

Krok 13: Eksploracja danych

W tej części omówiono następujące tematy:

- * Rzeczy do rozważenia przy eksploracji danych
- * Tradycyjne techniki analizy kontra eksploracja danych
- * Znaczenie eksploracji danych
- * Źródła danych do eksploracji danych
- * Pięć najpopularniejszych technik eksploracji danych: wykrywanie powiązań, wykrywanie wzorców sekwencyjnych, klasyfikacja, grupowanie i prognozowanie
- * Operacje eksploracji danych, takie jak modelowanie predykcyjne i klasyfikacyjne, analiza łączy, segmentacja bazy danych i wykrywanie odchyleń
- * Zastosowania eksploracji danych w obszarach zarządzania rynkiem, wykrywania oszustw, zarządzania ryzykiem, usług finansowych i dystrybucji
- * Krótkie opisy działań związanych z eksploracją danych, rezultatów wynikających z tych działań oraz zaangażowanych ról
- * Ryzyko niewykonania kroku 13

Rzeczy do rozważenia

Pytania marketingowe

- * Czy wiemy, jakie mamy ogólne klasy klientów?
- * Czy istnieją podklasy klientów o podobnych wzorcach behawioralnych? Czy możemy używać ukierunkowanych komunikatów marketingowych dla tych klientów?
- * Czy wiemy, co łączy naszych najlepszych klientów? Czy jest wzór, który zweryfikuje nasze przeczucie, że jeśli zaoferujemy nowe produkty lub usługi „just in time” naszym najlepszym klientom, to je kupią?
- * Czy wiemy, jak zatrzymać naszych najlepszych klientów? Jak możemy przewidzieć, którzy klienci z większym prawdopodobieństwem nas opuszczą?
- * Jak możemy sprzedawać więcej naszym obecnym klientom? Którzy z naszych klientów chętniej kupią od nas więcej produktów lub usług?
- * Czy mamy klientów, którzy kosztują nas pieniądze?
- * Czy podejrzewamy nieuczciwe działania, które należy wykryć?

Dane

- * Czy nasze dane są wystarczająco czyste do eksploracji danych?
- * Czy dane są zrozumiałe? Czy będzie poprawnie używany i interpretowany?
- * Czy jest prawidłowo zakodowany do eksploracji danych?

* Czy dane są prawidłowo zorganizowane do eksploracji danych?

Narzędzie do eksploracji danych

* Jaki rodzaj narzędzia do eksploracji danych jest odpowiedni dla naszej organizacji?

* Jakie kryteria należy wziąć pod uwagę, oceniając narzędzia do eksploracji danych?

* Jak określimy zwrot z inwestycji (ROI) dla danych?

Inicjatywa eksploracji i narzędzie do eksploracji danych?

Rekrutacja

* Czy mamy statystyków lub wykwalifikowanych analityków, którzy potrafią zinterpretować?

wyniki eksploracji danych?

* Czy będziemy musieli zatrudnić dodatkowych statystyków do eksploracji danych?

* Czy administrator bazy danych będzie dostępny do tworzenia i ładowania baz danych eksploracji danych? Czy będziemy potrzebować pełnoetatowego administratora bazy danych? W niepełnym wymiarze godzin?

Wiele organizacji zgromadziło ogromne ilości danych w swoich systemach operacyjnych. Dane te stanowią potencjalne źródło cennych informacji biznesowych, które można wydobyć. Modele analityczne można generować w celu znalezienia wzorców w danych i umożliwienia wykorzystania informacji w celu uzyskania przewagi konkurencyjnej. Daje to menedżerom i kadrze kierowniczej informacje potrzebne do podjęcia działań, umożliwiając im zwiększanie zysków, redukcję kosztów, tworzenie innowacyjnych strategii produktowych i zwiększanie udziału w rynku.

Definiowanie eksploracji danych

Możliwości eksploracji danych nie są czymś, co można kupić z półki. Eksploracja danych wymaga zbudowania aplikacji wspomagającej podejmowanie decyzji BI, w szczególności aplikacji do eksploracji danych, przy użyciu narzędzia do eksploracji danych. Aplikacja do eksploracji danych może następnie wykorzystywać wyrafinowaną mieszankę klasycznych i zaawansowanych komponentów, takich jak sztuczna inteligencja, rozpoznawanie wzorców, bazy danych, tradycyjne statystyki i grafika, aby prezentować ukryte relacje i wzorce znalezione w puli danych organizacji. Eksploracja danych to analiza danych z zamiarem odkrycia klejnotów ukrytych informacji w ogromnej ilości danych, które zostały przechwycone w normalnym toku prowadzenia firmy. Eksploracja danych różni się od konwencjonalnej analizy statystycznej, jak wskazano w tabeli 13.1. Obaj mają mocne i słabe strony.

Analiza statystyczna:

* Statystycy zwykle zaczynają od hipotezy.

* Statystycy muszą opracować własne równania, aby dopasować je do swojej hipotezy.

