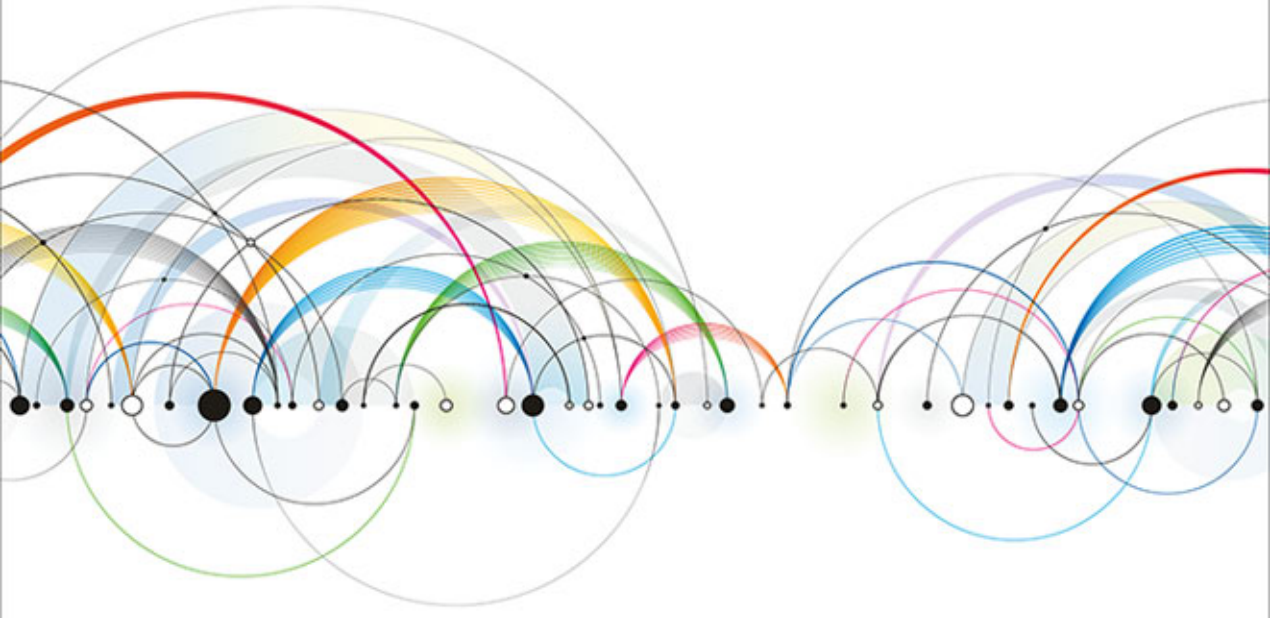
Foster Provost, Tom Fawcett

Analiza danych *w* biznesie

Sztuka podejmowania skutecznych decyzji



one press

Helion

Tytuł oryginału: Data Science for Business

Tłumaczenie: Leszek Sielicki

ISBN: 978-83-246-9610-9

Wydawnictwo HELION
ul. Kościuszki 1c, 44-100 GLIWICE
tel. 32 231 22 19, 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *http://helion.pl* (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
http://helion.pl/user/opinie/andabi
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Printed in Poland.

Spis treści

ROZDZIAŁ 2.

Problemy biznesowe a rozwiązania z zakresu nauki o danych

Podstawowe pojęcia: *Zbiór kanonicznych zadań związanych z eksploracją danych; Proces eksploracji danych; Nadzorowana i nienadzorowana eksploracja danych.*

Ważną zasadą nauki o danych jest to, że eksploracja danych jest *procesem* o stosunkowo dobrze zdefiniowanych etapach. Niektóre z nich wymagają stosowania technologii informatycznych, takich jak zautomatyzowane wykrywanie i ewaluacja wzorców z danych, podczas gdy inne wiążą się głównie z kreatywnością, wiedzą biznesową i zdrowym rozsądkiem analityka. Zrozumienie całego procesu pomaga w ujmowaniu projektów eksploracji danych w ramy strukturalne, a więc stają się one raczej usystematyzowanymi analizami niż heroicznymi przedsięwzięciami napędzanymi w dużej mierze przez przypadek i wnikliwość badaczy.

Ponieważ proces eksploracji danych rozbija ogólne zadanie wyszukania wzorców w danych na zestaw dokładnie zdefiniowanych podzadań, jest on także przydatny do strukturyzacji dyskusji o nauce o danych. W tej książce będziemy wykorzystywać ten proces jako ogólną platformę dla naszej dyskusji. W tym rozdziale przedstawimy proces eksploracji danych, ale najpierw wprowadzimy dodatkowy kontekst, omawiając typowe zadania z zakresu eksploracji danych. Przedstawienie ich pozwoli nam bardziej konkretnie zaprezentować cały proces i inne pojęcia w kolejnych rozdziałach.

Rozdział kończy omówienie szeregu innych istotnych zagadnień z zakresu analityki biznesowej, które nie są tematem tej książki (ale o których napisano wiele innych przydatnych książek), takich jak bazy danych, magazynowanie danych i podstawy statystyki.

Od problemów biznesowych do zadań eksploracji danych

Każdy problem decyzyjny w firmie, której funkcjonowanie opiera się na danych, jest wyjątkowy, zawiera własną kombinację celów, pragnień, ograniczeń, a nawet osobowości. Tak jak w przypadku inżynierii, istnieją jednak zbiory typowych zadań, które leżą u podstaw problemów biznesowych. We współpracy z decydentami w firmach badacze danych rozkładają problem biznesowy na podzadania. Rozwiązania podzadań mogą następnie zostać połączone

w celu rozwiązania problemu ogólnego. Niektóre z tych podzadań są wyjątkowe i dotyczą tylko jednego konkretnego problemu biznesowego, ale inne to typowe zadania eksploracji danych. Nasz problem z odpływem klientów jest na przykład wyjątkowy dla MegaTelCo: ma on specyficzne cechy, które odróżniają go od problemów związanych z odpływem klientów innych firm telekomunikacyjnych. Podzadaniem, które będzie jednak prawdopodobnie częścią rozwiązania każdego problemu odpływu abonentów, będzie oszacowanie na podstawie danych historycznych prawdopodobieństwa odejścia klienta rezygnującego z usług firmy wkrótce po wygaśnięciu umowy. Kiedy niepowtarzalne dane MegaTelCo zostały zestawione w określony format (co opiszemy w następnym rozdziale), oszacowanie prawdopodobieństwa zaczęło wyglądać jak jedno z bardzo typowych zadań związanych z eksploracją danych. Wiemy dużo o rozwiązywaniu typowych zadań dotyczących eksploracji danych, zarówno w kontekście naukowym, jak i praktycznym. W kolejnych rozdziałach będziemy również prezentować należące do sfery nauki o danych platformy, które pomogą nam w rozkładaniu problemów biznesowych i zestawianiu rozwiązań z podzadań.

W nauce o danych podstawowa jest umiejętność rozkładania problemu z zakresu analityki danych na części w taki sposób, że każda część odpowiada znanemu zadaniu, do wykonania którego dostępne są niezbędne narzędzia. Rozpoznawanie znanych problemów i ich rozwiązywanie sprawia, że unikamy marnowania czasu i zasobów na ponowne wynajdywanie koła. Pozwala nam również skoncentrować się na bardziej interesujących elementach procesu, które wymagają zaangażowania ze strony człowieka — na elementach, które nie zostały zautomatyzowane, a więc w ich przypadku w grę wchodzą kreatywność i inteligencja.

Pomimo wielkiej liczby konkretnych algorytmów eksploracji danych, które opracowano przez lata, istnieje tylko kilka fundamentalnie różniących się typów zadań, których te algorytmy dotyczą. Warto te zadania jasno zdefiniować. W kolejnych kilku rozdziałach będziemy wykorzystywać pierwsze dwa z nich (klasyfikację i regresję), aby zilustrować kilka podstawowych pojęć. W dalszej części książki określenie „jednostka" będzie odnosić się do podmiotu, dla którego dostępne są dane, takiego jak klient lub konsument, czy też podmiotu nieożywionego, takiego jak firma. Bardziej precyzyjnie opiszemy to pojęcie w rozdziale 3. W wielu projektach ze sfery analiz biznesowych zależy nam na znalezieniu „korelacji" między konkretną zmienną opisującą daną jednostkę a innymi zmiennymi. Możemy na przykład dysponować opartą na danych historycznych informacją, którzy klienci zrezygnowali po wygaśnięciu umów. Możemy zechcieć ustalić, jakie inne zmienne będą korelować z odpływem klientów w najbliższej przyszłości. Znajdowanie takich korelacji to najbardziej podstawowe przykłady zadań z zakresu klasyfikacji i regresji.

1. **Klasyfikacja** i **szacowanie prawdopodobieństwa** klas próbują prognozować, dla każdego osobnika w populacji, do którego z (małego) zbioru klas ten osobnik należy. Zazwyczaj klasy wykluczają się wzajemnie. Przykładowe pytanie odnoszące się do klasyfikacji mogłoby brzmieć: „Którzy spośród wszystkich klientów MegaTelCo prawdopodobnie odpowiedzą na złożoną ofertę?". W tym przykładzie dwie klasy można byłoby nazwać `zareagują` i `nie zareagują`.

   W zadaniu klasyfikacji procedura eksploracji danych tworzy model, który dla danego nowego osobnika określa, do której klasy ten osobnik należy. Ściśle powiązanym zadaniem jest ***scoring*** lub **szacowanie prawdopodobieństwa** klasy. Model scoringowy zastosowany dla danego osobnika podaje zamiast klasy wynik określający prawdopodobieństwo przynależności danego osobnika do każdej z klas. W naszym scenariuszu reakcji klienta mo-

del scoringowy będzie w stanie dokonać oceny każdego pojedynczego klienta i określić, z jakim prawdopodobieństwem zareaguje on na ofertę. Klasyfikacja i *scoring* są bardzo ściśle powiązane; jak się przekonamy, model klasyfikacyjny można zwykle zmodyfikować, aby przeprowadził *scoring*, i odwrotnie.

