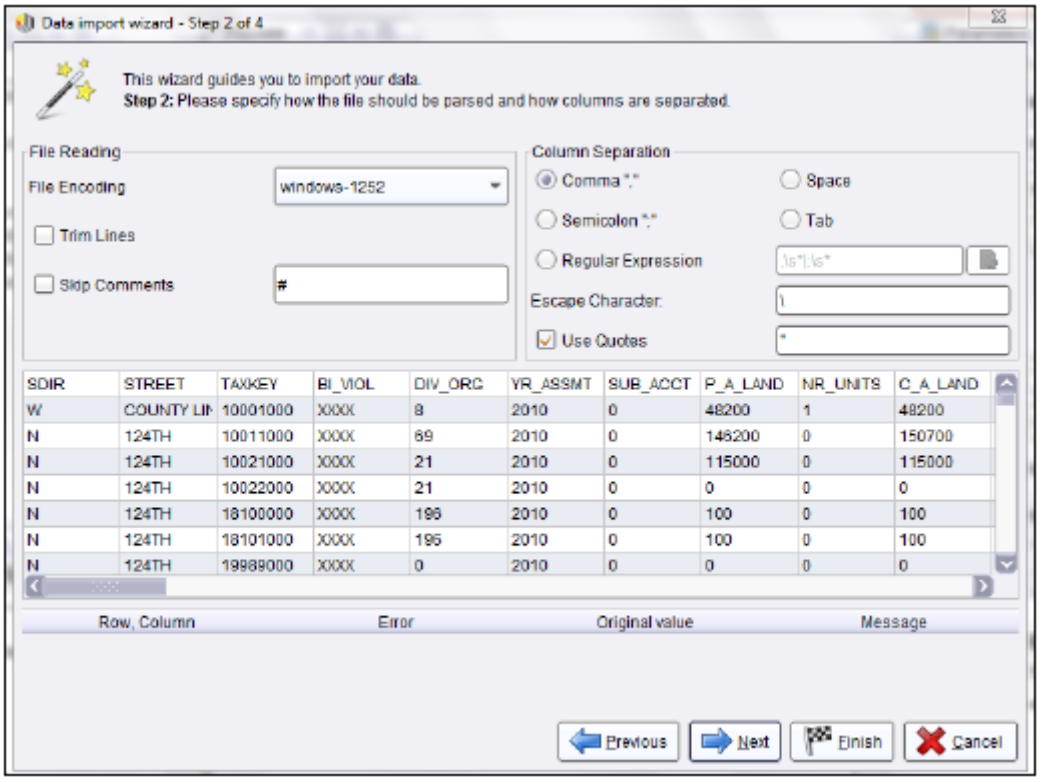
* Jakość danych jest słaba: wiele brakujących przypadków lub nieprawidłowe wartości

* Nie zmienia się: wszystkie przypadki mają tę samą wartość (niekoniecznie problem z jakością danych)

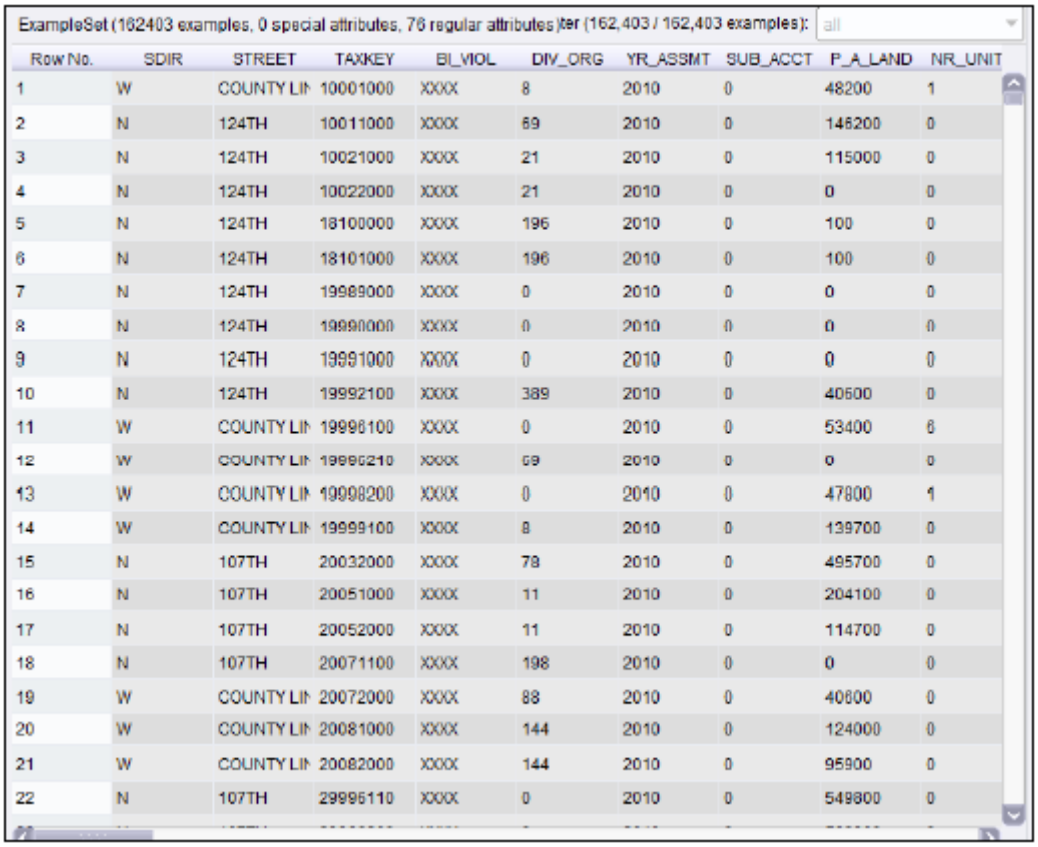
Pracujesz ze specjalistycznym oprogramowaniem do eksploracji danych. Chociaż możesz wykonywać te same operacje za pomocą innych rodzajów narzędzi, oprogramowanie do eksploracji danych zostało zaprojektowane tak, aby ułatwić przeglądanie etapów procesu i szybką pracę, łącząc ze sobą sekwencję operacji reprezentowaną przez małe ikony. Każda ikona to narzędzie z określoną funkcją oraz własnymi opcjami i ustawieniami. Nazywa się to wizualnym interfejsem programowania. Plik danych właściwości jest dość duży, więc na początek zaimportujesz plik danych właściwości, usuniesz zmienne, których nie możesz użyć, a resztę zapiszesz w nowym (nieco mniejszym) pliku. Najpierw wybierasz narzędzie do odczytu danych i umieszczasz je w głównym obszarze roboczym oprogramowania do eksploracji danych, jak pokazano na rysunku



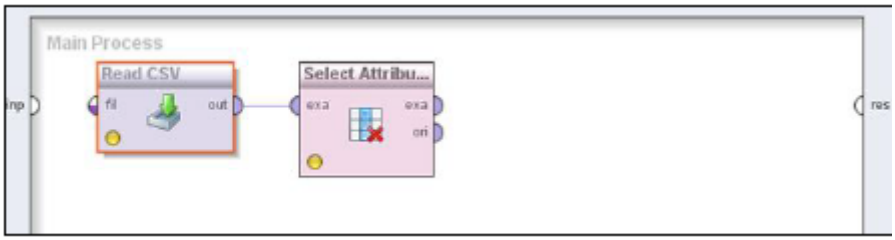
Kreator (specjalny interfejs użytkownika, który upraszcza złożone zadania) pomaga poprawnie zaimportować dane. Jeden krok kreatora pokazano.



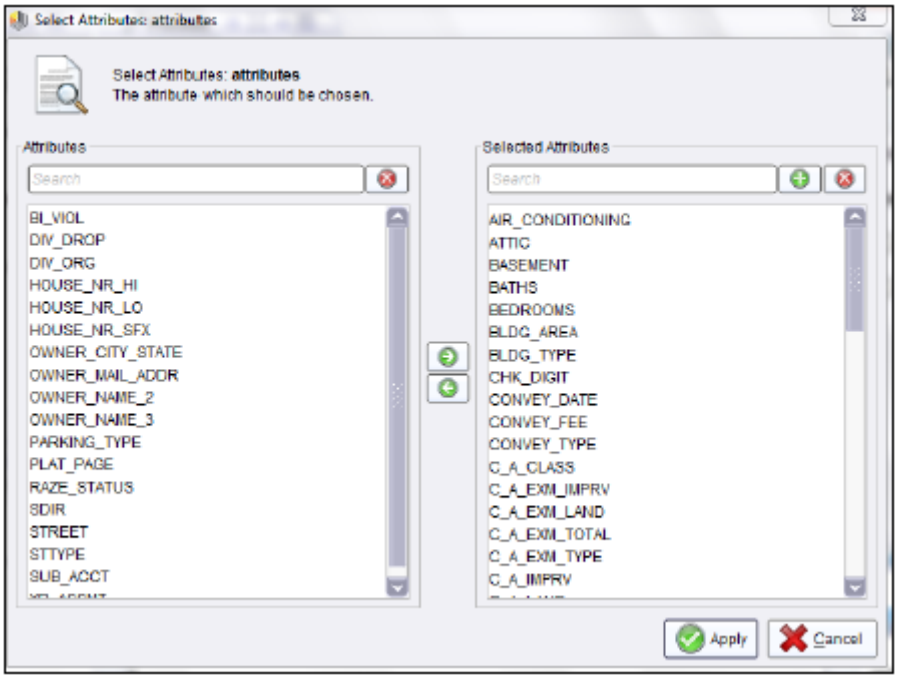
Po zaimportowaniu danych możesz je wyświetlić i sprawdzić, czy wyglądają prawidłowo.



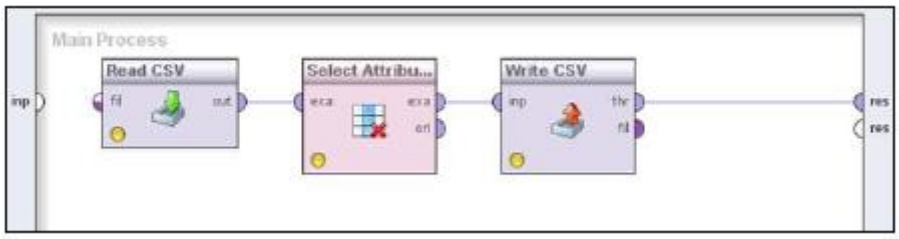
Dodajesz kolejne narzędzie do obszaru roboczego, aby wybrać zmienne, które mają być przechowywane w danych.



Konfiguracja nie jest skomplikowana. Narzędzie wyświetla listę zmiennych w danych, a Ty wybierasz te, które chcesz zachować. Rysunek przedstawia konfigurację. Lista po prawej stronie zawiera wszystkie zmienne wybrane do zachowania



Jeszcze jedno narzędzie

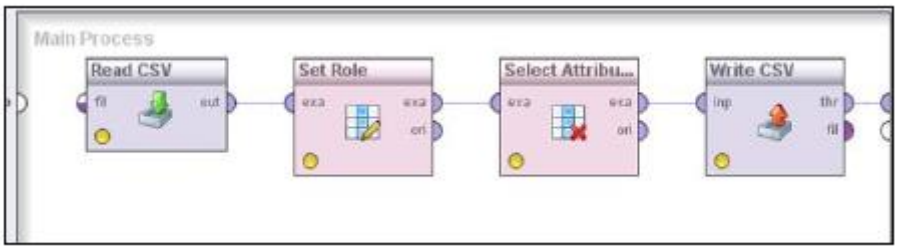


pozwala zapisać wybrane zmienne w nowym pliku. Oprogramowanie do eksploracji danych użyte w tym przykładzie zawiera wiele z tych specjalistycznych narzędzi. Na przykład ma inne narzędzie dla każdego z typów plików, które może odczytać, i dla każdego typu, który może zapisać. Nie każdy produkt ma takie podejście; inni mogą mieć jedno narzędzie, które może zapisać wybór kilku typów plików.

Przygotowanie wskaźnika zmiany właścicielskiej

Matt ułatwił ci pracę, wyprowadzając zmienną wskazującą, które właściwości zmieniły właściciela, a które nie. Będzie to zmienna zależna lub docelowa do modelowania. Nadal będziesz musiał wykonać

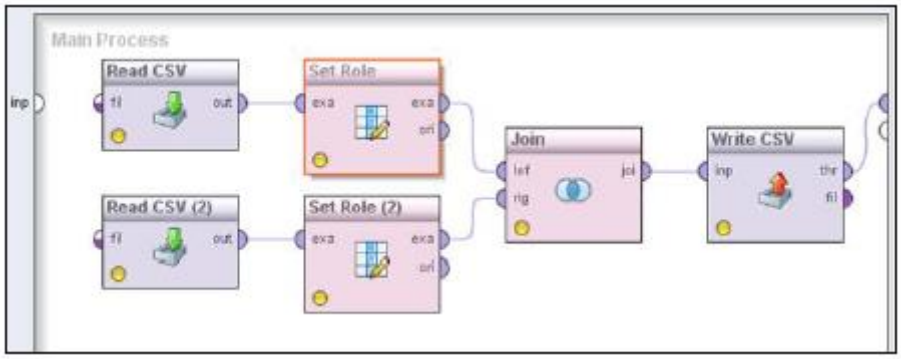
pewne przygotowania z tą częścią danych, w szczególności wybierając odpowiednie ustawienia oprogramowania do eksploracji danych, aby zidentyfikować zmienną docelową. Tworzysz sekwencję podobną do tej, której użyłeś w poprzedniej sekcji dla danych właściwości. Zaimportujesz dane, wybierzesz zmienną do zachowania i zapiszesz ją w nowym pliku. Ale na rysunku 2-9 widać, że tym razem istnieje inne narzędzie między importem danych a ich selekcją.



Dzięki niemu wskazujesz, która zmienna jest celem, ustawiając właściwości narzędzia, jak pokazano na rysunku.

Łączenie zbiorów danych

Masz dane właściwości w jednym pliku, a dane o usługach, które zmieniły właściciela, w innym. Musisz połączyć te dwa elementy. Proces pokazano na rysunku



Wczytujesz się w każdym z plików, które utworzyłeś wcześniej. Dla każdego podajesz nazwę zmiennej identyfikującej właściwość, ustawiając właściwości odpowiedniego narzędzia .

Zmienna identyfikacyjna kieruje połączeniem dwóch plików, dopasowując ogólne dane dla każdej właściwości do wyników: Czy właściwość zmieniła właściciela?

Wyprowadzanie nowych zmiennych

Oprogramowanie do eksploracji danych wykonuje za Ciebie dużo pracy, ale nic nie zastąpi Twojej wiedzy biznesowej. Rozumiesz, że jedna zmienna reprezentuje cenę zapłaconą za nieruchomość, inna inwestycje mające na celu ulepszenie nieruchomości, a trzecia szacowaną wartość - ale oprogramowanie tego nie robi. Oprogramowanie widzi tylko liczby, kategorie i tekst, a nie znaczenie. Rozumiesz, że poza wszystkimi innymi interesujesz się związkami między tymi trzema zmiennymi. Oprogramowanie nie. Tego rodzaju wiedzę biznesową integrujesz ze swoją analizą, wykorzystując ją do uzyskiwania odpowiednich nowych zmiennych do modelowania.

Wybór punktu wyjścia

Virginia i Matt powiedzieli wam o wielu czynnikach, które mogą być dobrymi wskaźnikami zbliżających się zmian własności nieruchomości. Sugestie te są wynikiem badań i wywiadów, które przeprowadzili we wcześniejszych projektach. Niektóre z tych czynników to

* Właściciele nie mieszkają w okolicy.

* Właściciel jest samorządem terytorialnym.

* Podatki nie są opłacone.

* Nieruchomość jest pusta.

* Podział na strefy i faktyczne wykorzystanie nie są dopasowane.

* Wartość nie jest zgodna z oceną.

* Ulepszenia są niewielkie w stosunku do wartości nieruchomości.

* Naruszenia przepisów budowlanych są jawne.

* Wiele przypadków naruszenia przepisów budowlanych zostało zamkniętych.

* Wiele zgłoszeń serwisowych zostało zamkniętych.

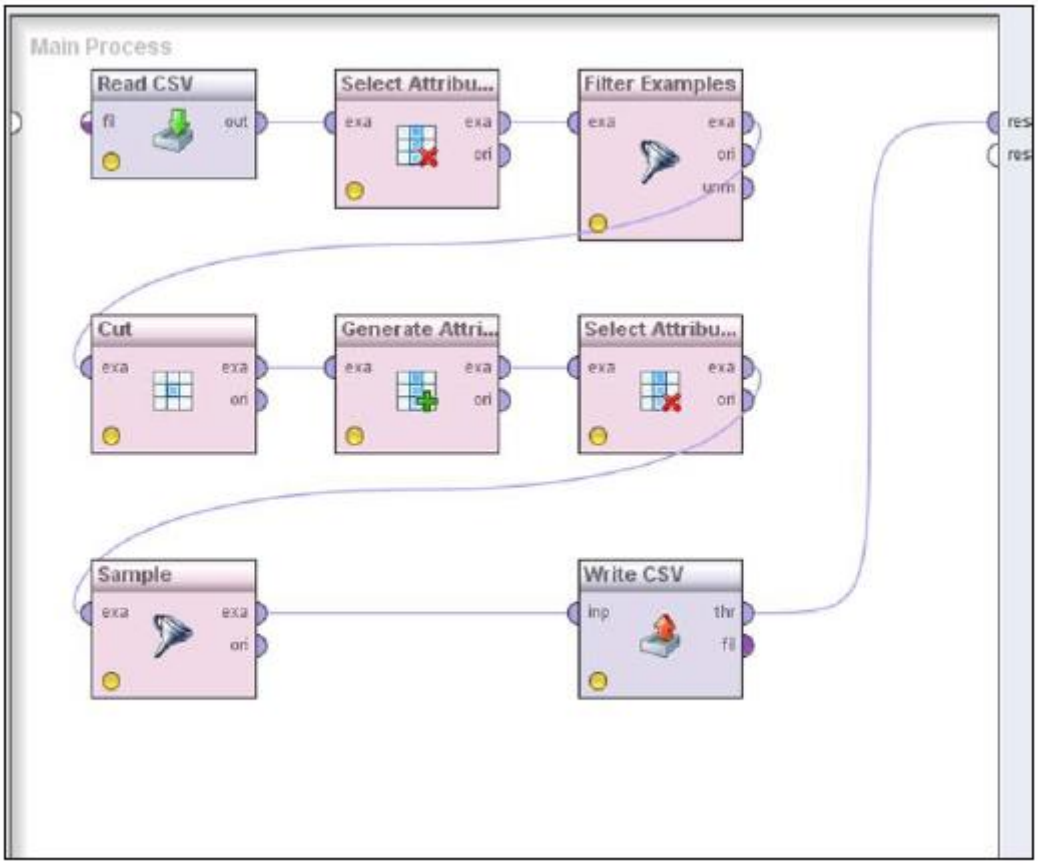
* Nieruchomość jest wystawiana na wynajem lub sprzedaż.

* Nieruchomość jest przejęta.

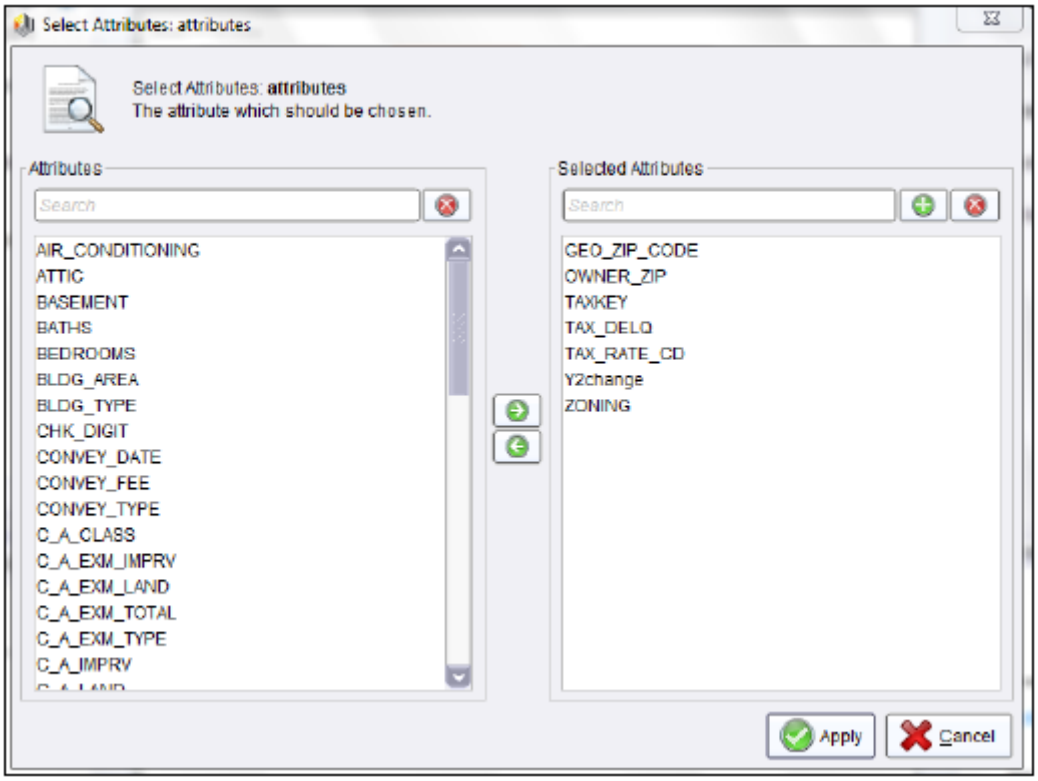
Chociaż masz dobre powody, by sądzić, że każdy z tych czynników jest ważny, nikt jeszcze nie potwierdził ich wartości, budując i testując model predykcyjny. Chciałbyś zbadać każdy z nich - a także inne. Ale nie masz odpowiednich danych dla niektórych, a inne będą wymagały wysiłku w celu przygotowania danych. Celem tego projektu nie jest opracowanie możliwie największego modelu, ale wykorzystanie danych do wykazania, że co najmniej jedna zmienna ma wartość do przewidywania zmian własności nieruchomości. Chodzi o to, aby szybko dostarczyć konkretnych dowodów na to, że modelowanie predykcyjne jest wykonalne. Aby zwiększyć szybkość, najpierw wybierasz kilka pozycji z tej listy, aby spróbować. (Jeśli nie działają, możesz wrócić i wypróbować inne). W pierwszej kolejności wybierasz nieruchomości z właścicielami, którzy nie mieszkają w okolicy, i nieruchomości z niezapłaconymi podatkami. Twoje powody są proste: masz odpowiednie dane dla tych zmiennych, a wymagane przygotowanie jest dość proste.

Wykonywanie obliczeń

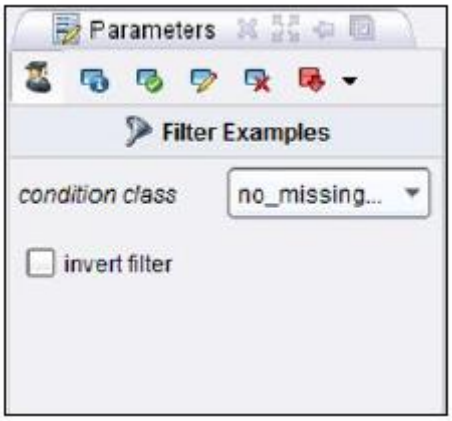
Ta część procesu jest bardziej złożona niż kroki, które zostały podjęte do tej pory.



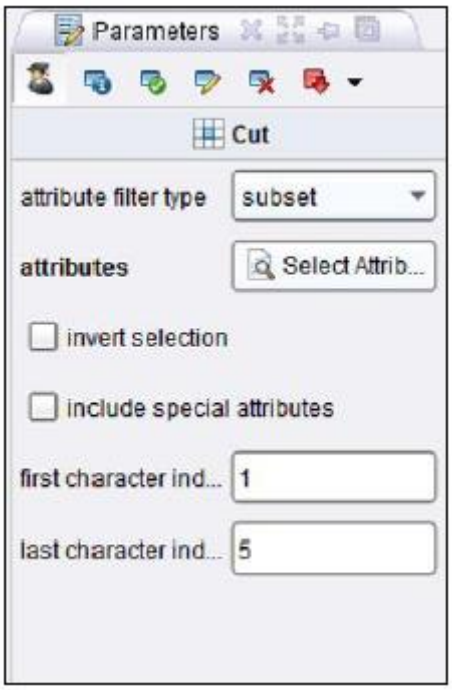
Utworzysz dwie nowe zmienne, wybierzesz podzbiór obserwacji do modelowania i usuniesz wszystkie obserwacje, które nie mają wystarczających danych do wykorzystania w procesie modelowania. Przed utworzeniem jakichkolwiek nowych zmiennych należy trochę uporządkować. Chociaż w danych znajduje się wiele zmiennych, zdecydowałeś się użyć tylko kilku zmiennych w swoim pierwszym modelu, więc wybierasz tylko te z danych.

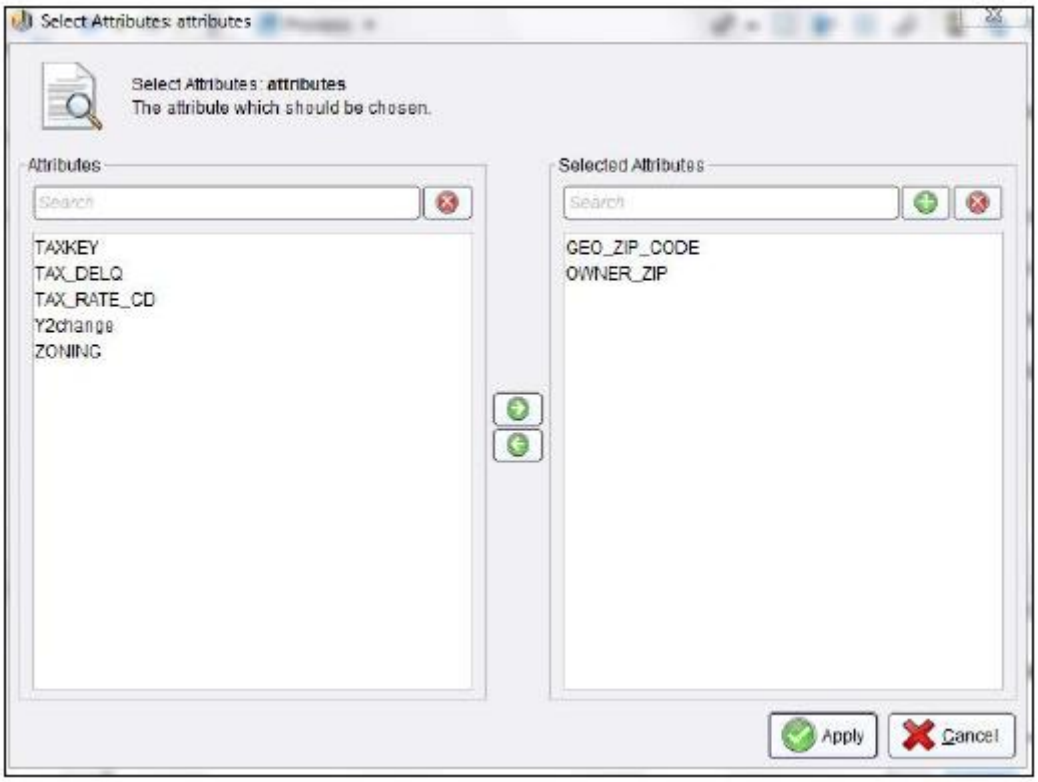


Narzędzia do modelowania, a nawet niektóre narzędzia do przygotowywania danych, nie działają dobrze i mogą w ogóle nie działać, jeśli w danych brakuje wartości, więc odfiltrowujesz przypadki z brakami danych. Odpowiednią konfigurację narzędzia filtrującego przedstawia Rysunek.



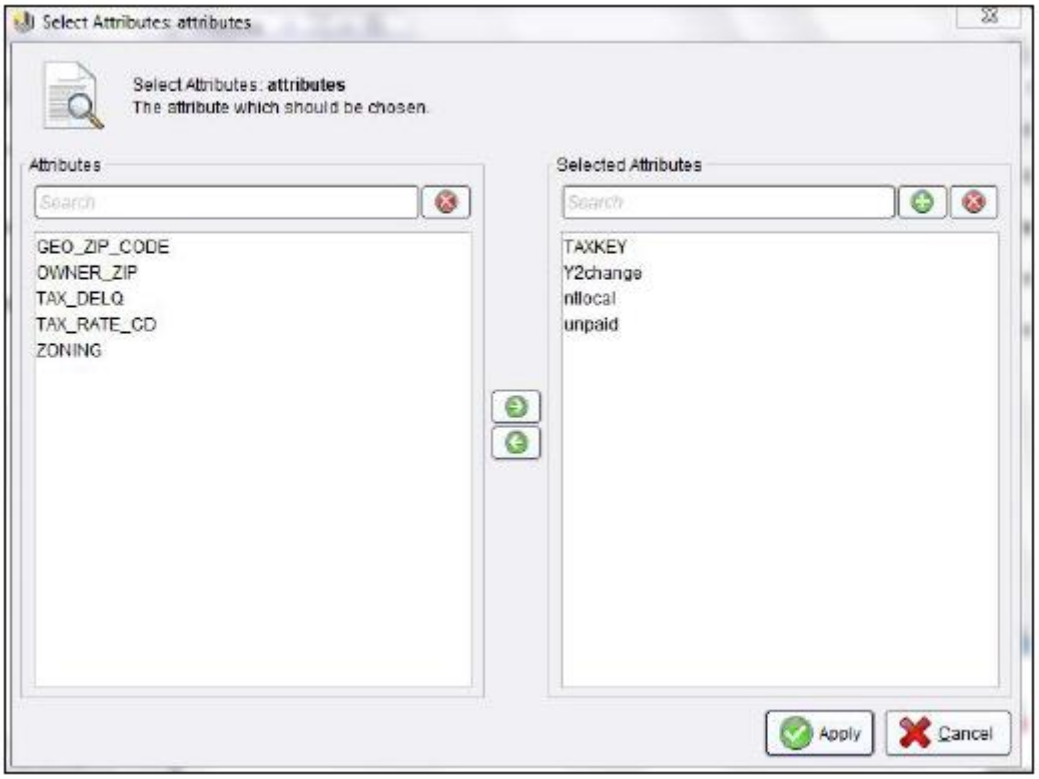
Aby zidentyfikować właścicieli nieruchomości, którzy nie mieszkają w swoich nieruchomościach lub są bardzo blisko ich nieruchomości, należy porównać domowy kod pocztowy właściciela z kodem pocztowym nieruchomości. Masz dane dla każdego z nich, ale istnieją pewne wyzwania związane z ich porównaniem. Niektóre kody pocztowe są zapisywane jako pięć cyfr; inne są w dłuższych formatach. Dlatego przed utworzeniem zmiennej wskaźnikowej dla nieruchomości, których właściciele nie są lokalni, musisz ustawić wszystkie kody pocztowe w spójnym formacie. Ustawiasz zmienną cut tak, aby zachować pierwsze pięć znaków z dwóch zmiennych kodu pocztowego.



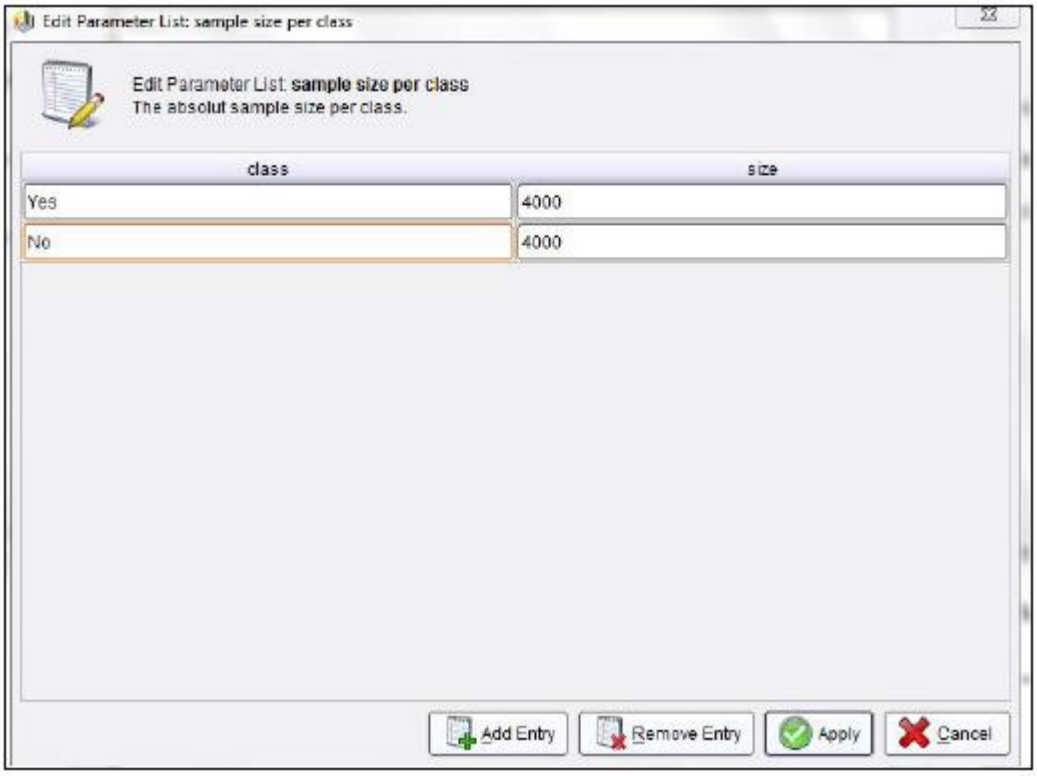


W danych nieruchomości istnieje już zmienna wskazująca, które nieruchomości mają niezapłacone podatki, ale nie jest w dobrej formie do modelowania. Ta zmienna ma wartość 1, jeśli podatki są niezapłacone, ale „NA” w przeciwnym razie. Narzędzia do modelowania tego nie lubią! Utworzysz więc ładną, nową zmienną o wartości 1, jeśli podatki są niezapłacone, i 0 w innym przypadku. Konfiguracja tworzenia obu nowych zmiennych wskaźnikowych jest pokazana na Rysunku.

Masz jeszcze kilka kroków, zanim przejdziesz do fazy modelowania. Teraz, gdy masz już nowe zmienne, nie będziesz potrzebować starych, więc możesz użyć narzędzia do wybierania zmiennych



aby zachować to, czego potrzebujesz. Użyjesz narzędzia do próbkowania danych, aby zrównoważyć dane i wybrać próbkę z mniej więcej równymi proporcjami właściwości, które zmieniły właściciela i nie. Rysunek przedstawia konfigurację równoważenia zbioru danych.



Żądasz około 4000 przypadków w każdej grupie, ale rozumiesz, że rzeczywiste rozmiary próbek mogą się nieco różnić. Wow, co za dużo kroków! Przygotowanie danych do tego przykładu jest prostsze niż w większości. Dlatego trzecia ustawa o eksploracji danych stwierdza, że przygotowanie danych to ponad połowa każdego procesu eksploracji danych.

Modelowanie danych

Modele predykcyjne to nic innego jak równania, które pomagają w dokonywaniu przemyślanych przypuszczeń w metodyczny, spójny sposób, w oparciu o dane. Ludzie przez cały czas formułują nieformalne prognozy, w domu i w pracy:

* Kupowanie artykułów spożywczych: szacowanie zużycia na podstawie ostatnich doświadczeń i przewidywanych zmian, takich jak goście w domu lub nadchodząca podróż

* Budżetowanie: planowanie potrzeb finansowych na podstawie informacji, takich jak przeszłe wydatki, znane nadchodzące wydarzenia i szacunkowe zapotrzebowanie na fundusze awaryjne

* Prognozowanie sprzedaży: przewidywanie przyszłej sprzedaży na podstawie wyników historycznych, przewidywanych transakcji, nastawienia do gospodarki i być może tylko odrobiny pobożnych życzeń

Ponieważ te nieformalne prognozy są tworzone w niespójny, nieudokumentowany i subiektywny sposób, trudno je poprawić. Jako eksplorator danych tworzysz niezawodne modele predykcyjne oparte na faktach i dokumentujesz proces, aby móc aktualizować i ulepszać modele w przyszłości.

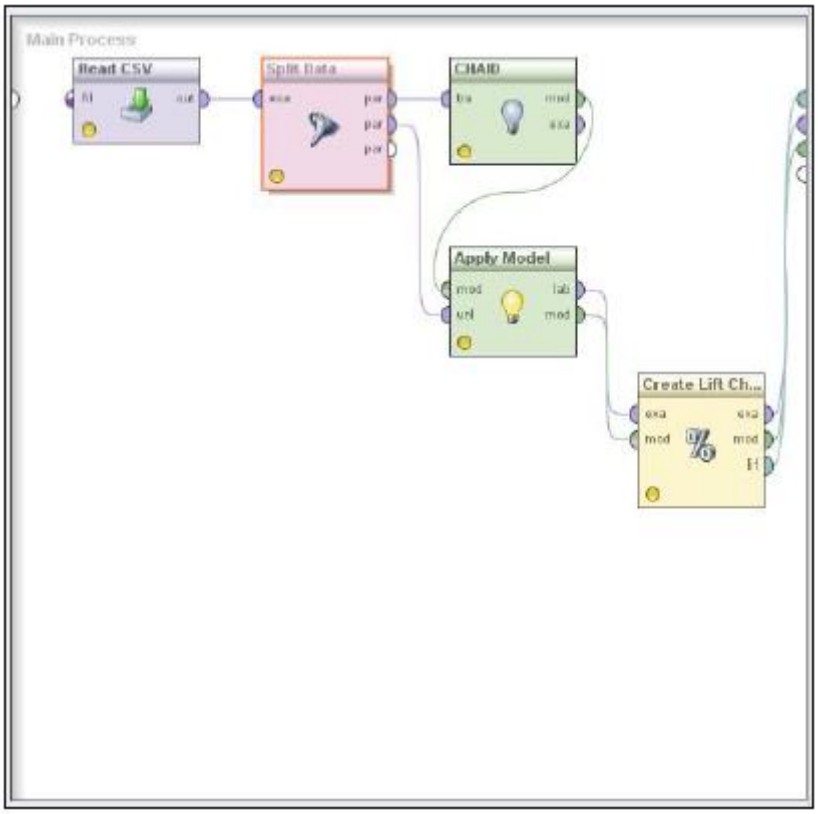
Korzystanie z wyważonych danych

Na etapie przygotowania danych pobrałeś specjalny rodzaj próbki z danych nieruchomości. Próba była zbilansowana, to znaczy obejmowała z grubsza równą liczbę przypadków dla nieruchomości, które zmieniły właściciela w określonym czasie, i dla nieruchomości, które się nie zmieniły. Teraz, gdy jesteś już znanym eksploratorem danych, robisz to z przyzwyczajenia. Równoważenie danych często wydaje

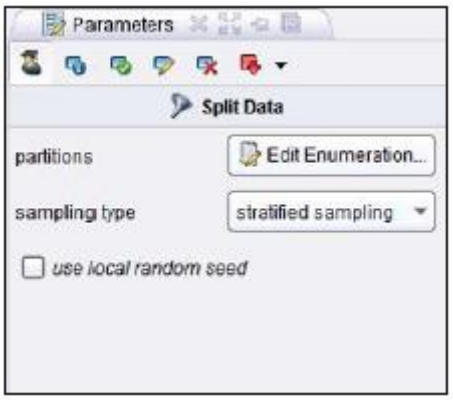
się dziwne lub niewłaściwe nowicjuszom w eksploracji danych. Nie jest oczywiste, dlaczego kopacze danych mieliby używać danych reprezentujących równe proporcje zdarzeń, które nie występują z taką samą częstotliwością w prawdziwym życiu. Na przykład w danym roku tylko niewielka część nieruchomości zmienia właściciela. Po co nadawać temu wydarzeniu reprezentację równą znacznie częstszemu przypadkowi, w którym nieruchomość pozostaje w tych samych rękach? Dzieje się tak, ponieważ celem modelu jest rozróżnienie tych dwóch zdarzeń na podstawie wzorców w danych. Aby skonstruować model, który może rozróżniać te wzorce, potrzebujesz przykładów każdego z nich i nadając każdemu typowi wzorca równe znaczenie w modelowaniu, nadając mu jednakową częstotliwość w danych.

Dzielenie danych

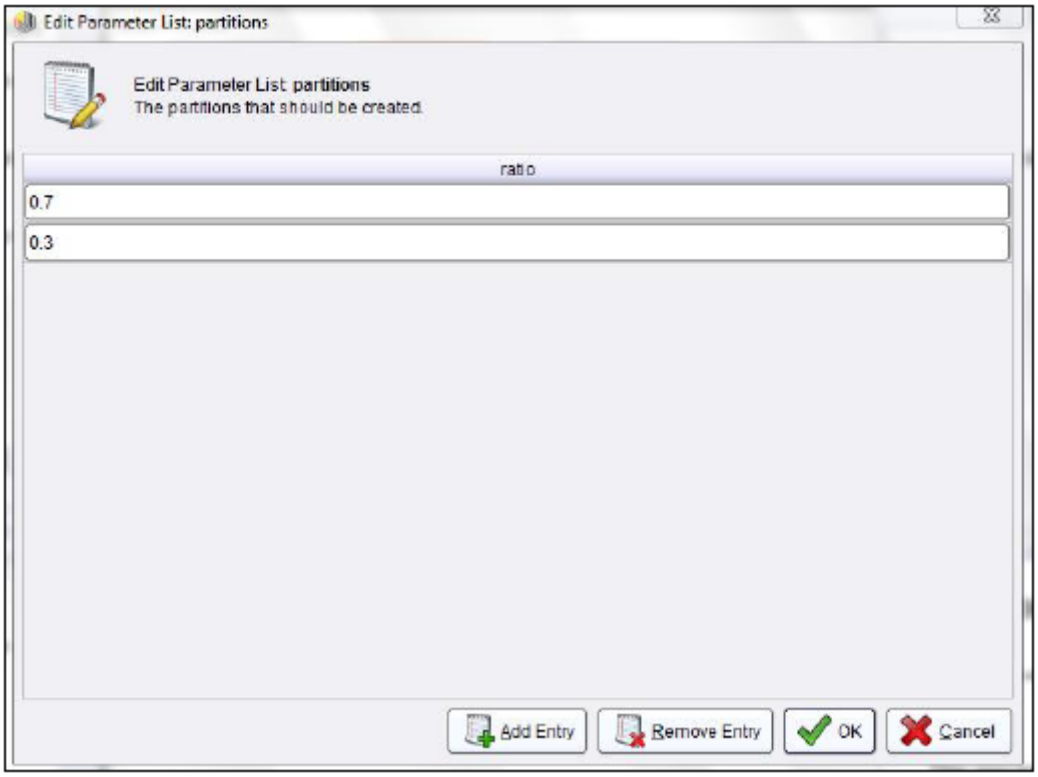
Niektóre techniki uczenia maszynowego, które są szeroko stosowane w eksploracji danych, takie jak drzewa decyzyjne i sieci neuronowe, wymagają jeszcze jednego przygotowania danych przed skonstruowaniem modelu. (Przygotowanie danych trwa i trwa, prawda?) Eksploratorzy danych nie zawsze mogą wykorzystać teorię, aby znaleźć jeden najlepszy model z danych, jak robią to klasyczni statystycy. Dlatego eksploratorzy danych oceniają modele, testując, testując i testując. Część testów jest ukryta w procesie dopasowywania modelu, automatyczna i (prawie) niezauważalna podczas pracy. Niektóre testy są przeprowadzane w terenie poprzez wdrażanie na małą lub pełną skalę. Część z nich jest wykonywana przez oddzielenie części danych (nazywanych danymi testowymi lub wstrzymanymi) przed modelowaniem i użycie modelu do przewidywania wyników dla tych danych, aby można było porównać te przewidywania z tym, co faktycznie się wydarzyło. Twój proces pracy związany z dzieleniem danych, budowaniem modelu i rozpoczęciem oceny jest pokazany na rysunku.



Aby podzielić dane, użyj specjalnego narzędzia do pobierania próbek i określ dwie rzeczy: metodę próbkowania



oraz proporcje danych, które mają być użyte do uczenia i testowania modelu

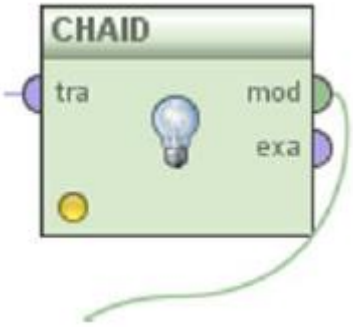


Określasz próbkowanie warstwowe, które zachowuje równowagę proporcji właściwości, które zmieniły właściciela lub nie zmieniły właściciela w próbkach uczących i testowych. Decydujesz się użyć 70% danych do trenowania i 30% do testów.

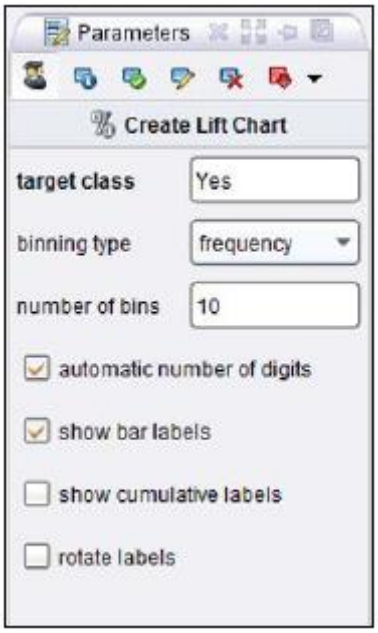
Budowanie modelu

W porównaniu z całą pracą, którą zainwestowałeś w przygotowanie danych, utworzenie pierwszego modelu dla tych danych nie wymaga wiele wysiłku. Do tej pory masz tylko dwie zmienne predykcyjne gotowe do wypróbowania w modelu. Jedna wskazuje, czy właściciel nieruchomości jest lokalny (adres właściciela ma ten sam kod pocztowy co nieruchomość), czy nie. Drugi wskazuje, czy istnieją niezapłacone podatki od nieruchomości. Obie są zmiennymi kategorialnymi, co zawęża wybór technik modelowania. Wybierasz model automatycznego detektora interakcji chi-kwadrat (CHAID), typ modelu drzewa decyzyjnego, na pierwszą próbę, ponieważ dobrze nadaje się do pracy ze zmiennymi kategorialnymi. Jest łatwy w użyciu. Po prostu dodajesz narzędzie do procesu i przepuszczasz dane,

aby zbudować model, nawet bez zmiany jakichkolwiek parametrów. Później możesz zmienić ustawienia, ale nie jest to konieczne przy pierwszej próbie.



Przed uruchomieniem modelu łączysz dwa narzędzia z wcześniej podzielonymi danymi. Narzędzie CHAID wykorzysta 70 procent danych, które umieściłeś na partycji szkoleniowej, a 30% danych, które odłożysz na testy, połączy się z innym narzędziem. To narzędzie zastosuje model CHAID do danych testowych. Na koniec dodajesz ostatni element do swojego procesu. Wykres pomoże Ci zwizualizować wyniki testu modelu. Narzędzie wykresów wymaga niewielkiej konfiguracji . Określasz kategorię, której przewidywanie najbardziej interesuje Cię. W tym przypadku jest to kategoria „Tak".

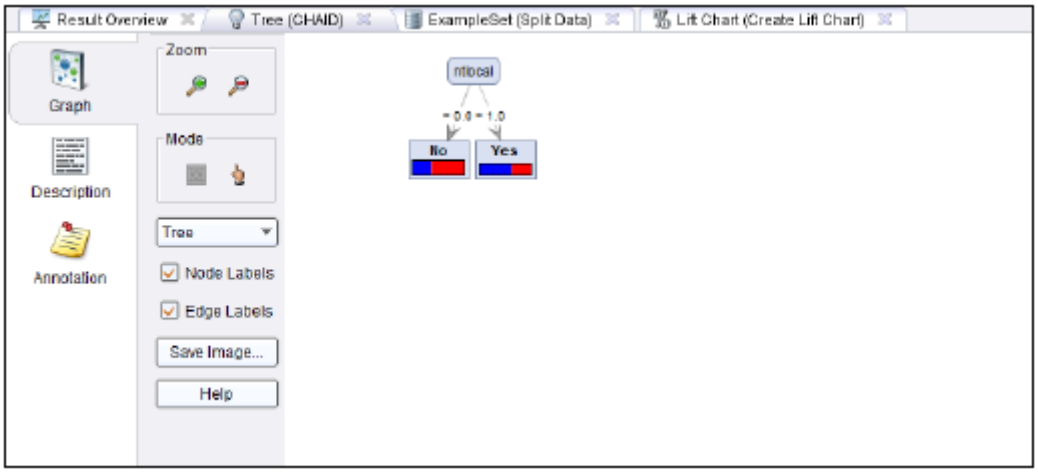


Ocena wyników

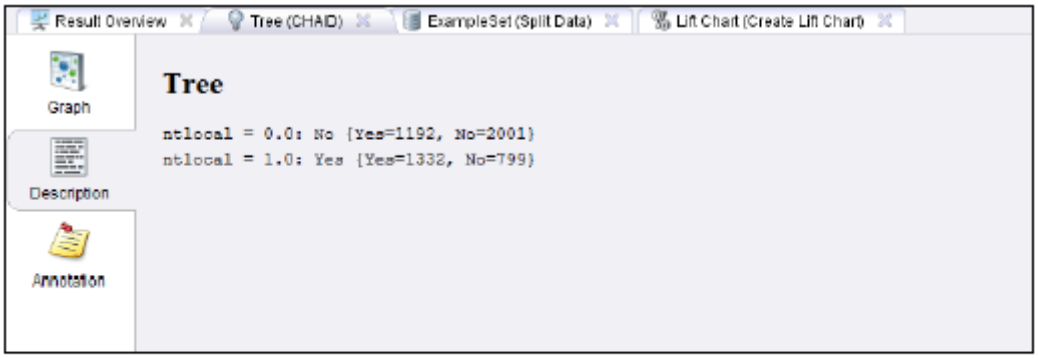
Po utworzeniu modelu nadszedł czas, aby przyjrzeć się modelowi, zobaczyć, jak działa, i wybrać kolejne kroki.

Badanie drzewa decyzyjnego

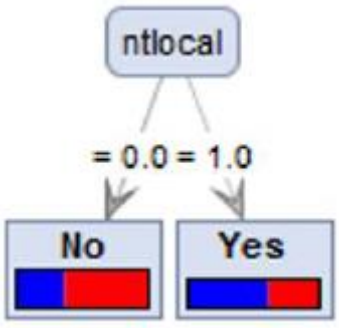
Podczas pierwszej próby modelowania wypróbowałeś tylko dwie predykcyjne zmienne, więc nie spodziewasz się skomplikowanych wyników. Najważniejsze pytanie brzmi, czy okaże się, że nawet jedna z tych zmiennych ma wartość predykcyjną. Oprogramowanie do eksploracji danych wyświetla model CHAID jako diagram drzewa decyzyjnego w interaktywnej przeglądarce wyników



Na początku wyświetlana jest tylko pierwsza gałąź. Narzędzia po lewej stronie przeglądarki umożliwiają rozwijanie drzewa, powiększanie obszarów zainteresowania i wprowadzanie innych zmian w sposobie wyświetlania drzewa. Masz również alternatywę obejrzenia modelu w inny sposób: napisany prostym tekstem



Drzewo



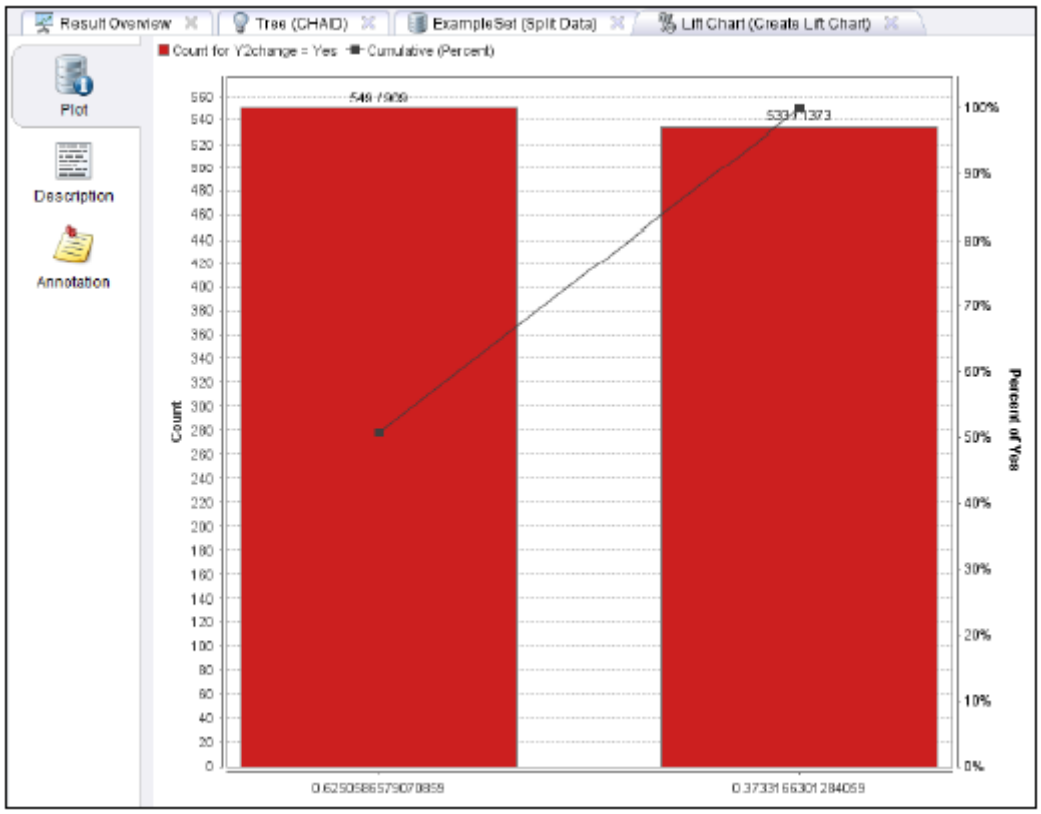
pokazuje, że zmienna lokalna właściciela jest najważniejszym predyktorem. Dane rozgałęziają się na dwie grupy. Lokalni właściciele (ntlocal = 0) są wskaźnikami dla kategorii „Nie”; większość zachowała swoją własność. Właściciele nielokalni (ntlocal = 1) są wskaźnikami dla kategorii „Tak”; byli bardziej skłonni do sprzedaży. W tym przykładzie większość nieruchomości, których właściciele nie są lokalni, zmieniła właściciela; widać to na małym wykresie słupkowym na gałęzi drzewa. (Ale różnice nie muszą

być aż tak dramatyczne, aby utworzyć gałąź w drzewie decyzyjnym. Znacznie bardziej subtelne różnice można wykryć, jeśli w danych istnieje wystarczająco silny wzorzec). Używasz wskaźnika i klikasz gałęzie drzew. Nie rozszerzają się. Lokalna zmienna właściciela jest jedyną zmienną w drzewie. Rzut oka na opis modelu

pokazuje to samo w inny sposób. Dlaczego drugi predyktor, niezapłacona zmienna podatkowa, nie pojawił się w modelu? Być może naprawdę nie jest to dobry wskaźnik zmiany własności nieruchomości. Być może ma to jakąś wartość, ale wybrany typ modelu lub użyte ustawienia (wszystkie pozostawiono z wartościami domyślnymi) nie były odpowiednie do wykrywania związku między niezapłaconymi podatkami a zmianami własności nieruchomości. To wszystko, co na razie wiesz.

Korzystanie z wykresu diagnostycznego

Wykresy diagnostyczne pomagają zrozumieć, jak skutecznie model tworzy dokładne prognozy na podstawie dostępnych danych. (Nie dotyczy to wyłącznie eksploracji danych; klasyczni statystycy również używają wykresów diagnostycznych). Istnieje wiele różnych wykresów diagnostycznych. Wybierasz je na podstawie tego, co jest dostępne w Twoim oprogramowaniu do eksploracji danych, i własnych preferencji. Używasz wykresu wzrostu , który porównuje przewidywania modelu z wyborem losowym.



Wykres jest oparty na prognozach modelu dla 30 procent danych, które zostały odłożone do celów testowych przed utworzeniem modelu. Słupek po lewej stronie pokazuje grupie, że model daje największe zaufanie do „Tak", zmiany właściciela. Z analizy drzewa decyzyjnego wiesz, że ta grupa to nielokalni właściciele nieruchomości. Model przewiduje, że każdy członek grupy będzie „Tak", zmianą właściciela. Dla tej grupy przewidywania są poprawne w 62,5% przypadków. (Poziom ufności odnotowany u podstawy każdego słupka jest taki sam, jak odsetek poprawnych prognoz). W tym modelu na wykresie widoczne są tylko dwa słupki, ale wykresy wzrostu dla bardziej złożonych modeli często mają wiele słupków. Grupa o największej pewności jest zawsze pierwszym słupkiem po lewej stronie, a każdy kolejny słupek ma następną największą pewność siebie. Korzystając z modelu, możesz wybrać 909 z 2282 przypadków (909 nielokalnych + 1373 lokalnych właścicieli) w testowym zbiorze danych, aby przewidzieć w kategorii „Tak", a 62,5% z nich, 549 przypadków, będzie prawdziwymi zmianami własności nieruchomości . Linia przechodząca przez słupki pokazuje, że wybranie losowo 909 przypadków spowodowałoby tylko około 280 prawdziwych zmian właściwości. Tak więc model prawie podwaja twoją skuteczność w przewidywaniu prawdziwych zmian własności. Znajdziesz kilka rodzajów wykresów wzrostu. Wszystkie przedstawiają zalety korzystania z modelu, a nie losowego wyboru, ale mogą różnić się organizacją i wyglądem.

Ocena stanu modelu

Twoim celem eksploracji danych było wykazanie wykonalności wykorzystania modelowania predykcyjnego w odniesieniu do własności zmiany własności poprzez wykazanie, że co najmniej jedna zmienna ma mierzalną wartość predykcyjną do tego celu. Ściśle mówiąc, cel został osiągnięty. ale jeśli nadal masz czas przed upływem terminu, powinieneś wykorzystać ten czas na ulepszenie tego, co zrobiłeś. Osiągnąłeś minimum, które zamierzałeś zapewnić. Ale nie chcesz robić tylko minimum, więc pracujesz dalej. Możesz spróbować tych rzeczy:

* Wróć i przygotuj dane potrzebne dla kilku innych czynników, które zasugerowali Virginia i Matt.

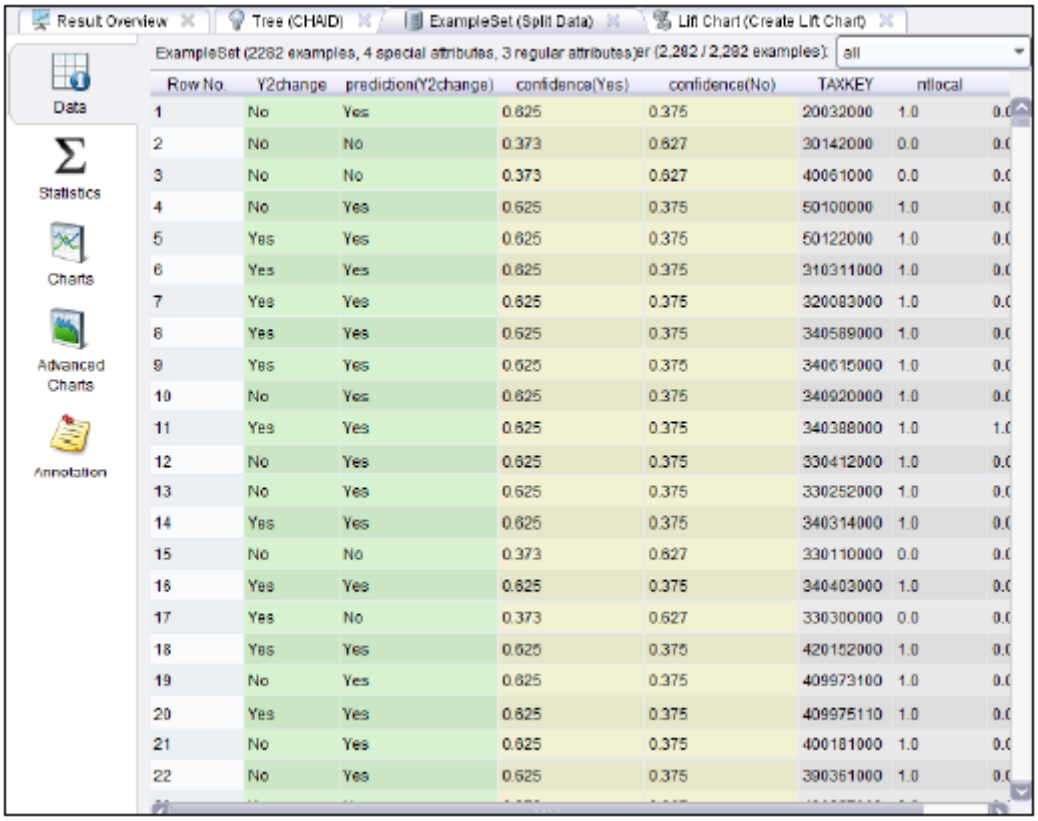
* Eksperymentuj z alternatywnymi typami modeli.

* Udoskonal ustawienia modelu.

Dokumentujesz swoje dotychczasowe osiągnięcia, a następnie wracasz do pracy, aby zbudować najlepszy możliwy model przed terminem zakończenia projektu.

Wprowadzanie wyników w życie

W ciągu jednego dnia nie udało Ci się zbudować modelu, który byłby gotowy do użycia w codziennej działalności. W porządku; to nigdy nie było twoim celem. Ale już pokazałeś, że modelowanie predykcyjne jest wykonalne, a to cholernie dobre jak na jeden dzień. Ponieważ pokazałeś, że modelowanie jest realistyczną opcją, istnieje prawdopodobieństwo, że klient będzie chciał, abyś kontynuował i zbudował najlepszy model, jaki możesz. Kiedy będzie gotowy, uruchomisz go, wykonując prognozy. Zaczniesz od sporządzenia list usług, które prawdopodobnie zmienią właściciela. Właściwie to już stworzyłeś jeden z nich. Znajduje się w danych wyjściowych narzędzia do tworzenia wykresów.



Row No.	Y2change	prediction(Y2change)	confidence(Yes)	confidence(No)	TAXKEY	nllocal	
1	No	Yes	0.625	0.375	20032000	1.0	0.0
2	No	No	0.373	0.627	30142000	0.0	0.0
3	No	No	0.373	0.627	40061000	0.0	0.0
4	No	Yes	0.625	0.375	50100000	1.0	0.0
5	Yes	Yes	0.625	0.375	50122000	1.0	0.0
6	Yes	Yes	0.625	0.375	310311000	1.0	0.0
7	Yes	Yes	0.625	0.375	320083000	1.0	0.0
8	Yes	Yes	0.625	0.375	340589000	1.0	0.0
9	Yes	Yes	0.625	0.375	340615000	1.0	0.0
10	No	Yes	0.625	0.375	340920000	1.0	0.0
11	Yes	Yes	0.625	0.375	340388000	1.0	1.0
12	No	Yes	0.625	0.375	330412000	1.0	0.0
13	No	Yes	0.625	0.375	330252000	1.0	0.0
14	Yes	Yes	0.625	0.375	340314000	1.0	0.0
15	No	No	0.373	0.627	330110000	0.0	0.0
16	Yes	Yes	0.625	0.375	340403000	1.0	0.0
17	Yes	No	0.373	0.627	330300000	0.0	0.0
18	Yes	Yes	0.625	0.375	420152000	1.0	0.0
19	No	Yes	0.625	0.375	409973100	1.0	0.0
20	Yes	Yes	0.625	0.375	409975110	1.0	0.0
21	No	Yes	0.625	0.375	400181000	1.0	0.0
22	No	Yes	0.625	0.375	390361000	1.0	0.0

Dla każdej właściwości wymienionej w danych istnieje prognoza. W przyszłości możesz skorzystać z innych opcji, aby prognozować takie jak te poza oprogramowaniem do eksploracji danych, a nawet zintegrować funkcje przewidywania ze zwykłymi aplikacjami biznesowymi

Współpraca, aby osiągnąć swoje cele

Jako eksplorator danych maksymalnie wykorzystasz swoją wiedzę biznesową. Twoje zrozumienie pochodzenia i implikacji danych jest nieocenionym atutem. Nie oznacza to jednak, że powinieneś pracować sam. Jesteś tylko jedną osobą. Nie możesz zrobić wszystkiego i nie możesz wiedzieć wszystkiego. Łącząc własną wiedzę i umiejętności z talentami rówieśników w uzupełniających się rolach, możesz lepiej zrozumieć stojące przed Tobą wyzwania i uzyskać bardziej realistyczny obraz tego, jakie rozwiązania mogą (lub nie) być możliwe. A odpowiedni zespół może zapobiec wpadnięciu w pułapki związane z procesami biznesowymi, etyką, a nawet prawem. Ten rozdział przedstawia korzyści płynące ze współpracy z innymi.

Nic nie może być lepsze niż eksploracja danych

Jeśli przeczytałeś kilka doniesień prasowych na temat eksploracji danych, możesz odnieść wrażenie, że jest to bardziej złożone niż operacja mózgu. Nie jest. Być może słyszałeś, że eksploratorzy danych mogą dowiedzieć się o Tobie rzeczy, których sam nie znasz. To mało prawdopodobne. Być może słyszałeś, że potrzebujesz doktoratu i ryzy danych, aby rozpocząć eksplorację danych, a to śmieszne.

Możesz być eksploratorem danych

Eksploracja danych to coś, co ludzie w wielu zawodach włączyli do swojej pracy, aby uzyskać lepsze informacje do podejmowania codziennych decyzji biznesowych. Eksplorację danych można zastosować w dowolnej dziedzinie, a wielu prawdziwych eksploratorów danych przyniosło pozytywne zwroty ze swoich pierwszych projektów. Więc kto może być eksploratorem danych? Możesz. Eksploracja danych nie jest wyłączną domeną osób z zaawansowanymi stopniami. Nie musisz być ekspertem od statystyk ani mieć na wyciągnięcie ręki ogromnej ilości danych. Eksploracja danych jest przeznaczona dla osób, które dobrze rozumieją własny biznes i związane z nim wyzwania, nie mają nic przeciwko zwykłym informacjom (takim jak korzystanie z aplikacji biurowych i innego oprogramowania biznesowego) i dobrze orientują się w liczbach (takich jak umiejętność poprawnie interpretować wykresy i tabele). Eksplorator danych również potrzebuje cierpliwości i czasu, aby poświęcić się temu procesowi. Eksploracja danych jest szybka w porównaniu z alternatywami, ale nie jest natychmiastowa. Zainspiruj się tymi prawdziwymi sukcesami w eksploracji danych:

* Bezpieczeństwo publiczne: Straż Pożarna Nowego Jorku wykorzystuje eksplorację danych do identyfikacji czynników, które narażają budynki na ryzyko pożaru. Eksperci danych zidentyfikowali dziesiątki tych czynników ryzyka i opracowali model, aby uzyskać ocenę ryzyka pożaru dla ponad 300 000 budynków w Nowym Jorku. Inspektorzy wykorzystują te wyniki, aby zdecydować, które budynki skontrolować jako pierwsze. Ich celem jest zmniejszenie liczby pożarów i ochrona życia nowojorczyków.

* Handel detaliczny: Amazon.com wykorzystuje eksplorację danych dzięki swoim rozległym zasobom danych, aby zapewnić zindywidualizowane rekomendacje produktów każdemu ze swoich klientów. Ten gigant handlu detalicznego wykorzystuje dane nie tylko do decydowania o oferowanych produktach. Testuje również każdy funkcjonalny i kosmetyczny aspekt swojej strony internetowej i poczty e-mail, aby odkryć szczegóły, które zwiększają sprzedaż.

* Badania medyczne i ankietowe: Palenie zagraża życiu i zdrowiu milionów Amerykanów. Partnerstwo zainteresowań naukowych i handlowych Centers for Disease Control wykorzystało eksplorację danych w połączeniu z badaniami ankietowymi, aby zidentyfikować komunikaty, które mogłyby skutecznie zniechęcić młodzież do palenia, i wykorzystało te informacje jako podstawę kampanii reklamowej przeciwko paleniu.

Wykorzystując posiadaną wiedzę

Aby zostać eksploratorem danych, odkryjesz nowe rzeczy. Znajdziesz nowe metody analizy danych, proces eksploracji danych oraz sposoby oceny i testowania swoich odkryć. Wypróbujesz nowe narzędzia. Poszerzysz swoje zasoby pozyskiwania danych, niezależnie od tego, czy tworzysz je nowe, czy otrzymujesz je ze źródła rządowego lub komercyjnego. Ale masz już najcenniejszy zasób do samodzielnej eksploracji danych , znajomość Twojej firmy. Wiesz, kto co i jak robi. Wiesz, jak pozyskiwane są Twoje dane. Wiesz dużo o możliwych rozwiązaniach Twoich problemów. Żadna matematyka, komputer czy oprogramowanie nie zastąpi tych informacji. Wiesz też coś o tym, kto jest kim w Twojej organizacji. A to oznacza, że możesz skorzystać z jeszcze bardziej obszernego repozytorium odpowiedniej wiedzy biznesowej, wiedzy przechowywanej w umysłach Twoich współpracowników i innych kolegów. To najcenniejszy zasób dostępny do eksploracji danych, który już należy do Ciebie.

Poszukiwacze danych dobrze bawią się z innymi

Skuteczna eksploracja danych wymaga współpracy z odpowiednikami z całej organizacji. Będziesz potrzebować informacji i zasobów od wielu innych osób, a oni również będą na Tobie polegać. Jest to część tego, co sprawia, że eksploracja danych jest tak interesująca i co sprawia, że jest ona kluczowa dla sukcesu i wpływu Twoich projektów.

Współpraca to konieczność

Możesz pracować sam i robić wszystko po swojemu, jeśli:

* Nie oczekuj zapłaty za swoją pracę

* Nie będziesz potrzebować żadnych zasobów, których jeszcze nie posiadasz

* Nie oczekuj, że ktoś poważnie potraktuje Twoje wyniki

Ale żaden prawdziwy eksplorator danych nigdy nie był w takiej sytuacji.Świat eksploratora danych jest interdyscyplinarny. Nie chodzi tylko o to, że będziesz integrować własną wiedzę biznesową z informatyką i analizą. Będziesz wspierać i polegać na innych w całej swojej organizacji w stopniu, jakiego nigdy wcześniej nie doświadczyłeś. Będziesz potrzebować danych. Będziesz potrzebować zasobów obliczeniowych. Musisz wiedzieć, jak działają inne jednostki biznesowe i dlaczego. Co najważniejsze, będziesz potrzebować uwagi i szacunku decydentów, aby wprowadzić swoje odkrycia w życie. A to oznacza, że będziesz musiał zdobyć pomoc, szacunek i zaufanie innych. Eksploracja danych nie jest dla samotników; to przedsięwzięcie zespołowe.

Zróżnicowany zespół pomaga Ci odnieść sukces

Firma ubezpieczeniowa zaprosiła mnie do przedstawienia prezentacji na temat wartości eksploracji danych i zademonstrowania, co mogę zrobić z ich danymi. Firma dostarczyła przykładowe dane i niewiele więcej. Nie miałem żadnych informacji o konkretnych problemach, z jakimi się borykałem, żadnych informacji o kosztach lub potencjale przychodów, ani możliwości aby porozmawiać z pracownikami firmy ubezpieczeniowej na początku projektu. Przy tak małej ilości informacji moje szanse na zademonstrowanie ekscytujących wyników były niewielkie. Wiedziałem, że mój klient rozmawiał również z innymi konsultantami i że wszyscy skupią się na wykrywaniu oszustw. Oszustwa to duży problem w branży ubezpieczeniowej, a wielu konsultantów i dostawców oprogramowania specjalizuje się w wykrywaniu oszustw. Ale jest to trudny przykład do analizy, zwłaszcza gdy nie masz możliwości sprawdzenia na pewno, które z przykładowych przypadków są fałszywe. Chciałem znaleźć inny sposób tworzenia wartości. Wpadłem na pomysł, jak można wykorzystać dane do wykazania

możliwości ulepszeń istniejących procesów biznesowych. Ale nie wiedziałem, czy kierownictwo firmy ubezpieczeniowej będzie się przejmować. Zrobiłem więc rozeznanie i rozmawiałem z informatorem z branży, który potwierdził moje podejrzenia, że codzienne koszty rozpatrywania roszczeń przewyższają straty w wyniku oszustw. (W rzeczywistości zrobił to w bardzo mocnych i barwnych słowach, które nie będą tutaj cytowane). Następnie potrzebowałem niewielkiej pomocy, aby uzyskać oszacowanie kosztów związanych z obsługą roszczeń ubezpieczeniowych. W tym celu zwróciłem się do zasobu, który zaniedbuje wielu eksploratorów danych: lokalnego bibliotekarza referencyjnego. Mając te informacje pod ręką, przygotowałem historię do mojej prezentacji, która stanowi przekonującą argumentację za dobrym zwrotem z inwestycji w eksplorację danych. Podczas gdy wszyscy moi konkurenci mówili o oszustwach, oszustwach i oszustwach, byłem jedynym eksploratorem danych, który przedstawił historię o tym, jak eksplorację danych można wykorzystać do usprawnienia rutynowych procesów biznesowych w celu obniżenia kosztów i poprawy obsługi wszystkich klientów. Różnorodne umiejętności i wiedza mojego zespołu (eksploratora danych, eksperta branżowego i bibliotekarza) sprawiły, że moja prezentacja była wyjątkowa. Zróżnicowany zespół usprawnia twoją pracę, wprowadzając różne umiejętności, doświadczenie i punkty widzenia, dodając głębi i szerokości do informacji, które możesz uzyskać, zwiększając świadomość potencjalnych pułapek, których możesz nie przewidzieć samodzielnie, i stymulując kreatywne myślenie.

Och, ludzie, których spotkasz!

Jako eksplorator danych Twoje miejsce na schemacie organizacyjnym może znajdować się w specjalnej grupie poświęconej analityce lub w dowolnej konwencjonalnej jednostce biznesowej. Bez względu na to, gdzie się znajdujesz, czy zajmujesz się eksploracją danych, czy zajmujesz się tym na pełen etat, będziesz najbardziej produktywny, jeśli znasz role innych jednostek biznesowych i będziesz w dobrych stosunkach z odpowiednimi pracownikami w każdym z nich.

Marketing i sprzedaż

Kiedy firmy decydują się na próbę eksploracji danych, siłą napędową zwykle jest marketing. Dla większości eksploratorów danych pierwszy projekt to projekt marketingowy. (A marketing to nie tylko biznes. Organizacje non-profit i agencje rządowe pełnią podobne role). Poznaj zakres funkcji marketingowych i sprzedażowych, w których pracujesz. W większości przypadków marketerzy są odpowiedzialni za przekształcanie członków ogółu w potencjalnych klientów, a sprzedawcy są odpowiedzialni za przekształcanie tych potencjalnych klientów w zamknięte sprzedaże. Można jednak znaleźć odmiany, szczególnie dla firm zajmujących się sprzedażą online lub katalogową. Wielu marketerów ma pewne doświadczenie z tradycyjną analizą danych, więc nie będzie dla nich wielkim skokiem zrozumienie ważnych koncepcji eksploracji danych. Zrozumieją Twoje pytania i ich rozumowanie. I prawdopodobnie wskażą pewne problemy, które byś przeoczył. Możesz oczekiwać, że będą zadawać trudne pytania dotyczące twoich procesów, i powinieneś poważnie rozważyć te pytania. Marketerzy są zazwyczaj również dobrymi komunikatorami, więc możesz dowiedzieć się od nich wiele o firmie.

Administracja biznesowa i finanse

Ludzie zajmujący się finansami nie zbliżają się do eksploracji danych prawie tak często, jak marketerzy, ale kiedy to robią, zwracaj uwagę. To są ludzie, którzy kontrolują przepływ pieniędzy w organizacji. Eksperci finansowi mogą być szczególnie cenni, pomagając zrozumieć, które potencjalne rozwiązania problemu są możliwe, a które nie. Mogą dostrzec problemy z przepływem pieniężnym, księgowością i problemami prawnymi, które mogą nie być oczywiste dla innych. Chociaż finanse i inne jednostki administracyjne nie są często sponsorami wykonawczymi projektów data-miningu, dbają o wyniki. Dyrektorzy finansowi (CFO) mogą być silnymi zwolennikami eksploracji danych, jeśli widzą konkretny

związek ze zwiększonymi przychodami, oszczędnościami kosztów lub lepszym przepływem środków pieniężnych.

Rozwój produktu

Twórcy produktów mogą być twórcami produktów fizycznych lub wirtualnych, a nawet usług. Mogą to być inżynierowie, programiści, projektanci, menedżerowie produktu lub dowolna z długiej listy innych specjalności. Twórcy produktów posiadają bezcenną wiedzę! Wiedzą, co mają , co mogą i nie mogą zrobić i dlaczego. Wiedzą, ile czasu zajmuje wyprodukowanie rzeczy. Wiedzą, ile pracy jest w to zaangażowane i jakie umiejętności są wymagane. Znają zasady związkowe i inne zasady pracy. Wiedzą, dlaczego coś zostało zrobione w określony sposób i czy zmiany są technicznie wykonalne. Przekonasz się również, że członkowie zespołu programistów mają wiedzę o danych, których nie jesteś świadomy. Inżynier produktu mógł poświęcić wiele godzin na przeglądanie roszczeń gwarancyjnych. Projektant mógł przeprowadzić wywiady (lub nawet nagrać wideo) z użytkownikami w terenie. Inżynier oprogramowania może prowadzić osobisty plik żądań nowych funkcji.

Technologia informacyjna

Eksperci danych absolutnie, zdecydowanie muszą mieć konstruktywne partnerstwo robocze z zespołem ds. technologii informacyjnej (IT), aby dobrze wykonać swoją pracę. To smutne, że w wielu przypadkach te dwie role nie współgrają ze sobą. Nie jest niczym niezwykłym, że technologia informacyjna i eksploracja danych (lub jakikolwiek inny zespół zajmujący się analizą danych) mają wręcz wrogie relacje. Ludzie opierają się pracy w IT z wielu powodów. Dostęp do danych za pośrednictwem zatwierdzonych kanałów jest często wolniejszy niż by sobie tego życzyli analitycy. Dział IT może narzucić zasady dotyczące dostępu do danych, ich wykorzystywania lub udostępniania. I mogą wymagać trochę elektronicznej papierkowej roboty. To wszystko, co wielu analityków danych postrzega jako stratę czasu. A IT też nie zawsze umiera, żeby się z nami zająć. Eksploracja danych czasami wymaga dużej ilości danych (przedstawiciel obsługi klienta otwiera jedną sprawę na raz, podczas gdy eksploracja danych może użyć tysięcy lub więcej). Jedno duże zapytanie od eksploratora danych może sparaliżować codzienne operacje. Dlatego eksploratorzy danych na całym świecie stosują obejścia, aby uniknąć zajmowania się IT. Uzyskują dane z dowolnego źródła, często bez jasnego zrozumienia źródła lub problemów z jakością. Wtedy nie dzielą się wynikami. Dlaczego nie? Nie chcą, żeby ktokolwiek zadawał pytania. W jaki sposób to zachowanie wspiera podejmowanie decyzji w oparciu o dane? Słabo, bardzo słabo. Jeśli eksploracja danych ma mieć naprawdę znaczący wpływ, eksploratorzy danych muszą radzić sobie z IT, ponieważ

* Kierownictwo nie może zastosować wyników analizy, jeśli szczegóły nie są dostępne dla osób zarządzających Twoim działem IT.

* Dane i analizy nie są twoją własnością osobistą. Należą do twojej organizacji. Musisz się podzielić.

* Magazyny danych są coraz większe. Możesz potrzebować dużej ilości danych, których nie można potajemnie przechowywać w osobistym pliku.

* Kierownictwo (przynajmniej w niektórych miejscach) staje się na tyle sprytne, by pytać o szczegóły. Będziesz musiał udokumentować, gdzie, kiedy i w jaki sposób Twoje dane zostały pozyskane.

* Omijanie IT naraża Cię na ryzyko naruszenia przepisów dotyczących prywatności danych lub niewywiązania się z innych ważnych zobowiązań biznesowych. .

Planując projekty związane z eksploracją danych, porozmawiaj z działem IT o tym, co chcesz osiągnąć. Uzyskaj informacje zwrotne na temat problemów z zarządzaniem danymi, z którymi będziesz się

borykać, oraz swoich zobowiązań dotyczących prywatności danych i innych kwestii. Jeśli ktoś z działu IT powie Ci, że jest problem z uzyskaniem potrzebnych danych, zapytaj o przyczyny i uważnie wysłuchaj odpowiedzi. Możesz poprosić o zrobienie czegoś, co narusza prawo lub zobowiązanie umowne. Wyjaśnij swoje cele i zapytaj o możliwe alternatywy. Rozpocznij te rozmowy na początku procesu, abyś nie musiał podejmować zobowiązań, których później odkryjesz, że nie możesz dotrzymać.