* Analiza statystyczna wykorzystuje tylko dane liczbowe.

* Podczas analizy statystycy mogą wyszukiwać i filtrować brudne dane.

* Statystycy interpretują własne wyniki i przekazują je menedżerom i dyrektorom biznesowym.

Eksploracja danych:

- * Eksploracja danych nie wymaga hipotezy.
- * Algorytmy eksploracji danych mogą automatycznie tworzyć równania.
- * Eksploracja danych może wykorzystywać różne typy danych (np. tekst, głos), a nie tylko dane liczbowe.
- * Eksploracja danych zależy od czystych, dobrze udokumentowanych danych.
- * Wyniki eksploracji danych nie są łatwe do interpretacji. Statystyk nadal musi być zaangażowany w analizę wyników eksploracji danych i przekazywanie wyników menedżerom i dyrektorom biznesowym.

Przykład analizy oszustw ubezpieczeniowych

Tradycyjna technika analizy

* Analityk zauważa wzorec zachowania, który może wskazywać oszustwa ubezpieczeniowe. W oparciu o tę hipotezę analityk tworzy zestaw zapytań w celu ustalenia, czy to zaobserwowane zachowanie rzeczywiście stanowi oszustwo. Jeśli wyniki nie są rozstrzygające, analityk zaczyna od zmodyfikowanej lub nowej hipotezy i większej liczby zapytań. Ten proces jest nie tylko czasochłonny, ale również zależy od subiektywnej interpretacji wyników przez analityka. Co ważniejsze, proces ten nie znajdzie żadnych wzorców oszustw, których analityk już nie podejrzewa.

Eksploracja danych oparta na odkrywaniu

* Analityk konfiguruje aplikację do eksploracji danych, a następnie „uczy” ją w poszukiwaniu wszelkich nietypowych wzorców, trendów lub odchyleń od normy, które mogą stanowić oszustwo ubezpieczeniowe. Wyniki eksploracji danych ujawniają różne sytuacje, które analityk może dokładniej zbadać. W ramach dochodzenia uzupełniającego analityk może następnie użyć zapytań opartych na weryfikacji. Razem te wysiłki mogą pomóc analitykowi zbudować model przewidujący, którzy klienci lub potencjalni klienci mogą popełnić oszustwo.

Przykład analizy segmentacji rynku

Tradycyjna technika analizy

* Analityk chce zbadać zachowania zakupowe znanych klas klientów (np. emerytowanych nauczycieli szkolnych, młodych specjalistów z miast) w celu zaprojektowania ukierunkowanych programów marketingowych. Najpierw analityk wykorzystuje znane cechy charakterystyczne tych klas klientów i próbuje je posortować w grupy. Po drugie, bada zachowania zakupowe wspólne dla każdej grupy. Analityk powtarza ten proces, dopóki nie będzie zadowolony z końcowego pogrupowania klientów.

Eksploracja danych oparta na odkrywaniu

* Narzędzie do eksploracji danych „bada” bazę danych przy użyciu techniki grupowania w celu zidentyfikowania wszystkich grup klientów o różnych wzorcach zakupowych. Po wydobyciu danych i prezentowaniu są grupy, analityk może używać różnych zapytań, narzędzia raportowania i analizy wielowymiarowej do analizy wyników.

Znaczenie eksploracji danych

Eksploracja danych oparta na odkrywaniu pozwala znaleźć odpowiedzi na pytania, których decyzyci nie potrafią zadać. Ze względu na tę potężną funkcję eksploracja danych jest ważnym elementem analizy biznesowej. Można nawet powiedzieć, że eksploracja danych, zwana także odkrywaniem

wiedzy, to przełom w dostarczaniu analityki biznesowej decydującym strategicznym. Na pierwszy rzut oka to twierdzenie może wydawać się przesadne. W końcu wiele obecnych aplikacji wspomagających podejmowanie decyzji zapewnia analizę biznesową i wgląd.

* Systemy informacji dla kadry kierowniczej (EIS) umożliwiają menedżerom wyższego szczebla monitorowanie, badanie i zmianę wielu aspektów ich działalności biznesowej.

* Narzędzia do tworzenia zapytań i raportowania dają analitykom biznesowym możliwość badania wydajności firmy i zachowania klientów. *Narzędzia statystyczne umożliwiają statystykom przeprowadzanie wyrafinowanych badań zachowania firmy.

* Nowe narzędzia wielowymiarowego przetwarzania analitycznego online (OLAP) zapewniają możliwość wykonywania analiz typu „co, jeśli” i przyglądania się dużej liczbie współzależnych czynników związanych z problemem biznesowym.