2. **Regresja** („szacowanie wartości") próbuje dla każdego osobnika oszacować czy też przewidzieć wartość liczbową jakiejś zmiennej dotyczącej tego osobnika. Przykładowe pytanie odnoszące się do regresji mogłoby brzmieć: „W jakim stopniu dany klient będzie korzystał z usługi?". Właściwość (zmienna), która tutaj ma zostać przewidziana, to *wykorzystanie usługi*, a model mógłby zostać wygenerowany na podstawie obserwacji innych podobnych osobników w ramach populacji i historycznego wykorzystywania przez nich usługi. Procedura regresji tworzy model, który, biorąc pod uwagę osobnika, szacuje wartość danej zmiennej, specyficznej dla tego osobnika.

   Regresja jest powiązana z klasyfikacją, ale się od niej różni. Ujmując to nieformalnie, klasyfikacja przewiduje, *czy* coś się stanie, natomiast regresja przewiduje, *ile* tego czegoś się stanie. To rozróżnienie stanie się bardziej przejrzyste w dalszej treści książki.

3. **Dopasowywanie podobieństw** (ang. *similarity matching*) próbuje *identyfikować* podobne jednostki na podstawie danych o nich. Dopasowywanie podobieństw może być stosowane bezpośrednio, w celu znajdowania podobnych osobników. Na przykład firma IBM jest zainteresowana znalezieniem firm podobnych do swoich najlepszych klientów biznesowych, aby skoncentrować wysiłki swoich handlowców na najlepszych potencjalnych okazjach biznesowych. Wykorzystuje dopasowanie podobieństw, którego podstawą są dane „firmograficzne", opisujące charakterystyczne cechy różnych firm. Zestawianie podobieństw leży u podstaw jednej z najbardziej popularnych metod rekomendowania produktów (znajdowanie osób, które są podobne do nas z punktu widzenia produktów, które im się podobały lub które zostały przez nie zakupione). Miary podobieństwa leżą u podstaw szeregu rozwiązań innych zadań z zakresu eksploracji danych, takich jak klasyfikacja, regresja i klastrowanie. Podobieństwo i jego zastosowania omawiamy dokładnie w rozdziale 6.

4. **Klastrowanie** próbuje *grupować* jednostki w populacji na podstawie podobieństw, ale nie jest to podyktowane konkretnym celem. Przykładowe pytanie związane z klastrowaniem mogłoby brzmieć: „Czy klienci tworzą naturalne grupy lub segmenty?". Klastrowanie jest przydatne we wstępnej eksploracji domeny w celu sprawdzenia, jakie naturalne grupy w niej istnieją, ponieważ grupy te z kolei mogą zasugerować inne zadania z zakresu eksploracji danych lub inne podejścia. Klastrowanie służy także jako wstęp do procesów decyzyjnych, koncentrujących się na takich kwestiach jak: *Jakie produkty powinniśmy zaoferować lub rozwinąć? Jaką strukturę powinny mieć nasze zespoły obsługi klienta (czy też zespoły sprzedażowe)?* Klastrowanie omawiamy szczegółowo w rozdziale 6.

5. **Grupowanie współwystąpień** (ang. *co-occurence grouping*, znane również jako odkrywanie zbiorów częstych, odkrywanie zależności i analiza koszykowa rynku) próbuje znajdować *powiązania* pomiędzy jednostkami na podstawie transakcji z ich udziałem. Przykładowe pytanie z zakresu grupowania współwystąpień mogłoby brzmieć: „Jakie przedmioty są powszechnie kupowane razem?". O ile klastrowanie zajmuje się podobieństwami pomiędzy obiektami na podstawie atrybutów tych obiektów, to grupowanie współwystąpień uwzględnia podobieństwo obiektów na podstawie ich łącznego pojawiania się w transakcji. Na przykład, analizując ewidencję zakupów w supermarkecie, możemy zauważyć, że mielone mięso jest kupowane razem z pikantnym sosem znacznie częściej, niż można by się spodziewać. Zdecydowanie, jakie działania należy podjąć w związku z tym odkryciem,

może wymagać nieco kreatywności, ale być może wskazane byłoby zaproponowanie specjalnej promocji, nowego sposobu prezentacji produktów lub oferty kombinowanej. Współwystępowanie produktów w ramach zakupów to popularny rodzaj grupowania, znany jako analiza koszykowa rynku. Niektóre systemy *rekomendacyjne* również przeprowadzają pewnego rodzaju grupowanie spowinowacone, wyszukując na przykład pary książek, które są często kupowane przez te same osoby („osoby, które kupiły X, kupiły też Y").

Wynikiem grupowania współwystąpień jest opis elementów, które występują razem. Opisy te zwykle zawierają dane statystyczne dotyczące częstości współwystępowania i oszacowanie, na ile jest to zaskakujące.

6. **Profilowanie** (znane także jako opis zachowania) próbuje charakteryzować typowe zachowania jednostki, grupy lub populacji. Przykładowe pytanie z zakresu profilowania mogłoby brzmieć: „Jaki jest typowy poziom wykorzystania telefonów komórkowych w tym segmencie klientów?". Opis zachowania nie zawsze bywa łatwy; profilowanie wykorzystania telefonów komórkowych może wymagać skomplikowanego opisu przeciętnej aktywności w godzinach nocnych i w weekendy, wykorzystania telefonu w rozmowach międzynarodowych, opłat za roaming, korzystania z SMS-ów i tak dalej. Zachowanie można opisać ogólnie dla całej populacji lub z coraz większą szczegółowością, do poziomu małych grup lub nawet poszczególnych osób.

   Profilowanie jest często wykorzystywane do tworzenia norm zachowania dla aplikacji wykrywających anomalie, służących na przykład do wykrywania oszustw lub monitorowania włamań do systemów komputerowych (gdy ktoś na przykład włamuje się na nasze konto w iTunes). Jeżeli wiemy, jakie zakupy dana osoba zazwyczaj robi, płacąc kartą kredytową, to możemy określić, czy nowe obciążenie karty pasuje do tego profilu czy nie. Możemy wykorzystać stopień niedopasowania jako wskaźnik określający, na ile podejrzana jest ta sytuacja, i wszcząć alarm, jeśli będzie on zbyt wysoki.

7. **Predykcja połączeń** próbuje przewidzieć połączenia pomiędzy elementami danych, zazwyczaj poprzez zasugerowanie, że połączenie powinno istnieć, a czasem także szacowanie siły połączenia. Predykcja połączeń jest powszechna w systemach społecznościowych: „Skoro ty i Karen macie dziesięciu wspólnych znajomych, to może chcesz być znajomym Karen?". Predykcja połączeń może też szacować siłę połączenia. Na przykład, aby zarekomendować klientom filmy, moglibyśmy pomyśleć o wykresie łączącym klientów i filmy, które obejrzeli lub ocenili. Na wykresie szukamy połączeń, które pomiędzy klientami i filmami *nie* istnieją, ale przewidujemy, że powinny istnieć i być silne. Te połączenia stanowią podstawę dla rekomendacji.

8. **Redukcja danych** próbuje duże zbiory danych zastępować mniejszymi, które zawierają większość istotnych informacji zbiorów większych. Mniejszy zbiór danych może być łatwiejszy do obróbki lub przetwarzania. Co więcej, mniejszy zbiór danych może umożliwiać lepszy wgląd w informacje. Ogromny zbiór danych dotyczących preferencji klientów w kwestii oglądania filmów można na przykład zredukować do znacznie mniejszego zbioru danych, ujawniających preferencje gatunkowe konsumentów ukryte w danych związanych z oglądalnością (np. preferencje widza dotyczące gatunków filmowych). Redukcja danych prawie zawsze związana jest z utratą informacji. Taki kompromis bywa jednak korzystny, bo umożliwia lepsze zrozumienie istoty problemu.

9. **Modelowanie przyczynowe** próbuje zrozumieć, jakie zdarzenia lub działania faktycznie *wpływają* na inne. Załóżmy na przykład, że używamy modelowania predykcyjnego w celu kierowania reklam do klientów i zauważamy, że poziom zakupów klientów stargetowa-

nych staje się wyższy po skierowaniu do nich reklam. Czy stało się tak, bo reklamy wpłynęły na klientów, skłaniając ich do zakupu? A może modele predykcyjne po prostu dobrze się spisały, identyfikując klientów, którzy i tak dokonaliby zakupu? Wśród technik modelowania przyczynowego istnieją takie, które wymagają poważnych inwestycji w dane, w rodzaju randomizowanych kontrolowanych eksperymentów (np. tak zwanych testów A/B), oraz zaawansowanych metod wyciągania wniosków przyczynowych z zaobserwowanych danych. Zarówno eksperymentalne, jak i obserwacyjne metody modelowania przyczynowego ogólnie mogą być postrzegane jako analiza „kontrfaktyczna": starają się one zrozumieć, jaka byłaby różnica pomiędzy sytuacjami — z których miejsce może mieć tylko jedna — gdyby „badane" zdarzenie (np. prezentacja reklamy określonej jednostce) zaszło i nie zaszło.