Rozpoczęcie rozmowy z IT

Jeśli masz szczęście, nie ma historii konfliktów między analitykami danych (eksploratorami danych, statystykami, marketerami lub innymi rolami) a działem IT w Twojej firmie. Ale możesz nie mieć tyle szczęścia. Tak czy inaczej, postaraj się dotrzeć do budowania atmosfery współpracy i szacunku z IT. Od tego zależy Twoje utrzymanie. Zacznij od obiadu. Tak, dobrze to przeczytałeś. Zacznij od lunchu, swojej uczty. Zaproś swoich współpracowników IT na lunch z zespołem, zamów kilka pizzy i porozmawiaj. Nie musisz rozmawiać o interesach za pierwszym razem; po prostu bądź miły i daj im szansę, aby zrobili to samo. Po obiedzie poproś o wycieczki po obszarze IT, porozmawiaj o funkcjach IT, prywatności danych i o tym, co musisz wiedzieć. Zapytaj o problemy, z jakimi borykają się Twoi odpowiednicy w IT, a zobaczysz je inaczej. (Pamiętasz, że nie mogłeś uzyskać danych? Ktoś z działu IT zostałby zwolniony, gdybyś to zrobił.) Zaproponuj, że zrobisz to samo, aby personel IT mógł zrozumieć, co robisz. Okazuj szacunek i rozmawiaj jak cywilizowani ludzie, a zbudujesz dobre partnerstwo.

Rozwiązanie problemu prywatności danych

Morgan Hunter, COO i współzałożyciel Intreis, integratora rozwiązań specjalizującego się w zarządzaniu usługami i automatyzacji zgodności, rozumie, dlaczego uzyskanie dostępu do żądanych danych może być tak trudne. W jej własnych słowach: „Przez większość czasu, gdy pracowałam w IT, koncentrowałam się na zarządzaniu ryzykiem i zgodności. W gruncie rzeczy jestem osobą, która wymyśla wszystkie te szalone formy, zasady i procesy. Pracując w wielu firmach zajmujących się badaniem rynku, jestem szczególnie wyczulony na to, jak firmy wykorzystują dane, w szczególności, dane klienta. Istnieją setki przepisów regulujących postępowanie z danymi chronionymi (opieka zdrowotna, finanse, prywatność danych itp.), a kiedy zbierasz i przetwarzasz te dane w imieniu swoich klientów, musisz mieć kontrolę (przeczytaj zasady), aby odpowiednio chronić dane i musisz śledzić i zarządzać tym, kto ma dostęp do tych danych.” Morgan wyjaśnia, że rozmowa na temat uzyskiwania potrzebnych danych może oznaczać omówienie takich kwestii:

* Jakich danych potrzebujesz? Opisz swoje wymagania jak najdokładniej.

* Kto jest pierwotnym właścicielem danych? Na przykład, czy te dane zostały nam dostarczone przez klienta do wykorzystania w projekcie? Co to był za projekt? Jakie są wymagania dotyczące prywatności danych określone w umowie? Jeśli dane zostały zebrane wewnętrznie, czy zawierają jakiekolwiek informacje umożliwiające identyfikację osoby (PII)?

* Kto w Twoim zespole zajmowałby się tymi danymi? Mogą istnieć ograniczenia dotyczące tego, kto może obsługiwać jakie dane.

* Dokąd trafiają dane? Laptop, Internet, chmura, przekraczanie granic międzynarodowych, e-mail – wszystko to ma legalne i inne implikacje.

* Jeśli dane zostaną przeniesione, w jaki sposób kopie zostaną zniszczone po zakończeniu projektu? (Tak, możesz mieć obowiązek prawny, aby to zrobić.)

Współpraca z kadrą kierowniczą

Większość eksploratorów danych zajmuje stanowiska kierownicze lub menedżerskie pierwszej linii. Możesz przejść przez całe życie, nie napotykając kierownika, który jest praktycznym eksploratorem danych. Dlaczego? Czy eksploracja danych nie przemawia do kadry kierowniczej? Przeciwnie. Dyrektorzy, którzy widzą pokazy użycia narzędzi do eksploracji danych, są często bardzo zaangażowani. Lubią przepływ pracy i intuicyjną atrakcyjność grafiki. Rzadko jednak sami zajmują się eksploracją danych. Sprowadza się to do podstawowych wymagań, które eksploratorzy danych muszą mieć cierpliwość i czas, aby poświęcić proces eksploracji danych. Nie każdy dyrektor jest cierpliwy i żaden z nich nie ma wolnego czasu. Dlatego cię potrzebują. Firmy płacą za analitykę, aby wspierać podejmowanie decyzji, dostarczać im informacji, które dają firmie najlepsze szanse na największe zyski. Jednak dyrektorzy wyższego szczebla często ignorują lub nie doceniają dostępnych im analiz. Eksperci danych robią dwie rzeczy, aby zniechęcić kadrę kierowniczą i możesz uniknąć obu z nich. Po pierwsze, eksploratorzy danych nie zawsze skupiają się na problemach, które dotyczą kadry zarządzającej, a po drugie, przedstawiają swoje wyniki w niewłaściwy sposób. Następne dwie sekcje pomogą ci dostosować koncentrację do swojego kierownika. Następnie idziesz do przodu i dowiadujesz się, jak maksymalnie prezentować wyniki wpływ.

Pozdrowienia i prośby

Czy kiedykolwiek wszedłeś do sklepu i zauważyłeś, że sprzedawca próbował Ci coś sprzedać, zanim zadał Ci pytanie o to, czego chciałeś? Niektórzy dealerzy samochodowi pokazują każdej kobiecie lusterko do makijażu samochodu, nawet kobietom, które nie noszą żadnego makijażu. Niektórzy sprzedawcy komputerów nie mogą się powstrzymać przed pokazaniem najnowszej „naprawdę słodkiej" maszyny do gier wszystkim, nawet osobom, które nie wyraziły najmniejszego zainteresowania grami. Okropne, co? Oto coś okropnego: niektórzy eksploratorzy danych traktują kierownictwo w podobny sposób. Nie pytają, co jest ważne; zakładają. I zwykle źle zakładają. Każde śledztwo w zakresie eksploracji danych powinno rozpocząć się od dobrej, szczerej rozmowy między tobą a twoim kierownikiem, osobą, która będzie podejmować decyzje w przyszłości. Tylko nie nazywaj tego rozmową od serca do serca. Nazwij to odprawą projektową lub coś takiego korporacyjnego. To spotkanie pokazuje twój szacunek dla autorytetu decydenta i robi coś jeszcze ważniejszego. Zapewnia zrozumienie tego, co jest naprawdę cenne dla tego menedżera. To zrozumienie pokieruje twoją pracą i zapewni, że otrzymasz informacje o prawdziwej wartości. Zadawaj pytania i poważnie słuchaj odpowiedzi. Wyjaśnij od razu rzeczy, których nie rozumiesz, ponieważ możesz nie mieć kolejnej okazji, aby ponownie porozmawiać z dyrektorem przed zakończeniem pracy. Zadawaj pytania otwarte, aby dać dyrektorowi szansę poruszenia kwestii, które zaniedbałeś. A oto dobra sztuczka: Zapytaj, czy są pytania, które powinieneś był zadać, a nie zadałeś. Co zrobić, jeśli nie masz możliwości spotkania się z decydentem od samego początku lub jeśli potrzebujesz informacji później, ale nie możesz skontaktować się z decydentem? Następnie porozmawiaj z dobrze poinformowanymi osobami z całej organizacji, zadając pytania w ten sam sposób.

Ustal swoje priorytety

Po zrozumieniu zagadnień, które najbardziej interesują decydenta, możesz rozpocząć planowanie projektu eksploracji danych. Pomysł tutaj jest dość prosty: włóż swój wysiłek w kwestie, które dyrektor uważa za ważne. Dopasuj swoje priorytety do decydentów. Chociaż ten pomysł nie jest skomplikowany, wielu eksploratorom danych jest to trudne. Być może nie udaje im się odkryć preferencji decydenta, nie zgadzają się z priorytetami wykonawczymi lub głęboko angażują się w odkrytą po drodze kwestię poboczną. To są formuły na krótką karierę w eksploracji danych. Ale co, jeśli odkryjesz coś nieoczekiwanego, coś, co może być naprawdę ważne? Najpierw zrób to, do czego się

zobowiązałeś, i wykonaj pracę, którą obiecałeś, pracę, której chce dyrektor. Jeśli masz wolny czas, wykorzystaj go do zbadania nowego problemu, który znalazłeś. Jeśli dokonasz wspaniałego odkrycia, poczekaj, aż główne zadanie zostanie przedstawione i kierownik będzie zadowolony, a następnie przedstaw swoje dodatkowe odkrycie jako bonus. Takie podejście oznacza, że jesteś wyjątkowym eksploratorem danych.

Rozmowa o eksploracji danych z kadrą kierowniczą

Nie zajmujesz się eksploracją danych tylko po to, by bawić się liczbami. Chcesz działania. Chcesz, aby wszystko zostało zrobione dobrze i rozumiesz, że ważne jest, aby decyzje biznesowe opierać na solidnych dowodach z danych. Ale to nie ty masz prawo do podejmowania decyzji. Musisz więc wywierać wpływ na ludzi, którzy mają władzę. Nie oczekuj, że nauczą się Twojego języka; to do Ciebie należy nauczenie się ich. I nie wystarczy mówić prostym językiem. Musisz także docenić uczucia dyrektora. Kierownicy nie wygrywają pracy na loterii. Walczą o nich. Są motywowani do podnoszenia się do ról wykonawczych, ponieważ są ludzie, którzy potrzebują czuć się

* Ważni

* Pewni siebie

* Potężni

Musisz więc grać z tymi elementami osobowości wykonawczej. Rozważ te cztery zasady rozmawiania o eksploracji danych z kierownictwem:

* Jedyne liczby, które interesują kadrę kierowniczą, to liczby ze znakiem dolara z przodu. (Funty, jeny, euro itd. są również dopuszczalne.) Podczas prezentacji dla kadry kierowniczej nie wspominaj o miarach dopasowania modelu, analizie wrażliwości, istotności lub innych terminach technicznych związanych z eksploracją danych. Używaj prostych słów i wyjaśnij wszystko, co możesz, jeśli chodzi o pieniądze:

• Źle: „Opracowałem stabilny model ze zoptymalizowanymi parametrami do oceniania historii kredytowej klientów. Pseudo r-kwadrat wynosi 0,5519, a wszystkie wartości p dla zmiennych wejściowych są mniejsze niż 0,02”. Dlaczego używanie języka technicznego jest złe? Twój menedżer nie zrozumie tego w pełni i nie będzie czuł się pewnie. Twoja wiadomość nie zabrzmi jako solidna podstawa do podejmowania działań.

• Dobrze: „Możemy zidentyfikować dodatkowe 2 procent bazy danych klientów, którzy spełniają nasze kryteria dopuszczalnego ryzyka kredytowego, oferując potencjał w wysokości 2 mln USD dodatkowego przychodu w następnym roku podatkowym”. Teraz mówisz językiem, który dyrektor w pełni rozumie! To budzi zaufanie.

* Kierownicy mają bardzo krótki czas skupienia uwagi: Przejdź do sedna. Najskuteczniejszym sposobem, aby dyrektor znalazł czas dla Ciebie i Twojej wiadomości, jest założenie, że nie będzie miał czasu dla Ciebie. Dźwięk do tyłu? Wiele osób konkurujących o czas i możliwość wyboru najciekawszych i najbardziej fascynujących zajęć sprawia, że dyrektor czuje się ważny. Zmuś kierownika, aby został z Tobą, szybko oferując przekonującą wiadomość. Wyobraź sobie, że dyrektor będzie miał z tobą tylko 60 sekund. Co byś powiedział? Tylko rzeczy, na których najbardziej Ci zależy, prawda? I najbardziej przekonujące powody, by ci uwierzyć. Gdybyś miał pięć minut, dodałbyś więcej informacji, ale i tak byłyby one starannie wybrane. Mając to na uwadze, zaplanuj prezentację. Otwórz z 60 sekundami najważniejszych i przekonujących myśli, które masz do zaoferowania. Następnie krótko dodaj mały szczegół. Przygotuj się na kontynuację kroków. Przestań odpowiadać na pytania. Znaj dobrze swój

materiał i bądź przygotowany na zmianę kolejności lub zagłębianie się w mniej lub bardziej szczegółowe informacje w odpowiedzi na zainteresowanie dyrektora.

* Uważaj na szczegóły.

Znajdziesz dwa typy kierowników: tych, którzy nie interesują się szczegółami i tych, którzy interesują się szczegółami zbyt mocno. Pierwszy typ łatwo się nudzi; jeśli podasz zbyt wiele szczegółów, całkowicie stracisz zainteresowanie dyrektora. Drugi łatwo zboczyć z tropu; dyrektor może zauważyć coś małego i silnie się zaangażować, ale nie w punkcie, w którym próbujesz zrobić. Zachowaj swoje prezentacje, a zwłaszcza wizualizacje, oszczędne i proste. Skoncentruj się na tym, co robisz i informacjach, które go wspierają. Nie wprowadzaj rozpraszających szczegółów. Jeśli próbujesz coś powiedzieć w oparciu o zachowanie 90 procent kupujących, nie wyświetlaj ogromnego wykresu rozrzutu obejmującego 100 procent kupujących na ekranie za Tobą. Wybierz inny rodzaj obrazu, taki, który nie będzie zwracał uwagi na odstające i skrajne wartości i nie odwracał uwagi od twojego punktu widzenia.

* Striptiz przykuwa uwagę lepiej niż pełne ujawnienie. Nic nie sprawia, że dyrektor czuje się silniejszy niż odkrycie. Dyrektor chce czuć się mądry, a nawet genialny. Nie będzie dreszczykiem emocji, gdy będzie świadkiem zrzutu informacji. Na początku ujawnij tylko trochę i pozwól swojemu kierownikowi poczuć się silnym, zadając pytania i traktując odpowiedzi jako odkrycia. Planując prezentację warstwową, daj kierownikowi możliwość zadawania pytań. Przewidziałeś pytania, stworzyłeś ścieżki prowadzące do pytań, przygotowałeś odpowiedzi na pytania (w rzeczywistości pragniesz tych pytań) i wiesz, że odpowiedzi doprowadzą wykonawcę do wniosku, który masz na myśli. Ale nigdy, przenigdy nie ujawniaj swojego planu. Zadawanie pytań otwartych i uzyskanie na nie odpowiedzi umożliwia menedżerowi wyciągnięcie wniosków, wniosków, które należą do samego menedżera. Wiesz, że twoja prezentacja doprowadziła dyrektora do jedynego rozsądnego wyboru, ale zachowaj to w tajemnicy. Niech menedżer czuje się ważny, pewny siebie, a przede wszystkim potężny.

Nauka praw eksploracji danych

Muzycy mają nuty, skale i teorię muzyki. Kierowcy mają przepisy ruchu drogowego. Fizycy mają prawa ruchu Newtona. Każdy zawód ma swoje zasady przewodnie, idee, które nadają strukturę i kierunek w codziennej pracy. Eksploracja danych nie jest wyjątkiem. W tej części poznasz dziewięć podstawowych pomysłów, które pomogą Ci zabrać się do pracy i zostać eksploratorem danych. Oto 9 praw eksploracji danych, które zostały pierwotnie określone przez pioniera eksploracji danych, Thomasa Khabazę. Tutaj pokazujemy, co każde z tych praw oznacza dla Twojej codziennej pracy.

1. Prawo: Cele biznesowe Business

Oto pierwsze prawo eksploracji danych lub „prawo celów biznesowych": Cele biznesowe są źródłem każdego rozwiązania do eksploracji danych. Analizujemy dane, aby znaleźć informacje, które pomogą nam lepiej prowadzić firmę. Czy nie powinno to być mantrą wszystkich analiz danych biznesowych? Oczywiście, że powinno! Jednak początkujący eksploratorzy danych często koncentrują się na technologii i innych szczegółach, które mogą być interesujące, ale nie są zgodne z potrzebami i celami decydentów wykonawczych. Musisz wyrobić w sobie nawyk identyfikowania celów biznesowych przed zrobieniem czegokolwiek innego i koncentrowania się na tych celach na każdym etapie procesu eksploracji danych. To ważne, że to prawo jest na pierwszym miejscu. Każdy powinien zrozumieć, że eksploracja danych to proces mający cel. Prawdziwi górnicy nie bawią się w błocie; podążają metodycznym procesem, aby odkryć określony cenny materiał. Eksperci danych postępują również zgodnie z metodycznymi procesami, aby wyszukać konkretne informacje, których potrzebują.

2. Prawo: Wiedza biznesowa

Oto drugie prawo eksploracji danych lub „prawo wiedzy biznesowej": Wiedza biznesowa jest kluczowa na każdym etapie procesu eksploracji danych. Eksploracja danych daje moc ludziom -przedsiębiorcom - którzy wykorzystują swoją wiedzę biznesową, doświadczenie i spostrzeżenia wraz z metodami eksploracji danych, aby znaleźć sens w danych. Nie musisz być wyrafinowanym statystykiem, aby eksplorować dane, ale musisz wiedzieć coś o tym, co oznaczają dane i jak działa firma. Tylko wtedy, gdy zrozumiesz dane i problem, który musisz rozwiązać, procesy eksploracji danych pomogą Ci odkryć przydatne informacje i wykorzystać je. Eksploracja danych daje użyteczne wyniki tylko w kontekście dostępnych danych. Musisz wiedzieć, co oznaczają dane. (Jeżeli ktoś przesyła ci dane bez etykiet, wyjaśnij, że jesteś eksploratorem danych, a nie magikiem. Musisz wiedzieć, jakie są pola i przypadki). Eksploracja danych nie zastąpi zrozumienia biznesowego. Twoja własna wiedza biznesowa ma większą wartość niż jakiekolwiek narzędzie do eksploracji danych. Narzędzia same w sobie nic nie znaczą; dodają jedynie szybkości i mocy, aby wspomóc twój własny proces myślowy. Jeśli nie wiesz nic o dziedzinie problemu, musisz połączyć siły z kimś, kto ma tę wiedzę.

Model za milion dolarów, którego nikt nie używał

Jeden dobrze nagłośniony przykład pokazuje wiele o tym, co może pójść nie tak, gdy projekt analityczny nie jest zaplanowany tak, aby odpowiadał wszystkim potrzebom organizacji. Zaczęło się od pomysłu, który wydawał się rozsądny: zdefiniuj metrykę, udostępnij dane i zaoferuj nagrodę za algorytm, który spełnia określone kryterium wydajności. Do 2006 roku Netflix, internetowa wypożyczalnia płyt DVD, miała prawie 10 milionów subskrybentów, z których wielu zgłosiło się na ochotnika do ocen oglądanych filmów. Netflix opracował model do przewidywania ocen, ale zastanawiał się, czy inni mogliby stworzyć jeszcze skuteczniejsze modele. Dlatego zasponsorował konkurs, otwarty dla każdego, oferując milion dolarów na pierwszy program, który okazał się co najmniej o 10 procent dokładniejszy niż własny Netflix. Ludzie z Netflixa nie są głupkami. Mają w domu ekspertów analityków i mają duże doświadczenie w korzystaniu z analityki w organizacji. Jednak doświadczenie z nagrodą Netflix

pokazuje nam, że nawet mądrzy ludzie nie są wystarczająco inteligentni, aby przewidywać każdy problem i zapobiegać każdemu problemowi. Netflix udostępnił dane użytkowników wszystkim zainteresowanym rywalizacją o nagrodę. Obrońcy prywatności wskazali, że dane nie były tak anonimowe, jak powinny, a w 2009 roku grupa subskrybentów złożyła pozew zbiorowy przeciwko firmie. I istniały inne problemy. Zasady konkursu nagradzały dokładność, a nie prostotę, więc modele, które otrzymali, nie były ani proste, ani łatwe w obsłudze. W 2009 roku międzynarodowy zespół naukowców o nazwie „BellKor's Pragmatic Chaos" osiągnął 10-procentową poprawę i zdobył nagrodę. Jednak Netflix nigdy nie używał tego algorytmu, po części dlatego, że był zbyt skomplikowany do praktycznego zastosowania, a po części dlatego, że charakter działalności Netflix zmienił się w ciągu ostatnich lat, przez co oceny były mniej ważne niż na początku konkursu. Jak uniknąć kosztownych i krępujących wpadek w analityce? Nie miej złudzeń, że jesteś tak mądry, że wiesz czego wszyscy potrzebują. Wyjdź i porozmawiaj ze wszystkimi zainteresowanymi stronami. Dowiedz się, czego ludzie oczekują od wyników. W jaki sposób kierownictwo wykorzysta te informacje do podejmowania decyzji? Co będzie musiało zrobić dział IT, aby wdrożyć model? Kto może wyjaśnić istotne kwestie dotyczące prywatności? Mapuj wymagania, zaczynając od pożądanego wyniku końcowego, przechodząc wstecz do procesu zbierania danych. Podziel się swoim procesem i zaproś komentarze i krytykę. Możesz być w stanie zaangażować specjalistów, którzy są ekspertami ds. procesów, takich jak menedżerowie produktu lub analitycy biznesowi. Łatwo uwierzyć, że wiesz, co jest ważne i możesz samodzielnie określić najlepsze podejście. Ale mądrzej jest zaakceptować własne ograniczenia, otworzyć umysł i na początku każdego projektu analitycznego dotrzeć do informacji od innych.

3 Prawo: Przygotowanie danych

Oto trzecie prawo eksploracji danych lub „ustawa o przygotowywaniu danych": Przygotowanie danych to ponad połowa każdego procesu eksploracji danych. Tradycyjni statystycy często mają możliwość zebrania nowych danych, aby odpowiedzieć na konkretne pytania badawcze. Mogą stosować rygorystyczne procesy do planowania eksperymentów, projektowania kwestionariuszy badań ankietowych lub w inny sposób gromadzić wysokiej jakości dane, które są dobrze ukierunkowane na określone cele badawcze. Jednak po tym wszystkim nadal spędzają dużo czasu na czyszczeniu i przygotowywaniu danych do analizy. Z drugiej strony, eksploratorzy danych prawie zawsze muszą pracować z dowolnymi dostępnymi danymi. Korzystają z istniejących rejestrów biznesowych, danych publicznych lub danych, które mogą kupić. Możliwe, że wszystkie te dane zostały zebrane w innym celu niż eksploracja danych i bez rygorystycznego planu lub starannego procesu gromadzenia danych. Tak więc eksploratorzy danych spędzają dużo czasu na przygotowaniu danych. Ile czasu? Prawie każdy eksplorator danych przyzna, że poświęca więcej czasu na przygotowanie danych niż na analizę. Niektórzy twierdzą, że 80 lub 90% czasu poświęcają na przygotowanie danych. To nie jest efektowne, ale jest istotnym elementem procesu.

4 Prawo: Właściwy model

Oto czwarte prawo eksploracji danych lub „NFL-DM": odpowiedni model dla danej aplikacji można odkryć tylko eksperymentalnie. To prawo jest również znane pod skrótem NFL-DM, co oznacza, że nie ma darmowego lunchu dla eksploratora danych. Po pierwsze, czym jest model? To równanie, które reprezentuje wzorzec obserwowany w danych. Przynajmniej reprezentuje wzór w szorstki sposób. Matematyczne modele rzeczywistych rzeczy nigdy nie są doskonałe! Jest to fakt życiowy i dotyczy zarówno fizyków jądrowych, jak i eksploratorów danych. Fizyk jądrowy może mieć teorie o mechanizmie leżącym u podstaw konkretnego procesu z życia codziennego. Teorie te mogą skłonić fizyka do wybrania określonego typu modelu matematycznego jako najbardziej odpowiedniego dla konkretnej sytuacji. Jednak eksploratorzy danych nie działają w ten sposób. W eksploracji danych modele są wybierane metodą prób i błędów. Będziesz eksperymentować z różnymi typami modeli.

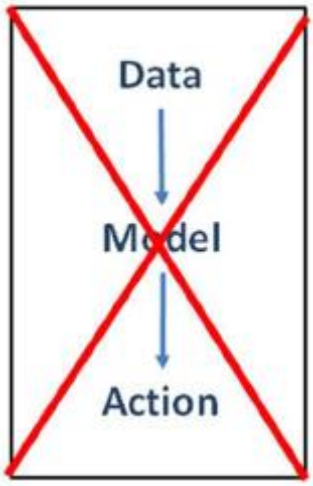
Wybór modeli, które wypróbujesz, będzie zależał od charakterystyki zaangażowanych zmiennych (Czy zmienne są kategoryczne czy liczbowe? Ile masz kategorii?) oraz opcji modelowania dostępnych za pośrednictwem dostępnych narzędzi. Nie będziesz w stanie obronić swojego wyboru modelu w oparciu o teorię. Zamiast tego przetestujesz. Najpierw przetestujesz modele, używając danych, które zarezerwowałeś tylko do testowania. Następnie użyjesz swojego modelu w terenie na małą skalę i uzyskasz nowe dane, aby ocenić, jak dobrze model radzi sobie w świecie rzeczywistym.

5 Prawo: Wzór

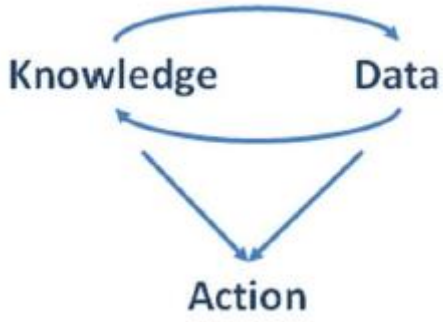
Oto piąte prawo eksploracji danych: zawsze istnieją wzorce. Pomyśl o jakimkolwiek znanym odkrywcy, a zdasz sobie sprawę, że udana eksploracja zaczyna się od celu. Frederick A. Cook i Robert E. Peary zbadali najbardziej wysunięty na północ region planety w poszukiwaniu bieguna północnego. Richard Burton i John Speke badali Afrykę w poszukiwaniu źródła Nilu. Jako eksplorator danych będziesz eksplorować dane w poszukiwaniu przydatnych wzorców. Innymi słowy, będziesz szukał znaczących relacji między zmiennymi w danych. Zrozumienie tych relacji zapewnia lepsze zrozumienie biznesu i lepsze prognozy tego, co wydarzy się w przyszłości. Co najważniejsze, zrozumienie wzorców w danych pozwala wpływać na to, co wydarzy się w przyszłości. Oto przykład: sprzedawca komputerów chciałby zwiększyć marże zysku poprzez kultywowanie sprzedaży dodatkowej. Dealer może zarobić więcej pieniędzy, jeśli kupujący komputer kupią również urządzenia peryferyjne (takie jak drukarki i klawiatury), oprogramowanie i małe przedmioty, takie jak wycieraczki do ekranu komputera. Badasz dane w celu zrozumienia cech klientów, którzy kupują te produkty. Być może odkryjesz, że ludzie, którzy kupują komputery marki Acme, kupują również wiele dodatkowych przedmiotów, więcej niż kupujący jakiejkolwiek innej marki komputerów. To jest wzór, który kieruje poczynaniami krupiera. Zachowaj nabywców komputerów Acme jako klientów, uzyskaj więcej sprzedaży dodatkowej. Przynajmniej takie jest oczekiwanie. Aby to udowodnić, musisz przetestować. Zawsze znajdziesz wzory. Dane zawsze mają ci coś do powiedzenia. Czasami potwierdza, że to, co robisz, jest słuszne. To może nie wydawać się ekscytujące, ale przynajmniej mówi ci, że byłeś na dobrej drodze. W inne dni dane mogą wskazywać, że Twoje obecne praktyki biznesowe nie działają. To ekscytujące i choć na krótką metę może nie być przyjemne, poznanie prawdy jest ważnym krokiem w kierunku poprawy. Podobnie jak inni wielcy odkrywcy, zawsze będziesz mieć na uwadze konkretny cel. Skup się i nie spędzaj dużo czasu na badaniu wzorców, które nie są związane z Twoim celem biznesowym. Krzysztof Kolumb badał oceany w poszukiwaniu lepszej drogi do Azji, ale nigdy jej nie znalazł. W takim przypadku jego kierownictwo i tak było bardzo szczęśliwe. Nie licz na to, że sam będziesz miał takie samo szczęście. Eksploracja danych to wciąż młoda dziedzina, coś zupełnie nowego dla większości ludzi. Możesz być pionierem w swojej dziedzinie, wykorzystując eksplorację danych do badania ważnych dla Ciebie problemów. (I w przeciwieństwie do innych odkrywców wspomnianych w tej sekcji, możesz być odkrywcą we własnym, bezpiecznym, ciepłym biurze.)

6 Prawo: Wzmocnienie

Oto szóste prawo eksploracji danych lub „prawo wglądu”: eksploracja danych wzmacnia percepcję w domenie biznesowej. Tak, sformułowanie tego prawa to rodzaj fantazyjnego schmancy. Ujmę to inaczej: metody eksploracji danych pozwalają lepiej zrozumieć Twój biznes niż mogłeś się bez nich obejść. Jeśli ważne informacje zostały napisane drobnym drukiem, być może będziesz w stanie sam je przeczytać, ale przy pomocy lupy byłoby to łatwiejsze. Jeśli odcisk był bardzo mały, możesz go w ogóle nie zobaczyć, chyba że masz mikroskop. Metody eksploracji danych pomagają Ci jak lupa lub mikroskop, umożliwiając odkrycie efektów, które byłyby trudne lub niemożliwe do wykrycia poprzez zwykłe raportowanie. Eksploracja danych nie jest natychmiastowa.



Odkrywanie i uczenie się poprzez eksplorację danych to proces interaktywny .



Dokonasz odkryć, dowiesz się trochę o każdym z nich i wykorzystasz to, co odkryłeś, aby podjąć działanie. Wyniki każdego działania, które spróbujesz, dadzą więcej danych, a te dane pozwolą Ci zrozumieć coś więcej. To cykl odkrywania, który trwa tak długo, jak długo będziesz eksplorować i eksperymentować.

7. Prawo: Przewidywanie

Oto siódme prawo eksploracji danych lub „prawo przewidywań”: Przewidywanie zwiększa informacje lokalnie poprzez uogólnianie. Tak, kolejny fantazyjny. Oto inny sposób sformułowania tego prawa: Eksploracja danych pomaga nam wykorzystać to, co wiemy, do lepszego przewidywania (lub szacunków) rzeczy, których nie znamy. Klient wchodzi do Twojego sklepu. Ile wyda ten klient? Jeśli nie znasz żadnych szczegółów na temat klienta, najlepszym oszacowaniem jest to, że klient wyda średnią kwotę, którą wydają inni klienci. Ale może wiesz coś więcej. Klient kieruje się do działu elektroniki. To może prowadzić do oczekiwania wyższego poziomu wydatków. A może klient idzie do toalety, co prowadzi do oczekiwania, że nie ma go tam, aby dokonać zakupu. Eksploracja danych wykorzystuje dane i metody modelowania, aby zastąpić nieformalne oczekiwania opartymi na danych, spójnymi i dokładniejszymi szacunkami.

8 Prawo: Wartość

Oto ósme prawo eksploracji danych lub „prawo wartości”: Wartość wyników eksploracji danych nie jest określona przez dokładność ani stabilność modeli predykcyjnych. Eksperci danych nie zajmują się teorią. Jako eksplorator danych możesz nawet nie znać teorii kryjącej się za używanymi modelami statystycznymi. Może tak samo ,cóż, ponieważ w eksploracji danych będziesz używać tych modeli w sposób, który niekoniecznie jest zgodny z teorią, która za nimi stoi. Statystycy zajmują się teorią. W

tym kontekście sensowna jest ocena modeli w oparciu o dokładność (dopasowanie modelu do danych eksperymentalnych) i stabilność (tworzenie spójnej struktury modelu z różnych próbek danych). Dokładność i stabilność to dobre rzeczy, ale model może być zarówno dokładny, jak i stabilny, ale nie oferuje dużej wartości dla firmy. Ty, eksplorator danych, musisz zastosować inne podejście. Będziesz szukał modeli, które dają prawidłowe przewidywania (a do oceny tego użyjesz testowania, a nie teorii statystycznej), tak. Ale możesz bardziej martwić się innymi kwestiami, takimi jak to, czy model ma sens biznesowy, czy Cię oświeci o nieoczekiwanych czynnikach prognostycznych lub praktycznych w Twoim miejscu pracy.

9. Prawo: Zmiana

Oto dziewiąte prawo eksploracji danych, czyli „prawo zmian": wszystkie wzorce mogą ulec zmianie. Świat ciągle się zmienia. Model, który dziś daje świetne prognozy, jutro może być bezużyteczny. To fakt dla wszystkich analityków danych, nie tylko dla eksploratorów danych.

0 Prawo eksploracji danych

Duncan Ross, inny szanowany eksplorator danych, zasugerował dodanie do 9 praw eksploracji danych. Aby zrozumieć Prawo Zero Rossa, potrzebujesz trochę tła. Weźmy pod uwagę naukowca danych, nowy tytuł zawodowy analityka, który jest stosowany w niektórych organizacjach, zwłaszcza w niektórych większych firmach internetowych. Tytuł oznacza różne rzeczy dla różnych osób. Czasami jest to osoba, która ma stopień naukowy w dziedzinie statystyki, ale częściej nie. Zastosowania, doświadczenie, szkolenia i narzędzia są różne. Jedyną stałą jest to, że te role opierają się na umiejętnościach programowania. Niektórzy opisują ich jako po części statystyka, po części programistę i po części gawędziarza, a czasami potrzeba kilku dodatkowych części, tworząc nierealistyczny ideał dla zawodu. Ale w każdym razie tytuł naukowca danych i koncepcja nauki o danych są gorące. Oto Prawo Zero: 9 praw eksploracji danych jest równie istotnych dla nauki o danych.

P: Czy Prawo Zero jest prawdziwe?

O: W przybliżeniu.

Większość z 9 praw jest uniwersalna dla każdego rodzaju analizy danych. Pomyśl o pierwszym prawie, ustawie o celach biznesowych. To podstawa dla każdego analityka danych. Klasyczni statystycy i badacze operacji mogą pracować z myślą o konkretnych celach biznesowych, podobnie jak eksploratorzy danych. Badacze stosujący metody klasyczne mogą mieć miejsce na dyskusję na temat czwartego prawa, które mówi, że właściwy model można znaleźć tylko poprzez eksperymenty. I na pewno masz kilka przemyśleń na temat ósmego prawa, które bagatelizuje dopasowanie i stabilność modelu. Statystycy i naukowcy lubią, gdy ich modele są stabilne, i mają powody do takiej preferencji. Stopień, w jakim zaakceptujesz te części 9 praw, zależy od twojego podejście do analizy danych, a nie nazwę, którą to nazywasz.

Obejmowanie procesu eksploracji danych

Eksploracja danych nie ma oficjalnych zasad. Masz ogromną elastyczność w definiowaniu i udoskonalaniu własnych metod pracy. Mimo to odniesiesz korzyści ze zrozumienia i stosowania podejść, które sprawdzają się w przypadku innych. Międzybranżowy standardowy proces eksploracji danych (CRISP-DM) jest dominującą strukturą procesu eksploracji danych. To otwarty standard; każdy może z niego korzystać.

Czyj to właściwie jest standard?

Model procesu CRISP-DM to krok po kroku podejście do eksploracji danych, które zostało stworzone przez eksploratorów danych dla eksploratorów danych. Uczestnicy z ponad 200 organizacji (głównie zróżnicowana grupa przedsiębiorstw zainteresowanych wewnętrznym wykorzystaniem eksploracji danych lub promowaniem szeroko zakrojonego wykorzystania eksploracji danych) wnieśli wkład do opracowania ram, które określają kluczowe zadania eksploracji danych w kategoriach biznesowych i pozostawiają użytkownicy mogą swobodnie dokonywać własnych wyborów dotyczących określonych podejść matematycznych i obliczeniowych oraz innych kwestii technicznych. Wyjaśnienie procesu CRISP-DM jest bardzo ściśle zgodne z oryginalną opublikowaną wersją. Istnieją jednak różnice, takie jak zmiany terminologii lub diagramu, mające na celu uczynienie informacji bardziej przejrzystymi dla nowicjuszy w eksploracji danych.

Podejście do procesu etapami

Model procesu CRISP-DM ma sześć podstawowych faz. Są to:

1. Zrozumienie biznesu: uzyskaj jasne zrozumienie problemu, który zamierzasz rozwiązać, jego wpływu na Twoją organizację oraz celów związanych z jego rozwiązaniem.

2. Zrozumienie danych: Przejrzyj posiadane dane, udokumentuj je i zidentyfikuj problemy związane z zarządzaniem danymi i jakością danych.

3. Przygotowanie danych: Przygotuj dane do użycia w modelowaniu.

4. Modelowanie: Użyj technik matematycznych, aby zidentyfikować wzorce w swoich danych.

5. Ocena: Przejrzyj wykryte wzorce i oceń ich potencjał do zastosowań biznesowych.

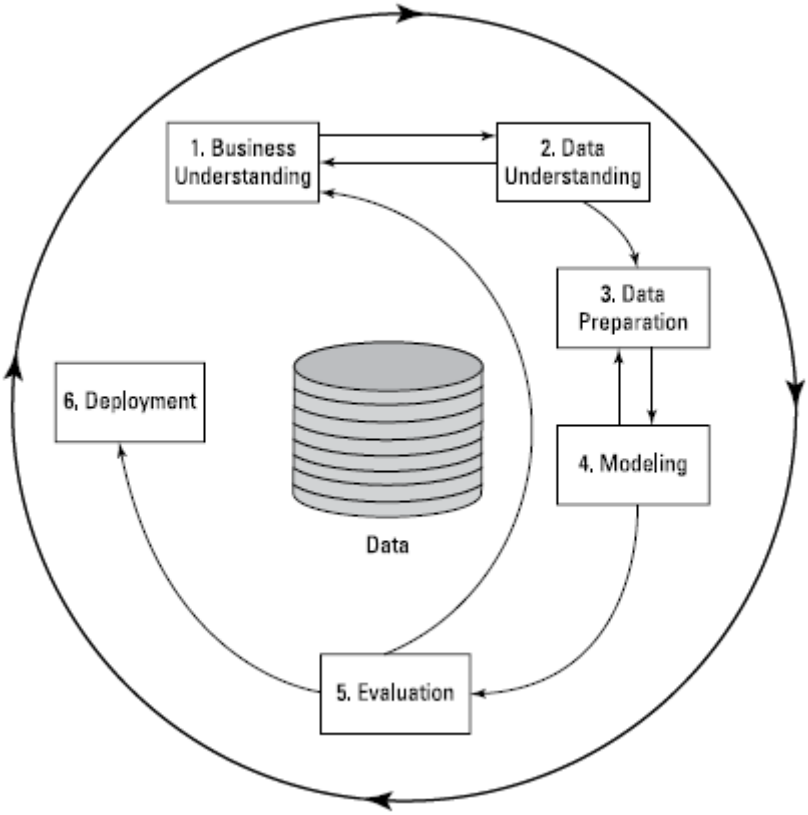
6. Wdrożenie: Wykorzystaj swoje odkrycia w codziennej pracy.

Każda z tych faz obejmuje kilka głównych zadań, a każde zadanie wymaga kilku wyników - przede wszystkim raportów podsumowujących wykonaną pracę i informacje zdobyte w tej fazie procesu eksploracji danych. Jednak CRISP-DM nie definiuje szablonów dla tych wyników. Musisz je zaplanować i stworzyć tak, aby odpowiadały specyficznym potrzebom i stylowi własnego miejsca pracy. CRISP-DM definiuje proces eksploracji danych przede wszystkim z biznesowego punktu widzenia. Mówi dużo o tym, co musisz zrobić, ale nie przedstawia wszystkich szczegółów technicznych.

Cyklicznie przez fazy i projekty

Eksploracja danych nie jest czymś, co robisz raz, a potem zapominasz. To ciągły cykl działań. W każdym projekcie możesz zająć się tylko małym elementem dużego i ważnego problemu, ale będziesz do niego wracać z nowymi projektami. Ponieważ Twoja praca może być również zastosowana w nowych projektach, będziesz często wracać do swoich poprzednich projektów, aby sprawdzić, czy modele, które opracowałeś w przeszłości, są nadal skuteczne i szukać możliwości ulepszenia tego, co zrobiłeś. Recykling pracy w ten sposób minimalizuje wysiłek i pomaga uniknąć zamieszania. Model procesu

CRISP-DM (nie model matematyczny, ale zestaw wytycznych do pracy z eksploracją danych) to cykl często reprezentowany przez diagram podobny do pokazanego na rysunku



Każdy projekt rozpoczyna się od zrozumienia biznesu i przechodzi przez każdą z pięciu faz procesu. W ramach cyklu znajdują się mniejsze cykle, więc możesz wykonać kilka przejść tam i z powrotem, pracując nad zrozumieniem biznesu i danych lub przygotowując dane i budując modele. Cykl powtarza się, gdy ocena projektu i doświadczenie podczas wdrażania zwiększają zrozumienie firmy i inspirują do nowych projektów.

Dokumentowanie Twojej pracy

Kiedy jesteś w trakcie projektu, głęboko zaangażowany w swoje dane i problemy, które zamierzasz rozwiązać, możesz łatwo tak zaznajomić się ze szczegółami, że wydają się one oczywiste. Nie możesz niczego zapisywać, ponieważ nie widzisz żadnej potrzeby. Jednak później, gdy przejdziesz do innych projektów, spędzisz czas na myśleniu o różnych zestawach danych i różnych problemach, a potem wrócisz do tego projektu, przekonasz się, że te szczegóły wcale nie są oczywiste. Będziesz się zastanawiać, co oznaczają twoje rzadkie notatki, nie będziesz mieć pewności, gdzie i w jaki sposób uzyskałeś i przygotowałeś dane, a także znajdziesz inne dziury w swoich informacjach. Nieodpowiednia dokumentacja prowadzi do wielu problemów. Możesz skończyć powtarzając pracę i zmuszając innych do powtarzania pracy. Możesz nie wykryć błędów w swojej pracy. Twój zarząd lub współpracownicy mogą być sfrustrowani (a nawet źli), jeśli nie przygotowałeś potrzebnej dokumentacji. Brak udokumentowania powodów podjęcia określonych decyzji lub dowodu, że spełniłeś obowiązki w zakresie ochrony danych, może mieć nawet konsekwencje prawne. Dlatego duża część modelu procesu CRISP-DM koncentruje się na raportach i innych dokumentach, które eksploratorzy danych tworzą w trakcie swojej pracy. Te dokumenty są Twoim sposobem na zachowanie informacji o tym, co zrobiłeś, dzięki czemu Ty i inni nie będziecie się później zastanawiać.

Zrozumienie biznesowe

W pierwszej fazie projektu eksploracji danych, zanim podejdziesz do danych lub narzędzi, określasz, co chcesz osiągnąć i powody, dla których chcesz osiągnąć ten cel. Faza zrozumienia biznesu obejmuje cztery zadania (działania podstawowe, z których każde może obejmować kilka mniejszych części). Są to

✓ Identyfikacja celów biznesowych

✓ Oceń swoją sytuację

✓ Definiowanie celów eksploracji danych

✓ Tworzenie planu projektu

Zadanie: Identyfikacja celów biznesowych

Pierwszą rzeczą, którą musisz zrobić w każdym projekcie, to dowiedzieć się dokładnie, co chcesz osiągnąć! To mniej oczywiste, niż się wydaje. Wielu eksploratorów danych zainwestowało czas w analizę danych tylko po to, by odkryć, że ich kierownictwo nie było szczególnie zainteresowane kwestią, którą badali. Musisz zacząć od jasnego zrozumienia

✓ Problemu, którym kierownictwo chce się zająć

✓ Celów biznesowych

✓ Ograniczeń (ograniczenia tego, co możesz zrobić, rodzaje rozwiązań, które można zastosować, kiedy praca musi zostać ukończona itp.)

✓ Wpływu (jak problem i możliwe rozwiązania pasują do biznesu)

Wyniki tego zadania obejmują trzy elementy (zwykle krótkie raporty skupiające się tylko na głównych punktach):

✓ Kontekst: Wyjaśnij sytuację biznesową, która napędza projekt. Ta pozycja, podobnie jak wiele następnych, obejmuje tylko kilka akapitów. Oto przykład elementu tła:

Nasz klient, regionalna komisja planowania, stara się wpłynąć na wykorzystanie nieruchomości, aby poprawić jakość życia mieszkańców. Komisja ds. planowania posiada obszerną kartę, która pozwala jej rozważać szeroko zakrojone kwestie, w tym zatrudnienie, rekreację, środowisko i wiele innych aspektów życia społeczności; jednak rola komisji jest czysto doradcza. Ma dużą swobodę w wyborze zagadnień do badania, prowadzeniu badań i formułowaniu zaleceń dotyczących polityki lokalnym prawodawcom i pracownikom, ale nie ma niezależnych uprawnień do ustanawiania przepisów ani wpływania na właścicieli nieruchomości. Członkowie komisji (oraz inne osoby z samorządów lokalnych i organizacji obywatelskich) uważają, że najlepszą okazją do wpływania na sposób użytkowania nieruchomości jest to, że nieruchomość przechodzi z rąk do rąk. Oznacza to, że wysiłki władz lokalnych w zakresie planowania mogą osiągnąć największy wpływ, koncentrując się na nieruchomościach, które wkrótce zmienią właściciela. To stwarza problem: najlepszy czas na działanie to czas, zanim nieruchomość zmieni właściciela, ale samorząd lokalny nie ma wiarygodnych informacji o tym, które nieruchomości mogą zostać przekazane. (Wykazy nieruchomości komercyjnych mogą być przydatne, ale nie obejmują wszystkich przeniesień własności, a najlepszym momentem na działanie może być to, zanim nieruchomość zostanie umieszczona na liście). obejmują one między innymi własność nielokalną, wielokrotne naruszenia kodeksu budowlanego i wykluczenie. Chociaż komisarze mają

powody, by sądzić, że czynniki te wpływają na prawdopodobieństwo zmiany właściciela majątku, ich skutki nie zostały określone ilościowo.

✓ Cele biznesowe: Określ, co Twoja organizacja zamierza osiągnąć w ramach projektu. Jest to zwykle szerszy cel, niż Ty, jako eksplorator danych, możesz osiągnąć niezależnie. Na przykład celem biznesowym może być zwiększenie sprzedaży z kampanii reklam świątecznych o 10 procent rok do roku.

✓ Kryteria sukcesu biznesowego: Określ sposób pomiaru wyników. Postaraj się uzyskać jasno określone ilościowe kryteria sukcesu. Jeśli musisz użyć kryteriów subiektywnych (wskazówka: terminy takie jak uzyskanie wglądu lub radzenie sobie z subiektywnymi kryteriami), przynajmniej uzyskaj zgodę na to, kto dokładnie będzie oceniał, czy kryteria te zostały spełnione.

Zadanie: Ocena twojej sytuacji

W tym miejscu możesz bardziej szczegółowo omówić problemy związane z Twoimi celami biznesowymi. Teraz zagłębisz się w proces ustalania faktów, budując znacznie bardziej mięsiste wyjaśnienie kwestii nakreślonych w zadaniu celów biznesowych. Wyniki tego zadania obejmują pięć szczegółowych raportów:

✓ Inwentaryzacja zasobów: Lista wszystkich zasobów dostępnych dla projektu. Mogą to być osoby (nie tylko eksploratorzy danych, ale także osoby posiadające specjalistyczną wiedzę na temat problemu biznesowego, menedżerowie danych, wsparcie techniczne i inne), dane, sprzęt i oprogramowanie.

✓ Wymagania, założenia i ograniczenia: Wymagania będą obejmować harmonogram realizacji, zobowiązania prawne i dotyczące bezpieczeństwa oraz wymagania dotyczące akceptowalnych ukończonych prac. To jest punkt, aby sprawdzić, czy będziesz miał dostęp do odpowiednich danych!

✓ Ryzyka i nieprzewidziane okoliczności: Zidentyfikuj przyczyny, które mogą opóźnić zakończenie projektu i przygotuj plan awaryjny dla każdego z nich. Na przykład, jeśli przerwa w dostępie do Internetu w twoim biurze może stanowić problem, być może twoją ewentualnością może być praca w innym biurze do czasu zakończenia przerwy.

✓ Terminologia: Stwórz listę terminów biznesowych i terminów związanych z eksploracją danych, które są istotne dla Twojego projektu i zapisz je w słowniku z definicjami (i być może przykładami), aby wszyscy zaangażowani w projekt mogli rozumieć te terminy .

✓ Koszty i korzyści: Przygotuj analizę kosztów i korzyści dla projektu. Spróbuj podać wszystkie koszty i korzyści w dolarach (euro, funtach, jenach itd.). Jeśli korzyści nie przekraczają znacząco kosztów, zatrzymaj się i ponownie rozważ tę analizę i swój projekt.

Decydenci często czują się bardziej komfortowo przeznaczając zasoby na projekty, które zmniejszają koszty, niż te, które mają na celu zwiększenie przychodów, dlatego zawsze szukaj potencjalnych oszczędności i przedstaw możliwości oszczędności w raporcie kosztów i korzyści.

Zadanie: Zdefiniuj swoje cele związane z eksploracją danych

Osiągnięcie celu biznesowego często wymaga działania wielu osób, nie tylko eksploratora danych. Więc teraz musisz zdefiniować swoją małą część w szerszym obrazie. Jeśli celem biznesowym jest na przykład zmniejszenie utraty klientów, Twoimi celami eksploracji danych może być identyfikacja współczynników utraty danych dla kilku segmentów klientów i opracowanie modeli do przewidywania, którzy klienci są najbardziej zagrożeni. Materiały dostarczane do tego zadania obejmują dwa raporty:

✓ Cele eksploracji danych: Zdefiniuj wyniki eksploracji danych, takie jak modele, raporty, prezentacje i przetworzone zbiory danych.

✓ Kryteria sukcesu eksploracji danych: Zdefiniuj techniczne kryteria eksploracji danych niezbędne do wsparcia kryteriów sukcesu biznesowego. Spróbuj zdefiniować je w kategoriach ilościowych (takich jak dokładność modelu lub predykcyjna poprawa w porównaniu z istniejącą metodą). Jeśli kryteria muszą być jakościowe, określ osobę, która dokonuje oceny.

Zadanie: Opracowanie planu projektu

Teraz określasz każdy krok, który ty, eksplorator danych, zamierzasz podjąć, dopóki projekt nie zostanie ukończony, a wyniki zostaną zaprezentowane i przejrzane. Materiały dostarczane do tego zadania obejmują dwa raporty:

✓ Plan projektu: nakreśl swój krok po kroku plan działania dla projektu. Rozwiń zarys harmonogramu wykonania każdego kroku, wymaganych zasobów, danych wejściowych (takich jak dane lub spotkanie z ekspertem w danej dziedzinie) i wyników (takich jak oczyszczone dane, model lub raport) dla każdego etapu oraz zależności (kroki, których nie można rozpocząć przed zakończeniem tego kroku). Wyraźnie zaznacz, że pewne kroki muszą zostać powtórzone (na przykład modelowanie i ocena zwykle wymagają kilku powtórzeń w tę i z powrotem).

✓ Wstępna ocena narzędzi i technik: Zidentyfikuj wymagane zdolności do realizacji celów eksploracji danych i oceń narzędzia i zasoby, które posiadasz. Jeśli czegoś brakuje, musisz rozwiązać ten problem na bardzo wczesnym etapie procesu.

Zrozumienie danych

W drugiej fazie projektu data-miningu, prowadzonego po zdefiniowaniu celów i sporządzeniu planu, pozyskujesz dane i sprawdzasz, czy są one odpowiednie dla Twoich potrzeb. Możesz zidentyfikować problemy, które powodują, że wracasz do zrozumienia biznesu i poprawiasz swój plan. Możesz nawet odkryć błędy w zrozumieniu biznesu, co jest kolejnym powodem do przemyślenia celów i planów. Faza zrozumienia danych obejmuje cztery zadania. Są to

✓ Zbieranie danych

✓ Opisywanie danych

✓ Eksploracja danych

✓ Weryfikacja jakości danych

Zadanie: Zbieranie danych

Właśnie wyznaczyłeś cele i zdefiniowałeś plan eksploracji danych. Każdy krok planu zależy od posiadania odpowiednich danych. Lepiej upewnij się, że naprawdę masz te dane! Dla tego zadania istnieje tylko jeden produkt: wstępny raport dotyczący gromadzenia danych. W swoim raporcie musisz zweryfikować, czy uzyskałeś dane lub przynajmniej uzyskałeś dostęp do danych, przetestowałeś proces dostępu do danych i zweryfikowałeś, że dane istnieją. Będziesz także musiał załadować dane do wszelkich narzędzi, których będziesz używać do eksploracji danych, aby sprawdzić, czy narzędzia są kompatybilne z danymi. Możesz wykonać dużo pracy, aby zebrać potrzebne dane, zanim będziesz mógł napisać ten raport. Najpierw ułożysz swój plan w następujący sposób:

✓ Zarys wymagań dotyczących danych: Utwórz listę typów danych niezbędnych do realizacji celów eksploracji danych. Rozwiń listę o szczegóły, takie jak wymagany zakres czasu i formaty danych.

✓ Sprawdź dostępność danych: Potwierdź, że wymagane dane istnieją i możesz z nich korzystać. Jeśli niektóre z żądanych danych są niedostępne, zdecyduj, w jaki sposób rozwiążesz ten problem. Rozważ alternatywy, takie jak

• Zastąpienie alternatywnym źródłem danych

• Zawężenie zakresu projektu

• Zbieranie nowych danych

✓ Zdefiniuj kryteria wyboru: Zidentyfikuj konkretne źródła danych (bazy danych, pliki, dokumenty itd.), których będziesz używać. W tych źródłach określ tabele, pola i zakresy spraw, które są istotne dla tego projektu. Po wykonaniu tych kroków musisz faktycznie uzyskać dane. Na tym etapie zaimportuj dane do platformy do eksploracji danych, której będziesz używać w projekcie, aby potwierdzić, że jest to możliwe i że rozumiesz proces. W trakcie tej wersji próbnej możesz odkryć ograniczenia oprogramowania (lub sprzętu), których nie przewidziałeś, takie jak:

✓ Ograniczenia liczby przypadków lub pól lub ilości pamięci, której możesz użyć

✓ Brak możliwości odczytania formatów danych Twoich źródeł

✓ Trudności w radzeniu sobie z niedoskonałościami danych (na przykład możesz napotkać produkty, które nie będą importować lub analizować niekompletnych zbiorów danych)

Na koniec podsumuj proces zbierania w raporcie. Raport powinien opisywać Twoje wymagania i szczegółowo wyjaśniać, jakie dane zebrałeś i z jakich źródeł. Tutaj potwierdzasz, że faktycznie uzyskałeś dane i że są one kompatybilne z Twoją platformą do eksploracji danych. Jeśli napotkasz trudności, wyjaśnisz, czym były i jak sobie z nimi radziłeś (korzystając z alternatywnych źródeł, poprawiając plany, zmieniając formaty). Rezultatem tego zadania jest po prostu prosty raport, ale praca, którą musisz wykonać, zanim będziesz mógł napisać ten raport, nie będzie prosta! Dostęp do danych może być jedną z najtrudniejszych i najbardziej frustrujących części procesu eksploracji danych, obfitującego zarówno w wyzwania techniczne, jak i biznesowe.

Zadanie: opisywanie danych

Teraz, gdy masz dane, przygotuj ogólny opis tego, co masz. Produktem dostarczanym do tego zadania jest raport z opisem danych. W nim opisujesz źródło i formaty danych, liczbę spraw, liczbę i opisy pól oraz wszelkie inne ogólne informacje, które mogą być ważne. Dokonujesz również krótkiej oceny przydatności danych do celów eksploracji danych. Na przykład sprawdź, czy dane zawierają pola, których oczekujesz i których musisz tam być, oraz czy jest wystarczająca liczba przypadków do analizy.

Zadanie: Eksploracja danych

W tym zadaniu dokładniej przyjrzysz się danym. Dla każdej zmiennej przyjrzysz się zakresowi wartości i ich rozkładom. Będziesz korzystać z prostej manipulacji danymi i podstawowych technik statystycznych do dalszego sprawdzania danych. Eksploracja danych wspiera kilka celów:

✓ Zapoznaj się z danymi.

✓ Dostrzegaj oznaki problemów z jakością danych.

✓ Ustaw scenę dla kroków przygotowania danych.

Wynikiem tego zadania jest raport z eksploracji danych. To miejsce, w którym można udokumentować wszelkie hipotezy lub wstępne ustalenia, które zostały opracowane podczas eksploracji danych. Raport ten powinien zawierać bardziej szczegółowy opis danych niż raport z opisem danych, w tym rozkłady, podsumowania i wszelkie oznaki problemów z jakością danych.

Zadanie: Weryfikacja jakości danych

Masz dane i zbadałeś je, a teraz musisz określić, czy są wystarczająco dobre, aby wspierać Twoje cele. Często będziesz mieć problem z jakością do rozwiązania, ale nadal będziesz w stanie iść do przodu, ale czasami jakość danych jest tak słaba, że nie jest w stanie obsłużyć twojego planu i będziesz musiał szukać alternatyw. Niektóre z najgorszych problemów z danymi obejmowałyby

✓ Dane, których potrzebujesz, nie istnieją. (Czy nigdy nie istniała, czy została odrzucona? Czy te dane można gromadzić i zapisywać do wykorzystania w przyszłości?)

✓ Istnieje, ale nie możesz go mieć. (Czy można przezwyciężyć to ograniczenie?)

✓ Znajdujesz poważne problemy z jakością danych (wiele brakujących lub nieprawidłowych wartości, których nie można poprawić).

Wynikiem tego zadania jest raport jakości danych. Zawiera podsumowanie posiadanych danych, wykrytych drobnych i poważnych problemów z jakością oraz możliwych rozwiązań problemów z jakością lub rozwiązań alternatywnych (takich jak korzystanie z alternatywnego zasobu danych). Jeśli napotykasz naprawdę poważne problemy z jakością danych i nie możesz znaleźć odpowiedniego rozwiązania, być może będziesz musiał zalecić ponowne rozważenie celów lub planów.

Przygotowywanie danych

Eksperci danych spędzają większość czasu na trzeciej fazie procesu eksploracji danych: przygotowaniu danych. Większość danych wykorzystywanych do eksploracji danych została pierwotnie zebrana i zachowana do innych celów i wymaga pewnego doprecyzowania, zanim będzie gotowa do użycia w modelowaniu. Faza przygotowania danych obejmuje pięć zadań. Są to

✓ Wybór danych

✓ Czyszczenie danych

✓ Konstruowanie danych

✓ Integracja danych

✓ Formatowanie danych

Przewodnik krok po kroku CRISP-DM nie wymienia wyraźnie zestawów danych jako produktów dostarczanych dla każdego zadania przygotowania danych, ale te zestawy danych znacznie lepiej istniały i były odpowiednio archiwizowane i dokumentowane. Zbiory danych nie będą odpowiadać jeden do jednego z zadaniami, ale informacje o użytych danych powinny być zawarte w każdym dostarczanym raporcie.

Zadanie: Wybór danych

Teraz zdecydujesz, która część danych, które posiadasz, zostanie faktycznie wykorzystana do eksploracji danych. Rezultatem tego zadania jest uzasadnienie włączenia i wykluczenia. W nim

wyjaśnisz, jakie dane będą, a jakie nie będą wykorzystywane do dalszych prac związanych z eksploracją danych. Wyjaśnisz powody włączenia lub wyłączenia każdej części danych, które posiadasz, w oparciu o znaczenie dla Twoich celów, jakość danych i problemy techniczne – takie jak ograniczenie liczby pól lub wierszy, które mogą obsłużyć Twoje narzędzia, lub przydatność formatów danych do Twoich potrzeb.

Zadanie: Czyszczenie danych

Dane, których używasz, prawdopodobnie nie będą idealnie czyste (bezbłędne). Wprowadzisz zmiany, być może śledząc źródła, aby wprowadzić określone poprawki danych, wyłączając niektóre przypadki lub pojedyncze komórki (elementy danych) lub zastępując niektóre elementy danych wartościami domyślnymi lub zamiennikami wybranymi za pomocą bardziej zaawansowanej techniki modelowania. Możesz zdecydować się na użycie tylko podzbiorów danych do wszystkich lub niektórych prac związanych z eksploracją danych. Produktem dostarczanym do tego zadania jest raport z czyszczenia danych, który szczegółowo dokumentuje każdą decyzję i działanie użyte do oczyszczenia danych. Raport ten powinien obejmować i odnosić się do każdego problemu z jakością danych, który został zidentyfikowany w zadaniu weryfikacji jakości danych w fazie rozumienia danych w procesie. W zgłoszeniu należy również uwzględnić potencjalny wpływ wyborów dokonanych podczas czyszczenia danych na wyniki.

Zadanie: Konstruowanie danych

Może być konieczne wyprowadzenie nowych pól (na przykład użycie daty dostawy i daty złożenia zamówienia przez klienta, aby obliczyć, jak długo klient czekał na otrzymanie zamówienia), zagregowanie danych lub utworzenie w inny sposób nowej formy danych. Materiały dostarczane do tego zadania obejmują dwa raporty:

✓ Atrybuty pochodne: raport opisujący, jakie nowe pola (kolumny) utworzyłeś, jak to zrobiłeś i dlaczego.

✓ Wygenerowane rekordy: Raport opisujący, jakie nowe przypadki (wiersze) utworzyłeś, jak to zrobiłeś i dlaczego.

Chociaż zadania scalania danych i formatowania danych są wymienione jako ostatnie w tej fazie procesu, nie zawsze są ostatnie i mogą nie pojawić się tylko raz. Być może trzeba będzie wykonać scalanie lub ponowne formatowanie na początku fazy przygotowania danych.

Zadanie: Integracja danych

Twoje dane mogą teraz znajdować się w kilku różnych zestawach danych. Aby przygotować się do fazy modelowania, musisz połączyć niektóre lub wszystkie te odmienne zestawy danych. Rezultatem tego zadania są scalone dane. (I nie zaszkodzi udokumentowanie, jak przeprowadzono scalenie).

Zadanie: Formatowanie danych

Dane często przychodzą do Ciebie w formatach innych niż te, które są najwygodniejsze do modelowania. (Zmiany formatu są zwykle zależne od projektu Twoich narzędzi). Konwertuj więc teraz te formaty. Elementem dostarczanym do tego zadania są Twoje przeformatowane dane. (I warto zamieścić mały raport opisujący wprowadzone zmiany.) Fazę przygotowania danych w procesie eksploracji danych należy zakończyć zestawem danych gotowym do modelowania i dokładnym raportem opisującym zestaw danych.

Modelowanie

Jest to część procesu, którą większość eksploratorów danych lubi najbardziej. Twoje dane są już w dobrym stanie, a teraz możesz wyszukiwać przydatne wzorce w swoich danych. Faza modelowania obejmuje cztery zadania. To są

✓ Dobór technik modelowania

✓ Test(y) projektowania

✓ Model(e) budynku

✓ Ocena modelu(ów)

Zadanie: Wybór technik modelowania

Wspaniały świat eksploracji danych oferuje mnóstwo technik modelowania, ale nie wszystkie z nich będą odpowiadać Twoim potrzebom. Zawęź listę w oparciu o rodzaje zaangażowanych zmiennych, wybór technik dostępnych w Twoich narzędziach oraz wszelkie kwestie biznesowe, które są dla Ciebie ważne. (Na przykład wiele organizacji preferuje metody, których wyniki są łatwe do zinterpretowania, więc drzewa decyzyjne lub regresja logistyczna mogą być akceptowalne, ale sieci neuronowe prawdopodobnie nie będą akceptowane). Materiały dostarczane do tego zadania obejmują dwa raporty:

✓ Technika modelowania: Określ techniki, których będziesz używać.

✓ Założenia modelowania: Wiele technik modelowania opiera się na pewnych założeniach. Na przykład typ modelu może być przeznaczony do użytku z danymi, które mają określony typ dystrybucji. Udokumentuj te założenia w niniejszym raporcie.

Statystycy są dobrze poinformowani, surowi i wybredni w kwestii założeń. Niekoniecznie dotyczy to eksploratorów danych i nie jest to wymagane, aby zostać eksploratorem danych. Jeśli masz głęboką wiedzę statystyczną i rozumiesz założenia stojące za wybranymi modelami, możesz być surowy i wybredny w kwestii założeń. Jednak wielu eksploratorów danych, zwłaszcza początkujących, nie przejmuje się zbytnio założeniami. Alternatywą jest testowanie - wiele, wiele testów - waszych modeli.

Zadanie: Projektowanie testów

Test w tym zadaniu to test, którego użyjesz do określenia, jak dobrze działa Twój model. Może to być tak proste, jak podzielenie danych na grupę przypadków do uczenia modelu i inną grupę do testowania modelu. Dane uczące są używane do dopasowania formularzy matematycznych do modelu danych, a dane testowe są wykorzystywane podczas procesu uczenia modelu, aby uniknąć nadmiernego dopasowania: tworząc model, który jest idealny dla jednego zestawu danych, ale dla żadnego innego. Możesz również użyć danych wstrzymania, które nie są używane podczas procesu uczenia modelu, do dodatkowego testu. Produktem dostarczanym do tego zadania jest projekt testu. To nie musi być skomplikowane, ale powinieneś przynajmniej zadbać o to, aby twoje dane treningowe i testowe były podobne i aby nie wprowadzać żadnych błędów do danych.

Zadanie: Budowanie modeli

Modelowanie jest tym, co wiele osób wyobraża sobie jako całe zadanie eksploratora danych, ale to tylko jedno zadanie z dziesiątek! Niemniej jednak modelowanie pod kątem określonych celów biznesowych jest sercem zawodu eksplorującego dane. Materiały dostarczane do tego zadania obejmują trzy elementy:

✓ Ustawienia parametrów: Podczas budowania modeli większość narzędzi daje możliwość dostosowania różnych ustawień, a te ustawienia mają wpływ na strukturę ostatecznego modelu. Zapisz te ustawienia w raporcie.

✓ Opisy modeli: opisz swoje modele. Podaj typ modelu (np. regresja liniowa lub sieć neuronowa) oraz użyte zmienne. Wyjaśnij, jak interpretowany jest model. Dokumentuj wszelkie trudności napotkane w procesie modelowania.

✓ Modele: Ten element dostawy to same modele. Niektóre typy modeli można łatwo zdefiniować za pomocą prostego równania; inne są zbyt złożone i muszą być przesyłane w bardziej wyrafinowanym formacie.

Zadanie: Ocena modeli

Teraz przejrzysz modele, które utworzyłeś, z technicznego, a także biznesowego punktu widzenia (często z wkładem ekspertów biznesowych z Twojego zespołu projektowego). Materiały dostarczane do tego zadania obejmują dwa raporty:

✓ Ocena modelu: podsumowuje informacje opracowane w ramach przeglądu modelu. Jeśli stworzyłeś kilka modeli, możesz je uszeregować na podstawie oceny ich wartości dla konkretnego zastosowania.

✓ Zmienione ustawienia parametrów: Możesz dostroić ustawienia, które zostały użyte do zbudowania modelu i przeprowadzić kolejną rundę modelowania i spróbować poprawić swoje wyniki.

Eksploracja danych, taka jak cebula, tort Dobos lub skała osadowa, ma wiele warstw. Gdy dopiero zaczynasz eksplorację danych, możesz zacząć od pozostawienia ustawień parametrów z ich wartościami domyślnymi (w rzeczywistości możesz nawet nie zauważyć opcji, chyba że spróbujesz ich poszukać). Gdy poczujesz się komfortowo w swojej nowej karierze w eksploracji danych, warto poznać parametry modeli i wiedzieć, jak z nich korzystać. Twoje opcje będą się znacznie różnić w zależności od typu modelu i konkretnego narzędzia, którego .

Ocena

Eksplorowałeś dane i znalazłeś wzorce, a teraz musisz zapytać: czy wyniki są dobre? Ocenisz nie tylko modele, które tworzysz, ale także proces, w którym je stworzyłeś, oraz ich potencjał do praktycznego wykorzystania. Faza zrozumienia danych obejmuje trzy zadania. Są to

✓ Ocena wyników

✓ Przegląd procesu

✓ Ustalenie kolejnych kroków

Zadanie: Ocena wyników

Na tym etapie ocenisz wartość swoich modeli dla realizacji celów biznesowych, które rozpoczęły proces eksploracji danych. Poszukasz powodów, dla których model nie nadawałby się do użytku biznesowego. Jeśli to możliwe, przetestujesz model w praktycznym zastosowaniu, aby sprawdzić, czy działa tak dobrze w miejscu pracy, jak w twoich testach. Materiały dostarczane do tego zadania obejmują dwa elementy:

✓ Ocena wyników (dla celów biznesowych): Podsumuj wyniki w odniesieniu do kryteriów sukcesu biznesowego, które ustaliłeś w fazie rozumienia biznesu. Wyraźnie określ, czy osiągnąłeś cele biznesowe zdefiniowane na początku projektu.

✓ Zatwierdzone modele: obejmują wszystkie modele, które spełniają kryteria sukcesu biznesowego.

Zadanie: Przegląd procesu

Teraz, gdy już zapoznałeś się z danymi i opracowałeś modele, poświęć trochę czasu na przejrzenie swojego procesu. Jest to okazja do zauważenia problemów, które mogłeś przeoczyć i które mogą zwrócić twoją uwagę na błędy w pracy, którą wykonałeś, gdy wciąż masz czas na rozwiązanie problemu przed wdrożeniem. Zastanów się również, w jaki sposób możesz ulepszyć swój proces w przyszłych projektach. Rezultatem tego zadania jest przegląd raportu z procesu. Powinieneś w nim opisać proces przeglądu i ustalenia oraz podkreślić wszelkie obawy, które wymagają natychmiastowej uwagi, np. kroki, które zostały przeoczone lub które należy ponownie odwiedzić.

Zadanie: Ustalenie kolejnych kroków

Faza oceny kończy się Twoimi zaleceniami dotyczącymi następnego ruchu. Model może być gotowy do wdrożenia lub możesz uznać, że lepiej byłoby powtórzyć niektóre kroki i spróbować go ulepszyć. Twoje odkrycia mogą zainspirować nowe projekty eksploracji danych. Materiały dostarczane do tego zadania obejmują dwa elementy:

✓ Lista możliwych działań: Opisz każde alternatywne działanie wraz z najsilniejszymi argumentami za i przeciw.

✓ Decyzja: Podaj ostateczną decyzję w sprawie każdego możliwego działania, wraz z uzasadnieniem decyzji.

Wdrożenie

Wdrożenie jest tam, gdzie eksploracja danych się opłaca. Nie ma znaczenia, jak genialne mogą być Twoje odkrycia lub jak doskonale Twoje modele pasują do danych, jeśli tak naprawdę nie używasz tych rzeczy do ulepszania sposobu, w jaki prowadzisz działalność. Faza rozmieszczania obejmuje cztery zadania. Są to

✓ Planowanie wdrożenia (Twoje metody integracji odkryć opartych na eksploracji danych)

✓ Planowanie monitorowania i konserwacji

✓ Raportowanie wyników końcowych

✓ Przeglądanie wyników końcowych

Zadanie: Planowanie wdrożenia

Kiedy Twój model będzie gotowy do użycia, będziesz potrzebować strategii, jak wykorzystać go w swojej firmie. Produktem dostarczanym do tego zadania jest plan wdrożenia. To jest podsumowanie Twojej strategii wdrażania, wymaganych kroków oraz instrukcji wykonywania tych kroków.

Zadanie: Planowanie monitorowania i konserwacji

Praca w zakresie eksploracji danych to cykl, więc oczekuj aktywnego zaangażowania w swoje modele, ponieważ są one zintegrowane z codziennym użytkowaniem. Produktem dostarczanym do tego zadania jest plan monitorowania i konserwacji. To jest podsumowanie Twojej strategii ciągłego przeglądu wydajności modelu. Musisz upewnić się, że jest on stale używany prawidłowo i że każdy spadek wydajności modelu zostanie wykryty.

Zadanie: Raportowanie wyników końcowych

Materiały dostarczane do tego zadania obejmują dwa elementy:

✓ Raport końcowy: Raport końcowy podsumowuje cały projekt poprzez zestawienie wszystkich raportów utworzonych do tego momentu oraz dodanie przeglądu podsumowującego cały projekt i jego wyniki.

✓ Prezentacja końcowa: Podsumowanie raportu końcowego prezentowane jest na spotkaniu z kierownictwem. Jest to również okazja do odpowiedzi na wszelkie pytania otwarte.

Zadanie: Przejrzyj projekt

Na koniec zespół ds. eksploracji danych spotyka się, aby omówić, co zadziałało, a co nie, co byłoby dobrze zrobić ponownie, a czego należy unikać! Ten krok również ma wynik, chociaż jest przeznaczony tylko dla zespołu zajmującego się eksploracją danych, a nie menedżera (lub klienta). To raport z dokumentacji doświadczenia. W tym miejscu powinieneś opisać wszelkie metody pracy, które działały szczególnie dobrze, aby zostały udokumentowane do ponownego użycia w przyszłości, a także wszelkie ulepszenia, które mogą zostać wprowadzone w twoim procesie. Jest to również miejsce do dokumentowania problemów i złych doświadczeń, wraz z zaleceniami dotyczącymi unikania podobnych problemów w przyszłości. Eksploracja danych to działanie zespołowe. Jeśli więc wydaje się, że ten proces obejmuje wiele kroków, zdaj sobie sprawę, że wykonanie każdego z nich może nie być twoją osobistą odpowiedzialnością i że zawsze właściwe jest proszenie o pomoc innych, kiedy jej potrzebujesz. (Na początku projektu stworzyłeś listę osób, które są zasobami projektu eksploracji danych. To twój mały katalog pomocników!)

Planowanie sukcesu eksploracji danych

Mam dobre i złe wieści na temat eksploracji danych. Zła wiadomość jest taka, że programy do eksploracji danych mogą i zawodzą. W rzeczywistości organizacje dość często zgłaszają, że nie osiągnęły pozytywnych zwrotów z inwestycji w eksplorację danych. Dobrą wiadomością jest to, że tego rodzaju niepowodzeniu można zapobiec. Wykonanie kilku prostych, zdroworozsądkowych kroków na początku pozwoli Ci osiągnąć sukces w eksploracji danych i udokumentowaną wartość. Eksperci danych, którzy konsekwentnie i rozważnie planują sukces, rutynowo osiągają wyniki, które prowadzą do pozytywnego zwrotu z inwestycji i zadowolenia kadry kierowniczej. Przygotowanie uzasadnienia biznesowego maksymalizuje Twoje szanse na uzyskanie informacji, które decydenci mogą i wykorzystają, aby uzyskać dobre wyniki. Planowanie pozwala zmniejszyć ryzyko dla firmy, jednocześnie budując własną wiarygodność. Ten rozdział pokazuje, jak zbudować uzasadnienie biznesowe i ustalić plan powodzenia w eksploracji danych.

Ustalanie kursu za pomocą formalnych przypadków biznesowych

Eksploracja danych wiąże się z kosztami - kosztami oprogramowania, kosztami pracy, kosztami serwerów i być może również kosztami pozyskania danych. Wydatki mogą sięgać dziesiątek tysięcy dolarów, a nawet setek tysięcy dolarów. Aby uzasadnić płacenie za to wszystko, może być konieczne przygotowanie uzasadnienia biznesowego. Przypadek biznesowy przedstawia konkretny problem biznesowy, proponowany plan jego rozwiązania oraz związane z nim korzyści i koszty. Uzasadnienie biznesowe może pomóc decydentom firmy lub organizacji zrozumieć sytuację i podjąć odpowiednią decyzję. Każdy rodzaj organizacji może używać formalnych przypadków biznesowych wspierających podejmowanie decyzji. Notowane na giełdzie lub duże przedsiębiorstwa i agencje rządowe zazwyczaj wymagają ich przy wszystkich znaczących wydatkach. Przypadki biznesowe również mogą Ci pomóc. Przygotowanie dobrego uzasadnienia biznesowego wyjaśnia Twoje własne myślenie o prawdopodobnych celach i drodze do ich osiągnięcia. Dokumentuje uzasadnienie Twojego planu. Możesz też wykorzystać swoje uzasadnienie biznesowe do ustalenia realistycznych oczekiwań wobec kierownictwa i współpracowników. Tak więc, nawet jeśli uzasadnienie biznesowe nie jest wymogiem formalnym, nadal dobrym pomysłem jest przygotowanie go dla programu lub projektu do eksploracji danych.

Zadowolenie szefa

Twoje uprawnienia do ustalania priorytetów i przydzielania zasobów są ograniczone. Prawie wszyscy eksploratorzy danych odpowiadają przełożonemu. Nawet osoby pracujące na własny rachunek mają klientów do zaspokojenia. Jeśli chcesz uzyskać dostęp do znacznego lub drogiego zasobu, potrzebujesz zgody swojego szefa, aby go uzyskać. Jak uzyskać ten dostęp? Sprawa biznesowa. Twoje uzasadnienie biznesowe to sposób na uzyskanie zgody (pozwolenia i zasobów) na eksplorację danych. Ale po co tyle zamieszania? Twój szef cię zna i ci ufa. Dlaczego po prostu nie powiedzieć swojemu szefowi, którą opcję polecasz i zostawić to na tym? Chodzi o zmniejszenie ryzyka. Możesz myśleć, że dobrym rozwiązaniem jest pewne podejście, a twój szef może ci uwierzyć. Ale twój szef też ma szefa, a szef twojego szefa nie tak dobrze cię zna lub tak samo ci ufa. Żaden z tych szefów nie chce ryzykować złej decyzji. Jeśli pojawią się pytania o to, dlaczego zasoby zostały przeznaczone na konkretny projekt (a nie na inny) lub jeśli coś pójdzie nie tak, nie będą mogli bronić swoich decyzji bez udokumentowanej sprawy. Więc pomóż im! Twój szef potrzebuje Twojego uzasadnienia biznesowego, aby

✓ Dokładnie zrozumieć swoją rekomendację i jej uzasadnienie

✓ Rozwiązać problemy menedżerów wyższego szczebla

✓ Mieć odpowiednią dokumentację, aby rozwiązać problemy, które mogą pojawić się później

Minimalizowanie własnego ryzyka

Uzasadnienie biznesowe nie tylko zmniejsza ryzyko dla twojego szefa. Zmniejsza również twoje ryzyko. Proces przygotowywania uzasadnienia biznesowego wymaga poświęcenia przemyśleń i badań nad problemem biznesowym, z którym chcesz się zmierzyć. Uzasadnienie biznesowe przygotujesz przed rozpoczęciem eksploracji danych – a właściwie przed zaangażowaniem jakichkolwiek zasobów w projekt. Gdy to zrobisz, będziesz miał okazję dostrzec problemy, których mogłeś nie rozważyć w inny sposób, pytania, które może zadać kierownictwo, oraz sposoby ulepszenia własnych pomysłów. Więc kiedy pójdziesz do swojego szefa z prawdziwym uzasadnieniem biznesowym, a nie tylko z rekomendacją, będziesz bardziej pewny siebie i lepiej przygotowany do obrony swojej rekomendacji. Opracowanie uzasadnienia biznesowego zmniejsza ryzyko, ponieważ

✓ Będziesz bardziej przekonujący, wyjaśniając swoją propozycję i jej korzyści.

✓ Niektóre potencjalne problemy zostaną już zidentyfikowane, a propozycja będzie zawierała kroki mające im zapobiec.

✓ Dokumentacja uzasadnienia biznesowego będzie ważna, jeśli później pojawią się pytania. Decyzja zostanie podjęta na podstawie uzasadnienia biznesowego, a nie Twojej osobistej opinii.

Budowanie uzasadnień biznesowych

Jako eksplorator danych potrzebujesz narzędzi do eksploracji danych, czasu, który możesz poświęcić na wartościowy projekt eksploracji danych, a może po prostu możliwości zrobienia czegoś nowego i innego niż zwykła rutyna. Ale nic z tego nie ma znaczenia dla twojego kierownictwa, akcjonariuszy ani klientów. Mają problem, który kosztuje ich pieniądze, ingeruje w ich życie, a może jedno i drugie. Chcą tylko rozwiązania tego problemu. Tu właśnie pojawia się Twój business case. Twój business case będzie poświęcony temu problemowi: co to jest, jak duży jest i jak bardzo boli. Ma to sprawić, że poczują ten ból i zrozumieją, jak tylko twoja propozycja może go złagodzić. W Twoim przypadku biznesowym nie zamierzasz zmuszać nikogo i wszystkich do eksploracji danych. Zamierzasz przekonać konkretną grupę ludzi, że ich ból jest zbyt duży, aby z nim żyć, że twój plan może sprawić, że ból zniknie i że można ci zaufać, że to zrobisz. Musisz przekonać ich, że Twoja propozycja jest cenniejsza niż pieniądze, które muszą wydać

Elementy uzasadnienia biznesowego

Elementami uzasadnienia biznesowego eksploracji danych są . . . elementy uzasadnienia biznesowego. Nic nie jest wyjątkowe w uzasadnieniu biznesowym związanym z eksploracją danych poza faktem, że proponowane rozwiązanie będzie wymagało eksploracji danych. Jeśli twoje kierownictwo nie jest zaznajomione z eksploracją danych, przygotuj się na podanie wystarczających szczegółów, aby przekonywać, że twoja propozycja zadziała. Ale i tak powinieneś to robić! Ponieważ uzasadnienie biznesowe jest dokumentem opartym na dowodach, będziesz musiał zebrać pewne dowody. Zanim usiądziesz do pisania uzasadnienia biznesowego, zbierz materiały, które będą Ci potrzebne. Musisz wyjaśnić problem i jego wpływ na organizację. Poszukaj raportów, zapisów skarg i innej dokumentacji opisującej naturę i skalę problemu. Być może zapoznałeś się już z niektórymi opcjami rozwiązania problemu. Czy przeprowadziłeś wywiady, pozyskałeś literaturę od dostawców lub znalazłeś raporty branżowe, które wykorzystasz do wydania rekomendacji? Zbierz wszystkie te materiały razem, gdzie można je łatwo znaleźć w celach informacyjnych. Nie martw się o to, czego jeszcze brakuje. Zawsze możesz uzyskać dodatkowe informacje podczas przechodzenia przez ten proces. Tabela przedstawia elementy uzasadnienia biznesowego ułożone w kolejności, w jakiej powinny się pojawić. To są

informacje, które powinieneś zbierać i pytania, które powinieneś zadawać podczas prowadzenia badań.