Wiele z tych narzędzi współpracuje z aplikacjami BI i może przeszukiwać ogromne ilości danych. Biorąc pod uwagę tę obfitość narzędzi, czym różni się eksploracja danych oparta na odkrywaniu? Duża różnica polega na tym, że tradycyjne techniki analizy, nawet te wyrafinowane, polegają na tym, że analityk wie, czego szukać w danych. Analityk tworzy i uruchamia zapytania w oparciu o pewne hipotezy i domysły dotyczące możliwych relacji, trendów i korelacji, o których sądzi się, że są obecne w danych. Podobnie, kadra kierownicza polega na widokach biznesowych wbudowanych w narzędzie EIS, które mogą badać tylko te czynniki, do których przeglądu narzędzie zostało zaprogramowane. Ponieważ problemy stają się coraz bardziej złożone i wymagają analizy większej liczby zmiennych, te tradycyjne techniki analizy mogą okazać się niewystarczające. W przeciwieństwie do tego, eksploracja danych oparta na odkrywaniu wspiera bardzo subtelne i złożone badania.

Źródła danych do eksploracji danych

Docelowe bazy danych BI są popularnymi źródłami aplikacji do eksploracji danych. Zawierają bogactwo danych wewnętrznych, które zostały zebrane i skonsolidowane ponad granicami biznesowymi, zweryfikowane i oczyszczone w procesie ekstrakcji/transformacji/ładowania (ETL). Bazy danych docelowych BI mogą również zawierać cenne dane zewnętrzne, takie jak regulacje, dane demograficzne czy informacje geograficzne. Połączenie danych zewnętrznych z wewnętrznymi danymi organizacyjnymi stanowi doskonałą podstawę do eksploracji danych. Wadą wielowymiarowych docelowych baz danych BI jest to, że ponieważ dane zostały podsumowane, ukryte wzorce danych, relacje danych i asocjacje danych często nie są już widoczne w tej puli danych. Na przykład narzędzie do eksploracji danych może nie być w stanie wykonać typowego zadania eksploracji danych, jakim jest analiza koszyka rynkowego (nazywana również wykrywaniem powiązań, opisana w następnej sekcji) na podstawie podsumowanych danych sprzedaży, ponieważ niektóre szczegółowe wzorce danych dotyczące każdej sprzedaży mogą zgubić się w podsumowaniu. Dlatego też pliki operacyjne i bazy danych są również popularnymi źródłami aplikacji do eksploracji danych, zwłaszcza dlatego, że zawierają szczegółowe dane na poziomie transakcji z mnóstwem ukrytych wzorców danych, relacji danych i skojarzeń danych.

Zachowaj ostrożność przy wyciągach z systemów operacyjnych, ponieważ dane mogą zawierać wiele duplikatów, niespójności i błędów oraz mogą zniekształcać wyniki eksploracji danych.

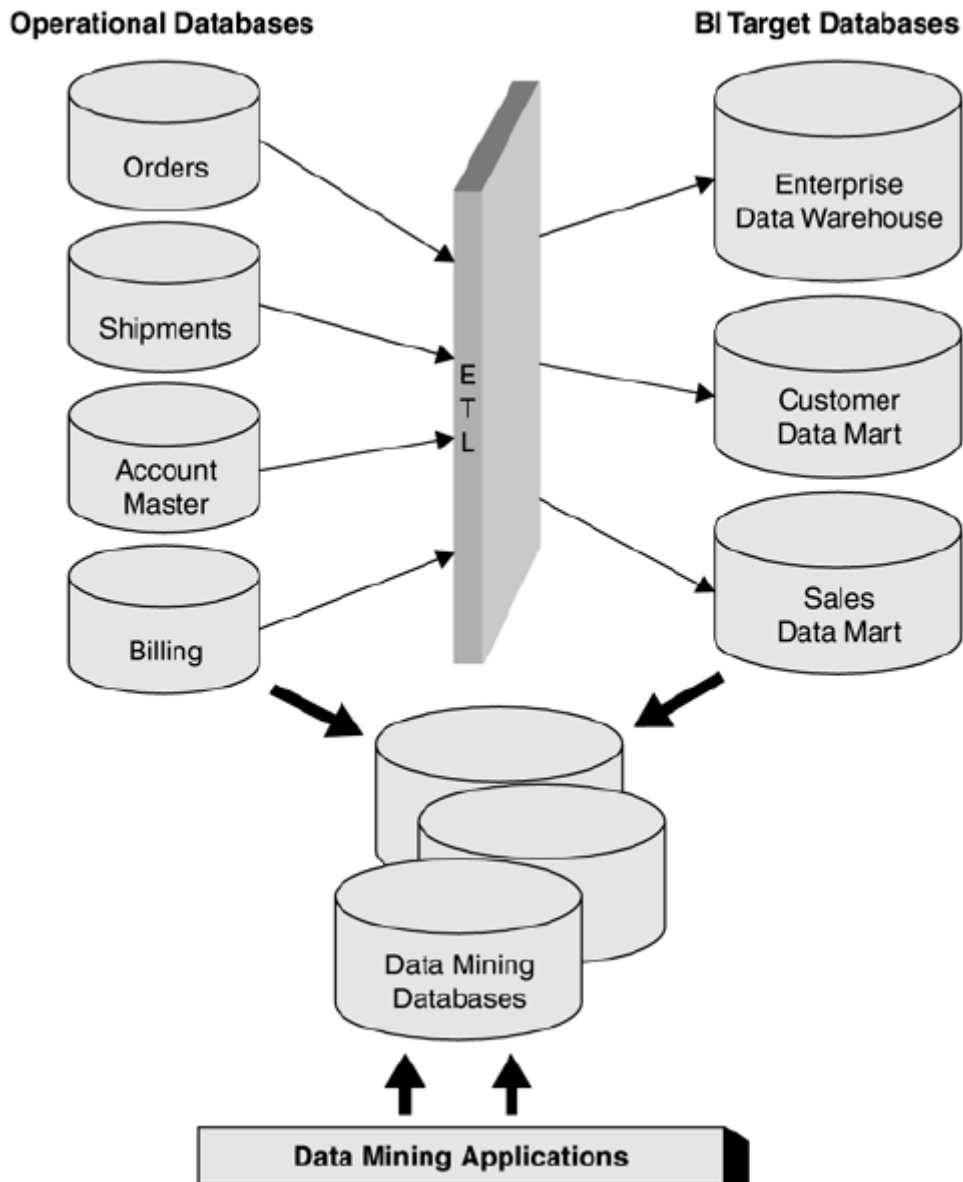
Narzędzia do eksploracji danych mogą teoretycznie uzyskiwać dostęp do operacyjnych baz danych i docelowych baz danych BI bez uprzedniego budowania baz danych eksploracji danych, o ile struktury bazy danych są obsługiwane przez narzędzie (np. relacyjne, takie jak Oracle, hierarchiczne, takie jak IMS, lub nawet płaskie pliki, takie jak VSAM). Nie jest to jednak wskazana praktyka z kilku powodów.