W każdym takim przypadku ostrożny badacz danych, wyciągając wniosek przyczynowy, powinien zawsze podać dokładne założenia, które są niezbędne, aby wniosek przyczynowy był prawdziwy (takie założenia istnieją *zawsze* — zawsze o nie pytaj). Podejmując się modelowania przyczynowego, firma musi rozważyć kompromis pomiędzy zwiększeniem inwestycji, aby zredukować przyjęte założenia, i zadecydowaniem, że wnioski są wystarczająco trafne, biorąc pod uwagę założenia. Nawet w najbardziej starannie zrandomizowanym, kontrolowanym procesie eksperymentalnym dokonuje się założeń, które mogą spowodować, że wnioski odnoszące się do przyczynowości będą niewłaściwe. Odkrycie w medycynie efektu placebo ilustruje znaną powszechnie sytuację, w której w dokładnie zaprojektowanym, zrandomizowanym eksperymencie przeoczono założenie.

Szczegółowe omówienie wszystkich tych zadań wymagałoby wielu książek. W tej przedstawiamy zbiór najbardziej podstawowych zasad nauki o danych, które łącznie stanowią fundament dla wszystkich rodzajów tych zadań. Zasady te będziemy ilustrować, posługując się głównie klasyfikacją, regresją, dopasowywaniem podobieństw i klastrowaniem, a inne omówimy, gdy będą istotną ilustracją podstawowych zasad (w końcowej części książki).

Zastanówmy się, które z tych typów zadań mogłyby pasować do naszego problemu z prognozowaniem odpływu abonentów. Praktycy często traktują prognozowanie odpływu abonentów jako problem związany ze znajdowaniem *segmentów* klientów, których odejście jest mniej lub bardziej prawdopodobne. Ten problem związany z segmentacją wygląda na problem klasyfikacji, lub ewentualnie klastrowania, a nawet regresji. Aby wybrać najlepszą formułę, musimy najpierw wprowadzić kilka istotnych rozróżnień.

Metody nadzorowane i nienadzorowane

Zastanówmy się nad dwoma podobnymi pytaniami, które moglibyśmy zadać populacji naszych klientów. Pierwsze z nich brzmi: „Czy nasi klienci w naturalny sposób należą do różnych grup?". Tutaj grupowanie nie ma określonego celu czy też *wielkości docelowej*. Jeżeli nie ma takiej wielkości docelowej, to problem eksploracji danych określa się jako **nienadzorowany**. Porównajmy to z nieco innym pytaniem: „Czy możemy znaleźć grupy klientów, w przypadku których istnieje szczególnie duże prawdopodobieństwo rezygnacji z usług naszej firmy po wygaśnięciu umowy?". W tym miejscu określona została konkretna wielkość docelowa: czy klient zrezygnuje po wygaśnięciu umowy? W tym przypadku segmentacja jest przeprowadzana z konkretnego powodu: aby podjąć działanie oparte na prawdopodobieństwie rezygnacji. Taki problem eksploracji danych określa się jako **nadzorowany**.

Uwaga o terminologii: uczenie nadzorowane i nienadzorowane

Terminy *nadzorowane* i *nienadzorowane* pochodzą z dziedziny uczenia maszynowego. Ujmując rzecz metaforycznie, nauczyciel „nadzoruje" ucznia, starannie dostarczając informacji o wielkości docelowej, wraz z zestawem przykładów. Zadanie związane z nienadzorowanym uczeniem się może zawierać ten sam zestaw przykładów, ale nie zawiera informacji o wielkości docelowej. Uczeń nie zostaje poinformowany o celach uczenia się i ma sformułować własne wnioski dotyczące tego, co przykłady mają wspólnego.

Różnica między tymi kwestiami jest subtelna, ale istotna. Jeśli istnieje konkretna wielkość docelowa, to problem można określić jako nadzorowany. Nadzorowane zadania wymagają innych technik niż nienadzorowane, a wyniki często bywają o wiele bardziej przydatne. W technice nadzorowanej grupowanie ma określony cel — predykcję wielkości docelowej. Klastrowanie, zadanie nienadzorowane, tworzy grupy oparte na podobieństwach, ale nie ma gwarancji, że te podobieństwa są znaczące i będą przydatne do jakiegoś konkretnego celu.

Z technicznego punktu widzenia nadzorowana eksploracja danych wymaga spełnienia jeszcze jednego warunku: muszą istnieć *dane* dotyczące wielkości docelowej. Nie wystarczy, że istnieją o niej informacje jako takie, muszą one również występować w danych. Dobrze byłoby na przykład wiedzieć, czy dany klient będzie korzystał z usług firmy przez co najmniej sześć miesięcy, ale jeżeli w danych historycznych takie informacje nie istnieją lub są niekompletne (na przykład dlatego, że dane są przechowywane tylko przez dwa miesiące), to wartości wielkości docelowej nie da się określić. Pozyskiwanie danych o wielkości docelowej często bywa kluczową inwestycją w ramach nauki o danych. Wartość zmiennej docelowej jednostki jest często nazywana *etykietą* tej jednostki, podkreślając fakt, że często (choć nie zawsze) opatrywanie danych etykietami wiąże się z pewnym wydatkiem.

Zadania klasyfikacji, regresji i modelowania przyczynowego są zazwyczaj rozwiązywane przy pomocy metod nadzorowanych. Dopasowywanie podobieństw, predykcja połączeń i redukcja danych mogą należeć do obu grup. Klastrowanie, grupowanie współwystąpień i profilowanie zazwyczaj są nienadzorowane. Podstawowe zasady eksploracji danych, które będziemy przedstawiać, leżą u podstaw wszystkich rodzajów tych technik.

Dwie główne podklasy *nadzorowanej* eksploracji danych, klasyfikacja i regresja, wyróżnia rodzaj wielkości docelowej. Regresja wiąże się z liczbową wielkością docelową, podczas gdy klasyfikacja odnosi się do wielkości docelowej kategorycznej (często binarnej). Zastanówmy się nad poniższymi, podobnymi do siebie pytaniami, które moglibyśmy zadać w ramach nadzorowanej eksploracji danych:

Czy ten klient nabędzie usługę S1, jeśli otrzyma zachętę I?
: To problem klasyfikacji, ponieważ ma binarną wielkość docelową (klient albo kupi, albo nie).

Który pakiet usług (S1, S2 lub żaden) prawdopodobnie nabędzie klient, jeśli otrzyma zachętę I?
: To także problem klasyfikacji, o trójwartościowej wielkości docelowej.

W jakim stopniu ten klient będzie korzystał z usługi?
: To problem regresji, bo ma liczbową wielkość docelową. Zmienną docelową jest poziom wykorzystania usługi (rzeczywisty lub prognozowany) dla danego klienta.

Pytania te zawierają pewne subtelności, o których warto wspomnieć. W zastosowaniach biznesowych często pożądana jest liczbowa *predykcja* zamiast kategorycznej wielkości docelowej. W przykładzie z odpływem abonentów podstawowa predykcja typu tak/nie dotycząca tego,

czy klient nadal będzie korzystał z usługi, może nie być wystarczająca; chcemy zamodelować *prawdopodobieństwo*, że klient nadal będzie korzystał z usługi. I tak uznajemy to za modelowanie raczej klasyfikacyjne niż regresyjne, ponieważ jego wielkość docelowa jest kategoryczna. Tam, gdzie jest to konieczne w celu zapewnienia przejrzystości, będziemy to nazywać „szacowaniem prawdopodobieństwa klas".

W początkowych etapach procesu eksploracji danych istotną rolę pełni podjęcie decyzji, czy będziemy podchodzić do problemu w sposób nadzorowany czy nienadzorowany, a jeśli miałby to być sposób nadzorowany, to konieczne jest stworzenie precyzyjnej definicji zmiennej docelowej. Ta zmienna musi być konkretną wielkością, na której koncentrować się będzie eksploracja danych (i dla której możemy uzyskać wartości jakichś przykładów z danych). Wrócimy do tego w rozdziale 3.

Eksploracja danych i jej wyniki

Istnieje jeszcze inne istotne rozróżnienie odnoszące się do eksploracji danych. Chodzi o różnicę pomiędzy: (1) eksploracją danych w celu znalezienia wzorców i zbudowania modeli, oraz (2) *wykorzystywaniem* wyników eksploracji danych. Studenci, poznając naukę o danych, często mylą te dwa procesy, a menedżerowie czasami je robią to samo przy omawianiu analiz biznesowych. Wykorzystywanie wyników eksploracji danych powinno wpływać na sam proces eksploracji danych i go przenikać, ale te dwie kwestie należy rozróżnić.

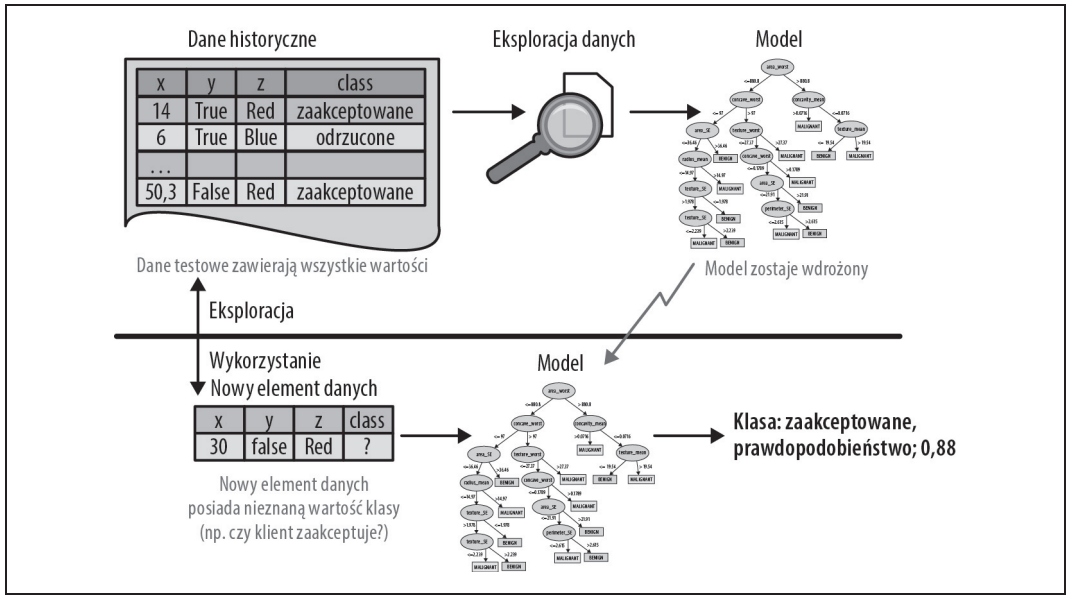
W naszym przykładzie odpływu abonentów zastanówmy się nad scenariuszem wdrożenia, w którym wykorzystane zostaną wyniki. Chcemy wykorzystać model, aby przewidzieć, który z naszych klientów odejdzie. Załóżmy zwłaszcza, że eksploracja danych wytworzyła model oszacowania prawdopodobieństwa klasy *M*. Każdy istniejący klient został opisany z wykorzystaniem zbioru cech charakterystycznych; *M* traktuje te cechy jako dane wejściowe i podaje wskaźnik czy też oszacowanie prawdopodobieństwa odejścia klienta. To jest *wykorzystanie* wyników eksploracji danych. Eksploracja danych tworzy model *M* z innych danych, często historycznych.