Elementy uzasadnienia biznesowego

Temat: Pytania do zadawania

Sytuacja

Kontekst : Jakie organizacje są zaangażowane? Jaka jest jego działalność?

Stwierdzenie problemu: Co się dzieje? Kiedy to się zaczęło? Kogo to dotyczy? Czy przyczyna jest znana? Czy to jest powszechne czy niezwykłe? Rodzaj problemu? Czy my lub inni naprawiliśmy już takie problemy?

Opcje

Alternatywy działania : Jakie rozwiązania zostały zaproponowane w celu rozwiązania problemu? Jakie są zalety i wady każdej alternatywy?

Preferowane działanie : Którą alternatywę uważasz za najlepszą? Przedstaw to swoje stanowisko ponad wszystkie inne.

Powiązanie preferowanego działania z celami strategicznymi: Niektóre organizacje podkreślają takie cele, a inne nie. Jeśli tak,

musisz wyjaśnić, że twoja propozycja jest zgodna z co najmniej jednym z tych celów i w jaki sposób.

Korzyści

Oczekiwane korzyści: Jaką wartość przyniesie proponowane rozwiązanie? Powinno to być wyrażone w dolarach, nawet jeśli świadczenie nie jest pieniężne. Jeśli przewidujesz na przykład oszczędność pracy, uzyskaj oszacowanie kosztów pracy na godzinę i pomnóż przez przewidywane zaoszczędzone godziny, aby przeliczyć korzyści na wartość w dolarach.

Mechanizm: W jaki sposób proponowane działanie przyniesie oczekiwane korzyści?

Metryki: Jak będą mierzone korzyści? (Innymi słowy, w jaki sposób kierownictwo może później sprawdzić wyniki?)

Koszty

Koszty : Ile będzie kosztować proponowane działanie? Podaj szczegółowy budżet z kosztami gotówkowymi i niepieniężnymi oraz harmonogram.

Koszty niepodejmowania działań: Ile będzie kosztował problem, jeśli w ogóle nie podejmiemy żadnych działań?

Co zrobić, jeśli brakuje Ci ważnej informacji? Nie pomijaj tego! Idź i dowiedz się, co musisz wiedzieć. Jeśli nie możesz znaleźć dokładnie tego, czego szukasz, zdobądź najlepszy zamiennik, jakim możesz zarządzać. (Na przykład, jeśli nie wiesz, jak długo coś zajmie, oszacuj rozsądny zakres osobogodzin na wykonanie zadania.)

Zapisuję to

Zbieranie informacji to najbardziej czasochłonna część tworzenia uzasadnienia biznesowego. Po zebraniu informacji pomocniczych i poświęceniu czasu na przemyślenie każdego elementu

uzasadnienia biznesowego, spisanie ich powinno być szybkie i proste w porównaniu. Możesz wpisać te same informacje, uporządkowane w tej samej kolejności, które właśnie przeglądałeś. Nie musisz być fantazyjny. Piszesz uzasadnienie biznesowe, a nie poezję. Masz również pewną swobodę w zmianie struktury uzasadnienia biznesowego podczas jego pisania. Jeśli uważasz, że sprawa zostałaby wyrażona jaśniej z pewnymi zmianami w organizacji, to w porządku. Na przykład możesz chcieć umieścić koszt problemu na początku raportu, wraz z opisem problemu. Lub możesz chcieć podać koszty i korzyści natychmiast po każdej alternatywnej opcji, którą opisujesz. Dopóki wszystkie elementy są uwzględnione, a uzasadnienie biznesowe jest łatwe do odczytania i zrozumienia, masz się dobrze. Każdy przypadek biznesowy, który ma więcej niż kilka stron, musi zawierać streszczenie. To jednostronicowe (już nie) podsumowanie kluczowych punktów, zwłaszcza konkluzji (rekomendacji i korzyści). Streszczenie powinno być pierwszą stroną Twojego uzasadnienia biznesowego. Nie zapominaj o tym, ponieważ może to być jedyna część, którą czyta zajęty dyrektor.

Podstawy świadczeń

W przypadku biznesowym wszystko sprowadza się do pieniędzy. Nawet jeśli korzyści, które oferuje Twoja analiza biznesowa, nie są w oczywisty sposób związane z pieniędzmi – to znaczy, jeśli Twoja analiza biznesowa tylko oszczędza czas, poprawia warunki pracy lub dostarcza lepszych informacji – upewnij się, że określiłeś tę korzyść w kategoriach finansowych. Musisz wyrobić w sobie nawyk określania wartości pieniężnej korzyści, jakie oferuje Twoja analiza biznesowa i we wszystkich dyskusjach z decydentami. Patrząc na to w ten sposób, dostrzegasz tylko dwa rodzaje korzyści:

✓ Zwiększone przychody

✓ Zmniejszone koszty

Kluczową częścią przygotowania uzasadnienia biznesowego jest zatem zidentyfikowanie i oszacowanie wszystkich sposobów, w jakie proponowane rozwiązanie wykona jedną z tych dwóch rzeczy. To mało znana, brudna prawda, że przypadek biznesowy, który równoważy koszty projektu z oszczędnościami, jest zdecydowanie bardziej przekonujący niż ten, który obiecuje wzrost przychodów. Dzieje się tak głównie dlatego, że szacunki dotyczące wzrostu przychodów wymagają więcej domysłów, a często po drodze dzieje się więcej rzeczy. Wielu menedżerów nauczyło się z doświadczenia, że do osiągnięcia wzrostu przychodów potrzebna jest większa współpraca z innymi osobami, a Ty możesz nie uzyskać takiej współpracy, jakiej oczekujesz. Nie oznacza to, że wzrost przychodów nie jest pożądany. W rzeczywistości we wzroście przychodów często istnieje większa potencjalna wartość. Możesz jednak zaoszczędzić tylko tyle, ile wydajesz, co jest czynnikiem ograniczającym. Z drugiej strony dochody nie mają granic. Jednak priorytetem jest zidentyfikowanie potencjalnych oszczędności i zbudowanie na tym uzasadnienia biznesowego, nawet jeśli Twoim osobistym celem jest zwiększenie przychodów. (Jeśli odniesiesz sukces, nikt nie odrzuci premii!)

Unikanie opcji niepowodzenia

Jeśli przejrzysz najnowsze doniesienia biznesowe, będziesz mieć dużą szansę na znalezienie kilku historii sukcesu biznesowego, które obejmują eksplorację danych. Awarie nieczęsto trafiają do wiadomości. Nikt nie zatrudnia publicysty, aby umieścić historię niepowodzenia w mediach. Ale jeśli porozmawiasz prywatnie z ludźmi, którzy wypróbowali eksplorację danych, wielu przyzna, że nie poszło im to tak dobrze. Idealnie byłoby rozpocząć dyskusję na temat dowodów przeciwko eksploracji danych od ładnych, solidnych danych. Gdyby istniały tylko odpowiednie badania, aby z całą pewnością stwierdzić, że pewna liczba firm zainwestowała w eksplorację danych w zeszłym roku i że x procent odnotowało znaczące zyski, y procent osiągnął próg rentowności, a z procent nie osiągnął progu rentowności. Jednak te dane nie istnieją. Przeprowadzono szereg ankiet na ten temat. Niektórzy

malują różowy obraz zwrotu z inwestycji analityków. Inne przyniosły bardzo różne wyniki, wskazując, że tylko niewielka część tych, którzy inwestują w eksplorację danych (lub analitykę predykcyjną lub podobną koncepcję) osiągają próg rentowności, nie mówiąc już o osiąganiu znacznych zysków. Dlaczego tak różne wyniki? Niektóre ankiety dotyczą tylko klientów określonych dostawców, a respondentami są wszyscy klienci wybrani przez tych dostawców. Inne mogą być względnie niezależne od wpływu dostawcy. Jednak badania te niekoniecznie były prowadzone zgodnie z najwyższymi standardami badań ankietowych. Niektóre wyniki ankiety w obiegu nigdy nie podają źródła innego niż „ostatnie badanie branżowe".

Ponieważ nie znajdziesz wiarygodnego źródła twardych danych, oto ocena oparta na rozległym osobistym doświadczeniu z eksploracją danych i branżą eksploracji danych. Sprawę przeciwko eksploracji danych można podsumować jednym zdaniem:

Większość organizacji, które inwestują w eksplorację danych, nigdy nie osiąga rentowności.

Dlaczego to prawda? Co to oznacza dla Ciebie? Czy eksploracja danych jest odpowiednia tylko w niektórych szczególnych sytuacjach? Czy niektórzy ludzie urodzili się eksploratorami danych, a inni nie? Nie, to znacznie prostsze. To takie proste, że aż głupie. Oto najczęstsza przyczyna niepowodzenia eksploracji danych (proszę o bęben):

Brak planowania.

Dość antyklimatyczne, co?

Więc tajemnica się skończyła. Eksperci danych zawodzą, ponieważ nie planują sukcesu. Większość nieudanych programów do eksploracji danych nigdy nie rozpoczęła się z żadnym planem sukcesu. Możesz czuć się rozczarowany tym objawieniem. To tak, jakby powiedzieć, że większość angielskich kierunków nie zostaje najlepiej sprzedającymi się pisarzami, ponieważ nie napisali żadnych książek. Ale to też wspaniała wiadomość! Ponieważ masz teraz ogromną przewagę nad większością początkujących eksploratorów danych. Wiesz to, czego oni nie wiedzą: aby zmaksymalizować swoje szanse na sukces w eksploracji danych, musisz stworzyć plan, który połączy punkty od problemu, przez eksplorację danych, do rozwiązania. Czy to brzmi znajomo? Brzmi jak uzasadnienie biznesowe, prawda? Wcześniej dowiedziałeś się, jak i dlaczego rozwijać swoje uzasadnienie biznesowe. Uzasadnienie biznesowe to nie cały plan. Będziesz także musiał przygotować plan pracy krok po kroku, mapę drogową do realizacji zobowiązań podjętych w swoim uzasadnieniu biznesowym. Ta koncepcja również może brzmieć znajomo. Jeśli nie, zapoznaj się z rozdziałem 5, aby przeczytać o procesie CRISP-DM do eksploracji danych.

Przygotowanie do właściwego oprogramowania

Osoby zajmujące się eksploracją danych często mają wiele pytań dotyczących oprogramowania. Na konferencjach i prezentacjach poświęconych eksploracji danych prawie każdy mówca jest pytany: „Jakich narzędzi używasz?" Wiele mówi się o cenach i negocjacjach z dostawcami. Relacje medialne dotyczące eksploracji danych i innych rodzajów analiz również kładą duży nacisk na oprogramowanie. Przy tak dużym skupieniu uwagi na oprogramowaniu może się wydawać, że Twoja kariera w dataminingu zaczyna się od zdobycia odpowiednich narzędzi, ale to nie jest dobry początek. Uzyskasz lepsze wyniki i bardziej pozytywne wrażenia, stawiając na pierwszym miejscu cele biznesowe i pozwalając potrzebom biznesowym kierować procesem wyboru oprogramowania. I tu właśnie wkracza ten rozdział. W tym rozdziale dowiesz się, jak na pierwszym miejscu postawić własne cele biznesowe i wykorzystać je jako przewodnik przygotowujący do eksploracji danych. Skupiamy się tutaj na procesie: ocenie potrzeb, ustalaniu priorytetów, porównywaniu produktów i współpracy z dostawcami oprogramowania.

Spojrzenie na narzędzia do eksploracji danych z perspektywy

Gdybyś potrzebował miejsca do życia, nie zacząłbyś od biegania do sklepu po nowe elektronarzędzia i nie przeniósłbyś się po prostu w dowolne miejsce, w którym mógłbyś dostać wolny pokój. Zacząłbyś od rozważenia swoich potrzeb. Pomyślałbyś o tym, kto będzie mieszkał w twoim domu i stylu życia, którego pragniesz. Możesz przejrzeć mocne i słabe strony obecnego domu, aby zaspokoić te potrzeby, i sformułować kilka pomysłów na to, co jest naprawdę wymagane do zaspokojenia twoich potrzeb. Mając zdefiniowane wymagania, ułożyłbyś plan. Budżet i inne czynniki mogą sprawić, że Twój plan będzie trochę inny od tego, co sobie wyobrażałeś na początku. Być może zacząłbyś od skromnego mieszkania, które spełnia Twoje najpilniejsze potrzeby, pozostawiając miejsce na przyszłe dodatki i ulepszenia. Nie ma sensu wybierać narzędzi, materiałów lub usług, dopóki nie masz planu, ponieważ plan pomaga ci zrozumieć, czego będziesz potrzebować i kiedy zbudować swój dom. Twoje projekty eksploracji danych również zaczynają się od planów. Dopóki nie określisz, co zamierzasz osiągnąć i własnych wymagań dotyczących pracy, nie skupiaj się na oprogramowaniu.

Unikanie zagrożeń związanych z oprogramowaniem

Zastanów się, co mogłoby się stać, gdybyś wybrał dom bez dokładnego zastanowienia się nad wyborem. Kiedy się wprowadziłeś, może się okazać, że układ nie pasował do Twojego stylu życia. Możesz znaleźć równie ładne domy w tej samej okolicy na sprzedaż po niższych cenach. Albo może się okazać, że utrzymanie domu jest trudniejsze, niż się spodziewałeś. Pośpieszny wybór oprogramowania wiąże się z podobnym ryzykiem. Dobre przygotowanie chroni przed tymi typowymi zagrożeniami:

✓ Nieodpowiednie możliwości oprogramowania: nowi eksploratorzy danych często stwierdzają, że wybrane przez nich oprogramowanie nie ma pełnego zestawu funkcji, których potrzebują. Chociaż kuszące jest zrzucanie winy na dostawcę oprogramowania, ta wymówka prawdopodobnie nie zadowoli twojego szefa lub klienta. Pierwszym i najważniejszym sposobem uniknięcia wyboru produktów, które nie spełniają Twoich wymagań, jest opóźnienie wyboru oprogramowania do czasu zakończenia dokładnej oceny w celu ustalenia, jakie będą te wymagania. Gdy już dobrze zrozumiesz, czego potrzebujesz, warto zaplanować odpowiednie testy przed podjęciem decyzji o zakupie. Zawsze testuj oprogramowanie, zanim zdecydujesz się na zakup produktu. Oceń różnorodne produkty. Darmowe produkty są oczywiście łatwo dostępne, a większość komercyjnych dostawców oprogramowania umożliwia testowanie ich produktów przez krótki czas bez żadnych opłat. Wykorzystaj w pełni te próby, przygotowując plan testów i określając kryteria sukcesu przed rozpoczęciem. Jeśli test oprogramowania wymaga znacznej pomocy technicznej ze strony dostawcy, na przykład

przyprowadzenia osoby z pomocy technicznej do Twojej witryny na kilka dni, prawdopodobnie zostaniesz poproszony o uiszczenie opłaty za tę usługę. (Miej to na uwadze, rozważając wolne oprogramowanie. Firmy, które rozpowszechniają swoje oprogramowanie za darmo, często zarabiają na sprzedaży usług.) Sprzedawcy czasami stosują te opłaty przy zakupie oprogramowania. Jeśli nie jest to oferowane, zapytaj. Czasami problem nie polega na tym, że oprogramowanie nie ma niezbędnych możliwości, ale na tym, że eksplorator danych nie nauczył się wystarczająco, jak z niego korzystać. Zaplanuj szkolenie dla dowolnego zakupionego produktu i rozpocznij szkolenie, gdy tylko wprowadzisz narzędzie do użytku.

✓ Przepłacanie: Ludzie często nawet nie zdają sobie sprawy, że kupili droższy produkt, niż potrzebowali. Chociaż najbardziej opłacalnym wyborem dla Ciebie niekoniecznie jest ten z najniższą ceną, nie ma sensu płacić za funkcje produktu, z których nie będziesz korzystać. Najlepszą obroną przed przepłacaniem jest zrobienie zakupów porównawczych – z listą wymagań pod ręką. Jeśli przedstawiciel handlowy zachęca Cię do wybrania droższego produktu niż myślisz, że potrzebujesz, zadawaj pytania. Zapytaj, jakie dodatkowe możliwości są dostępne w droższym produkcie i jakie są dla Ciebie istotne. Uzyskaj wystarczającą ilość informacji, aby określić, czy aktualizacja będzie rzeczywiście bardziej wartościowa dla Twojej organizacji, czy po prostu zwiększy prowizję przedstawiciela handlowego.

✓ Niepotrzebna złożoność: Kiedy eksploratorzy danych siadają do realizacji projektu z nowym narzędziem, czasami okazuje się, że korzystanie z niego jest trudniejsze niż się spodziewali. Czasami jasne jest, że narzędzie może zrobić to, co jest potrzebne, ale nie jest to łatwe. W innych przypadkach może się wydawać, że produkt po prostu nie spełnia wymagań. Jednym ze sposobów uniknięcia niepotrzebnej złożoności jest wypróbowanie produktów przed dokonaniem wyboru. Zawsze przetestuj oprogramowanie, zanim zdecydujesz się używać go w prawdziwej aplikacji. Dobre narzędzia do eksploracji danych mają pomóc użytkownikom biznesowym w szybkim odkrywaniu przydatnych wzorców w danych. Przynajmniej taki jest cel. Prawda jest jednak taka, że zrozumienie nowych narzędzi i kompletnych projektów nadal wymaga wysiłku.

Koncentracja na celach biznesowych, a nie na narzędziach

Ważne jest, aby najpierw skupić się na celach biznesowych i przez cały proces eksploracji danych. Twój szef nigdy nie odpowie na przekonującą prezentację, pytając: „Jakich narzędzi używałeś?" Nigdy. Twój szef jest zainteresowany zwiększeniem sprzedaży produktów, zmniejszeniem roszczeń gwarancyjnych, uzyskaniem większej liczby problemów z pomocą techniczną rozwiązywanych przy pierwszym kontakcie lub zapobieganiem problemom, aby pomoc techniczna nie była potrzebna w pierwszej kolejności – takie rzeczy. Więc mów o takich rzeczach. I pomyśl o takich rzeczach. Dobrzy eksploratorzy danych często uzyskują doskonale użyteczne wyniki przy użyciu narzędzi, które nie są najlepsze. Możesz mieć najbardziej wyszukane oprogramowanie na rynku, ale nie przynosić dobrych rezultatów. Twoja uwaga na cele biznesowe i myśl, którą wkładasz w proces eksploracji danych, są o wiele ważniejsze niż zabawki, eee, narzędzia, które masz pod ręką. Zbyt wczesne skupienie się na oprogramowaniu może prowadzić do wyboru niewłaściwego oprogramowania, ale to najmniejszy z twoich problemów. Prawdziwym problemem z mentalnością „najpierw narzędzia" jest to, że nie jest to mentalność „najpierw cele biznesowe". Zajmowanie się celami biznesowymi jest tym, na czym polega eksploracja danych.

Dlatego miej oko na nagrodę i pozwól, aby twoje wybory dotyczące oprogramowania oraz każdy wybór, którego dokonujesz podczas eksploracji danych, kierował się celami biznesowymi, które wyznaczyłeś od samego początku, w oparciu o wkład kierownictwa i własną wiedzę biznesową. Pamiętaj o tych zasadach:

✓ Skoncentruj się na kwestiach, które są ważne dla firmy i Twojego kierownictwa. Narzędzia, techniki i ciekawe wzorce w danych, które nie są związane z celem biznesowym, mogą Cię fascynować, ale nic nie znaczą dla menedżera pod presją rozwiązania konkretnego problemu biznesowego.

✓ Przestrzegaj terminów. Najpierw rób pierwsze rzeczy. Pominięcie kamienia milowego oznacza, że uniemożliwiasz innym wykonanie swojej pracy na czas.

✓ Wykorzystaj w pełni swoje zasoby. Wybór oprogramowania to tylko jeden obszar, w którym przezorność i planowanie zapobiega marnotrawstwu. Twój czas i czas innych członków zespołu są jeszcze cenniejsze.

✓ Krok po kroku buduj wiarygodność. Za każdym razem, gdy pomagasz kierownictwu osiągnąć ważny cel biznesowy, budujesz zaufanie. Menedżerowie szanują analityków, którzy mówią w ich języku, więc rozmawiaj o dolarach i centach, a nie o wymyślnej matematyce. Prezentacje powinny być krótkie i na temat. Prezentuj tylko te informacje, które są bezpośrednio związane z celem biznesowym.

Określanie, czego potrzebujesz

Nie możesz dokonać racjonalnego wyboru oprogramowania, dopóki nie przemyślisz potrzeb biznesowych. Dokładny przegląd Twoich celów, środowiska pracy oraz potrzeb i preferencji zespołu stanowi podstawę do dokonania mądrego wyboru. Ustalenie, czego potrzebujesz Nie możesz dokonać racjonalnego wyboru oprogramowania, dopóki nie przemyślisz potrzeb biznesowych. Dokładny przegląd Twoich celów, środowiska pracy oraz potrzeb i preferencji zespołu stanowi podstawę do dokonania mądrego wyboru. Dane same się nie wydobywają, a świetne modele nie będą wyglądać tak wspaniale, jeśli nikt ich nie użyje. Uzyskiwanie informacji od wszystkich zaangażowanych w proces pomaga zrozumieć, jakich funkcji oprogramowania będziesz potrzebować i dlaczego, ale co ważniejsze, pomaga budować mosty i dowiedzieć się, jak współpracować z innymi, aby uruchomić eksplorację danych. Chociaż możesz mieć tylko jedną lub tylko kilka osób bezpośrednio korzystających z oprogramowania do eksploracji danych, wiele innych przyczynia się do tego procesu. Na przykład będziesz potrzebować danych, więc porozmawiaj z personelem informatycznym o dostępie do danych. Modele są bezużyteczne, chyba że można je zintegrować z operacjami biznesowymi. Dowiedz się, jak to zrobić i jakie są techniczne (a także prawne i polityczne) wymagania dotyczące tego procesu. Rozmowa z członkami personelu o procesach, których używają, może być świetnym sposobem na nawiązanie kontaktów. Oto kilka wskazówek, o których warto pamiętać:

✓ Poszukuj informacji o charakterze inkluzywnym. Dotrzyj do osób pełniących wszystkie funkcje, które mogą wymagać wsparcia Twojego procesu. Szukaj różnorodności w doświadczeniu zawodowym, wykształceniu i punktach widzenia. To nie tylko uprzejmość. Uzyskiwanie informacji od zróżnicowanej grupy ludzi pomaga uniknąć martwych punktów, które mogą powodować nieprzewidziane problemy i stanąć między Tobą a Twoimi celami.

✓ Okazuj szacunek. Ważne jest, aby okazywać szacunek każdej osobie, z którą przeprowadzasz wywiad i uważnie słuchać. Jednak słuchanie nie wystarczy. Większość ludzi nie jest ekspertami w komunikacji. Mogą nie wiedzieć dokładnie, jak wyjaśnić swoje obawy w sposób łatwy do zrozumienia. Więc od Ciebie zależy, czy zadasz dobre pytania. Jeśli nie rozumiesz jasno odpowiedzi, powiedz to (uprzejmie) i poproś o wyjaśnienie. Wyjaśnij, czego potrzebujesz od każdej osoby i zapytaj, czego oczekuje się w zamian. Jeśli są to dane, których potrzebujesz, opisz, czego potrzebujesz i zapytaj, czego potrzeba, aby je zdobyć. Rozpoczęcie od ogólnych pytań może pomóc w odkryciu problemów, o których sam byś nie pomyślał (może myślisz o specyfikacjach technicznych, ale nie zdajesz sobie sprawy, że będziesz

potrzebować autoryzacji kierownictwa). W miarę postępów pracuj stopniowo nad bardziej szczegółowymi pytaniami.

✓ Udostępnij swoje dane. Po zakończeniu przeglądu celów i wywiadów z pracownikami podsumuj i podziel się swoimi spostrzeżeniami. Rozpowszechnianie podsumowania daje każdemu szansę na poprawienie wszystkiego, co błędnie zinterpretowałeś, i wskazanie problemów (lub osób), które mogłeś przeoczyć. Poinformuj zespół, że z zadowoleniem przyjmujesz te poprawki. Lepiej dowiedzieć się o tym teraz, niż później, gdy projekt jest w toku i natkniesz się na problem.

Porównywanie narzędzi

Teraz nadchodzi część zabawna w zależności od osobistego gustu. Czas sporządzić listę wymagań dotyczących oprogramowania w oparciu o cele biznesowe, warunki pracy i preferencje zespołu. Opracuj listę kontrolną oprogramowania. Przejrzyj swoje ustalenia i określ, w jaki sposób oprogramowanie może pomóc Ci osiągnąć Twoje cele. Pamiętaj o następujących kwestiach:

✓ Funkcje: Do czego potrzebujesz swojego oprogramowania? Poniższa lista przedstawia niektóre z bardziej typowych funkcji oprogramowania do eksploracji danych:

• Dostęp do danych i import: nie osiągniesz punktu wyjścia, jeśli Twoje narzędzie nie może importować danych z bazy danych lub plików, w których są przechowywane.

• Przygotowanie danych: potrzeby w zakresie manipulacji danymi nie zawsze są oczywiste od samego początku, należy więc szukać narzędzi oferujących szerokie możliwości manipulacji danymi.

• Techniki eksploracyjne: musisz tworzyć tabele i różnorodne wykresy.

• Modelowanie: nie wiesz jeszcze dokładnie, który model będzie dla Ciebie najlepszy, ale wiesz, w oparciu o wymagania biznesowe, jakiego rodzaju zmienne (pola) będą danymi wejściowymi (predyktory, zmienne niezależne) i wyjściowymi (cele, zmienne zależne). Twoje narzędzie powinno oferować nie tylko jeden, ale kilka odpowiednich typów modeli.

• Raportowanie: zaoszczędzisz czas, jeśli wyniki z narzędzia można łatwo włączyć do pisemnych raportów i prezentacji.

• Eksport danych: Eksploratorzy danych tworzą nowe dane. Będziesz potrzebować dobrego procesu eksportowania tego z narzędzia do wykorzystania w innym miejscu w firmie.

• Wdrożenie: dobry model nic nie znaczy, jeśli nie można go wdrożyć w codziennej działalności. Niektóre narzędzia oferują więcej lub łatwiejsze opcje dla wdrożenia niż inne.

• Śledzenie tego, co zrobiłeś: Porównaj możliwości tworzenia ścieżek audytu, organizowania pracy i współpracy.

✓ Interfejs: Najbardziej odpowiedni interfejs użytkownika do eksploracji danych to taki, który wykorzystuje programowanie wizualne, w którym główne etapy (takie jak importowanie danych lub konstruowanie modelu) pojawiają się jako ikony, które zastępują wiele wierszy napisanego kodu. Pomocne może być posiadanie produktu, który oferuje opcję użycia kodu do określonych zadań.

✓ Usługi: Upewnij się, że Twoje potrzeby w zakresie szkoleń, wsparcia technicznego i obsługi klienta mogą zostać zaspokojone. Możesz również potrzebować pomocy doradczej, aby dowiedzieć się, jak za pomocą tego narzędzia sprostać konkretnym potrzebom biznesowym.

✓ Zasoby informacyjne: Poszukaj zasobów, które pomogą Ci zrozumieć i używać swojego narzędzia, w tym dokumentacji, grup użytkowników i książek.

Zakupy oprogramowania

Pamiętaj o tych wskazówkach, gdy kupujesz oprogramowanie:

✓ Nadaj priorytety wymaganiom. Jest prawdopodobne, że nie będziesz w stanie od razu uzyskać wszystkich pożądanych umiejętności. Przygotuj się na dokonywanie dobrych wyborów, zastanawiając się nad tym, co jest najważniejsze, a co może się poślizgnąć.

✓ Przetestuj narzędzia, które rozważasz. Instalacja oprogramowania do eksploracji danych nie powinna być skomplikowana, a zaimportowanie niewielkiej ilości danych i eksperymentowanie z kilkoma funkcjami powinno być dość łatwe.

✓ Nalegaj na uzyskanie satysfakcjonujących odpowiedzi na pytania techniczne. Sprzedawcy powinni dysponować kompetentnym technicznie personelem, aby dokładnie odpowiadać na pytania.

✓ Nie zakładaj, że przedstawiciele handlowi faktycznie korzystali ze sprzedawanych przez siebie produktów. W większości przypadków przedstawiciele handlowi mają niewielkie lub żadne praktyczne doświadczenie z narzędziami do eksploracji danych.

✓ Nie polegaj na żadnym dostawcy w kwestii informacji o ofertach konkurencyjnego dostawcy. Sprzedawcy na ogół nie mają szczegółowych, aktualnej wiedzy o konkurentach. Twierdzenia dotyczące konkurencyjnych produktów są często nieaktualne, mylące lub po prostu błędne. To nie jest celowe; po prostu trudno nadążyć.

✓ Nie używaj dostawców jako jedynego źródła informacji, nawet jeśli chodzi o informacje o własnych produktach dostawcy. Zadaniem sprzedawcy jest zrobienie z ciebie klienta. Odpowiedzialność za uzyskanie dobrych i bezstronnych informacji spoczywa na Tobie.

✓ Nie żałuj szkolenia. Nie uzyskasz dobrej wartości z narzędzi, których nie umiesz dobrze używać.

✓ Nie ignoruj ostrzeżeń. Jeśli przedstawiciel poinformuje Cię, że dany produkt nie nadaje się do Twojego zastosowania lub jest słabo dopasowany, poważnie rozważ to ostrzeżenie.

Ocena oprogramowania

Kupując samochód, patrzysz nie tylko na cenę zakupu. Bierzesz pod uwagę oszczędność paliwa, a także koszty niezbędnych rzeczy, takich jak konserwacja i ubezpieczenie. Możesz zrezygnować z taniego samochodu na rzecz droższego, jeśli spodziewasz się, że droższy będzie łatwiejszy w utrzymaniu lub będzie zużywał mniej paliwa. Przy wyborze oprogramowania racjonalnie jest brać pod uwagę całkowity koszt posiadania. Wybór oprogramowania działa tak samo, jak wybór samochodu. Wybierz coś, co pozwoli Ci dotrzeć tam, gdzie chcesz, za rozsądną cenę. Weź pod uwagę wszystkie koszty, nie tylko cenę początkową. Wśród najczęstszych kosztów, które napotkasz, są

✓ Wsparcie techniczne: Zapytaj, czy dostępne jest wsparcie techniczne, czy jest zawarte w oprogramowaniu i czy będzie wystarczające dla Twoich potrzeb. Czy dostępne są wyższe poziomy wsparcia, a jeśli tak, to jaki jest koszt?

✓ Szkolenie: Przejrzyj dostępne opcje szkoleniowe - mogą one obejmować wbudowane samouczki, biblioteki szkoleniowe online lub zajęcia na żywo. Znajdź abysprawdzić, ile kosztuje każda opcja. Jeśli masz kilka osób, które będą potrzebowały szkolenia, zapytaj o zniżki dla wielu uczestników.

✓ Praca: Niektóre narzędzia wymagają znacznie więcej czasu i wysiłku do wykonania typowych zadań niż inne. Koszt wykwalifikowanej siły roboczej jest ważnym elementem całkowitego kosztu posiadania oprogramowania.

Obecnie dostępnych jest wiele bezpłatnych pakietów oprogramowania, w tym kilka produktów do eksploracji danych, a także powiązanych narzędzi, takich jak języki programowania statystycznego. To brzmi jak okazja i pod pewnymi względami tak jest. Jednak wykonywanie pracy zawsze oznacza coś więcej niż tylko pobieranie oprogramowania. Nadal będziesz potrzebować oprogramowania o odpowiednich możliwościach do wykonywania swojej pracy, czegoś, co jest dość łatwe w użyciu i zgodne z innymi używanymi narzędziami. Będziesz potrzebować pomocy technicznej i szkolenia. Oceń więc wolne oprogramowanie tak, jak każdą inną opcję.

Darmowe oprogramowanie nie zawsze jest darmowe!

Czekaj, co? Zgadza się, produkty, które są bezpłatne w niektórych sytuacjach, mogą nie być bezpłatne w innych. Może być konieczne uiszczenie opłaty za licencję, aby móc korzystać z pewnych funkcji o dużej wartości, korzystać z oprogramowania w celach komercyjnych lub z innych powodów. Przeczytaj umowy użytkownika i upewnij się, że je rozumiesz i szanujesz. Upewnij się, że wiesz, jakiej licencji będziesz potrzebować, i rozlicz koszty licencji przy wyborze oprogramowania do swoich projektów.

Nie zakochuj się (w swoim oprogramowaniu)

Wiele osób tak bardzo przywiązuje się do swojego oprogramowania, że staje się ono centralnym elementem ich tożsamości. Możesz na przykład spotkać programistów, którzy identyfikują się na podstawie języka, w którym programują, lub architektów baz danych, którzy identyfikują się z określonym rodzajem bazy danych. Nie daj Boże, że pewnego dnia najlepszym narzędziem do zadania będzie, no cóż, inne. Eksplorator danych prawdopodobnie nie podejdzie do Ciebie i nie powie: „Jestem eksploratorem danych [Marki X]", ale to nie znaczy, że lojalność jest mniej silna. Ludzie przywiązują się do swoich narzędzi i czasami opierają się zmianom. Jeśli zdobędziesz narzędzia, które od początku będą odpowiadać Twoim potrzebom, prawdopodobnie będziesz mógł z nich korzystać przez długi czas. Zmiana narzędzi wymaga wysiłku, więc nie chcesz dokonywać częstych zmian. Ale warunki też się zmieniają. Twoje obecne oprogramowanie może nie wyglądać na tak świetny za kilka lat, gdy Twoje potrzeby staną się bardziej złożone. Konkurencyjne produkty mogą wprowadzać ulepszenia, które są dla Ciebie atrakcyjne. Być może sprowadzisz nowych pracowników, którzy potrzebują czegoś łatwiejszego w obsłudze lub bardziej elastycznego. Dlatego miej otwarty umysł i od czasu do czasu przeglądaj swoje potrzeby i produkty, które są na rynku. Są też inne powody, aby nie przywiązywać się zbytnio do swojego ulubionego oprogramowania. Twój pracodawca może nalegać na zmianę lub możesz zmienić pracodawców i musisz dostosować się do nowego sposobu działania. Twoi koledzy mogą nie podzielać Twojego entuzjazmu dla konkretnego narzędzia, a Twoje osobiste preferencje mogą nie zwyciężyć. Uparte przywiązanie do konkretnego produktu może zawęzić Twoje możliwości, sprawić, że będziesz wyglądać głupio i niepotrzebnie irytować współpracowników. Pamiętaj, nie istnieje doskonałe oprogramowanie i zawsze znajdziesz więcej niż jeden sposób na wykonanie pracy. To, co wiesz i robisz, zawsze będzie ważniejsze niż narzędzia, których używasz. Jeśli zastanawiasz się nad trudnym testem swojego oprogramowania (często nazywanym „weryfikacją koncepcji"), takim jak ukończenie małego projektu, opracuj kryteria sukcesu przed rozpoczęciem testu. Zdefiniuj swoje wymagania dotyczące pomyślnego testu: Określ, jakie zadania należy wykonać i jakie wyniki należy osiągnąć, aby test został nazwany sukcesem. Te wymagania, znane jako „kryteria sukcesu", pomagają utrzymać test na ścieżce i pozwalają przejść przez etap testowy i przejść do produktywnej pracy. Kryteria sukcesu są również przydatne, jeśli prosisz dostawców o dostarczenie kopii ewaluacyjnych

kosztownych produktów lub o wsparcie ich personelu podczas testu, ponieważ zdefiniowanie i udostępnienie kryteriów sukcesu wskazuje, że jesteś poważnym potencjalnym nabywcą.

Współpraca z przedstawicielami handlowymi

Kiedy zaczynasz eksplorację danych, możesz być znacznie bardziej niż zwykle zaangażowany w wybór i pozyskiwanie swoich narzędzi. Pracodawca może nie dać Ci żadnego wyboru, jakiego komputera lub aplikacji biurowych użyć, ale jeśli chodzi o eksplorację danych, użytkownicy końcowi zwykle aktywnie angażują się w identyfikowanie produktów do rozważenia, testowania i oceny, a nawet w negocjacje ze specjalistami ds. sprzedaży. Będziesz mieć bardziej produktywne interakcje z dostawcami oprogramowania, jeśli zrozumiesz, co dzieje się na końcu ich rozmowy. Chwila, o co chodzi ze specjalistami ds. sprzedaży? Po co zajmować się sprzedawcami, skoro oprogramowanie jest dostępne bezpłatnie? Dlatego:

✓ Licencjonowanie: licencje na bezpłatne oprogramowanie niekoniecznie obejmują wszystkie zastosowania. Produkt, który lubisz, może nie być darmowy do użytku, który masz na myśli.

✓ Możliwości: bezpłatne produkty często mają ograniczoną funkcjonalność. Te ograniczenia mogą być strategicznymi wyborami mającymi na celu zmuszenie do zapłaty za uaktualnienie.

✓ Usługi: dostawcy oferujący bezpłatne oprogramowanie często zarabiają na opłatach za usługi.

Możesz pominąć resztę tego rozdziału, jeśli znalazłeś darmową aplikację do eksploracji danych, którą kochasz, i

✓ Nie potrzebujesz wsparcia technicznego, szkolenia ani gwarancji.

✓ Masz pewność, że z tą aplikacją możesz pracować tak samo wydajnie, jak z każdą inną.

✓ Zasady IT organizacji pozwalają na korzystanie z tego produktu.

Możesz mieć długoterminowe cele i silne osobiste zaangażowanie w rodzaj wykonywanej pracy. Twój pracodawca może mocno zachęcać do długoterminowej perspektywy, a jeśli nie, to Twoi rówieśnicy. Przedstawiciele handlowi są prawie zawsze oceniani na podstawie wyników krótkoterminowych. Tak więc wszyscy przedstawiciele handlowi są bardzo zainteresowani możliwościami, które mogą szybko doprowadzić do sprzedaży, a im większa sprzedaż, tym lepiej. Naprawdę dobry przedstawiciel handlowy dołoży starań, aby zrozumieć Twój biznes i długoterminowe cele. Czemu? Wydajna, długoterminowa relacja biznesowa maksymalizuje potencjał uzyskania wielu sprzedaży z biegiem czasu. Życie zawodowe handlowca przebiega cyklicznie – z celami na każdy kwartał, a także na cały rok obrotowy. Żadne cele nie są ustalane z wyprzedzeniem większym niż rok. Przedstawiciele handlowi często zmieniają role, często zmieniając terytoria w firmie co roku i przenosząc się z jednej firmy do drugiej co kilka lat. Bądź świadomy cykli sprzedaży w kontaktach z dostawcami. Przedstawiciele są zajęci i pod dużą presją, aby osiągnąć limity sprzedaży w ostatnim miesiącu każdego kwartału (marzec, czerwiec, wrzesień i grudzień). Jeśli będziesz badać produkty przez kilka miesięcy, lepiej rozpocząć rozmowy z dostawcami na początku kwartału, kiedy przedstawiciele będą mieli więcej czasu i cierpliwości. Jeśli to możliwe, zaplanuj sfinalizowanie zakupów pod koniec kwartału, kiedy naciskani przedstawiciele mogą być bardziej zmotywowani do oferowania rabatów. Przedstawiciele handlowi są zobowiązani do wypełniania limitów sprzedaży do końca każdego kwartału roku, więc im bliżej końca kwartału, tym większe prawdopodobieństwo, że dostaniesz dobrą ofertę. Koniec marca, czerwca, września i grudnia to najlepszy czas na uzyskanie rabatów, ale aby je otrzymać, musisz być gotowy i mieć możliwość natychmiastowego zakupu.

Mantra sprzedawcy - BANT

Przedstawiciele handlowi zadają wiele pytań. Rozmowa przedstawiciela handlowego twojego dostawcy to nie tylko rozmowa. Celem jest zrównoważenie dwóch interesów – Twoich potrzeb i dostawcy. Żadna sprzedaż nie nastąpi, dopóki nie będziesz przekonany, że produkt spełnia Twoje wymagania, więc przedstawiciel poszuka informacji o Twoich celach i oczekiwaniach. A sprzedawca musi wiedzieć, czy jesteś poważnym potencjalnym klientem. Połączenie Budżetu, Uprawnień (Authority), Potrzeb (Needs) i Ram czasowych (Time frame) (BANT) jest powszechnie akceptowane jako standard „kwalifikacji" potencjalnego klienta. Innymi słowy, BANT to metoda używana przez sprzedawców do określenia, czy istnieje uzasadniona okazja do sprzedaży czegoś. Spodziewaj się odpowiedzi na pytania dotyczące tych tematów:

✓ Budżet: Musisz wiedzieć, czy na zakup został przydzielony budżet i ile pieniędzy jest dostępnych. Dokładne budżety mogą być czymś, co chcesz zachować w tajemnicy. Nie powinno to stanowić problemu, ale powinieneś zaplanować udostępnienie wystarczającej ilości informacji, aby poinformować przedstawiciela, czy realnie możesz sobie pozwolić na zakup omawianych produktów.

✓ Uprawnienia (Authority): Zapoznaj się z procesem zakupu oprogramowania i powiązanych usług w Twojej organizacji przed rozpoczęciem rozmów z przedstawicielami handlowymi. Zakupy oprogramowania są często opóźniane, gdy osoby, które sądziły, że mają uprawnienia do samodzielnego dokonania zakupu, odkrywają, że ich pracodawca nie wystawi zamówienia bez takich elementów, jak uzyskanie konkurencyjnych ofert lub zatwierdzenie z działu technologii informatycznych.

✓ Potrzeby (Needs): Przedstawiciele handlowi pytają o Twoje potrzeby, aby mogli zidentyfikować i zaoferować odpowiednie produkty oraz udzielić Ci odpowiednich informacji. Ale znajdziesz inną stronę dyskusji o potrzebach. Przedstawiciel wie, że brak potrzeby oznacza brak sprzedaży, a oferowane produkty i usługi muszą być dostosowane do skali problemu, z którym się zmagasz. Nie ma sensu oferować rozwiązania problemu o wartości 100 000 $ za milion dolarów.

✓ Ramy czasowe (Time frame): Twoje ramy czasowe wskazują, czy i kiedy zamierzasz dokonać zakupu. W większości przypadków im bliżej terminu, tym poważniej się pojawisz. Docelowa data zakupu większa niż 12 miesięcy w przyszłości wskazuje, że twoje zainteresowanie jest oparte na ciekawości, a nie na poważnym zainteresowaniu, ponieważ budżety oprogramowania rzadko są ustalane z wyprzedzeniem większym niż rok.

Możesz ulec pokusie, aby w ogóle nie udostępniać żadnych informacji. Chociaż możesz mieć uzasadnione obawy dotyczące prywatności i ustalenia pozycji negocjacyjnej, całkowita tajność nie jest najlepszą strategią ochrony Twoich interesów przy zakupie oprogramowania. Bez odpowiednich informacji o Twoich potrzebach dostawcy mogą nie być w stanie zidentyfikować najbardziej odpowiednich dla Ciebie narzędzi lub schematów cenowych, a bez dowodu, że jesteś poważnym potencjalnym klientem, mogą nie być w stanie zaoferować potrzebnego wsparcia.

Zagłębianie się w Twoje dane

Wiele organizacji posiada górę danych, które zostały zebrane w trakcie rutynowych działań biznesowych i każdego dnia dodają nowe dane. Jako eksplorator danych użyjesz tych wewnętrznych danych jako podstawowego zasobu naturalnego. Ten rozdział koncentruje się na określeniu problemu i znalezieniu odpowiednich danych w istniejących zasobach. Jeśli masz pod ręką więcej danych, niż wiesz, co zrobić, znajdujesz się w sytuacji, w której stworzono eksplorację danych. Ale z drugiej strony, jeśli Twoje zasoby danych wydają się skąpe, nie martw się. Idee zawarte w tej Części nadal odnoszą się do ciebie. Wykorzystaj w pełni to wszystko co masz!

Koncentrowanie się na problemie

Projekt eksploracji danych rozpoczyna się, gdy zidentyfikujesz konkretny problem biznesowy do zbadania. Im węższe i lepiej zdefiniowane pytanie, tym skuteczniej można na nie odpowiedzieć. Im jaśniej zdefiniowano pytanie, tym wyraźniej można zrozumieć wymagania dotyczące danych, a także ograniczenia odpowiedzi. Jeśli masz do czynienia z bardzo ogólnym problemem (takim jak „Dlaczego nie sprzedajemy wystarczająco dużo?"), warto najpierw podzielić pytanie na łatwe do opanowania części. Nie musisz od razu omawiać całego tematu; po prostu weź jedną wąską część wielkiego problemu i zacznij od tego. Weźmy na przykład pierwsze pytanie jednego sprzedawcy: „Ile powtarzających się transakcji dostajemy?" Na pierwszy rzut oka brzmi to jak proste, proste pytanie, ale w rzeczywistości jest to szerokie pytanie, które obejmuje wiele mniejszych, bardziej szczegółowych pytań, takich jak te:

✓ Ilu nowych klientów wraca?

✓ Ilu klientów po raz drugi wraca po raz trzeci?

✓ Czy klienci, którzy jako pierwsi kupują garnitury, wracają po buty?

✓ Czy ci, którzy dokonają małego zakupu, zwracają się za większe zakupy?

Aby odpowiedzieć na te pytania, trzeba tylko liczyć. Po prostu nie jest trudno obliczyć, ilu klientów wraca po raz drugi lub trzeci, jeśli masz jakiś sposób na identyfikację osób. To łatwe w przypadku sklepów internetowych, w których kupujących można śledzić za pomocą loginu do konta lub adresu e-mail. Tradycyjni sprzedawcy detaliczni mogą identyfikować klientów za pomocą domowych kart kredytowych lub kart lojalnościowych, chociaż nie każdy klient z nich korzysta. Od tego momentu, gdy sprzedawca bardziej zaznajomił się z eksploracją danych i potencjałem analiz predykcyjnych i narzędzi do eksploracji danych, jego pytania stały się bardziej wyrafinowane i zorientowane na działanie:

✓ Jak kwota wydana na pierwszej wizycie ma się do długoterminowych wydatków?

✓ Jakie zachowania lub cechy są wskaźnikami wysokich przyszłych wydatków? Jeśli tak, to czym one są?

✓ Czy dodatkowe informacje (na przykład dane demograficzne) poprawiłyby naszą zdolność przewidywania zachowań zakupowych klienta?

Celem eksploracji danych jest wyjście poza zwykłą wiedzę o tym, co już się wydarzyło i zrozumienie, w jaki sposób możesz wpłynąć na to, co wydarzy się w przyszłości.

Opierając się na wiedzy biznesowej

Najbardziej podstawowymi danymi potrzebnymi do każdego projektu eksploracji danych nie są dane przechowywane w plikach elektronicznych. To wiedza biznesowa, którą Ty i inni członkowie Twojego zespołu zgromadziliście na podstawie własnego doświadczenia i szkoleń. Nie musisz być czołowym ekspertem w dziedzinie, którą badasz, ale musisz rozumieć podstawy biznesu. Musisz znać definicje pól w danych i trochę o tym, jak dane są zbierane i jakie wady mogą wystąpić w danych. Jeśli wiesz więcej, tym lepiej.

Nierealistyczne oczekiwania jednego klienta

Niektórzy ludzie mają nierealistyczne oczekiwania dotyczące eksploracji danych. Mówiąc prościej, myślą, że to magia i oczekują rezultatów, które tylko magia może zapewnić. Ale jest to naprawdę praktyczny proces, który pomaga wykorzystać wiedzę biznesową i kilka dobrych narzędzi do szybkiego wyodrębniania przydatnych informacji z danych, dzięki czemu można je wykorzystać do rozwiązania konkretnego problemu biznesowego. Oto przykład z życia wzięty nierealistycznego oczekiwania co do eksploracji danych i tego, jak przeszkodziło to w procesie eksploracji danych. Duża firma ubezpieczeniowa wysłała mi próbkę danych i poprosiła o wyniki. Ale firma ubezpieczeniowa nie zgłosiła żadnych problemów biznesowych, którymi należałoby się zająć. W rzeczywistości nie oznaczył nawet żadnych danych. Spojrzałem na plik danych i stwierdziłem, że to tylko nieoznaczone kolumny i wiersze danych. Pracownicy firmy ubezpieczeniowej uważali, że to wszystko, czego potrzebuję. Takie sytuacje nie są rzadkie i stanowią wyzwanie, z którym eksploratorzy danych muszą cierpliwie się uporać. Skontaktowałem się z klientem, aby wyjaśnić, że nie mogę nic zrobić, nie wiedząc, jakie są zmienne w danych. Wyjaśniłem, że nie, eksploracja danych naprawdę nie działa w ten sposób, poprosiłem klienta o dostarczenie brakujących informacji, a następnie poczekałem, aż zostaną zebrane i przesłane do mnie. Wszystko to wymagało czasu. Nawet wtedy klient nie chciał ujawnić żadnego konkretnego problemu biznesowego do rozwiązania, więc musiałem zgadywać. Podejrzewałem, że przetwarzanie roszczeń będzie poważnym problemem związanym z kosztami i satysfakcją klienta, i potwierdziłem to z ekspertem branżowym, co zajęło więcej czasu. Kiedy praca została wykonana (skoncentrowana analiza czasu potrzebnego na rozpatrzenie roszczeń, identyfikacja urzędów, które rozpatrywały roszczenia szybciej niż inne, w celu zbadania praktyk stosowanych przez te urzędy jako modeli dla usprawnienia procesów gdzie indziej), nie byłem t pewność, że klient doceni wyniki, ponieważ klient nigdy nie wyrażał rzeczywistego zainteresowania wprowadzeniem informacji w życie. Pomyśl, o ile szybszy i lepszy byłby ten proces, gdyby klient zaczął od realistycznych oczekiwań i od początku współpracował ze mną, aby omówić kwestie biznesowe.

Zakres zarządzania

Zadawanie pytań i badanie danych może być świetną zabawą. Teraz, gdy jesteś eksploratorem danych, przekonasz się, że możesz zadawać i odpowiadać na pytania, które wcześniej były poza Twoim zasięgiem. Znajdowanie odpowiedzi jest motywujące. Wymyślisz jeszcze więcej pytań. Być może odkryjesz coś tak fajnego, że zechcesz o tym wszystkim opowiedzieć. To wszystko jest tak ekscytujące, że łatwo może wymknąć się spod kontroli! Nie tylko twoje własne zainteresowania mogą spowodować rozszerzenie zakresu projektu. Podczas pracy będziesz dyskutować ze współpracownikami i wszyscy będą mieli pomysły i pytania, które zainspirują do dalszych poszukiwań. Nie ma ograniczeń co do źródeł inspiracji dostępnych dla eksploratora danych. Ale istnieje limit twojego dostępnego czasu. Podczas pracy musisz mieć na uwadze konkretne cele, a także realistyczny plan ich realizacji. Cele muszą być zdefiniowane w kategoriach biznesowych, które odpowiadają potrzebom Twojego menedżera lub klienta. Twój plan to pewność, że wyprodukujesz coś wartościowego, a nie tylko coś, co uznasz za interesujące. Twój plan jest Twoim przewodnikiem przy podejmowaniu decyzji, które pytania należy rozwiązać teraz, a które należy odłożyć na później. Skoncentruj się więc na konkretnych celach, odnieś się do swojego planu i nie pozwól, aby zakres projektu rozszerzał się lub wędrował, zanim zrealizujesz

swoje cele. (Szczegółowe omówienie planowania eksploracji danych znajduje się w rozdziale 6.) Co zrobić, jeśli zrealizowałeś cele projektu i nadal masz czas przed terminem? Fantastyczny! Teraz masz możliwość zbadania jednego lub więcej najlepszych nowych pytań, które przyszły Ci do głowy, i dodania wartościowego dodatku do końcowej prezentacji.

Jak sprzedawca był podekscytowany eksploracją danych

Zanim ludzie zaczną entuzjastycznie podchodzić do eksploracji danych, zwykle są źli lub sfrustrowani czymś innym. Jako eksplorator danych najbardziej satysfakcjonujące chwile i najlepsze okazje do stworzenia lojalnych fanów Twojej pracy leżą w rozwiązywaniu problemów, które sprawiają, że menedżerowie nie mogą spać w nocy. Wcześniej opowiedziałem ci o sprzedawcy, który zaczął od zadawania prostych pytań, takich jak „Ilu klientów, którzy po raz pierwszy wracają na drugą wizytę?" i stopniowo ewoluowała w bardziej wyrafinowane, zorientowane na działanie pytania, takie jak: „Jakie cechy klientów wiążą się z wysokimi poziomami wydatków długoterminowych?" O czym myślał ten sprzedawca, gdy proces się rozpoczął? Tylko jedno: brak pewnych pożądanych raportów. Sprzedawca zainwestował w bardzo drogie oprogramowanie w celu tworzenia rutynowych raportów na temat kilku prostych wskaźników, takich jak liczba nowych klientów powracających na drugą wizytę, ale oprogramowanie nie było skuteczne, a raporty nigdy się nie pojawiły. Zarząd nie był zadowolony. Kiedy więc detalista szukał lepszego rozwiązania, kierownictwo chciało udowodnić, że nowe rozwiązanie rzeczywiście generuje te raporty. Tylko raporty. Nikt nie prosił o eksplorację danych. Nikt nie myślał o zadaniu pytań, które mogłyby dać lepsze wskazówki do działania. Po prostu przekaż nam nasze raporty, proszę. Jeśli jako eksplorator danych słyszałeś o sytuacji tego sprzedawcy, możesz pokusić się o krzyk: „Zapomnij o tych starych raportach; eksploracja danych jest lepsza. Zobaczysz, że eksploracja danych jest znacznie potężniejsza niż jakikolwiek raport!" Ale to byłby zły sposób na zdobycie sprzedawcy. Oto stare powiedzenie: musisz być równy, zanim będziesz mógł być lepszy. Tak więc, jeśli twój menedżer lub klient chce czegoś konkretnego, musisz najpierw spełnić tę potrzebę. Pokażesz, że spełniasz wymagania. Kiedy to zrobisz, zdobędziesz szacunek wymagany, aby iść naprzód, zapewnić coś dodatkowego i nieoczekiwanego oraz być… . . lepszy. I tak właśnie stało się z tym sprzedawcą. Po ukończeniu raportu i spełnieniu wymagań sprzedawcy danych eksplorator danych mógł swobodnie kopać nieco głębiej. Zauważyła, że dane zawierały pewne informacje uzyskane za pośrednictwem programu lojalnościowego sprzedawcy, podstawowe informacje o klientach, a także szczegóły dotyczące ich domów i zainteresowań. Ale ta informacja często była pusta. Zastanawiała się: „Czy warto zbierać te informacje?" Dlatego szybko eksperymentowała z modelami drzewa decyzyjnego do przewidywania wydatków konsumentów. Testowała kombinacje danych behawioralnych (co klient kupił) i danych demograficznych (informacje o klientach zebrane podczas rejestracji w programach lojalnościowych). Odkryła, że dane programu lojalnościowego są nie tylko przydatne, ale łącząc je z informacjami sprzedażowymi przy pierwszym zakupie klienta, udało jej się również opracować zaskakująco dobrą prognozę długoterminowych wydatków klienta. Kiedy nadszedł czas na prezentację, eksplorator danych najpierw przedstawił raport, o który detalista poprosił na początku. Dopiero gdy sprzedawca sprawdził i poczuł się w pełni zadowolony z tego raportu, eksplorator danych pokazał coś więcej. I wow! Model wydatków klientów był świetnym finałem. Sprzedawca natychmiast stał się fanem eksploracji danych.

Korzystanie z własnych danych organizacji

Eksplorator danych nie ma nic bez danych. A jeśli pracujesz w dużej organizacji, będziesz mieć setki, a może tysiące istniejących zasobów danych potencjalnie dostępnych do eksploracji danych. Każda aktywność generuje rekordy, a te rekordy mogą stać się Twoim surowcem. Tabela 8-1 przedstawia różnorodność powszechnie gromadzonych danych w wielu działaniach biznesowych.

Działalność gospodarcza: Zebrane dane

Badania: informacje eksperymentalne i testowe informacji o produkcie konkurencji;

Produkcja: dane procesowe; ewidencja zamówień ewidencja produkcji ewidencja kontroli i testów,

Marketing : informacje marketingowe konkurencji i dane dotyczące sprzedaży, dane o kosztach marketingowych kampanii

Sprzedaż: Działalność sprzedażowa

dane sprzedażowe

Informacje dla klientów

Realizacja: zapisy dotyczące opakowań

ewidencja wysyłkowa

reklamacje wysyłkowe

Obsługa klienta: rekordy interakcji z klientem

reklamacje produktów i usług

problemy z obsługą

Wsparcie techniczne: prośby o wsparcie

raporty o problemach z produktem

projekt i inne sugestie dotyczące produktów

Szkolenie: Zapisy szkoleń personelu

rekordy szkoleń klientów

Certyfikacja i inne zapisy uwierzytelniające

Księgowość: Płatności rachunków

zapisy audytu

zebrane i zapłacone podatki

To dość długa lista, ale tak naprawdę to tylko niewielka próbka działań i powiązanych danych, które już czekają gdzieś w Twojej firmie. Ale świadomość, że dane istnieją, to nie to samo, co możliwość uzyskania do nich dostępu i wykorzystania ich do eksploracji danych. Po pierwsze, będziesz potrzebować znacznie bardziej szczegółowych informacji o tym, jakie dane wewnętrzne są istotne dla konkretnego problemu biznesowego, który badasz. Kto je zbiera? Kto kontroluje dostęp? Jakie zmienne (pola) są rejestrowane i dla jakiego zakresu czasu lub aktywności? Gdzie można znaleźć dokumentację?

Docenianie własnych danych

Ty i Twój menedżer możecie wybierać spośród wielu opcji przy wyborze projektu, który ma zostać rozwiązany za pomocą eksploracji danych. Zawsze masz wybór narzędzi. Ale jeśli chodzi o dane, możesz

nie mieć żadnego wyboru: korzystasz z danych dostępnych dla Ciebie lub Twojej firmy w tej chwili. Możesz mieć wątpliwości co do tych danych. Na pewno wiesz coś o jego wadach. Być może słyszałeś o innych organizacjach, które mają większe ilości danych lub inne typy danych niż twoje własne. Niemniej jednak, dane wewnętrzne Twojej organizacji, informacje gromadzone w toku codziennej działalności, są Twoim najcenniejszym zasobem. To najlepsze dane, jakie możesz mieć do eksploracji danych. Pod wieloma względami przewyższa wszystkie źródła zewnętrzne:

✓ Wyjątkowe znaczenie: dane dotyczą Twojej firmy, ze wszystkimi jej charakterystycznymi cechami. Chodzi o twoich własnych klientów, twoje własne produkty, twoje własne praktyki biznesowe. Cokolwiek odkryjesz w tych danych, z pewnością będzie miało również znaczenie dla firmy. Nikt nie będzie mógł odrzucić twoich wyników, ale nasza wymówka to inna wymówka.

✓ Przejrzystość: Znasz (lub możesz dowiedzieć się) źródła własnych danych. Nie powinno być żadnych tajemnic dotyczących definicji zmiennych, metod zbierania danych, czasu, miejsca czy zaangażowanych osób.

✓ Szczegóły: Będziesz mieć surowe dane, zebrane na możliwie najlepszym poziomie szczegółu.

✓ Zasięg: Twoje zasoby danych obejmują pełen zakres działalności prowadzonej w Twojej firmie.

✓ Przewaga konkurencyjna: Tylko Ty masz własne dane wewnętrzne. Nie jest dostępny dla Twoich obecnych lub przyszłych konkurentów.

✓ Potencjał rozwojowy: Możesz budować na własnych danych w sposób, który nie byłby możliwy w przypadku danych z jakiegokolwiek zewnętrznego źródła. Jeśli chcesz zintegrować informacje z wielu źródeł, Twoje dane będą zawierać identyfikatory, których potrzebujesz. Jeśli chcesz dowiedzieć się więcej o klientach, masz ich nazwiska i dane kontaktowe, możesz odwołać się do innych rekordów, przeprowadzić ankietę, a nawet zadzwonić i przeprowadzić osobistą rozmowę. Jeśli potrzebujesz bardziej szczegółowych lub dodatkowych danych, możesz zmienić praktykę gromadzenia danych.

Kolejna fajna rzecz dotycząca twoich danych: jesteś ich właścicielem. Wszelkie koszty zbierania danych pokryła jednostka biznesowa, która wygenerowała dane w pierwszej kolejności. Nie zapłacisz żadnych opłat i nie będziesz mieć problemów licencyjnych do rozważenia podczas używania i ponownego wykorzystywania danych. (Możesz napotkać problemy z przechowywaniem danych i innymi problemami z zarządzaniem danymi, ale dotyczy to każdego źródła danych). Twoje własne zasoby danych nie będą doskonałe pod każdym względem. Możesz odkryć, że niektóre dane, których chcesz użyć, nie zostały zebrane lub zostały odrzucone. Na pewno napotkasz pewne problemy z jakością danych. I oczywiście dane wewnętrzne mają ograniczenia – mówią o Twojej organizacji, ale nie o konkurencji. Mimo to dane wewnętrzne zawsze będą Twoimi podstawowymi i najważniejszymi cennymi zasobami danych.

Obchodzenie się z danymi z szacunkiem

Eksploracja danych, podobnie jak wszelkiego rodzaju analiza danych lub raportowanie, wykorzystuje dużo danych, znacznie więcej niż większość codziennych czynności biznesowych. Gdy uzyskujesz dostęp do danych i przeprowadzasz analizy, musisz zachować ostrożność, aby robić to w sposób zgodny z wytycznymi Twojej firmy i nie zakłócający rutynowych procesów biznesowych. Zasoby danych mogą być równie cenne i tak samo prywatne jak gotówka. Rozpocznij właściwy start w eksploracji danych, traktując dane z szacunkiem i odkrywając właściwe praktyki w zakresie zarządzania danymi i zarządzania, które mają wpływ na Twoją pracę. Nieprzestrzeganie prawnych i dobrych praktyk biznesowych w zakresie zarządzania danymi może prowadzić do poważnych problemów. Ważne jest,

aby dane nie były dostępne dla osób, które nie powinny ich używać, aby rekordy nie były nieprawidłowo zmieniane lub niszczone, a nowe dane, które tworzysz, były odpowiednio archiwizowane. Dokumentacja to konieczność. Wiele wymagań prawnych i dobrych praktyk biznesowych będzie istotnych dla Twojej pracy w eksploracji danych.

To może nie być proste. Będziesz musiał dowiedzieć się, jakie dane są dostępne, jak uzyskać dostęp i jak właściwie obchodzić się z danymi, aby nie przeszkadzać innym. Krótko mówiąc, będziesz musiał zaangażować się w nowe rzeczy i nowych ludzi. A będzie warto, bo dzięki temu zrobisz więcej i poszerzysz własne horyzonty. Będziesz musiał dowiedzieć się nowych rzeczy, ale nie będziesz musiał zostać ekspertem od zarządzania danymi. Możesz polegać na innych w swojej organizacji, którzy są ekspertami w zarządzaniu danymi i zarządzaniu danymi. Współpracuj z nimi w sposób konstruktywny, a pomogą Ci one przestrzegać prawa i stosować dobre praktyki zarządzania danymi. Eksperci danych i specjaliści ds. zarządzania danymi nie zawsze dobrze ze sobą współpracują. Eksperci danych są znani z tego, że nie zgadzają się na kontrolę dostępu do danych i czasami uciekają się do skomplikowanych schematów, aby uniknąć grania zgodnie z regułami dostępu do danych. Eksperci ds. zarządzania danymi nie zawsze rozumieją, dlaczego eksploratorzy danych muszą korzystać z tak dużej ilości danych; znane są z przeciągania. Frustrujące doświadczenia z przeszłości mogą wpływać na sposób, w jaki jedna grupa radzi sobie z drugą. Tak więc, gdy zaczynasz swoją karierę w eksploracji danych, postaraj się dotrzeć do osób z działu technologii informatycznych. Porozmawiaj z nimi o swojej pracy związanej z eksploracją danych i omów, jakie korzyści przyniesie to organizacji. Zapytaj o kwestie związane z zarządzaniem danymi, które dotyczą Twojej pracy i pokaż, że zależy Ci na dobrym zarządzaniu danymi. Ten okaz szacunku jest sposobem na rozpoczęcie pozytywnego i absolutnie niezbędnego partnerstwa roboczego między zespołami eksploracji danych i zarządzania danymi.

Tworzenie nowych danych

Najlepsze dane to Twoje własne dane. Twoje własne rzeczy są bardziej istotne dla Twojej organizacji i klientów niż jakiekolwiek dane, które możesz kupić, i często są bogatsze w szczegóły. I tylko ty to masz! Im lepsze są Twoje prywatne zasoby danych, tym większa jest Twoja przewaga informacyjna nad konkurencją. (Więcej informacji na temat cudów własnych wewnętrznych źródeł danych znajduje się w rozdziale 8.) Ale co, jeśli potrzebujesz danych, których jeszcze nie masz? Czy to Twoja wskazówka, aby szukać dostawcy danych? Być może . . . ale prawdopodobnie nie. Kiedy w pełni wykorzystasz posiadane dane, następnym krokiem jest budowanie na tym, czego się nauczyłeś, dodając dodatkową głębię i szczegóły. Chcesz wiedzieć więcej o swoich klientach, kim są, jak się zachowują i jak myślą. Potrzebujesz informacji dotyczących problemów biznesowych, które chcesz rozwiązać. W większości przypadków żaden dostawca nie ma danych, które odpowiadają na te konkretne potrzeby. Kiedy nie masz potrzebnych danych i nikt nie oferuje ich do sprzedaży, czas zacząć zbierać własne nowe dane. I tu właśnie pojawia się ta Część

Zgłębianie programów lojalnościowych

Detaliści wykorzystują eksplorację danych, aby uzyskać wskazówki dotyczące pozyskiwania i utrzymywania klientów oraz zachęcania ich do większych zakupów w bardziej dochodowy sposób. Eksplorację danych można również wykorzystać do usprawnienia procesów biznesowych w handlu detalicznym, obniżenia kosztów i poprawy obsługi klienta. Ale nie możesz tego zrobić bez odpowiednich danych. Jako eksplorator danych w sektorze detalicznym zaczniesz od zbadania zachowań klientów w prosty i ogólny sposób, a z czasem zaczniesz kopać głębiej, stopniowo dodając więcej szczegółów. Możesz zacząć od sprawdzenia, jakie produkty sprzedajesz iw jakiej ilości. Każdy sprzedawca prowadzi te rejestry. Następnie możesz zbadać kombinacje produktów sprzedawanych razem w ramach poszczególnych transakcji. Większość sprzedawców posiada te informacje, przynajmniej dotyczące ostatnich transakcji. (Ten rodzaj szczegółów danych jest czasami odrzucany przez personel IT, który błędnie uważa, że nie jest potrzebny. Jeśli dzieje się tak w Twojej firmie, usiądź z zespołem IT i wyjaśnij swoje potrzeby.) Następnym krokiem byłoby śledzenie poszczególnych klientów w czasie aby zrozumieć ich zachowania i wzorce zakupowe. Jednak wielu sprzedawców detalicznych nie zebrało i nie zachowało niezbędnych do tego danych. W jaki sposób detaliści mogą uzyskać dane potrzebne do zrozumienia zachowań klientów poza pojedynczymi transakcjami? Najczęściej stosowanym rozwiązaniem jest program lojalnościowy.

Zrozumienie koncepcji lojalności

Program lojalnościowy to umowa między firmą a jej klientami. Klienci zgadzają się, aby firma mogła śledzić zakupy (i ewentualnie także inne działania), a w zamian firma oferuje nagrody. Typowe nagrody to niższe ceny lub darmowy produkt lub usługa. Możesz być teraz zaangażowany w kilka programów lojalnościowych jako klient. Programy lojalnościowe dla linii lotniczych to programy lojalnościowe. Podobnie jak członkostwa w klubach hurtowych, karty preferowanych klientów, a nawet karty perforowane w kawiarniach. Każdy program lojalnościowy uzależniony jest od współpracy klienta. Klient musi najpierw zdecydować się na udział w programie, a następnie śledzić każdą transakcję. W najprostszych przypadkach klient może mieć przy sobie zwykłą papierową kartę dziurkowaną, którą firma zaznacza za każdym razem, gdy dokonuje zakupu. Po dokonaniu przez klienta wymaganej liczby zakupów (np. 10 lub 12) kartę można przekazać firmie w zamian za darmowy przedmiot. Ale karty dziurkowane nie dostarczają informacji potrzebnych do eksploracji danych! Bardziej wyrafinowane programy lojalnościowe dostarczają klientowi kartę przypominającą kartę kredytową, którą można zeskanować elektronicznie. Gdy klient zapomni swojej karty, może mieć alternatywę podania numeru członkowskiego, numeru telefonu lub nazwiska w celu zlokalizowania właściwego konta. Programy te

są ważne dla eksploratorów danych, ponieważ umożliwiają firmie szczegółowe śledzenie zachowań zakupowych klienta. Niektóre firmy używają tego typu śledzenia specjalnie do pozyskiwania danych do analizy. Ale inni mają inny motyw: papierowe karty i kupony można łatwo podrobić. Dlatego niektóre firmy wybierają te skomputeryzowane metody śledzenia przede wszystkim w celu ochrony przed stratami spowodowanymi oszustwami. Aplikacje na smartfony pozwalają klientom na identyfikację oraz otrzymywanie kuponów i innych ofert bez konieczności posiadania kart członkowskich. Te aplikacje są popularne wśród kupujących, którzy uważają je za wygodne. Dla sprzedawcy aplikacje na smartfony mogą zapewnić głębię informacji, które nie są dostępne w żadnym tradycyjnym programie lojalnościowym, w tym przeglądanie w czasie rzeczywistym i dane geograficzne. Programy lojalnościowe na smartfony ułatwiają również udostępnianie użytkownikom. Łatwiej jest udostępnić telefonowi świetną promocję w mediach społecznościowych niż przekazywać na przykład papierowe kupony. Sprzedawcy internetowi mają szczególne zalety. Ich klienci zwykle zakładają konta podczas pierwszego zakupu i dostarczają informacje, które mogą być wykorzystane do śledzenia, takie jak adres e-mail. Sprzedawcy ci niekoniecznie obiecują nagrody za założenie konta, ale mogą zbierać dane o transakcjach i wszelkiego rodzaju zachowaniach online. Mogą również śledzić odwiedzających za pomocą dzienników internetowych i plików cookie, śledząc informacje związane z przeglądaniem stron internetowych. Wiedzą, kiedy klient odwiedza witrynę i jakie produkty ogląda, nawet jeśli nie dochodzi do zakupu. Użytkownicy mogą dobrowolnie dodawać do swoich kont dodatkowe informacje, takie jak recenzje produktów, listy życzeń i profile użytkowników. Każdy bit tych informacji ma wartość dla eksploracji danych.

Twoja bonanza danych

Oto niektóre elementy danych, które mogą być dostępne dla Ciebie jako eksploratora danych w sektorze detalicznym:

✓ Lokalizacja klienta

✓ Zakupione produkty

✓ Kombinacje zakupionych razem produktów

✓ Ceny zapłacone

✓ Ceny katalogowe lub codzienne (często różne od ceny, którą zapłacił klient)

✓ Wykorzystano kupon lub inną ofertę rabatową

✓ Czas, w którym dokonano zakupu

✓ Szczegółowe opisy produktów

✓ Przeglądane strony/produkty

✓ Czas na miejscu

✓ Terminy wizyt na miejscu

✓ Recenzje produktów i udostępnianie informacji

✓ Odesłania (na przykład, czy klient przeszedł bezpośrednio na stronę, czy przez link z innej strony lub link w e-mailu?)

✓ Oferty lub reklamy, które oglądał klient

✓ Dane sieci społecznościowej, takie jak osoby, które klient zna

Ta informacja to skarbnica marketingu! Jako eksplorator danych trudno marzyć o bardziej wartościowym źródle danych, które pomoże firmie zrozumieć, co i jak sprzedawać osobom fizycznym. Ale bądź ostrożny. Istnieją granice właściwego wykorzystania tych danych osobowych. Jedne są określone przez prawo, inne przez wrażliwość społeczną i preferencje indywidualnego klienta. (Zapoznaj się z rozdziałem 3, aby uzyskać więcej informacji na temat współpracy z zespołem ds. technologii informatycznych w celu rozwiązania problemów związanych z prywatnością danych).

Wykorzystanie danych lojalnościowych do pracy

Teraz, gdy masz program lojalnościowy i dane, które generuje, co masz z nim zrobić? Jako eksplorator danych Twoim zadaniem jest dostarczanie decydentom analiz, które wspierają biznes. Niektórzy dyrektorzy rozumieją programy lojalnościowe i mogą prosić o określone informacje, być może więcej, niż masz na to godziny. Ale wielu innych nie pyta. Niektórzy dyrektorzy nie ufają danym, inni ich nie lubią, a wielu ich nie rozumie, ale najczęstszym powodem, dla którego dyrektorzy nie proszą Cię o informacje, jest to, że mają po prostu wiele innych rzeczy na głowie . Kiedy kierownictwo nie prosi o analizę, nie czekaj na telefon. To dobra okazja do podjęcia aktywnej roli. To więcej niż okazja; to konieczność! Twoja organizacja może mieć wielu dyrektorów, ale musisz traktować każdego indywidualnie. Twoja firma może wytwarzać 101 rodzajów przekąsek, ale osoba odpowiedzialna za chipsy kukurydziane chce tylko usłyszeć o chipsach kukurydzianych. On nie ma prawa do podejmowania decyzji dotyczących czekolady, krakersów czy bułek z owocami, a on też nie ma czasu o nich myśleć. Skoncentruj się na czymś, co jest ważne dla konkretnego decydenta. Jeśli nie znasz jeszcze priorytetów dyrektora, oto jak możesz to rozgryźć. Zacznij od zrozumienia obowiązków dyrektora. Mogą one być definiowane przez elementy, takie jak określone linie produktów lub geografia. Dyrektor będzie miał określone cele strategiczne i musisz wiedzieć, jakie one są. Następnie dowiedz się, jakie wskaźniki są najważniejsze dla przetrwania dyrektora. Na przykład wynagrodzenie kadry kierowniczej jest często powiązane z miernikami wyników biznesowych. Kiedy wiesz, jakie wskaźniki definiują wynagrodzenie dyrektora, będziesz dokładnie wiedział, na czym skoncentrować swoje wysiłki w zakresie eksploracji danych. Weź pod uwagę, że chcesz przedstawić przydatne analizy dla dyrektora odpowiedzialnego za marketing chipsów kukurydzianych w Kanadzie. Znając te obowiązki, znacznie zawęziłeś zakres; nie musisz brać pod uwagę żadnych linii produktów z wyjątkiem chipsów kukurydzianych ani żadnej lokalizacji poza Kanadą. Następnie spójrz na cele. Być może dyrektor ma na celu zwiększenie sprzedaży o 7 procent w tym roku. Oto klucz: dyrektorzy już wiedzą, co się stało i potrzebują, abyś pokazał im, w jaki sposób mogą wpłynąć na to, co stanie się dalej. Więc nie idź do dyrektora i nie wyjaśniaj, że sprzedaż chipsów kukurydzianych wzrosła o 4 procent. Jak dotąd w tym roku. Ktoś inny już to zrobił. Zamiast tego eksploruj dane, aby dowiedzieć się, jakie działania mogą zwiększyć sprzedaż. Zostań bohaterem eksploratora danych swojego decydenta, odkrywając przydatne informacje, takie jak

✓ Charakterystyka klientów kupujących duże ilości produktu

✓ Charakterystyka klientów, którzy zwiększają kupowaną kwotę

✓ Rosnące segmenty klientów

✓ Kombinacje produktów, które często kupowane są razem

✓ Promocje, które działają lepiej niż inne

✓ Kanały marketingowe, które są bardziej opłacalne niż inne

✓ Wzorce zachowań kupujących (w sklepie i online), które wpływają na sprzedaż

✓ Nieoczekiwane czynniki (lub kombinacje czynników) wpływające na sprzedaż

Eksploracja danych maksymalizuje zyski klubu magazynowego

Być może robiłeś zakupy w jednym z klubów magazynowych, sieci sklepów oferujących członkom - tylko zakupy w dużych, bez dodatków, sklepach. Kluby magazynowe mają gołe betonowe podłogi, proste funkcjonalne półki i ograniczony wybór produktów i rozmiarów opakowań. Ich stanowiska kasowe nie oferują toreb, nie mówiąc już o pakowaczach, do spakowania zakupów. Kluby magazynowe odróżniają się od typowych sprzedawców detalicznych, otwierając swoje drzwi tylko dla kupujących, którzy są gotowi płacić roczne składki członkowskie. Po co tworzyć tę barierę wejścia? Niektórzy zwracają uwagę, że członkostwo tworzy więź między kupującym a sklepem, motywację do powrotu i maksymalizację wartości zwracanej składki członkowskiej. A potem masz dane.

Ponieważ kupujący w klubach magazynowych muszą przedstawić karty członkowskie, aby dokonać zakupu, sprzedawcy ci dokładnie wiedzą, kto kupuje i co. Mogą szczegółowo śledzić każdą transakcję. Znają tożsamość kupującego, ponieważ potencjalni członkowie muszą zapewnić dowód tożsamości. Wiedzą, co kupuje kupujący. Znają czas i miejsce każdego zakupu. Wiedzą, jakie ceny zapłacił kupujący i czy w grę wchodziły jakieś specjalne promocje.

Tak więc kluby magazynowe mają dokładniejsze i pełne informacje o swoich kupujących niż jakiekolwiek inne sklepy fizyczne. W rzeczywistości mogą mieć lepsze informacje niż ich internetowa konkurencja. Bogate zasoby danych o zakupach konsumenckich, tożsamościowych i demograficznych umożliwiają hurtowniom wydobywanie ich danych i dostarczanie wyjątkowo wysokiej jakości informacji wspierających podejmowanie decyzji. Dane kupujących kopalnie mogą się ujawnić

✓ Charakterystykaę kupujących o wysokich wydatkach: Jak często i kiedy robią zakupy, jakie produkty kupują i inne dane demograficzne detale.

✓ Powiązania produktów: Grupy produktów często kupowanych razem.

✓ Relacje między różnymi ofertami: Czy ludzie, którzy przychodzą po benzynę, zostają w pobliżu, aby kupić artykuły spożywcze? Czy wydają więcej czy mniej niż inni? Czy kupują podobne czy różne produkty? A co z tymi, którzy kupują gaz, okulary lub leki na receptę? Która transakcja jest pierwsza i czy to mówi coś o kolejnych wzorcach zakupów?

✓ Dane geograficzne: Gdzie mieszkają kupujący? Jak daleko dojeżdżają do sklepu? Jak preferencje dotyczące produktów i wzorce zachowań różnią się w zależności od regionu?

Dobre praktyki gromadzenia i eksploracji danych zapewniają sklepom magazynowym dokładne i szczegółowe informacje o zachowaniu kupujących, które mogą wykorzystać do podejmowania świadomych decyzji o tym, jakie produkty oferować w każdym sklepie, jakie ceny mają pobierać, i innych kwestiach. Mogą również łączyć dane kupujących z innymi danymi biznesowymi, aby dowiedzieć się o produktywności, doskonaleniu procesów i jakości produktu. (Korzyści wykraczają poza eksplorację danych, gdy dane są wykorzystywane do informowania klientów o produkcie wycofuje lub w celu uproszczenia zwrotów i innych spraw związanych z obsługą klienta. Niektóre dane – jak na przykład zagregowane dane demograficzne nabywców związane z określonymi kategoriami produktów – można nawet sprzedać, aby stworzyć dodatkowy strumień przychodów.) Co to oznacza finansowo dla klubu hurtowni? Sieć klubów magazynowych Costco ma obecnie ponad 70 milionów członków i odnotowała przychody w wysokości ponad 100 miliardów dolarów za rok podatkowy 2013. Nikt nie twierdzi, że eksploracja danych jest jedynym powodem (Costco publicznie podkreśla znaczenie

dobrego zatrudniania, dobrego traktowania pracowników oraz szkolenia i awansu od wewnątrz), jednak eksploracja danych umożliwia Costco opieranie się na tych podstawach w oparciu o szczegółowe informacje o zachowaniu klientów i preferencje na poziomie lokalnym, a nawet indywidualnym.