* Pula danych musi być w stanie zmieniać się dla różnych przebiegów eksploracji danych, takich jak porzucenie regionu sprzedaży lub ograniczenie typu produktu do określonego celu eksploracji. Zmiana zawartości danych operacyjnych lub docelowych baz danych BI nie jest możliwa.

* Operacje eksploracji danych miałyby wpływ na wydajność operacyjnych i docelowych baz danych BI. Jest to niedopuszczalne w przypadku operacyjnych baz danych i niepożądane w przypadku docelowych baz danych BI.

* Operacja eksploracji danych może wymagać szczegółowych danych historycznych. Operacyjne bazy danych nie przechowują danych historycznych, a docelowe bazy danych BI często nie mają pożądanego poziomu szczegółowości. W celu wyodrębnienia żądanych danych konieczne może być przywrócenie i połączenie taśm archiwalnych.

Dlatego organizacje często w razie potrzeby wyodrębniają dane do eksploracji danych ze swoich docelowych baz danych BI oraz z plików operacyjnych i baz danych do baz danych specjalnego przeznaczenia



Techniki eksploracji danych

Techniki eksploracji danych to specyficzne implementacje algorytmów wykorzystywanych w operacjach eksploracji danych. Pięć najpopularniejszych technik eksploracji danych zostało krótko opisanych poniżej.

Odkrywanie skojarzeń. Ta technika eksploracji danych służy do identyfikacji zachowania określonych zdarzeń lub procesów. Odnajdywanie skojarzeń łączy wystąpienia w ramach jednego zdarzenia. Przykładem może być odkrycie, że mężczyźni, którzy kupują kawę najwyższej jakości, trzy razy częściej kupują importowane cygara niż mężczyźni, którzy kupują kawę standardowych marek. Wykrywanie skojarzeń opiera się na regułach, które są zgodne z tą ogólną formą: „Jeśli element A jest częścią zdarzenia, to X procent czasu (współczynnik ufności), element B jest częścią tego samego zdarzenia”. Na przykład:

* Jeśli klient kupuje przekąski, istnieje 85-procentowe prawdopodobieństwo, że kupi również napoje bezalkoholowe lub piwo.

* Jeśli dana osoba kupuje wakacyjne bilety lotnicze dla całej rodziny, istnieje 95-procentowe prawdopodobieństwo, że wynajmie pełnowymiarowy samochód w miejscu wakacji.

Za pomocą skanerów sklepy detaliczne wykorzystują tę technikę eksploracji danych, aby znaleźć wzorce zakupów w sklepach spożywczych. Ze względu na kontekst sklepu spożywczego odkrywanie skojarzeń jest czasami nazywane analizą koszyka rynkowego.