Rysunek 2.1 przedstawia te dwie fazy. Eksploracja danych tworzy model szacowania prawdopodobieństwa, co widać w górnej części rysunku. W fazie wykorzystania (dolna połowa) model zostaje zastosowany do nowego, nieznanego przypadku i generuje dla niego oszacowanie prawdopodobieństwa.

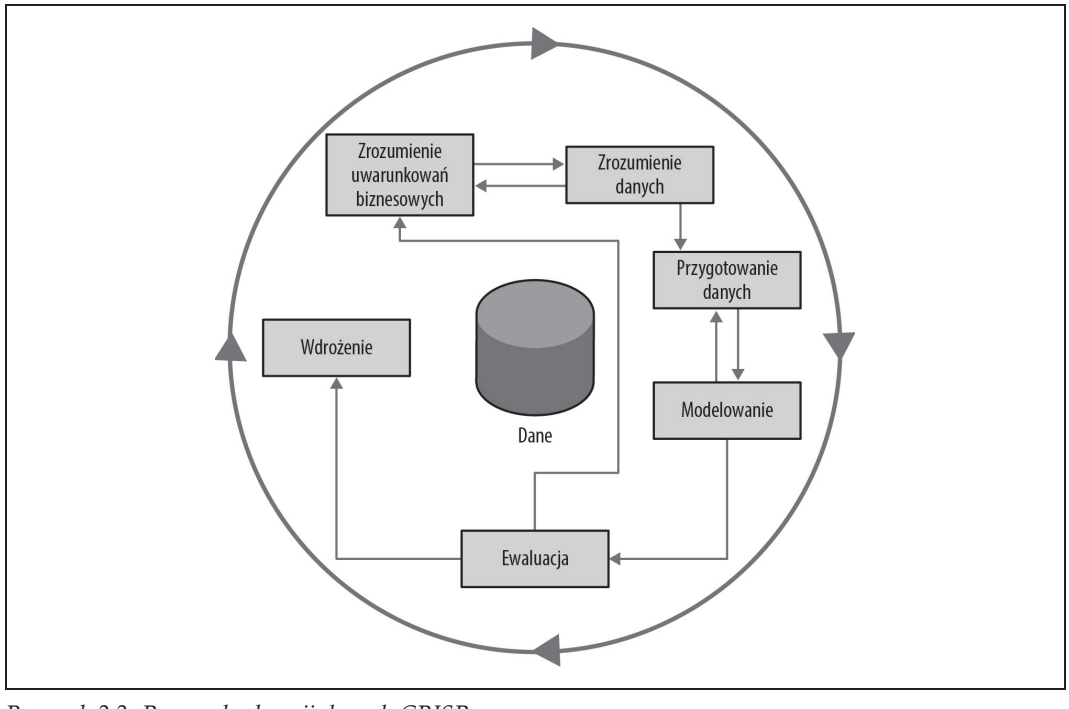
Proces eksploracji danych

Eksploracja danych jest rzemiosłem. Wymaga wykorzystywania w znaczącym stopniu nauki i technologii, ale prawidłowe posługiwanie się nią zawiera pierwiastek sztuki. Tak jak w przypadku wielu dojrzałych rzemiosł, istnieje tutaj jednak zrozumiały proces, który nadaje problemowi określoną strukturę, umożliwiając osiągnięcie odpowiedniej spójności, powtarzalności i obiektywizmu. Przydatną kodyfikację procesu eksploracji danych zawiera schemat *Cross Industry Standard Process for Data Mining* (CRISP-DM Project, 2000; Shearer, 2000), przedstawiony na rysunku 2.2[1].

[1] Patrz także strona Wikipedii o procesie CRISP-DM (*http://pl.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining*).



Rysunek 2.1. Eksploracja danych a wykorzystanie wyników eksploracji danych. Górna połowa rysunku przedstawia eksplorację danych historycznych, aby wytworzyć model. Co ważne, w danych historycznych wartość wielkości docelowej („klasy") jest określona. Dolna połowa pokazuje wynik eksploracji w działaniu; model zostaje zastosowany do nowych danych, dla których nie znamy wartości klasy. Model przewiduje zarówno wartość klasy, jak i prawdopodobieństwo tego, że zmienna klasy przyjmie tę wartość



Rysunek 2.2. Proces eksploracji danych CRISP

Schemat ten wyraźnie pokazuje, że iteracja jest tutaj raczej regułą niż wyjątkiem. Przeprowadzenie tego procesu jeden raz i nieznalezienie rozwiązania problemu nie jest, ogólnie rzecz biorąc, niepowodzeniem. Przeprowadzenie tego procesu często bywa źródłem danych i po pierwszej iteracji zespół badaczy danych wie znacznie więcej. Następna iteracja może więc być znacznie bardziej trafna. Przedyskutujmy teraz szczegółowo kolejne jego etapy.

Zrozumienie uwarunkowań biznesowych

Przede wszystkim podstawowe znaczenie ma zrozumienie problemu, który ma zostać rozwiązany. Może się to wydawać oczywiste; projekty biznesowe rzadko funkcjonują jako jasne i jednoznaczne problemy z zakresu eksploracji danych. Często przekształcenie problemu i opracowanie rozwiązania jest iteracyjnym procesem odkrywania. Schemat pokazany na rysunku 2.2 obrazuje to raczej jako cykle wewnątrz cyklu, a nie prosty proces linearny. Wstępne sformułowanie niekoniecznie bywa kompletne czy też optymalne, więc wielokrotne powtórzenia mogą być niezbędne do sformułowania możliwego do przyjęcia rozwiązania.

Etap zrozumienia uwarunkowań biznesowych stanowi ten element rzemiosła, w którym bardzo dużą rolę odgrywa kreatywność analityka. Nauka o danych ma tutaj, jak się przekonamy, także coś do powiedzenia, ale często kluczem do wielkiego sukcesu jest twórcze sformułowanie problemu przez analityka, określające sposób przedstawienia problemu biznesowego jako jednego lub kilku problemów z zakresu nauki o danych. Wysoki poziom wiedzy o podstawach pomaga kreatywnym analitykom biznesowym dostrzegać nowatorskie formuły.

Dysponujemy zestawem skutecznych narzędzi do rozwiązywania poszczególnych problemów eksploracji danych: podstawowe zadania eksploracji danych omawialiśmy w podrozdziale „Od problemów biznesowych do zadań eksploracji danych". Zwykle wczesne etapy przedsięwzięcia obejmują opracowywanie rozwiązania, które wykorzystuje te narzędzia. Może to oznaczać ujmowanie (opracowywanie) problemu w taki sposób, że jeden lub kilka problemów częściowych wiąże się z budowaniem modeli do celów klasyfikacji, regresji, szacowania prawdopodobieństwa itd.

W ramach tego pierwszego etapu *zespół projektowy powinien dokładnie przemyśleć scenariusz przypadków użycia*. To jedno z najważniejszych założeń nauki o danych, któremu poświęcamy całe dwa rozdziały (rozdział 7. i rozdział 11.). Co dokładnie chcemy zrobić? Jak chcemy to zrobić? Jakie elementy tego scenariusza użycia kreują możliwe modele eksploracji danych? Omawiając te kwestie bardziej szczegółowo, zaczniemy od uproszczonego scenariusza użycia, ale w dalszym toku naszych rozważań wrócimy do podstaw i zrozumiemy, że scenariusz często należy dostosowywać, aby lepiej odzwierciedlał rzeczywiste potrzeby biznesowe. Przedstawimy narzędzia koncepcyjne wspomagające rozumowanie w tym zakresie, w tym umieszczenie problemu biznesowego w kontekście wartości oczekiwanej, które może pozwolić nam na systematyczne rozłożenie go na zadania eksploracji danych.

Zrozumienie danych

Jeśli rozwiązanie problemu biznesowego jest celem, to dane są dostępnym surowcem, z którego zbudowane zostanie rozwiązanie. Ważne jest zrozumienie zalet i ograniczeń związanych z danymi, bo rzadko dokładnie pokrywają się one z problemem. Dane historyczne są często gromadzone w celach niezwiązanych z bieżącym problemem biznesowym lub w ogóle bez

wyraźnego celu. Baza danych klientów, baza danych transakcji czy baza danych marketingowych może obejmować różne przenikające się wzajemnie populacje. Poza tym, bazy te mogą być w różnym stopniu wiarygodne.