Testowanie, testowanie . . .

Pamiętasz te zajęcia z nauk ścisłych, które brałeś dawno temu? Wiesz, co robią naukowcy, gdy potrzebują danych do swoich badań, prawda? Tworzą potrzebne dane, przeprowadzając eksperymenty i rejestrując wyniki. Prawdopodobnie sam przeprowadziłeś wtedy kilka eksperymentów. Być może eksperymentowałeś z wpływem temperatury na wzrost grzybów lub z wpływem katalizatora na reakcję chemiczną. Naukowcy, tacy jak chemicy, biolodzy i fizycy, cały czas przeprowadzają kontrolowane eksperymenty. Chemicy i fizycy mają dużą kontrolę nad warunkami swoich eksperymentów. Biologom, a jeszcze trudniej każdemu, kto prowadzi badania na ludziach, trudniej jest mieć doskonałą kontrolę nad eksperymentem. Ale na szczęście dla nas, my, eksploratorzy danych, nie wymagamy perfekcji. Czasami bierzemy dane, które już posiadamy i które nie zostały zebrane w kontrolowanych warunkach, i analizujemy je jak dane eksperymentalne. Jeśli odkryjemy coś interesującego, próbujemy tego samego podejścia z inną próbką danych lub nowymi danymi i sprawdzamy, czy otrzymujemy spójne wyniki. Jeśli nie masz odpowiednich danych, aby odpowiedzieć na konkretne pytanie lub chcesz uzyskać świeższe lub lepsze dane niż te, które masz, nadal możesz zrobić to, co robiłeś w szkole: przeprowadzić eksperyment.

Eksperymentowanie w marketingu bezpośrednim

Kiedy myślisz o eksperymencie, możesz wyobrazić sobie osobę ubraną w biały fartuch laboratoryjny i gogle, która patrzy na probówkę pełną jakiejś tajemniczej substancji. Ten obraz oraz brak fartuchów laboratoryjnych, okularów i probówek w twoim miejscu pracy mogą prowadzić do przekonania, że eksperymenty to coś, co robią inni ludzie w innych miejscach, ludzie, którzy różnią się od ciebie. Ale eksperymenty są przeprowadzane codziennie przez ludzi, którzy w ogóle nie pasują do tego obrazu, ludzi, którzy są ci bliżsi, niż możesz sądzić. Być może najczęstszym zastosowaniem do eksperymentów, legalnych eksperymentów kontrolowanych, podobnych do tych, których używają naukowcy, jest marketing bezpośredni. Kiedy widzisz reklamę w telewizji, na billboardzie lub w czasopiśmie, jest to marketing masowy, dostarczający szeroką wiadomość do szerokiej publiczności, bez dotarcia do konkretnych osób lub nawet wiedzy, kim są. Marketing bezpośredni polega na kontaktowaniu się z poszczególnymi osobami. Gdy otrzymujesz SMS-a lub e-maila od sprzedawcy, jest to marketing bezpośredni. Tradycyjne katalogi wysyłkowe, telefony od organizacji charytatywnych i listy wyborcze od kandydatów politycznych to wszystkie formy marketingu bezpośredniego. Skuteczni marketingowcy bezpośredni to agresywni eksperymentatorzy. Mogą nazywać swoje eksperymenty testami A/B, testami dzielonymi lub po prostu zwykłymi testami, ale są to po prostu terminologia branżowa dla kontrolowanych eksperymentów. Oto prosty i powszechny przykład testu marketingu bezpośredniego: sprzedawca internetowy wysyła e-maile, aby skontaktować się z klientami, którzy oglądali określony produkt, ale nie kupili go po 24 godzinach. Czy zmiany w wiadomości e-mail poprawią odpowiedź? Być może inny wiersz tematu skłoniłby więcej klientów do otwarcia wiadomości. Tę teorię można przetestować, pobierając próbkę klientów, dzieląc ją na dwie grupy, które są jak najbardziej podobne, i wysyłając do jednej grupy wiadomość, która jest już używana, podczas gdy druga grupa otrzymuje wiadomość testową, która jest identyczna, z wyjątkiem temat. Analiza odpowiedzi na każdą wiadomość pokazuje, czy istniała jakakolwiek różnica w działaniu dwóch tematów, a jeśli tak, to które zadziałały lepiej i o ile.

Szpiegowanie możliwości testowych

Wiele codziennych działań biznesowych w handlu, organizacjach non-profit i niektórych pracach rządowych sprowadza się do marketingu bezpośredniego. Jeśli wzywasz określone osoby (nawet jeśli są ich miliony) do podjęcia określonych działań, robisz marketing bezpośredni. Marketing bezpośredni służy nie tylko do sprzedaży, ale także do takich zastosowań:

✓ Pozyskiwanie funduszy

✓ Uzyskać głos

✓ Oddzielanie surowców wtórnych od innych śmieci

✓ Promowanie zdrowia publicznego

✓ Promowanie korzystania z usług publicznych

✓ Usuwanie nieefektywnych urządzeń z sieci energetycznej

✓ Pobieranie podatków

Zawsze, gdy masz listę osób i działanie, które chcesz, aby podjęli, pomyśl o testowaniu. Każdy aspekt marketingu bezpośredniego, który możesz zmienić a kontrola jest testowalna. Typowe przykłady rzeczy, które możesz przetestować, obejmują

✓ Kopiowanie: krótkie kontra długie, wiadomości lub różne sformułowania

✓ Układ: obrazy, odstępy, czcionki

✓ Koperta: kolory, kształt, papier

✓ Drukowanie: kolor, tekst, obrazy

✓ Załączniki: prezenty (takie jak etykiety adresowe, kartki z życzeniami, gotówka lub inne wartościowe przedmioty), koperty zwrotne, podziękowania, historia wręczania/zakupów

✓ Wiersze tematu: temat, sformułowanie, użycie nazwy odbiorcy

✓ Oferta: Cena, opakowanie, wysyłka

Testowanie online

Środowiska online oferują eksploratorom danych unikalną kombinację wyzwań i zalet związanych z gromadzeniem i analizą danych. Oto zła wiadomość. Formaty danych sieci Web mogą być trudne do importowania i manipulowania w aplikacjach do eksploracji danych. Systemy obsługujące strony internetowe są często słabo zintegrowane z systemami śledzenia sprzedaży, co utrudnia identyfikację powiązań między doświadczeniem odwiedzającego a wynikającymi z tego działaniami. Projektanci stron internetowych i webmasterzy nie zawsze mają na uwadze testowanie podczas opracowywania projektów lub wybierania technologii internetowych. Nawet jeśli nie ma wielkich wyzwań technicznych, ludzie mogą niechętnie otwierać platformy internetowe do eksperymentów. Są też dobre wieści. Dostępne są specjalne narzędzia, które znacznie upraszczają proces prawidłowego udostępniania stron internetowych do testowania, a także zapewniają możliwości analizy. Tak więc, bez względu na to, jakich narzędzi używasz do eksploracji danych, rozważ użycie specjalnego narzędzia do testowania stron internetowych. (Możesz znaleźć informacje na ten temat, wyszukując terminy, takie jak narzędzie do testowania A/B, narzędzie do testowania na wielu odmianach i narzędzie do testowania podzielonego.) Tylko jedno: aby testować online, musisz najpierw współpracować z

osobami odpowiedzialnymi za witrynę internetową Twojej organizacji . Jeśli nie masz tego teraz, czas otworzyć tę dyskusję. Narzędzia zaprojektowane do eksploracji danych niekoniecznie są idealne do testowania A/B wiadomości e-mail lub stron internetowych, a ich integracja może stanowić wyzwanie. Ale możesz być w stanie całkowicie ominąć te problemy. Wielu dostawców usług poczty e-mail oferuje wbudowaną funkcję testowania A/B. Jeśli korzystasz z którejś z głównych usług poczty e-mail, prawdopodobnie masz to już teraz dostępne. Do testowania projektowania stron internetowych , użyj ulubionej wyszukiwarki i słów kluczowych Testowanie A/B lub testowanie na wielu odmianach.

Mikrotargetowanie w celu wygrania wyborów

Większość kampanii politycznych polega na tym, że konsultanci przeprowadzają badania wyborców, albo radzą sobie z bardzo nieformalną oceną postaw wyborców i zainteresowania głosowaniem na konkretnego kandydata (lub głosowaniem w ogóle). Jednak w ostatnich latach niektóre kampanie polityczne, w tym zarówno kandydujące, jak i emisyjne, zaczęły wykorzystywać mikrotargetowanie, zorganizowane programy badań ankietowych i testowania komunikatów w celu opracowania i dostarczania spersonalizowanych komunikatów kampanii dostosowanych do indywidualnych wyborców.

Traktowanie wyborców jako jednostek

Pomyśl o różnicy między zakupami w centrum handlowym a zakupami w ulubionym sklepie internetowym. W centrum handlowym wszyscy widzą te same szyldy, te same ulotki, te same przedmioty na wystawie. Kupujący ma dostęp do wszystkiego, co jest dostępne, ale musi podjąć wysiłek, aby znaleźć najbardziej odpowiednie produkty. Twój ulubiony sklep internetowy nie wyświetla wszystkim tych samych reklam i produktów. Wykorzystuje Twoją przeszłą historię, aby dostosować prezentację. W sklepie internetowym widzisz reklamy, które zostały wybrane specjalnie dla Ciebie na podstawie takich czynników, jak produkty, które kupiłeś wcześniej, produkty, które oglądałeś, oraz produkty kupione przez inne osoby, których historia zakupów lub przeglądania przypomina Twoją. I że sklep internetowy testuje każdy element prezentacji (ofertę, tekst, obrazy, układ itp.), aby dowiedzieć się, co działa najlepiej. Typowa kampania polityczna może wykorzystywać program ankiet politycznych do identyfikacji kluczowych komunikatów dla wyborców jako całości lub duże segmenty wyborców, takie jak kobiety, seniorzy czy młodzież. Microtargeting analizuje każdego wyborcę indywidualnie i, podobnie jak sklep internetowy, wykorzystuje informacje o osobach do personalizacji kampanii.

Patrząc na przykład

Weź pod uwagę, że dwóch kandydatów, Fred Mertz i Lucy McGillicuddy, ubiega się o urząd. Fred użyje tradycyjnych technik kampanii. Kampania Lucy będzie wykorzystywała mikrotargetowanie. Czym będą się różnić te dwie kampanie? Obie kampanie będą wykorzystywać publicznie dostępne dane wyborców. Te zapisy zapewniają każdemu kandydatowi listę zarejestrowanych wyborców, adresy i historię głosowania. Zapisy nie ujawniają, jak głosują poszczególne osoby! Samo głosowanie jest zawsze tajne. Informują jednak, czy i kiedy ktoś głosował, i mogą zawierać szczegóły, takie jak przynależność do partii (niektóre regiony wymagają tej informacji do głosowania w prawyborach). Nawet Fred, tradycjonalista, rozumie, że nie wszyscy są identyczni. Ale ma niewiele informacji na temat postaw jednostek lub nie ma ich wcale. Wyśle więc te same broszury do wszystkich w okręgu i użyje tych samych kilku wiadomości we wszystkich swoich reklamach. Wyjdzie, żeby uścisnąć dłoń w każdym sąsiedztwie. Chociaż rozumie, że ludzie w różnych dzielnicach mają różne obawy, będzie miał tylko intuicję, która poprowadzi go w rozmowach z poszczególnymi osobami. W najlepszym razie będzie zgadywał, co powiedzieć każdej osobie.

Zwiększanie danych wyborców

Jak zmieni się mikrotargetowana kampania Lucy? Dla niej publiczne rejestry wyborców są tylko sercem zasobów danych. Baza danych wyborców Lucy będzie zawierać wiele informacji, których Fred nie posiada, takich jak

✓ Demografia

✓ Zawód

✓ Historia datków na cele polityczne i charytatywne

✓ Członkostwa

✓ Status własności domu, samochodu i łodzi

✓ Zezwolenia i licencje

✓ Prenumeraty czasopism

✓ Historia wolontariatu politycznego i inne wskaźniki poglądów politycznych

Skąd Lucy zdobywa te wszystkie informacje? Jej partia polityczna, prywatne źródła i jej własny zespół wzbogacają bazę danych wyborców o dodatkowe informacje o każdym indywidualnym wyborcy. Niektóre z tych informacji są dostępne w rejestrach publicznych, a niektóre można nabyć od komercyjnych dostawców danych, ale najcenniejsze informacje dla kampanii politycznej pochodzą z bezpośredniego kontaktu z potencjalnymi wyborcami.

Uzyskanie przewagi informacyjnej

Zbudowanie bazy wyborców Lucy wymaga dużo pracy! Nawet kandydat, który ma pieniądze i wiedzę, aby uzyskać dane z różnych źródeł i dopasować je do indywidualnych wyborców, nadal potrzebowałby dużo cierpliwości i pracy, aby zintegrować źródła danych. Niewiele kampanii ma takie zasoby. Na szczęście dla Lucy jej partia opracowała już ulepszoną bazę danych wyborców, z której może skorzystać jako punkt wyjścia. Rozpoczyna więc kampanię ze znaczną przewagą informacyjną nad Fredem. Jednak przed kampanią ukierunkowaną na mikroukierunkowanie wciąż jest dużo gromadzenia danych. A co z Fredem? Dlaczego jego partia nie dostarcza danych, aby pomóc jego kampanii? Główne partie polityczne w Stanach Zjednoczonych, Kanadzie i Wielkiej Brytanii ,teraz wszyscy mają bazy danych wyborców dla swoich kandydatów, więc wielu kandydatów może rozpocząć kampanie z zasobami danych podobnymi do zasobów Lucy. Jeśli Fred tego nie robi, być może nie zdaje sobie sprawy z tego, co oferuje jego partia, nie wie, jak tego użyć lub po prostu nie docenia wartości danych.

Opracowywanie własnych danych testowych

Chociaż Lucy teraz dużo wie o wyborcach, nie ma jeszcze informacji potrzebnych do dostosowania wiadomości do poszczególnych wyborców. Aby to osiągnąć, Lucy i jej zespół kampanijny muszą prowadzić ciągły program opracowywania i testowania konkretnych wiadomości. Najpierw jej zespół zidentyfikuje kilka głównych segmentów wyborców, korzystając z rozszerzonej bazy danych. Być może mają dane z niektórych wstępnych badań, które wykorzystają do podzielenia wyborców na trzy grupy: zdecydowanych zwolenników Freda Mertza, zdecydowanych zwolenników Lucy McGillicuddy i niezdecydowanych (lub przekonujących) wyborców. Wśród niezdecydowanych wyborców mogą następnie wybrać węższy segment, na przykład latynoskie matki pracujące. Następnie mogli przeprowadzić burzę mózgów na temat problemów i wiadomości, które mogą przemawiać do pracujących latynoskich matek. Zespół marketingowy opracowuje kilka przykładowych skryptów, z których każdy koncentruje się na jednym konkretnym problemie i przekazie. Komunikaty muszą być

spójne ze stanowiskiem kandydata w danej sprawie, ale będzie dostępnych wiele opcji zbadania, które kwestie należy podkreślić, a które komunikaty są najbardziej przekonujące. Jedynym sposobem, aby dowiedzieć się, co działa, jest przetestowanie. W typowym scenariuszu testowym ochotnicy otrzymaliby listy wyborców do wywołania i alternatywne skrypty do wykorzystania, takie jak jeden skupiający się na szkołach publicznych, a drugi na przychodniach zdrowia. Wolontariusze czytali scenariusz, a także zadawali pytania dotyczące prawdopodobieństwa głosowania na Lucy. Pod koniec testu kampania Lucy będzie miała nowe dane, odpowiedzi z ankiety zebrane podczas tych rozmów testowych. Kampania Lucy ma teraz unikalne dane z ankiet. Korzystając z tych danych, kampania Lucy odkrywa teraz, która z wiadomości testowych była najbardziej przekonująca dla określonej grupy wyborców. Ankieta może ujawnić szczegóły, które nie były konkretnie częścią testu. Wyborcy mogą w swoich komentarzach ujawnić coś nieoczekiwanego. Niektórzy z tej grupy mogą wspomnieć, że nie przejmują się tak bardzo szkołami publicznymi, ponieważ posyłają swoje dzieci do szkół parafialnych. Te informacje pomagają kandydatowi zrozumieć, dlaczego niektóre komunikaty działają lepiej niż inne. Wskazuje również na możliwość jeszcze głębszego zrozumienia jednostek. Kolejne badanie może porównać pracujące latynoskie matki, których dzieci chodzą do szkół publicznych, z pracującymi latynoskimi matkami, których dzieci chodzą do szkół parafialnych lub innych szkół prywatnych.

Dokonywanie odkryć na ścieżce kampanii

Teraz, gdy Lucy wie coś o przesłaniu, które przemawia do pracujących latynoskich matek, ona i jej wolontariusze zamierzają wykorzystać te informacje we wszystkim, od przemówień Lucy, przez ulotki, po punkty rozmów dla wolontariuszy przeszukujących sąsiedztwo. Zespół Lucy przeprowadzi podobne ankiety przez telefon, e-mail i twarzą w twarz przez całą kampanię. Jej baza danych będzie na bieżąco uzupełniana o nowe dane wyborców i nowe testy. W miarę pogłębiania się informacji Lucy, ona i jej wolontariusze będą z każdym dniem coraz lepiej przygotowani do przedstawiania poszczególnym wyborcom wiadomości, które są istotne i atrakcyjne dla danego wyborcy. Brzmi jak dużo pracy! A to dużo pracy. Poważna kampania mikrotargetowania przeprowadza codziennie ankiety i testy oraz wykorzystuje wyniki do codziennego informowania o działaniach kandydata, personelu kampanii i wolontariuszy w miarę postępu kampanii.

Historia sukcesu testów A/B wartych miliardy dolarów

Jaką wartość mogą dostarczyć dane testowe Twojej organizacji? Może być wart miliard dolarów lub coś jeszcze cenniejszego. Wybory prezydenckie w Stanach Zjednoczonych w 2012 r. były droższe niż jakiekolwiek poprzednie. Budżety kampanii rosły ponad dziesięciolecia, a te wybory okazały się kontrowersyjne: w 2008 roku prezydentura przesunęła się z kontroli republikańskiej na demokratyczną, podczas gdy w 2010 roku Republikanie zwiększyli swoją liczbę w Kongresie. Kampania reelekcji urzędującego prezydenta USA Baracka Obamy wyznaczyła bezprecedensowy cel pozyskiwania funduszy: miliard dolarów. Poprzednie kampanie Obamy były znane z pewnych mocnych stron: przyciąganie nowych wyborców, angażowanie darczyńców i wolontariuszy, którzy wcześniej nie byli aktywni politycznie, wnoszenie niewielkich wkładów od wielu osób oraz skuteczne łączenie tradycyjnych metod kampanii, takich jak wizyty od drzwi do domu, z nową taktyką w mediach społecznościowych . Aby osiągnąć swój cel pozyskiwania funduszy, kampania Obamy opierałaby się na istniejących mocnych stronach w przyciąganiu nowych i małych darczyńców poprzez agresywny program nagabywania przez e-mail. Innymi słowy, wykorzystali nowoczesną kampanię marketingu bezpośredniego. Kampania zmaksymalizowała wartość ofert e-mailowych za pomocą techniki, która była ulubioną przez sprzedawców bezpośrednich od prawie wieku: testów A/B. Podobnie jak sprzedawca testuje wiersze tematu, treść, ofertę i inne aspekty każdej reklamy, zespół ds. kampanii Obamy testował elementy jego e-maili dotyczących pozyskiwania funduszy. Odkryli, że więcej wiadomości oznacza więcej pieniędzy, ta kopia ma znaczenie i że wiele osób nie może się oprzeć

otwarciu wiadomości od Baracka Obamy z prostym tematem „Hej”. Ostatecznie kampania zebrała 1,1 miliarda dolarów, ładnie przekraczając swój i tak już bezprecedensowy cel. Połączył wyjątkowe pozyskiwanie funduszy z ukierunkowaną kampanią polityczną i nowatorskim wykorzystaniem analiz, aby zmaksymalizować wpływ wydatków na reklamy. A jeśli nie słyszeliście, Obama został ponownie wybrany w 2012 roku. Analitycy nie mogą kupić nikomu prezydentury, ale mogą pomóc każdej organizacji w maksymalnym wykorzystaniu jej zasobów.

Badanie krajobrazu publicznego

Ankiety mogą być najczęstszym i najbardziej znanym podejściem do uzyskiwania własnych unikalnych danych od ludzi. Każdy może napisać kilka pytań, przedstawić je komuś i gotowe. . . Ankieta. Dobre ankiety wymagają jednak przemyślenia i wysiłku.

Pozyskiwanie informacji za pomocą ankiet

W badaniach ankietowych ludzie proszeni są o odpowiedzi na pytania, zwykle o sobie. Typowe pytania ankiety dotyczą

✓ Demografia: wiek, płeć, zawód

✓ Zachowanie: Kupowanie lub używanie określonych produktów, wzorce wydatków, uczestnictwo w zajęciach towarzyskich lub sportowych

✓ Intencje: Głosuj lub nie głosuj, kandydat A lub B, zatrudnij mniej lub więcej nowych pracowników w przyszłym roku

✓ Postawy: Często dotyczące bieżących kwestii politycznych lub społecznych

Jeśli dane, które chcesz, mają związek z odczuciami lub działaniami ludzi, które mogą podjąć w przyszłości, ankieta może być jedyną opcją uzyskania danych. Ale ankiety są również wykorzystywane do uzyskiwania informacji, które są po prostu łatwiejsze do uzyskania za pomocą ankiety niż inne opcje. Na przykład, istnieje wiele informacji o wydatkach zawartych w rejestrach kart kredytowych, ale uzyskanie dostępu do tych rejestrów za pośrednictwem indywidualnego właściciela konta lub banków jest prawie niemożliwe (i nie bez powodu). Ale możesz zapytać ludzi, ile wydali na Twój produkt (lub konkurenta lub konkretną klasę produktów) w zeszłym roku lub ile zamierzają wydać w przyszłym roku, a wiele osób Ci powie. Dobre badania ankietowe oferują korzyści z

✓ Elastyczności: zadajesz pytania na dowolnie wybrany temat. Dzięki temu zawsze możesz uzyskać istotne informacje, nawet w przypadku tematów, w których żadne inne źródła danych nie są dla Ciebie dostępne.

✓ Szybkości: Ankiety można szybko skonfigurować i przeprowadzić, dzięki czemu Twoje dane będą aktualne.

✓ Głębokości: Użyj ankiet, aby wypełnić luki informacyjne, które inne dane pozostają otwarte.

✓ Prywatności: nie masz obowiązku dzielić się wynikami z nikim. Będziesz mieć przewagę informacyjną nad konkurencją.

Ale to nie jest tak proste, jak zapisanie kilku pytań i poproszenie kilku osób o odpowiedź. Pytania muszą być napisane poprawnie, aby były zrozumiałe i aby uzyskać odpowiedzi, które są dokładne i odpowiednie do Twoich potrzeb. I musisz otrzymać odpowiedzi od ludzi, którzy są reprezentatywni dla tych, których chcesz zrozumieć.

Korzystanie z ankiet

Ankiety są przydatne do zbierania danych o niemal każdym aspekcie ludzkiego życia. Ankiety możesz zignorować tylko wtedy, gdy Twój zawód nie ma nic wspólnego z ludźmi, na przykład astrofizyką. Z drugiej strony astrofizycy potrzebują ludzi do finansowania swoich badań i chcą, aby ludzie odwiedzali planetaria, więc mogą potrzebować również ankiet! Oto przykłady różnych zastosowań ankiet:

✓ Rząd: Oceń stan ekonomiczny, fizyczny i psychiczny ludzi i firm, aby wesprzeć działalność rządu. Na całym świecie istnieje prawie 200 krajowych agencji statystycznych, z których wszystkie wykorzystują badania ankietowe, aby lepiej zrozumieć swoich ludzi. Stany, hrabstwa i miasta przeprowadzają lokalne ankiety w tych samych celach.

✓ Psychologia: Studiuj zdrowie psychiczne i pracę ludzkiego umysłu.

✓ Socjologia i politologia: Zrozum nastawienie społeczne dotyczące bieżących problemów.

✓ Zdrowie publiczne: Dowiedz się, jak ludzie dbają o siebie, jakie dokonują wyborów dotyczących opcji zdrowotnych i dlaczego.

✓ Marketing i reklama: Mierz świadomość marki, preferencje produktowe i inne czynniki, które wpływają na zachowania zakupowe.

✓ Rzecznictwo i zarządzanie kampanią: Zidentyfikuj cechy zwolenników i krytyków oraz przetestuj opcje przesyłania wiadomości w ramach kampanii.

✓ Media: Uzyskuj informacje o nastrojach społecznych, które należy uwzględnić w raportach i przewidzieć wyniki wyborów.

✓ Obsługa klienta: Oceń satysfakcję klienta oraz zidentyfikuj problemy i możliwe rozwiązania.

Rozwijanie pytań

Dobre pytanie w ankiecie powinno być:

✓ Konkretny: Zajmij się tylko jednym pomysłem.

✓ Wąski: Ograniczony do określonych ram czasowych, lokalizacji lub innego zakresu, który jest odpowiedni dla Twoich potrzeb.

✓ Neutralny: Sformułowanie nie powinno prowadzić respondenta do odpowiedzi.

✓ Przejrzysty: łatwo zrozumiały dla każdego, kto mógłby wziąć udział w ankiecie.

Często warto oferować opcje odpowiedzi. Opcje odpowiedzi powinny być

✓ Proste: Aby wszyscy respondenci mogli je zrozumieć.

✓ Spójny: Wszystkie opcje odpowiedzi powinny mieć taką samą strukturę.

✓ Kompletny: Należy uwzględnić pełny zakres opcji.

✓ Różne: opcje nie mogą się pokrywać.

Te dwie ostatnie pozycje, obejmujące wszystkie możliwości i nie zachodzące na siebie, są często zaniedbywane. Rezultatem jest zamieszanie i frustracja dla respondentów, a dla Ciebie błędne dane. Dlatego zadbaj o rozwój swoich kwestionariuszy ankietowych.

Przeprowadzanie ankiet

Teraz, gdy masz już opracowane pytania ankietowe i zebrane w kwestionariusz, możesz skontaktować się z ludźmi w celu uzyskania odpowiedzi. Będziesz musiał wybrać co najmniej jeden kanał, aby dotrzeć do respondentów:

✓ Twarzą w twarz: Ankieter spotyka się osobiście z respondentem, zadaje pytania i rejestruje odpowiedzi. Ta metoda jest często stosowana w przypadku złożonych ankiet, takich jak ankiety medyczne lub ankiety, w których respondenci niechętnie udzielają odpowiedzi, jak w przypadku niektórych ankiet rządowych.

✓ Papier lub kiosk: Respondent otrzymuje formularz (lub po prostu odbiera formularz pozostawiony w dogodnym miejscu) lub jest kierowany do kiosku elektronicznego w celu wypełnienia ankiety. Często używane w ankietach dotyczących obsługi klienta.

✓ Poczta: Ankieta jest wysyłana pocztą do respondenta, który wypełnia ją i odsyła pocztą.

✓ Telefon: Ankieter dzwoni do respondenta, zadaje pytania i rejestruje odpowiedzi.

✓ Internet: Respondenci są rekrutowani i wypełniają ankietę online. Twój wybór spośród tych opcji zależy od wielu czynników. Pożądani respondenci mogą preferować niektóre kanały od innych. Niektóre kosztują więcej niż inne: rozmowa twarzą w twarz w domu respondenta kosztuje znacznie więcej niż przedstawienie tych samych pytań w Internecie. A czas potrzebny na wypełnienie ankiety różni się w zależności od sposobu jej przeprowadzenia.

Rozpoznawanie ograniczeń

Pomimo wielu pożądanych aspektów badań ankietowych, napotykasz również ograniczenia. Trudno jest uzyskać dobre dane, gdy tematem są ludzie, bez względu na to, jak się do tego zabierzesz. Nawet badacze naukowi, którzy dokładają wszelkich starań, aby prowadzić kontrolowane badania, nie mogą kontrolować warunków doświadczalnych na ludziach, tak jak robią to zwierzęta laboratoryjne. Dotarcie do odpowiednich respondentów w ankiecie nie zawsze jest łatwe. Do niektórych osób trudno jest dotrzeć; inni niechętnie uczestniczą. Ludzie, którzy są dostępni i chętni do odpowiedzi, mogą, ale nie muszą, mieć takie samo zachowanie i nastawienie jak ci, którzy nie są. Gdy masz satysfakcjonującą pulę respondentów, nie myśl, że Twoje kłopoty się skończyły. Możesz nie uzyskać odpowiedzi na wszystkie swoje pytania. Ludzie nie zawsze znają odpowiedzi. Pytanie, które wydaje Ci się proste, respondentowi może nie wydawać się proste. Być może pytałeś o dochody respondenta. Czy miałeś na myśli dochód jednej osoby czy gospodarstwa domowego? Czy obejmowałoby to dochód dzieci? Dochód niepodlegający opodatkowaniu i podlegający opodatkowaniu? A co ze stratami? Co jeśli dochód się zmienia? Respondent może się zastanawiać, czy właściwą odpowiedzią byłby poziom dochodów, które ostatnio miała, spodziewa się w najbliższym czasie lub zazwyczaj zarabia, a to mogą być trzy różne rzeczy. Badacze ankiet czasem zdają się zapominać, że respondenci są mniej zainteresowani tematem ankiety niż oni sami. Możesz być bardzo zainteresowany ketchupem, nawykami kupowania ketchupu, preferencjami smakowymi i konsystencji ketchupu oraz wszystkimi rzeczami związanymi z ketchupem, ale większość ludzi nie jest. Więc nawet chętny respondent może nie być w stanie odpowiedzieć na wszystkie szczegółowe pytania dotyczące ketchupu. Może kupić keczup, ale nie pamięta, kiedy i jak często, jaką cenę zapłacił lub jaką markę kupił, nie mówiąc już o tym, jak ta marka porównuje smak i konsystencję z każdą z głównych konkurencyjnych marek ketchupu. Więc nie oszukujmy się, jakiego poziomu głębokości można się spodziewać w odpowiedziach na ankietę. A potem najgorszy ze wszystkich problemów ankietowych jest to: nie zadawanie właściwych pytań. Możesz zadać każde pytanie, jakie przyjdzie Ci do głowy, a mimo to nie trafisz w cel. Najważniejszą kwestią w umyśle klienta

(lub pacjenta, części składowej lub członka) może być kwestia, o której po prostu nie pomyślałeś. Dlatego wiele ankiet kończy się pytaniem otwartym, takim jak „Czy jest coś jeszcze, co możemy zrobić, aby poprawić Twoje wrażenia?" Chociaż to dobry pomysł aby zadawać takie pytania, nie masz gwarancji, że otrzymasz wszystkie potrzebne informacje.

Sprowadzanie pomocy

Jesteś zajęty eksploracją danych. Jeśli jesteś jak większość eksploratorów danych, oznacza to, że pracujesz już w swoim zawodzie i dowiadujesz się o eksploracji danych, która pomoże Ci lepiej wykonywać swoją pracę. Jesteś zajęty! Masz ograniczenia co do czasu i energii, które możesz poświęcić na prowadzenie badań ankietowych, aby uzyskać potrzebne dane. Rób proste rzeczy samodzielnie, jeśli chcesz, ale gdy robi się trudno, sprowadź fachową pomoc. Firmy badawcze zajmujące się badaniami ankietowymi występują w dużych i małych rozmiarach, typach ogólnego przeznaczenia i niszowych oraz lokalnych i międzynarodowych. Nazwy i szczegóły ciągle się zmieniają, ale zawsze będziesz na dobrej drodze, jeśli szukasz dobrego punktu wyjścia w stowarzyszeniach zawodowych branży badań ankietowych.

Wejście w pole

Nie masz niektórych danych i nie możesz ich pobrać, kupić ani uzyskać, prosząc o nie. Czasami jedynym sposobem na uzyskanie potrzebnych danych jest wyjście w świat, obserwacja i pomiary. Albo poproś kogoś, żeby zrobił to za ciebie.

Udać się tam, gdzie nie dotarł jeszcze żaden eksplorator danych

Teraz, gdy jesteś eksploratorem danych, jesteś także głównym badaczem. Brzmi bardziej naukowo, prawda? Twoje badania są najważniejsze, ponieważ zaczniesz od surowych (podstawowych, nieprzetworzonych) danych i przeanalizujesz je, aby dodać coś nowego do światowej wiedzy. Prawdopodobnie włączysz też do swojej pracy dodatkowe badania. Innymi słowy, wykorzystasz również analizy wykonane przez Ciebie lub kogoś innego wcześniej. Możesz przeglądać dokumenty wewnętrzne, aby być na bieżąco z tym, co zostało zrobione w Twojej organizacji, lub odwiedzić bibliotekę, aby przeczytać artykuły naukowe. Możesz uzyskać dane, które nie są w stanie surowym, ale zostały już poddane pewnej analizie przetwarzania. Na przykład, gdy korzystasz z ostatnich danych spisowych, nie otrzymujesz informacji o osobach, ale raczej dane zbiorcze opisujące grupy ludzi w określonym regionie geograficznym. Wykorzystywanie tego rodzaju danych jest również formą badań wtórnych. W rzeczywistości nigdy nie powinieneś wychodzić, aby w żaden sposób zbierać nowe dane, dopóki nie zapoznasz się z dostępnymi opcjami – nie tylko w celu uzyskania surowych danych, ale także skorzystanie z jakiejkolwiek analizy twojego tematu, która została już wykonana. Badania podstawowe to to, co robisz, gdy nikt inny nie wykonał jeszcze pracy, aby uzyskać potrzebne informacje. Wyrób sobie nawyk przeglądania źródeł wewnętrznych i zewnętrznych przed rozpoczęciem jakiegokolwiek projektu, aby nie marnować zasobów na ponowne tworzenie informacji, które już istnieją.

Robić więcej niż prosić

To niezwykłe, ile danych można uzyskać, prosząc o nie. Programy lojalnościowe opierają się na pytaniu uczestników o ich dane, a miliony ludzi na całym świecie zgadzają się na udział. Ankiety to nic innego jak proszenie ludzi o dane, a trudno byłoby znaleźć osobę dorosłą, która nie odpowiedziała na co najmniej jedną ankietę. A to są cenne podejścia do badań pierwotnych. Ale nadal istnieją pewne rodzaje danych, których albo nie możesz uzyskać, albo których nie uzyskasz dobrej jakości, prosząc o informacje. Paco Underhill, autor książki Why We Buy: The Science of Shopping (wydanej przez Simon & Schuster), opisał swoje badania dotyczące zachowań klientów w sklepach. Mógł pytać ludzi o ich

ścieżki zakupowe lub powody, dla których zwlekali lub omijali określone miejsca. Ale gdyby zapytano cię, wchodząc do sklepu, jaką ścieżką masz zamiar obrać, jak dobrze myślisz, że możesz to przewidzieć? Jeśli zadano ci to samo pytanie przy wyjeździe, czy byłbyś w stanie zapamiętać? Ile cierpliwości miałbyś, żeby to przemyśleć i wyjaśnić szczegóły? Underhill i jego koledzy postanowili nie pytać kupujących o ich zachowanie, ale obserwować je bezpośrednio, za pomocą wideo i innych metod. San Francisco Chronicle ogłosiło go „Sherlockiem Holmesem dla detalistów".

Projektanci oprogramowania muszą wiedzieć, jak użytkownicy widzą ich projekty. Mogą chcieć wiedzieć, na czym użytkownicy skupiają swoją uwagę na stronie, czy są w stanie wykonywać zadania w zamierzony sposób i ile czasu im to zajmuje. Projektanci mogliby zapytać, ale użytkownikom prawdopodobnie trudno byłoby przekazać naprawdę skuteczną informację zwrotną na temat tych szczegółów. Dlatego projektanci oprogramowania obserwują ludzi korzystających z ich oprogramowania. Mogą to zrobić bezpośrednio, tworząc stacje, na których mogą oglądać użytkowników na żywo i osobiście. Lub mogą użyć opcji zdalnych opartych na naciśnięciach klawiszy przez użytkownika lub nagrywanie wideo za pomocą kamer internetowych. Jeśli twoja agencja rządowa chce wiedzieć, gdzie na świecie może mieć miejsce produkcja bomb, szkolenie wojskowe lub inna podejrzana działalność, z pewnością nie uzyskasz informacji, dzwoniąc do wszystkich szefów rządów na świecie i prosząc o nie. Dlatego szpiedzy kochają satelity! Możesz nie potrzebować niczego tak wymyślnego jak satelita, aby uzyskać potrzebne dane, ale możesz skorzystać z przyziemnych technik gromadzenia danych, takich jak te:

✓ Chcesz wiedzieć, skąd pochodzą kupujący w Twoim centrum handlowym? Przejdź przez parking i zwróć uwagę na stany na tablicach rejestracyjnych.

✓ Nie wierzysz szacunkom dotyczącym ruchu pieszego z lokalnej izby handlowej? Stań na ulicy i policz przechodniów.

✓ Dążysz do tego, aby ubrania lepiej pasowały do kobiet? Wyjmij miarkę . . . Proszę!

Nowe spojrzenie na swój problem

Kiedy przeanalizowałeś i przeanalizowałeś wszystkie dane, które możesz uzyskać, ale nadal nie możesz rozwiązać problemu, może nadszedł czas, aby wyjść z biura i poszukać inspiracji u innych. Grupa pielęgniarek chirurgicznych właśnie to zrobiła i dokonała odkrycia, które doprowadziło do 50-procentowego wzrostu wydajności ich placówki, bez utraty jakości. Te pielęgniarki nie były obce dane. Rozumieli i stosowali przyjęte metody statystyczne do poprawy jakości w opiece zdrowotnej. Ale byli też gotowi zrobić coś, czego wielu ludzi by nie zrobiło: opuścili szpital i udali się w teren, aby obserwować wybitnych profesjonalistów z zupełnie innej branży. Poszli na tor wyścigów samochodowych, aby obejrzeć załogę w akcji. Ekipy pitowe pracują bardzo szybko, a jednocześnie bardzo dobrze. Nawet jedna zmarnowana sekunda może zadecydować o wygranej lub przegranej kierowcy . Ale jakości nie można poświęcić na rzecz szybkości, ponieważ słaba jakość pracy może doprowadzić do śmierci kierowcy. Ćwiczą ekipy serwisowe. Tak jak muzycy grają na wagach, a baleriny ćwiczą przy drążku, tak ekipa pit-stopowa ćwiczy w kółko zadania, aby rozwijać szybkość i umiejętności. Pielęgniarki wykonywały swoją pracę, ale nigdy nie praktykowały. Postanowili spróbować. Praktyki umożliwiły zespołowi pielęgniarskiemu wykonywanie pracy szybciej, ale równie dobrze. W rezultacie wydajność ich placówki wzrosła z dwóch zabiegów dziennie do trzech dziennie. Pacjenci skrócili czas oczekiwania na operację, szpital uzyskał większe przychody, a pielęgniarki otrzymały świetne oceny wyników, a wszystko to dlatego, że pielęgniarki otworzyły swoje umysły i przetestowały świeże podejście do swojej pracy.

Jedno wyzwanie, wiele podejść

Czasami istnieją przeszkody w uzyskaniu danych, które chcesz rozwiązać w konkretnym przypadku biznesowym. Możesz napotkać wyzwania techniczne, problemy prawne lub wysokie koszty. Kiedy napotykasz dużą przeszkodę, postaraj się nie poświęcać dużo czasu na jej pokonanie. Zamiast tego poszukaj alternatywnych (i łatwiejszych) sposobów rozwiązania swojego problemu. Być może chcesz wiedzieć, ile zarabiają Twoi klienci, ale nie możesz uzyskać dostępu do tych danych. Może istnieje, ale nie możesz uzyskać pozwolenia na jego wykorzystanie, może polityka firmy zabrania zadawania tego pytania w ankiecie, a może klienci po prostu nie powiedzą. Nie martw się o powody; po prostu poszukaj innego kąta. Możesz być w stanie oszacować dochód na podstawie danych ze spisu ludności lub szacunków zakupu od dostawcy danych. Co więcej, możesz ponownie rozważyć, czy dochód jest naprawdę tak cenną informacją. Co naprawdę chcesz wiedzieć? Może nie jesteś tak zainteresowany tym, co ludzie robią, co tym, co wydają. Pytanie ludzi, ile pieniędzy zarobili w zeszłym roku, jest drażliwym pytaniem; pytanie, ile wydali na mydło do naczyń, nie jest. Do każdego projektu eksploracji danych można podejść na więcej niż jeden sposób. Większość rzeczy można zmierzyć na wiele sposobów. Jeśli Twój projekt się opóźnia, użyj swojego najlepszego kreatywnego myślenia, aby rozważyć alternatywy dotyczące rodzaju danych, których możesz użyć i sposobów, w jakie możesz je uzyskać. Ale twoje metody nie muszą być oryginalne, więc zwróć uwagę na to, w jaki sposób inni uzyskują dane, a nie tylko inni eksploratorzy danych. Naukowcy, marketerzy, szpiedzy – wszyscy zbierają dane, dzięki czemu od każdego można odkryć nowe podejście. Rozważmy następujący przypadek: Badacze z firmy Ansell (producent produktów opieki zdrowotnej), Uniwersytetu Indiany i tureckiego Ministerstwa Zdrowia stanęli przed tym samym wyzwaniem dotyczącym danych. Każdy zespół naukowców miał inny cel – w tym udoskonalenie projektu produktu, zrozumienie problemów związanych z wizerunkiem ciała i ochronę zdrowia publicznego, wszystkie ważne sprawy – ale nie istniało wiarygodne źródło danych, których potrzebowali. Następnie każdy zespół rozpoczął tworzenie nowych danych i każdy podchodził do tego w inny sposób. Czego szukali? Każda z tych organizacji potrzebowała pomiarów ludzkiego penisa. Możesz nigdy nie potrzebować danych, które są tak nieuchwytne, tak emocjonalnie naładowane lub tak niezręczne do zmierzenia, jak to. I o to chodzi. Jeśli znajdziesz wiele dobrych sposobów na uzyskanie tych danych, istnieje wiele dobrych sposobów na uzyskanie potrzebnych danych. Oto jak to zrobili:

✓ Ansell: Jako producent prezerwatyw Lifestyles, celem firmy Ansell było zapewnienie, że jego produkty będą zarówno funkcjonalne, jak i wygodne. Firma Ansell chciała, aby pomiary wykonane przez personel medyczny były dokładne. Pracownicy ustawili prywatne namioty przed klubem nocnym i rekrutowali ochotników, których mierzył personel medyczny. (Wyciągnięta lekcja: To może nie być takie wspaniałe podejście. Wielu mężczyzn nie było w stanie zrobić tego, co należało zrobić w celu pomiaru. Można się zastanawiać, ile musieli wypić w klubie nocnym i czy czuli się komfortowo robiąc to w namiot.)

✓ Uniwersytet Indiana: Zespół Centrum Promocji Zdrowia Seksualnego współpracował z Churchem i Dwightem, producentami prezerwatyw Trojan, więc byli również zainteresowani dopasowaniem prezerwatyw. Ale w tym przypadku w grę wchodziło większe badanie. Badanie obejmowało szereg zagadnień, w tym praktyki antykoncepcyjne i zapobiegawcze chorobom, a także obraz ciała. Zespół ten wierzył, że mężczyźni mogą uczciwie i dokładnie zgłaszać się do siebie, i zaoferował do tego zachętę – dopasowaną prezerwatywę.

✓ Ministerstwo Zdrowia, Turcja: Badanie to koncentrowało się na obrazie ciała i obawach, że mężczyźni mogą szukać niepotrzebnych interwencji medycznych. Badanie zostało przeprowadzone pod nadzorem szpitalnej komisji rewizyjnej, a badani zostali przebadani w celu wykluczenia osób z pewnymi problemami medycznymi. Wszystkie pomiary zostały wykonane przez tego samego lekarza (dla spójności) w kontrolowanym środowisku (temperatura, oświetlenie itd.) przy użyciu techniki

opracowanej na potrzeby wcześniejszych badań przez amerykańskiego urologa i badacza medycznego. Osobom tym zaoferowano możliwość oglądania filmów, aby pomóc przygotować się do pomiaru.

Każdy z tych zespołów zebrał dane w najlepszy sposób, jaki mógł zidentyfikować na potrzeby swoich konkretnych potrzeb. Możesz zauważyć, że w zestawach danych, które każdy z nich zgromadził, występują niedoskonałości i ograniczenia. Jednak fakt, że każdy zespół jasno zdefiniował i udokumentował swoje praktyki pomiarowe oraz podjął znaczące kroki w celu zapewnienia spójności pomiarów, oznacza, że jakość i użyteczność tych zestawów danych jest lepsza niż wielu, które można napotkać w środowiskach biznesowych. Inne zespoły na całym świecie również gromadzą podobne dane. I znaleźli jeszcze więcej sposobów podejścia do problemu. Niektórzy zatrudniają profesjonalistów do pomiaru, ale nie są profesjonalistami medycznymi. Podczas gdy badacze w Turcji oferowali filmy wideo, niektórzy z ich odpowiedników w Stanach Zjednoczonych stosują środki farmaceutyczne. Jeśli istnieje wiele sposobów uzyskiwania takich danych, dostępnych jest również wiele sposobów uzyskiwania danych spełniających Twoje potrzeby! Myśl kreatywnie, szukaj pomysłów u innych i pogódź się z niedoskonałością. Eksploracja danych polega na uzyskiwaniu przydatnych informacji już teraz, a nie doskonałych informacji w przyszłości.

Wyszukanie publicznych źródeł danych

Jeśli potrzebujesz danych, których jeszcze nie posiadasz, najpierw poszukaj źródeł publicznych. Źródła te są nie tylko liczne i różnorodne, ale w wielu przypadkach żaden podmiot komercyjny nie byłby w stanie samodzielnie zebrać tych samych informacji. A dane publiczne są zwykle dostępne bezpłatnie lub po niskich kosztach.

Patrząc na ukształtowanie terenu

Dane publiczne to przede wszystkim dane rządowe. Agencje rządowe zbierają i udostępniają dane o ludziach, działaniach i zasobach. W Stanach Zjednoczonych masz prawo żądać informacji od dowolnej części rządu federalnego (chociaż nie wszystkie części są równie responsywne). Każdy stan, hrabstwo i miasto gromadzi i przechowuje dane, które mogą być dla Ciebie dostępne. Kraje na całym świecie mają własne agencje statystyczne, podobnie jak wiele międzyrządowych organizacji. Dane publiczne są produktem ubocznym codziennej pracy rządu; żaden rząd nie gromadzi danych w celu udostępniania ich eksploratorom danych. Uzyskanie odpowiednich danych rządowych w formie, z której możesz skorzystać, nie zawsze jest łatwe. Możesz dobrze wyczuć, czego się spodziewać, jeśli spróbujesz zrozumieć, dlaczego i w jaki sposób rządy zbierają i udostępniają dane:

✓ Dlaczego rządy zbierają dane: Każda firma prowadzi rejestry swojej działalności, takie jak kontrakty, zakupy i sprzedaż; płatności na rzecz pracowników i dostawców; i interakcje z klientami. Te rejestry są potrzebne do wspierania codziennej działalności, ponieważ to, co robimy dzisiaj, zależy od tego, co zrobiliśmy i co uzgodniliśmy wczoraj, a prowadzenie rejestrów gwarantuje, że będziemy mieć jasne informacje o przeszłych działaniach, kiedy ich potrzebujemy. Rządy mają te same potrzeby, więc rządy również prowadzą dokumentację swoich działań. W rzeczywistości prowadzenie rejestrów może być nawet ważniejsze w rządzie niż w przemyśle, ponieważ każdy wyborca jest zainteresowany tym, co rząd robi (lub czego nie robi). Ten typ danych jest również znany jako dane transakcyjne i większość danych rządowych jest tego typu. Eksperci danych (i innego rodzaju analitycy danych) są często bardziej zainteresowani danymi o ludziach: kim są, czym się zajmują i jak żyją. Dysponując tego rodzaju danymi, możesz odkryć wzorce zachowań i inne aspekty ludzkiego życia, które są istotne dla Twoich celów biznesowych. Rządy również gromadzą tego rodzaju dane, ponieważ muszą również wiedzieć o swoich wyborcach, ich życiu i potrzebach w zakresie usług rządowych. Być może pamiętasz, jak w przeszłości wypełniałeś formularz ankiety do spisu ludności. Twoje odpowiedzi stają się danymi rządowymi. Spis jest tylko jednym z wielu badań, które rząd wykorzystuje do pozyskiwania danych do analizy. Ten rodzaj danych jest również nazywany danymi statystycznymi. Celem tych ankiet i innych badań rządowych jest dostarczenie danych, które pracownicy rządowi potrafią analizować w celu dostarczenia informacji prawodawcom i innym urzędnikom. Dane statystyczne stanowią stosunkowo niewielką część wszystkich danych rządowych, ale są ważne.

✓ Dlaczego rządy udostępniają dane: Rządy czasami udostępniają dane, aby wykonać zadanie. Jeśli istnieje inicjatywa zachęcająca do ćwiczeń, pomocne może być omówienie danych z ankiet dotyczących aktualnych wzorców ćwiczeń. Dzielą się również danymi, aby przekonać. Jeśli potrzebujesz finansowania na budowę mostu, ta sprawa jest bardziej przekonująca, jeśli udostępnisz dane wskazujące na potrzebę budowy mostu, konkurencyjne koszty budowy i tak dalej. I wreszcie, rządy dzielą się danymi, ponieważ muszą. Oczekują tego wyborcy, a wymaga tego prawo. Nie wszystkie dane rządowe są dostępne publicznie. Niektóre informacje są chronione ze względów bezpieczeństwa, a niektóre są utrzymywane w tajemnicy, aby chronić prywatność obywateli. Kiedy odpowiadasz na spis, Twoje indywidualne odpowiedzi nie są udostępniane, ale zbiorcze informacje o grupach osób stają się publiczne. (Tak więc dochód jednostki jest prywatny, ale średni dochód społeczności jest publiczny).

Eksploracja publicznych źródeł danych

Zasoby danych publicznych są ogromne, ale zostały dostosowane do konkretnych celów rządowych, a nie Twoich potrzeb. Gdy szukasz potrzebnych danych w publicznych źródłach, musisz być przygotowany. Zanim rozpoczniesz wyszukiwanie, upewnij się, że najpierw

✓ Określasz swoje potrzeby: Musisz dokładnie wiedzieć, jakich danych potrzebujesz iw jakiej formie. Musisz być przygotowany do wyjaśnienia tego pisemnie lub ustnie w trakcie poszukiwań.

✓ Znajdujesz właściwe źródło: zapoznasz się z agencjami rządowymi i dowiesz się, które z nich są prawdopodobnymi źródłami potrzebnych danych.

✓ Dowiedziałeś się, jak uzyskać dane: Czasami uzyskanie potrzebnych danych będzie proste; po prostu go pobierzesz lub uzyskasz raport. Ale jeśli potrzebne dane nie są już dystrybuowane tak prostymi kanałami, ich uzyskanie może wymagać złożonego i powolnego procesu.

Setki rządów i organizacji quasi-rządowych na całym świecie gromadzą, analizują i udostępniają dane. Mogą być udostępniane jako dane do odczytu maszynowego w plikach lub za pośrednictwem interfejsu programowania aplikacji (API); można go znaleźć w pisemnych raportach, wraz z wyrafinowaną analizą; i może nawet pojawić się w małych fragmentach znalezionych w wiadomościach. Nie myśl o danych tylko jako o polach i przypadkach, z którymi możesz pracować w swoim oprogramowaniu do analizy danych. Dowolna forma danych, surowych lub przeanalizowanych, może być przydatna, aby pomóc Ci lepiej zrozumieć problemy biznesowe, z którymi się borykasz. Może się nawet okazać, że jakaś agencja lub organizacja zbadała już pewne problemy, które Cię dotyczą, zebrała dane, przeanalizowała je i udostępniła raport. Jeśli tak, skorzystaj z tego, co już zostało zrobione, odkryj i przejdź do następnego etapu.

Dane czy statystyki?

W codziennych sytuacjach mówię o pozyskiwaniu i wykorzystywaniu danych. Ale, ściśle rzecz biorąc, nie zawsze jest to właściwe określenie. Często używam statystyk. W większości przypadków możesz użyć dowolnego terminu i nie ma to znaczenia, ale czasami będziesz musiał znać różnicę. Być może mierzysz wzrost każdego dziecka w szkole. Poszczególne pomiary są danymi. A ponieważ są właśnie tym, co zmierzyłeś i nie zostały w żaden sposób przetworzone, pomiary te można również nazwać surowe dane. Możesz obliczyć średni wzrost wszystkich uczniów w klasie lub wszystkich uczniów w szkole; te średnie (wartości obliczone) są statystykami. Ale ludzie nadal nazywają je danymi. I przez większość czasu nie spowoduje to żadnego zamieszania. Ale może przeglądasz witrynę agencji rządowej. Możesz zobaczyć, że agencja udostępnia statystyki. Ale wydaje się, że się nie dzielą danymi. Możesz też wdać się w rozmowę z pracownikiem agencji i spierać się o to, czy agencja udostępnia dane. Personel agencji może być znacznie bardziej rygorystyczny niż ty w doborze słów. Pamiętaj więc, że właściwym określeniem tego, czego chcesz, mogą być statystyki, a nie dane.

Rządy na całym świecie

Stany Zjednoczone są tylko jednym z wielu rządów, które udostępniają dane opinii publicznej. Chociaż nie znajdziesz dokładnie tego samego zakresu lub typów danych z każdego kraju, okaże się, że większość krajów ma pewne dane do udostępnienia. Ta sekcja obejmuje również niektóre organizacje międzyrządowe i non-profit, które oferują międzynarodowe zasoby danych.

OFFSTATS

Baza danych OFFSTATS Uniwersytetu w Auckland to portal do źródeł agencji statystycznych na całym świecie, podobnie jak międzynarodowa wersja portalu FedStats w Stanach Zjednoczonych. Zawiera

linki uporządkowane według kraju, regionu i tematu. (Międzynarodowe agencje prowadzą interesy w lokalnych językach, a wiele z nich nie ma wersji anglojęzycznych).

Organizacja Współpracy Gospodarczej i Rozwoju

Organizacja Współpracy Gospodarczej i Rozwoju (OECD) (dostępna za pośrednictwem portalu statystycznego pod adresem www.oecd.org/statistics) ma na celu promowanie polityk mających na celu poprawę dobrobytu ludzi na świecie. OECD mierzy produktywność, handel światowy i inwestycje. Analizuje dane dotyczące handlu i życia codziennego. OECD oferuje również zasoby przeznaczone do użytku przez statystyków na stronie www.oecd.org/statistics/statisticalresources.htm. Są one również cenne dla eksploratorów danych.

Organizacja Narodów Zjednoczonych

Organizacja Narodów Zjednoczonych (ONZ) to najbardziej wpływowa organizacja międzyrządowa na świecie.

Unia Europejska

Unia Europejska, do której należy większość krajów Europy Zachodniej, posiada portal do swoich źródeł statystycznych.

Open Data Institute promuje udostępnianie i wykorzystywanie otwartych danych na całym świecie. Jest to kluczowe źródło wiadomości na temat otwartych danych, a także centrum badań i edukacji.

Kupowanie danych

Nawet jeśli masz dostęp do ogromnych ilości danych w wewnętrznych bazach danych i bezcennych zasobów danych publicznych, czasami nadal warto kupować dane od komercyjnych dostawców. Gdy potrzebne dane są niedostępne lub gdy są dostępne, ale nie nadają się do tabaki, zwracanie się do prywatnych źródeł danych może mieć wiele sensu. Oto niektóre z zalet danych handlowych:

✓ Dostępność: Niektóre dane są dostępne wyłącznie ze źródeł prywatnych: dane generowane w ramach działalności gospodarczej lub innej działalności pozarządowej, na przykład takie jak historie kredytowe konsumentów lub firm lub listy osób o określonych powiązaniach lub zainteresowaniach.

✓ Przygotowanie: Zbieranie danych z różnych źródeł publicznych, organizowanie ich w spójne formaty i sprawowanie odpowiedniej kontroli zarządczej nad tym procesem może być czasochłonne i kosztowne. Jeśli na przykład potrzebujesz danych o transakcjach dotyczących nieruchomości ze społeczności w całych Stanach Zjednoczonych, możesz je uzyskać, żądając rekordów transakcji bezpośrednio od tysięcy samorządów w całym kraju i samodzielnie scalając wszystkie wyniki, ale korzystając z usługi nieruchomości usługa danych z pewnością byłaby znacznie łatwiejsza - i prawdopodobnie też tańsza.

✓ Ulepszenia: Niektóre prywatne źródła danych oferują ulepszenia, które mogą zwiększyć wartość zasobu danych. Ratingi kredytowe dla konsumentów i przedsiębiorstw mogą być najbardziej znanym rodzajem wzbogacania danych. Inne obejmują ocenę sympatii do określonych przyczyn politycznych lub prawdopodobieństwa dokonania zakupu oraz identyfikację języka lub przedmiotu danych tekstowych.

Żadne dane nie są takie jak Twoje własne dane. Dane generowane w ramach Twojej firmy są dla Ciebie wyjątkowo istotne, a Ty masz najwyższą moc, aby zapewnić, że są one odpowiednio zarządzane. Uzyskaj większą wartość ze swoich danych, korzystając selektywnie ze źródeł zewnętrznych w celu uzyskania informacji, które uzupełniają lub wyjaśniają posiadane dane.

Przeglądanie danych konsumenckich

Aby przedstawić rodzaje informacji konsumenckich dostępnych u dostawców komercyjnych, przyjrzę się szczegółowemu przykładowi. Tabela zawiera wszystkie dane zebrane na temat jednego konsumenta przez Axciom, głównego dostawcę danych marketingowych dla konsumentów. Ten sprzedawca dostarcza dane marketingowe dotyczące indywidualnych konsumentów i gospodarstw domowych, w których ci konsumenci mieszkają, w następujący sposób:

✓ Indywidualni konsumenci: Dla każdej osoby sprzedawca dzieli informacje na dwie kategorie danych:

• Charakterystyka: Dane demograficzne, takie jak wiek, stan cywilny, poziom wykształcenia i to, czy konsument ma dzieci. Można tu również zamieścić dane o członkach gospodarstwa domowego, którzy dzielą nazwisko konsumenta.

• Dom: Informacje o miejscu zamieszkania konsumenta, czy jest to mieszkanie jednorodzinne czy wielorodzinne, czy konsument wynajmuje lub posiada oraz długość pobytu.

✓ Gospodarstwa domowe: Sprzedawca śledzi cztery kategorie danych gospodarstw domowych:

• Pojazd: Szczegóły dotyczące własności i ubezpieczenia samochodu, w tym liczba pojazdów, marki i modele oraz daty odnowienia ubezpieczenia.

• Ekonomiczne: Informacje o działalności finansowej gospodarstwa domowego. Szacunkowy dochód, preferowane metody wydawania i aktywność wydatkowa różnymi kanałami.

• Zakupy: Informacje o zwyczajach zakupowych gospodarstwa domowego, online i offline. Może zawierać informacje o typach najczęściej kupowanych produktów – kategoriach, kwotach i częstotliwościach.

• Zainteresowania: Hobby i inne zainteresowania, takie jak gotowanie, sport i majsterkowanie.

Characteristic data	
Date of Birth	01/23/1945
Gender	Female
Education	Completed Graduate School
Marital Status	Single
Small or Home Business	True
Home	
Home Information	No Data Found
Vehicle	
Auto Policy Renewal	October
Economic	
Estimated Household Income Range	$75,000–$99,999
Presence of Credit Card	Credit Card Holder – Unknown Type
Credit Card Use – American Express	Regular
Credit Card Use – Discover	Regular
Online Purchasing Activity	True
Number of Purchases – Cash	2
Number of Purchases – Credit Card	1
Number of Purchases – AMEX	20
Number of Purchases – Discover	1
Number of Purchases – Visa	1
Number of Purchases – Other	11
Purchases	
Mail Order Responder	Mail Order Responder
Mail Order Buyer	Mail Order Buyer
Gardening Products	Purchased
General Merchandise	Purchased
Total Dollars Spent	1502
Total Number of Purchases	9
Average Dollars Spent Per Offline Purchase	157
Total Offline Dollars Spent	1394

Purchases	
Total Number of Offline Purchases	31
Total Offline Purchases – Under $50	25
Total Offline Purchases – $50–$99.99	6
Total Offline Purchases – $250–$499.99	1
Average Dollars Spent Per Online Purchase	101
Total Online Dollars Spent	304
Total Number of Online Purchases	3
Total Online Purchases – Under $50	3
Total Online Purchases – $50–$99.99	1
Total Online Purchases – $100–$249.99	2
Interests	
Interests	Fashion, Children's Items, Cooking, Gourmet Cooking, Health/Medical, Current Affairs/Politics, Crafts, Home Furnishings/Decorating, Home Improvement, Gardening, Other Pet Ownership, Reading, Reading Magazines, Aerobics

To tylko jeden przykład danych marketingowych dostępnych w sprzedaży. (Kilka pól zostało nieznacznie zmienionych ze względu na ochronę prywatności; w przeciwnym razie przykład zawiera wszystkie pełne dane pobrane z Axciom.) Inny przykład - nawet ten od tego samego dostawcy - może wyglądać inaczej, z różnymi polami, dodatkowymi informacjami o rodzinie lub dokładniejszymi (lub niedokładne) wyniki. A nawet pełne zapisy mogą się nie liczyć, jeśli konsument zrezygnował z udostępniania danych. Możesz zapoznać się z niektórymi udostępnianymi o Tobie danymi. Dane przedstawione w tym przykładzie zostały dostarczone przez firmę Axciom, głównego dostawcę danych marketingowych dla konsumentów. Za pośrednictwem strony internetowej About the Data (https://aboutthedata.com) Axciom umożliwia konsumentom przeglądanie własnych danych, uzyskiwanie informacji o sposobie gromadzenia i wykorzystywania danych, edytowanie danych lub rezygnację z udostępniania danych. Poświęć chwilę na przemyślenie źródeł informacji dostępnych dla dostawcy danych. Dostawca musi zebrać swoje profile marketingowe konsumentów z publicznych lub legalnie udostępnianych prywatnych źródeł. Wiele źródeł danych - osobiste wypłaty, dokumenty bankowe, deklaracje podatkowe i wiele innych - jest niedostępnych. Przykład w tabeli został skompilowany z trzech rodzajów źródeł:

✓ Źródła publiczne: Należą do nich

• Informacje rządowe, takie jak akta majątkowe i akta rzeczoznawcy oraz zapisy licencyjne

• Źródła publicznie dostępne, takie jak książki telefoniczne i internetowe

✓ Badania ankietowe: Ankiety i kwestionariusze, które konsumenci zdecydowali się wypełnić. Chociaż dane tutaj są ograniczone do tych konsumentów, którzy uczestniczyli, czasami są wykorzystywane do oszacowania danych dla innych.

✓ Komercyjne dane opt-in: Informacje zebrane przez źródła komercyjne, które uzyskały zgodę (zgodę) od konsumenta na wykorzystanie danych.

Na rynku wszelkie dane dotyczące osób mogą być określane jako konsumenckie dane marketingowe, niezależnie od tego, czy interesuje Cię aspekt „konsumencki" tych danych, czy też zamierzasz wykorzystać te dane do celów marketingowych. Pamiętaj jednak, w jaki sposób zamierzasz

wykorzystać dane i upewnij się, że umowy z dostawcą danych są zgodne z zamierzonym wykorzystaniem. Źródła te mogą mieć wiele niedoskonałości. Dane mogą być nieaktualne. Możesz znaleźć błędy lub niekompletne dane. Osoby mogą nie być odpowiednio dopasowane do innych członków gospodarstwa domowego. Tak jak musisz ocenić jakość i przydatność swoich wewnętrznych źródeł danych do danego zastosowania, musisz również dokładnie ocenić komercyjne źródła danych. Ale chociaż możesz podjąć działania w celu poprawy jakości wewnętrznego źródła danych, prawdopodobnie nie będziesz mieć takiej opcji w przypadku źródła komercyjnego. Jeśli jakość danych lub dokumentacji jest bardzo słaba, nie marnuj pieniędzy na ich zakup. Poszukaj alternatywnych dostawców, rozważ zebranie własnych danych lub po prostu żyj bez nich.

Poza danymi konsumenckimi

Nie wszystkie dane, których możesz potrzebować, dotyczą ludzi. Być może bardziej interesują Cię firmy lub organizacje non-profit. Może interesują Cię burze, ananasy lub mosty. Nie ma problemu. Źródła komercyjne mogą dostarczyć danych dla wszystkich tych rzeczy i wielu innych. Jeśli dostępne są dane, które cenisz na tyle, aby za nie zapłacić, prawdopodobnie ktoś jest gotowy do sprzedaży. Dotyczy to danych dotyczących osób i organizacji, a także danych dotyczących niezliczonych innych rzeczy. Niektóre powszechnie używane kategorie obejmują dane dotyczące

✓ Geografia i lokalizacje

✓ Zasoby i produkty

✓ Pogoda i klimat

Wiele rodzajów danych, których używamy do zrozumienia ludzi i ich zachowań, ma odpowiedniki dla firm i innych organizacji. Podstawowe opisowe fakty dotyczące ludzi, takie jak wiek, płeć i dochód, nazywane są danymi demograficznymi. Podobne informacje o organizacjach nazywamy firmografią. Informacje o statusie finansowym organizacji, transakcje finansowe lub powiązania z ludźmi i rzeczami są również często dostępne za pośrednictwem źródeł komercyjnych. Dla organizacji często dostępne są bogatsze dane niż dla pojedynczych osób. Organizacje, zwłaszcza korporacje publiczne i organizacje non-profit, są często zobowiązane do upubliczniania informacji o swoich finansach i działalności, a nawet wiele prywatnych firm decyduje się na udostępnianie niektórych informacji. Ponadto dostawcy danych mogą mieć większą swobodę w zakresie badania i udostępniania informacji biznesowych niż informacji o ludziach, a nawet ci, którzy oferują konsumentom opcje rezygnacji, mogą tego nie robić w przypadku organizacji.

Desperacko szukając źródeł

Możesz dowiedzieć się więcej o wykorzystywaniu dostępnych komercyjnie danych do marketingu biznesowego i konsumenckiego, kontaktując się z marketerami i badaczami rynku, którzy podzielają Twoje zainteresowania, a także z dostawcami danych. Te stowarzyszenia zawodowe są dobrym punktem wyjścia do nawiązywania kontaktów:

✓ Amerykańskie Stowarzyszenie Marketingu

✓ Stowarzyszenie Marketingu Bezpośredniego

✓ Fundacja Badań nad Reklamą

Lista głównych dostawców danych i rodzaje dostarczanych przez nich danych. Chociaż ta lista stanowi tylko niewielką część spośród setek dostawców danych działających na dzisiejszym rynku, nawet ci nieliczni oferują szeroką gamę ofert, obejmującą miliony osób. Źródła danych dotyczących konkretnych

rzeczy nie zawsze są oczywiste, ale zazwyczaj można znaleźć oczywiste miejsce, w którym można zacząć dociekać. Jeśli potrzebujesz danych o produkcie – czy jest to surowiec, plon, towar, czy gotowy produkt markowy – możesz znaleźć stowarzyszenie branżowe dla osób zaangażowanych w wytwarzanie i sprzedaż tego produktu. Niektóre stowarzyszenia branżowe zlecają lub przeprowadzają własne badania rynkowe i bezpośrednio sprzedają raporty i dane. Nawet jeśli ten, którego potrzebujesz, nie, jego personel lub członkowie nadal będą dobrymi kontaktami, aby zapytać o źródła danych. Dostawcy usług specjalistycznych oferują szereg innych typów danych. Niektóre kategorie są dość łatwe do znalezienia za pomocą prostego wyszukiwania w Internecie i mogą być zdominowane przez jednego lub tylko kilku wpływowych dostawców. Na przykład dane mapowe i geograficzne są zdominowane przez firmy takie jak ESRI i MapInfo, a dane pogodowe przez The Weather Company. Znalezienie mniej popularnych typów danych może wymagać sporo wysiłku. Jeśli masz trudności ze znalezieniem dostawców poprzez wyszukiwanie lub skierowania, rozważ skorzystanie z dobrego, staromodnego źródła, którego możesz zaniedbywać: z bibliotekarzem referencyjnym, zwłaszcza jeśli możesz znaleźć takiego, który regularnie zajmuje się zapytaniami biznesowymi.

Ocena jakości i przydatności

Mała próbka danych pokazana wcześniej ma kilka oczywistych wad. Nie pobrano żadnych danych o domu konsumenta. Musi gdzieś mieszkać! Widzisz miesiąc odnowienia ubezpieczenia, ale nic o samym samochodzie. W rzeczywistości wiele elementów zawartych w danych nie odzwierciedla realistycznego obrazu tego konsumenta. Jeśli zastosowałeś się do tej wskazówki i uzyskałeś własne dane marketingowe konsumentów za pośrednictwem Aboutthedata.com, prawdopodobnie zauważyłeś pewne niespodzianki w swoich własnych danych. Upewnij się, że dokładnie rozumiesz, co sprzedawca oferuje do sprzedaży. Będziesz chciał wiedzieć

✓ Co oznaczają dane: Co reprezentuje każde pole (zmienna) i wiersz (przypadek)? Jaka dokumentacja jest dostarczana? Idealna dokumentacja szczegółowo wyjaśnia źródło danych i wszelkie udoskonalenia wprowadzone przez dostawcę. W praktyce dokumentacja danych jest często skąpa.

✓ W jaki sposób oferowane są dane: Czy dane są oferowane na zasadzie abonamentu czy jednorazowo? Dostaniesz tylko wyciąg z zasobów dostawcy, czy elastyczny dostęp do źródła danych?

✓ Jakie zastosowania są dozwolone: Czy warunki świadczenia usług przez dostawcę zezwalają na zastosowanie, o którym myślisz? Czy będzie można zatrzymać dane lub wykorzystać je tylko raz?

✓ Sposób dostarczania danych: Czy otrzymasz wyciąg danych w postaci pliku do pobrania lub w innej formie (na przykład na płycie CD-ROM, DVD, a nawet na papierze), użyj samoobsługowego interfejsu API lub innego interfejsu ? Czy dostępna jest pomoc techniczna, jeśli masz trudności z używaniem tych interfejsów?

✓ Jaka jest struktura danych: Czy poziom danych (surowych lub zagregowanych) jest odpowiedni do analizy, o której myślisz? Czy format przechowywania (baza danych lub format pliku) jest dla Ciebie wygodny w użyciu? Jeśli nie, czy dostawca dokona konwersji dla Ciebie? Czy tabele bazy danych są zorganizowane intuicyjnie zgodnie z Twoimi potrzebami? Jeśli nie, czy dostawca może zapewnić interfejs, który uprości zapytania o dane? (Spróbuj zminimalizować restrukturyzację danych, którą będziesz musiał zrobić później).

Przed dokonaniem znaczącej inwestycji w jakiekolwiek źródło danych postaraj się ocenić jakość danych i zweryfikuj, czy nadają się one do zamierzonego celu. Zadawać pytania! Dowiedz się, jak pozyskiwane są dane. Jakie są źródła? Czy w gromadzenie i przygotowywanie danych zaangażowany jest przeszkolony personel badawczy? Czy widzisz, jak sformułowano i uporządkowano pytania ankietowe?

Upewnij się, że rozumiesz warunki użytkowania dostawcy i czy zezwalają one na użycie, którego potrzebujesz. W wielu przypadkach uzyskanie niewielkiej ilości danych jest dobrym pierwszym krokiem, aby można było zbadać i przetestować używane dane. Pamiętaj, aby sprawdzić dokumentację pod kątem danych, która czasami jest nieodpowiednia, a czasami nie istnieje.

Zapoznaj się ze swoimi danymi

Zanim francuska szefowa kuchni przygotuje olśniewające danie, przygotowuje wszystkie składniki i narzędzia. Sprawdza, czy składniki są świeże i dobre, a narzędzia działają prawidłowo. Nie zaczyna gotować, dopóki nie ułoży wszystkiego na swoim miejscu. Eksplorator danych nie jest inny. Zanim stworzysz olśniewający model predykcyjny, zapoznaj się z danymi, z których będziesz korzystać. Umieszczasz go tam, gdzie go potrzebujesz. Upewniasz się, że rozumiesz, jakie dane posiadasz, w jaki sposób są one uporządkowane i przechowywane oraz czy są kompletne i poprawne. Tu dowiesz się, jak analizować i oceniać swoje dane.

Organizowanie danych dla górnictwa

Eksploracja danych ma bardzo surowe wymagania dotyczące organizacji danych. Nie są to wymagania egzotyczne, złożone ani trudne do spełnienia, ale są surowe. Posłużę się przykładem, aby pokazać, jak dane muszą być zorganizowane w celu eksploracji danych. Rysunek poniżej przedstawia próbkę danych w postaci tabeli w oprogramowaniu do eksploracji danych. Każdy wiersz reprezentuje jedną działkę nieruchomości. Informacje o działkach są uporządkowane w kolumnach. Pierwsza kolumna zawiera numer identyfikacji podatkowej (TAXKEY), druga kolumna zawiera oszacowaną wartość gruntu z wcześniejszej wyceny (P_A_LAND) i tak dalej. Każdy wpis w jednym rzędzie dotyczy jednej konkretnej działki. Każdy wpis w jednej kolumnie to ten sam typ informacji. Żadne wiersze ani kolumny nie są puste z powodów związanych ze stylem i czytelnością. Dane te są odpowiednio zorganizowane do badania różnic pomiędzy działkami nieruchomości.

ExampleSet (162403 examples, 0 special attributes, 58 regular attributes)03 / 162,403 examples): all

Row No.	TAXKEY	P_A_LAND	NR_UNITS	C_A_LAND	LAND_USE	C_A_CLASS	C_A_TOTAL	CHK_DIGIT
1	10001000	48200	1	48200	8810	1	229600	3
2	10011000	146200	0	150700	5093	3	602800	8
3	10021000	115000	0	115000	1794	2	384000	2
4	10022000	0	0	0	8880	9	0	8
5	18100000	100	0	100	6	4	37000	7
6	18101000	100	0	100	6	4	37000	2
7	19989000	0	0	0	4010	9	0	X
8	19990000	0	0	0	4010	9	0	5
9	19991000	0	0	0	4010	9	0	0
10	19992100	40600	0	40600	4010	2	40600	2
11	19996100	53400	6	53400	8830	7	179600	4
12	19996210	0	0	0	8885	9	0	8
13	19998200	47800	1	47800	8810	1	153700	1
14	19999100	139700	0	139700	4225	2	268800	0
15	20032000	495700	0	495700	5171	4	15729000	X
16	20051000	204100	0	204100	5171	4	475000	3
17	20052000	114700	0	114700	4225	4	120000	9
18	20071100	0	0	1734900	5172	4	12638000	9

Jeśli zamiast nieruchomości zbadasz ludzi, każda osoba byłaby reprezentowana przez jeden wiersz w danych, a wszystkie szczegóły dotyczące osób byłyby zorganizowane w kolumny. Jeśli zbadasz zdjęcia rentgenowskie klatki piersiowej, każde zdjęcie rentgenowskie klatki piersiowej będzie reprezentowane w danych w jednym wierszu, a wszystkie szczegóły dotyczące zdjęć rentgenowskich klatki piersiowej zostaną zorganizowane w kolumny. W terminologii analizy danych rzeczy, które studiujesz – rzeczy w wierszach – nazywane są przypadkami lub zapisami. A szczegóły na ich temat, które znajdują się w kolumnach, nazywane są zmiennymi. Usłyszysz także kolumny zwane polami, zwłaszcza w kontekście

baz danych. Tak więc eksploracja danych wymaga danych zorganizowanych w jednym wierszu dla każdego przypadku i jednej kolumnie dla każdej zmiennej. Wiele źródeł danych jest już zorganizowanych w ten sposób. Statystycy organizują dane w ten sposób z przyzwyczajenia. Specjaliści od baz danych mogą nie używać tego podejścia w większości swojej pracy, ale zazwyczaj zrozumieją, czego chcesz, jeśli nazwiesz to płaską tabelą. Znajdziesz subtelne różnice w strukturze danych. Niektóre typy oprogramowania wykorzystują informacje opisowe w nagłówku przed danymi, takie jak niektóre specjalne formaty związane z aplikacjami do eksploracji danych Orange i Weka. Niektóre złożone procedury analityczne mają dodatkowe lub nieco zróżnicowane wymagania (są to dość nietypowe). Jednak rdzeń danych nadal zawiera obserwacje w wierszach i zmienne w kolumnach.

Pobieranie danych stamtąd do tego miejsca

Pierwszym praktycznym krokiem z danymi jest dostarczenie ich z dowolnego miejsca do miejsca, w którym ich potrzebujesz. Czynności, które podejmiesz, aby zaimportować dane do wykorzystania w eksploracji danych, mogą się znacznie różnić w zależności od sytuacji. Twoje własne umiejętności, styl pracy, zasady i procedury firmy oraz specyfika konkretnego projektu mogą mieć wpływ na sposób, w jaki uzyskujesz dostęp do danych. Do najważniejszych wpływów należą:

✓ Format danych: format danych. Przykłady obejmują relacyjną bazę danych, bazę danych NoSQL, plik tekstowy, arkusz kalkulacyjny, XML lub inne.

✓ Organizacja danych: Struktura Twoich danych. Struktura danych może być wygodna do eksploracji danych (i konkretnego projektu) lub nie.

✓ Oprogramowanie: Każdy produkt ma własne procedury importowania danych, a różnice istnieją nawet w obrębie pojedynczego produktu.

Pliki tekstowe

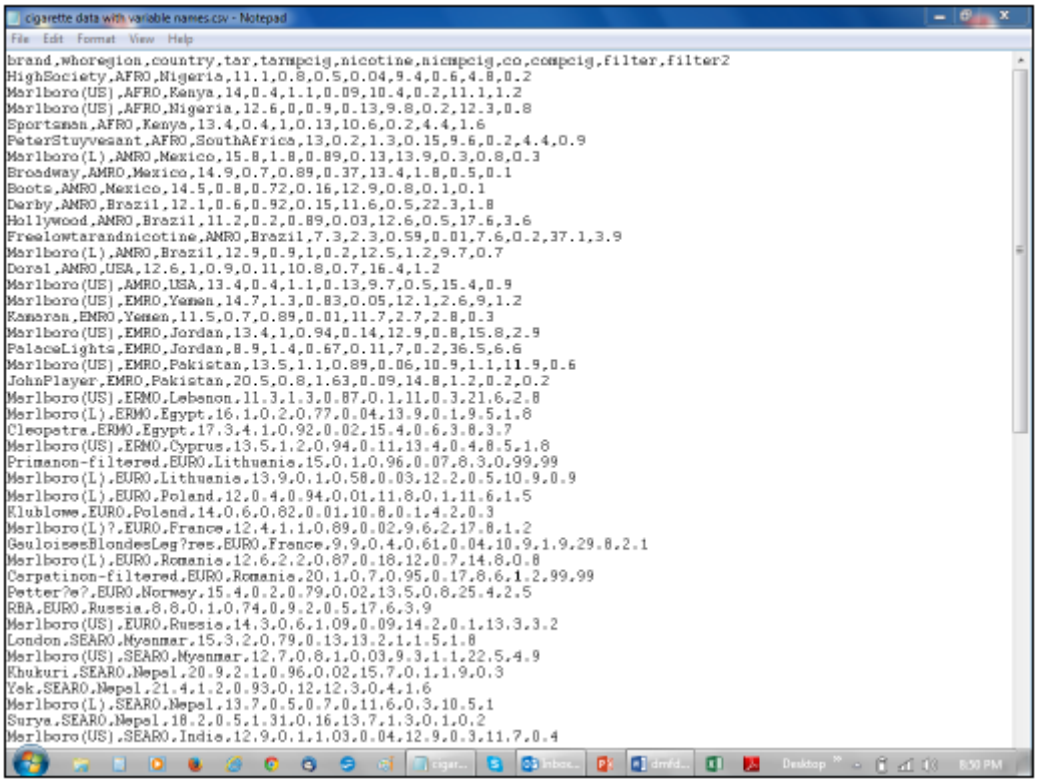
Formaty tekstowe są powszechne i prawdopodobnie często się z nimi spotykasz. Znajdziesz kilka odmian, ale niektóre z najczęstszych to wartości rozdzielane przecinkami (.csv), rozdzielane tabulatorami i tekst o stałej kolumnie. Większość publicznych źródeł danych, w tym źródła rządowe i agencje non-profit, oferuje dane w postaci plików tekstowych. Wielu badaczy uwielbia pliki tekstowe, ponieważ nie są powiązane z konkretnymi produktami lub platformami i są zwarte (czyli zajmują minimalną ilość miejsca na dane, które zawierają). Oto nowości dotyczące plików tekstowych:

✓ Dobra wiadomość: Każda aplikacja do eksploracji danych może importować dane z plików tekstowych.