Sekwencyjne wykrywanie wzorców. Ta technika eksploracji danych jest podobna do wykrywania skojarzeń, z tą różnicą, że sekwencyjne wykrywanie wzorców łączy zdarzenia w czasie i określa, w jaki sposób elementy są ze sobą powiązane w czasie. Na przykład sekwencyjne wykrywanie wzorców może przewidywać, że osoba, która kupuje pralkę, może również kupić suszarkę do ubrań w ciągu sześciu miesięcy z prawdopodobieństwem 0,7. Aby zwiększyć szanse powyżej przewidywanego 70-procentowego prawdopodobieństwa, sklep może zaoferować każdemu kupującemu 10-procentową zniżkę na suszarkę do ubrań w ciągu czterech miesięcy od zakupu pralki.

Klasyfikacja. Technika klasyfikacji jest najczęstszym zastosowaniem eksploracji danych. Klasyfikacja dotyczy zachowania i atrybutów z góry określonych grup. Do grup mogą należeć osoby często podróżujące samolotami, osoby wydające dużo pieniędzy, lojalni klienci, osoby, które odpowiadają na bezpośrednie kampanie pocztowe lub osoby z częstymi problemami z kręgosłupem (np. osoby, które codziennie jeżdżą na długich dystansach). Narzędzie do eksploracji danych może przypisywać klasyfikacje do nowych danych, analizując istniejące dane, które zostały już sklasyfikowane, i wykorzystując te wyniki do wywnioskowania zestawu reguł. Zestaw reguł jest następnie stosowany do wszystkich nowych danych, które mają zostać sklasyfikowane. Ta technika często wykorzystuje nadzorowaną indukcję, która wykorzystuje mały zestaw szkoleniowy już sklasyfikowanych rekordów w celu określenia dodatkowych klasyfikacji. Przykładem takiego zastosowania jest odkrycie cech klientów, którzy są (lub nie są) skłonni do zakupu określonego rodzaju produktu. Ta wiedza zaowocowałaby obniżeniem kosztów promocji i mailingów bezpośrednich.

Grupowanie

Technika grupowania służy do wykrywania różnych grupowań w danych. Klastrowanie jest podobne do klasyfikacji, z tym wyjątkiem, że na początku działania narzędzia do eksploracji danych nie zdefiniowano jeszcze żadnych grup. Technika grupowania często wykorzystuje sieci neuronowe lub metody statystyczne. Grupowanie dzieli elementy na grupy na podstawie podobieństw znalezionych przez narzędzie do eksploracji danych. W ramach klastra członkowie są bardzo podobni, ale same klastry są bardzo odmienne. Klastrowanie jest używane do rozwiązywania problemów, takich jak wykrywanie defektów produkcyjnych lub znajdowanie grup koligacji dla kart kredytowych.

Prognozowanie

Prognozowana technika eksploracji danych ma dwa rodzaje: analiza regresji i odkrywanie sekwencji czasowych.

* Analiza regresji wykorzystuje znane wartości danych do przewidywania przyszłych wartości lub przyszłych zdarzeń na podstawie historycznych trendów i statystyk. Na przykład wielkość sprzedaży akcesoriów do samochodów sportowych można prognozować na podstawie liczby sprzedanych samochodów sportowych w zeszłym miesiącu.

* Odkrywanie sekwencji czasowych różni się od analizy regresji tym, że prognozuje tylko wartości danych zależne od czasu. Na przykład określa wskaźniki wypadków w okresie wakacyjnym na podstawie liczby wypadków, które miały miejsce w tym samym okresie wakacyjnym w poprzednich latach. Własnością czasu może być:

- Tydzień pracy a tydzień kalendarzowy

- Wakacje

- Pory roku

- Zakresy dat i interwały dat

Operacje eksploracji danych

Narzędzia do eksploracji danych umożliwiają statystykom budowanie modeli analitycznych, które następnie wykorzystują podczas operacji eksploracji danych. Silnik predykcyjny prosi o listę kryteriów wejściowych i podąża za krokami oraz relacjami z modelu analitycznego w celu określenia najbardziej prawdopodobnych predykcji. Wynikiem operacji eksploracji danych są tabele i pliki załadowane danymi analitycznymi, do których można uzyskać dostęp za pomocą narzędzi zapytań i raportowania. Poniżej opisano cztery główne operacje eksploracji danych.

Modelowanie predykcyjne i klasyfikacyjne

Ta operacja eksploracji danych służy do prognozowania konkretnego zdarzenia. Zakłada, że analityk ma konkretne pytanie, które chce zadać. Model dostarcza odpowiedzi poprzez przypisanie rang, które wskazują prawdopodobieństwo wystąpienia pewnych klas. Na przykład, jeśli analityk bankowy chce przewidzieć, którzy klienci prawdopodobnie odejdą, musi przygotować się do modelowania predykcyjnego, wprowadzając dane o dwóch typach klientów do narzędzia do eksploracji danych.

1. Dane klientów, które wskazują, którzy klienci już wyszli. Te dane są nazywane „złymi” danymi.

2. Dane klientów, które wskazują, którzy klienci zostali i są klientami długoletnimi. Te dane są nazywane „dobrymi” danymi.