Typowe jest także zróżnicowanie *kosztów* danych. Niektóre dane są dostępne praktycznie za darmo, podczas gdy zdobycie innych może wymagać wysiłku. Pewne dane można kupić. Jeszcze inne po prostu nie istnieją i konieczne są całe projekty pomocnicze, aby zorganizować ich zebranie. Podstawowa część fazy zrozumienia danych to oszacowanie kosztów oraz korzyści wiążących się z każdym źródłem danych i podjęcie decyzji, czy dalsze inwestowanie jest uzasadnione. Nawet po pozyskaniu wszystkich zbiorów danych ich zestawienie może wymagać dodatkowego wysiłku. Na przykład dane ewidencyjne klientów i identyfikatory produktów powszechnie bywają niejednoznaczne i zaszumione. Oczyszczanie i dopasowywanie danych klientów, aby mieć pewność, że każdemu klientowi odpowiada tylko rekord, jest samo w sobie skomplikowanym problemem analitycznym (Hernández i Stolfo, 1995; Elmagarmid, Ipeirotis i Verykios, 2007).

W reakcji na postęp procesu zrozumienia danych kierunek mogą zmieniać także drogi wiodące do rozwiązania problemu, a działania zespołu mogą nawet zacząć podążać różnymi torami. Ilustrację takiej sytuacji można znaleźć w sferze wykrywania oszustw. Eksploracja danych jest szeroko wykorzystywana do wykrywania oszustw i wiele problemów wykrywania oszustw zawiera klasyczne zadania nadzorowanej eksploracji danych. Zastanówmy się nad zadaniem wykrycia oszustwa z wykorzystaniem karty kredytowej. Obciążenia pojawiają się na rachunku każdego klienta, więc obciążenia będące wynikiem oszustwa są zwykle wykrywane — jeśli nie od razu przez firmę, to w późniejszym terminie przez klienta po sprawdzeniu historii rachunku. Możemy założyć, że prawie wszystkie oszustwa są identyfikowane i opatrywane wiarygodną etykietą, ponieważ uprawniony klient i osoba popełniająca oszustwo to różne osoby, mające przeciwstawne cele. Transakcje przy użyciu kart kredytowych mają więc wiarygodne etykiety (*oszustwo* i *uprawnione użycie*), które mogą służyć jako wielkości docelowe dla techniki nadzorowanej.

Rozważmy teraz problem związany z wykrywaniem oszustw w sferze ubezpieczeń zdrowotnych. W Stanach Zjednoczonych to ogromny problem, którego koszt wynosi miliardy dolarów rocznie. Choć może się wydawać, że mamy tu do czynienia z tradycyjnym problemem wykrywania oszustw, to kiedy uwzględnimy relację problemu biznesowego do danych, zdamy sobie sprawę, że problem jest zupełnie inny. Sprawcy oszustw — dostawcy usług medycznych, którzy składają fałszywe zgłoszenia, a czasem ich pacjenci — są uprawnionymi usługodawcami i użytkownikami systemu rozliczeniowego. Sprawcy oszustw są podzbiorem zbioru uprawnionych użytkowników; nie ma odrębnej, niezainteresowanej strony, która mogłaby określić, jakie dokładnie powinny być „właściwe" opłaty. W związku z tym dane rozliczeniowe systemu ubezpieczeń zdrowotnych nie posiadają wiarygodnej zmiennej docelowej wskazującej oszustwo i nie może tutaj zostać zastosowane podejście nadzorowane, które mogłoby być skuteczne w przypadku oszustw związanych z kartami kredytowymi. Taki problem wymaga zwykle podejścia nienadzorowanego, na przykład profilowania, klastrowania czy grupowania współwystąpień.

To, że oba powyższe problemy dotyczą wykrywania oszustw, jest tylko powierzchownym podobieństwem, które tak naprawdę jest mylące. W procesie zrozumienia danych musimy drążyć głęboko, aby odkryć strukturę problemu biznesowego i dane, które są dostępne, a następnie dopasować je do jednego lub większej liczby zadań eksploracji danych, dla których

dysponujemy znaczącym zasobem nauki i technologii. W przypadku problemu biznesowego nie jest niczym niezwykłym, że zawiera on szereg zadań eksploracji danych, często należących do różnych typów, i konieczne będzie połączenie ich rozwiązań (patrz rozdział 11.).

Przygotowanie danych

Technologie analityczne, których możemy użyć, są bardzo skuteczne, ale nakładają pewne wymogi na dane, które wykorzystują. Często wymagają, żeby dane miały inną postać niż ich naturalna forma, niezbędne będzie więc pewne ich przekształcenie. Dlatego faza przygotowania danych przebiega jednocześnie z fazą ich zrozumienia; danymi się manipuluje i przekształca je do postaci, w której przyniosą lepsze wyniki.

Typowe przykłady przygotowania danych to ich konwersja do postaci tabeli, usuwanie lub dedukcja brakujących wartości i konwertowanie danych na inne ich typy. Niektóre techniki eksploracji danych są przeznaczone dla danych symbolicznych i kategorycznych, inne operują wyłącznie na wartościach liczbowych. Dodatkowo wartości liczbowe często muszą być normalizowane czy też skalowane, aby były porównywalne. Istnieją standardowe techniki i ogólne zasady umożliwiające przeprowadzanie takich konwersji. W rozdziale 3. omawiamy bardziej szczegółowo formaty danych najbardziej typowe dla procesu eksploracji.

Ogólnie jednak w tej książce nie będziemy koncentrować się na technikach przygotowania danych, które same w sobie mogą być tematem osobnej książki (Pyle, 1999). W kolejnych rozdziałach zdefiniujemy podstawowe formaty danych, a szczegółami związanymi z przygotowywaniem danych będziemy zajmować się tylko wtedy, jeśli będą one miały związek z podstawowymi zasadami nauki o danych lub będą niezbędne, aby przedstawić konkretny przykład.

Mówiąc bardziej ogólnie, we wstępnej fazie procesu badacze danych poświęcają zwykle sporo czasu na zdefiniowanie zmiennych, które będą wykorzystywane w dalszym jego toku. To jeden z głównych punktów, w których istotne znaczenie mają kreatywność, zdrowy rozsądek i wiedza fachowa. Wartość rozwiązania z zakresu eksploracji danych często opiera się na tym, na ile dobrze analitycy usystematyzują problemy i określą zmienne (a czasem bywa im zaskakująco trudno się do tego przyznać).

Bardzo istotnym ogólnym problemem związanym z przygotowywaniem danych jest unikanie „wycieków". (Kaufman i in., 2012). Wyciek to sytuacja, w której zmienna zawarta w danych historycznych niesie informację o zmiennej docelowej — informację, która pojawia się w danych historycznych, ale nie jest faktycznie dostępna, gdy należy podjąć decyzję. Przykładem może być sytuacja prognozowania, czy w określonym momencie odwiedzająca stronę internetową osoba zakończy sesję czy będzie kontynuować przeglądanie, przenosząc się na inną stronę. Zmienna „całkowita liczba stron odwiedzonych w trakcie sesji" jest tutaj możliwa do ustalenia. Całkowita liczba stron odwiedzonych w trakcie sesji będzie jednak znana dopiero po jej zakończeniu (Kohavi i in., 2000), gdy wartość zmiennej docelowej będzie znana! Innym przykładem może być prognozowanie, czy klient *będzie miał* „szeroki gest"; informacje o kategoriach zakupionych przedmiotów (lub, co gorsza, kwota zapłaconego podatku) są wyjątkowo predyktywne, ale nie są znane w momencie podejmowania decyzji (Kohavi i Parekh, 2003). Nad wyciekiem należy się uważnie zastanowić w trakcie przygotowywania danych, ponieważ ma ono zwykle miejsce po fakcie — z danych historycznych. Bardziej szczegółowy przykład rzeczywistego wycieku, który trudno było znaleźć, prezentujemy w rozdziale 14.

Modelowanie

Modelowanie jest tematem kilku następnych rozdziałów i w tym miejscu nie będziemy zajmować się nim szczegółowo. Powiemy tylko, że wynikiem modelowania jest pewnego rodzaju model lub wzorzec wychwytujący prawidłowości w danych.

Etap modelowania jest zasadniczym miejscem, w którym do danych stosowane są techniki eksploracji danych. Istotne jest, aby w pewnym stopniu rozumieć podstawowe koncepcje eksploracji danych, w tym rodzaje istniejących technik i algorytmów, ponieważ jest to ta część rzemiosła, w której w największym stopniu wykorzystywane są nauka i technologia.

Ewaluacja

Celem etapu ewaluacji jest rygorystyczna ocena wyników eksploracji danych przed przejściem dalej w celu zyskania pewności, że są one prawidłowe i wiarygodne. Jeśli przyjrzymy się uważnie jakiemukolwiek zbiorowi danych, to znajdziemy wzorce, ale mogą one nie przetrwać starannej i krytycznej analizy. Chcielibyśmy mieć pewność, że modele i wzorce wydobyte z danych są rzeczywistymi prawidłowościami, a nie tylko osobliwościami lub wadami próbek. Możliwe jest wdrożenie wyników bezpośrednio po przeprowadzeniu eksploracji danych, ale nie jest to wskazane; zwykle o wiele łatwiej, taniej, szybciej i bezpieczniej jest najpierw przetestować model w kontrolowanych warunkach laboratoryjnych.