✓ Zła wiadomość: Każda aplikacja do eksploracji danych ma swój własny sposób importowania danych z plików tekstowych, a niektóre z nich są dość trudne w użyciu.

✓ Jeszcze gorsza wiadomość: Niektóre aplikacje do eksploracji danych mogą importować niektóre rodzaje plików tekstowych, ale nie inne.

Rozważ przykład. Rysunek przedstawia dane w pliku tekstowym. Dane są w formacie wartości rozdzielanych przecinkami .csv. Pierwszy wiersz zawiera nazwy zmiennych oddzielone przecinkami. Wszystkie pozostałe wiersze zawierają dane, po jednym wierszu dla każdej marki papierosów. Dane obejmują nazwę marki, region, w którym jest sprzedawana, zawartość smoły i inne zmienne. Te wartości są oddzielone przecinkami. Te dane są dobrze zorganizowane do eksploracji danych. Jak wygląda proces otwierania tych danych?



Oto, jak to się robi w czterech przykładowych aplikacjach do eksploracji danych. Przejrzyj te procedury, a zaczniesz rozumieć, jak te aplikacje wyglądają i jak są używane. Aby otworzyć przykładowe dane w KNIME:

1. Uruchom KNIME.

2. Znajdź czytnik CSV w repozytorium węzłów (menu). Jest zgrupowany z innymi narzędziami do importowania danych.

3. Przeciągnij CSV Reader do obszaru roboczego.

4. Kliknij prawym przyciskiem myszy i wybierz Konfiguruj. Przeglądaj, aby znaleźć dane papierosów.

5. Dostosuj ustawienia. Upewnij się, że wybrałeś właściwe ograniczniki (przecinki) i zaznacz, że nagłówki kolumn (nazwy zmiennych) znajdują się w pierwszym wierszu danych

6. Kliknij przycisk Wykonaj (pokazany na rysunku 12-3), aby zaimportować dane. Czytnik CSV pokaże zielony wskaźnik, gdy dane zostaną zaimportowane.

Aby otworzyć przykładowe dane w Orange, wykonaj następujące kroki:

1. Uruchom Orange canvas.

2. Znajdź widżet Plik. Znajduje się w grupie Dane, jedynym narzędziu do importowania danych.

3. Kliknij widżet Plik jeden raz, aby umieścić go w obszarze roboczym.

4. Kliknij prawym przyciskiem myszy i wybierz Otwórz. Przeglądaj, aby znaleźć dane papierosów. Ups! Lista rozwijana typów plików nie oferuje opcji dla formatu .csv. Będziesz musiał przekonwertować dane na inny format, zanim będziesz mógł je otworzyć w tej aplikacji do eksploracji danych.

Aby otworzyć przykładowe dane w RapidMiner, wykonaj następujące kroki:

1. Uruchom RapidMiner Studio.

2. Znajdź operator Czytaj CSV. Jest zgrupowany z innymi narzędziami do importowania danych.

3. Przeciągnij operator Czytaj CSV do obszaru roboczego.

4. Kliknij operator Czytaj CSV. Ustawienia operatora Czytaj CSV zostaną wyświetlone w obszarze Parametry

5. W obszarze Parametry kliknij przycisk Kreator konfiguracji importu i użyj kreatora, aby wyszukać dane papierosów.

6. Dostosuj ustawienia. Kreator podaje wskazówki, które pomagają w prawidłowym ustawieniu ustawień. Kliknij przycisk Zakończ, aby powrócić do obszaru roboczego.

7. Kliknij przycisk Wykonaj, aby zaimportować dane. Operator odczytu CSV pokaże okrągły zielony wskaźnik, gdy dane zostaną zaimportowane.

Aby zaimportować przykładowe dane w Weka, wykonaj następujące kroki:

1. Uruchom Weka KnowledgeFlow.

2. Znajdź CSVLoader na pasku narzędzi Projekt. Jest zgrupowany z innymi narzędziami do importowania danych.

3. Kliknij CSVLoader, a następnie kliknij w obszarze roboczym, aby umieścić CSVLoader w obszarze roboczym.

4. Kliknij prawym przyciskiem myszy i wybierz Konfiguruj. Przeglądaj, aby znaleźć dane papierosów.

5. Dostosuj ustawienia.

6. Kliknij przycisk Uruchom proces (pokazany na rysunku 12-20), aby zaimportować dane. Obszar stanu aktualizuje się po zaimportowaniu danych.

Wygląd aplikacji, organizacja narzędzi i szczegóły konfiguracji różnią się, ale główne kroki są dość podobne. Dopóki Twoja aplikacja będzie mogła odczytać Twój format, wyniki będą takie same.

Bazy danych

Dane gromadzone przez duże organizacje w toku codziennej działalności zwykle przechowywane są w bazach danych. Jednak administratorzy baz danych mogą nie chcieć zezwalać eksploratorom danych na bezpośredni dostęp do tych źródeł danych, a bezpośredni dostęp może również nie być najlepszą opcją z Twojego punktu widzenia. Bezpośredni dostęp do operacyjnych (używanych do rutynowych operacji biznesowych) baz danych może być złym pomysłem, ponieważ

✓ Eksperci danych wykorzystują dużo danych. Możesz nieumyślnie zablokować zasoby i ingerować w zwykłe operacje biznesowe.

✓ Liczą się zobowiązania prawne i inne obowiązki biznesowe. Możesz nieumyślnie naruszyć przepisy dotyczące prywatności danych lub inne wymagania dotyczące zarządzania danymi, jeśli dostęp do danych nie będzie odpowiednio kontrolowany.

✓ Operacyjne bazy danych nie są zorganizowane do eksploracji danych. Możesz spędzić dużo czasu, próbując uzyskać potrzebne dane, a mimo to nie mieć pewności, czy zrobisz to dobrze.

Gdy potrzebujesz danych z operacyjnej bazy danych (i masz odpowiednią zgodę na wykorzystanie danych), powinieneś omówić swoje potrzeby z administratorem odpowiedzialnym za te dane. Musisz dokładnie wyjaśnić, jakich danych potrzebujesz, jakiego formatu potrzebujesz do eksploracji danych oraz czy potrzebujesz danych tylko raz, czy na bieżąco. Najlepszym rozwiązaniem w przypadku jednorazowych żądań jest często wyodrębnienie przez administratora danych za Ciebie i dostarczenie ich w pliku tekstowym lub innym akceptowalnym formacie. Inną sprawą jest stały dostęp do danych. Administrator może nie chcieć udostępniać wyciągów danych w kółko, a udzielanie bezpośredniego dostępu do systemów biznesowych jest ryzykowne. Powszechnym rozwiązaniem jest stworzenie analitycznej bazy danych. Jest to zwykła relacyjna baza danych, oddzielona od konwencjonalnych systemów biznesowych. Dane są rutynowo (i automatycznie) przesyłane z systemów biznesowych do analitycznej bazy danych, a eksploratorzy danych mogą uzyskać do nich dostęp w dowolnym momencie. Jeśli korzystasz z analitycznej bazy danych, upewnij się, że jest ona odpowiednio zorganizowana, aby wspierać eksplorację danych. Pomóż administratorowi bazy danych, szkicując diagram, taki jak pierwszy rysunek, aby pokazać, jak dane muszą być zorganizowane. Jeśli administrator bazy danych upiera się, że danych nie można przechowywać w ten sposób, zapytaj, czy można utworzyć widok (przechowywane zapytanie, które można odpytywać tak, jak gdyby była to konwencjonalna tabela danych) z organizacją, której potrzebujesz. Wiele produktów do eksploracji danych potrafi odczytywać dane z baz danych. Wymagane kroki różnią się w zależności od

✓ Projekt aplikacji do eksploracji danych

✓ Struktura źródłowej bazy danych

✓ Middleware, zwykle nazywane sterownikiem (sterownik ODBC, sterownik JDBC), specjalne oprogramowanie pośredniczące między bazą danych a oprogramowaniem aplikacyjnym

Dokumentacja aplikacji do eksploracji danych powinna zawierać informacje, czy może ona odczytywać dane z bazy danych, a jeśli tak, jakiego narzędzia lub funkcji użyć i w jaki sposób. Administrator, który konfiguruje bazę danych analitycznych, może podać szczegóły dotyczące dostępu do bazy danych. Jeśli już czujesz się komfortowo w pracy z bazami danych i innymi aplikacjami, nie znajdziesz nic zaskakującego w robieniu tego samego z aplikacją do eksploracji danych. Jeśli bazy danych są dla Ciebie nowe, poproś kompetentną osobę z Twojej organizacji, aby poprowadziła Cię przez proces za pomocą własnej bazy danych i aplikacji do eksploracji danych.

Arkusze kalkulacyjne, XML i specjalne formaty danych

Może być konieczne użycie danych znajdujących się w arkuszu kalkulacyjnym, XML (rozszerzalny język znaczników) lub dowolnym z dziesiątek mniej popularnych formatów. Kluczowe pytanie zawsze będzie brzmiało: Czy Twoja aplikacja do eksploracji danych importuje dane w tym formacie? Tak długo, jak Twoja aplikacja do eksploracji danych ma narzędzie do odczytywania formatu danych, którego potrzebujesz, proces będzie prosty — tylko niewielka odmiana przykładów, które możesz przeczytać w sekcji „Pliki tekstowe” we wcześniejszej części tego rozdziału. Może być konieczne wybranie innego narzędzia do importowania danych lub zmiana kilku ustawień, ale proces będzie bardzo podobny. Jeśli Twoja aplikacja do eksploracji danych nie może zaimportować danych w określonym formacie, wypróbuj te alternatywne metody:

✓ Sprawdź swoje źródło danych pod kątem innych formatów. Wiele źródeł oferuje wybór.

✓ Samodzielnie przekonwertuj format danych. Niektóre konwersje są łatwe, a inne trudne.

✓ Użyj innej aplikacji do eksploracji danych. Możliwość importowania danych jest ważnym czynnikiem przy wyborze oprogramowania do eksploracji danych, ale jeśli już jesteś zaangażowany w konkretny produkt, zmiana może być niepraktyczna.

Płynność w eksploracji danych

Zawód eksploracji danych ma swoje własne słownictwo. Tradycyjni analitycy danych nazywają coś, co chcesz przewidzieć, zmienną zależną, ale eksplorator danych może nazwać to celem lub wynikiem. Nazwa tradycyjnego analityka danych dla czegoś, co może mieć wpływ na zmienną zależną, to zmienna niezależna, ale eksplorator danych może preferować predyktor, dane wejściowe lub atrybut. Rodzaje używanych zmiennych wpływają na opcje manipulacji danymi i modelowania. Terminy te są używane zarówno przez tradycyjnych analityków danych, jak i eksploratorów danych: Typy zmiennych kategorialnych obejmują

✓ Nominalne: Nazwy lub kategorie bez kolejności (takie jak Mężczyzna i Kobieta).

✓ Porządkowe: klasyfikowane lub uporządkowane kategorie, takie jak oceny literowe lub gwiazdki w recenzji produktu. (Miary porządkowe nie są przeznaczone do stosowania w operacjach matematycznych, nawet jeśli są reprezentowane przez liczby. Jednak ludzie cały czas łamią tę zasadę. Czasami wyniki są przydatne. Często tak nie jest.)

Typy zmiennych ciągłych obejmują

✓ Interwał: Miary takie jak czas i temperatura Fahrenheita, które są odpowiednie do użycia w niektórych operacjach matematycznych, ale nie we wszystkich, ponieważ skale pomiarowe nie mają wyraźnej wartości zerowej. Na przykład (0 stopni Fahrenheita nie oznacza braku wszelkiego ciepła, ale jest to dość nieprzyjemne).

✓ Ratio: Miary, takie jak wagi, długości i temperatura Kelvina, które mogą być używane w operacjach matematycznych i które mają wyraźną wartość zero.

Zakres terminów używanych w oprogramowaniu do eksploracji danych jest duży i zróżnicowany, być może zbyt zróżnicowany. Na przykład wiele aplikacji do eksploracji danych wykorzystuje programowanie wizualne. Oznacza to, że funkcje są reprezentowane przez małe ikony, które można przenieść w puste miejsce na ekranie i połączyć ze sobą w celu zdefiniowania procesu eksploracji danych. My nazywam te ikony narzędziami, a niektóre produkty również używają tego terminu. Ale inni nazywają to samo węzłem, operatorem lub inną nazwą. W tej książce puste miejsce jest nazywane obszarem roboczym lub obszarem roboczym, ale w aplikacji do eksploracji danych może być nazywane inaczej, na przykład płótnem.

Badanie danych

Po zaimportowaniu zestawu danych do aplikacji do eksploracji danych następnym krokiem jest przeglądanie zmiennych jeden po drugim. W swojej recenzji zbadasz zmienne, aby upewnić się, że rozumiesz, co każda z nich reprezentuje, aby dowiedzieć się, czy dane są kompletne i aby ocenić jakość danych, które posiadasz. Przegląd pomaga określić, czy Twoje dane są adekwatne do realizacji celów eksploracji danych. Przegląd danych jest częścią fazy rozumienia danych w procesie CRISP-DM do eksploracji danych. Więcej informacji na temat tego procesu można znaleźć w Części 4, a o przykładowym przeglądzie danych można przeczytać w rozdziale 2. Potrzebne będą podsumowania dla każdej zmiennej, np.

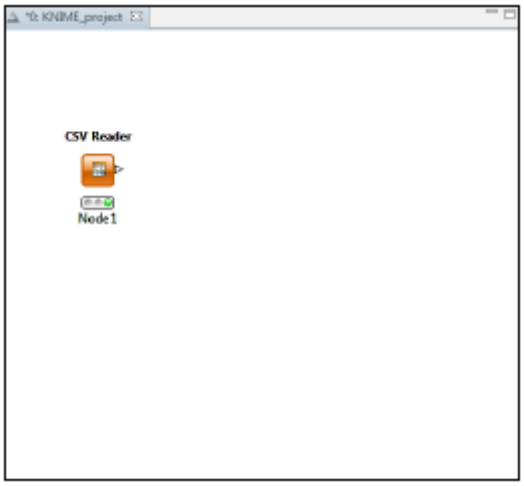
✓ Liczba brakujących spraw

✓ Wartości minimalne i maksymalne

✓ Średnie i odchylenia standardowe (miary zmienności)

✓ Wartości zmiennych kategorycznych

Niektóre platformy udostępniają podsumowania danych dla wielu zmiennych w jednym kroku. Inni będą wymagać wielu kroków, aby uzyskać te informacje. Jeden przykład podsumowania danych znajduje się w rozdziale 2. Oto inny, który pochodzi z importu danych w KNIME pokazanego wcześniej w tym rozdziale. Rysunek przedstawia proces tuż po zaimportowaniu danych.



Zajmowanie się szczegółami graficznymi

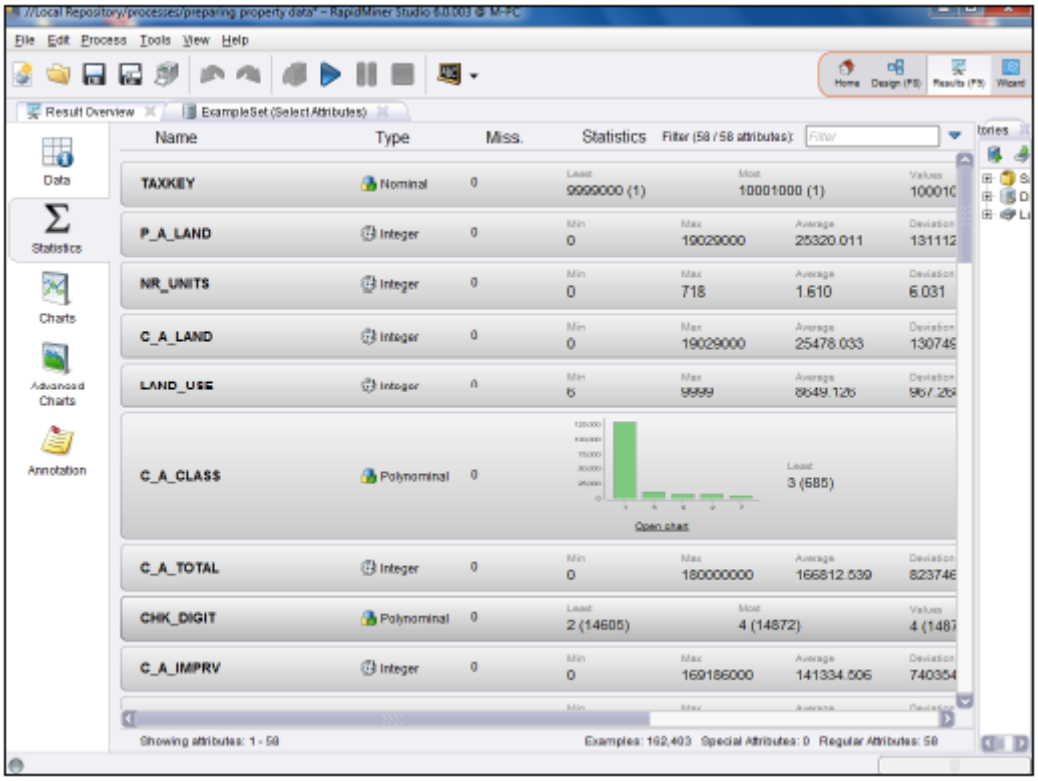
Być może jesteś nowy w eksploracji danych, ale znasz już kilka ważnych narzędzi handlu. Wykresy, takie jak wykresy słupkowe, histogramy i wykresy rozrzutu, są ważnymi narzędziami do eksploracji danych. Eksploratorzy danych używają tych konwencjonalnych wykresów w konwencjonalny i niekonwencjonalny sposób! A teraz, gdy jesteś eksploratorem danych, możesz poszerzyć swój repertuar o specjalne wykresy, które pomogą Ci upakować więcej informacji na stronie (bez utraty głównych pomysłów), wykryć typowe wzorce lub ocenić modele predykcyjne. Ta część zawiera wprowadzenie do arsenału wykresów i narzędzi do eksploracji danych. Przekonasz się, że wykresy są jednym z najłatwiejszych sposobów na rozpoczęcie eksploracji danych, zwłaszcza że eksploratorzy danych często korzystają z rodzajów wykresów (lub ich odmian), z których prawdopodobnie korzystałeś już gdzie indziej.

Rozpoczęcie proste

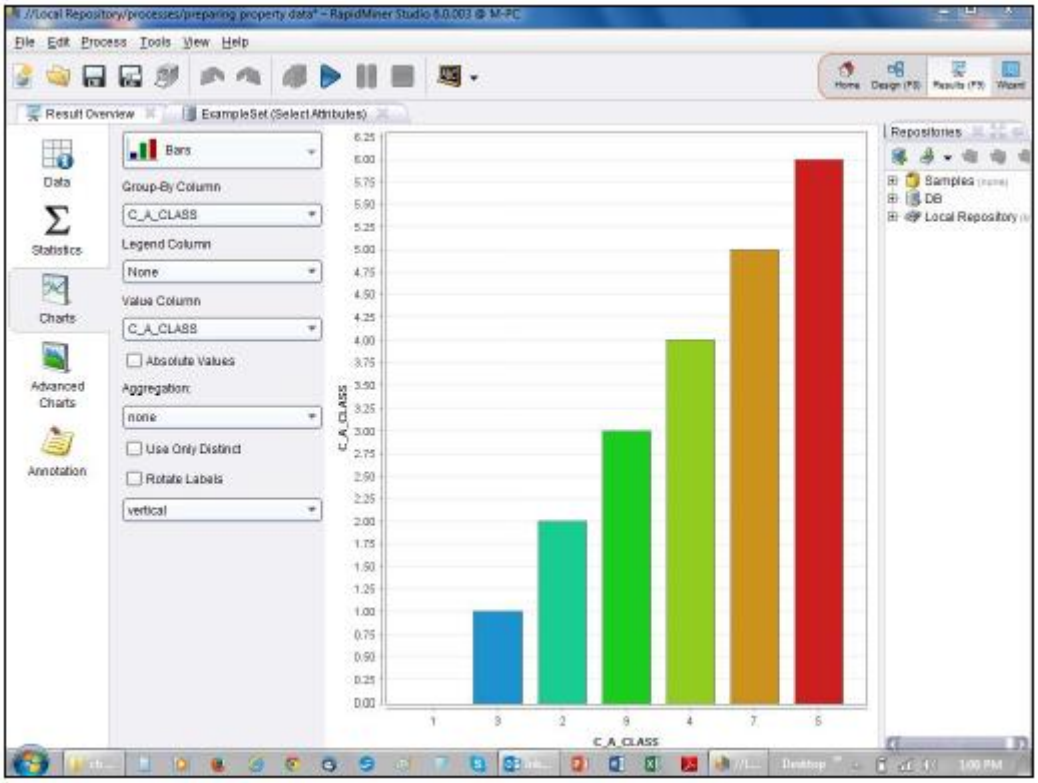
Wszyscy eksploratorzy danych używają wykresów, a wszystkie aplikacje do eksploracji danych oferują pewne możliwości graficzne. Niektóre aplikacje do eksploracji danych oferują tylko wykresy, które możesz zapamiętać z czasów szkoły podstawowej, takie jak wykresy słupkowe i wykresy rozrzutu. Dzieje się tak, ponieważ te proste wykresy są najczęściej używane przez eksploratorów danych.

Zmienne eyeballing z wykresami słupkowymi i histogramami

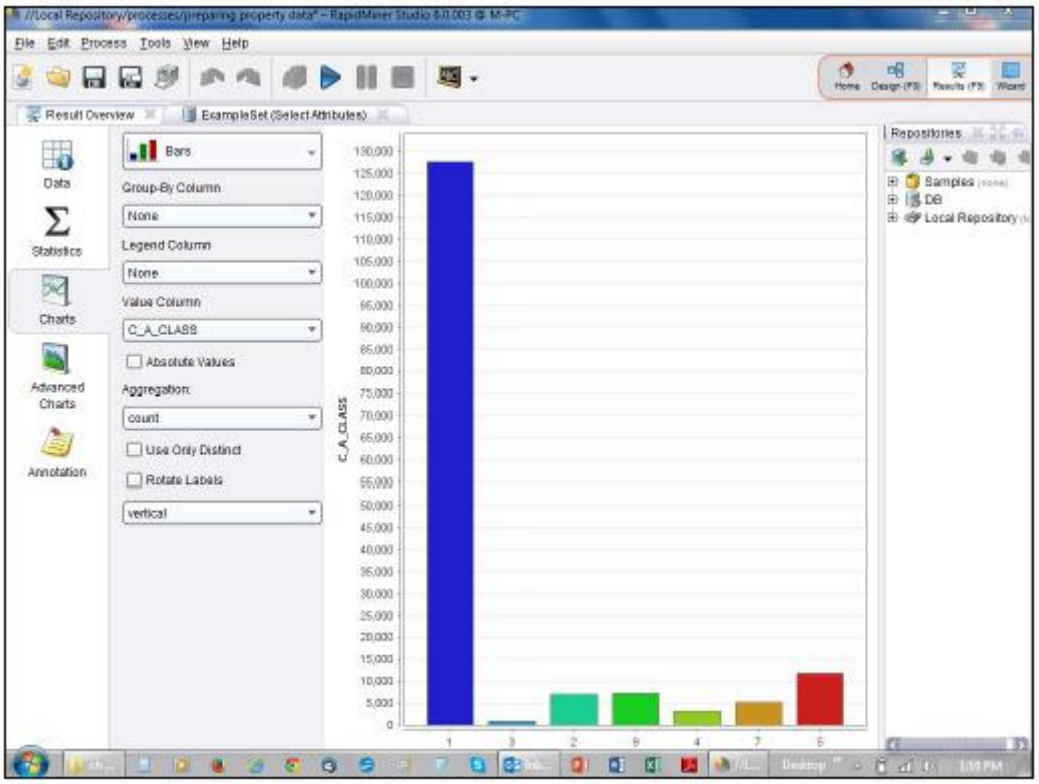
Podstawową częścią fazy zrozumienia danych w procesie eksploracji danych polega na badaniu zmiennych pojedynczo, przeglądaniu ich rozkładu i sprawdzaniu oczywistych problemów z jakością danych. Wykresy słupkowe i histogramy to wizualne podsumowania, które ułatwiają i przyspieszają zrozumienie rozkładów zmiennych. Te dwa typy wykresów są bardzo podobne. Jeśli zmienna jest jakościowa, użyj wykresu słupkowego; będzie miał jeden pasek dla każdej kategorii, a wysokość paska pokazuje częstotliwość każdej kategorii. Jeśli zmienna jest ciągła, użyj histogramu. Na histogramie każdy słupek reprezentuje zakres wartości zmiennej. Twoja aplikacja do eksploracji danych może bardzo ułatwić uzyskanie tych wykresów. Są one często uwzględniane w wynikach ogólnych narzędzi podsumowujących dane, takich jak przykład pokazany na rysunku.



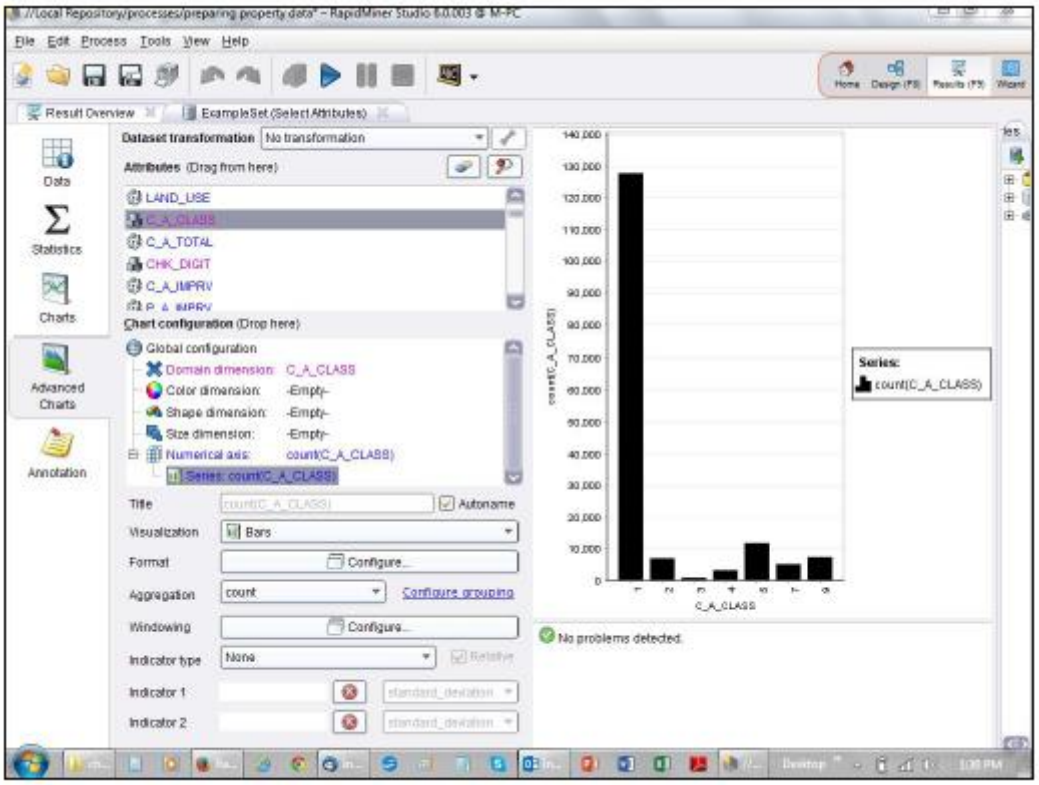
Ale uzyskanie odpowiedniego wykresu nie zawsze jest proste. Przyjrzyj się uważnie rysunkowi powyżej, a zobaczysz frazę Otwórz wykres pod wykresem słupkowym. Kliknięcie tego linku otwiera edytor wykresów. Spodziewałbyś się, że zobaczysz wykres, który jest identyczny z tym w podsumowaniu danych, który zostanie otwarty w edytorze, prawda? Rysunek pokazuje edytor wykresów, jak wygląda po otwarciu w ten sposób.



Nieidentyczny! Będziesz musiał zajmować się konfiguracją, aby wrócić do tego samego punktu.



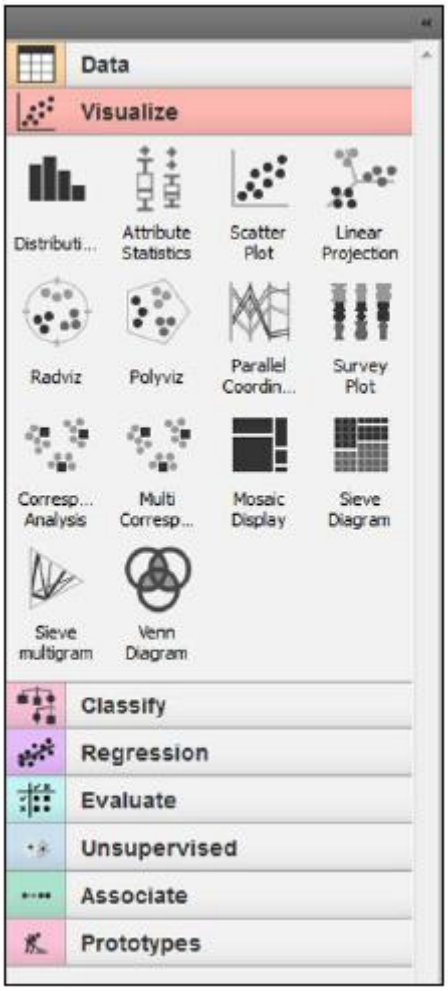
Ale ten edytor wykresów oferuje wartość na inne sposoby. Daje więcej opcji, takich jak tworzenie bardziej wyrafinowanej struktury wykresu lub kontrolowanie elementów kosmetycznych, takich jak kolor.



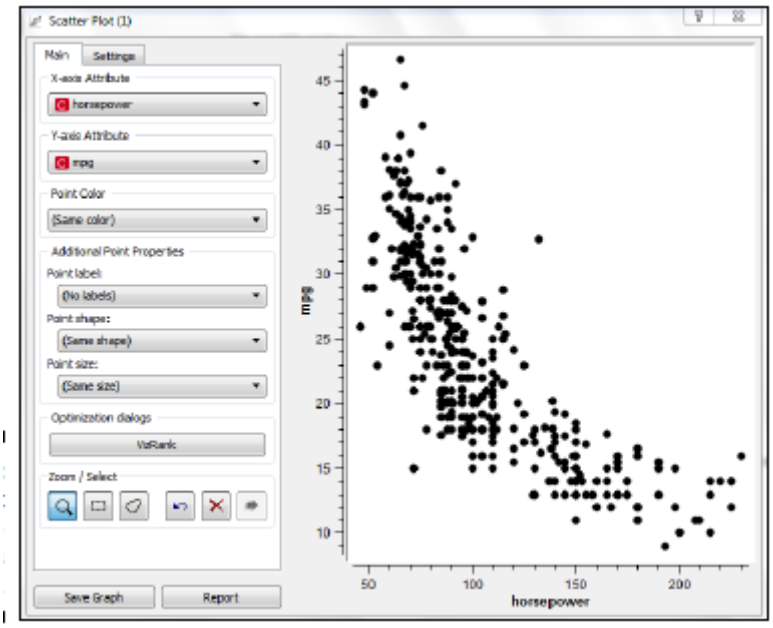
Edytory wykresów zapewniają również ścieżki do eksportowania wykresów do wykorzystania w raportach lub prezentacjach. Złożoność konfiguracji wykresu widoczna w tej sekcji jest kwestią projektowania produktu. Aplikacja do eksploracji danych może sprawić, że niektóre operacje będą bardzo łatwe, a inne niezwykle złożone lub niemożliwe. Żaden magiczny produkt nie przyćmiewa wszystkich innych pod względem łatwości użytkowania, ale jeden z nich może lepiej pasować do Twojego stylu pracy niż inne. Tak więc, zanim zdecydujesz się na produkt do użycia, przetestuj go dokładnie pod kątem rodzaju pracy, którą musisz wykonać.

Powiązanie jednej zmiennej z drugą za pomocą wykresów rozrzutu

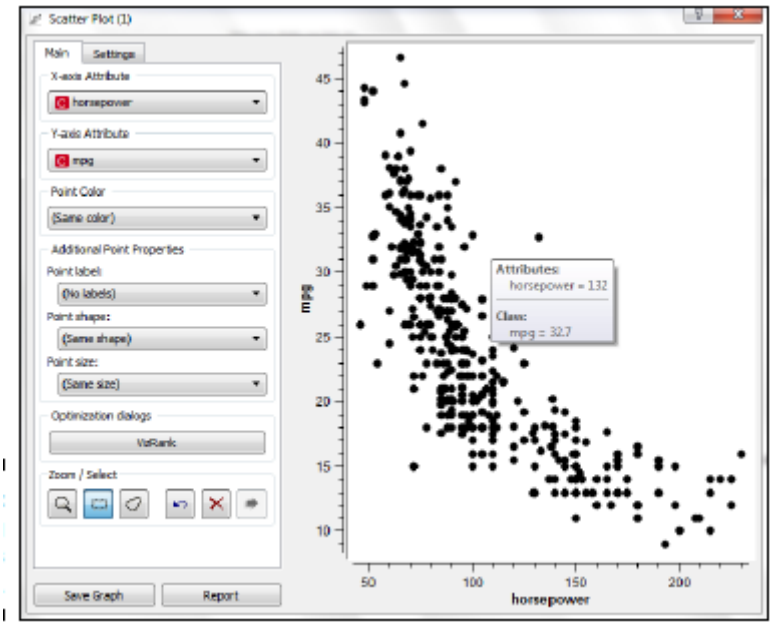
Pierwszym krokiem w kierunku modelowania predykcyjnego jest wzajemne powiązanie zmiennych. Prostym, niezwykłym narzędziem do tego jest wykres rozrzutu. Służy do powiązania jednej ciągłej miary z drugą. Eksperci danych czasami rozciągają reguły i używają ich również ze zmiennymi kategorialnymi. Oś pozioma (x) wykresu przedstawia wartości jednej zmiennej; oś pionowa (y) reprezentuje drugą zmienną. Możesz nie mieć pojęcia, która zmienna jest niezależna, a która zależna dla każdej pary zmiennych. Jeśli tak, zmienna niezależna powinna znajdować się na osi poziomej. Każdy punkt na wykresie reprezentuje współrzędne, parę wartości dla dwóch zmiennych w jednym przypadku. (Te pary są czasami nazywane parami xy). Znajdź swoje narzędzie do wykresu rozrzutu



i skonfiguruj podstawowe narzędzie do tworzenia wykresu rozrzutu, wybierając dwie zmienne, których chcesz użyć. Przykład na rysunku przedstawia interaktywny wyświetlacz; wykres rozrzutu pojawia się natychmiast.



W innym narzędziu do wykonania i utworzenia wykresu mogą być potrzebne dodatkowe kroki. Przykład wykresu rozrzutu na powyższym rysunku odnosi przebieg automatyczny do mocy silnika. Niska moc wiąże się z dużymi przebiegami, a im wyższa moc, tym niższy przebieg. Możesz łatwo zobaczyć ten wzór w danych. Możesz zauważyć kształt, nie liniowy, ale nieco zakrzywiony. Może to dostarczyć wskazówek dotyczących typów modeli, które należy wypróbować później. Aplikacje do eksploracji danych często mają pewne interaktywne funkcje wyświetlania wykresów. Na przykład rysunek pokazuje, że najechanie kursorem myszy na punkt pokazuje dokładne wartości dwóch zmiennych dla tego punktu. Jest to łatwiejsze niż próba odczytania wartości z osi!



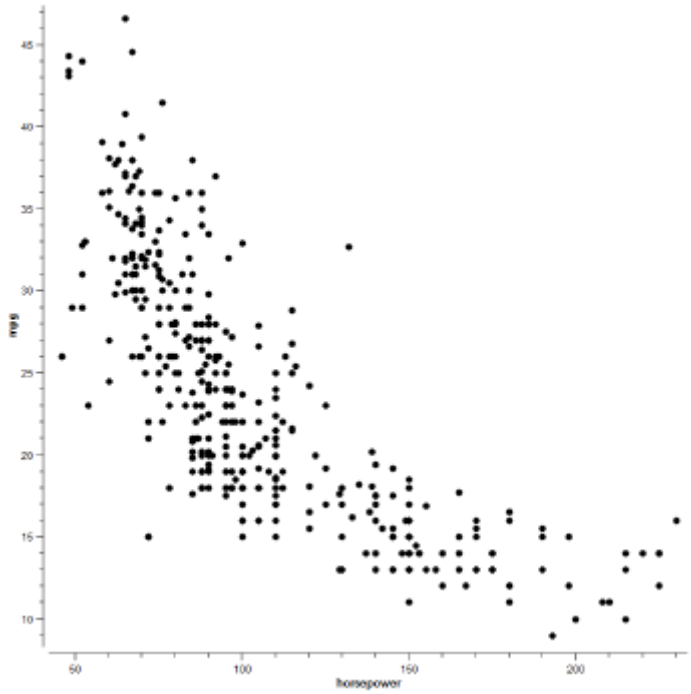
Po prostu powiedz „Nie” wykresom kołowym

Eksperci danych często używają wykresów słupkowych, histogramów i wykresów rozrzutu, ale możesz nigdy nie spotkać kogoś, kto używa wykresów kołowych. Co sprawia, że wykres słupkowy tak bardzo różni się od wykresu kołowego? Oba reprezentują względne częstotliwości kategorii. Wykres słupkowy jest liniowy, a widzowie wizualnie porównują długości słupka. Wykres kołowy nie jest liniowy. Na wykresie kołowym względne częstotliwości są reprezentowane przez obszar na wykresie powierzchnia. Ludziom trudno jest dokładnie porównać obszary. Ludzie są całkiem dobrzy w porównywaniu długości (jeden wymiar), a gorzej w porównywaniu obszarów (dwa wymiary). Nie chcesz nawet wiedzieć, co dzieje się z objętościami (trzy wymiary) lub skalami wykładniczymi. Skąd mam to wiedzieć? Badania! (Kto przeprowadził badania? Ja, ja to zrobiłem. No więc!) Co gorsza, wykres kołowy jest okrągły. Jeśli ludzie nie potrafią dokładnie porównać obszarów prostokątów (co wiem z badań), mógłbym podejrzewać, że jeszcze gorzej porównują obszary w kształcie plasterków ciasta. Nie znajdziesz splendoru w korzystaniu z wybrednych wykresów. Unikaj wszelkiego rodzaju reprezentacji nieliniowych. Nie używaj ciast, uroczych kształtów ani nieliniowych łusek. Nie używaj trójwymiarowych wykresów słupkowych. Niech Twoje wykresy będą proste i pouczające, aby uzyskać dobre wyniki i wspierać dobre decyzje biznesowe.

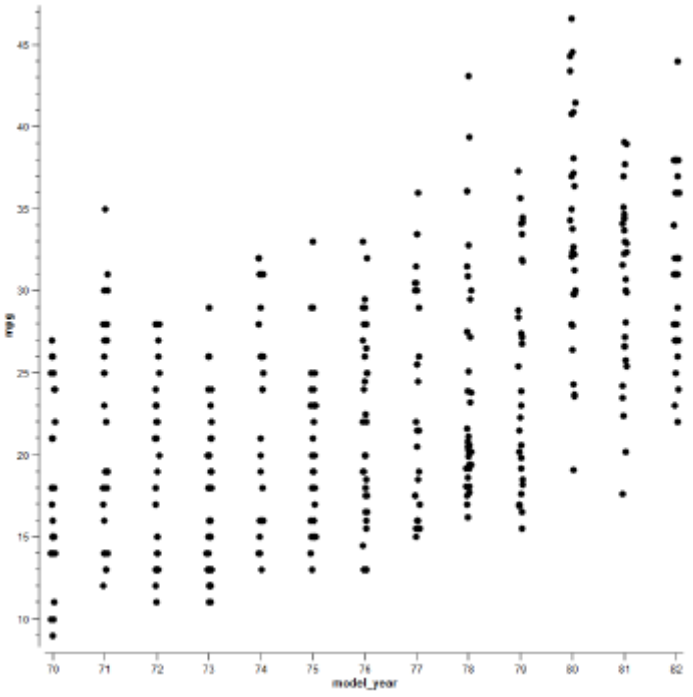
Opierając się na podstawach

Eksperci danych często korzystają ze specjalnych funkcji, aby umieścić więcej informacji na prostych wykresach. Etykiety, nakładki i interaktywny wybór to cechy charakterystyczne aplikacji do eksploracji danych, specjalne funkcje, które pozwalają zwiększyć produktywność. Sprawianie, że wykresy rozrzutu mówią więcej

Rysunek szósty przedstawia wykres rozrzutu, który łączy przebieg automatyczny z mocą silnika. Widać, że przebieg zmniejsza się wraz ze wzrostem mocy. To pierwszy krok w zrozumieniu czynników, które decydują o przebiegu. Przebieg zmniejsza się wraz ze wzrostem mocy, jak widać na rysunku.

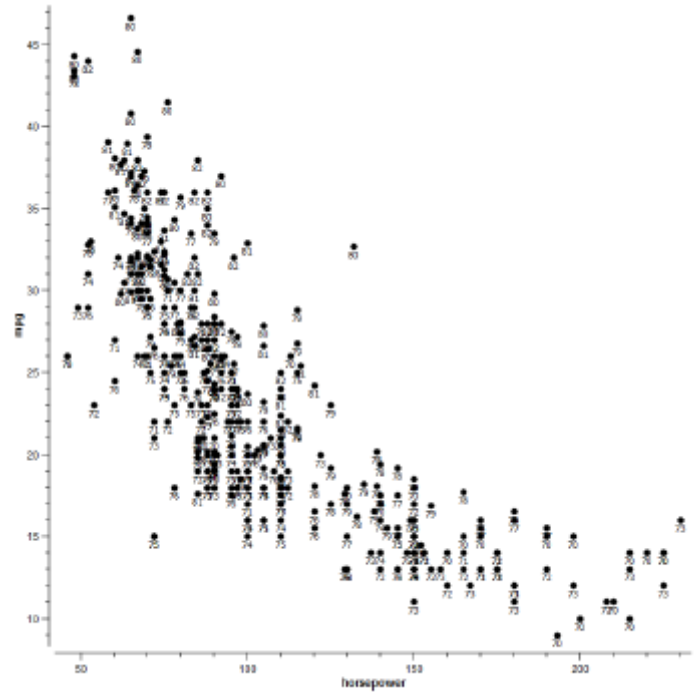


Przebieg zwiększa się z czasem, jak widać na rysunku, wykresie rozrzutu przebiegu w porównaniu z rokiem modelowym. Pomocne byłoby zebranie tych dwóch pomysłów na jednym wykresie.

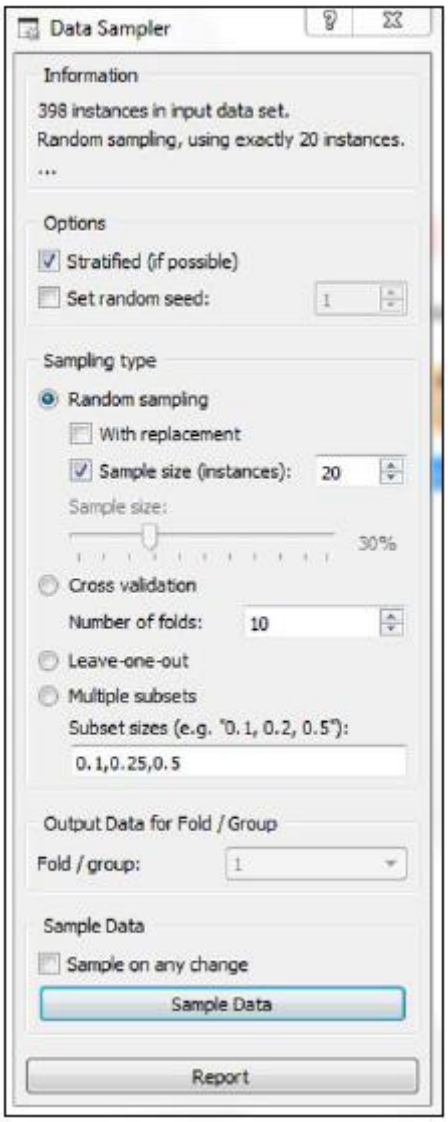


Typowe podejścia do eksploracji danych do integrowania więcej niż dwóch zmiennych na wykresie obejmują

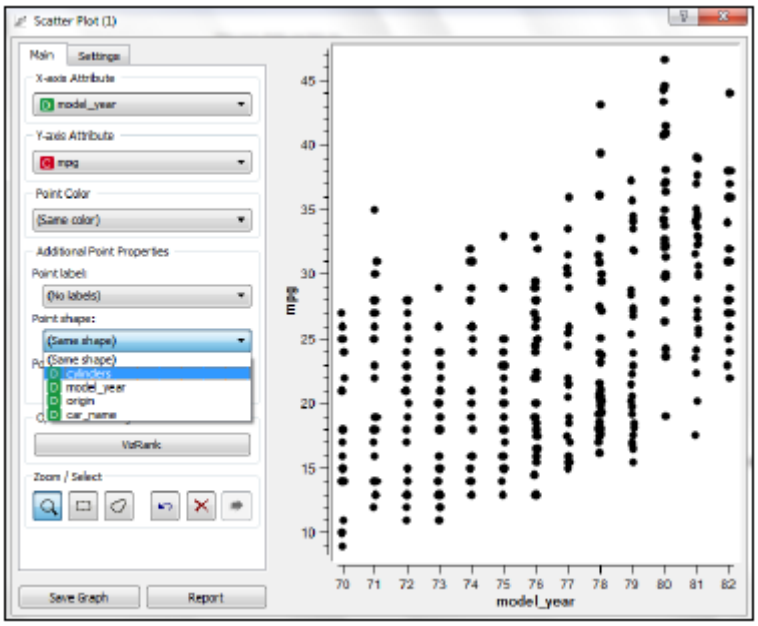
✓ Etykiety: Etykiety to wartości ciągu znaków lub zmiennej kategorialnej, które zostały nałożone na wykres rozrzutu. Rysunek przedstawia wykres rozrzutu oznaczony rokiem modelowym samochodu.

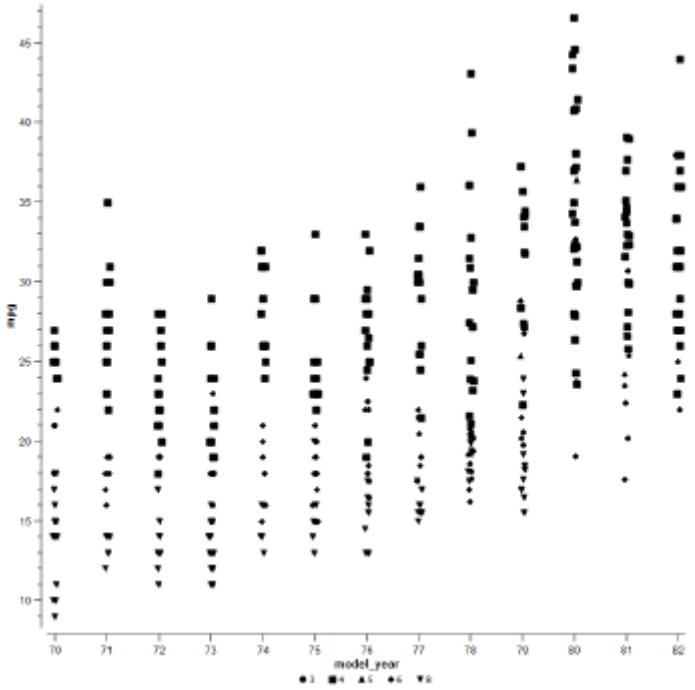


(Zestawy danych z wieloma punktami lub długimi etykietami mogą jednak spowodować, że te wykresy będą nieczytelne! Rozwiązaniem jest użycie tylko próbki danych. Konfiguracja tego rodzaju próbkowania jest pokazana na rysunku)



✓ Nakładki: W przypadku nakładek wartości zmiennej kategorycznej definiują kształt lub kolor punktów. Rysunek przedstawia konfigurację, w której wykres rozrzutu nakłada się na rok modelowy na wykresie rozrzutu w zależności od liczby koni mechanicznych, a wyeksportowany wykres rozrzutu nakładany jest na rysunku.





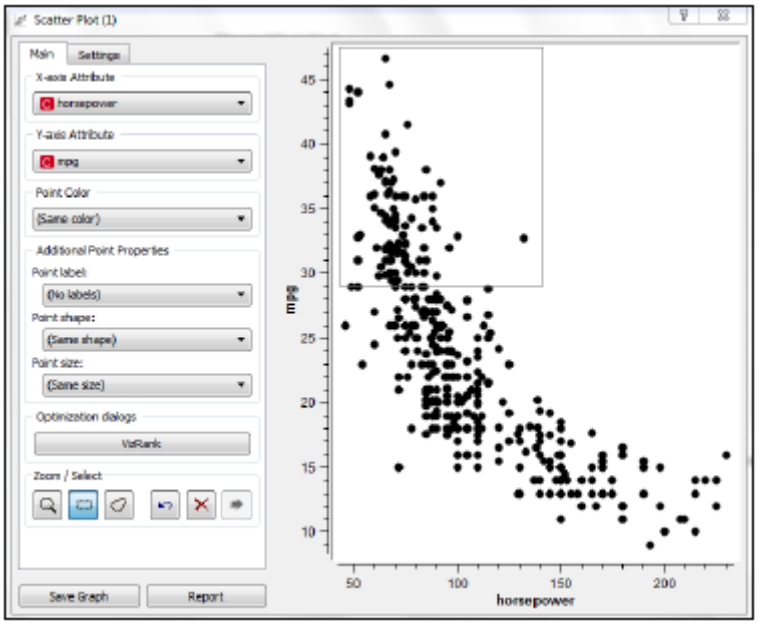
Odczytywanie kolorowych nakładek może być łatwiejsze niż nakładek z kształtami punktów. Konfiguracja jest zwykle taka sama.

Kolejna rzecz, o której należy pamiętać w przypadku wykresów rozrzutu: możesz mieć wiele punktów przypadających na to samo miejsce! Jeśli tak, możesz nie być w stanie odróżnić punktu dla jednego przypadku od punktu dla 100 przypadków. Rozwiązaniem jest sprawdzenie opcji wyświetlania wielu instancji. Poszukaj opcji rozmiaru punktu lub jittera (przesuwa punkty nieco poza ich prawdziwe położenie, aby wszystkie były widoczne).

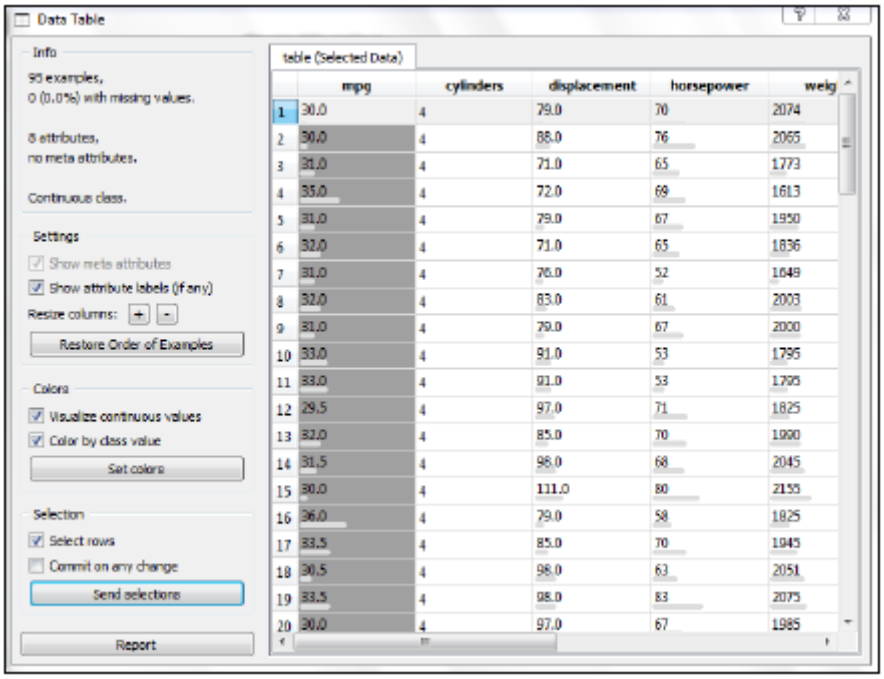
Interakcja z wykresami rozrzutu

Interaktywne wykresy rozrzutu to świetna oszczędność czasu dla eksploratorów danych. Załóżmy, że widzisz na wykresie interesującą grupę przypadków i chcesz dokładniej zbadać tylko te przypadki. Jeśli

patrzysz tylko na jeden lub dwa punkty, możesz uzyskać potrzebne informacje, najeżdżając kursorem na , ale nie jest to satysfakcjonujące, gdy interesuje Cię więcej niż kilka punktów. Narzędzia do zaznaczania danych na interaktywnych wykresach rozrzutu dają więcej możliwości wybierania danych. Rysunek przedstawia tę samą konfigurację wykresu, ale z grupą punktów wybranych przez kliknięcie i przeciągnięcie wokół nich myszy.

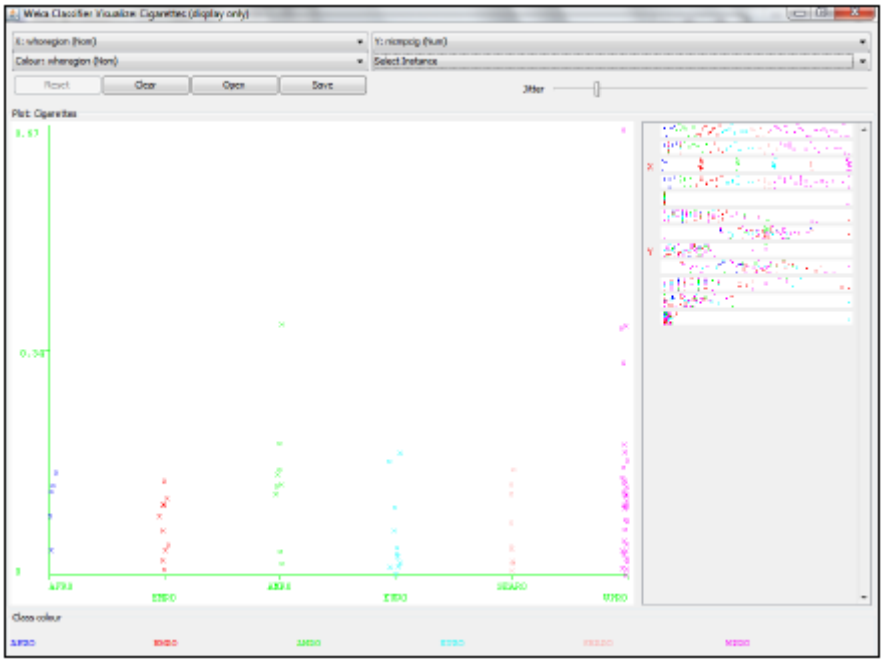


To nie tylko funkcja wizualna. Wybrane punkty można wyeksportować jako nowy zestaw danych.

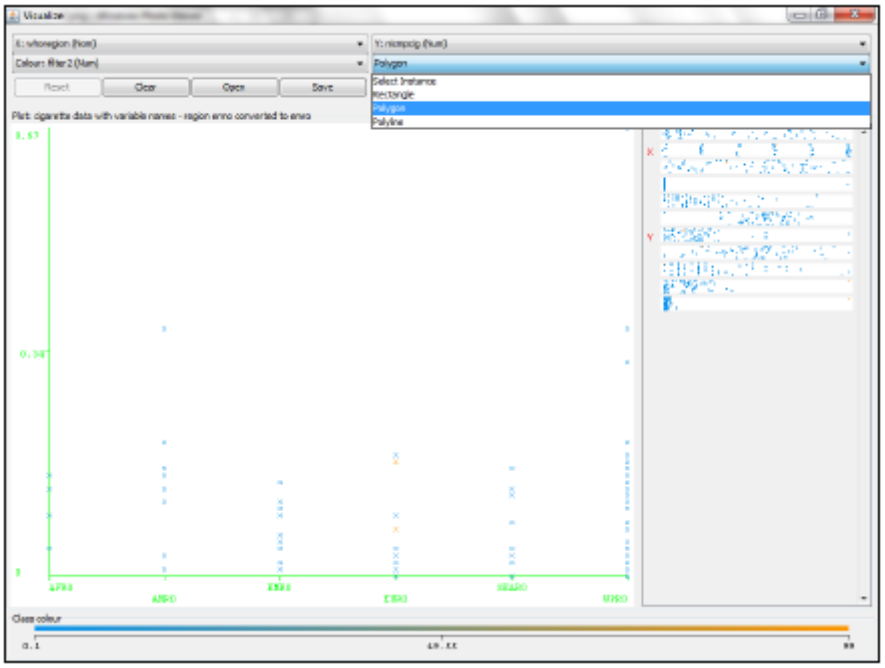


To bardzo przydatne i szybkie! Jeśli potrzebne ci punkty nie pasują do prostokątnego zaznaczenia, masz inne opcje. Zobacz obszar Zoom/Select na rysunku. Możesz zobaczyć przycisk z prostokątem do zaznaczenia prostokąta, a drugi z zaokrąglonym kształtem do swobodnego wyboru. Oto przykład swobodnego wyboru z wykorzystaniem danych dotyczących zawartości nikotyny w papierosach

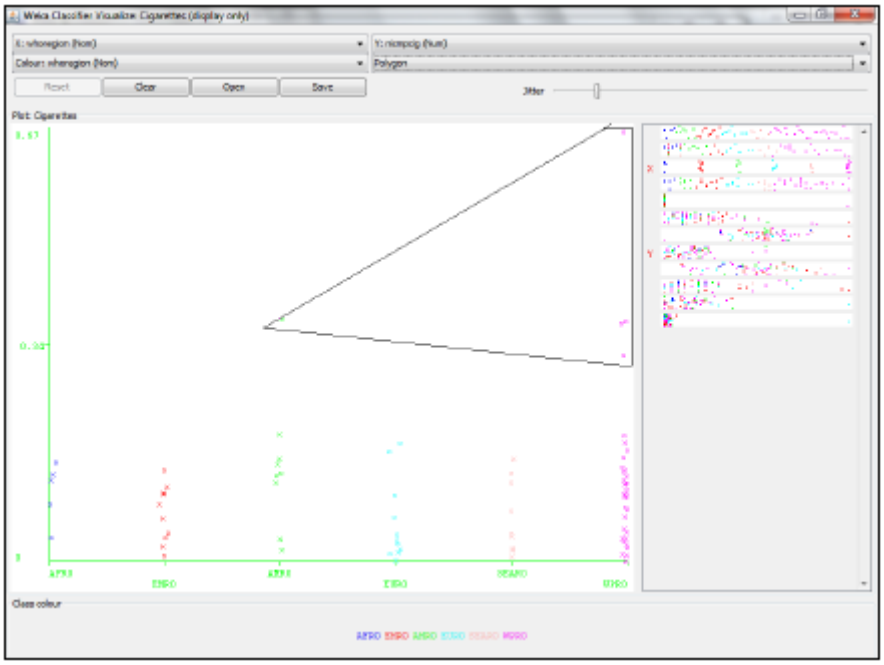
sprzedawanych w różnych częściach świata. Ten wykres rozrzutu przedstawia nikotynę na papieros dla próbek z sześciu regionów Organizacji Narodów Zjednoczonych.



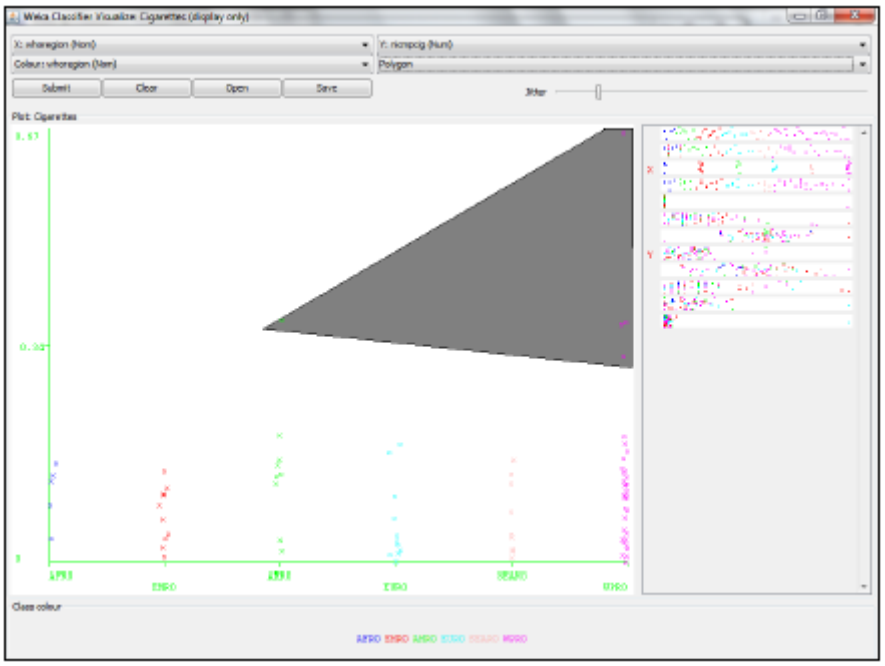
(Jest to nietradycyjne użycie wykresu rozrzutu, ponieważ region nie jest zmienną ciągłą; jest kategoryczny. Eksperci danych często używają tradycyjnych narzędzi w nietradycyjny sposób.) Punkty w regionie nie układają się w idealną linię pionową. Niewielkie przesunięcia (jitter) w lewo i w prawo mają na celu wyłącznie czytelność i wygląd. Kilka papierosów ma wyjątkowo wysoki poziom nikotyny i chcesz wybrać te przypadki. Rozwijane menu oferuje opcje wyboru.



Zaznaczanie wielokątów umożliwia zaznaczenie na wykresie rozrzutu obszaru o dowolnym kształcie. Aby zaznaczyć, kliknij wykres, aby utworzyć punkt początkowy, a następnie klikaj raz za razem wokół grupy punktów, które chcesz, aż uzyskasz pożądany kształt.



Kliknięcie prawym przyciskiem myszy oznacza, że dokonałeś wyboru; widać to po podświetleniu na wykresie.



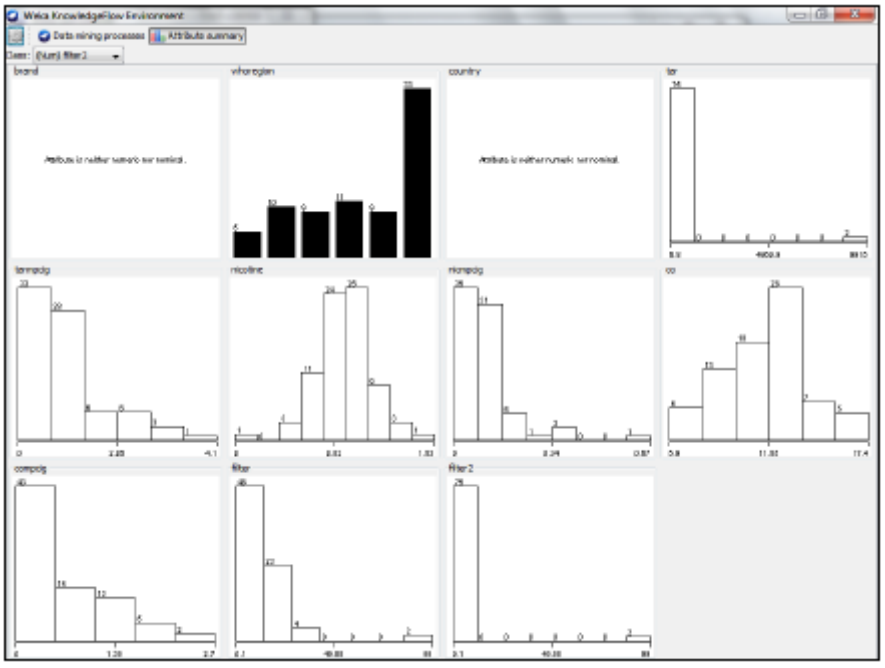
Szybka praca z wykresami Galore

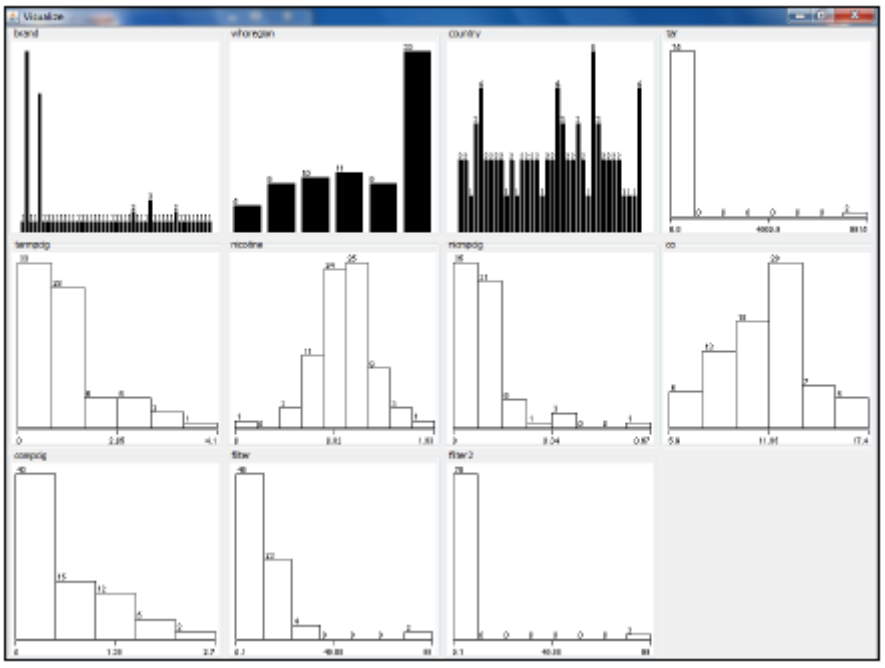
Eksperci danych działają szybko. Jednym ze sposobów na zwiększenie produktywności jest pełne wykorzystanie narzędzi, które pozwalają robić kilka rzeczy naraz.

Konfigurowanie kilku wykresów oddzielnie, po jednym na raz, jest czasochłonne (i nudne). Dlatego używaj tych alternatyw, kiedy tylko możesz:

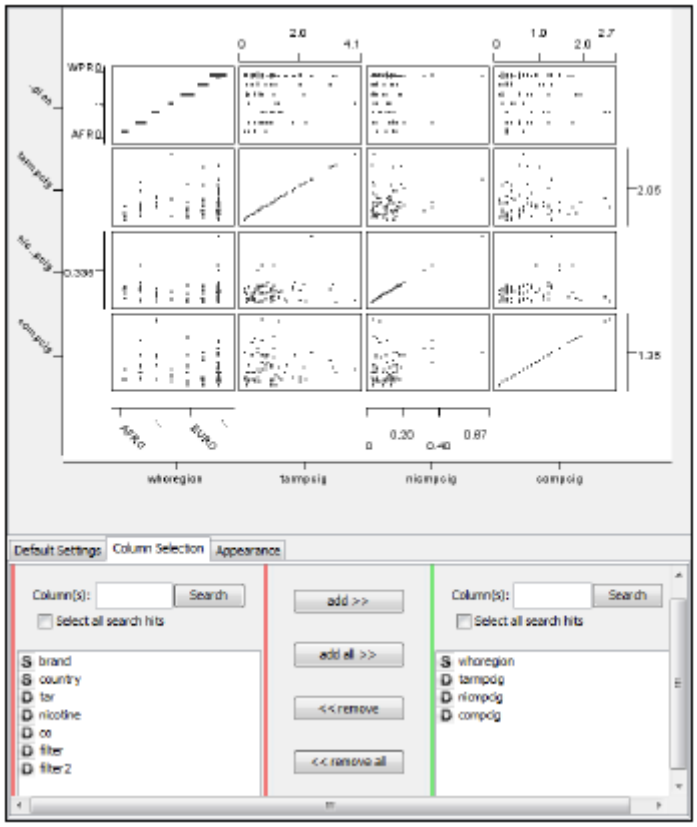
✓ Podsumowania danych: narzędzia, które pozwalają szybko poprosić o podsumowania wielu zmiennych i uzyskać wszystkie podsumowania naraz. Wyniki często zawierają wykresy słupkowe i histogramy.

✓ Macierz wykresu: Wynikiem jest siatka (macierz) małych wykresów słupkowych i histogramów, dzięki czemu można szybko przeglądać wiele rozkładów danych. Rysunki poniżej pokazują dwa przykłady. Zostały stworzone z różnych produktów, ale są bardzo podobne pod względem funkcji i wyglądu.





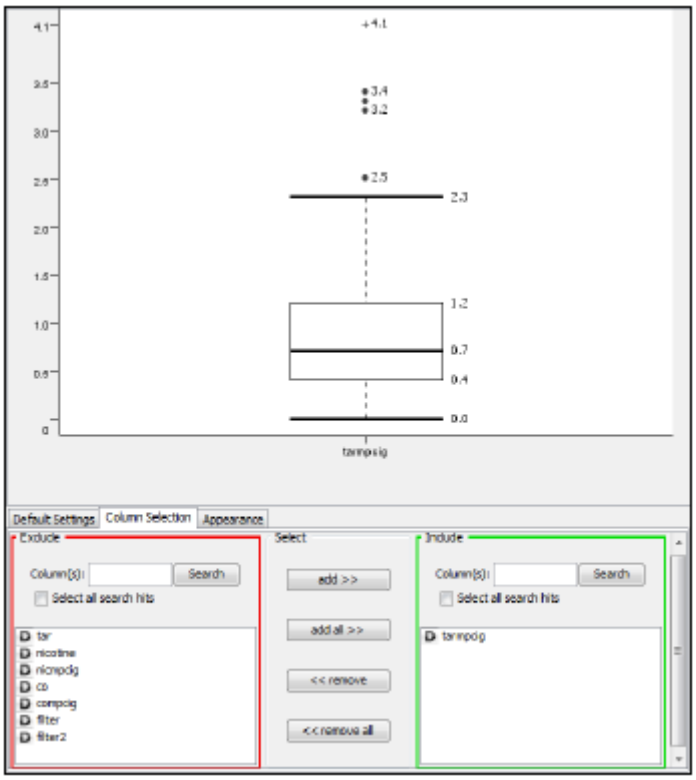
✓ Macierz wykresów rozrzutu: siatka małych wykresów rozrzutu. Każdy mały wykres przedstawia zależność dla jednej pary zmiennych (zazwyczaj używa się zmiennych ciągłych). Wprowadzasz listę zmiennych, a macierz wykresu rozrzutu pokazuje wszystkie możliwe pary.



Rozszerzenie zakresu grafiki

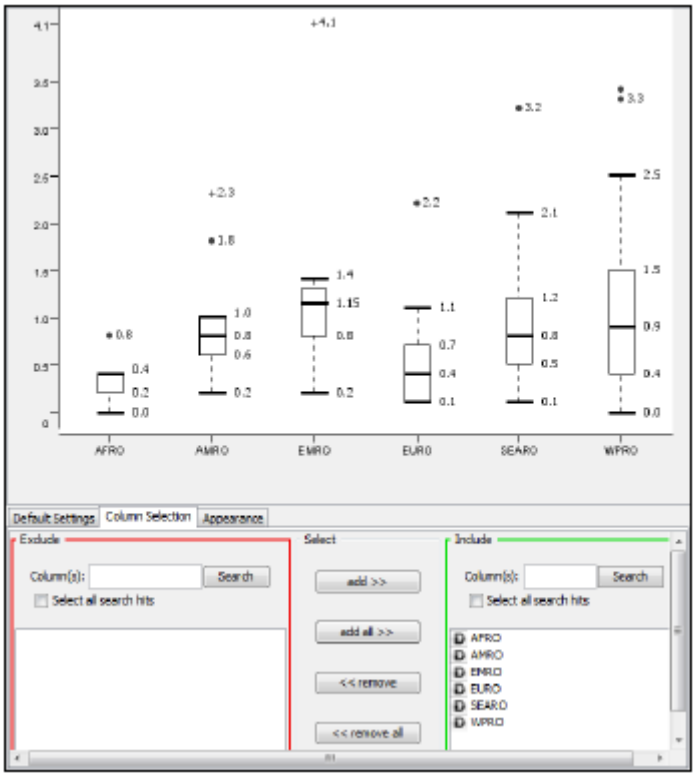
Ponieważ eksploratorzy danych opierają się głównie na podstawowych wykresach, niektóre aplikacje do eksploracji danych oferują niewiele lub nic więcej. Inne zapewniają szeroką gamę opcji wykresów, od zwykłych po egzotyczne. Nie ma potrzeby używania ich wszystkich (przeczytaj, co jeden z ekspertów ma do powiedzenia na ten temat w pobliskim pasku „Umieszczanie grafiki w kontekście: Wywiad z Laurą Kippen”), ale możesz skorzystać, wybierając i używając kilku, które najbardziej Ci odpowiadają. wymagania. Eksperci danych często używają tych wykresów:

✓ Wykres pudełkowy (zwany również pudełkowym i wąsami): Histogramy opisują rozkłady zmiennych ciągłych, ale mają ograniczoną wartość do pokazywania szczegółów. Alternatywą jest wykres pudełkowy.

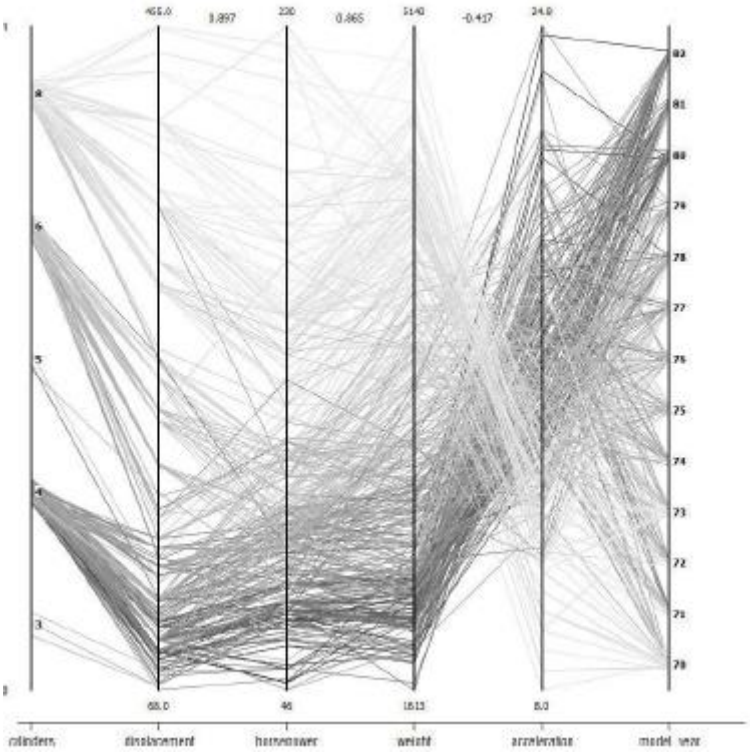


Sercem obrazu jest pudełko; stanowi to połowę danych, pobranych w środku jego zakresu. Środek pudełka to mediana wartości zmiennej, a dolny i górny koniec pudełka reprezentują odpowiednio 25. i 75. percentyl. Wąsy rozciągają się pod i nad polem, reprezentując zakres większości danych. Punkty poza wąsami są traktowane jako wartości odstające, wartości wysoce nietypowe (niektóre wykresy wskazują również skrajności, które są wartościami odstającymi wśród wartości odstających).

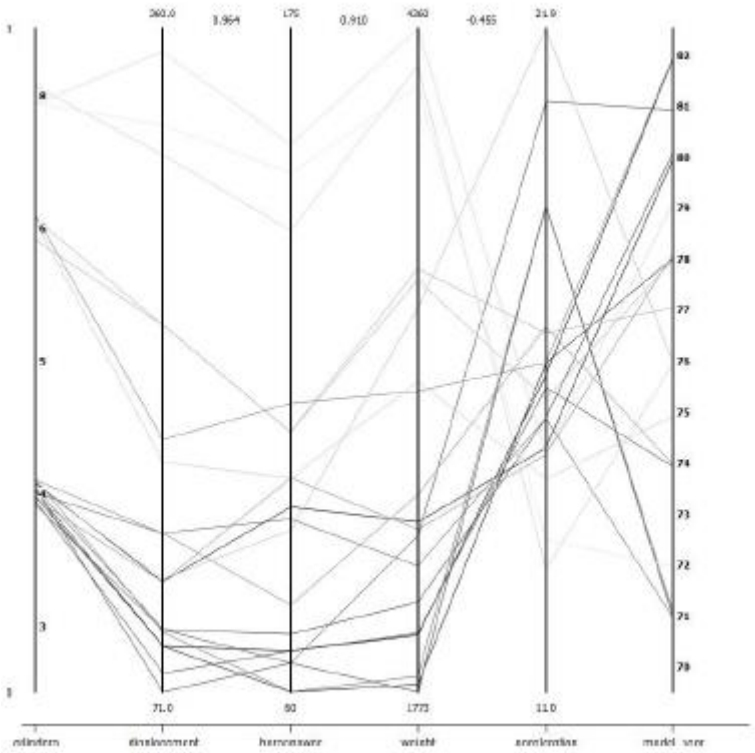
✓ Warunkowy wykres pudełkowy: wykresy pudełkowe dla kilku grup (takich jak regiony geograficzne) można umieścić obok siebie na jednym wykresie w celu łatwego porównania.



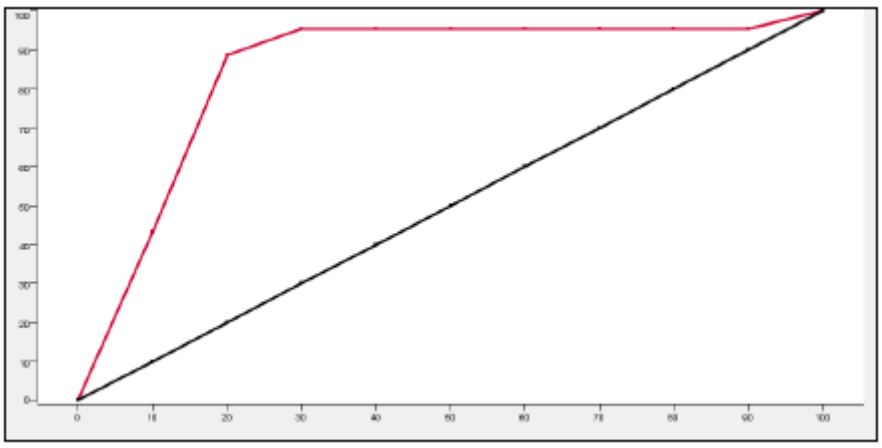
✓ Współrzędne równoległe: Wykresy przedstawiają wartości kilku zmiennych razem na jednym wykresie, przy czym wartości dla każdego przypadku są połączone segmentami linii. Wspólne kombinacje wyróżniają się od reszty. Na przykład spójrz na rysunek który pokazuje kilka zmiennych związanych z samochodami i zużyciem paliwa.



Wiele przypadków ma wspólne wartości, dokładnie lub w przybliżeniu tworząc ciemne wzory z wielu linii podążających podobnymi ścieżkami na wykresie. Na przykład kufry do samochodów czterocylindrowych o małej pojemności skokowej, duży przebieg i późne lata modelowe tworzą bardzo ciemny i rzucający się w oczy wzór. (Wykresy współrzędnych równoległych są trudne do odczytania, gdy używanych jest zbyt wiele przypadków. Jeśli tak się stanie, weź losową próbkę danych, jak pokazano na rysunku, i utwórz nowy wykres.)



✓ Wykresy zysków (nazywane również skumulowanymi zyskami): Wykres zysków pokazuje, jak bardzo model predykcyjny poprawia wyniki w porównaniu z losowym próbkowaniem.



Niektórzy ludzie są bardziej skłonni do działania (kupują produkt, głosują na kandydata, łamią prawo...) niż inni. Jeśli nic nie wiesz o grupie ludzi, najlepsze, co możesz powiedzieć, to to, że skontaktowanie się z połową osób przyniesie połowę tych, którzy podejmą działania. Ale model predykcyjny może ci

powiedzieć, którzy ludzie są najlepszymi perspektywami, więc możesz użyć modelu, aby wybrać połowę (lub 10 procent lub 60 procent itd.) i uzyskać więcej działań. O ile więcej? Na wykresie na rysunku powyżej widać ukośną linię, w której wartości x i y są zawsze takie same; reprezentuje to, co otrzymasz, wybierając losowo potencjalnych klientów. Druga linia reprezentuje model. Różnica w wartościach y między modelem a wyborem losowym pokazuje, jak bardzo model poprawia wynik. Odczytaj linię modelu wykreśloną na wykresie, i porównaj ją z linią do losowego pobierania próbek.

✓ Wykresy wzrostu: Wykresy wzrostu są bardzo podobne do wykresów zysków. Kluczową różnicą jest to, że dane są znormalizowane, dzięki czemu losowe próbkowanie jest zawsze reprezentowane jako wartość 1, a wyniki modelu są wyświetlane proporcjonalnie do losowego próbkowania.