Następnie narzędzie przesiewa dane, aby odkryć zmienne, które identyfikują klasy profili typowych klientów, którzy odchodzą, oraz klasy profili typowych klientów, którzy zostają. Wyniki analizy mogą brzmieć: „Klientka w wieku powyżej 40 lat, która ma dochód większy niż 150 000 USD rocznie i posiada własny dom, ma 35 procent szans na opuszczenie banku”. Typowe pytania sondujące do predykcyjnego eksploracji danych to te, które szukają powiązań, wzorców, trendów i faktów w celu podejmowania decyzji. Na przykład:

* Które oferty zachęcą klientów do większych zakupów? (Tendencja)

* Do jakich klientów powinien trafić nowy produkt? (Stowarzyszenie)

* Jakie są oznaki nieuczciwej działalności? (Wzorzec)

* Którzy klienci mają większe ryzyko kredytowe? (Fakt)

Analiza linków

Operacja eksploracji danych analizy łączy to zbiór algorytmów matematycznych i technik wizualizacji, które identyfikują i wizualnie przedstawiają powiązania między poszczególnymi rekordami w bazie danych. Jest to związane z technikami wyszukiwania skojarzeń i sekwencyjnego wykrywania wzorców. Na przykład analiza linków może określić, które produkty zwykle sprzedają się razem (np. płatki zbożowe i mleko).

Segmentacja bazy danych

Ta operacja eksploracji danych to zestaw algorytmów, które grupują podobne rekordy w jednorodne segmenty. Jest to związane z techniką eksploracji danych w klastrach. Takie grupowanie jest często pierwszym etapem selekcji danych, przed rozpoczęciem innych operacji eksploracji danych. Na przykład segmentacja bazy danych może grupować pasażerów linii lotniczych jako pasażerów często podróżujących samolotami lub pasażerów okazjonalnych.

Wykrywanie odchyleń

Operacja eksploracji danych wykrywania odchyleń to zestaw algorytmów, które wyszukują rekordy, które wykraczają poza pewne oczekiwania lub normy, a następnie sugerują przyczyny anomalii. Podczas gdy wykrywanie odchyleń służy głównie do wykrywania oszustw, inne zastosowania obejmują śledzenie potencjalnych przyczyn spadków liczby klientów lub sprzedaży. Na przykład: „Klienci, którzy dokonywali częstych zakupów, ale od dłuższego czasu niczego nie kupowali, zostali przeniesieni przez swoje firmy lub wyprowadzili się z okolicy”.

Zastosowania eksploracji danych

Należy pamiętać, że pomimo wszystkich olśniewających technologii, eksploracja danych musi być napędzana silnymi potrzebami biznesowymi, aby uzasadnić nakłady czasu i pieniędzy. Jednym z typowych bodźców biznesowych angażujących się w eksplorację danych jest zdobywanie udziału w rynku. Można to osiągnąć wprowadzając nowe produkty lub odbierając udział w rynku konkurentom. W obu przypadkach aplikacja do eksploracji danych może pomóc Ci zdecydować, jak najlepiej osiągnąć swoje cele. Istnieje wiele rodzajów aplikacji do eksploracji danych. Pięć najczęstszych z nich zostało krótko opisanych poniżej.

*** Zarządzanie rynkiem**

- Sprzedaż krzyżowa: Zidentyfikuj możliwości sprzedaży krzyżowej wśród obecnych klientów, którzy prawdopodobnie dokonają zakupu w wyniku kampanii i promocji wysyłanych pocztą bezpośrednią (a tym samym zminimalizują koszty sprzedaży).
- Klienci oszukujący: określ, którzy klienci prawdopodobnie zmienią markę, korzystając z analizy podatności, która tworzy predykcyjny model zachowania, aby firma mogła opracować strategię zatrzymania tych klientów.
- Promocje i kampanie: Rozróżnij naturalne grupy na rynku, takie jak kluczowe okresy sprzedaży dla danych produktów, wykonując analizę segmentacji rynku w celu dopracowania promocji i kampanii.
- Poszukiwanie: Klasyfikuj grupy potencjalnych klientów, aby znaleźć sposoby prowadzenia marketingu docelowego dla każdej grupy.
- Analiza koszyka rynkowego: Oceń, które przedmioty ludzie kupują razem podczas wizyty w supermarkecie lub sklepie, korzystając z analizy koszyków rynkowych na podstawie danych z punktów sprzedaży. Następnie użyj tych informacji, aby pogrupować produkty na wystawach sklepowych, dostosować zapasy oraz wycenić i promować produkty.

*** Wykrywanie oszustw**

- Oszustwa związane z kartami kredytowymi: Izoluj oszustwa związane z kartami kredytowymi, identyfikując znaczące wzorce transakcji, a także odstępstwa od tych wzorców. Użyj tego modelu, aby przewidzieć wiarygodność wnioskodawcy.
- Oszustwa związane z kartami telefonicznymi: Określ sytuacje, które wydają się podejrzane i mogą wskazywać na oszustwo.