Co równie ważne, etap ewaluacji służy także do tego, aby upewnić się, że model spełnia zakładane cele biznesowe. Przypomnijmy, że głównym celem nauki o danych dla firm jest wspieranie procesu podejmowania decyzji i że rozpoczęliśmy proces, koncentrując się na problemie biznesowym, który chcielibyśmy rozwiązać. Zazwyczaj rozwiązanie z zakresu eksploracji danych jest tylko elementem szerszego rozwiązania i jako takie musi podlegać ewaluacji. Ponadto, nawet jeśli model przejdzie rygorystyczne testy ewaluacyjne w „warunkach laboratoryjnych", to mogą istnieć uwarunkowania zewnętrzne, które sprawią, że będzie on niepraktyczny. Na przykład typową wadą rozwiązań z zakresu wykrywania nadużyć (np. wykrywania oszustw, wykrywania spamu i monitorowania włamań do systemów informatycznych) jest to, że generują one zbyt wiele fałszywych alarmów. Model może być bardzo dokładny (>99%) według standardów laboratoryjnych, ale jego ewaluacja w rzeczywistym kontekście biznesowym może ujawnić, że nadal powoduje zbyt wiele fałszywych alarmów, aby być ekonomicznie wykonalny. (Ile kosztowałoby zatrudnienie personelu, który miałby obsługiwać wszystkie te fałszywe alarmy? Jaki byłby koszt w kategoriach niezadowolenia klientów?).

Ewaluacja wyników eksploracji danych obejmuje zarówno oceny ilościowe, jak i jakościowe. Różni decydenci miewają różne interesy w procesie podejmowania decyzji biznesowych, który będzie przeprowadzany lub wspierany przez otrzymane modele. W wielu przypadkach tacy decydenci muszą „zaakceptować" wykorzystanie modelu i aby to zrobić, muszą być usatysfakcjonowani jakością jego decyzji. Określenie „usatysfakcjonowani jakością" może mieć różne znaczenia w przypadku różnych zastosowań, ale często decydenci poszukują modelu, który przyniesie więcej pożytku niż szkody, a zwłaszcza takiego, w przypadku którego prawdopodobieństwo popełnienia katastrofalnych błędów jest niewielkie[2]. Aby ułatwić taką ocenę jako-

[2] W ramach pewnego projektu z zakresu eksploracji danych stworzono na przykład model diagnozowania problemów w sieciach telefonii lokalnej i wysyłania techników do prawdopodobnych miejsc występowania tych problemów. Przed wdrożeniem zespół decydentów firmy telekomunikacyjnej zażądał przekonstruowania modelu tak, aby mógł on zawierać pewne nietypowe rozwiązania, charakterystyczne dla szpitali.

ściową, badacz danych musi pamiętać, aby model był *zrozumiały* dla decydentów (a nie tylko dla badaczy danych). A jeżeli same modele nie są zrozumiałe (np. model jest bardzo złożonym wzorem matematycznym), to badacze danych powinni wiedzieć co zrobić, aby zrozumiałe było zachowanie tych modeli.

Istotna jest wreszcie zrozumiała platforma ewaluacyjna, ponieważ uzyskanie szczegółowych informacji o skuteczności wdrożonego modelu może być trudne lub niemożliwe. Często istnieje tylko ograniczony dostęp do środowiska wdrożeniowego, więc dokonanie kompleksowej ewaluacji „w działaniu" jest skomplikowane. Wdrożone systemy zawierają zazwyczaj wiele „ruchomych części" i ocena funkcjonowania pojedynczej części jest trudna. Firmy posiadające wysokiej klasy zespoły badaczy danych rozsądnie tworzą środowiska pomiarowe, które najlepiej jak to możliwe odzwierciedlają warunki operacyjne, aby można było uzyskać jak najbardziej realistyczne ewaluacje przed podjęciem ryzyka wdrożenia.

W pewnych przypadkach możemy jednak zechcieć rozszerzyć ewaluację na środowisko wdrożeniowe, na przykład za pomocą odpowiedniego ustawienia funkcjonującego systemu, aby móc przeprowadzać randomizowane eksperymenty. Gdybyśmy w naszym przykładzie z odpływem abonentów uznali na podstawie testów laboratoryjnych, że model stworzony na podstawie eksploracji danych da nam lepszy wskaźnik zmniejszenia odpływu abonentów, to być może chcielibyśmy przejść do ewaluacji „in vivo", w której system rzeczywisty (produkcyjny) losowo stosuje model do niektórych klientów, a inni klienci stanowią grupę kontrolną (przypomnij sobie nasze omówienie modelowania przyczynowego w rozdziale 1.). Takie eksperymenty muszą być starannie zaprojektowane, a szczegóły techniczne pozostają poza zakresem tej książki. Zainteresowani czytelnicy mogą zacząć od krótkich artykułów Rona Kohavi i współautorów (Kohavi i in., 2007, 2009, 2012), które zawierają wnioski z zakończonych projektów. Moglibyśmy także zechcieć ewaluować już wdrożone systemy, aby upewnić się, że świat nie zmienia się ze szkodą dla podejmowanych przez model decyzji. Na przykład w pewnych przypadkach, takich jak oszustwa lub spam, zachowanie może ulec zmianie w bezpośredniej reakcji na wdrożenie modeli. Dodatkowo wynik generowany przez model zależy przede wszystkim od danych wejściowych; mogą się one zmieniać pod względem formatu i treści, często nie wzbudzając niepokoju w zespole badaczy danych. Raeder i in. (2012) przedstawiają szczegółowe omówienie aranżacji systemu tak, aby móc dać sobie radę z takimi i innymi problemami z zakresu ewaluacji w trakcie wdrożenia.

Wdrożenie

W fazie wdrożenia wyniki eksploracji danych — oraz w coraz większym stopniu techniki eksploracji danych jako takie — zaczynają być stosowane w warunkach rzeczywistych w celu realizacji określonego zwrotu z inwestycji. Najczystszymi przypadkami wdrażania jest wprowadzenie modeli predykcyjnych do jakiegoś systemu informatycznego lub procesu biznesowego. W naszym przykładzie z odpływem abonentów model do prognozowania prawdopodobieństwa ich utraty mógłby zostać zintegrowany z biznesowym procesem zarządzania odpływem klientów — na przykład poprzez wysyłanie ofert do klientów, w przypadku których przewiduje się, że istnieje największe ryzyko rezygnacji. (Omówimy to szerzej w dalszej części książki). Nowy model wykrywania oszustw może zostać wbudowany w system informatyczny zarządzający personelem, w celu monitorowania rachunków i tworzenia „przypadków", które będzie można przedstawić analitykom ds. oszustw.

Coraz częściej wdrażane są techniki eksploracji danych jako takie. Na przykład przy targetowaniu określonych odbiorców reklam internetowych wdrażane są systemy, które automatycznie

budują (oraz testują) modele w warunkach produkcyjnych, gdy pojawia się nowa kampania reklamowa. Dwoma głównymi powodami wdrażania raczej samego systemu eksploracji danych, a nie modeli tworzonych przez system eksploracji danych są następujące fakty: (i) świat może zmieniać się szybciej niż możliwości dostosowania się do tych zmian ze strony zespołu badaczy danych, jak w przypadku oszustw i wykrywania włamań, oraz (ii) firma ma dla swojego zespołu nauki o danych zbyt dużo zadań związanych z modelowaniem danych, aby ręcznie doprecyzowywać każdy model z osobna. W takich przypadkach najlepszym rozwiązaniem może być włączenie fazy eksploracji danych w proces produkcji. Bardzo istotne jest wtedy takie zorganizowanie procesu, aby ostrzegał on zespół do spraw nauki o danych o wszelkich możliwych anomaliach i zapewniał bezawaryjne działanie (Raeder i in., 2012).

Wdrożenie może być także znacznie mniej „techniczne". Słynny jest przypadek, kiedy eksploracja danych ujawniła zbiór zasad, które mogły pomóc w szybkim zdiagnozowaniu i naprawieniu powszechnego błędu w sferze druku przemysłowego. Wdrożenie polegało po prostu na przyklejeniu taśmą klejącą kartki papieru, zawierającej zasady, do bocznych ścian drukarek (Evans i Fisher, 2002). Wdrożenie może być także znacznie bardziej subtelne, obejmując zmianę procedur pozyskiwania danych lub zmianę strategii, zasad marketingu albo funkcjonowania, wynikającą z wniosków wyciągniętych z eksploracji danych.

Wdrażanie modelu w systemie produkcyjnym wymaga zwykle przekodowania modelu do potrzeb środowiska produkcyjnego, zazwyczaj w celu zwiększenia prędkości lub kompatybilności z istniejącym systemem. Może się to wiązać ze znaczącymi wydatkami i inwestycjami. W wielu przypadkach zespół nauki o danych jest odpowiedzialny za stworzenie funkcjonującego prototypu, wraz z jego ewaluacją. Przekazywane są one zespołowi tworzącemu oprogramowanie.

Praktycznie rzecz ujmując, istnieje ryzyko związane z transferami „przez ścianę" z obszaru nauki o danych do obszaru tworzenia oprogramowania. Pomocne może tu być zapamiętanie maksymy: „Twój model nie jest tym, co zaprojektowali badacze danych, tylko tym, co zbudowali inżynierowie". Z punktu widzenia zarządzania wskazane jest, aby członkowie zespołu programistów zaangażowali się odpowiednio wcześnie w projekt z dziedziny nauki o danych. Mogą zacząć jako doradcy, dostarczając krytycznych spostrzeżeń zespołowi analityków. W praktyce coraz częściej ci konkretni programiści to „inżynierowie nauki o danych" — inżynierowie oprogramowania, którzy mają konkretną wiedzę zarówno w zakresie systemów produkcyjnych, jak i nauki o danych. Ci deweloperzy w miarę postępu projektu stopniowo przejmują coraz większą odpowiedzialność. W pewnym momencie deweloperzy przejmą prowadzenie i własność produktu. Ogólnie rzecz biorąc, badacze danych powinni nadal pozostać zaangażowani w projekt, do jego ostatecznego wdrożenia, jako doradcy lub deweloperzy, w zależności od ich umiejętności.