Możesz zobaczyć kilka różnych typów wykresów, zwanych wykresami wzrostu. Niektóre kumulują się, a inne nie. Niektóre mogą być nawet wykresami zysków (opisanych wcześniej).

Pokazywanie Twoich danych, kto tu rządzi

Ekspert ds. urody opowiedział historię rozchwytywanej wizażystki, której umiejętność sprawiania, by ludzie wyglądali jak najlepiej, przyciąga mnóstwo sławnych klientów. Wyjaśniła, że jeśli ta wizażystka spędziła godzinę z klientem, 50 minut poświęcono na przygotowanie, dokładne oczyszczenie skóry, nawilżenie i podjęcie innych kroków, aby położyć podwaliny pod doskonałe rezultaty. Jako eksplorator danych również poświęcisz większość czasu na przygotowanie. Twój wysiłek zostanie nagrodzony, ponieważ dobre przygotowanie danych to podstawa dla doskonałych wyników eksploracji danych. Przygotowanie danych wymaga czasu i cierpliwości. Wiele osób poświęca znacznie więcej czasu niż to konieczne na przygotowanie danych lub uzyskuje słabe wyniki, ponieważ nigdy nie mieli cierpliwości, aby poświęcić czas na opanowanie umiejętności eksploracji danych.

Słabe umiejętności w zakresie przygotowywania danych mogą spowodować niepowodzenie w zbudowaniu skutecznego modelu predykcyjnego (być może wtedy, gdy konkurencja odniesie sukces) i zmarnować godziny, a nawet dni cennego czasu na każdy projekt. Jeśli chcesz zostać wybitnym eksploratorem danych, rutynowo inwestuj swój czas i cierpliwość, aby rozwijać swoje umiejętności w zakresie przygotowywania danych. Będziesz potrzebować dwóch rodzajów informacji:

✓ Rodzaje manipulacji danymi: Jeśli nie wiesz, co jest możliwe i potencjalnie przydatne, nie będziesz ich szukać ani nie wiesz, co to jest, gdy je zobaczysz.

✓ Możliwości i metody w Twoich narzędziach: Większość aplikacji do eksploracji danych oferuje szeroki zakres technik manipulacji danymi, ale wymagane kroki różnią się znacznie w zależności od produktu.

Żadna książka nie może podać instrukcji krok po kroku dla każdej metody manipulacji danymi w każdym produkcie; które zajęłoby wiele tysięcy stron i wymagałoby aktualizacji prawie każdego dnia. Ale jeśli rozumiesz, co jest możliwe i dlaczego każda technika jest przydatna, będziesz wiedział, czego szukać w każdej aplikacji do eksploracji danych. Weź udział w odpowiednim, bezpośrednim szkoleniu dotyczącym aplikacji do eksploracji danych. Nie oszczędzaj na szkoleniu! Pieniądze, które wydajesz na oprogramowanie i czas spędzony na jego używaniu, mogą zostać zmarnowane, jeśli nie wykorzystasz w pełni tego, co masz. (Jeśli nie możesz uczestniczyć w szkoleniu na żywo, zdobądź podręczniki do samodzielnej nauki i samouczki do aplikacji do eksploracji danych, przeczytaj je i wykonaj ćwiczenia!) I nie poprzestawaj na tym. Od czasu do czasu przeglądaj materiały szkoleniowe i zapoznaj się z dokumentacją i samouczkami, aby znaleźć szczegółowe informacje, które pozwolą pogłębić swoją wiedzę. W tym rozdziale przedstawiono szeroki zakres metod manipulacji danymi, których używają eksploratorzy danych do przygotowania danych do modelowania. Ważne jest, aby zapoznać się z nimi wszystkimi, aby wiedzieć, czego szukać w aplikacjach do eksploracji danych i dlaczego.

Porządkowanie danych

Zanim zbudujesz modele lub wykonasz obliczenia dowolnego typu, powinieneś ułożyć dane tak, aby praca była jak najłatwiejsza. Możesz kontrolować format i wyświetlanie danych oraz sposób, w jaki dane będą traktowane podczas analizy.

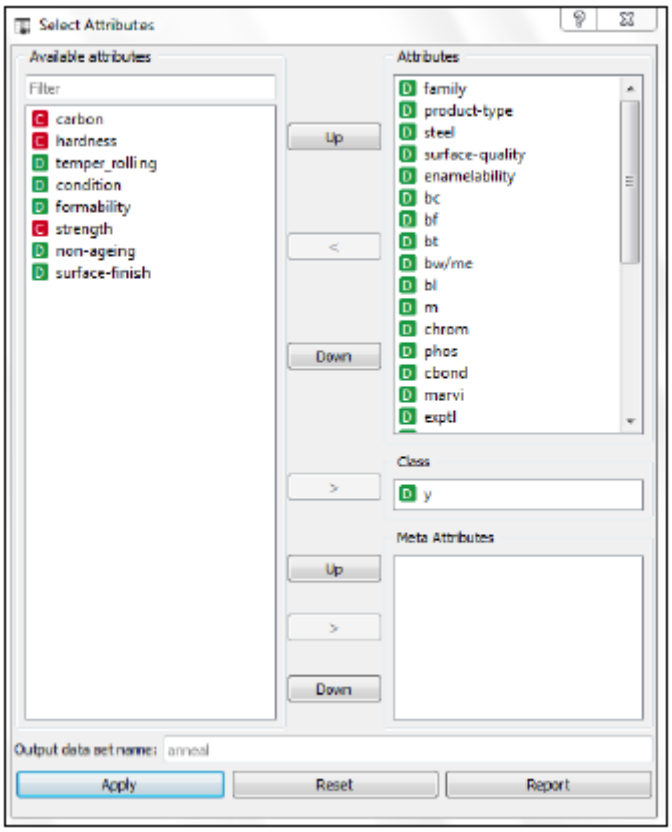
Kontrolowanie zmiennej kolejności

Kolejność zmiennych (kolumn) w zestawie danych jest zwykle kwestią ich ułożenia w pliku źródłowym lub zapytania bazy danych użytego do ich zaimportowania. Taki układ może nie być dla ciebie wygodny. Jeśli masz wiele zmiennych, może być trudno znaleźć te, które chcesz zobaczyć. A może jakiś porządek ma dla Ciebie sens i chciałbyś, aby zmienne były ułożone w ten sposób. Aplikacje do eksploracji danych często umożliwiają zmianę kolejności zmiennych, ale instrukcje rzadko są wyraźnie widoczne w menu lub w pomocy. Te funkcje są zwykle ukryte w narzędziach, które służą szerszym celom. Poszukaj takich

subtelnych opcji w przeglądarkach danych i oknach dialogowych dla innych procedur (zwłaszcza tych dotyczących manipulacji danymi i eksportu danych):

✓ Przeciągnij i upuść: Tabele wyświetlające dane lub metadane (takie jak nazwy i formaty zmiennych) mogą być interaktywne, co pozwala na zmianę kolejności zmiennych poprzez przeciąganie i upuszczanie kolumn.

✓ Przyciski W górę i W dół: Przyciski oznaczone słowami W górę i W dół lub strzałkami skierowanymi w górę lub w dół umożliwiają przesuwanie zmiennych w górę iw dół w ramach list w celu zmiany kolejności. Oba typy pokazano na rysunku.



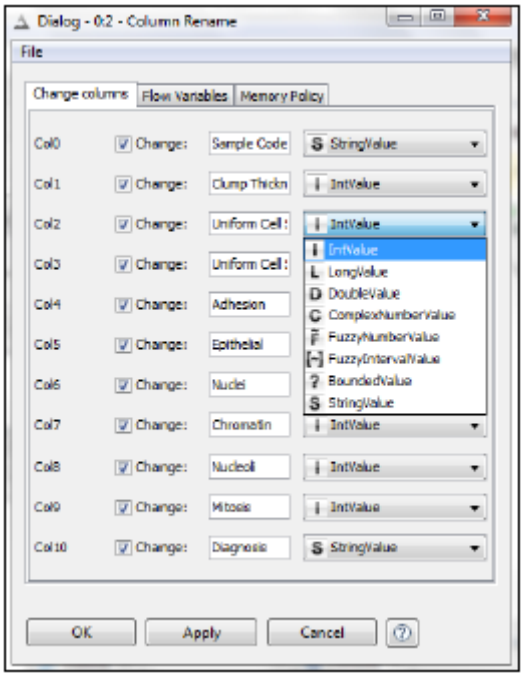
✓ Kolejność wyboru: Kiedy wybierasz zmienne z listy (ponieważ w danej procedurze używa się tylko kilku zmiennych z zestawu danych), kolejność, w jakiej je wybierasz, może pozostać w kolejnych operacjach.

✓ Gesty sortowania: Podczas przeglądania list zmiennych (metadanych) możesz mieć opcje sortowania, które umożliwiają zmianę kolejności zmiennych, na przykład w kolejności alfabetycznej lub według typu.

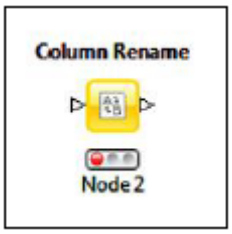
Możesz spotkać się z sytuacjami, w których rozmieszczenie zmiennych nie jest czysto kosmetyczne. (Na przykład aplikacja do eksploracji danych Orange oczekuje, że zmienna zależna będzie ostatnią zmienną w zbiorze danych). Jeśli nie możesz znaleźć sposobu na zmianę kolejności zmiennych w aplikacji do eksploracji danych, wróć do danych źródłowych i zmień format z innym narzędziem; następnie ponownie zaimportuj dane do swojej aplikacji do eksploracji danych.

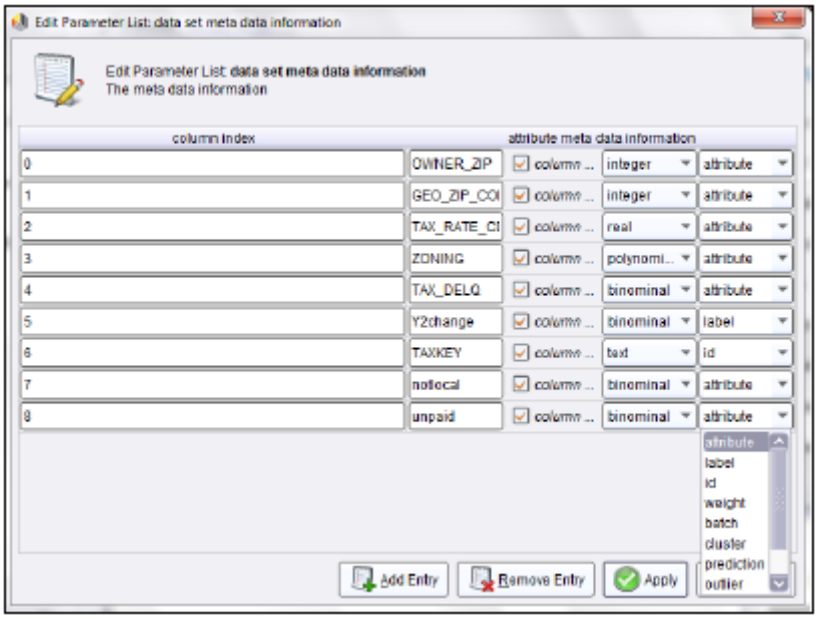
Prawidłowe formatowanie danych

Kiedy widzisz listę kodów pocztowych, nie próbujesz ich dodawać ani odejmować. Wiesz, że reprezentują miejsca. Rozumiesz to, ponieważ masz duże doświadczenie w obserwowaniu i rozpoznawaniu kodów pocztowych. Ludzie wykorzystują doświadczenie, gdy interpretują dane, które widzą, ale komputery nie. Twoje oprogramowanie do eksploracji danych dołoży wszelkich starań, aby zidentyfikować rodzaj danych w każdej kolumnie, ale typy danych są często niejednoznaczne. Może więc interpretować kod pocztowy jako liczbę całkowitą lub miarę ciągłą. Ostatecznie to do Ciebie należy zdefiniowanie odpowiedniego formatu. Funkcje do ustawiania formatów danych i ról (takie jak oznaczanie zmiennej zależnej do modelowania) mogą być ukryte w różnych miejscach w aplikacji do eksploracji danych. Możesz zdefiniować formaty i rolę zmiennych w pliku danych jeszcze przed otwarciem aplikacji do eksploracji danych (natywne formaty danych dla Orange i Weka pozwalają na to), w ramach importu lub jakiś czas później.



Możesz mieć narzędzia stworzone do tego celu, takie jak narzędzia pokazane na rysunkach , lub możesz zdefiniować te właściwości w ramach innych procedur

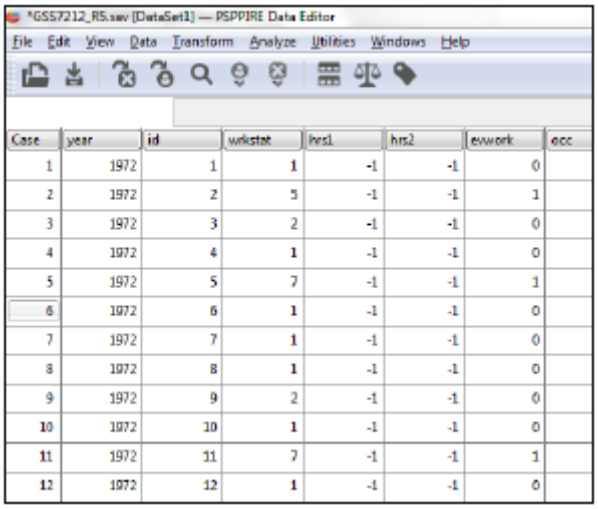




Każda aplikacja do eksploracji danych ma własny zestaw typów zmiennych i własne ograniczenia dotyczące sposobu użycia każdego typu. Niektóre z tych ograniczeń są oparte na teorii. Na przykład możesz dodawać i odejmować tylko liczby, a nie litery. Ale inne mogą być tylko kwestią sposobu zaprojektowania aplikacji. Na przykład może się okazać, że określone narzędzie do modelowania w jednej aplikacji pozwala przewidywać zarówno zmienne kategoryczne, jak i ciągłe, ale podobne narzędzie w innej aplikacji może umożliwiać modelowanie tylko jednego lub drugiego.

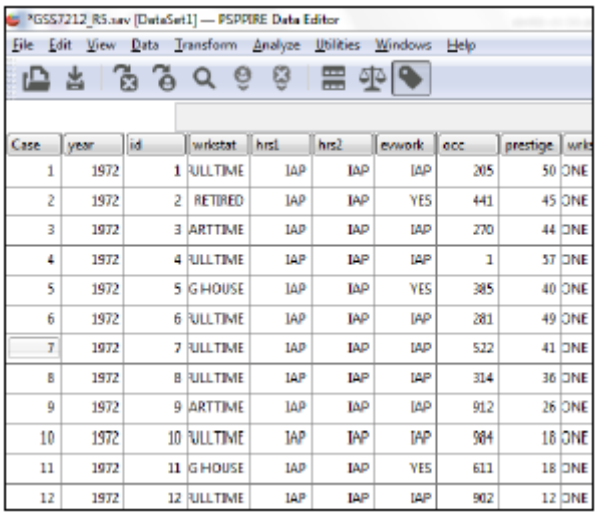
Dane do etykietowania

Możesz mieć zmienną regionu geograficznego z czterema możliwymi wartościami: Wschód, Zachód, Północ i Południe. Informacje można przechowywać w postaci ciągów znaków, wpisywanych pełnych słów lub w krótszym kodzie, na przykład 1 dla Wschodu, 2 dla Zachodu i tak dalej. Użycie kodu skraca czas wprowadzania danych, zapobiega błędom i zmniejsza wymagania dotyczące pamięci do przechowywania danych. Ale kody nie mają znaczenia, chyba że masz dokumentację lub etykiety wyjaśniające ich znaczenie. Niektóre formaty danych pozwalają czerpać korzyści z używania kodów przy jednoczesnym zachowaniu informacji o znaczeniu kodów w tym samym pliku. Nie są to typowe w eksploracji danych — jest bardziej prawdopodobne, że zobaczysz je w produktach do analizy statystycznej — ale niektóre aplikacje do eksploracji danych mogą używać tych oznaczonych formatów danych. Oto jak działają. Rozważmy rysunek który pokazuje zbiór danych otwarty w aplikacji do analizy statystycznej PSPP.



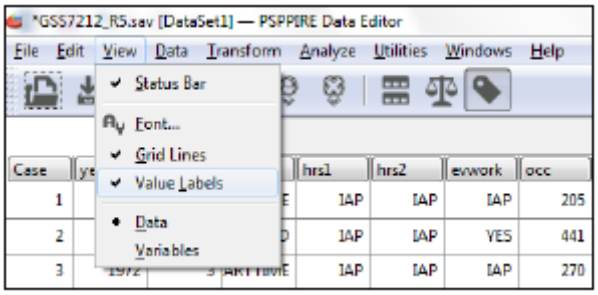
Case	year	id	wrkstat	hrs1	hrs2	evwork	occ
1	1972	1	1	-1	-1	0	
2	1972	2	5	-1	-1	1	
3	1972	3	2	-1	-1	0	
4	1972	4	1	-1	-1	0	
5	1972	5	7	-1	-1	1	
6	1972	6	1	-1	-1	0	
7	1972	7	1	-1	-1	0	
8	1972	8	1	-1	-1	0	
9	1972	9	2	-1	-1	0	
10	1972	10	1	-1	-1	0	
11	1972	11	7	-1	-1	1	
12	1972	12	1	-1	-1	0	

(Jeśli chcesz poeksperymentować z PSPP, możesz go znaleźć na www.gnu.org/software/pspp). Dane wydają się zawierać tylko liczby, ale te liczby są kodami wartości zmiennych kategorycznych. Rysunek przedstawia ten sam zestaw danych z etykietami zamiast kodów numerycznych.

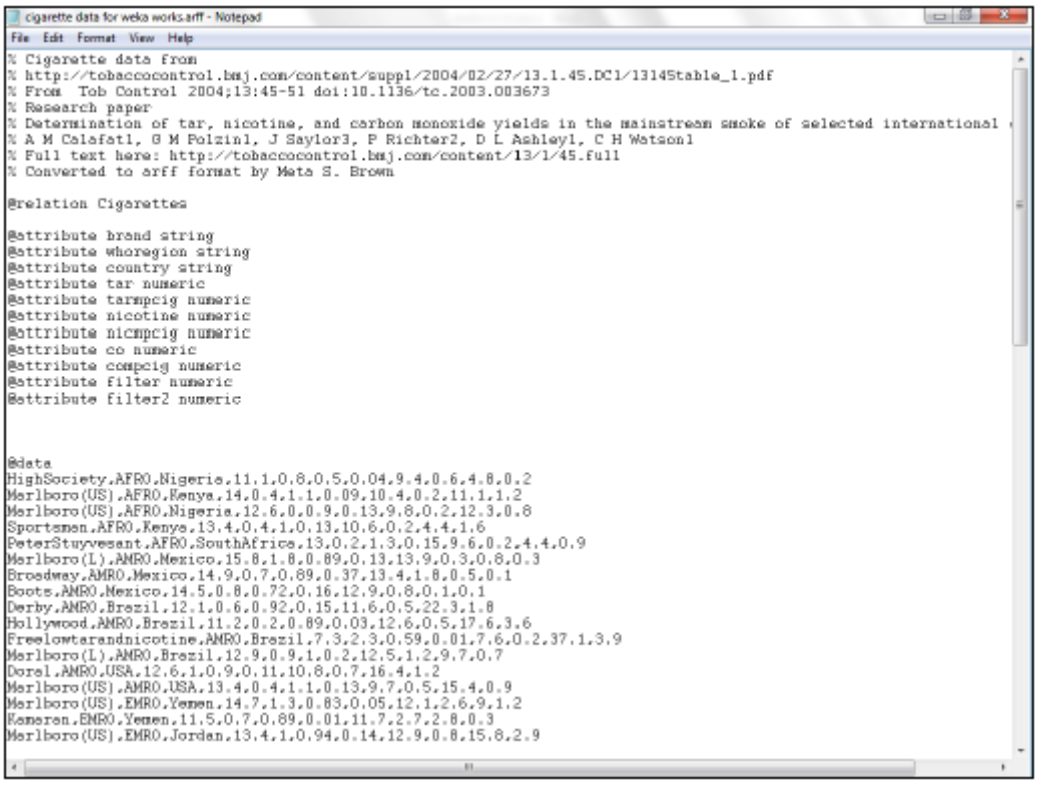


Case	year	id	wrkstat	hrs1	hrs2	evwork	occ	prestige	wrks
1	1972	1	FULLTIME	IAP	IAP	IAP	205	50	ONE
2	1972	2	RETIRED	IAP	IAP	YES	441	45	ONE
3	1972	3	PARTTIME	IAP	IAP	IAP	270	44	ONE
4	1972	4	FULLTIME	IAP	IAP	IAP	1	57	ONE
5	1972	5	G HOUSE	IAP	IAP	YES	385	40	ONE
6	1972	6	FULLTIME	IAP	IAP	IAP	281	49	ONE
7	1972	7	FULLTIME	IAP	IAP	IAP	522	41	ONE
8	1972	8	FULLTIME	IAP	IAP	IAP	314	36	ONE
9	1972	9	PARTTIME	IAP	IAP	IAP	912	26	ONE
10	1972	10	FULLTIME	IAP	IAP	IAP	984	18	ONE
11	1972	11	G HOUSE	IAP	IAP	YES	611	18	ONE
12	1972	12	FULLTIME	IAP	IAP	IAP	902	12	ONE

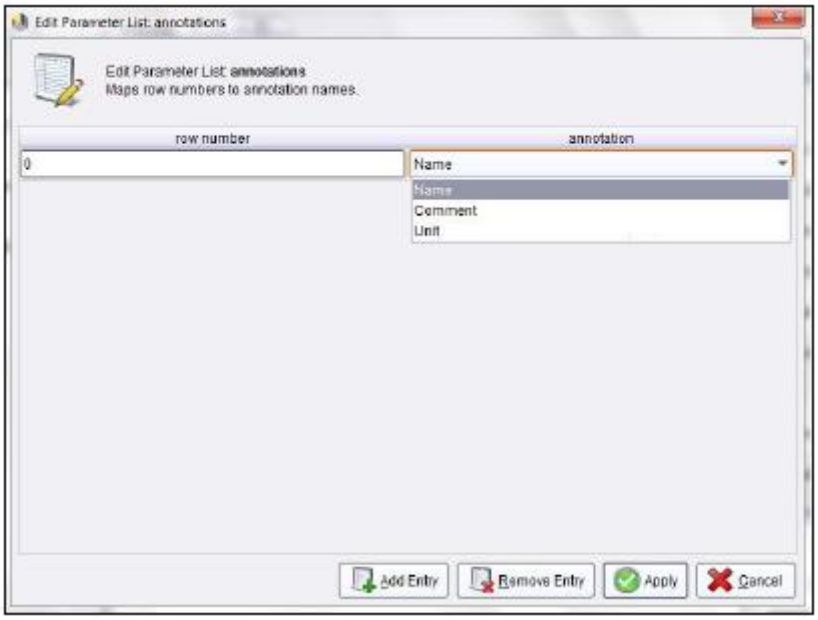
Możesz przełączać się między tymi dwiema opcjami wyświetlania za pomocą menu, jak pokazano na rysunku.



Chociaż dane są przechowywane jako liczby, etykiety pozwalają zobaczyć, co oznaczają dane, niezależnie od tego, czy przeglądasz je w edytorze danych, przeprowadzasz analizę lub przeglądasz wyniki. Możesz również znaleźć inne typy etykiet danych w aplikacjach do eksploracji danych. Natywny format danych dla Weka umożliwia dołączanie komentarzy do zestawu danych;



jest to dobre miejsce do umieszczania adnotacji o źródle danych i innych ważnych szczegółów. RapidMiner posiada również opcję adnotacji; możesz użyć graficznego interfejsu użytkownika, aby wprowadzić adnotacje dla poszczególnych wierszy danych.

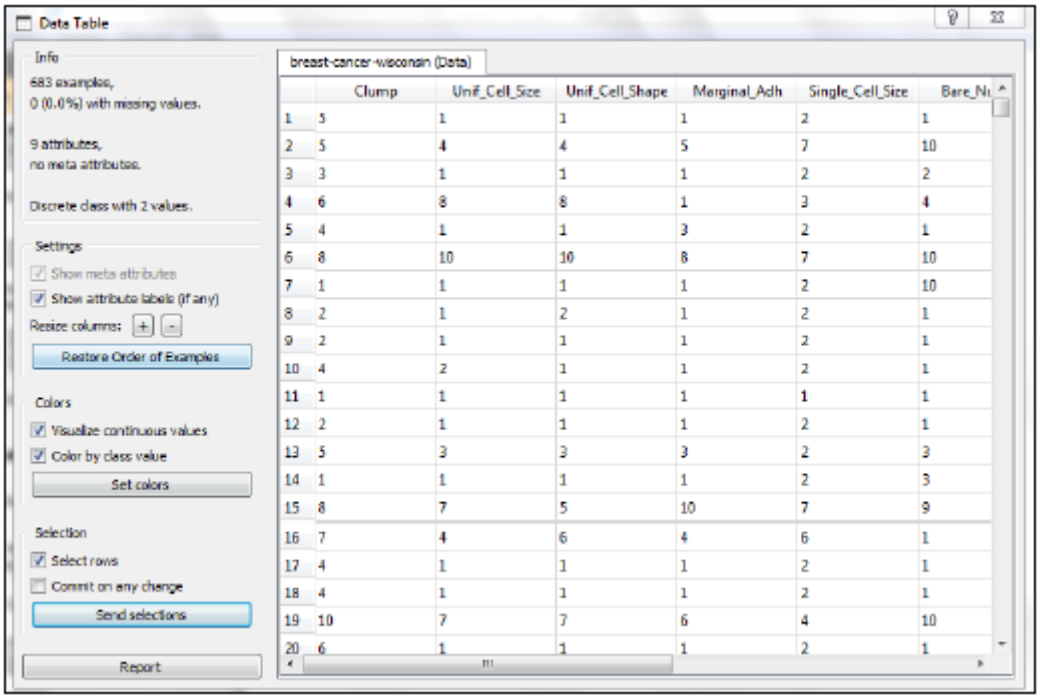


Kontrolowanie kolejności spraw

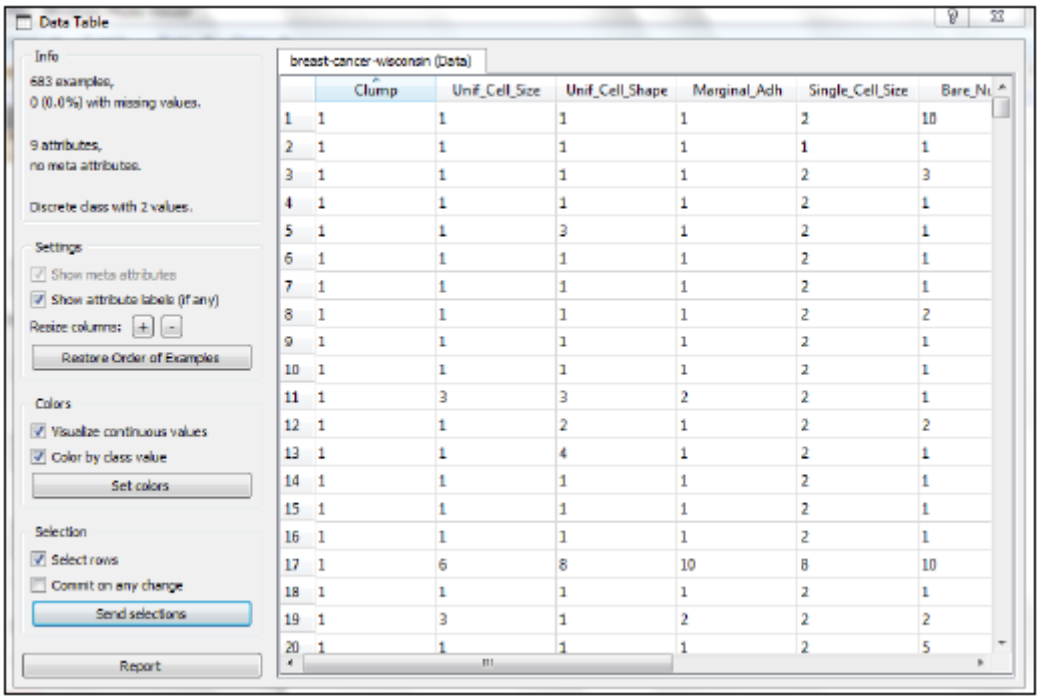
Eksperci danych często sortują sprawy (zmieniają kolejność wierszy), aby uzyskać bardziej przejrzystą organizację przeglądania danych lub eksportu. Lub możesz mieć funkcjonalny powód do sortowania. Na przykład niektóre aplikacje wymagają sortowania danych przed scaleniem (łączenie kolumn z

różnych źródeł danych). Kroki sortowania różnią się znacznie w zależności od aplikacji. Oto typowe opcje, które możesz znaleźć w swojej aplikacji do eksploracji danych:

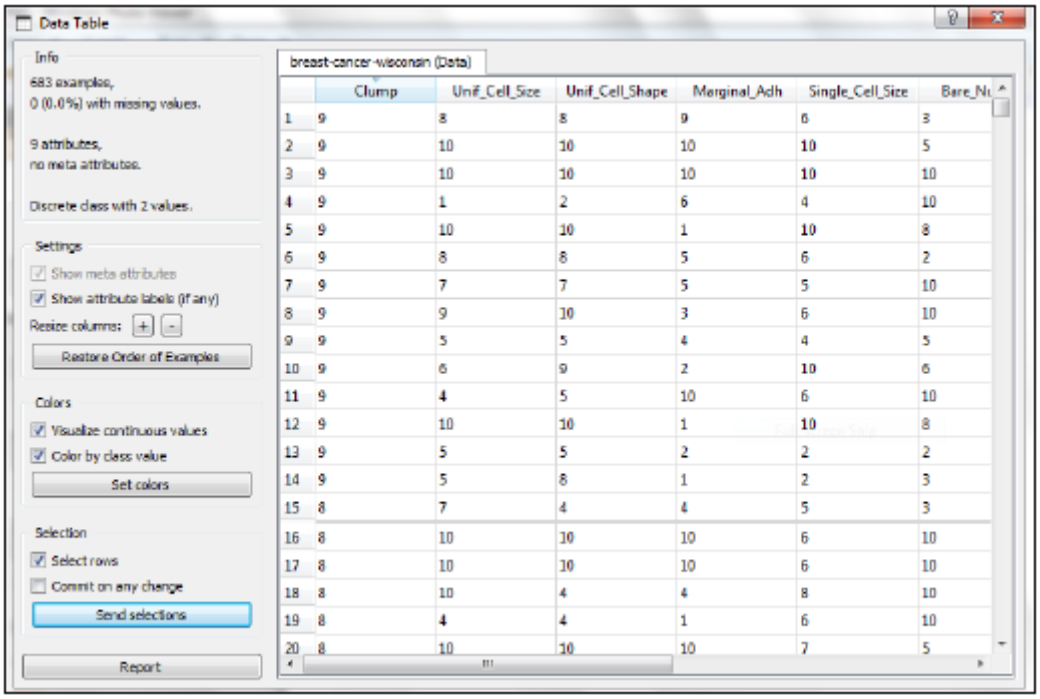
✓ Interaktywne sortowanie w przeglądarce danych: Spójrz na pierwszą zmienną na rysunku, Clump.



Wartości zmieniają się w zależności od przypadku, bez określonej kolejności. Jedno kliknięcie na nagłówku kolumny (gdzie widać nazwę zmiennej) sortuje dane w porządku rosnącym tej zmiennej, jak pokazano na rysunku.

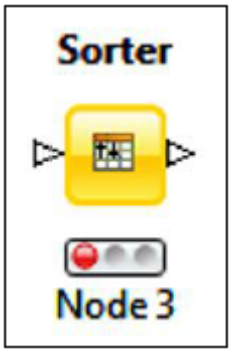


Drugie kliknięcie ponownie sortuje sprawy w kolejności malejącej.

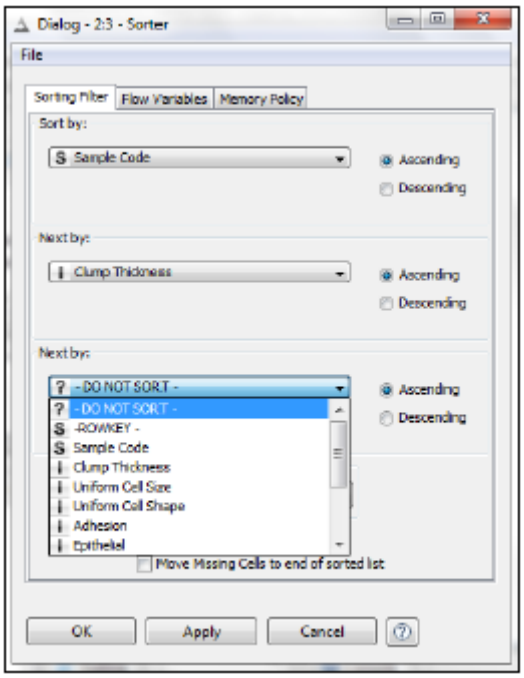


Przycisk Przywróć kolejność przykładów przywraca dane do oryginalnej kolejności.

✓ Specjalistyczne narzędzia do sortowania: Aplikacje do eksploracji danych często posiadają narzędzia do sortowania. Szukaj ich w menu do manipulacji danymi lub transformacji. W aplikacjach z wieloma narzędziami może być konieczne skorzystanie z wyszukiwania, aby znaleźć odpowiednie narzędzie; po prostu wyszukaj sortowanie. Rysunek przedstawia przykład; ta ikona reprezentuje funkcję wyszukiwania w procesie.

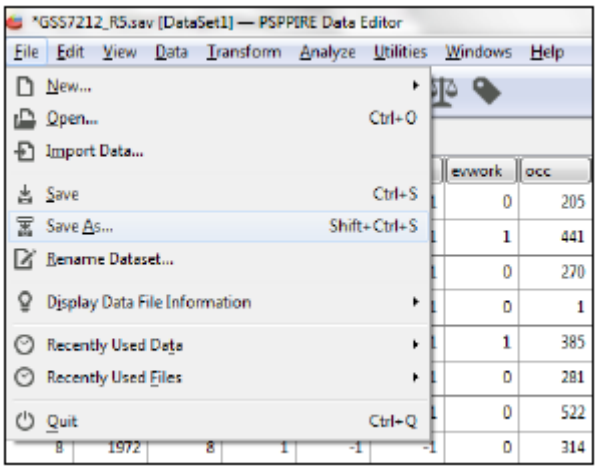


Narzędzia te często obsługują złożone schematy sortowania zagnieżdżonego obejmujące kilka zmiennych.



Umieszczanie danych tam, gdzie ich potrzebujesz

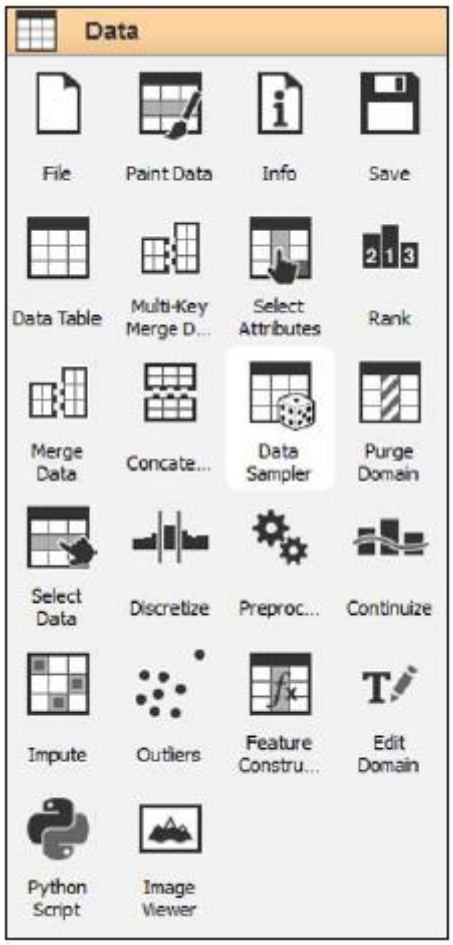
Po zdobyciu danych i wprowadzeniu zmian zgodnie z własnymi potrzebami musisz je gdzieś umieścić. Zapisałeś już dane za pomocą arkuszy kalkulacyjnych i innych aplikacji, więc wiedza zostanie przeniesiona do eksploracji danych, prawda? Cóż, nie do końca. W większości aplikacji możesz zapisywać dane, znajdując tę opcję w menu, tak jak pokazano na rysunku, i używając okna dialogowego przeglądarki plików, aby wskazać, gdzie chcesz umieścić plik.



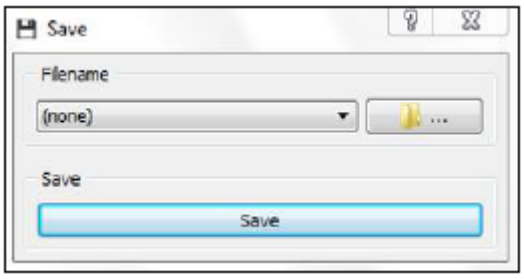
Jednak aplikacje do eksploracji danych z wizualnymi interfejsami programistycznymi nie są zorganizowane w ten sposób. W swoich aplikacjach do eksploracji danych będziesz musiał znaleźć odpowiednie narzędzie do zapisywania danych. Większość aplikacji do eksploracji danych ma kilka narzędzi do zapisywania danych, a ich nazwy różnią się w zależności od produktu, więc być może będziesz musiał trochę polować, aby znaleźć narzędzie, którego potrzebujesz. Oto, gdzie szukać eksportu danych w kilku aplikacjach do eksploracji danych. Poniższe przykłady dostarczają wskazówek dotyczących znajdowania funkcji eksportu danych w dowolnej aplikacji do eksploracji danych:

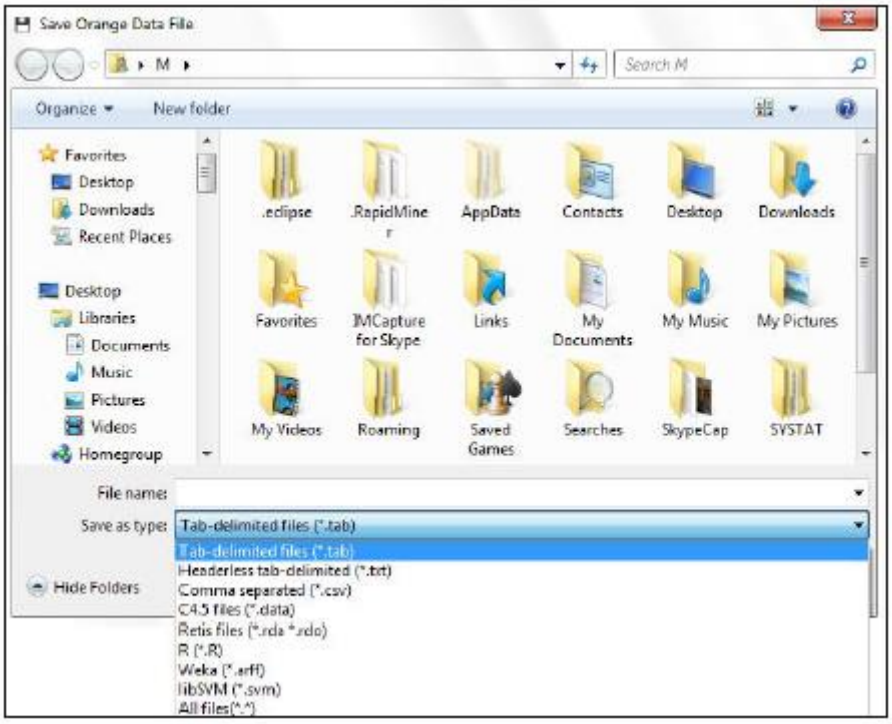
✓ KNIME: Narzędzia do eksportu danych znajdują się w Repozytorium węzłów pod IO . . . Napisz (patrz rysunek 14-16). Są one pogrupowane z narzędziami do eksportu modeli oraz danych.

✓ Pomarańczowy: masz tylko jedno narzędzie do zapisywania danych w tym produkcie i nazywa się ono Zapisz. Znajdź go w menu Widget w obszarze Dane na rysunku.

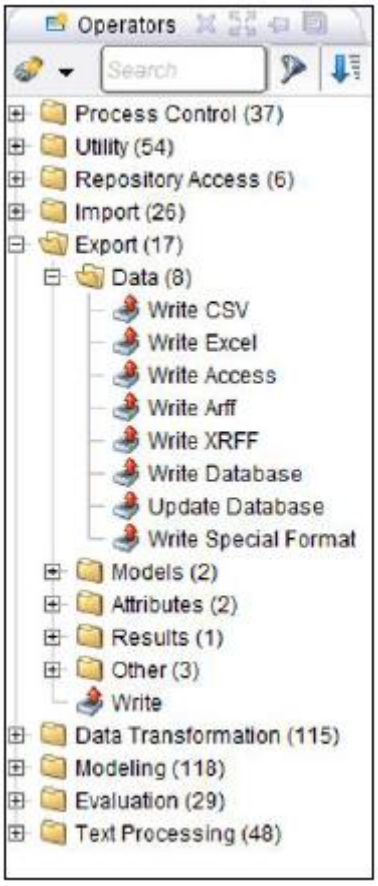


Samo okno dialogowe Zapisz jest wręcz minimalistyczne, ale kliknięcie przycisku Przeglądaj (ten z obrazem otwartego folderu) powoduje wyświetlenie konwencjonalnego okna dialogowego przeglądarki plików . Rozwijane menu zawiera opcje formatu eksportu danych.

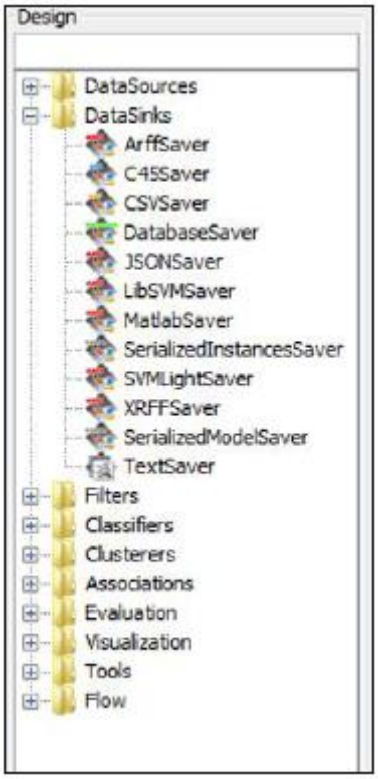




✓ RapidMiner: Narzędzia do eksportu danych znajdują się w menu Operatorzy pod opcją Eksport . . . Dane na rysunku. Dla każdej opcji formatu danych znajdziesz osobne narzędzie.



✓ Weka: Poszukaj eksportu danych w menu Design, pod DataSinks, gdzie wygrywa nagrodę za najbardziej geekową terminologię w aplikacji do eksploracji danych.

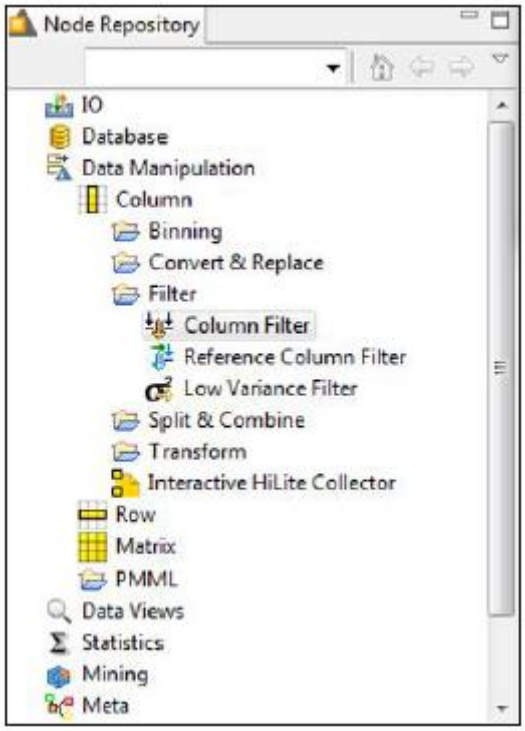


Odsiewanie potrzebnych danych

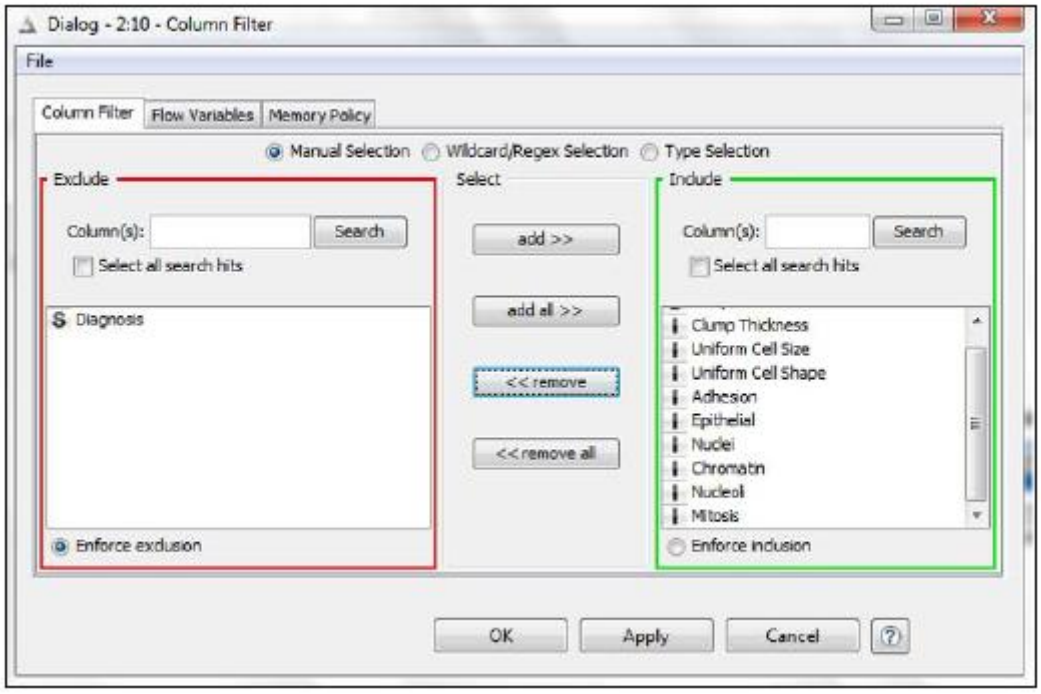
Czasami będziesz mieć więcej danych niż potrzebujesz do danego projektu. Oto, jak ograniczyć się do tego, czego potrzebujesz.

Zawężenie pól

Gdy w zbiorze danych znajduje się wiele zmiennych, znalezienie lub zobaczenie tych, które Cię interesują, może być trudne. A jeśli twoje zbiory danych są duże i nie potrzebujesz wszystkich zmiennych, przechowywanie dodatków niepotrzebnie pochłania zasoby. Tak więc czasami trzeba zachować niektóre zmienne i porzucić inne. Rysunek pokazuje przykład w KNIME, gdzie właściwe narzędzie nazywa się Filtrem Kolumnowym.



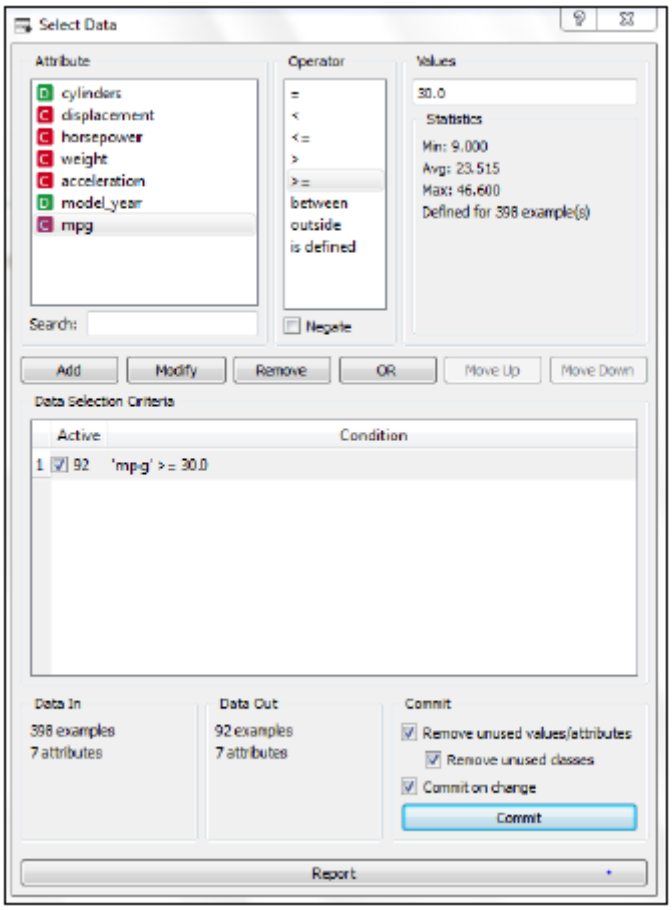
Przykład konfiguracji tego narzędzia pokazano na rysunku.



Aby zawęzić pola, poszukaj narzędzia do selekcji zmiennych w swojej aplikacji do eksploracji danych; można je znaleźć w innych narzędziach do manipulacji danymi. Podobnie jak w przypadku innych narzędzi do eksploracji danych, nazwy różnią się w zależności od produktu. Poszukaj odmian kolumny słów, zmiennej lub pola oraz selekcji lub filtrowania.

Wybór odpowiednich przypadków

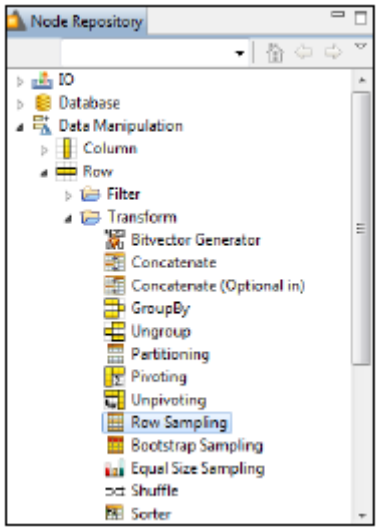
W rozdziale 2 widzieliście proces budowania bardzo prostego modelu predykcyjnego zmiany własności nieruchomości. W tym przykładzie przypadki z niekompletnymi danymi zostały odfiltrowane przed zbudowaniem modelu. Jednym z typowych przykładów selekcji lub filtrowania danych jest usuwanie niekompletnych spraw. Gdyby ten przykład został przeprowadzony szczegółowo w celu stworzenia użytecznego modelu, wymagałoby to znacznie większej selekcji danych. Prawie na pewno uzyskasz lepsze wyniki, jeśli weźmiesz pod uwagę tylko jedną klasę własności na raz. Oddzielisz na przykład nieruchomości mieszkalne od przemysłowych. Możesz porównać zachowanie lokalnych właścicieli do osób spoza miasta. Ale jak wybrać tylko odpowiednie przypadki dla każdego segmentu, który Cię interesuje? Użyłbyś narzędzia do selekcji danych. Zobacz rozdział 2, aby zobaczyć jeden przykład usuwania niekompletnych przypadków. Rysunek przedstawia narzędzie do selekcji danych w innej aplikacji do eksploracji danych, a rysunek drugipokazuje, jak skonfigurować to narzędzie do innego rodzaju selekcji, tym razem opartego na wartości zmiennej.



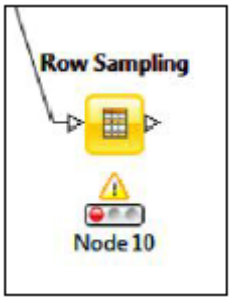
Powszechnie używa się tego rodzaju wyboru danych, a niektóre aplikacje udostępniają różnego rodzaju wbudowane funkcje, które pomogą Ci zdefiniować dokładnie takie przypadki, jakie chcesz. Ten ma kilka wyjątkowych cech; wyświetla statystyki podsumowujące dla zmiennej i dokładnie informuje, ile przypadków spełnia kryteria wyboru. Większość aplikacji do eksploracji danych ma narzędzia do wybierania tylko potrzebnych przypadków. Poszukaj w menu (lub wyszukaj), aby wybrać lub filtrować.

Próbkowanie

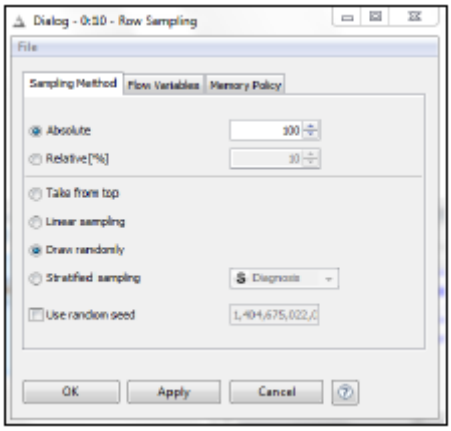
Popularnym obecnie pojęciem jest to, że więcej danych to lepsze dane. To nie jest nowy pomysł. Aplikacje do eksploracji danych zawsze były tworzone do pracy z dużymi ilościami danych. Nawet nazwa „eksploracja danych” sugeruje duże ilości. Często jednak praca z próbką danych dostarcza informacji, które są równie przydatne, ułatwiają pracę oraz oszczędzają czas i zasoby. Próbkowanie odgrywa ważną rolę w eksploracji danych. Dwa z nich opisano w Części 2. Po pierwsze, dane są zrównoważone. Oznacza to, że model używał równej liczby przypadków w każdej z porównywanych grup (w tym przykładzie grupy były właściwościami, które zmieniły właściciela, a właściwościami, które nie zmieniły), mimo że jedna grupa miała znacznie więcej przypadków niż druga w oryginale dane. Później dane zostały podzielone, podzielone na jeden podzbiór do wykorzystania do trenowania modelu, a drugi do testowania. W Części 13 widziałeś, jak użycie tylko próbki danych na równoległym wykresie współrzędnych może ułatwić przeglądanie i interpretację. (Wykresy punktowe z tysiącami punktów mogą być niemożliwie trudne do odczytania!) Być może najważniejsze jest to, że próbkowanie po prostu zmniejsza ilość danych, dzięki czemu wszystko działa szybciej. Nie zawsze jest oczywiste, gdzie znaleźć narzędzia do pobierania próbek w aplikacjach do eksploracji danych. Poszukaj nazw, które są wariacjami na temat próbki lub podziału słów. Rysunek pokazuje jeden przykład, w którym trzy warstwy zostały ukryte w menu w KNIME.



Znalezienie tego może być najtrudniejsze. Samo narzędzie wygląda podobnie do innych i działa jak każde inne.



Konfiguracja może wyglądać nieco inaczej w zależności od aplikacji, ale wszystkie umożliwiają pobieranie prostych losowych próbek z danych.



Zebranie danych razem

Gdy Twoje dane znajdują się w więcej niż jednym miejscu, potrzebujesz sposobów na umieszczenie ich wszystkich razem.

Scalanie

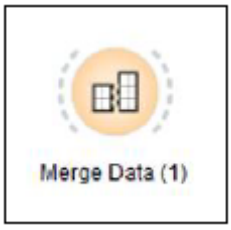
Łącząc dwa zbiory danych z różnymi zmiennymi, łączysz dane. Scalanie to powszechna operacja. Widziałeś przykład łączenia danych w Części 2, kiedy zbiór danych zawierający ogólne informacje o działkach nieruchomości został połączony z innym zawierającym informacje o tym, które nieruchomości zmieniły właściciela (a które nie). Scalanie jest często używane w eksploracji danych, łącząc połączone dane, takie jak

✓ Ewidencja klientów i dane kampanii marketingowych

✓ Przed i po wynikach badań

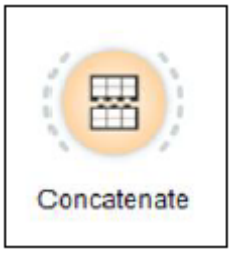
✓ Dane wewnętrzne i dostawcy

Aby scalić zestawy danych, musisz mieć zmienną, która identyfikuje przypadki do dopasowania; nazywa się to kluczem lub zmienną identyfikatora. Być może będziesz musiał określić jeden z zestawów danych jako podstawowy; tabela podstawowa musi mieć tylko jedną wielkość liter dla dowolnej wartości zmiennej kluczowej. Niektóre aplikacje do eksploracji danych mają więcej niż jedno narzędzie do łączenia zestawów danych: Rysunek pierwszy przedstawia narzędzie do podstawowych połączeń, a Rysunek drugi przedstawia narzędzie do konfigurowania bardziej złożonych kryteriów scalania.



Dołączanie

Jeśli źródła danych zawierają te same zmienne (mniej więcej; dopasowanie nie musi być identyczne), ale różne przypadki, łączenie ich nazywa się dołączaniem lub konkatenacją. Podobnie jak scalanie, jest to powszechna operacja. Jest używany, gdy masz nowe sprawy dotyczące czegoś, co już śledziłeś.



Trudną częścią znalezienia odpowiedniego narzędzia jest często ustalenie, jak się nazywa. Poszukaj w menu (lub wyszukaj) dołączania, łączenia lub scalania wierszy.

Tworzenie nowych danych ze starych danych

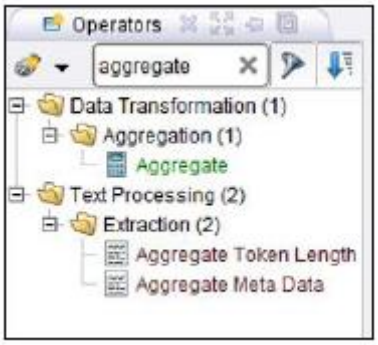
Poniższe sekcje dotyczą wykorzystania wiedzy biznesowej do tworzenia odpowiednich nowych konstrukcji danych z istniejących danych.

Wyprowadzanie nowych zmiennych

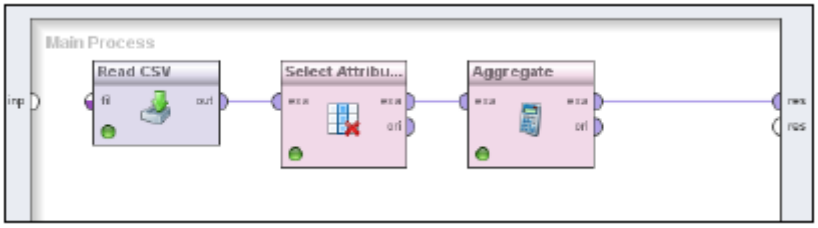
Chcesz dodać zmienną A plus zmienną B, aby utworzyć zmienną C? Wyprowadzasz nową zmienną. Obliczenia, których potrzebujesz, aby utworzyć nową zmienną, mogą być o wiele bardziej złożone niż zwykłe dodanie dwóch kolumn, ale to jest w porządku. Możesz użyć specjalnych funkcji, aby pomóc w tym procesie: dostępne są standardowe funkcje matematyczne, funkcje statystyczne, manipulacja ciągami i inne. Zapoznaj się z Częścią 2, aby zapoznać się z przykładem manipulacji ciągami, w którym skróciłeś kody pocztowe +4 do wersji pięcioznakowych. Gdy masz zmienną ciągłą i musisz pogrupować wartości (takie jak grupując wyniki testu w przedziały centylowe lub wyniki modelu w decyle), potrzebna funkcja to binning. Możesz go znaleźć w tym samym narzędziu, co inne podobne funkcje lub w specjalnym narzędziu zaprojektowanym specjalnie do tego celu.

Zbiór

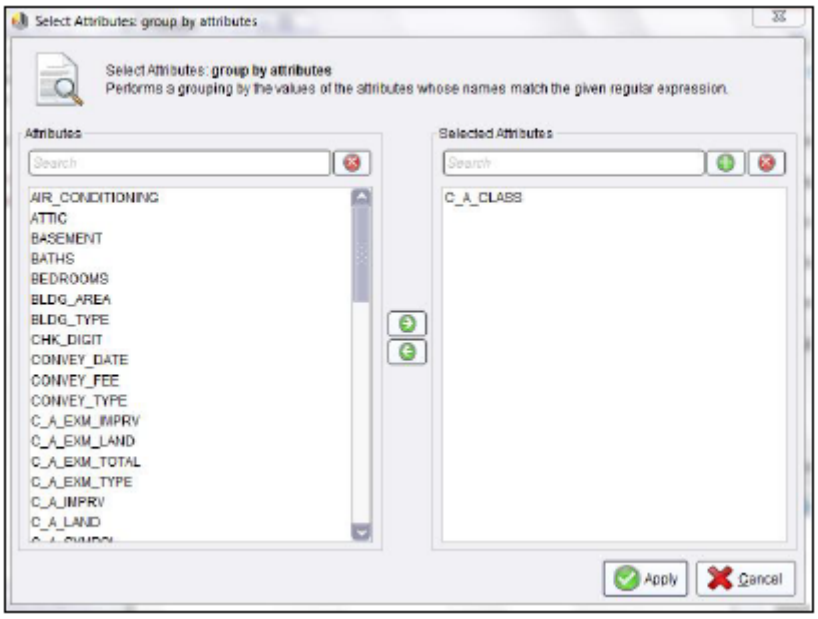
Podsumowywanie danych, znajdowanie sum, obliczanie średnich i innych miar opisowych prawdopodobnie nie jest dla Ciebie niczym nowym. Gdy potrzebujesz podsumowań w postaci nowych danych, a nie raportów, proces ten nazywa się agregacją. Zagregowane dane mogą stać się podstawą do dodatkowych obliczeń, połączone z innymi zbiorami danych, wykorzystywanymi w jakikolwiek sposób, w jaki wykorzystywane są inne dane. Oto przykład procesu agregacji danych. Zbiór danych (wykorzystywany również w przykładach przedstawionych w rozdziale 2) zawiera ogólne informacje o ponad 160 000 działek nieruchomości. Dane te obejmują różne sposoby użytkowania gruntów. Co zrobić, jeśli chcesz zobaczyć średnią oszacowaną wartość gruntu w każdej kategorii użytkowania gruntów? Oto, jak to zrobisz. Narzędzie do agregacji danych znajdziesz w swojej aplikacji do eksploracji danych. Możesz użyć wyszukiwania, aby go znaleźć, jak pokazano na rysunku.



Możesz dodać narzędzie do procesu i połączyć je ze źródłowym zbiorem danych.



W narzędziu do agregacji danych należy wybrać zmienną grupującą. W tym przypadku jest to zmienna Land Use C_A_CLASS.



Następnie możesz zdefiniować podsumowania, które chcesz. Aby uzyskać średnią oszacowaną wartość gruntu, wybierz zmienną z ocenami do podsumowania i wybierz średnią funkcję.

Kiedy agregacja jest wykonywana, wynikiem jest nowy zestaw danych z jednym wierszem dla każdego rodzaju użytkowania gruntów i nową zmienną dla obliczonych średnich.

ExampleSet (7 examples, 0 special attributes, 2 regular attributes)

Row No.	C_A_CLASS	average(C_A_LAND)
1	1	17769.491
2	2	59417.886
3	3	223198.540
4	4	266981.687
5	5	10903.431
6	7	69426.847
7	9	0

Wcześniej czy później będziesz musiał zagregować cały zbiór danych. Ale kiedy chcesz zsumować lub uśrednić wszystkie dane w zbiorze danych, możesz napotkać problem: Jaka jest twoja zmienna grupująca? Sztuką jest użycie zmiennej ze stałą wartość dla całego zbioru danych. Dlatego utwórz zmienną, w której każda wartość jest taka sama, a następnie użyj jej jako zmiennej grupującej.

Oszczędzanie czasu

Przygotowanie danych zajmuje dużo czasu. Twoja aplikacja do eksploracji danych może mieć już wbudowane i czekające na Ciebie charakterystyczne narzędzia oszczędzające czas, ale nie dowiesz się o tym, jeśli nie sprawdzisz, co tam jest. Spraw, aby proces eksploracji danych był łatwiejszy i szybszy, szukając i używając specjalnych narzędzi do takich celów jak:

✓ Zaadresowanie brakujących wartości: Jeśli dane nie są kompletne, może być konieczne usunięcie niekompletnych przypadków lub dokonanie inteligentnych podstawień pustych miejsc. (Jeśli właściwe narzędzie nie jest nazywane czymś oczywistym, np. brakiem wartości, poszukaj imputacji).

✓ Ważenie: Potrzebujesz niektórych przypadków (lub pól), aby liczyć więcej niż inne? Poszukaj ważenia.

✓ Podziel plik: musisz wykonać te same czynności dla każdej wartości zmiennej? Poszukaj funkcji podziału pliku lub funkcji pętli.

Twoja ekscytująca kariera w modelingu

Teraz naprawdę docierasz do gadżetów związanych z eksploracją danych. Modelowanie jest ścieżką eksploratora danych do poznania nieznanego lub przynajmniej dokonania dobrze poinformowanych, wartościowych domysłów na temat nieznanego. Eksperci danych są znani z tego, że skutecznie przewidują, którzy nowi klienci staną się dużymi wydatkami, jakie warunki procesu doprowadzą do poważnych problemów produkcyjnych, a które roszczenia ubezpieczeniowe wskazują na oszustwo. Możesz zostać kolejnym eksploratorem danych, który zajrzy w nieznane i dokona ekscytujących odkryć. Nie będziesz potrzebować mocy psychicznych ani kryształowej kuli, ale będziesz potrzebować dobrej znajomości podstaw eksploracji danych, takich jak drzewa decyzyjne, sieci neuronowe i klastry. Ta część wprowadzi Cię w te podstawowe i pomoże Ci na drodze do kariery w modelowaniu danych! Jest to część pracy, którą eksploratorzy danych często uwielbiają najbardziej, ponieważ to tutaj odkryjesz informacje, które pomagają podejmować dobre decyzje biznesowe.

Pojmowanie koncepcji modelowania

Jako eksplorator danych szukasz przydatnych wzorców w danych. Twoim celem jest odkrycie czegoś, co zdarzyło się wiele razy w przeszłości i co do których możesz rozsądnie oczekiwać, że powtórzy się w przyszłości. Robisz to już nieformalnie w swoim codziennym życiu. Życie jest pełne małych wzorów. Bez wątpienia zauważyłeś, że szare niebo i wiatr to znaki, że za chwilę zacznie padać. To jest wzór. Być może odkryłeś, że kiedy Twoje dziecko jest marudne i ma gorączkę, prawdopodobnie ma infekcję ucha. To też jest wzór. Jeśli twój szef chodzi na kręgle w każdy wtorek i przychodzi do pracy do późna w każdą środę, cóż, to też jest wzór. Modelowanie umożliwia:

✓ Odkrywaj wzorce, które są zbyt subtelne, aby je zidentyfikować poprzez nieformalną obserwację

✓ Podaj jednoznaczne opisy znalezionych wzorów

✓ Zastosuj wzorce, aby spójnie przewidywać

Przez całe życie nieformalnie wykorzystywałeś swoje obserwacje wzorców, aby przewidzieć, co wydarzy się w przyszłości, lub zgadnąć na temat faktów, których nie znasz na pewno. Te przewidywania kierują Twoimi działaniami. Kiedy niebo ciemnieje i wzmaga się wiatr, wchodzisz do domu. Kiedy dziecko robi się marudne i ciepłe, idziesz do lekarza. We wtorki sprawdzasz, czy szef przyniósł mu buty do kręgli. Jeśli tak, śpisz dodatkowe pół godziny w środę rano. Jako eksplorator danych robisz te same rzeczy, ale korzystasz z matematyki i komputerów, dzięki czemu możesz dokonywać większej liczby obserwacji i przewidywań. Ponieważ stajesz się eksploratorem danych, a nie statystykem czy profesorem, nie będziesz zajmować się teorią, aby to zrobić. Proces eksploracji danych może być znacznie bardziej formalny niż codzienne obserwacje i decyzje, ale nadal jest znacznie mniej formalny niż klasyczne statystyki i badania akademickie. W południe wychodzisz na zewnątrz. Jest ciemno i wietrznie. Otwierasz parasol. Właśnie użyłeś wzorca (dzień, ciemność, wietrznie), aby dokonać prognozy (deszcz) i użyłeś tej prognozy jako podstawy decyzji (użyj parasola). Znaczące jest to, że nie czułeś się zmuszony do identyfikowania mechanizmów, które powodują, że deszcz używa wzoru. Chociaż powinieneś szukać głębszego zrozumienia, kiedy możesz je uzyskać, nadal możesz używać wzorców, które uznasz za przydatne, nawet jeśli nie wiesz, dlaczego działają. Tak działają również eksploratorzy danych. Każda technika analizy, która wykorzystuje dane i matematykę do opracowania dobrze zdefiniowanych, spójnych reguł dokonywania prognoz, jest formą analizy predykcyjnej. Obejmuje to eksplorację danych, a także podejścia, takie jak klasyczne statystyki i badania operacyjne. Wszystkie te reguły, bez względu na to, czy są bardzo proste i intuicyjne do zrozumienia, czy też rozległe, niezrozumiałe sieci równań, są modelami. Tak więc na najprostszym poziomie model matematyczny jest maszyną do przewidywania. Używam szeroko pojętego przewidywania. W swojej

pracy możesz również spotkać się z terminem estymacja, który jest często używany w ten sam sposób, lub terminem klasyfikacja, który jest przewidywaniem czegoś, co jest kategoryczne, np. czy klient kupi czy nie, zagłosuj na kandydata A lub B albo wybierz plan Budżetowy, Standardowy lub Rodzinny dla usług telefonii komórkowej.

Kultywowanie drzew decyzyjnych

Drzewa decyzyjne są najpotężniejszym i najbardziej użytecznym narzędziem w zestawie narzędzi eksploratora danych. Możesz ich użyć do

✓ Zidentyfikowania najskuteczniejszych czynników do prognozowania

✓ Odkryciz kombinacje czynników, które definiują ważne segmenty, których inaczej mogłeś nigdy nie zauważyć

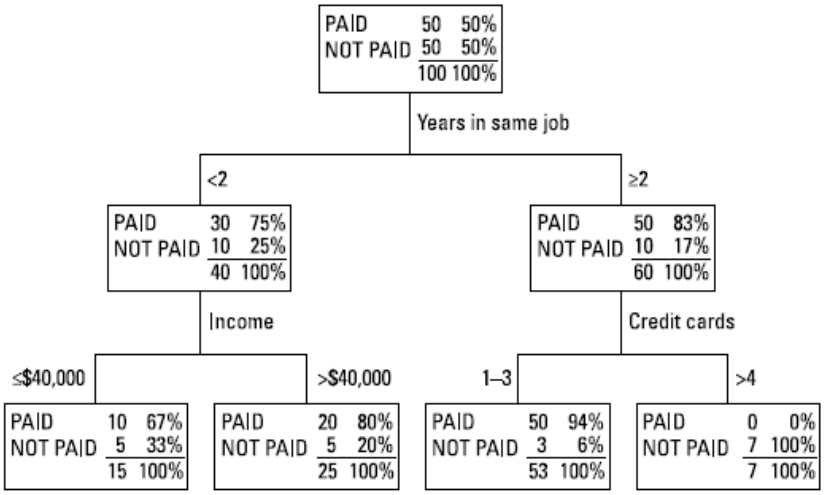
✓ Opracuj proste, zrozumiałe grafiki, które wyjaśnią Twoje wnioski menedżerom i klientom

✓ Twórz reguły, które są łatwe do zrozumienia i użytkowania

Przez długi czas możesz obejść się bez żadnej innej techniki modelowania, ale żaden eksplorator danych nie może obejść się bez drzew decyzyjnych.

Badanie drzewa decyzyjnego

Oto typowe wyzwanie związane z przewidywaniami: dana osoba ubiega się o pożyczkę. W jaki sposób pożyczkodawca może ustalić, czy wniosek o pożyczkę powinien zostać zatwierdzony? Dawniej pożyczkodawcy podejmowali decyzje kredytowe w oparciu o osobisty osąd i intuicję, ale te czasy już minęły. Teraz używają spójnych reguł opartych na danych, aby decydować, które wnioski kredytowe mają zostać zatwierdzone. Jak więc określić, które czynniki oddzielają wnioski kredytowe, które prawdopodobnie zostaną spłacone, od tych, które nie są spłacane? Możesz pobrać ewidencję poprzednich pożyczek i wykorzystać dane z aplikacji jako dane do analizy. Wiedziałbyś, które kredyty zostały spłacone w całości, a które nie, i wiedziałbyś dużo o kredytobiorcach, takich jak kwota i źródła ich dochodów, historia zatrudnienia i szczegóły ich historii kredytowej. Możesz mieć informacje o dziesiątkach czynników, które mają potencjał jako predyktory zdolności kredytowej. Drzewo decyzyjne wyprowadzone z tych danych może wyglądać tak, jak w przykładzie pokazanym na rysunku.



Na górze możesz zobaczyć podsumowanie wszystkich danych użytych do stworzenia drzewa. W tym przykładzie przeanalizowano łącznie 100 pożyczek — 50 spłacanych i 50 niespłacanych. Może to zabrzmieć dziwnie, ponieważ 50 procent nieopłacanych pieniędzy wydaje się strasznie nieopłacalne. Ale dane, których używasz do rozwijania swojego drzewa decyzyjnego, nie muszą mieć takich samych proporcji spłat i niespłaconych pożyczek, jak w prawdziwym życiu. Celem jest tutaj odróżnienie wzorców wskazujących na pożyczkę, która nie zostanie spłacona, od takiej, która zostanie spłacona. Tak więc powszechną praktyką eksploracji danych jest używanie równej liczby przypadków w każdej z grup, które chcesz porównać. Zasadniczo przypisujesz równą wagę wzorcom danych w każdej grupie. Następnie gałęzie drzewa dzielą się na dwie podgrupy w oparciu o czas, przez jaki pożyczkobiorca pozostaje w jednym zadaniu. Dlaczego uwzględniać ten czynnik, a nie inne, które były dostępne w dokumentacji kredytowej? Jest to jedyny czynnik, który ma najsilniejszy wpływ na rozdzielenie tych dwóch grup, zgodnie z algorytmami matematycznymi użytymi do skonstruowania drzewa. Podział doprecyzowuje dane, definiując nowe grupy o większej lub mniejszej częstotliwości spłaty. Grupy mogą się wielokrotnie dzielić, a zmienne definiujące podziały drugiego poziomu lub kolejne podziały mogą nie być takie same w całym drzewie. W tym przypadku drugi podział opiera się na dochodach jednej grupy i liczbie kart kredytowych dla drugiej. Każdy podział zawiera nieco głębszą definicję tego, które pożyczki są dobrymi zakładami, a które nie.

Wykorzystanie drzew decyzyjnych do wspomagania komunikacji

Interpretacja diagramu drzewa decyzyjnego na rysunku powyżej nie wymaga od naukowca rakietowego. Nie widzisz żadnych równań, żadnych tajemniczych statystyk i nic skomplikowanego do interpretacji. Może ci się to nie podobać, ale jego prostota jest piękna! Dlatego drzewa decyzyjne są tak ważne dla eksploracji danych. Kiedy używasz drzew decyzyjnych do predykcji, nie zawsze będziesz wiedzieć dokładnie, dlaczego określone czynniki są ważne, ale na pewno dowiesz się, które czynniki są ważne i w jakich kombinacjach. A diagramy drzew stanowią proste narzędzie, które możesz zaprezentować każdemu, aby pomóc w wyjaśnieniu tych ustaleń. Istnieje wiele różnych sposobów budowania i wyświetlania drzew decyzyjnych. Chociaż ten przykład ma tylko dwukierunkowe rozgałęzienia (nazywane również podziałami binarnymi), istnieją drzewa decyzyjne z wielokierunkowymi rozgałęzieniami. Będziesz mieć opcje kombinacji używanych zmiennych i sposobu ich obsługi. Możesz ozdobić wyświetlacz małymi wykresami słupkowymi. (Niektóre oferują również małe wykresy kołowe, ale wykresy kołowe nie są dobrym wyborem do eksploracji danych. Więcej informacji na temat wykresów kołowych można znaleźć w Części 13). Elastyczność jest również kluczem do użyteczności. Można znaleźć wiele innych typów technik modelowania i analizy danych, z których każda ma swoją unikalną wartość, ale żadna nie ułatwia prezentowania i wyjaśniania wyników analizy predykcyjnej szerokiemu gronu odbiorców tak łatwo, jak drzewa decyzyjne.

Konstruowanie drzewa decyzyjnego

Przeprowadzę Cię przez proces tworzenia przykładowego drzewa decyzyjnego. Dr William H. Wolberg, lekarz badający przyczyny raka piersi, zbadał wiele guzów piersi i odnotował szereg czynników, a także swoją diagnozę dla każdego z nich. Jaki jest potencjał wykorzystania niektórych lub wszystkich tych czynników do automatyzacji lub wspomagania procesu diagnozy? W rzeczywistych aplikacjach powinieneś być dobrze zaznajomiony ze znaczeniem zmiennych w twoich danych lub współpracować z kimś, kto jest. Odnalezienie wyjaśnienia, którego potrzebujesz, może wymagać dużo pracy, ale jest to absolutna konieczność w pracy z eksploracją danych. W tym przykładzie praktycznym i większości przykładów praktycznych opartych na publicznie udostępnionych zestawach danych wiele informacji (nawet informacji, które nie są prywatne) nie jest uwzględnionych w zestawie danych. W niektórych przypadkach więcej informacji można znaleźć w opublikowanych pracach badawczych, które

wykorzystują te dane, ale nie zawsze tak będzie, a nie wszystkie prace naukowe są łatwe do uzyskania lub zrozumienia.

Zrozumienie danych

Dr William H. Wolberg z University of Wisconsin Hospitals w Madison zarejestrował następujące cechy kilkuset guzów:

✓ Grubość grudek

✓ Jednolitość wielkości komórek

✓ Jednorodność kształtu komórki

✓ Przyczepność brzeżna

✓ Rozmiar pojedynczej komórki nabłonkowej

✓ Gołe jądra

✓ Bezbarwna chromatyna

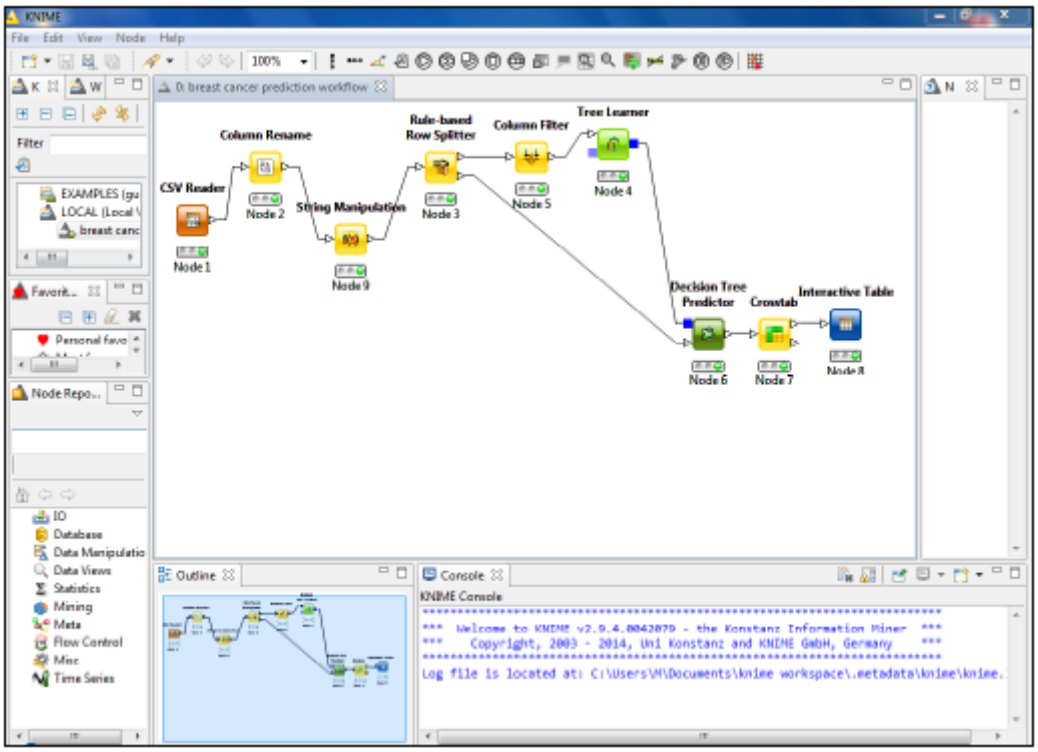
✓ Normalne jąderka

✓ Mitozy

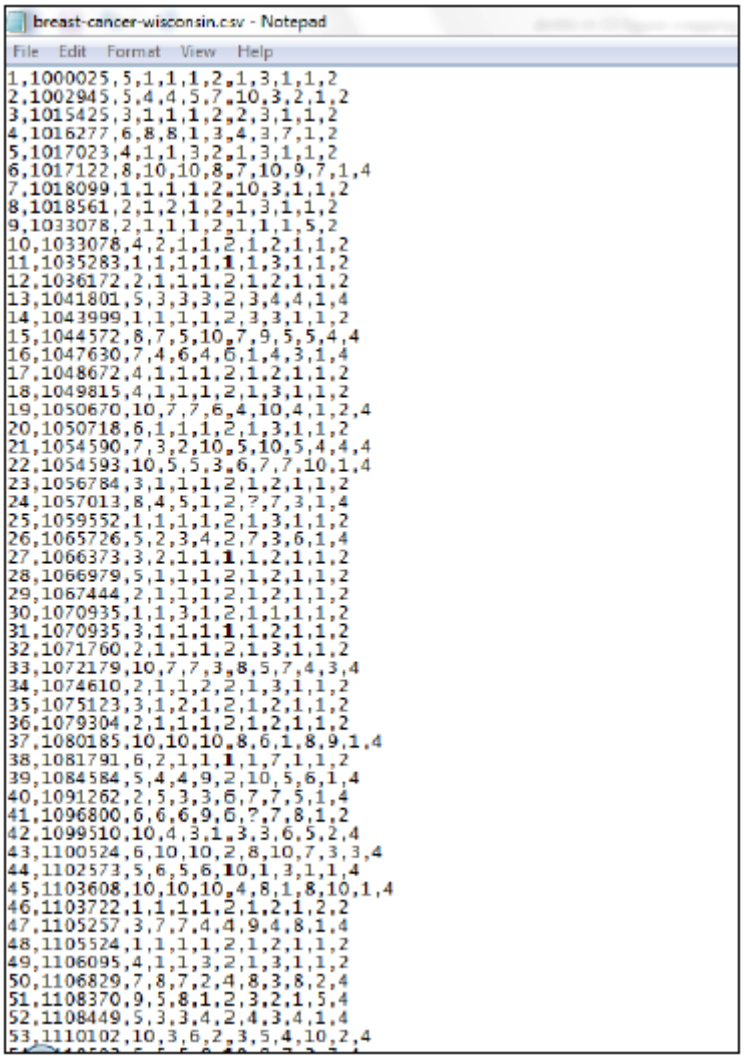
Każdy z nich został zapisany w skali od 1 do 10. (Niestety dla nas, ocumentacja nie wyjaśnia dokładnie, czym są te atrybuty ani w jaki sposób zdefiniowane są skale). 4 na złośliwe.