- Oszustwa ubezpieczeniowe: analizuj duże zbiory danych roszczeń ubezpieczeniowych, aby zidentyfikować możliwe oszustwa związane z ubezpieczeniem zdrowotnym, ubezpieczeniem samochodowym lub ubezpieczeniem majątkowym i wypadkowym.

* Zarządzanie ryzykiem

- Ryzyko kredytowe: Oceń ryzyko kredytowe potencjalnych wnioskodawców pożyczkowych na podstawie modelu predykcyjnego bazy danych, który szuka przyczyn i wzorców wpływających na ryzyko.

- Kontrola jakości: Znajdź wzorce problemów z jakością na liniach montażowych, aby zmniejszyć liczbę produktów zwracanych z powodu niskiej jakości.

* Usługi finansowe

- Utrzymanie klienta: Zidentyfikuj lojalnych klientów banku, którzy mają wiele kont z wysokim saldem i zapewnij każdemu z nich spersonalizowaną strukturę opłat. Dużo taniej jest utrzymać dotychczasowych klientów niż pozyskiwać nowych.

- Notowania akcji: Opracuj modele notowań w celu wsparcia zarządzania portfelem. Szukaj akcji, które zachowywały się w sposób podobny do niektórych wysokowydajnych papierów wartościowych.

* Usługi finansowe

- Utrzymanie klienta: Zidentyfikuj lojalnych klientów banku, którzy mają wiele kont z wysokim saldem i zapewnij każdemu z nich spersonalizowaną strukturę opłat. Dużo taniej jest utrzymać dotychczasowych klientów niż pozyskiwać nowych.

- Notowania akcji: Opracuj modele notowań w celu wsparcia zarządzania portfelem. Szukaj akcji, które zachowywały się w sposób podobny do niektórych wysokowydajnych papierów wartościowych.

* Dystrybucja

- Kontrola zapasów: Popraw kontrolę zapasów i dystrybucję, opracowując modele predykcyjne, które produkty lub części będą potrzebne w różnych punktach dystrybucji w różnych momentach.

Dobrym wskaźnikiem wartości eksploracji danych jest tajemnica otaczająca jej implementację. Wiele firm, które wdrożyły eksplorację danych, waha się, czy mówić o swoich sukcesach. Niektórzy nawet nie potwierdzają, że korzystają z tej technologii.

Działania związane z eksploracją danych

Czynności związane z eksploracją danych nie muszą być wykonywane liniowo. Rysunek 13.2 wskazuje, jakie czynności można wykonywać jednocześnie. Poniższa lista zawiera krótki opis czynności związanych z Krokiem 13, Eksploracja danych.

1 Przedstaw problem biznesowy. Ustal cele przed rozpoczęciem eksploracji danych i ustal cele priorytetowe (takie jak zwiększenie zysków, zmniejszenie kosztów, tworzenie innowacyjnych strategii produktów lub zwiększenie udziału w rynku). Aby osiągnąć którykolwiek z tych celów, trzeba zainwestować czas i pieniądze. Zarząd musi również zaangażować się w wdrożenie rozwiązania do eksploracji danych w organizacji.

2. Zbierz dane. Jedną z najbardziej czasochłonnych czynności eksploracji danych jest zbieranie odpowiednich typów i ilości danych. Aby mieć poprawną reprezentację, najpierw zidentyfikuj

wszystkie dane potrzebne do analizy. Obejmuje to dane przechowywane w operacyjnych bazach danych, dane z docelowych baz danych BI oraz wszelkie

dane zewnętrzne, które będą musiały zostać uwzględnione. Po zidentyfikowaniu danych źródłowych wyodrębnij wszystkie istotne elementy danych z tych różnych wewnętrznych i zewnętrznych źródeł danych.

3. Konsoliduj i oczyszczaj dane. Nadmiarowo przechowywane dane są bardziej normą niż wyjątkiem w większości organizacji. Dlatego dane z różnych źródeł muszą być skonsolidowane i oczyszczone. Jeżeli dane wewnętrzne mają być uzupełnione o pozyskane dane zewnętrzne, należy dopasować dane zewnętrzne do danych wewnętrznych i ustalić poprawną treść.

4. Przygotuj dane. Przed zbudowaniem analitycznego modelu danych należy przygotować dane. Częścią przygotowania danych jest klasyfikacja zmiennych. Zmienne mogą być dyskretne lub ciągłe, jakościowe lub ilościowe. Wyeliminuj zmienne z brakami danych lub zastąp je najbardziej prawdopodobnymi wartościami. Zapewnia doskonały wgląd w poznanie wartości maksymalnych, minimalnych, średnich, średnich, mediany i mody dla zmiennych ilościowych. Aby usprawnić proces przygotowania, rozważ zastosowanie przekształceń redukujących dane. Celem redukcji danych jest połączenie kilku zmiennych w jedną, aby zapewnić możliwość zarządzania zbiorem wyników do analizy. Na przykład połącz poziom wykształcenia, dochód, stan cywilny i kod pocztowy w jedną zmienną profilu.