Niezależnie od tego, czy wdrożenie się powiedzie, proces często wraca do fazy zrozumienia uwarunkowań biznesowych. Proces eksploracji danych umożliwia szczegółowy wgląd w problem biznesowy i trudności w jego rozwiązaniu. Druga iteracja może przynieść udoskonalone rozwiązanie. Samo doświadczenie w myśleniu o firmie, danych i celach zadaniowych często prowadzi do nowych pomysłów poprawy wyników biznesowych, a nawet do odkrywania nowych sfer działalności lub nowych przedsięwzięć.

Należy zauważyć, że do ponownego rozpoczęcia cyklu nie jest konieczna porażka w trakcie wdrażania. Etap ewaluacji może wykazać, że wyniki nie są wystarczająco dobre, aby wdrożyć model, i musimy doprecyzować definicję problemu czy pozyskać inne dane. To działanie przedstawia „skrót" powrotny z fazy ewaluacji do fazy zrozumienia uwarunkowań biznesowych w ramach schematu procesu. W praktyce powinny istnieć skróty powrotne z każdego etapu do każdego wcześniejszego, ponieważ proces zawsze zachowuje pewne aspekty eksploracyjne, a projekt powinien być na tyle elastyczny, aby można było zrewidować wcześniejsze kroki na podstawie dokonanych odkryć.[3]

Implikacje w sferze zarządzania zespołem nauki o danych

Kuszące — ale zazwyczaj błędne — jest postrzeganie procesu eksploracji danych jako cyklu rozwoju oprogramowania. Projekty eksploracji danych są rzeczywiście często traktowane i zarządzane jak projekty inżynieryjne, co jest zrozumiałe, gdy są inicjowane przez działy oprogramowania, gdzie dane generowane są przez ogromne systemy oprogramowania, a wyniki analiz trafiają do nich z powrotem. Menedżerowie znają zazwyczaj technologie oprogramowania i zarządzanie projektami dotyczącymi oprogramowania nie sprawia im problemów. Można ustalić cele częściowe, a sukces jest zwykle jednoznaczny. Przyglądając się cyklowi eksploracji danych CRISP (rysunek 2.2), zarządzający oprogramowaniem mogą myśleć, że wygląda on podobnie do cyklu rozwoju oprogramowania, więc powinni czuć się pewnie, zarządzając projektem analitycznym w ten sam sposób.

Może to być błąd, ponieważ eksploracja danych to przedsięwzięcie odkrywcze, bliższe działaniom z zakresu badań i rozwoju niż inżynierii. Cykl CRISP opiera się na eksploracji; powtarza raczej *podejścia* i *strategię*, a nie projekty oprogramowania. Wyniki są znacznie mniej pewne, a rezultaty danego etapu mogą zmieniać podstawowe rozumienie problemu. Konstruowanie rozwiązania z zakresu eksploracji danych tak, aby od razu je wdrożyć, może być kosztownym i przedwczesnym przedsięwzięciem. Zamiast tego projekty analityczne powinny przygotowywać do inwestowania w informacje w celu różnorodnego zmniejszania stopnia niepewności. Niewielkie inwestycje mogą być prowadzone poprzez badania pilotażowe i jednorazowe prototypy. Badacze danych powinni przeglądać literaturę, sprawdzając, co jeszcze zrobiono i na ile się to udało. Na większą skalę zespół może znacząco inwestować w budowę platform do testów eksperymentalnych, aby umożliwić prowadzenie bardziej rozbudowanych eksperymentów z zakresu modelowania zwinnego. Jeśli jesteś menedżerem działu oprogramowania, będzie Ci się to wszystko kojarzyć z badaniami i eksploracją w znacznie większym stopniu, niż jesteś przyzwyczajony, co może stać się źródłem dyskomfortu.

Umiejętności programistyczne kontra umiejętności analityczne

Chociaż eksploracja danych wiąże się z obecnością oprogramowania, to wymaga umiejętności, które niekoniecznie są powszechne wśród programistów. W dziedzinie inżynierii oprogramowania być może najważniejsza jest umiejętność pisania efektywnego, wysokiej jakości kodu na podstawie wymagań. Członków zespołu można oceniać za

[3] Specjaliści od oprogramowania być może rozpoznają tutaj podobieństwo do filozofii „Polegnij szybciej, żeby wygrać wcześniej" (Muoio, 1997).

pomocą wskaźników takich jak ilość napisanego kodu lub liczba zamkniętych raportów błędów. W sferze analityki bardziej istotna jest umiejętność właściwego formułowania problemów, szybkiego proponowania rozwiązań, przyjmowania rozsądnych założeń w obliczu niewłaściwie sformułowanych problemów, projektowania eksperymentów, które są dobrymi inwestycjami, i analizy wyników. Podczas budowania zespołu badaczy danych te właśnie cechy, a nie tradycyjna wiedza fachowa w zakresie inżynierii oprogramowania, to umiejętności, których należy poszukiwać.

Inne techniki i technologie analityczne

Analityka biznesowa polega na stosowaniu różnych technologii do analizy danych. Wiele z nich wykracza poza zakres tej książki, obejmujący myślenie w kategoriach analityki danych i zasady wydobywania z danych przydatnych wzorców. Istotne jest jednak, aby mieć świadomość istnienia tych pokrewnych technik, rozumieć ich cele, odgrywaną przez nie rolę i wiedzieć, kiedy konsultacja z ekspertami w ich dziedzinie może się okazać korzystna.

W tym celu prezentujemy sześć grup powiązanych technik analitycznych. Tam, gdzie to konieczne, dokonujemy porównań i przedstawiamy różnice pomiędzy nimi i eksploracją danych. Główną różnicą jest to, że eksploracja danych koncentruje się na *zautomatyzowanym* poszukiwaniu w danych *wiedzy*, *wzorców* czy też *prawidłowości*[4]. Dla analityka biznesowego istotna jest umiejętność rozpoznawania, jakiego rodzaju technika analityczna jest odpowiednia do rozwiązania konkretnego problemu.

Statystyka

Określenie „statystyka" ma dwa różne zastosowania w analizach biznesowych. Po pierwsze, jest ono stosowane jako zbiorczy termin odnoszący się do obliczania na podstawie danych konkretnych wartości liczbowych, które nas interesują (np. „Musimy zebrać statystyki wykorzystania naszych klientów w celu ustalenia, co jest nie tak"). Te wartości to często sumy, średnie, stopy itd. Nazwijmy je „miarami rozkładu". Często zależy nam na pogłębieniu wiedzy i obliczaniu miar rozkładu *warunkowo* dla jednego lub kilku podzbiorów populacji (np. „Czy poziom rezygnacji różni się w przypadku klientów płci męskiej i żeńskiej?" oraz „A co z klientami o wysokich dochodach z północnego wschodu?"). Miary rozkładu są podstawowym budulcem teorii i praktyki nauki o danych.

Miary rozkładu powinno się dobierać, zwracając szczególną uwagę na to, jaki problem biznesowy ma zostać rozwiązany (to jedna z podstawowych zasad, którą przedstawimy dalej), a także uwzględniając *rozkład* danych, których dotyczą. Na przykład średni (średnia arytmetyczna) roczny dochód w Stanach Zjednoczonych, zgodnie ze spisem ludności Census Bureau z 2004 roku, wynosił ponad 60 000 dolarów. Gdybyśmy mieli użyć go jako miary średniego dochodu w celu podejmowania decyzji politycznych, to sami wprowadzilibyśmy się w błąd. Rozkład dochodów w Stanach Zjednoczonych jest bardzo nierównomierny; wiele osób zarabia dość mało, ale są i takie, które zarabiają fantastycznie dużo. W takich przypadkach średnia arytmetyczna mówi nam stosunkowo niewiele o rzeczywistych poziomach zarobków. Zamiast niej

[4] Istotne jest, aby pamiętać, że odkrycie rzadko miewa charakter całkowicie zautomatyzowany. Ważnym czynnikiem jest fakt przynajmniej częściowego automatyzowania przez eksplorację danych procesu poszukiwania i odkrywania, a nie zapewnianie wsparcia technicznego dla poszukiwania i odkrywania prawidłowości ręcznie.

powinniśmy zastosować inną miarę „średniego" dochodu, taką jak mediana. Mediana dochodu — kwota, w stosunku do której połowa populacji zarabia więcej, a druga połowa mniej — wyniosła według spisu ludności w USA w 2004 r. tylko 44 389 dolarów, czyli znacznie poniżej średniej. Ten przykład może wydawać się oczywisty, bo jesteśmy przyzwyczajeni do słuchania o „medianie dochodu", ale samo rozumowanie odnosi się do wszelkich obliczeń związanych z miarami rozkładu: czy pomyślałeś o problemie, który chciałbyś rozwiązać, lub o pytaniu, na które chciałbyś odpowiedzieć? Czy wziąłeś pod uwagę rozkład danych i to, czy wybrana miara jest odpowiednia?

Inne znaczenie terminu „statystyka" określa nazywaną tak naukę. Statystyka jako nauka daje nam ogromną ilość wiedzy, która leży u podstaw analityki i może być traktowana jako składnik szerszej dziedziny nauki o danych. Statystyka pomaga nam na przykład zrozumieć różne rozkłady danych oraz to, jakie miary są właściwe do ich określenia. Statystyka pomaga zrozumieć, jak korzystać z danych do testowania hipotez i szacowania niepewności wniosków. W odniesieniu do eksploracji danych testowanie hipotez może nam pomóc w ustaleniu, czy obserwowany wzorzec może być przekonującą ogólną prawidłowością, a nie tylko przypadkowym wystąpieniem w jakimś konkretnym zbiorze danych. Najbardziej istotny z punktu widzenia tej książki jest fakt, że wiele technik wydobywania wzorców lub tworzenia modeli na podstawie danych ma swoje korzenie w statystyce.