Przeglądanie przepływu pracy

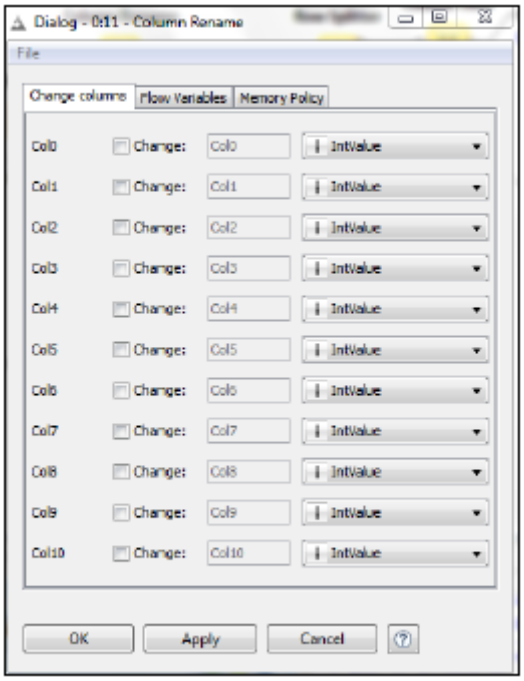
Jeden ze sposobów rozwiązania tego problemu pokazano na rysunku, który przedstawia przykładowy proces tworzenia drzewa decyzyjnego dla danych diagnozy raka piersi. (W tym przykładzie użyto platformy do eksploracji danych KNIME). Co dzieje się w każdym z tych kroków?

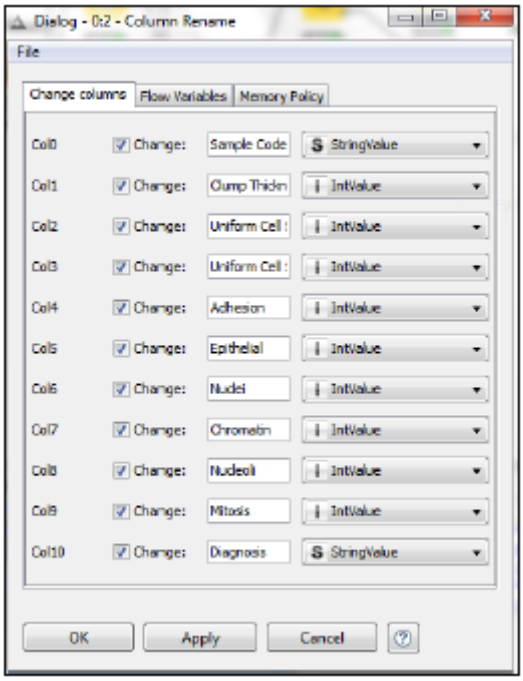


✓ Import danych: Dane znajdują się w pliku tekstowym w formacie zmiennej rozdzielanej przecinkami (.csv). Możesz zobaczyć, jak wyglądają dane na rysunku.

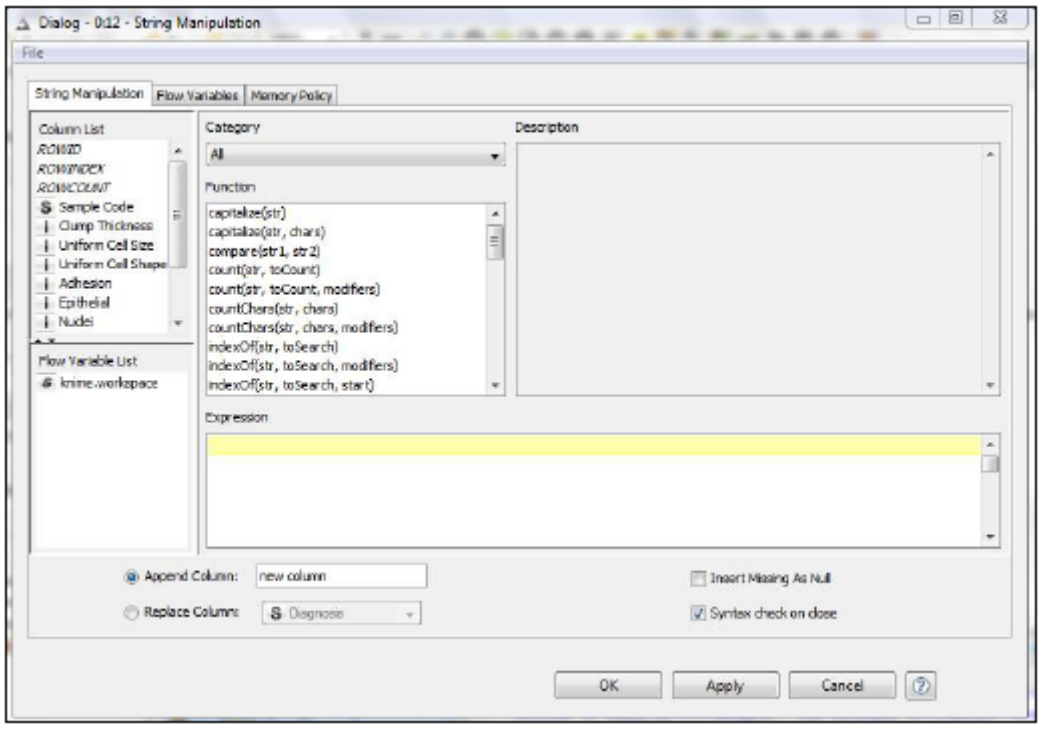


✓ Etykietowanie zmiennych: W pliku danych nie ma etykiet zmiennych, więc są one wpisywane ręcznie, przy użyciu informacji z oddzielnego pliku tekstowego zawierającego nazwy zmiennych i inną dokumentację. Zobacz ten węzeł przed i po oznaczeniu na rysunkach.

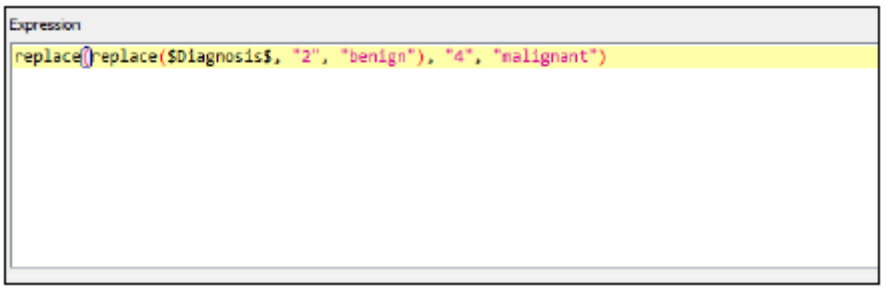




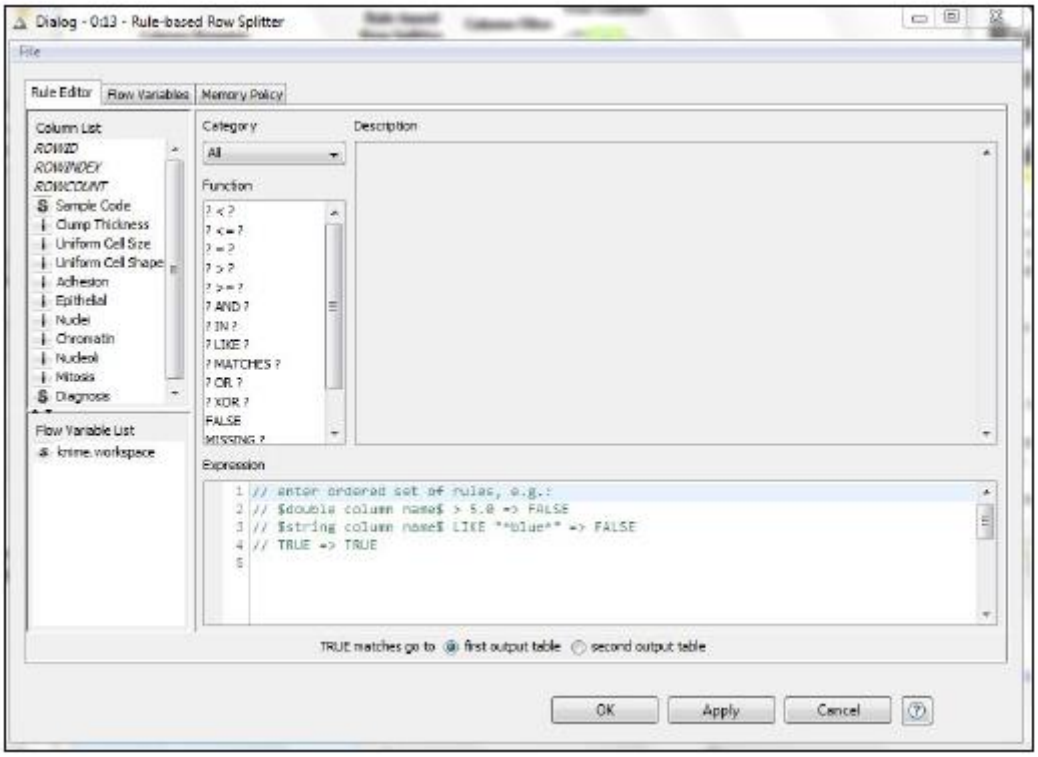
✓ Przygotowanie zmiennej do analizy: Narzędzie drzewa decyzyjnego użyte w tym przykładzie wymaga nominalnej zmiennej zależnej. Kody diagnostyczne są nominalne, ale ponieważ są to kody 2 i 4, oprogramowanie uważa, że Diagnoza jest liczbą. Tak więc kody muszą zostać przekonwertowane na ciągi. Bonus: i tak łatwiej jest zrozumieć słowa. Zobacz, jak wygląda to okno dialogowe na rysunku

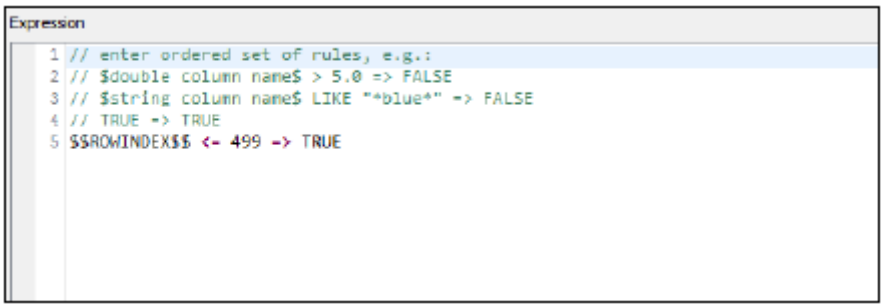


a instrukcje konwersji na rysunku.

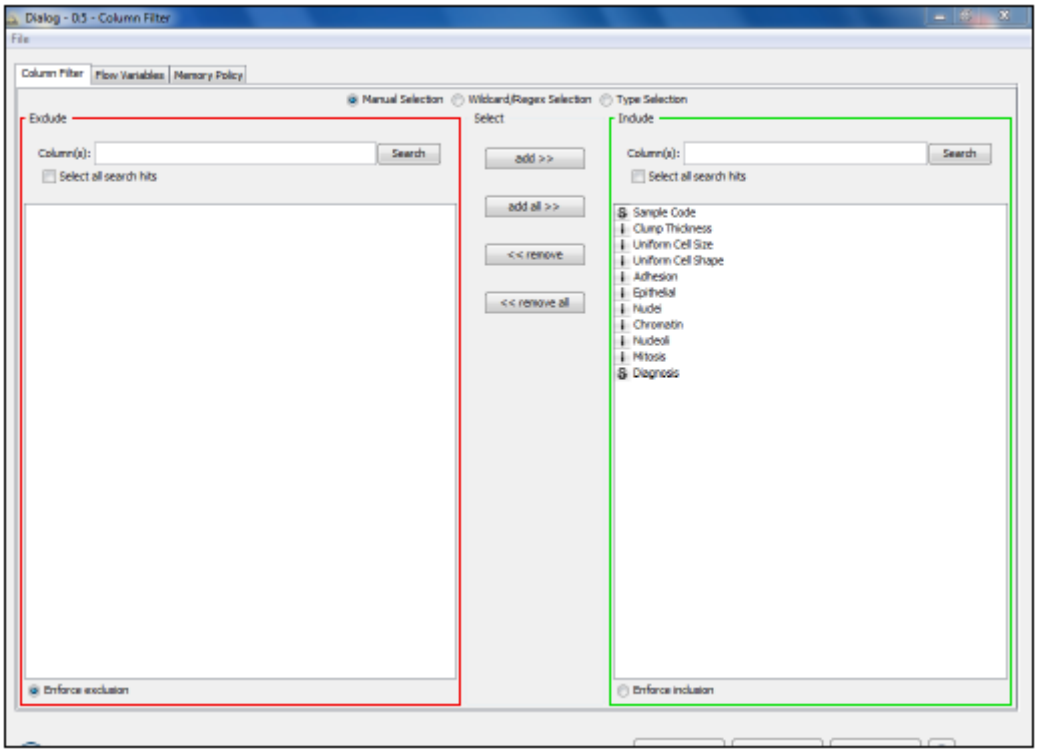


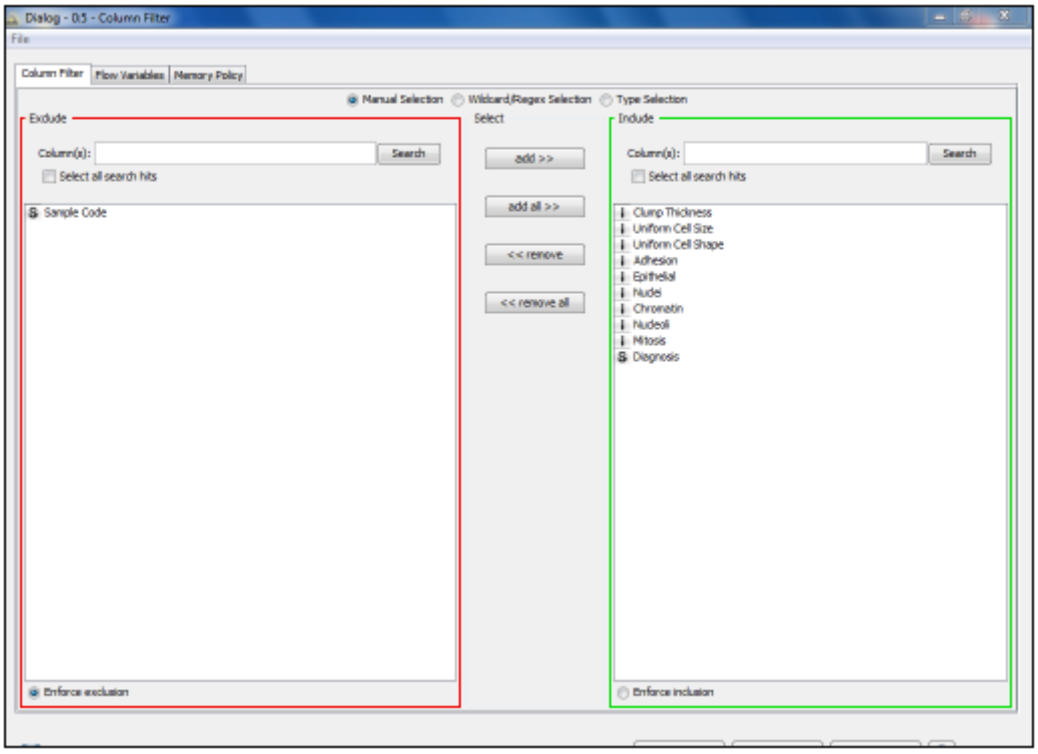
✓ Oddzielanie przypadków do uczenia się od przypadków do testowania: Użyj niektórych danych do utworzenia drzewa i zapisz niektóre, aby sprawdzić, jak dobrze algorytm predykcji działa na świeżych danych. (To wymaga przemyślenia. Czy masz jakieś specjalne uwagi dotyczące kolejności przypadków? Na przykład, jeśli plik danych zawiera wszystkie zdiagnozowane przypadki łagodne zgrupowane razem na początku i te zdiagnozowane jako złośliwe na końcu, potrzebujesz aby podjąć pewne kroki, aby upewnić się, że dane treningowe i testowe zawierały oba typy przypadków). Zobacz ten węzeł przed i po zdefiniowaniu podziału na rysunkach.



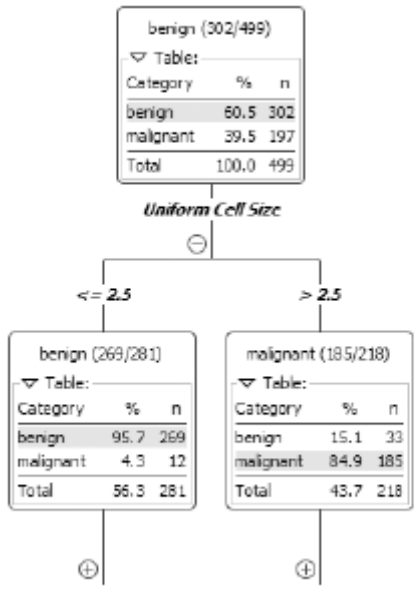


✓ Wykluczenie zmiennej: Jedną ze zmiennych jest kod, którego nie należy używać do przewidywania. Zobacz ten węzeł przed i po wykluczeniu tej zmiennej na rysunkach.

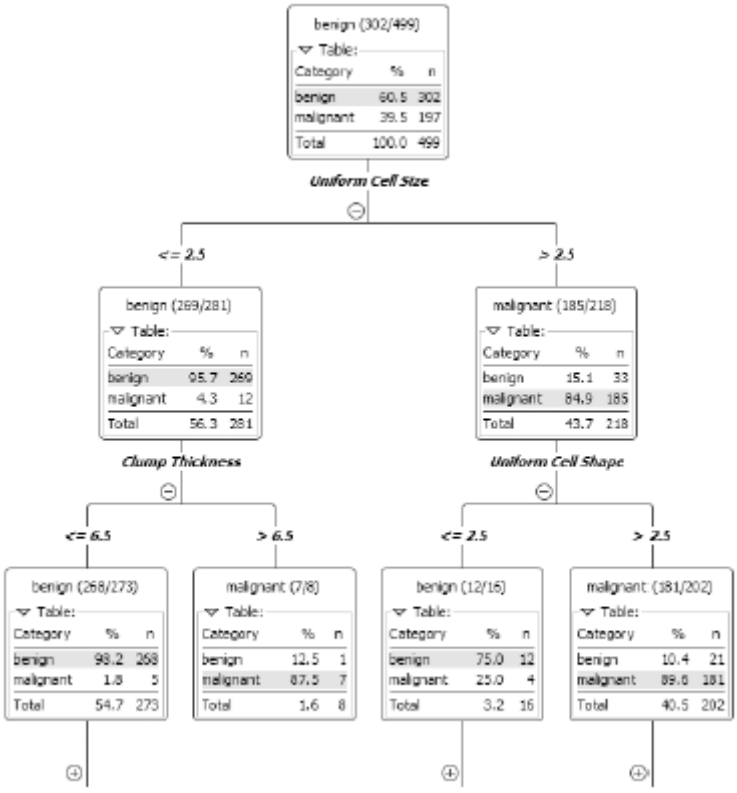




✓ Tworzenie drzewa: Po odpowiednim przygotowaniu danych można podłączyć narzędzie do drzewa decyzyjnego i uruchomić je, aby utworzyć drzewo i je wyświetlić. Chociaż większość narzędzi oferuje szereg ustawień, które możesz zmienić, zwykle możesz utworzyć swoje pierwsze drzewo bez dostosowywania (lub nawet oglądania) żadnego z nich. Rysunek przedstawia drzewo w takiej postaci, w jakiej jest wyświetlane na początku, z tylko jednym podziałem.

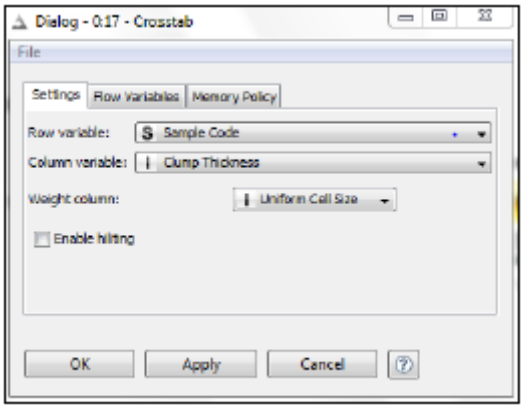


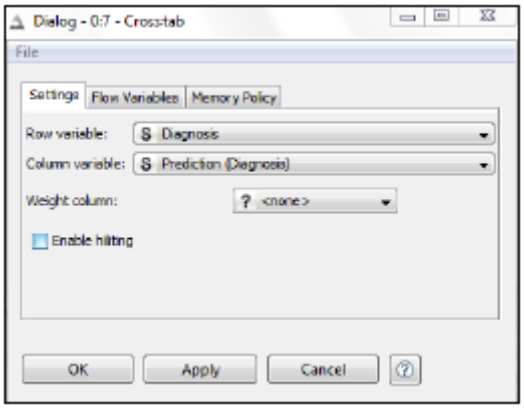
Kliknięcie znaku plusa rozwija tę część drzewa. Powiększony widok pokazano na rysunku.



✓ Przewidywanie nowych danych: Drzewo wyglądało dobrze, ale czy te same zasady będą działać w przypadku danych, które nie zostały użyte do stworzenia drzewa?

✓ Porównywanie przewidywań z rzeczywistością: Jest to dopasowanie przewidywania do rzeczywistych wyników. Zobacz widoki przed i po, aby ustawić tabelę przestawną na rysunkach.





✓ Badanie wyników: Do wyświetlenia wyników wymagany jest jeszcze jeden krok. Zobacz wyniki na rysunku.

Row ID	Diagnosis	Predicti...	Freque...	Expected	Deviation	Percent	Row Pe...	Colum
Row0	benign	benign	154	122.46	31.54	77	98.718	98.089
Row1	benign	malignant	2	33.54	-31.54	1	1.282	4.651
Row2	malignant	benign	3	34.54	-31.54	1.5	6.818	1.911
Row3	malignant	malignant	41	9.46	31.54	20.5	93.182	95.349

Trzecie prawo eksploracji danych mówi, że większość wysiłku w każdym projekcie eksploracji danych należy do przygotowań. Nawet w tym przykładzie, który jest prosty w porównaniu z większością projektów, wiele kroków jest wymaganych, zanim dane będą gotowe do wprowadzenia do modelowania. Większość z nich wiąże się z dostosowaniem formatu danych z tego, co masz, do tego, do czego nadaje się narzędzie do modelowania. Zbadanie drzewa ujawnia kilka przydatnych informacji. Pierwsza gałąź wskazuje, że jednolitość wielkości komórek jest najpotężniejszym czynnikiem różnicującym nowotwory łagodne od złośliwych i najlepszą wartością danych do dokonania tego podziału. Rozwinięcie drzewa ujawnia więcej czynników, interakcji i ponownie właściwe wartości dla podziałów. Jeśli te same reguły podziału zostaną zastosowane do danych, które nie zostały użyte do zbudowania drzewa, czy przewidywania będą prawidłowe? Tabela przestawna przedstawia prognozy w porównaniu z rzeczywistymi diagnozami dla danych wstrzymania. Jednak tabela przestawna nie jest łatwa do odczytania, ponieważ jest wyświetlana przez oprogramowanie. Tabela pokazuje te same informacje przedstawione w inny sposób.

Another way to present the crosstab			
	Predicted - benign	*Predicted - malignant*	*Total*
Actual - benign	154 (99%)	2 (1%)	156 (100%)
Actual - malignant	3 (7%)	41 (93%)	44 (100%)
Total	157	43	200

Każdy przypadek ma zarówno diagnozę przewidywaną (z modelu), jak i diagnozę rzeczywistą (od lekarza). W tabeli przedstawiono każdą z możliwych kombinacji diagnoz przewidywanych i rzeczywistych. Na przykład było 154 przypadków, które były zarówno przewidywane jako łagodne, jak i rzeczywiście łagodne; były to 99 procent ze 156 rzeczywistych łagodnych przypadków. Pozostałe dwa rzeczywiście łagodne przypadki były błędnie przewidywane jako złośliwe.Każdy etap procesu, reprezentowany przez własny węzeł, ma etykietę. W tym przypadku etykiety są po prostu numerami przypisanymi do każdego węzła automatycznie przez oprogramowanie do eksploracji danych. Ale numery węzłów nie pasują do kolejności, w jakiej pojawiają się w procesie tworzenia drzewa decyzyjnego. Jest to powszechne i dzieje się tak tylko dlatego, że możesz ustawić kilka kroków, a następnie stwierdzić, że coś pominąłeś. Węzły otrzymują numery w kolejności, w jakiej są dodawane do procesu, a nie w kolejności, w jakiej działają po zakończeniu procesu. Nie ma to żadnego znaczenia dla wyników; numery węzłów to tylko etykiety. Możesz je zmienić na dowolne nazwy, bez wpływu na obliczenia. Każdy węzeł ma również mały wyświetlacz z czerwonym, żółtym lub zielonym wskaźnikiem. To zawsze zacznie się od czerwonego, zmieni się na żółty, gdy węzeł zostanie skonfigurowany (prawidłowo lub nie), i zmieni kolor na zielony tylko wtedy, gdy ten krok zostanie pomyślnie wykonany.

Ocena wyników

Tabela przestawna porównująca przewidywane diagnozy z rzeczywistymi dla danych o wstrzymaniu podsumowuje wydajność tego drzewa decyzyjnego. Spośród 156 przypadków, które faktycznie były łagodne, 154 zostały prawidłowo przewidziane, czyli 99 procent. Spośród 44 przypadków, które faktycznie były złośliwe, 41 zostało prawidłowo przewidzianych, czyli 93 procent. Korzystanie z alternatywnej aplikacji do eksploracji danych często zapewnia alternatywne ścieżki do uzyskania potrzebnych informacji. Diagnostyczna tabela przestawna pokazana w tabeli 15-1 została utworzona przy użyciu modelu drzewa decyzyjnego i dwóch dodatkowych narzędzi do uzyskania wyników, jak opisano w poprzedniej sekcji. Wyniki nie były zorganizowane w sposób łatwy do odczytania. Dlatego ten proces może wydawać się nieco złożony; wykorzystał kilka narzędzi i kilka kroków tylko po to, aby stworzyć tabelę przestawną do oceny wyników modelu. Oto inny sposób na uzyskanie podobnych informacji. Zamiast przechodzić wieloetapowy proces pokazany w poprzedniej sekcji, możesz uzyskać tabelę przestawną modelu w jednym kroku, używając jednego narzędzia zaprojektowanego do tego celu. Punktator KNIME tworzy tabelę przestawną podobnie jak w powyższej tabeli.

Jeśli możesz uzyskać przydatne wyniki w jednym kroku, dlaczego miałbyś kiedykolwiek korzystać z dłuższego, bardziej złożonego procesu? Każdy proces ma pewne zalety. Jednoetapowy proces był szybki i prosty, ale dłuższy proces zapewniał dodatkowe informacje (procenty wierszy), które są przydatne do oceny modelu. Alternatywy, takie jak te, poszerzają twoje możliwości, zapewniając alternatywne formaty wyjściowe, dostosowując wyniki do zamierzonego zastosowania, a nawet pozwalając uzyskać potrzebne informacje nawet w przypadku napotkania błędu lub innego problemu z preferowanym narzędziem. Nabierz nawyku szukania alternatywnych ścieżek, takich jak te, do pracy związanej z eksploracją danych. Pomoże Ci to opanować aplikację do eksploracji danych i uzyskać precyzyjną kontrolę nad procesami i wynikami. Ponieważ tak wiele jest zagrożonych w diagnostyce medycznej, jest mało prawdopodobne, że maszyny zastąpią lekarzy wykonujących to konkretne zadanie w najbliższym czasie. Ale mogą istnieć inne praktyczne sposoby wykorzystania takiego modelu. Na przykład automatyczne diagnozy mogą być używane do ustalania priorytetów spraw do przeglądu, gdy zasoby są ograniczone. Lekarze i badacze mogą odkryć, że drzewo decyzyjne dostarcza użytecznych wskazówek, które mogą zainspirować dalsze badania, nawet jeśli nie wykorzystaliby bezpośrednio jego przewidywań. Nie każda aplikacja ma tak rygorystyczne wymagania, jak potencjalne decyzje medyczne dotyczące życia lub śmierci. Jeśli uzyskasz takie wyniki w modelu odpowiedzi na zaproszenie do marketingu bezpośredniego, weźmiesz ten model i pobiegniesz z nim!

Zapoznanie się z popularnymi typami drzew decyzyjnych

Chociaż możesz napotkać wiele nazw algorytmów drzew decyzyjnych, większość z nich to odmiany kilku głównych tematów. Oto kilka wielkich nazwisk, które powinieneś znać:

✓ CHAID: Jest to prawdopodobnie najbardziej znany i najczęściej używany rodzaj drzewa decyzyjnego. Jeśli uczęszczałeś na zajęcia ze statystyki w szkole, prawdopodobnie pierwszym testem statystycznym, którego się nauczyłeś, był test niezależności chi-kwadrat. CHAID jest czasami określany jako „chi-kwadrat na sterydach", ponieważ opiera się na wielu testach, takich jak te, których nauczyłeś się w Statistics 101, aby skonstruować diagram najważniejszych interakcji w zbiorze danych. Nazwa jest skrótem od Chi-squared Automatic Interaction Detector.

✓ C&RT: Oto kolejne drzewo decyzyjne oparte na technice, która może być Tobie znana. Ten oparty jest na regresji liniowej. C&RT jest skrótem od drzew klasyfikacyjnych i regresyjnych, ale właściwą nazwą dla tego rodzaju analizy jest binarne partycjonowanie rekurencyjne. C&RT tworzy drzewa tylko z podziałami binarnymi (dwukierunkowymi). (Niektórzy eksperci twierdzą, że C&RT jest najbardziej odpornym z drzew decyzyjnych. Innymi słowy, daje stosunkowo spójne wyniki z próbki do próbki.)

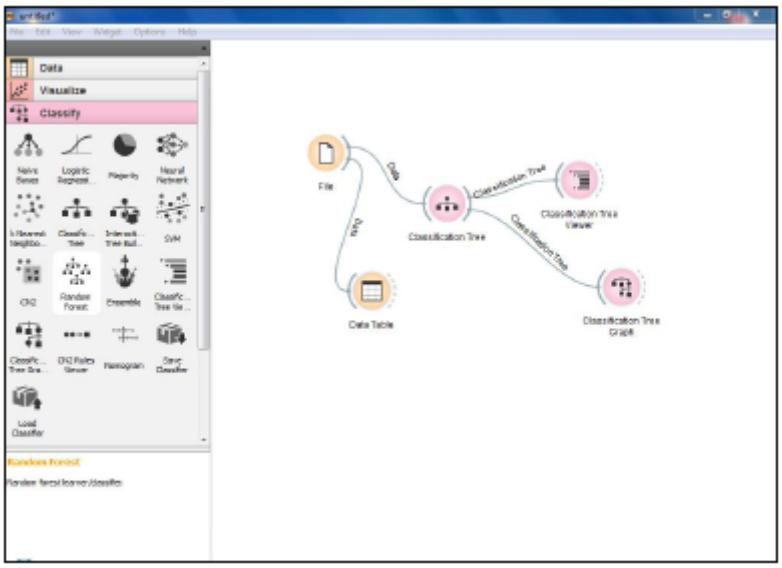
✓ C5.0 (znany również jako See 5.0): C5.0 to zastrzeżony algorytm, który większość początkujących eksploratorów danych napotka tylko w komercyjnym oprogramowaniu do eksploracji danych (na licencji dewelopera). Jednak można znaleźć poprzedników, w tym C4.5 i ID3, nawet w darmowych narzędziach, ponieważ szczegóły tych algorytmów są publikowane.

✓ QUEST: Drzewo decyzyjne zbudowane z myślą o szybkości. To może być Twoje ulubione, jeśli pracujesz z naprawdę dużą ilością danych. Nazwa oznacza szybkie, bezstronne, wydajne, statystyczne drzewo. Podobnie jak C&RT, QUEST tworzy drzewa tylko z podziałami binarnymi (dwukierunkowymi).

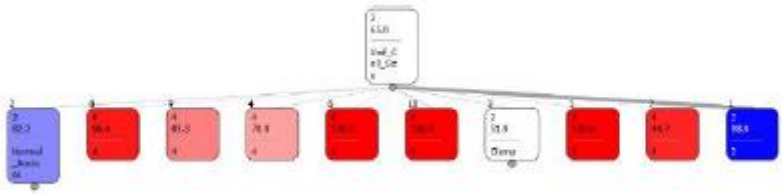
Wszystkie te typy drzew decyzyjnych można wykorzystać do przewidywania zarówno jakościowych, jak i ciągłych zmiennych wyjściowych, a wszystkie mogą wykorzystywać kombinację ciągłych i jakościowych zmiennych wejściowych. Ale fakt, że jest to możliwe, nie gwarantuje, że drzewa w narzędziach, których używasz, będą miały tak elastyczną strukturę. Niektóre narzędzia nie są tak elastyczne! Musisz więc sprawdzić, na co pozwalają Twoje własne narzędzia.

Dostosowywanie się do Twoich narzędzi

Proces użyty do stworzenia drzewa decyzyjnego diagnozy nowotworu został częściowo zdefiniowany przez konkretny produkt użyty do jego zbudowania, którym był KNIME. Możesz uzyskać takie same wyniki z wieloma innymi produktami do eksploracji danych; to kwestia dowiedzenia się, co potrafią Twoje narzędzia i jak z nich właściwie korzystać. Patrząc w innym narzędziu, proces ten może wydawać się na pierwszy rzut oka zupełnie inny. Podobny proces utworzony w Orange, jak pokazano na rysunku, wygląda na prostszy.



Wydaje się, że nie wymaga to wszystkich kroków przygotowania danych, które zostały uwzględnione w przykładzie KNIME. Chociaż w tym przykładzie etapy przygotowania nie są oczywiste, nie oznacza to, że nie jest wymagane żadne przygotowanie danych. Po prostu pojawia się w innych, mniej oczywistych miejscach. Na przykład Orange jest zaprojektowany do pracy z danymi w swoim specyficznym formacie i zdarza się, że dane wymagane do tego procesu były dostępne w tym formacie. Gdybyś musiał sam go przekonwertować, okazałoby się, że zaangażowanych było sporo kroków i mógłbyś spędzić sporo czasu na tym procesie. (I wydaje się, że ten strumień nie zapisuje żadnych danych o wstrzymaniu do oceny modelu. Być może jest to gdzieś w innym pliku. Więcej pracy w przyszłości.) Spójrz na początkowy diagram drzewa utworzony przez ten proces w Orange, pokazany na rysunku



Jest dość ładny, ale wyświetla minimum informacji. Ponownie, będziesz musiał poświęcić czas na odkrywanie narzędzia i wprowadzanie zmian w celu zaspokojenia własnych potrzeb. Każdy produkt ma swoje własne subtelne i niezbyt subtelne różnice w możliwościach i funkcjach, a najlepsze dopasowanie zależy od Twoich konkretnych celów, rodzaju posiadanych danych i interfejsów, z których korzystasz najwygodniej. To może być bardzo osobista sprawa – marzenie jednego eksploratora danych jest koszmarem innego. Biorąc pod uwagę wystarczającą wiedzę i czas, wszelkie wyniki, które można znaleźć dzięki eksploracji danych można było również znaleźć z narzędziami, które nie zostały zaprojektowane do eksploracji danych. Ale celem eksploracji danych jest oddanie analityki w ręce ludzi, którzy nie mają takiej wiedzy lub takiego czasu. Wybierz więc narzędzie, które sprawi, że poczujesz się najwygodniej i pomożesz zrobić jak najwięcej w dostępnym czasie.

Eksploratorzy danych potrzebują otwartych umysłów

Gdy zapoznasz się z metodami modelowania używanymi w eksploracji danych, przekonasz się, że niektóre lubisz bardziej niż inne. Uważaj jednak, aby Twoje preferencje nie przeszkodziły Ci w wypróbowaniu różnych alternatyw. Kiedyś na jednym z moich seminariów zostałem zastraszony przez dwóch dżentelmenów, którzy siedzieli razem z przodu publiczności. Mężczyźni byli najwyraźniej

wielkimi fanami logistycznej regresji, tak wielkimi fanami, że głośno sprzeciwiali się użyciu jakiegokolwiek alternatywnego modelu. Odważnie dyskutowali nad teorią statystyczną, wysuwając twierdzenia, które po prostu nie były prawdziwe. Przyszli na seminarium, ponieważ wymagali tego ich pracodawcy, ale wcale nie byli zainteresowani moimi sugestiami dotyczącymi zalet nowych technik modelowania.

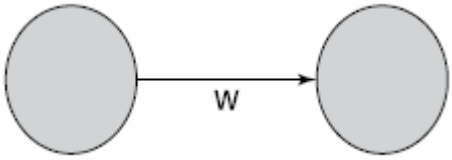
Ludzie rzadko spotykają się z seminariami na temat eksploracji danych, ale często zdarza się, że ludzie opierają się stosowaniu różnych technik modelowania w codziennej pracy. Jaka szkoda! Skuteczni eksploratorzy danych nie ograniczają swoich możliwości podczas budowania modeli. Eksploracja danych opiera się na eksperymentowaniu, porównywaniu alternatyw i wynikach testów. Podczas pracy staraj się wypróbowywać alternatywne typy modeli, kiedy tylko możesz. Porównywanie alternatyw pomaga znaleźć modele, które działają najlepiej, i często pozwala lepiej zrozumieć wzorce w danych. Miej otwarty umysł i eksperymentuj z technikami modelowania, gdy staną się dla Ciebie dostępne. Postaraj się poszerzać swoją wiedzę przez całą karierę, dowiedz się więcej o oprogramowanie, które już posiadasz, wypróbowywanie nowych funkcji po otrzymaniu aktualizacji produktów oraz odkrywanie nowych narzędzi i technik, kiedy tylko możesz. Będziesz lepszym eksploratorem danych za swój wysiłek i rozwiniesz umiejętności i elastyczność niezbędną do dostarczania swoim klientom najlepszych informacji, jakie możesz.

Sieci neuronowe do przewidywania

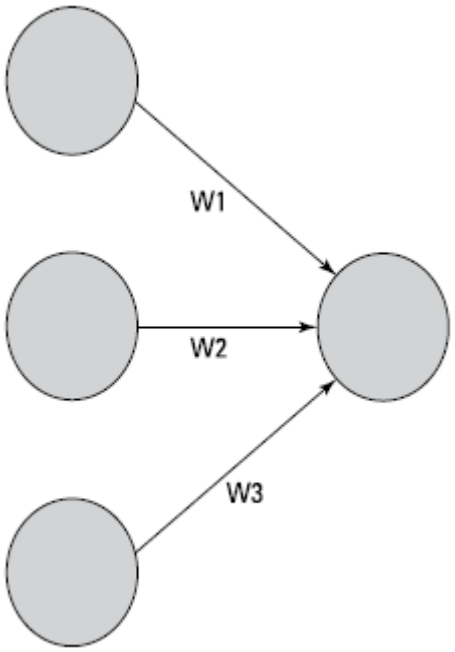
Sieci neuronowe to najbardziej tajemnicza z technik eksploracji danych. Mogą być cenne dla przewidywania i nie są szczególnie trudne w użyciu do tego celu, ale tworzą algorytmy przewidywania tak skomplikowane, że mogą zrobić niewiele lub nic w celu promowania zrozumienia.

Zaglądanie do wnętrza sieci neuronowej

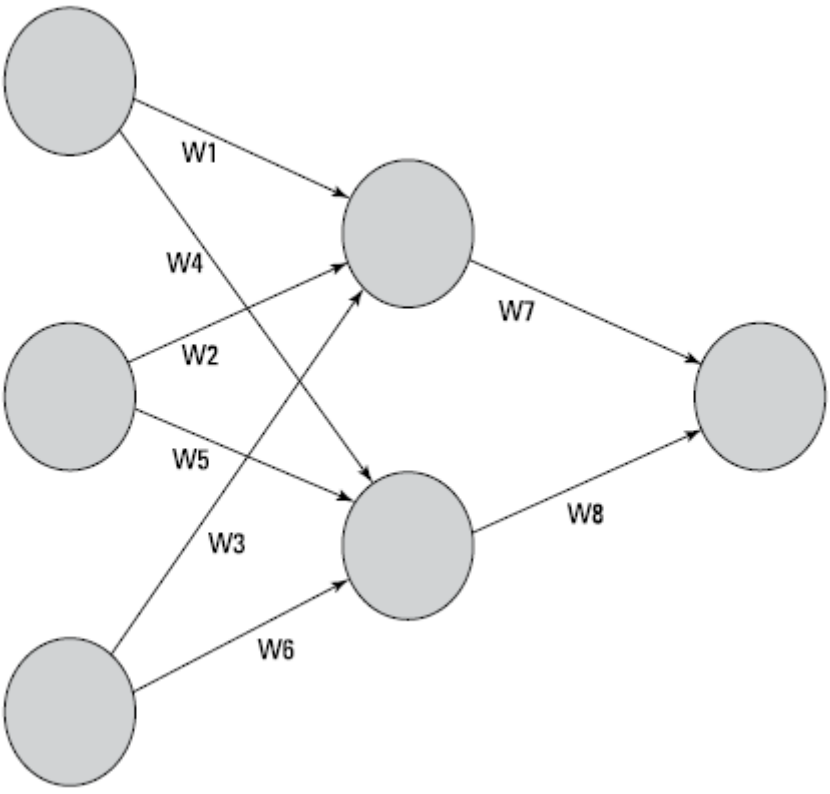
Sieci neuronowe powstały jako modele funkcjonowania ludzkiego mózgu. Neurony, rodzaj komórek, przenoszą informacje przez nasze mózgi i ciała, przesyłając sygnały elektryczne. Okazuje się, że te matematyczne modele nie są dobrymi reprezentacjami tego, jak działają nasze mózgi, ale okazały się całkiem przydatne w niektórych innych zastosowaniach. Aby zrozumieć strukturę sieci neuronowej, zacznę od przyjrzenia się prostemu modelowi matematycznemu i pokonuję krok po kroku. Zacznę od koncepcji linii prostej i zbuduję wspólny typ sieci neuronowej, wielowarstwowy perceptron. (Szybko, powiedz, że trzy razy szybciej! Założę się, że nie możesz. Nikt nie może.) Linia prosta, inaczej zwana modelem liniowym, pobiera dane wejściowe, mnoży je przez pewną wartość, zwaną wagą, i dodaje kolejną stałą wartość, nazywana stałą, aby uzyskać wynik. Prawdopodobnie zrobiłeś coś takiego na lekcji matematyki dawno temu. Zamiast równania, pomyśl o tym jako o prostym diagramie, takim jak ten pokazany na rysunku, który pokazuje jedno wejście połączone z jednym wyjściem, z wagą połączenia.



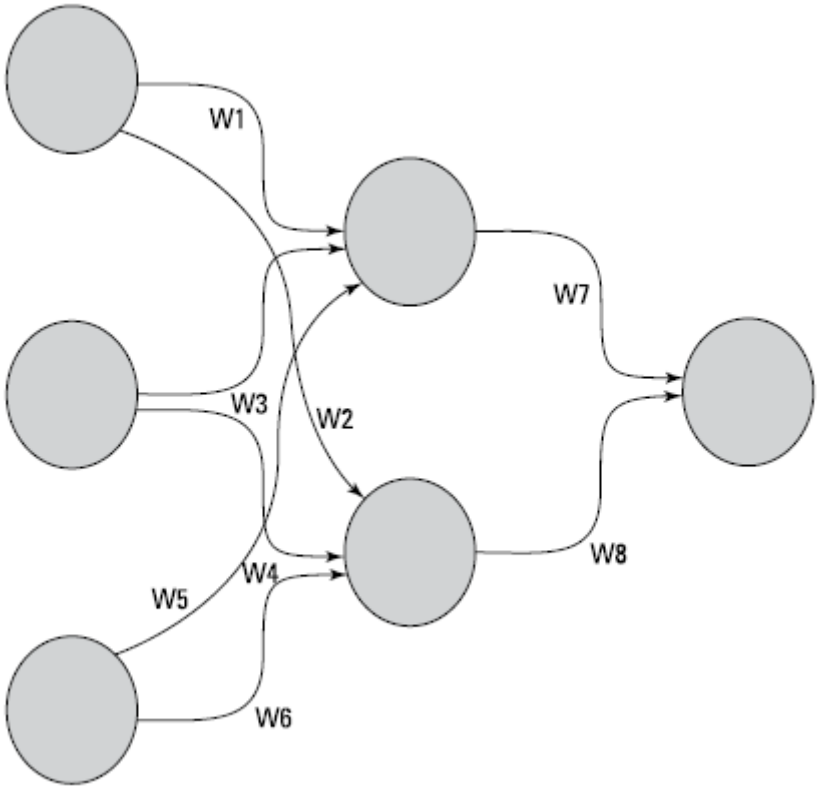
Modele liniowe mogą również mieć więcej niż jedno wejście. Każdy ma swoją wagę. Po prostu mnożysz każde wejście przez jego wagę, dodajesz je i dodajesz stałą, aby uzyskać wynik. Można to przedstawić na schemacie przedstawionym na rysunku.



Pamiętajcie teraz, że sieci neuronowe zostały wynalezione przez badaczy mózgu. Badacze ci ustalili, że niektórych rodzajów problemów nie można odpowiednio modelować za pomocą konwencjonalnych modeli liniowych. Więc wpadli na pomysł ukrytej warstwy. Nakłady byłyby wykorzystywane do obliczania wartości pośrednich, a te pośrednie byłyby pomnożone przez własne wagi w celu uzyskania wyniku, jak pokazano na rysunku.



Jeszcze jedno: modele liniowe nadal obejmują tylko ograniczony zakres sytuacji. Tak więc, zamiast prostych linii, funkcje są bardziej złożonymi formami nieliniowymi, stąd zawijasy na rysunku.



Możesz mieć więcej niż jedną ukrytą warstwę i wiele, wiele danych wejściowych do modelu sieci neuronowej. Jeśli diagram wygląda na skomplikowany, wyobraź sobie skomplikowane instrukcje wymagane do przeprowadzenia obliczeń, które przedstawia! Niektórzy twierdzą, że w sieci neuronowej nie istnieją żadne równania. To nieprawda, ale równania są tak skomplikowane, że większość ludzi nie próbuje na nie patrzeć. Nawet jeśli to zrobisz, nie będziesz w stanie ich zinterpretować tak, jak w przypadku mniej skomplikowanych modeli.

Problemy związane z modelami sieci neuronowych

W poprzedniej sekcji poznałeś, jak działają sieci neuronowe. To dość skomplikowane! A powstałe modele nie są łatwe do zrozumienia przez intuicję. Dlatego wiele osób opisuje sieć neuronową jako czarną skrzynkę. Możesz zrozumieć, że sieć neuronowa pobiera dane wejściowe, wykonuje obliczenia i generuje prognozy (wyjścia). Ale szczegóły tych obliczeń mogą być niezgłębione. Czy to ma znaczenie? Może, a może nie. Jeśli używasz własnych pieniędzy do handlu akcjami, a sieć neuronowa działa dla Ciebie, nie ma problemu, jeśli nie rozumiesz wszystkiego na temat sieci neuronowej. Dbasz tylko o to, czy to działa. Jednak niewielka zmiana w Twojej sytuacji może mieć duży wpływ na to, czy czujesz się komfortowo przy użyciu sieci neuronowych do przewidywania. Być może twoja praca wymaga handlu akcjami, nie własnymi, ale akcjami posiadanymi przez korporacyjny lub związkowy fundusz emerytalny. W takim przypadku każda transakcja może wymagać szczegółowego uzasadnienia. Czarna skrzynka, nawet gdyby jej przewidywania były konsekwentnie trafne, prawdopodobnie nie byłaby satysfakcjonująca. (Patrz Rozdział 4, aby zapoznać się z ósmym prawem eksploracji danych: Wartość wyników eksploracji danych nie jest określona przez dokładność lub stabilność modeli predykcyjnych.)

Twoi klienci prawdopodobnie nie będą Cię pytać o konkretne typy sieci neuronowych, z których korzystasz, ale może się przydać, aby zdawać sobie sprawę z kilku większych. Najpowszechniejszym typem jest perceptron wielowarstwowy (MLP), opisany wcześniej w tym rozdziale. Radial Basis Function (RBF) jest niewielką modyfikacją MLP; modyfikacja sprawia, że działa szybciej. Możesz również spotkać się z siecią Bayesa. Wszystkie te typy sieci neuronowych są wykorzystywane do predykcji. Jeden powszechny typ sieci neuronowej, sieć Kohonena, jest używany do grupowania, podobnie jak metody omówione w następnej sekcji.

Wyznania sieci neuronowej

Niektórzy eksploratorzy danych wierzą w sieci neuronowe. Należą do nich jedni z najbardziej wyrafinowanych praktyków w tej dziedzinie, ludzie z doktoratem i głęboką wiedzą w zakresie uczenia maszynowego, a także eksploratorzy danych z różnych ogrodów, którzy nigdy nie martwią się wewnętrznym działaniem procesu. Odnotowują wielki sukces z sieciami neuronowymi w zastosowaniach tak różnorodnych, jak prognozy giełdowe, rozpoznawanie znaków i włamania do bezpieczeństwa sieci. Inni nie będą dotykać sieci neuronowych. Świadomość, że wyniki są bardzo trudne do wytłumaczenia, wystarczy, aby uniemożliwić im nawet wypróbowanie tych technik. Dla osób pełniących określone funkcje rządowe lub w branżach ściśle regulowanych, w których ważna jest przejrzystość, sieci neuronowe są mało atrakcyjne. Można znaleźć nawet tych, którzy przedstawiają dość rozbudowane teoretyczne wyjaśnienia ograniczeń sieci neuronowych, ale te teoretyczne ograniczenia nie są ważne dla większości codziennych eksploracji danych. Ważne jest to, co działa lub nie działa dobrze dla Ciebie w Twoim własnym środowisku pracy. Sieci neuronowe, w głębi, to tylko grupy równań. To prawda, że równania są złożone i istnieje ich wiele, ale komputer chroni Cię przed tymi szczegółami. Tak więc w większości przypadków prawdziwe pytanie brzmi, czy sieć neuronowa będzie miała dobrą moc predykcyjną dla twojej aplikacji. Korzystam z sieci neuronowych od prawie dwóch dekad. Prowadziłem o nich zajęcia. Napisałem pierwszy podręcznik szkoleniowy dotyczący sieci neuronowych używany przez jednego z największych dostawców eksploracji danych. A oto moje wyznanie sieci neuronowej: w mojej własnej pracy nie przypominam sobie ani jednej rzeczywistej aplikacji klienckiej, w której sieć neuronowa byłaby najbardziej efektywnym modelem, jaki znalazłem. Technika modelowania, dowolna technika modelowania, powoduje, że równania tylko przybliżają wzorce w danych. Jeśli nie istnieją żadne silne wzorce lub jeśli ich forma nie jest dobrze dopasowana do rodzaju używanego modelu, wyniki nie będą miały dużej wartości. Sieci neuronowe są świetne, jeśli w danych istnieją pewne wzorce. Jeśli nie ma właściwych wzorców, cóż, po prostu ich tam nie ma. Dowiesz się tego tylko wtedy, gdy spróbujesz. Możesz uzyskać świetne wyniki, ale możesz nie.

Grupowanie

Kiedy byłeś dzieckiem, być może twoją ulubioną zabawką było pudełko z guzikami babci. Guziki były w wielu różnych kolorach i rozmiarach, niektóre miały dwie dziurki, inne cztery, a być może kilka nie miało żadnych dziurek, ale zamiast nich cholewki. Niektóre były wykonane z plastiku, a inne z drewna, kości lub metalu. Miałeś proste okrągłe i fantazyjne o kształtach, takich jak kwiaty lub klejnoty. Niektóre mogły wydawać się bardzo egzotyczne. Być może kilka zostało złamanych, ale i tak pozostały w kolekcji. Możesz bawić się tymi przyciskami przez całe popołudnie, grupując je według koloru lub kształtu, lub bardziej subtelne i złożone grupowanie według własnego pomysłu. Klastrowanie jest jak zabawa z pudełkiem przycisków, tylko z danymi zamiast przycisków.

Nauka nadzorowana i nienadzorowana

Kiedy myślisz o przewidywaniu, prawdopodobnie myślisz o przewidywaniu jakiegoś określonego wyniku. Jeśli myślisz o dzisiejszej pogodzie, możesz chcieć wiedzieć, czy będzie padać – tak lub nie. Prognoza pogody to rodzaj prognozy. Modele wykorzystywane do prognozowania pogody opierają się

na pogodzie historycznej. Tak więc wynik każdego dnia jest dobrze określony — padało lub nie padało. Budujesz modele, aby przewidywać w ten sposób, porównując dni, w których padało, aby zidentyfikować różnice między nimi i przewidzieć, co stanie się w przyszłości. Sytuacje takie jak ta, w których pogrupowania są jasno określone przez jakiś znany wynik, są zastosowaniami do nadzorowanego uczenia się. W nadzorowanym uczeniu się tworzysz grupy na podstawie jakiegoś znanego wyniku (lub wartości zmiennej). Przykłady podane wcześniej w tym rozdziale, przewidujące, czy pożyczka zostanie spłacona lub czy nowotwór zostanie zdiagnozowany jako łagodny lub złośliwy, były przykładami nadzorowanego uczenia się. Kiedy weźmiesz mieszaną partię przycisków i posortujesz je w grupy, nie masz predefiniowanej zasady tworzenia grup. Badasz przyciski, identyfikujesz interesujące funkcje i sortujesz według znalezionych podobieństw. Takie procesy nazywane są uczeniem nienadzorowanym. (Teraz wiesz, że kiedy widzisz dziecko zagubione w jakimś pozornie bezcelowym dążeniu, możesz być świadkiem uczenia się bez nadzoru). Techniki grupowania to matematyczne procesy uczenia się bez nadzoru, uporządkowanie spraw w Twoich danych w podobne grupy.

Klastrowanie w celu wyjaśnienia

Załóżmy, że wziąłeś garść guzików i posortowałeś je w grupy. Być może podzieliłbyś je na trzy partie:

✓ Małe, białe guziki z czterema otworami

✓ Duże, ciemne guziki

✓ Fantazyjne kształty

Jaki jest z tego pożytek? Możesz użyć tych grup, aby zdecydować, co zrobić z przyciskami. Pierwsza grupa brzmi dobrze do zastąpienia zagubionych guzików koszuli. Drugi może być odpowiedni dla płaszczy. A trzecia grupa, fantazyjne kształty, przydałaby się do celów dekoracyjnych. Grupowanie przycisków w grupy podobnych typów może pomóc w wyjaśnieniu tego, co masz i co dalej.

Korzystanie z person

Firmy często stają przed dużym wyzwaniem w zrozumieniu swoich klientów. Pomyśl o swoim lokalnym sklepie spożywczym. Jest otwarty na biznes z każdym, a jego klienci mają różnorodne potrzeby i pragnienia. Trudno jest zaplanować, kiedy musisz służyć wszystkim, a marketerom bardzo trudno zrozumieć, jak komunikować się ze wszystkimi. Marketerzy muszą zrozumieć swoich docelowych klientów, aby oferować odpowiednie produkty, dostosowywać komunikaty do obaw klientów i robić wszystko, od organizowania układu sklepu po ustalanie cen. Kiedy Twoim docelowym klientem są wszyscy, zrozumienie jest niemożliwe. Prostszą alternatywą jest użycie person, kilku modelowych typów klientów, które razem reprezentują większość bazy klientów. Niektórzy ludzie tworzą te persony klientów w oparciu o intuicję, ale lepiej oprzeć je na danych. Twoja intuicja może być nieaktualna, źle dopasowana do Twojej geografii lub po prostu niewłaściwa.

Oderwanie się od stereotypów

Ten przykład może być ci znajomy: Osobą, która przyciąga wiele uwagi, jest piłkarska mama, często opisywana jako gospodyni domowa z przedmieścia i matka, która jeździ minivanem i spędza większość czasu, prowadząc swoje dzieci na zajęcia, takie jak treningi piłkarskie. Ta postać jest często wspominana od co najmniej dekady. Ale czy ten opis naprawdę przypomina klientów wszystkich firm, które używają tej persony? Jeden z marketerów stanął na dużym zgromadzeniu kobiet, wiele z nich to matki z przedmieścia. Poprosiła matki piłkarskie, aby podniosły ręce. Niewiele kobiet podniosło ręce. Jej publiczność wiedziała o matkach piłkarskich i nie identyfikowała się z tą postacią; dla nich jest to raczej

stereotyp niż realistyczna reprezentacja ich życia. Dysponując odpowiednimi danymi, marketer mógłby odkryć, co naprawdę dzieje się w życiu tych kobiet. Może znaleźć takie grupy:

✓ Matki pracujące w pełnym wymiarze godzin poza domem, z małymi dziećmi w przedszkolu

✓ Matki z dziećmi w wieku szkolnym, które pracują w niepełnym wymiarze godzin i chcą przejść na pełny etat

✓ Bezdzietne kobiety, które interesują się jedzeniem i zdrowiem.

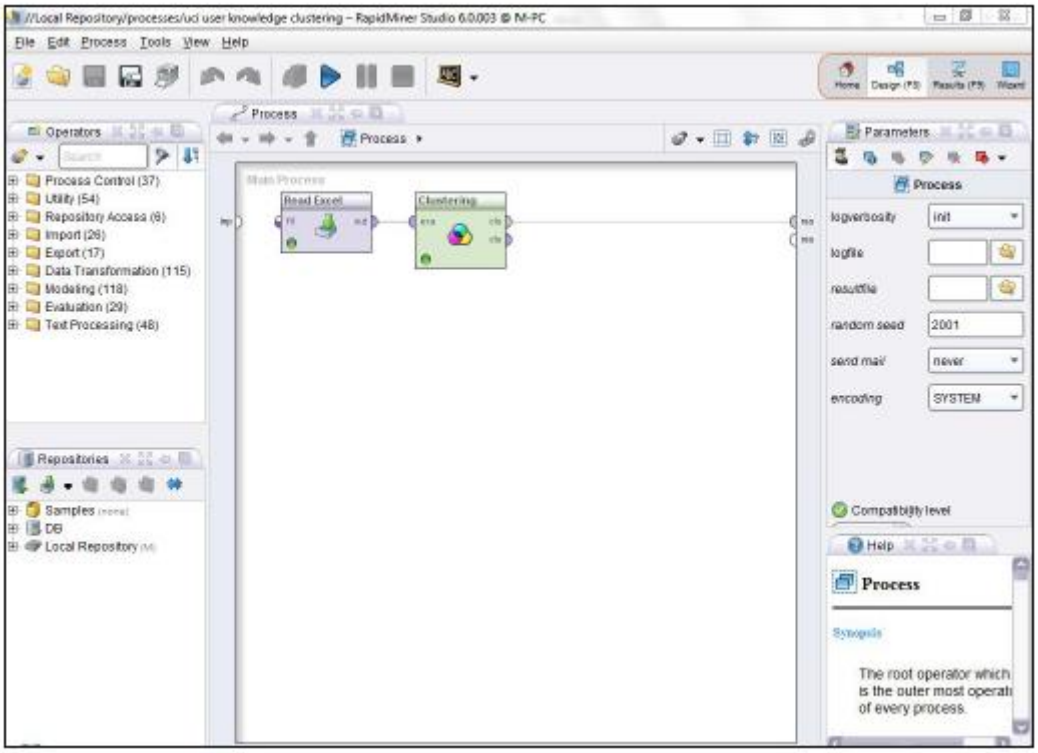
Niezależnie od tego, z jakich postaci korzystasz, ważne jest, aby realistycznie reprezentowały grupy klientów, których obsługujesz, a nie wiesz o tym, chyba że masz na to dane. Możesz stworzyć najbardziej szczegółowe persony, nadawać im imiona i obrazy i myśleć o nich we wszystkim, co robisz, ale to nic nie znaczy, chyba że realistycznie reprezentują Twoich klientów.

Odkrywanie podobieństw wśród uczniów

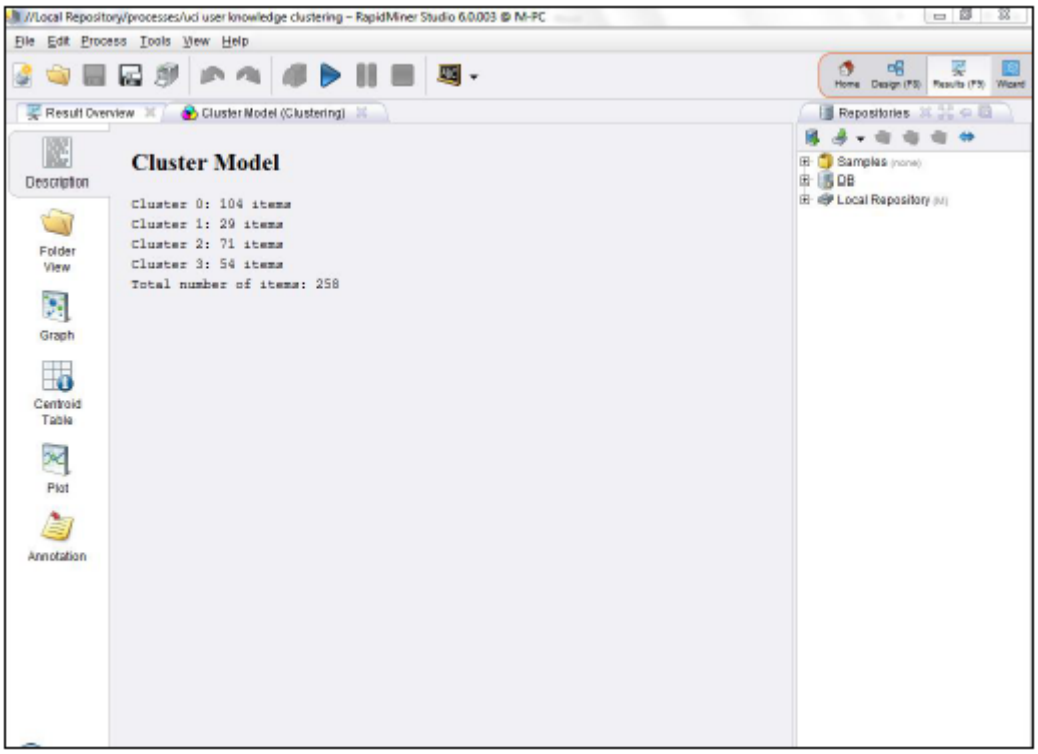
Nauczyciele często mają do czynienia z klasami pełnymi uczniów o zróżnicowanym pochodzeniu i potrzebach edukacyjnych. Dobry nauczyciel może wiele zrobić, aby ocenić sytuację ucznia podczas interakcji twarzą w twarz, ale nie każda sytuacja zapewnia odpowiednie zasoby do wielu interakcji jeden na jednego. Nawet jeśli czas jest dostępny, dostępne materiały dydaktyczne mogą nie być dobrze dopasowane do potrzeb każdego ucznia. Gdybyś przestudiował zróżnicowaną grupę uczniów i przyjrzał się ich nawykom i umiejętnościom, być może znalazłbyś kilka grup o dużych podobieństwach. Bardziej realistyczne byłoby myślenie, że można by dostosować materiały i metody nauczania do trzech lub czterech głównych kategorii niż na przykład do każdego z kilkuset uczniów. Naukowcy z Uniwersytetu Gazi zebrali dane na temat czasu, jaki poszczególni studenci spędzili na różnych rodzajach studiów oraz ich wyników w powiązanych testach. Czy te dane można wykorzystać do identyfikacji i scharakteryzowania grup podobnych uczniów? Jeśli tak, instruktorzy mogą uzyskać przydatne informacje i lepiej pomóc uczniom w osiągnięciu sukcesu.

Patrząc na proces

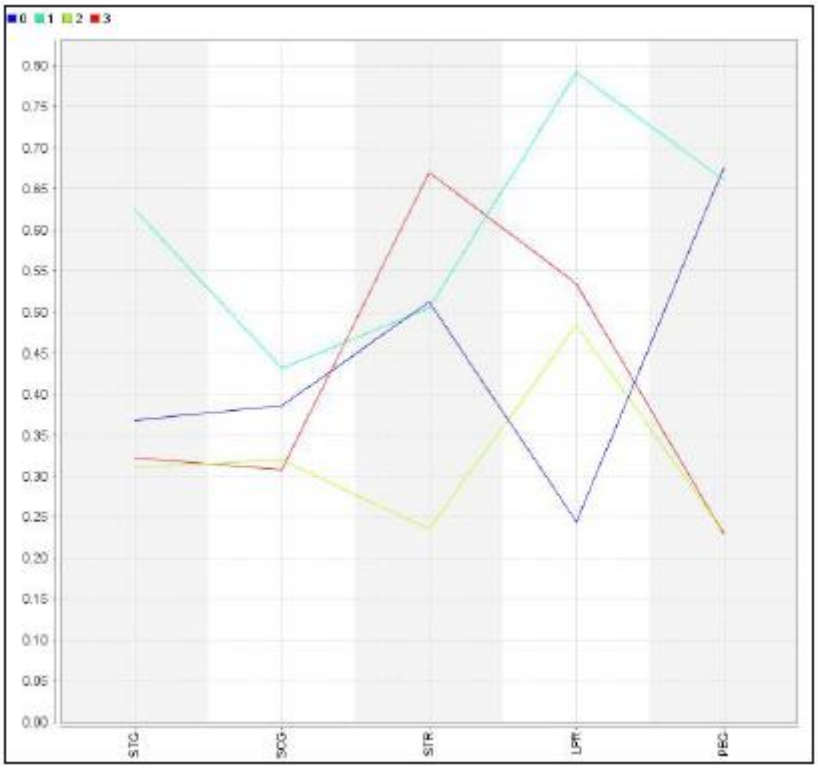
Jeden proces grupowania danych pokazano na rysunku.



Wygląda to strasznie prosto, tylko w dwóch krokach: import danych i klaster. Jednym z powodów, dla których w grę wchodzi tak niewiele kroków, jest to, że badacze, którzy zbierali i udostępniali dane, wykonali dobrą robotę. Zmienne są właściwymi typami (dane ciągłe) dla tej procedury grupowania (k-średnie, najpopularniejsza metoda grupowania), a dane są kompletne; brak brakujących wartości. Każdy z tych dwóch kroków wymagał trochę konfiguracji, np. określenia, którą część arkusza kalkulacyjnego należy zaimportować i ile klastrów należy utworzyć. Dane obejmują 258 studentów. Dla każdego masz zmienne dotyczące czasu spędzonego na trzech typach badań (nazwy zmiennych STG, SCG i STR) oraz wydajności w dwóch testach (nazwy zmiennych LPR i PEG). Wartości w danych są znormalizowane; innymi słowy, wszystkie zostały przekonwertowane do skali od 0 do 1. Normalizacja jest wymagana do korzystania z niektórych narzędzi, a matematycy ją uwielbiają, ale nie jest zbyt pouczająca dla zwykłych użytkowników biznesowych. Jeśli robiłeś to dla klienta, chciałbyś oryginalnych danych lub przynajmniej sposobu na konwersję danych. Jakie informacje znajdują się w wynikach? Rysunek 15-25 pokazuje, że powstały cztery klastry.



Nie są tego samego rozmiaru. Najmniejsza obejmuje tylko 29 uczniów, a największa 104. Wykres przedstawiony na rysunku 15-26 pokazuje typowe cechy każdej grupy.



Ten wykres skupień pokazuje wyniki testu dla czterech grup pokazanych:

✓ Grupa 0, największa grupa, poświęca dużo czasu na badanie, druga najwyższa dla każdego z trzech pomiarów badania. Można by się spodziewać, że opłaci się to wysokimi wynikami testów, a w

przypadku jednego testu tak się dzieje. Ta grupa uzyskała najwyższe wyniki w teście PEG. Ale najgorzej wypadli w teście LPR.

✓ Grupa 1, najmniejsza grupa, poświęca dużo czasu na wszystkie rodzaje badań i dobrze wypada w obu testach. (To są zwierzęta domowe nauczyciela.)

✓ Grupa 2 studiuje najmniej, ale nadal radzi sobie dobrze w jednym teście, LPR. Osiągają słabe wyniki w teście PEG.

✓ Grupa 3 inwestuje umiarkowaną ilość czasu we wszystkie obszary badań i osiąga najwyższe wyniki w teście STR. Ale radzą sobie gorzej niż jakakolwiek inna grupa w teście PEG.

Teraz, zamiast próbować dowiedzieć się, co się dzieje z każdą z 258 osób, nauczyciele mają cztery grupy, z których każda ma inną sytuację. Jest prawdopodobne, że nauczyciele, znający uczniów i treść testów, dostrzegliby zagadnienia, które odzwierciedlają te skupienia. Jeśli nie, to przynajmniej mają teraz możliwość wykorzystania tych informacji jako podstawy do dalszych badań. Im lepiej rozumieją, co może kryć się za różnicami grupowymi, tym lepiej mogą dostosować program edukacyjny, aby pomóc uczniom lepiej.

Eksploracja danych przy użyciu klasycznych metod statystycznych

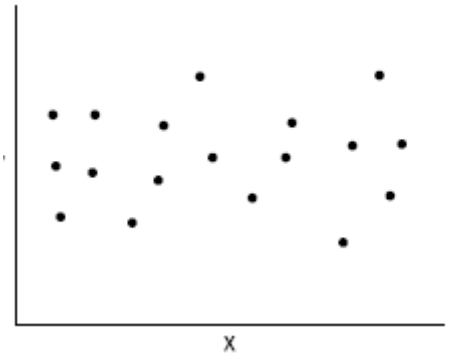
Eksperci danych nie są purystami, więc nie istnieje wyraźna granica między metodami wykorzystywanymi przez eksploratorów danych a metodami stosowanymi przez tradycyjnych analityków. Eksperci danych pożyczają od tradycjonalistów, kiedy jest to korzystne i praktyczne. Zestaw narzędzi do eksploracji danych zawiera kilka technik, które są znane nawet najbardziej rygorystycznym klasycznym statystykom. Wśród dawnych ulubionych technik eksploracji danych znajdują się korelacja, regresja liniowa i regresja logistyczna. Ten rozdział zawiera szczegółowe informacje na temat każdego z nich.

Zrozumienie korelacji

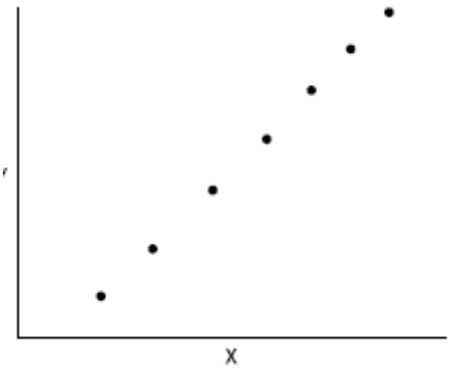
Czy owinąłbyś usta wokół rury wydechowej samochodu i wdychał pełne oparów? Oczywiście nie! Dlaczego nie? Bo wiesz, że wdychanie spalin nie jest zdrowe (a to wyglądałoby śmiesznie). Prawdopodobnie wiesz również, że wdychanie dymu papierosowego nie jest zdrowe. Dlaczego nie? Jednym z powodów jest to, że dym papierosowy zawiera tlenek węgla, ten sam materiał, który znajduje się w spalinach. Być może teraz o tym wiesz, ale kilkadziesiąt lat temu ludzie nie byli świadomi niebezpieczeństw związanych z paleniem. W rzeczywistości firmy tytoniowe reklamowały swoje produkty jako zdrowe. Jedna z reklam głosiła: „Więcej lekarzy pali wielbłądy niż jakikolwiek inny papieros". Inny powiedział: „20 679 lekarzy twierdzi, że szczęście jest mniej irytujące". Reklamy papierosów często zawierały zdjęcia lekarzy, naukowców, pielęgniarek, a nawet dentystów. Jak więc odkryliśmy, że palenie jest niezdrowe? Zaczęło się od korelacji. Ludzie zauważyli, że palacze kaszleli, a ich gardła były podrażnione. Być może zaczęli też podejrzewać związek między paleniem a niektórymi chorobami. Dlatego badacze zdrowia publicznego zebrali i przeanalizowali dane. Powiązali ilość palonej osoby z kaszlem, podrażnieniem gardła i występowaniem niektórych chorób. Dane potwierdziły, że kiedy palenie wzrosło, wszystkie te nieprzyjemne problemy zdrowotne również wzrosły. Kiedy rosnące wartości jednej zmiennej idą w parze z rosnącymi (lub malejącymi) wartościami innej, to jest to korelacja.

Obrazowanie korelacji

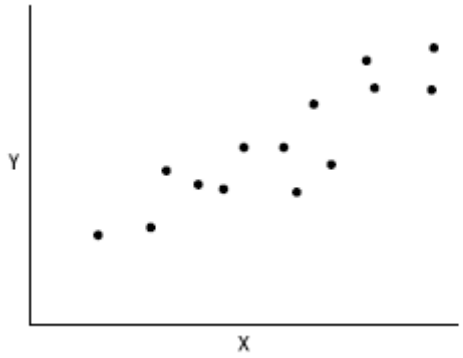
Być może zauważyłeś: niektórzy palacze często kaszlą. Zauważenie czegoś takiego może skłonić Cię do zastanowienia się: czy dane potwierdzają związek między ilością, jaką osoba pali, a ilością kaszlu? Aby to zbadać, musiałbyś szukać związku między dwiema ciągłymi zmiennymi – ilością wypalanych papierosów i częstotliwością kaszlu w ciągu dnia. Jak rozpoznać, czy dwie zmienne ciągłe są skorelowane? Jednym ze sposobów jest użycie prostego wykresu rozrzutu. Umieść jedną zmienną na osi X (poziomo), a drugą na osi Y (pionowo). Jeśli punkty przypominają poziomą chmurę na wykresie, jak pokazano na rysunku, te dwie zmienne nie są skorelowane.

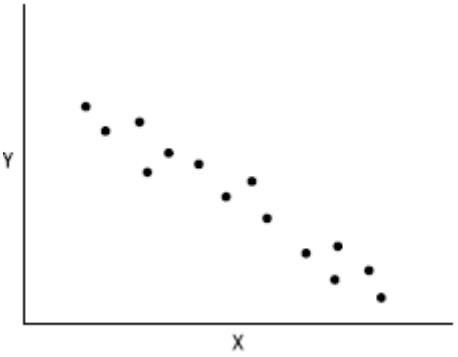


Zmienne są skorelowane, gdy wzrost jednej z nich wiąże się ze wzrostem (lub spadkiem) drugiej. Zupełnym przeciwieństwem zmiennych nieskorelowanych są doskonale skorelowane zmienne, takie jak para pokazana na rysunku. W takim przypadku widzisz spójną liniową zależność między dwiema zmiennymi, a każdy punkt leży na tej linii bez odchylenia.



W rzeczywistych przykładach z pewnością napotkasz pary zmiennych, które są skorelowane, ale nie oczekuj, że rzeczywiste wzorce będą wyglądać jak na rysunku 16-2. Realistyczne przypadki wyglądają bardziej jak te pokazane na rysunku, który pokazuje parę dodatnio skorelowanych zmiennych (gdy jedna idzie w górę, druga też) i na rysunku, który pokazuje ujemnie skorelowane zmienne.





Jeśli zmienne są skorelowane, zobaczysz, że punkty aproksymują linię. Im bliżej punkty zbliżają się do utworzenia linii prostej, tym silniejsza korelacja. Nachylenie linii nie ma znaczenia; może być w górę lub w dół, stromo lub płytko, o ile linia nie jest idealnie pozioma. (Gdyby linia była pozioma, oznaczałoby to, że zmienna zależna ma zawsze tę samą wartość. To jest stała, a nie korelacja).

Mierzenie siły korelacji

Korelacja jest określana jako wartość (nazywana współczynnikiem korelacji i często w skrócie literą r), która mieści się w zakresie od –1 do 1, w następujący sposób:

✓ Współczynnik korelacji równy 0 oznacza, że te dwie zmienne nie są skorelowane.

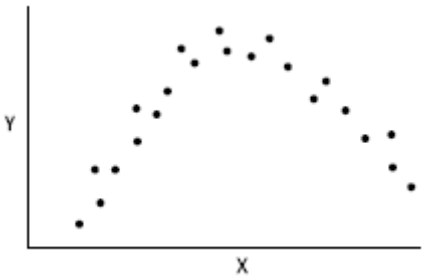
✓ Współczynnik korelacji równy 1 oznacza doskonałą korelację dodatnią.

✓ Współczynnik korelacji -1 oznacza doskonałą korelację ujemną.

Tak więc współczynnik korelacji bliski 0 oznacza niewielki lub żaden związek między zmiennymi. Im bliżej współczynnik zbliża się do 1 lub –1, tym silniejsza korelacja. Większość produktów do eksploracji danych i analizy statystycznej może obliczyć te współczynniki za Ciebie. Nawet niektóre kalkulatory mają wbudowane funkcje korelacji, które mogą być przydatne, jeśli pracujesz z bardzo małą ilością danych. Sztuczka polega na tym, aby dowiedzieć się, jak nazywa się współczynnik korelacji w produkcie, który posiadasz. Poszukaj korelacji, współczynnika korelacji, współczynnika korelacji liniowej, współczynnika korelacji Pearsona (lub po prostu zwykłego Pearsona) lub r.

Rysowanie linii w danych

Kiedy używasz funkcji korelacji w pracy związanej z eksploracją danych, szukasz liniowych relacji między zmiennymi. Ale jeśli w danych nie ma żadnego liniowego wzorca (innymi słowy, gdy współczynnik korelacji jest bliski 0 lub na wykresie rozrzutu nie widać żadnego liniowego wzorca), nie musi to oznaczać, że dane nie mają żadnego znaczącego wzorca do znalezienia. Znaczące wzory nie zawsze podążają za prostymi liniami. Wykres rozrzutu na rysunku przedstawia przykład wzoru nieliniowego. Wykres jest zakrzywiony, z wzrostami i spadkami. Zmienna y ma niską wartość, gdy wartość zmiennej x jest niska. Wraz ze wzrostem x zwiększa się również y, aż do pewnego punktu. Poza tym punktem y maleje, gdy x wzrasta.

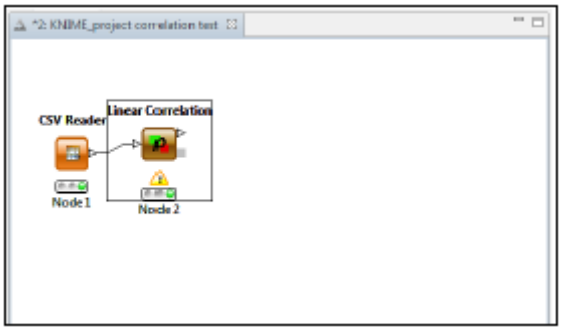


Korelacje liniowe nie są właściwym narzędziem do znajdowania lub mierzenia nieliniowych wzorców takich jak te. W rzeczywistości współczynnik korelacji dla tego wzoru wynosiłby 0. Jeśli nie zidentyfikujesz interesującego wzoru między dwiema zmiennymi za pomocą korelacji, nie oznacza to, że nie masz żadnego wzoru do znalezienia. Oznacza to po prostu, że nie istnieje żaden liniowy wzór.