5. Zbuduj analityczny model danych. Jedną z najważniejszych czynności data mining jest budowa analitycznego modelu danych. Analityczny model danych reprezentuje strukturę skonsolidowanych, zintegrowanych i zależnych od czasu danych, które zostały wybrane i wstępnie przetworzone z różnych wewnętrznych i zewnętrznych źródeł danych.

6. Po wdrożeniu model ten musi być w stanie kontynuować „uczenie się”, podczas gdy jest wielokrotnie używany przez narzędzie do eksploracji danych i dostrajany przez eksperta od eksploracji danych. Interpretuj wyniki eksploracji danych. Po uruchomieniu operacji eksploracji danych i uzyskaniu wyników kolejnym ważnym zadaniem jest ich interpretacja. Ważne rzeczy, które należy wziąć pod uwagę podczas tej interpretacji, to to, jak łatwo można zastosować wyniki i czy wyniki można przedstawić dyrektorom biznesowym w przekonujący, zorientowany na biznes sposób.

7. Przeprowadź zewnętrzną walidację wyników. Porównaj swoje wyniki z opublikowanymi statystykami branżowymi. Zidentyfikuj odchylenia od tych statystyk i określ przyczyny odchyłeń. Upewnij się, że korzystasz z aktualnych statystyk branżowych, ponieważ zmieniają się one od czasu do czasu. Porównaj kryteria wyboru swoich danych ze statystykami branżowymi i porównaj przedział czasu, w którym dane zostały wybrane, z przedziałem czasu objętym statystykami branżowymi. Kryteria wyboru i ramy czasowe modelu oraz statystyki branżowe muszą być podobne.

8. Monitoruj analityczny model danych w czasie. Statystyki branżowe są zwykle ustalane na podstawie bardzo dużych próbek. Ważne jest, aby w regularnych odstępach czasu weryfikować analityczny model danych ze statystykami branżowymi. Statystyki branżowe zmieniają się w czasie, a niektóre branże mają zmiany sezonowe. W takim przypadku dostosuj swój wewnętrzny model analityczny.

Rezultaty wynikające z tych działań

1. Baza eksploracji danych. Baza danych eksploracji danych została zaprojektowana i zbudowana dla określonego analitycznego modelu danych i określonego zestawu operacji eksploracji danych. Ta baza danych zostanie wypełniona danymi z systemu operacyjnego, docelowej bazy danych BI lub ich kombinacji.

2. Analityczny model danych. Analityczny model danych jest opracowywany i testowany, aby mógł być wykorzystywany przez algorytmy operacji eksploracji danych w narzędziu eksploracji danych.

Role zaangażowane w te działania

* Przedstawiciel firmy. Ostatecznie to przedstawiciel biznesowy (lub jego kierownictwo) skorzysta z wyników eksploracji danych. Dlatego przedstawiciel biznesowy musi bardzo ściśle współpracować z ekspertem w zakresie eksploracji danych (chyba że przedstawiciel biznesowy w tej aplikacji BI jest ekspertem w zakresie eksploracji danych), aby zrozumieć i zinterpretować wyniki eksploracji danych.

* Ekspert od eksploracji danych. Ekspert eksploracji danych to statystyk, który naprawdę zna dane i jest zaznajomiony z technikami eksploracji danych. Zwykle odpowiada za wybór najbardziej odpowiedniego narzędzia do eksploracji danych dla organizacji. Jest również główną osobą, która buduje analityczny model danych i analizuje wyniki eksploracji danych.

* Administrator bazy danych. Administrator bazy danych musi dobrze rozumieć zawartość danych, aby zaprojektować bazy danych eksploracji danych dla działań eksploracji danych. Współpracuje bardzo blisko z ekspertem od eksploracji danych.

* Ekspert merytoryczny. Ekspert w danej dziedzinie może pomóc w analizie, zdefiniowaniu, oczyszczeniu i przygotowaniu danych źródłowych do baz danych eksploracji danych. Będzie pracował pod kierunkiem eksperta ds. eksploracji danych.

Ryzyko niewykonania kroku 13

Większość organizacji siedzi na szczycie kopalni złota – „złoto” to wszystkie dane zebrane o ich klientach i produktach, które kupują. W tych danych zawarte są informacje o stylach wydawania pieniędzy, upodobaniach i niechęciach klientów oraz nawykach zakupowych. Te dane są zmarnowanym zasobem, jeśli nie są wydobywane w celu ujawnienia ukrytej analizy biznesowej. Ponadto kierownictwo musi zwracać uwagę na eksplorację danych przeprowadzaną przez konkurencję. Jeśli konkurenci zwiększą swoje zyski, zmniejszą koszty, stworzy bardziej innowacyjne strategie produktowe i powiększą swoje udziały w rynku, organizacja może bardzo szybko stracić klientów, co może zagrozić jej przetrwaniu.