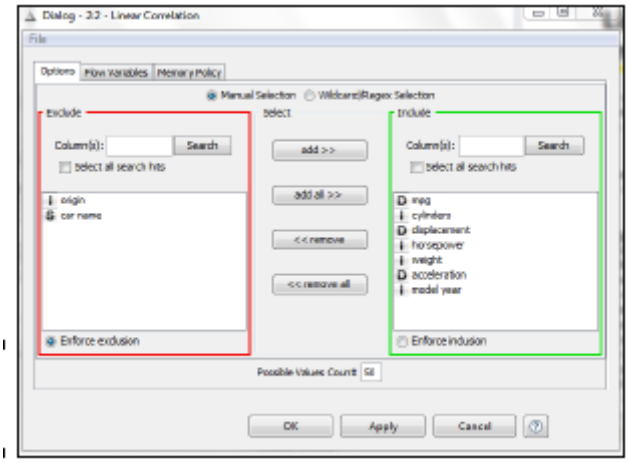
Próba korelacji

Generowanie współczynników korelacji z danych to zwykle kwestia znalezienia odpowiedniego narzędzia (w produkcie do eksploracji danych) lub pozycji menu (w konwencjonalnym produkcie do analizy statystycznej) i wybrania wszystkich zmiennych, dla których chcesz uzyskać korelacje. Korelacje dotyczą tylko dwóch zmiennych na raz, więc jeśli wybierzesz więcej niż dwie zmienne, oprogramowanie obliczy korelacje dla każdej pary z listy. Konstrukcja samochodu w dużej mierze decyduje o jego zachowaniu i ilości zużywanego paliwa. Na przykład można się spodziewać, że im więcej cylindrów ma samochód, tym gorszy będzie przebieg. Korelacje mogą pomóc w potwierdzeniu (lub odrzuceniu) takich oczekiwań i kwantyfikacji tych efektów. Rysunek 16-6 przedstawia przykładowy

przepływ pracy (w produkcie KNIME) do obliczania korelacji. Ma tylko dwa kroki, jeden do importowania danych do oprogramowania, a drugi do ustawiania korelacji.



Dokładna liczba kroków potrzebnych do wykonania danego zadania może różnić się od przykładów w tej książce. Przygotowanie danych, projekt Twojego produktu do eksploracji danych i Twoje własne specyficzne potrzeby mają wpływ na rodzaje i liczbę kroków wymaganych dla Twoich własnych procesów. Zazwyczaj podczas konfigurowania korelacji można wybrać dowolną liczbę zmiennych naraz. Rysunek 16-7 przedstawia konfigurację narzędzia korelacji, zawierającą kilka zmiennych i wykluczając inne. W tym przykładzie zmienna mpg to przebieg - mile na galon benzyny. Można oczekiwać, że inne czynniki, takie jak liczba cylindrów i waga samochodu, będą miały wpływ na przebieg.



Często będziesz mieć wybór formatów wyjściowych dla korelacji. Najpopularniejsza jest macierz korelacji, rodzaj tabeli, która pozwala odczytać współczynnik korelacji dla dowolnej pary zmiennych. Rysunek 16-8 przedstawia macierz korelacji dla przykładu na rysunku. (Odnieś się do poprzedniej sekcji, aby dowiedzieć się, jak interpretować współczynniki korelacji).

Table "Correlation values" - Rows: 7 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	mpg	cylinders	displac...	horsep...	weight	acceler...	model y...
mpg	1	-0.775	-0.804	-0.778	-0.832	0.42	0.579
cylinders	-0.775	1	0.951	0.843	0.896	-0.505	-0.349
displacement	-0.804	0.951	1	0.897	0.933	-0.544	-0.37
horsepower	-0.778	0.843	0.897	1	0.865	-0.689	-0.416
weight	-0.832	0.896	0.933	0.865	1	-0.417	-0.307
acceleration	0.42	-0.505	-0.544	-0.689	-0.417	1	0.288
model year	0.579	-0.349	-0.37	-0.416	-0.307	0.288	1

Jedna rzecz, którą możesz zauważyć w macierzy korelacji: powtarza informacje. Możesz znaleźć korelację między mpg a cylindrami w rzędzie mpg lub w rzędzie cylindrów. Jaki to ma sens? To tylko

kwestia wygody, abyś mógł zbadać korelacje w preferowany sposób. Możesz mieć alternatywę uzyskania korelacji w jakiejś formie graficznej. Niektóre produkty oferują macierz wykresów rozrzutu, zbiór wykresów rozrzutu podobny do tych przedstawionych na rysunkach, ale bardzo małe i zorganizowane w rzędy i kolumny podobnie do macierzy korelacji. Inny rodzaj grafiki wykorzystuje kolor do wskazania pozytywnych i negatywnych korelacji, a intensywność do odzwierciedlenia siły korelacji. Styl wydruku również różni się w zależności od produktu. Macierz korelacji pokazana na rysunku (utworzona za pomocą PSPP, narzędzia do analizy statystycznej opartej na menu) zawiera miary istotności. Odpowiadają one na pytanie, czy jakakolwiek korelacja zaobserwowana w danych jest znacząca, czy po prostu jest artefaktem naturalnej zmienności punktów danych. Wartości istotności bliskie 0 wskazują, że korelacja jest istotna, co oznacza, że dowody sugerują, że jest to spowodowane nie tylko przypadkową zmiennością. (Wartość N w macierzy to liczba przypadków wykorzystanych do obliczenia każdego współczynnika korelacji).

Korelacja i przyczynowość

Wcześniej czy później z pewnością usłyszysz to, co mówi: Korelacja nie oznacza związku przyczynowego. Możesz zauważyć, że zmienna B rośnie, gdy rośnie zmienna A. Albo że zmienna C maleje, gdy zmienna A rośnie. Te wzorce są korelacjami. Kiedy słyszysz, że korelacja nie implikuje związku przyczynowego, oznacza to, że te wzorce nie są dowodem na to, że zmienna A powoduje zmiany w zmiennych B i C. Korelacja występuje w kilku sytuacjach, m.in.

✓ Jedna zmienna jest przyczyną, a druga skutkiem.

✓ Zmienne wynikają ze wspólnej przyczyny.

✓ Zbieg okoliczności.

Idea, że korelacja nie implikuje związku przyczynowego, jest ważną koncepcją we wszelkiego rodzaju analizach danych. Ale jest używany przez niektórych jako wymówka do odrzucenia wszelkiego rodzaju statystyk. Możesz usłyszeć, jak jeden z nich mówi coś takiego: „Wskaźniki morderstw rosną, gdy rośnie sprzedaż lodów. Czy to oznacza, że lody powodują morderstwo? To prawda, że w letnim upale wskaźniki morderstw rosną wraz ze sprzedażą lodów, i to prawda, że nie jest to dowód na to, że lody są przyczyną morderstw. Ale znajdziesz inną stronę kwestii korelacji i przyczynowości. Jeśli jedno powoduje drugie, zostanie to odzwierciedlone w danych. Jeśli edukacja zapobiega ciąży nastolatek, nastolatki, które zdobywają więcej wykształcenia będą mieć mniej ciąż. Jeśli ekstremalne upały powodują śmierć osób starszych, podczas fal upałów nastąpi wzrost liczby zgonów osób starszych. Jeśli azot w glebie sprzyja wzrostowi soi, plony soi będą wyższe na polach z dużą ilością azotu w glebie niż na polach, na których gleba jest uboga w azot. Innymi słowy, korelacja może nie implikować związku przyczynowego, ale przyczynowość implikuje korelację. Jeśli więc podejrzewasz, że Rzecz 1 powoduje Rzecz 2, poszukaj danych i dowiedz się, czy dane potwierdzają twoją teorię.

Correlations

		mpg	cylinders	displacement	horsepower	weight	acceleration	model_year
mpg	Pearson Correlation	1.00	-.78	-.80	-.78	-.83	.42	.58
	Sig. (2-tailed)		.00	.00	.00	.00	.00	.00
	N	398	398	398	392	398	398	398
cylinders	Pearson Correlation	-.78	1.00	.95	.84	.90	-.51	-.35
	Sig. (2-tailed)	.00		.00	.00	.00	.00	.00
	N	398	398	398	392	398	398	398
displacement	Pearson Correlation	-.80	.95	1.00	.90	.93	-.54	-.37
	Sig. (2-tailed)	.00	.00		.00	.00	.00	.00
	N	398	398	398	392	398	398	398
horsepower	Pearson Correlation	-.78	.84	.90	1.00	.86	-.69	-.42
	Sig. (2-tailed)	.00	.00	.00		.00	.00	.00
	N	392	392	392	392	392	392	392
weight	Pearson Correlation	-.83	.90	.93	.86	1.00	-.42	-.31
	Sig. (2-tailed)	.00	.00	.00	.00		.00	.00
	N	398	398	398	392	398	398	398
acceleration	Pearson Correlation	.42	-.51	-.54	-.69	-.42	1.00	.29
	Sig. (2-tailed)	.00	.00	.00	.00	.00		.00
	N	398	398	398	392	398	398	398
model_year	Pearson Correlation	.58	-.35	-.37	-.42	-.31	.29	1.00
	Sig. (2-tailed)	.00	.00	.00	.00	.00	.00	
	N	398	398	398	392	398	398	398

Zrozumienie regresji liniowej

Korelacje mówią nam o liniowych wzorcach w danych. Współczynnik korelacji definiuje związek między dwiema zmiennymi ciągłymi, mówiąc nam, czy istnieje między nimi zależność liniowa, i wyrażający, jak silny jest ten związek. Następnym krokiem jest znalezienie równania prostej, która łączy jedną zmienną z drugą, proces zwany regresją liniową. Eksploratorzy danych używają tych równań do przewidywania wartości jednej zmiennej na podstawie wartości innej. Prognozy pomagają nam zrozumieć, w jaki sposób możemy kontrolować rzeczy, które chcemy kontrolować. A kiedy nie mamy kontroli, dobre prognozy pomagają nam planować. Po tym, jak odkryliśmy, jak znaleźć linię, która łączy jedną zmienną z drugą, już mały krok w górę jest znalezienie liniowych relacji między grupami więcej niż dwóch zmiennych. Nazywa się to wielokrotną regresją liniową.

Praca z liniami prostymi

Prawdopodobnie nauczyłeś się trochę o liniach w szkole. Linia łączy zmienną zależną ze zmienną niezależną za pomocą prostej formuły. W szkole prawdopodobnie wyrażałeś tę zależność jako wzór lub równanie, ale tutaj użyję słów, aby wyjaśnić ten proces. Wywołaj zmienną niezależną x i zmienną zależną y. Jednym ze sposobów uzyskania wartości y odpowiadającej danej wartości x jest pomnożenie x razy liczby zwanej nachyleniem, a następnie dodanie kolejnej liczby, zwanej stałą. Nachylenie określa nachylenie linii, a stała określa stały punkt początkowy linii (wartość y, gdy x wynosi zero). Często zapisuje się to jako równanie

$y = mx + b$

gdzie m to nachylenie, a b to stała, czyli punkt przecięcia z osią y.

Jeśli nauczyłeś się o prostych w algebrze lub w podstawowej klasie matematyki, być może spotkałeś się tylko z przykładami, w których każdy punkt idealnie pasuje do prostej. W eksploracji danych (jak w klasycznej statystyce) nie będziesz miał takiej perfekcji. Rzeczywiste dane nie układają się w idealne linie, ale czasami linia prosta jest użytecznym przybliżeniem naturalnego wzorca danych, jak pokazano na przykładach na rysunkach wcześniejszych .

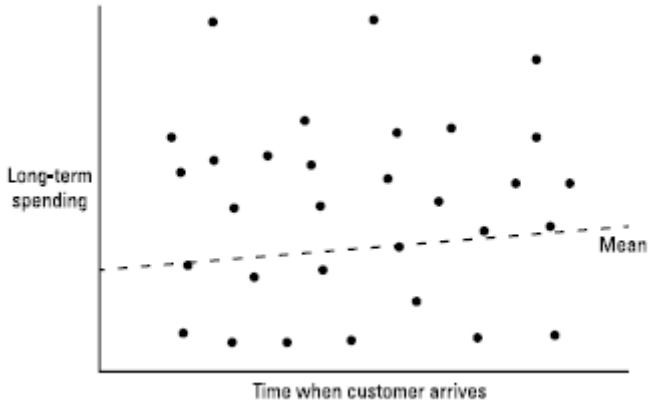
Konfrontacja ze zbyt wieloma wyborami

Być może w szkole nauczyłeś się, że wzór na linię prostą można znaleźć w zaledwie dwóch punktach. To prawda, ale nie jest to zbyt pomocne, gdy masz wiele punktów danych i nie tworzą one idealnej linii. Możesz wybrać jedną parę punktów i znaleźć linię, ale jeśli wybierzesz inną parę, otrzymasz inną linię. Prawie tyle możliwych linii istnieje jako punktów w twoich danych. Pytanie brzmi, której linii

najlepiej użyć do przewidywania? Na szczęście nie musisz zgadywać, jak znaleźć najlepszą linię dla swoich danych. Istnieje dobrze zdefiniowana metoda, która pozwala to zrobić za Ciebie, taka, która wykorzystuje wszystkie Twoje dane i daje Ci jedną linię, która najlepiej pasuje do danych. Ta metoda nazywa się regresją liniową.

Traktowanie każdego przypadku w ten sam sposób

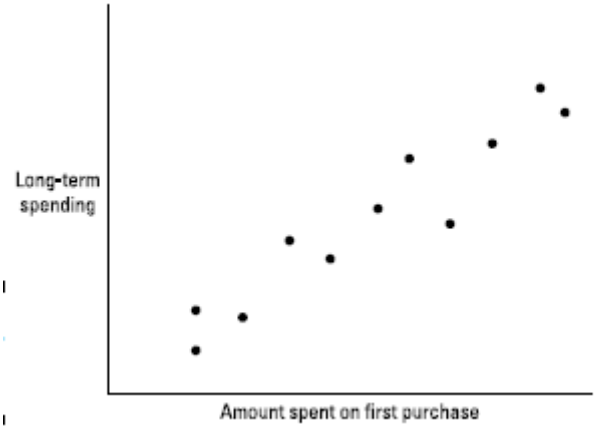
Być może prowadzisz sklep detaliczny i chciałbyś wiedzieć, ile pieniędzy nowy klient może wydać w dłuższej perspektywie. Jeśli nie wiesz nic o następnym kliencie, który wejdzie do drzwi, najlepiej, co możesz zrobić, to założyć, że wydatki tego klienta będą takie same, jak średni poziom wydatków (średnia) dla wszystkich Twoich klientów. Wiesz, że niektórzy klienci będą wydawać więcej, a inni mniej, ale oszacujesz taki sam, średni poziom wydatków dla wszystkich. Wyobraź sobie śledzenie zachowań zakupowych klientów wchodzących do Twojego sklepu. Możesz nadać każdej osobie, która wchodzi do sklepu, numer identyfikacyjny i odnotować kwotę, jaką każda osoba wydaje. Wykres rozrzutu wydatków w funkcji numeru identyfikacyjnego może wyglądać jak na rysunku.



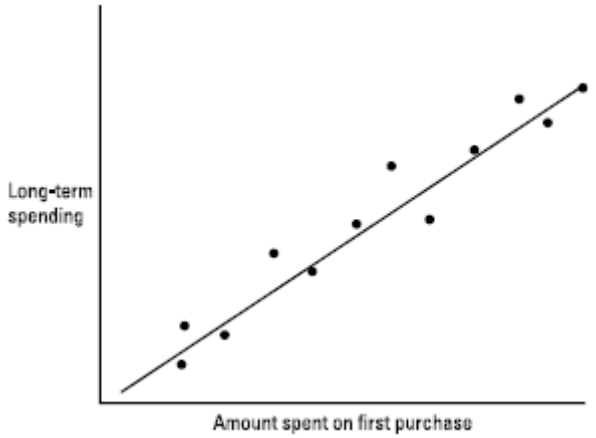
Nie istnieje żaden szczególny związek między numerem identyfikacyjnym a wydatkami, a wydatki różnią się w zależności od osoby, więc punkty tworzą rozproszoną chmurę na wykresie. Jeśli chcesz przewidzieć wydatki dowolnego pojedynczego klienta, najlepszym oszacowaniem, jakie możesz wykonać, byłoby przewidzenie, że klient wyda średnią (średnią) kwotę wydatków (obliczoną przez zsumowanie wszystkich wydatków klientów i podzielenie przez liczbę klientów). Zrobiłbyś taką samą prognozę dla każdego klienta, więc prognozy utworzyłyby poziomą linię na wykresie. Ta pozioma linia reprezentuje najlepszą prognozę, jaką możesz wykonać na podstawie posiadanych informacji. Daleko mu do ideału, jak pokazuje szeroki rozrzut punktów powyżej i poniżej linii.

Traktowanie jednostek jako jednostek

Jeśli wiesz trochę o kliencie, możesz lepiej oszacować. Co możesz wiedzieć o nowym kliencie? Wiesz, co klient kupił i jakie ceny zapłacił. Wiesz, czy klient zapłacił gotówką, kartą kredytową, debetową lub podarunkową. Wiesz, czy klient wykorzystał jakieś kupony. Jeśli masz program lojalnościowy, klient mógł dołączyć i przekazać Ci więcej informacji, w tym adres. Aby odpowiedzieć na pytanie, ile pieniędzy nowi klienci prawdopodobnie wydadzą w dłuższej perspektywie, dobrze byłoby zacząć od przyjrzenia się, jak wydatki na pierwszą wizytę klienta mają się do wydatków długoterminowych. (Jeśli masz program lojalnościowy, członkostwo klientów lub konta, powinieneś mieć dane niezbędne do tego.) Jeśli wykreślisz te dane za pomocą wykresu rozrzutu, może to wyglądać jak na rysunku.



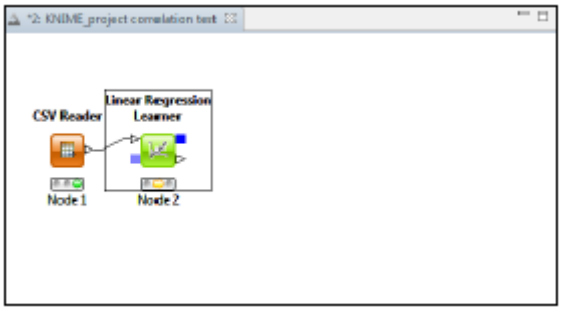
Kiedy nic nie wiedziałeś o klientach, użycie średniej mogło nie wyglądać na bardzo dobry sposób na przewidywanie, ale nie miałeś oczywistej lepszej opcji. Ale kiedy dodasz odpowiednie informacje (w tym przypadku wydatki na pierwszą wizytę), może pojawić się inna opcja. W tym przykładzie wykres rozrzutu pokazuje, że im więcej klienci wydają na pierwszą wizytę, tym więcej wydają w dłuższej perspektywie. To, czego teraz potrzebujesz, to dobry sposób na wykorzystanie tych informacji do tworzenia lepszych prognoz. Gdybyś mógł narysować właściwą linię przez dane, jak pokazano na rysunku, przypominałaby ona wzór punktów danych znacznie bardziej niż linia pozioma pokazana na rysunku wcześniej. W ten sposób regresja liniowa umożliwia dokonywanie lepszych przewidywań niż przy użyciu tylko średniej, tworząc linię, która najlepiej pasuje do danych. W tej sekcji średnia odnosi się do określonego rodzaju średniej, średniej. Jeśli odejmiesz średnią od każdej rzeczywistej wartości w zbiorze danych, aby obliczyć odchylenie, a następnie zsumujesz wszystkie odchylenia, suma wyniesie 0. Aby skorzystać ze zmiennej predykcyjnej i utworzyć linię, która pasuje do danych lepiej niż średnia, potrzebna jest nieco większa złożoność. Linia najlepszego dopasowania danych to ta, która minimalizuje sumę wszystkich kwadratów wartości odchyleń między każdą rzeczywistą wartością zmiennej zależnej a wartością przewidywaną przez linię. Teoria statystyczna dostarcza formuł do tych obliczeń, ale możesz nigdy ich nie widzieć. Twoje oprogramowanie zadba o to za Ciebie.



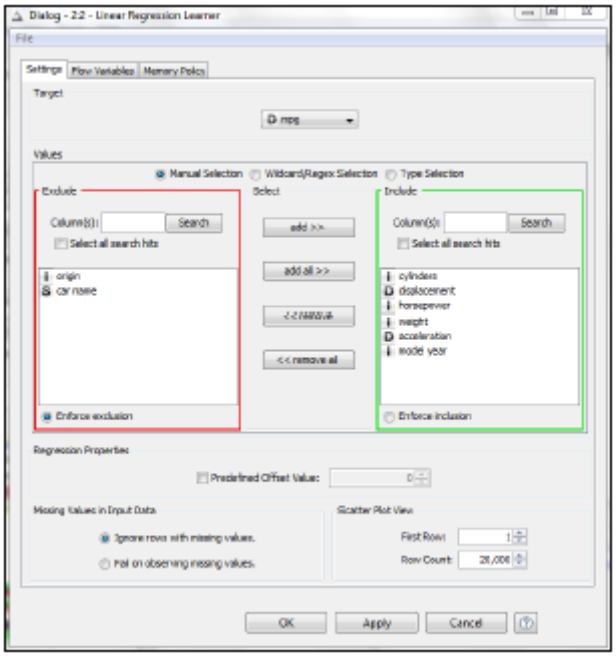
Znalezienie najlepszej linii

Regresja liniowa opiera się na koncepcji korelacji, aby uzyskać formułę, której można używać do przewidywania, oraz możliwość włączenia do tej formuły kilku zmiennych niezależnych (predyktorów) jednocześnie. Kroki, które podejmiesz, aby uzyskać ten wzór (wykonanie regresji liniowej) są tylko trochę bardziej skomplikowane niż to, co zrobiłeś dla korelacji. Przepływ pracy regresji liniowej

pokazany na rysunku wygląda bardzo podobnie do przykładu korelacji pokazanego na rysunku, ale tutaj drugi krok obejmuje narzędzie do regresji liniowej. Konfiguracja pokazana na rysunku jest nieco inna. W regresji liniowej musisz określić jedną zmienną zależną – zmienną, którą chcesz przewidzieć.

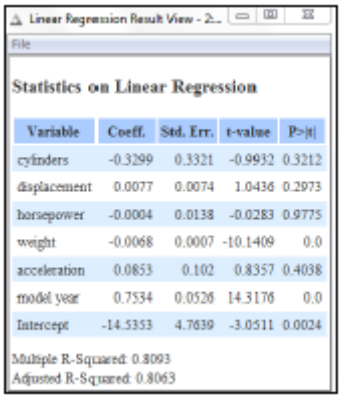


Jako eksplorator danych możesz robić rzeczy, których tradycyjny statystyk nigdy by nie zrobił. Konfiguracja regresji liniowej pokazana na rysunku zawiera kilka statystyk nie do przyjęcia: wiadomo, że kilka zmiennych niezależnych jest skorelowanych, a rok modelowy samochodu nie jest zmienną ciągłą; jest to kategoria reprezentowana przez liczbę. Bycie eksploratorem danych nie eliminuje problemów, które mogą się pojawić przy wykonywaniu tych czynności, ale zamiast wracać do teorii statystycznej w celu uzyskania poprawek, możesz najpierw pozwolić sobie na swobodę, a następnie ocenić wyniki, testując wstrzymanie i nowe dane. Jeśli to działa, to działa.



Korzystanie ze współczynników regresji liniowej

Wzór z regresji przebiegu auto pokazano na rysunku. Może to nie wygląda jak inne formuły, z którymi się spotkałeś, ale kiedy nauczysz się je czytać, przekonasz się, że to całkiem zwyczajne równanie.



Formuła składa się z dwóch części: współczynników regresji (takich jak nachylenie linii narysowanych w szkole) dla każdej zmiennej niezależnej oraz wyraz wolny (stała). Aby dokonać prognozy na podstawie zestawu wartości dla zmiennych niezależnych, pomnóż każdą zmienną przez jej współczynnik regresji, zsumuj wyniki dla wszystkich zmiennych, a następnie dodaj stałą (punkt przecięcia), aby uzyskać przewidywaną wartość zmiennej zależnej. Jeśli samochód spełnia ten opis:

✓ Cylindry: 4

✓ Przemieszczenie: 156

✓ Moc: 92

✓ Waga: 2585

✓ Przyspieszenie: 14,5

✓ Rok modelowy: 82

Jego przewidywany przebieg to

(–0,3299 × 4)

+ (0,0077 × 156)

+ (–0,0004 × 92)

+ (–0,0068 × 2585)

+ (0,0853 × 14,5)

+ (0,7534 × 82)

+ (–14.5353)

= 30,75

Model szacuje, że przebieg samochodu wyniesie 30,75 mil na galon. Czy będzie to prawda przez cały czas dla każdego samochodu, który spełnia powyższy opis? Nie, ale to najlepsze oszacowanie, jakie możesz wykonać na podstawie posiadanych informacji. Nie musisz używać ręcznych obliczeń, aby przewidywać! Twój produkt do eksploracji danych może to zrobić za Ciebie.

Interpretowanie statystyk modelu

Twoje wyniki regresji liniowej często zawierają więcej niż tylko współczynniki modelu. Możesz mieć statystyki modelu dla każdej zmiennej niezależnej. Kluczowa wartość, której należy szukać, będzie zwykle nazywana p lub znaczeniem (lub po prostu Sig). W ten sposób można interpretować znaczenie. Wartości mogą mieścić się w zakresie od 0 do 1. Niskie wartości (bliskie 0 i zwykle nie większe niż 0,05) sugerują, że czynnik jest ważny dla modelu. Wyższe wartości (powiedzmy 0,05 lub wyższe) sugerują, że czynnik nie dodaje wartości jako predyktor. Mówi się, że wynik jest bardzo istotny, gdy wartość istotności jest niska! Możesz usunąć z modelu czynniki, które nie są istotne i przebudować go. Daje to często prostszy, ale równie skuteczny model.

Stosowanie zdrowego rozsądku

Ze zdroworozsądkowego punktu widzenia wyprowadzony w poprzednim podrozdziale wzór na przewidywanie przebiegu gazu ma pewne problemy. Zdrowy rozsądek podpowiada, że każdy czynnik, który sprawia, że samochód jest mocniejszy lub szybszy, taki jak liczba cylindrów lub pojemność skokowa, powoduje również większe zużycie paliwa. Wraz ze wzrostem któregokolwiek z tych czynników wzrasta zużycie paliwa i zmniejsza się przebieg. Można więc oczekiwać, że współczynniki każdego z tych czynników będą ujemne. Ale ten model regresji liniowej (wzór) nie jest zgodny z tym zdroworozsądkowym oczekiwaniem. Współczynniki przemieszczenia i przyspieszenia są dodatnie.

Nie trać wiary w zdrowy rozsądek, gdy wyniki są zaskakujące. Wiedza biznesowa to podstawa każdego eksploratora danych, a Twoja wiedza na temat samochodów mówi Ci, że Twój model ma problem. (Możesz mieć momenty w swojej karierze, kiedy odkryjesz, że jedno z Twoich oczekiwań było po prostu błędne, ale potrzebujesz wielu dowodów, aby wyciągnąć taki wniosek, a nie tylko jednego zaskakującego modelu.) Te nieoczekiwane wyniki są spowodowane wielowspółliniowością, zmienne niezależne, które są ze sobą skorelowane. Masz kilka możliwości radzenia sobie z tym:

✓ Przetestuj swój model na nowych danych iw terenie, zaczynając od małej skali. Jeśli daje pożyteczne rezultaty, używaj go pomimo jego niedoskonałości.

✓ Spróbuj użyć mniejszej liczby zmiennych niezależnych i sprawdź, czy uzyskasz model, który wydaje się bardziej rozsądny.

✓ Dowiedz się więcej o tym, jak statystycy zajmują się kolinearnością i stosują niektóre z tych technik. Nie jest przestępstwem, że eksploratorzy danych pożyczają od czasu do czasu metody od innych analityków danych! (Książki, które szczegółowo wyjaśniają współliniowość i sposoby jej rozwiązywania, są dostępne w większości bibliotek uniwersyteckich lub dużych bibliotek publicznych.)

Rozpoznawanie wyzwania pomiarowego

Mierzenie wpływu ceny w kontekście, którego nie można w pełni kontrolować, jest jednym z najtrudniejszych (i najczęstszych) wyzwań pomiarowych. Oto przykład tego, jak to się dzieje, i kilka alternatyw dla rozwiązania problemu. Twoja firma staje przed pytaniem: czy podnosić ceny? Niektórzy menedżerowie sprzeciwiają się temu, wierząc, że jeśli ceny wzrosną, mniej ludzi będzie kupować, a przychody spadną. Ściśle wierzą w ekonomię podaży i popytu. Inni uważają, że cena Twoich produktów jest zbyt niska, a klienci nadal będą je kupować po wyższej cenie. Jeśli mają rację, podniesienie cen zwiększy przychody i zyski. Wy, menedżerowie, macie dwa sprzeczne pomysły:

✓ Wraz ze wzrostem cen spadają przychody.

✓ Gdy ceny rosną, rosną również przychody.

Ponieważ cena i dochód są zmiennymi ciągłymi, regresja liniowa jest odpowiednią techniką analityczną do pomiaru wpływu zmian cen na dochód. Jednak uzyskanie danych odpowiadających Twoim potrzebom może nie być tak łatwe. Jeśli masz środowisko sprzedaży, które możesz kontrolować, tak jak w przypadku bezpośredniej sprzedaży wysyłkowej, możesz przeprowadzić kontrolowany eksperyment. Podzielisz listę mailingową klientów na kilka grup i wyślesz każdemu ofertę na ten sam produkt, używając tej samej kopii i tej samej grafiki. Oferta jest dokładnie taka sama dla wszystkich, z jednym wyjątkiem: inna cena. Ale niektórych środowisk sprzedaży nie da się tak łatwo i skutecznie kontrolować. Jeśli Twój produkt jest sprzedawany w sklepach detalicznych, nie możesz oferować różnych cen osobom w tym samym sklepie. Zmiany cen związane ze specjalnymi promocjami w sklepach prawdopodobnie też nie są dobrym modelem w Twojej sytuacji, ponieważ zawsze wiążą się z czymś więcej niż tylko zmianą ceny; wpływ mają również reklamy i oznakowania w sklepach. Lepszymi alternatywami może być przejrzenie tego, co się stało z przeszłymi podwyżkami cen lub zbadanie różnic cenowych między sklepami i danych o sprzedaży.

Zrozumienie regresji logistycznej

Co się dzieje, gdy rzecz, którą chcesz przewidzieć, nie jest zmienną ciągłą? Często będziesz musiał przewidzieć, że coś się wydarzy lub nie. Klient, który odwiedzi Twoją witrynę, dokona lub nie dokona zakupu. Uczeń zda lub nie zda testu. Guz zostanie lub nie zostanie zdiagnozowany jako złośliwy. To wszystko są przykłady wyników binarnych (zmienna kategorialna z tylko dwiema możliwymi wartościami). Regresja liniowa nie jest odpowiednia do ich przewidywania. Drzewa decyzyjne i sieci neuronowe, dwa typy modeli wyjaśnione w rozdziale 15, mogą być używane do przewidywania takich zmiennych kategorialnych. Mimo to eksploratorzy danych często sięgają po bardziej tradycyjną technikę, regresję logistyczną.

Patrząc na regresję logistyczną

Regresja liniowa opiera się na liniach prostych, które są znane większości ludzi. Regresja logistyczna opiera się jednak na funkcjach logitowych, które nie są znane większości ludzi i nie są tak proste jak linia prosta. Możesz jednak znać logity pod inną nazwą: iloraz szans. Regresja logistyczna umożliwia przewidywanie kategorii, takich jak

✓ Pożyczka: spłacona lub niespłacona

✓ Guz: łagodny lub złośliwy

✓ Preferowana cola: Coca-Cola, Pepsi lub Royal Crown

Działa przy użyciu funkcji logitowej do obliczania szans (szansa, że dany wynik się wydarzy) dla każdej opcji na podstawie Twoich danych. Przewidywana kategoria to ta, która ma najkorzystniejsze kursy.

✓ Zła wiadomość: Dokonywanie prognozy za pomocą regresji logistycznej obejmuje więcej kroków i bardziej złożonych obliczeń niż te związane z prostszym modelem, takim jak regresja liniowa. I jeśli ty i twoi odbiorcy nie jesteście zaznajomieni zarówno z logarytmami, jak i szansami, wyjaśnienie wyników regresji logistycznej może być trudne.

✓ Nieco dobra wiadomość: Twoje oprogramowanie wykonuje obliczenia, aby tworzyć prognozy za Ciebie (a może nawet eksportować kod, który możesz zintegrować z innymi aplikacjami), dzięki czemu nie musisz zajmować się złożonością tworzenia prognoz. Ale twoje oprogramowanie nie ułatwi wyjaśnienia modelu!

Doceniając urok regresji logistycznej

Regresja logistyczna nie jest najłatwiejszym modelem do zrozumienia lub wyjaśnienia. Zawsze masz możliwość skorzystania z modelu drzewa decyzyjnego (pod warunkiem, że masz do tego odpowiednie narzędzie). Dlaczego więc jakikolwiek eksplorator danych miałby zamiast tego używać regresji logistycznej? Niektórzy ludzie wolą nawet regresję logistyczną od wszystkich innych opcji. Czemu? Regresja logistyczna może być atrakcyjna z wielu powodów:

✓ Eksperci danych rutynowo próbują każdy odpowiedni typ modelu dla danej aplikacji, a czasami regresja logistyczna okazuje się być typem modelu, który działa najlepiej.

✓ Niektórzy ludzie znają regresję logistyczną, a przynajmniej jej nazwę, z narażenia w miejscu pracy lub zaawansowanych zajęć statystycznych (i ufają znajomym rzeczom).

✓ Niektórzy odbiorcy są dobrze zaznajomieni z ilorazami szans i logarytmami (być może grupa inżynierów, którzy obstawiają sport), więc regresja logistyczna działa na ich strefę komfortu.

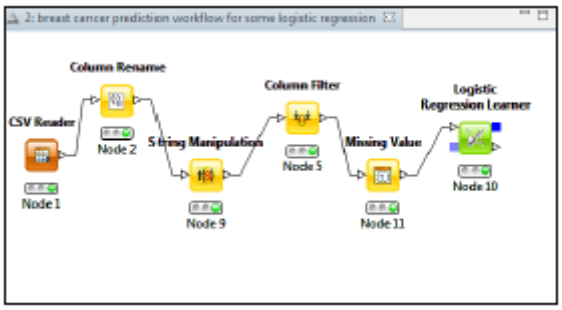
✓ Regresja logistyczna tworzy jedno równanie dla wyników, a niektórzy wolą to, nawet jeśli równanie jest dość złożone.

✓ Modele regresji logistycznej są zwykle stabilne (niezbyt wrażliwe na drobne zmiany danych), co jest pożądaną cechą i przewagą nad niektórymi innymi rodzajami modeli.

Ci, którzy najbardziej kochają regresję logistyczną, to zazwyczaj osoby, które mają spore przeszkolenie w klasycznej statystyce i ogólną nieufność do eksploracji danych.

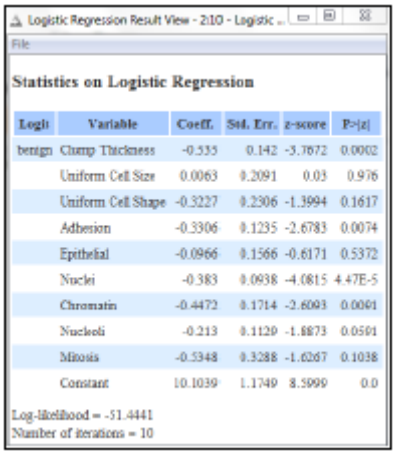
Patrząc na przykład regresji logistycznej

Rysunek przedstawia przepływ pracy dla przykładu regresji logistycznej. W tym przykładzie celem jest przewidzenie, czy nowotwór zostanie zdiagnozowany jako złośliwy na podstawie takich czynników, jak jednorodność rozmiaru i kształtu komórek w obrębie nowotworu.



Ten przepływ pracy obejmuje kilka kroków, które nie są uwzględnione w przykładach korelacji lub regresji liniowej omówionych wcześniej w tym rozdziale. Te kroki są wymagane do przygotowania danych, takich jak etykietowanie zmiennych i radzenie sobie z brakami danych w danych. Liczba takich kroków różni się w zależności od charakteru danych i projektu narzędzi, z których korzystasz. Współczynniki modelu przedstawiono na rysunku 16-17. Wygląda to podobnie do wyników regresji liniowej pokazanej na rysunku 16-15, ale znaczenie nie jest takie samo. W tym przypadku współczynniki są elementami funkcji logitowej.

Obliczenia wymagane do użycia danych wejściowych i tych współczynników do przewidzenia wyniku są bardziej złożone niż obliczenia ręczne, które pokazano wcześniej w tym rozdziale dla regresji liniowej. Można to zrobić ręcznie, ale jest to dość pracochłonne i łatwe do popełnienia błędu. Niech Twoje narzędzia zrobią to za Ciebie.


Logistic Regression Result View - 2:10 - Logistic ...
File
Statistics on Logistic Regression
Logit Variable Coeff. Std. Err. z-score P>|z|
benign Clump Thickness -0.535 0.142 -3.7672 0.0002
Uniform Cell Size 0.0063 0.2091 0.03 0.976
Uniform Cell Shape -0.3227 0.2306 -1.3994 0.1617
Adhesion -0.3306 0.1235 -2.6783 0.0074
Epithelial -0.0966 0.1566 -0.6171 0.5372
Nuclei -0.383 0.0938 -4.0815 4.47E-5
Chromatin -0.4472 0.1714 -2.6093 0.0091
Nucleoli -0.213 0.1129 -1.8873 0.0591
Mitosis -0.5348 0.3288 -1.6267 0.1038
Constant 10.1039 1.1749 8.5999 0.0
Log-likelihood = -51.4441
Number of iterations = 10

Wydobywanie danych dla wskazówek

Być może znasz powiedzenie: „Gdzie jest dym, tam jest ogień”. Co to znaczy? Dosłowne znaczenie jest właśnie tym, co mówi. Kiedy widzisz dym, to znak ognia. To ostrzeżenie, wskazówka. Wiesz, że to nie znaczy, że dym powoduje pożar. I wiesz, że to nie znaczy, że dym nie może istnieć bez ognia. Dym jest silnym wskaźnikiem ognia, a nie przyczyną i nie jest doskonałym dowodem. Ta idea dymu i ognia jest tak znacząca, że ludzie czasami używają tego samego wyrażenia w odniesieniu do rzeczy, które nie mają z nimi nic wspólnego. Często, gdy ktoś mówi: „Gdzie jest dym, tam jest ogień”, ma na myśli: „To ważna wskazówka”. Jeśli zobaczysz ciemne chmury i poczujesz szybki wiatr w letnie popołudnie, możesz podejrzewać, że będzie padać i zdecydujesz się nosić parasol. Jeśli zajrzysz do samochodu i zauważysz fotelik dla dziecka i torbę na pieluchy, możesz założyć, że właściciel jest dobrą perspektywą dla Twojej usługi opieki nad dziećmi i zostawić ulotkę na przedniej szybie. Szukasz i używasz tych wskazówek lub skojarzeń w codziennym życiu, aby pomóc sobie w podejmowaniu dobrych decyzji. To samo możesz zrobić ze swoimi danymi. W tym rozdziale dowiesz się, dlaczego asocjacje są ważne dla eksploratorów danych, jak je znaleźć w swoich danych oraz jak możesz skorzystać z opcji, które pozwalają dostosować reguły asocjacji do własnych potrzeb.

Kombinacje śledzenia

Przed ladą rybną w supermarkecie często można znaleźć półkę z butelkami sosu tatarskiego i koszem cytryn. Sklep ma alejkę z przyprawami i działem produktów z owocami, więc co te rzeczy robią w dziale rybnym? Chodzi o koszyk rynkowy, kombinację produktów kupionych razem w ramach jednego zakupu. Menedżerowie sklepów wiedzą, że ludzie często kupują ryby, sos tatarski i cytryny w połączeniu. Wyświetlanie tych trzech produktów razem przynosi korzyści kupującemu i sklepowi. Kupujący dostaje produkty szybko i łatwo, eliminując potrzebę chodzenia do trzech różnych działów lub wyszukiwania sosu tatarskiego wśród innych przypraw. Sklep sprzedaje więcej, ponieważ kupujący przy ladzie rybnej widzą i kupują przedmioty, które im się podobają, ale mogli zapomnieć lub brakowało im cierpliwości, aby znaleźć inne miejsca w sklepie. Takie ekspozycje mogą również zwiększyć zyski, gdy kupujący rybę, która jest w sprzedaży, kupują również dodatkowe przedmioty po pełnej cenie.

Gdyby menedżerowie supermarketów wiedzieli o innych kombinacjach kupowanych razem produktów, mogliby zastosować to samo podejście do zwiększenia przychodów i zysków w całym sklepie. Tutaj wkraczają eksploratorzy danych. Eksploratorzy danych mogą znaleźć kombinacje, takie jak to poprzez śledzenie powiązań w danych.

Znajdowanie skojarzeń w danych

Wiele rodzajów danych zawiera informacje o powiązaniach. Asocjacje można znaleźć w różnych źródłach danych, takich jak:

✓ Handel detaliczny: Produkty kupowane razem sugerują taktykę zwiększania przychodów.

✓ Medycyna: Objawy i wyniki badań występujące w połączeniu sugerują konkretne diagnozy.

✓ Usługi socjalne: Dane ogólne i ekonomiczne odzwierciedlają powszechne zapotrzebowanie na usługi.

Asocjacje są odzwierciedlone w danych przez wiele przypadków z identycznymi wartościami dla kilku zmiennych. Skojarzenie opisane w poprzedniej sekcji – ryba, sos tatarski i cytryny – pojawiało się w ewidencji sprzedaży supermarketu jako wiele przypadków, w których te trzy produkty zostały uwzględnione w jednej transakcji, niezależnie od tego, jakie inne produkty zostały lub nie zostały

zakupione . Koncepcja jest prosta, ale bardzo trudno byłoby spojrzeć na bazę danych zawierającą miliony przypadków i setki zmiennych i dostrzec te kombinacje.

Organizowanie zasad stowarzyszenia

Reguła asocjacji to stwierdzenie typu „Jeśli rzecz A, to rzecz B". Tak więc, gdybyśmy znaleźli powiązanie między dymem a ogniem, regułą asocjacyjną byłoby „Jeśli dym, to ogień". Oto jeden przykład:

herbatniki=t mrożonki=t owoce=t ogółem=wysoka 788 ==> chleb i ciasto=t 723

Oto, jak ten przykład mógłby brzmieć, gdyby został napisany słowami. Jeśli zakup obejmuje herbatniki, mrożonki i owoce, obejmuje również chleb i ciasto. Było to prawdą 723 z 788 razy w danych. Dokładna notacja różni się w zależności od wybranej aplikacji do eksploracji danych, ale idea zawsze będzie taka sama. W tym przykładzie herbatniki=t oznacza, że herbatniki zostały zakupione (t oznacza prawdę), a ==> zastępuje słowa If i then. Liczby są zliczane, ile razy każda rzecz się wydarzyła. Powinieneś także znać kilka pojęć, które są często używane z zasadami asocjacji:

✓ Zestawy przedmiotów: Zestawy przedmiotów to grupy rzeczy. Grupa składająca się z herbatników, mrożonek i owoców to zestaw przedmiotów. Podobnie jak ciastka, mrożonki, owoce, chleb i ciasto.

✓ Konsekwentny: Konsekwentny jest częścią Jeżeli reguły asocjacyjnej. W poprzednim przykładzie następstwem są ciastka, mrożonki, owoce.

✓ Poprzednik: Poprzednik jest częścią Następnie reguły stowarzyszenia. W tym przykładzie poprzednikiem jest chleb i ciasto.

Przygotowywanie się

Większość przykładów procesów eksploracji danych wykorzystuje wizualne interfejsy programowania. W programowaniu wizualnym używasz ikon (małych obrazków) do reprezentowania etapów procesu eksploracji danych. Programowanie wizualne ma kluczowe znaczenie dla eksploracji danych, ale nie jest to jedyny sposób, w jaki działają eksploratorzy danych. W tym przykładzie użyję innego rodzaju graficznego interfejsu użytkownika, aby wyprowadzić reguły skojarzeń dla zakupów w supermarkecie. Najpierw otwórz aplikację do eksploracji danych i zaimportuj dane. Poniższe kroki pokazują, jak.

1. Uruchom Weka Explorer.

2. Kliknij Otwórz plik i przeglądaj, aby znaleźć dane supermarketu.

3. Po zaimportowaniu danych możesz przeglądać podsumowania danych.

4. Kliknij przycisk Wizualizuj wszystko, aby zobaczyć wizualne podsumowania wszystkich zmiennych w danych.

Zakupy dla asocjacji

Teraz możesz tworzyć reguły asocjacji. Najpopularniejsza technika reguł asocjacyjnych nazywa się Apriori i właśnie jej tutaj użyjesz. Aby utworzyć reguły asocjacji, wykonaj następujące kroki:

1. Kliknij kartę Skojarz. W Associator zobaczysz, że Apriori jest domyślną metodą i że po nazwie następuje pewna tajemnicza notacja. Reprezentuje wybrane opcje.

Możesz wybrać metodę reguły asocjacji. Aby zobaczyć dostępne opcje, kliknij przycisk Wybierz. Eksperci danych najczęściej używają metody Apriori; dla naszych celów tutaj, jest to metoda, której

również powinieneś użyć. Jeśli ta metoda jest już wybrana, po prostu zamknij okno bez wprowadzania zmian.

2. Kliknij przycisk Start, aby utworzyć reguły. Wyniki pojawiają się w obszarze danych wyjściowych asocjatora po prawej stronie. W wynikach zobaczysz listę reguł, zaczynając od reguły użytej jako przykład w sekcji „Strukturyzacja reguł asocjacji” we wcześniejszej części tego rozdziału. Oto lista:

```
Best rules found:
 1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723
             <conf:(0.92)> lift:(1.27) lev:(0.03) [155] conv:(3.35)
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696
             <conf:(0.92)> lift:(1.27) lev:(0.03) [149] conv:(3.28)
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t
             705    <conf:(0.92)> lift:(1.27) lev:(0.03) [150] conv:(3.27)
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746
             <conf:(0.92)> lift:(1.27) lev:(0.03) [159] conv:(3.26)
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779
             <conf:(0.91)> lift:(1.27) lev:(0.04) [164] conv:(3.15)
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t
             725    <conf:(0.91)> lift:(1.26) lev:(0.03) [151] conv:(3.06)
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t
             701    <conf:(0.91)> lift:(1.26) lev:(0.03) [145] conv:(3.01)
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    <conf:(0.91)>
             lift:(1.26) lev:(0.04) [179] conv:(3)
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757
             <conf:(0.91)> lift:(1.26) lev:(0.03) [156] conv:(3)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877
             <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(2.92)
```

Opisałem, jak interpretować te reguły w sekcji „Strukturyzacja reguł skojarzeń”, ale co z wszystkimi uwagami po każdej regule? Co to wszystko znaczy? Spójrz na pierwszą zasadę:

1. ciastka=t mrożonek=t owoców=t ogółem=wysoka 788 ==> chleb i ciasto=t 723

<conf:(0.92)> winda:(1.27) poz:(0.03) [155] konw:(3.35)

Jest to ta sama zasada, którą pokazałem wcześniej plus kilka dodatkowych informacji. Reguła ma numer 1 w rankingu, co oznacza, że była najsilniejszą regułą opartą na metryce użytej do rankingu. W tym przykładzie reguły są klasyfikowane na podstawie ufności, czyli miary tego, jak często reguła okazała się prawdziwa w Twoich danych. Cały tekst po regule zawiera podsumowanie kilku wskaźników wydajności reguły. Nie musisz używać wszystkich metryk!.

Udoskonalanie wyników

Zmiana opcji zmieni Twoje wyniki. Zazwyczaj większość z tych opcji można pozostawić bez ustawień domyślnych, ale dwie, które często warto dostosować, to metryka reguł rankingu i maksymalna liczba reguł do wygenerowania. Aby zmienić te opcje, wykonaj następujące kroki:

1. W obszarze Associate kliknij w dowolnym miejscu białego pola, aby otworzyć edytor opcji.

2. Rozwijane menu dla metricType pozwala wybrać metrykę dla reguł rankingu. Otwórz to menu i wybierz Podnieś.

3. numRules to maksymalna liczba reguł, które chcesz. Wpisz 25. (Patrz rysunek 17-10.) Możesz zmienić to na dowolną liczbę, ale Twoje dane mogą ograniczać liczbę reguł, które faktycznie otrzymujesz.

4. Kliknij OK.

5. Kliknij przycisk Start na karcie Skojarz, aby utworzyć nowy zestaw reguł.

Tym razem poprosiłeś o maksymalnie 25 reguł i dostałeś wszystkie 25. Zmieniłeś także metrykę reguł rankingowych. Porównaj listę z tą, którą otrzymałeś wcześniej, a zobaczysz, że zasady są inne. Oto kilka pierwszych zasad:

```
Best rules found:
 1. fruit=t 2962 ==> bread and cake=t vegetables=t 1791    conf:(0.6) <
              lift:(1.22)> lev:(0.07) [319] conv:(1.27)
 2. bread and cake=t vegetables=t 2298 ==> fruit=t 1791    conf:(0.78) <
              lift:(1.22)> lev:(0.07) [319] conv:(1.63)
 3. bread and cake=t fruit=t 2325 ==> vegetables=t 1791    conf:(0.77) <
              lift:(1.2)> lev:(0.07) [303] conv:(1.56)
```

Reguły te różnią się od tych z tymi samymi rangami w pierwszym przykładzie, ponieważ zmieniono metrykę używaną do uszeregowania reguł, od pewności (miara tego, jak często reguła jest prawdziwa, na zwyżkę), miara tego, jak bardzo konkretna grupa różni się od typowych wzorów. (Więcej informacji na ten temat znajduje się w następnej sekcji.) Kliknięcie przycisku Dziennik wyświetla zapis tego, co zostało zrobione do tej pory w sesji Weka. Zobaczysz takie rzeczy:

Command: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Informacje te są przydatne podczas pracy z interfejsami programistycznymi. Ciąg liter i cyfr po nazwie polecenia reprezentuje ustawienia, takie jak liczba reguł do wygenerowania lub kryteria klasyfikacji reguł. Ta książka nie zajmuje się szczegółami programowania, ale odrobina programowania może przydać się później w twojej karierze, szczególnie w przypadku automatyzacji lub dokładnej kontroli nad procesami.

Zrozumienie metryk

Niektóre aplikacje do eksploracji danych oferują wybór kilku metryk. Są one używane na dwa sposoby:

✓ Diagnostyka: Oceń jakość reguły.

✓ Ranking: Uporządkuj zasady.

Najpopularniejsze metryki reguł asocjacji to wsparcie, zaufanie i wzrost. Oto jak są zdefiniowane:

✓ Wsparcie: Liczba spraw lub proporcja spraw, które obejmują określony zestaw elementów.

✓ Pewność: Proporcja przypadków, w których przewidywany wynik reguły jest prawidłowy.

✓ Wzrost: Stosunek częstotliwości wyniku w określonym zestawie pozycji do częstotliwości tego samego wyniku w pełnym zestawie danych.

Możesz również napotkać mniej popularne dane, takie jak:

✓ Dźwignia: Różnica między proporcją zdarzeń, jeśli zestaw pozycji w danych a proporcją zdarzeń, której można by się spodziewać, gdyby pozycje (poprzednie i następcze) wystąpiły niezależnie.

✓ Przekonanie: Podobny do windy. Stosunek oczekiwanej proporcji przypadków, w których reguła byłaby nieprawidłowa, gdyby pozycje były niezależne, do proporcji przypadków, w których reguła byłaby faktycznie niepoprawna.

Tak, te dwa ostatnie brzmią o wiele bardziej skomplikowanie niż pozostałe. Być może dlatego są rzadziej używane. Nie są trudne do obliczenia (w każdym razie komputer wykonuje całą pracę), ale ich zrozumienie i wyjaśnienie może być trudne. Istnieje jeszcze więcej metryk, które mogą pojawiać się w

aplikacjach do eksploracji danych. Nie przytłaczaj się metrykami! Wsparcie, pewność siebie i winda zajmą Ci długą, długą drogę.

Poszerz swoje horyzonty

Możesz być eksploratorem danych lub możesz być wybitnym eksploratorem danych. Różnica to tylko kwestia wysiłku. Zainwestuj myśl, studiuj i praktykuj w swoim nowym zawodzie, a staniesz się wybitnym praktykiem. W tym rozdziale dowiesz się, jak najlepiej wykorzystać swoje narzędzia, poszerzyć zakres analiz dzięki nowym technikom i poradzić sobie z dzisiejszymi dużymi i złożonymi źródłami danych.

Wyciskanie więcej z tego, co masz

Dostępne obecnie aplikacje do eksploracji danych, nawet te, które są bezpłatne, są niezwykle wydajne i pełne funkcji i opcji. Nie musisz używać ich wszystkich (nawet bardzo doświadczeni eksploratorzy danych rzadko, jeśli w ogóle, korzystają z każdej opcji), ale musisz zapoznać się z głównymi możliwościami i nigdy nie powinieneś przestać badać i próbować nowych rzeczy.

Opanowanie aplikacji do eksploracji danych

Dlaczego oprogramowanie do eksploracji danych przypomina niezmontowane meble? Nikt nie lubi czytać instrukcji dla żadnego z nich. Gdy po raz pierwszy użyjesz aplikacji do eksploracji danych, będzie to wyglądać dość tajemniczo. Każdy produkt ma swój własny, unikalny proces pracy i funkcje, więc nawet jeśli wcześniej korzystałeś z podobnego produktu (lub dwóch lub dziesięciu), nadal może być trudno to rozgryźć. Przykłady zawarte w tej książce dają wyobrażenie o tym, czego się spodziewać, ale nie podają wszystkich szczegółów, które musisz wiedzieć o wybranym produkcie. Jak więc dowiedzieć się, jak korzystać z oprogramowania? Szkolenie to jedno rozwiązanie. Weź udział w bezpośrednim szkoleniu produktowym przeznaczonym dla użytkowników takich jak Ty, jeśli możesz je znaleźć i sobie na to pozwolić. Jeśli jesteś nowicjuszem w eksploracji danych i bardziej ekspertem biznesowym niż ekspertem od analizy danych lub programowania, upewnij się, że zajęcia są przeznaczone dla Ciebie, a nie dla profesorów! Nie stanowi to dużego problemu, jeśli korzystasz z komercyjnego oprogramowania do eksploracji danych, ale wiele bezpłatnych narzędzi zostało stworzonych przez naukowców dla naukowców i możesz mieć problem ze znalezieniem odpowiedniej klasy. Jeśli nie możesz uczestniczyć w zajęciach, poszukaj materiałów do samodzielnej nauki. Większość dostawców udostępnia dokumentację i samouczki dołączone do oprogramowania lub dostępne do bezpłatnego pobrania. Niektóre oferują również pełnometrażowe książki i samouczki wideo. Szukaj online grup użytkowników i materiałów szkoleniowych udostępnionych przez innych użytkowników. Samouczki wideo i inne materiały szkoleniowe stworzone przez oddanych użytkowników są dostępne online dla każdego z produktów przedstawionych w tej książce.

Dostrajanie ustawień

Czasami domyślne ustawienia oprogramowania nie są właściwymi ustawieniami. Odpowiednie ustawienia decydują o różnicy między budowaniem modelu a budowaniem modelu, który jest dobrze dopasowany do konkretnych potrzeb biznesowych. Każde narzędzie w każdej aplikacji do eksploracji danych ma opcje. Możesz pracować przez całe życie i nie używać ich wszystkich, ale to nie znaczy, że nigdy nie powinieneś ich dotykać. Zapoznaj się z narzędziami do eksploracji danych jeden po drugim. Kiedy czujesz się komfortowo z podstawową obsługą narzędzia, spójrz na jego opcje. Jeśli masz szczęście, dokumentacja jasno wyjaśni każdą opcję, co robi, jak z niej korzystać i jak interpretować wyniki. Ale prawdopodobnie nie będziesz miał tyle szczęścia! (Dlatego mamy książki dla opornych.) Więc szukaj lepszych wyjaśnień w książkach, poradach online i innych eksploratorach danych.

Analizuję twoją analizę

Zanim użyjesz modelu, szukasz dowodów, że jest dobry. W rzeczywistości ocena jest jedną z sześciu głównych faz standardowego procesu eksploracji danych CRISP-DM.

Eksperci danych używają wielu metod, aby zrozumieć i ocenić swoje modele. Testowanie może być najważniejszym z nich, ponieważ eksploratorzy danych często lekceważą zasady modelowania, o których statystycy wiedzą, że są ważne. Eksploratorzy danych skąpią w teorii, ale robią duże wrażenie na testach. Będziesz więc testować na nowych danych, testować w terenie i monitorować wydajność swoich modeli również po wdrożeniu. Twoje narzędzia do modelowania będą oferować opcje metryk diagnostycznych, inny sposób zrozumienia i oceny modelu. Aplikacja do eksploracji danych może oferować dodatkowe metody zrozumienia i oceny modeli. Oferty te różnią się w zależności od produktu. Jedną szczególnie cenną funkcją jest analiza wrażliwości, która dostarcza informacji o zmiennych, które mają największy wpływ na predykcje modelu, oraz inne szczegóły dotyczące elementów modelu. Możesz użyć tych informacji, aby

✓ Uprościć modele: Eliminacja zmiennych, które mają niewielki wpływ

✓ Wzbogacić informacje, które przekazujesz decydentom: Zmienne reprezentują rzeczywiste rzeczy, na które decydenci mogą mieć wpływ poprzez swoje działania.

Analiza wrażliwości nie jest dostępna w bezpłatnych aplikacjach do eksploracji danych przedstawionych w tej książce ani w żadnym wolnym oprogramowaniu, które znalazłem. (Niektórzy eksploratorzy danych używają alternatywnej techniki zwanej walidacją krzyżową, ale to podejście jest dość skomplikowane i nie daje takich samych wyników).

Korzystanie z meta-modeli (modeli zespołowych)

Oto coś, czego eksploratorzy danych robią, czego nie robią tradycjonaliści: eksploratorzy danych używają meta-modeli. Meta-model lub model zespołowy jest tym, co otrzymujesz, gdy używasz dwóch lub więcej technik modelowania, aby stworzyć jeden wielki, duży model predykcyjny. Znajdziesz rzeczywiste przypadki modeli predykcyjnych utworzonych przez złożenie stu lub więcej zwykłych modeli. Czasami modele są połączone w łańcuch — dane wyjściowe z jednego modelu są używane jako dane wejściowe dla innego w kolejności — a czasami modele są wykorzystywane do tworzenia prognoz konsensusu w oparciu o zgodność między kilkoma indywidualnymi modelami. Meta-modele są jedną z najbardziej charakterystycznych cech praktyki eksploracji danych. Jeśli myślisz, że brzmi to skomplikowanie, masz rację. Meta-modele mogą być niezwykle złożone. (Patrz rozdział 4, aby zapoznać się z modelem konta, który okazał się zbyt skomplikowany w użyciu, nawet po tym, jak klient zapłacił fortunę, aby go zdobyć). Czy powinieneś spieszyć się z wypróbowaniem meta-modelowania? Może nie. Nie każdy eksplorator danych korzysta z tej techniki. (Nie jestem fanem.) Ale jest to dobrze znana i popularna technika dataminingu, więc powinieneś przynajmniej być jej świadomy.

Poszerzanie zasięgu

Potrafisz swobodnie używać i wyjaśniać kilka ważnych technik eksploracji danych, wiesz, jak korzystać z aplikacji do eksploracji danych i ukończyłeś kilka projektów. Gratulacje! Jesteś teraz odnoszącym sukcesy eksploratorem danych. Pójdź naprzód w swojej karierze w eksploracji danych, odkrywając nowe metody eksploracji danych.

Zwalczanie tekstu

Wiele rosnących zasobów danych na świecie nie jest danymi w tradycyjnym sensie (zmienne ciągłe lub kategorialne). Są tekstem, swobodnymi wyrażeniami ludzkiej myśli. Pomyśl o źródłach tekstowych, które możesz znaleźć w celu uzyskania przydatnych informacji. To tylko kilka z wielu:

✓ Posty w mediach społecznościowych

✓ Prośby o pomoc techniczną

✓ Reklamacje klientów

✓ Akta sądowe

✓ Doniesienia o nowościach

✓ Dokumentacja zdrowotna

✓ Odpowiedzi na pytania ankiety otwartej

✓ E-mail

✓ Wiadomości SMS (teksty)

✓ Życiorysy

Te formy danych nie pomogą w tworzeniu modeli predykcyjnych, ponieważ nie można ich używać jako danych wejściowych do modelowania — przynajmniej nie w ich oryginalnej formie. Jednak metody analizy tekstu umożliwiają konwersję tekstu na bardziej konwencjonalne formy danych, które można wykorzystać do modelowania. Najczęstsze zastosowania eksploracji danych dla tekstu to

✓ Analiza sentymentu: Identyfikacja sentymentu wyrażonego w tekście. Często jest to po prostu pozytywne/negatywne lub pozytywne/neutralne/negatywne, ale czasami używa się wielu subtelnych kategorii. Analiza sentymentu jest niezwykle popularna, ale ponieważ sentyment jest rzeczą nieuchwytną, wyniki są często niezadowalające.

✓ Klasyfikacja: Przypisanie tekstu do kategorii, zwykle na podstawie jego tematyki.

✓ Wyodrębnianie jednostek: Znajdowanie przydatnych fragmentów w tekście, takich jak nazwy lub miejsca.

Wszystkie te techniki umożliwiają tworzenie zmiennych kategorycznych opisujących tematykę lub inne cechy tekstu. Nowe zmienne można wykorzystać do modelowania i innych technik eksploracji danych. Wtedy eksploracja danych staje się eksploracją tekstu. Wiele osób odrzuca eksplorację tekstu, ponieważ wyniki są mniej niż doskonałe, ale to głupie. Nie szukaj perfekcji. Poszukaj praktycznych informacji, których możesz użyć do rozwiązania konkretnego problemu biznesowego związanego z wymiernymi możliwościami uzyskania przychodów lub oszczędności.

Aby jak najlepiej wykorzystać eksplorację tekstu, postępuj zgodnie z następującymi zasadami:

✓ Uważaj na spostrzeżenia: literatura produktowa, doniesienia medialne, a nawet wiele prezentacji konferencyjnych na temat eksploracji tekstu skupia się na uzyskaniu „wglądu”. Obietnica „wglądu” jest tak pociągająca. Zaczynasz myśleć: „Zajrzę głęboko do swoich danych, a ona ujawni mi i mnie samemu swoje najskrytsze sekrety”. Och, seksapil tego. Ale zdobycie wglądu jest zbyt niejasnym celem. Kuszące jest wiara, że można podejść do danych bez planu i wydobyć perły mądrości, ale to nierealne. Zamiast tego, na początku projektu eksploracji tekstów, postaraj się wybrać i określić ilościowo konkretny problem biznesowy do rozwiązania i określić, jakich informacji potrzebujesz, aby go rozwiązać. Badanie przeprowadzone w 2014 r. przez Altaplana Corporation wykazało, że tylko 42% respondentów, którzy korzystali z analizy tekstu, osiągnęło pozytywny zwrot z inwestycji. Jaka jest różnica między tymi, którzy uzyskują pozytywne zwroty, a tymi, którzy tego nie robią? Opierając się na moim długoletnim

doświadczeniu w tej dziedzinie, mogę powiedzieć, że powodem, dla którego większość organizacji nie uzyskuje dodatnich zwrotów z inwestycji, jest to, że nigdy nie zaczynały od realistycznego planu. (Patrz rozdział 6, aby uzyskać więcej informacji na temat planowania.)

✓ Myśl na małą skalę: używaj tylko takiej ilości danych, jaka jest uzasadniona w celu zaspokojenia określonej potrzeby biznesowej. Nie zawracaj sobie głowy kupowaniem rozwiązania Big Data, aby stworzyć wykres kołowy. Chcesz zbudować model predykcyjny? Świetnie. W większości przypadków nie musisz w tym celu wprowadzać każdego bitu danych do narzędzia analitycznego. Użyj tylko próbki - możesz zaoszczędzić fortunę w zasobach i szybciej uzyskać wyniki.

✓ Nie bądź sentymentalny: każdy chce analizy sentymentu. Na pewnym poziomie to sprytne. Wiedza o tym, ile osób wspomina o Twoim produkcie (lub jakimkolwiek innym temacie) nie ma większego znaczenia, jeśli nie wiesz również czegoś o tym, co mówią. Ale ocena sentymentu w tekście to trudna sprawa. Ludzie nie zgadzają się ze sobą konsekwentnie podczas oceny sentymentu tekstu. W rzeczywistości nawet jedna osoba poproszona o ocenę sentymentu wyrażonego w danym fragmencie tekstu przy kilku okazjach często udziela różnych odpowiedzi. Trudno nawet zrobić prezentację na ten temat, ponieważ publiczność niezmiennie daje się wciągnąć w wybieranie poszczególnych przypadków i debatowanie, czy oceny są akceptowalne. Gdzie jest w tym praktyczny wgląd? Zamiast kategorii nastrojów (takich jak pozytywne lub negatywne), poszukaj w swoich danych czegoś lepiej zdefiniowanego i bardziej wykonalnego. Weźmy za przykład Han-Sheong Lai z PayPal, który wykorzystuje eksplorację tekstów do identyfikacji klientów zamierzających zamknąć swoje konta. Czy szuka szerokich kategorii sentymentu pozytywnego i negatywnego? Nie. Szuka osób, które mówią takie rzeczy jak: „Zamknę moje konto”. Możesz się założyć, że znacznie ułatwia to dokładną ocenę ryzyka i kwantyfikację wyników.

Wykrywanie sekwencji

Sekwencja to wzór konkretnych rzeczy dziejących się w określonej kolejności. Oto przykład:

1. Obudź się.

2. Zdejmij piżamę.

3. Załóż odzież roboczą.

4. Załóż buty.

5. Wyjdź do pracy.

Sekwencja jest czymś w rodzaju zestawu przedmiotów, który jest grupą rzeczy, które dzieją się razem. Ale w zestawie przedmiotów kolejność nie ma znaczenia. W sekwencji musisz znać kolejność wydarzeń. Kupujący wchodzi do domu towarowego. Chce kupić czarne spodnie. Wchodzi do sklepu i rozgląda się. Gdzie jest dział odzieży męskiej? Nie widzi tego, więc szuka katalogu. Szuka w kilku niewłaściwych miejscach, zanim znajdzie katalog. Widzi, że po drugiej stronie sklepu znajduje się odzież męska, więc idzie tam, aby znaleźć czarne spodnie. Ten proces trwa i trwa. Nawet po znalezieniu części sklepu, która sprzedaje spodnie, musi znaleźć takie, jakie chce, we właściwym kolorze i rozmiarze. Potem będzie musiał przymierzyć spodnie w przymierzalni, a jeśli będą pasować, będzie musiał za nie zapłacić. W proces zaangażowanych jest wiele etapów i za każdym razem, gdy jeden z nich jest dla kupującego trudny, może wyjść bez kupowania czarnych spodni. Marketerzy bardzo dbają o takie sekwencje, ponieważ mogą mieć dramatyczny wpływ na sprzedaż. Musisz wiedzieć, że istnieje problem, zanim będziesz mógł go naprawić. Analiza sekwencji ujawnia problemy procesowe, a to wprawia w ruch proces rozwiązywania problemu. Możesz także użyć informacji o sekwencji, aby ulepszyć proces. Na

przykład, jeśli zidentyfikujesz wzorzec klientów wchodzących do sklepu, wybierających mięso i idących bezpośrednio do kasjera, aby zapłacić, możesz poszukać sposobów na zachęcenie tych klientów do kupienia czegoś więcej po drodze. Możesz spróbować wystawić sos do steków przy ladzie mięsnej lub popularne wina przy kasie. Sekwencje są również ważne między innymi w modelowaniu finansowym, wykrywaniu włamań i badaniach genetycznych. Jednym z prostych narzędzi do analizy sekwencji jest odmiana współrzędnych równoległych działkach. Jest to diagram przedstawiający kilka działań lub lokalizacji oraz kilka kroków w sekwencji. Każda sekwencja jest reprezentowana przez serię odcinków linii od punktu do punktu. Typowe sekwencje wyróżniają się podczas oglądania fabuły. Inne metody analizy sekwencji mogą nie być dostępne we wszystkich aplikacjach do eksploracji danych, ale często można je spotkać w specjalistycznych narzędziach do analizy dużych zestawów danych, zwłaszcza w przypadku danych internetowych.

Praca z szeregami czasowymi

Szereg czasowy to sekwencja, w której kroki odbywają się w określonych punktach w czasie. Analiza szeregów czasowych jest szeroko wykorzystywana do tworzenia modeli prognozowania sprzedaży i gospodarki, a do tego celu wykorzystuje się wiele metod z klasycznych statystyk. Analiza sygnałów, astronomia i epidemiologia są również ważnymi zastosowaniami do analizy szeregów czasowych. Analiza szeregów czasowych może być trudna. Modele te są często złożone. Co więcej, ilość i jakość danych dostępnych do ich budowania jest często mniejsza niż byś chciał. Nawet w erze Big Data nadal będziesz mieć chwile, kiedy nie będziesz mieć wystarczającej ilości danych. Bardzo łatwo jest stworzyć model szeregów czasowych, który nie ma większego sensu, więc zanim postawisz swój biznes na jeden, zdobądź w zespole kogoś, kto rozumie statystyki tych modeli. Eksploracja danych jest świetna, ale czasami nadal potrzebujesz statystyk.

Przejmowanie Big Data

W dobie komputerów elektronicznych zbieranie danych może być łatwe. Tak więc zbieramy ich dużo. Zbieramy tak dużo, że czasami trudno sobie z tym poradzić. Oto trzy kontra Big Data (jak po raz pierwszy stwierdził Doug Laney w 2001 roku):

✓ Jest go dużo (Głośność).

✓ Więcej nadchodzi szybko (Velocity).

✓ Występuje w wielu formach (różnorodność).

Big Data może być trudne, ale nie zakładaj, że za każdym razem, gdy usłyszysz źródło danych opisane jako Big Data, stajesz przed ogromnym wyzwaniem. Dzisiejsze popularne komputery i oprogramowanie mogą obsłużyć mnóstwo danych. Jeśli czujesz się przytłoczony danymi, oto opcje:

✓ Próbkowanie: często można uzyskać potrzebne informacje, korzystając tylko z niewielkiej części posiadanych danych.

✓ Staranny dobór oprogramowania i sprzętu: wybierz technologię, która pasuje do Twojej sytuacji. Twój ulubiony arkusz kalkulacyjny nie został zaprojektowany do zarządzania petabajtami danych, a najnowsze technologie Big Data to poważna przesada w przypadku 1-terabajtowego źródła danych.

✓ Ciężka praca: Nie, nie, nie ciężka praca! Poważnie, zanim zrobisz coś skomplikowanego, wróć i zastanów się nad innymi opcjami. Ale niektóre sytuacje wymagają pomysłowych rozwiązań.

Niektóre nowsze produkty do analizy danych są określane jako platformy Data Discovery, które zostały zaprojektowane w celu uproszczenia analizy ogromnych ilości (petabajtów) danych. Czy oferują one

nowe i inne rodzaje analizy danych, które nigdy wcześniej nie istniały? Nie! Ale platformy Data Discovery mogą Ci się przydać. Każdy z nich posiada własny, specjalny interfejs użytkownika, zaprojektowany tak, aby skrócić czas potrzebny na analizę bardzo dużych ilości danych i uczynić proces mniej skomplikowanym niż zwykle. Niektóre oferują graficzne interfejsy użytkownika; inne wymagają programowania, ale dostosowują popularne języki programowania (takie jak SQL) do niestandardowych wersji, które umożliwiają użytkownikom przeprowadzanie złożonej analizy danych za pomocą kilku linijek uproszczonego kodu.

Pogodzenie się z Big Data

Big Data jest cenne nie tylko ze względu na swoją ilość. W rzeczywistości bardzo duże zbiory danych zawsze wiążą się z bardzo dużymi problemami. Przechowywanie, konserwacja i zarządzanie bardzo dużymi zbiorami danych nie są proste. A posiadanie dużej ilości danych nie gwarantuje dużej wartości. Wartość dowolnego zestawu danych jest określana przez jakość informacji, które można z niego wyodrębnić. Kluczem do wartości w Big Data są szczegóły. Innymi słowy, wartość Big Data tkwi w małych rzeczach. Każda firma ma ogólne pojęcie o tym, ilu ma klientów, ile łącznie wydaje i być może średnie wydatki na klienta. Ale jeśli wszystko, co wiesz, to średnia, co zamierzasz zrobić - traktować każdego klienta jako przeciętnego? Gdybyś mógł osobiście poznać każdego klienta i poznać każdą osobę, nie pomyślałbyś o nikim jako o przeciętnym. Znałbyś nawyki każdej osoby. Wiesz, że Maria Perez robi zakupy dla siebie co tydzień i od czasu do czasu kupuje prezent, podczas gdy Laura Carter robi zakupy dla swojej pięcioosobowej rodziny, a Lily Yu sama nie używa twoich produktów, ale często kupuje je dla swoich rodziców. Znasz pory dnia, kiedy każda osoba woli robić zakupy, bez względu na to, czy robi to zrelaksowany, czy pospieszny klient, oraz produkty, które preferuje każda osoba. Ponieważ poznałbyś swoich klientów jako jednostki, traktowałbyś ich jak jednostki. Poinformuj Marię, że oferujesz pakowanie prezentów. Skierowałbyś Laurę do ekonomicznego, rodzinnego opakowania jej ulubionego produktu. Upewniłbyś się, że Lily wybrała pojemnik, który jej rodzice mogli łatwo otworzyć.

Prowadzenie analiz predykcyjnych za pomocą Big Data

Obietnica Big Data tkwi w szczegółach. Chcesz, aby dane dawały Ci informacje, które uzyskasz, gdybyś osobiście obserwował każdego klienta. Chcesz wiedzieć, co robi każda osoba. Chcesz wiedzieć, jak każdy z nich reaguje na różne rzeczy – oferowane produkty, ceny, prezentacje i tak dalej.

Zyskujesz wartość z danych tylko wtedy, gdy robisz z nimi coś wartościowego. W bezpośredniej interakcji z klientem wykorzystujesz swoją wiedzę o kliencie, aby przedstawiać odpowiednie sugestie, a im lepsze są sugestie, tym więcej klient kupuje, zwraca i poleca Cię innym. Najlepsze dane dostarczają przydatnych informacji. Informacja stwarza możliwości. Wartość jest wprowadzana, gdy używasz informacji do podjęcia sensownego działania. Więc co to mówi o wyborze zestawów danych dla aplikacji Big Data? Spójrz na proces. Po pierwsze, potrzebujesz celu. Co chcesz osiągnąć? Następnie musisz znać swoje możliwości działania. Czy możesz oferować nowe produkty lub zmienić ofertę, którą oferujesz, czy też musisz pracować w ramach tego, co masz teraz? Czy możesz opracować nowe reklamy, nowe oferty? Teraz wyobraź sobie, że masz ten sam cel i te same opcje w sytuacji twarzą w twarz. Jakich informacji potrzebujesz? Wiedząc o tym, jesteś gotowy do szukania źródeł danych, które odpowiadają Twoim potrzebom. Oto przykład. Twoje sklepy stacjonarne są zatłoczone w godzinach szczytu – tak zatłoczone, że klienci często odchodzą z frustracją – podczas gdy innym razem sklepy są prawie puste. Sprzedajesz poniżej swojego potencjału z powodu porzucania koszyka i nieprzyciągania klientów przez cały dzień. Jaki jest Twój cel? Zwiększ przychody dzięki lepszej dystrybucji aktywności w ciągu dnia. Masz budżet marketingowy, uprawnienia do wysyłania reklam drukowanych i e-mailowych oraz uprawnienia do składania ofert specjalnych za pomocą kuponów i innych programów

promocyjnych. Masz również pewien wpływ na planowanie personelu i procedury kasowe. Stwarza to możliwości działania. Teraz wyobraź sobie, że jesteś w sklepie i obserwujesz klientów. Jakie przydatne fakty możesz zaobserwować? Niektórzy kupujący zwykle robią zakupy poza godzinami szczytu. Kim oni są? Inni zwykle robią zakupy w godzinach szczytu. Czemu? Czy niektórzy kupujący różnią się porami, w których przychodzą do sklepu? Kto się poddaje i wychodzi? Co ci ludzie zamierzali kupić? Co możesz dowiedzieć się o przyczynach zachowania każdego kupującego? Musisz wprowadzić te informacje w czyn. Być może zauważyłeś, że niektórzy kupujący, którzy przychodzą w ruchliwych porach, po prostu nie są świadomi czasów, kiedy sklep nie jest tak zajęty. Kampania informacyjna może im pomóc. Może to być tak proste, jak umieszczenie znaków w sklepie lub dodanie tych informacji do regularnego okólnika. Inni mogą zostać nakłonieni do przeniesienia zakupów poza godziny szczytu, jeśli opłacisz im czas, korzystając z rabatu lub oferty specjalnej. Jeśli chodzi o ludzi, którzy już robią zakupy w godzinach ciszy, nie czerpiesz korzyści z oferowania im zachęt za to, co już robią, ale może możesz ich zmotywować do kupowania więcej. Jeśli wiesz, co kupują, możesz zaoferować kupon na produkt, którego nie wypróbowali, lub ofertę na większą ilość ulubionego produktu. Nie możesz osobiście rozmawiać z każdym klientem. Nie możesz śledzić wszystkich dookoła i obserwować. Ale możesz mieć dostęp do danych, które dostarczają wiele takich samych informacji. Jeśli masz do czynienia z wieloma osobami i wieloma szczegółami, mówisz o Big Data, rodzaju Big Data, który napędza zyskowną analitykę predykcyjną. Gdzie można znaleźć szczegółowe informacje o zachowaniu swoich klientów i potencjalnych klientów? Zacznij od danych, które już posiadasz. Twoje rekordy transakcji to skarbnica danych behawioralnych. Jeśli prowadzisz interesy online, będziesz mieć dzienniki internetowe zawierające szczegółowe informacje o zachowaniach zakupowych, w tym szczegóły dotyczące zachowania osób niekupujących. Dopiero po dokładnym zbadaniu możliwości wewnętrznych źródeł danych powinieneś spojrzeć poza swoje mury, być może w celu uzyskania danych demograficznych lub informacji o kredytach, posiadaniu domu lub innych czynnikach, które mogą wpływać na zachowanie. Kiedy masz jasne pojęcie o tym, co chcesz wiedzieć i ograniczeniach własnych danych, możesz selektywnie i sprytnie kupować dane, które wypełniają puste pola

Metody mieszania dla najlepszych rezultatów

Żadna technika analizy danych nie jest lepsza od wszystkich innych, ani żadna jedna klasa analizy danych (eksploracja danych, klasyczne statystyki, badania operacyjne itd.) nie jest lepsza od wszystkich innych. Najlepsze zrozumienie tematu uzyskasz, przyglądając się mu na wiele sposobów, łącząc różne techniki eksploracji danych z metodami z innych dyscyplin oraz udoskonalając swoje zrozumienie poprzez ciągłe dochodzenie w czasie. Nie ograniczaj swoich horyzontów tylko tym, co możesz zrobić za pomocą jednej aplikacji do eksploracji danych, jednego źródła danych lub jednej osoby!

✓ Szukaj nowych narzędzi: Być może wybrałeś swoją aplikację do eksploracji danych, ponieważ jest to ta, którą zapewnia Twój pracodawca, co możesz otrzymać za darmo lub która jest najbardziej znana. To wszystko są uzasadnione względy, ale nie powody, aby nie wiedzieć, co jest dostępne, stare lub nowe. Twoje potrzeby będą się zmieniać wraz z upływem czasu, podobnie jak opcje.

✓ Dowiedz się o technikach stosowanych przez innych analityków danych: analitycy danych używają wielu tytułów, a każdy z nich stosuje inny zestaw technik. Dowiedz się, co robią Twoi koledzy i dlaczego, ponieważ możesz użyć tych samych metod i uzyskać cenne nowe informacje, aby zaspokoić własne potrzeby biznesowe.

✓ Spotykaj się i współpracuj z innymi: Jesteś tylko jedną osobą. Zróżnicowany zespół może zastosować szeroki zakres wiedzy i umiejętności do Twojego problemu biznesowego. Graj ładnie z innymi, ponieważ eksploracja danych nie dotyczy ciebie; chodzi o rozwiązywanie problemów biznesowych. (Zrozumiesz i staniesz się znacznie lepszym eksploratorem danych!)