Badanie wstępne może na przykład sugerować, że w przypadku klientów z północnego wschodu Stanów Zjednoczonych wskaźnik utraty wynosi 22,5%, podczas gdy średnio w całym kraju jest to tylko 15%. Może to być po prostu przypadkowa fluktuacja, bo wskaźnik utraty nie ma stałej wartości; jest on zmienny w różnych częściach kraju i w czasie, należy więc oczekiwać różnic. Wartość wskaźnika dla północnego wschodu to jednak półtorej średniej w całych Stanach Zjednoczonych i wartość ta wydaje się niezwykle wysoka. Jaka jest szansa, że wynika to ze zmienności losowej? Do odpowiadania na takie pytania służy statystyczne testowanie hipotez.

Blisko spokrewniona z wcześniejszymi rozważaniami jest kwantyfikacja niepewności na przedziały ufności. Ogólny wskaźnik odpływu abonentów wynosi 15%, ale istnieje tutaj pewna różnica; tradycyjna analiza statystyczna może wykazać, że w 95% przypadków wskaźnik utraty wyniesie pomiędzy 13% a 17%.

Kontrastuje to z (komplementarnym) procesem eksploracji danych, który może być postrzegany jako *tworzenie* hipotezy. Przede wszystkim, czy możemy znaleźć wzorce w danych? Po stworzeniu hipotezy powinno nastąpić staranne jej przetestowanie (zazwyczaj na innych danych, por. rozdział 5.). Dodatkowo procedury eksploracji danych mogą tworzyć szacunki liczbowe, a nam często zależy również na tym, aby umieścić je w przedziałach ufności. Wrócimy do tego przy omawianiu ewaluacji wyników eksploracji danych.

W tej książce nie będziemy poświęcać więcej czasu na omawianie tych podstawowych pojęć statystycznych. Istnieje wiele książek wprowadzających do zagadnienia statystyki i statystyki biznesowej, a wszystko, co chcielibyśmy ewentualnie wtłoczyć w ramy tej książki, byłoby albo stanowczo zbyt wąskie, albo powierzchowne.

Mając na uwadze powyższe, musimy jednak stwierdzić, że istnieje termin statystyczny, który można często usłyszeć w kontekście analiz biznesowych. Termin ten to „korelacja". Na przykład: „Czy istnieją jakieś wskaźniki, które korelują z późniejszym odpływem klientów?". Podobnie jak statystyka, określenie „korelacja" ma zarówno znaczenie ogólne (zmiany danej wielkości mówią nam coś o zmianach innych), jak i konkretne znaczenie techniczne (np. korelacja

liniowa, oparta na konkretnym wzorze matematycznym). Pojęcie korelacji będzie punktem wyjścia dla pozostałej części naszej dyskusji o badaniu danych do celów biznesowych, poczynając od kolejnego rozdziału.

Zapytania do baz danych

Zapytanie to określone polecenie dostarczenia podzbioru danych lub statystyk dotyczących danych, sformułowane w języku technicznym i wprowadzane do systemu baz danych. Istnieje wiele narzędzi służących do odpowiadania na jednorazowe lub powtarzające się zapytania dotyczące danych zadawane przez analityka. Narzędzia te to zazwyczaj elementy interfejsu systemów baz danych oparte na SQL (Structured Query Language — strukturalny język zapytań) lub narzędzia z graficznym interfejsem użytkownika (GUI), pomagające w formułowaniu zapytań (np. technika Query By Example — QBE). Jeżeli na przykład analityk potrafi zdefiniować, że coś jest „dochodowe" w kategoriach operacyjnych, możliwych do obliczenia z pozycji w bazie danych, to narzędzie zapytań mogłoby odpowiedzieć na pytanie: „Którzy klienci są najbardziej dochodowi w Warszawie?". Analityk może następnie uruchomić zapytanie i otrzymać listę najbardziej dochodowych klientów, uszeregowanych w kolejności według dochodowości. To działanie różni się zasadniczo od eksploracji danych, ponieważ nie wiąże się z odkrywaniem wzorców lub modeli.

Zapytania do baz danych są właściwe, gdy analityk ma już koncepcję dotyczącą tego, co mogłoby być interesującą subpopulacją w ramach danych, i chce zbadać populację lub potwierdzić jakąś dotyczącą jej hipotezę. Na przykład, jeśli analityk podejrzewa, że mężczyźni w średnim wieku mieszkający w Warszawie przejawiają jakieś szczególnie interesujące zachowania związane z rezygnowaniem z określonych usług, to mógłby sformułować zapytanie SQL o treści:

```
SELECT * FROM KLIENCI WHERE WIEK > 45 and PLEC = 'M' and MIASTO = 'WARSZAWA'
```

Jeśli są to osoby, do których ma zostać skierowana oferta, to narzędzie obsługi zapytań może zostać wykorzystane do pobrania wszystkich informacji o nich (*) z tabeli `KLIENCI` w bazie danych.

W przeciwieństwie do powyższego, eksploracja danych może zostać wykorzystana, aby przede wszystkim określić cel tego zapytania — jako wzorzec lub prawidłowość w danych. Procedura eksploracji danych może zbadać wcześniejszych klientów, którzy zrezygnowali i nie zrezygnowali, i ustalić, że ten segment (charakteryzowany jako „WIEK powyżej 45 i PŁEĆ mężczyzna i MIASTO Warszawa") jest czynnikiem prognostycznym w odniesieniu do wskaźnika utraty klientów. Po przetłumaczeniu tego na zapytanie SQL narzędzie obsługi zapytań może następnie zostać wykorzystane do znalezienia pasujących rekordów w bazie danych.

Narzędzia zapytań mają na ogół możliwość posługiwania się zaawansowaną logiką, w tym obliczaniem statystyk podsumowujących dla subpopulacji, sortowaniem, łączeniem wielu tabel zawierających pokrewne dane i wieloma innymi kwestiami. Badacze danych często stają się biegli w formułowaniu zapytań w celu wydobycia danych, na których im zależy.

Przetwarzanie analityczne online (On-line Analytical Processing — OLAP) zapewnia łatwy w użyciu interfejs graficzny (Graphical User Interface — GUI) do eksplorowania dużych zbiorów danych, często połączony z magazynem danych. Koncepcja przetwarzania online oznacza, że jest ono wykonywane w czasie rzeczywistym, więc analitycy i decydenci w firmie mogą przeprowadzać analizy szybko i sprawnie. Inaczej niż w przypadku zapytań ad hoc, możliwych dzięki narzędziom typu SQL, w systemie OLAP rozmiary analizy muszą jednak zostać

wstępnie zaprogramowane. Jeśli ustaliliśmy, że chcemy zbadać wielkość sprzedaży w funkcji regionu i czasu, to możemy te trzy wartości zaprogramowane w systemie wprowadzić do populacji, często po prostu klikając, przeciągając i manipulując dynamicznymi wykresami.

Systemy OLAP zostały zaprojektowane w celu ułatwienia ręcznej lub wizualnej eksploracji danych przez analityków. OLAP nie wykonuje modelowania ani automatycznego wyszukiwania wzorców. Narzędzia eksploracji danych mogą z kolei, inaczej niż w przypadku OLAP, mówiąc ogólnie, włączać do procesu eksploracji nowe wymiary analizy. Narzędzia OLAP mogą być użytecznym uzupełnieniem narzędzi eksploracji danych w celu pozyskiwania użytecznych informacji z danych biznesowych.

Magazynowanie danych

Magazyny danych zbierają i łączą dane z obszaru całej firmy, często z wielu systemów przetwarzania transakcji, z których każdy ma własną bazę danych. Systemy analityczne mają dostęp do magazynów danych. Magazynowanie danych może być postrzegane jako technologia wspierająca eksplorację danych. Magazynowanie danych nie zawsze jest konieczne, ponieważ większość działań z zakresu eksploracji danych nie wymaga dostępu do magazynów danych, ale firmy, które decydują się zainwestować w budowę magazynów danych, często wykorzystują eksplorację danych w ramach organizacji znacznie szerzej. Jeżeli na przykład magazyn danych integruje rekordy z działów sprzedaży i fakturowania oraz z działu zasobów ludzkich, to można go wykorzystać do wyszukania wzorców charakterystycznych dla skutecznych sprzedawców.

Analiza regresji

Niektóre z metod omawianych w tej książce są także rdzeniem innego zbioru metod analitycznych, które często określa się wspólnym mianem **analizy regresji** i które są szeroko stosowane w dziedzinie statystyki, a także w innych dziedzinach opartych na analizie ekonometrycznej. W naszej książce koncentrujemy się na innym zestawie zagadnień niż ten, który zazwyczaj można znaleźć w książce lub na zajęciach poświęconych analizie regresji. Tutaj interesują nas nie tyle wyjaśnienia dotyczące określonego zestawu danych, ale wydobywanie wzorców, które mogą zostać uogólnione i zastosowane do innych danych w celu udoskonalenia określonych procesów biznesowych. Zazwyczaj będzie się to wiązało z szacowaniem lub przewidywaniem wartości przypadków niewystępujących w analizowanym zbiorze danych. Tak więc, na przykład, w tej książce mniej interesować nas będzie dochodzenie przyczyn odpływu abonentów (istotne samo w sobie) w określonym zbiorze danych historycznych, a bardziej prognozowanie, do których klientów spośród tych, którzy jeszcze nie zrezygnowali, powinniśmy się zwrócić, aby zapobiec ich rezygnacji w przyszłości. Dlatego poświęcimy nieco czasu na omówienie testowania wzorców na nowych danych w celu oceny stopnia ich ogólności oraz technik zmniejszania tendencji do znajdowania wzorców charakterystycznych dla określonego zestawu danych, ale nie uniwersalnych dla całej populacji, z której pochodzą dane.

Tematyka różnic między modelowaniem objaśniającym i modelowaniem predykcyjnym mogłaby wywołać gorącą debatę[5], wykraczającą daleko poza sferę naszego zainteresowania. Ważne

[5] Zainteresowanych czytelników zachęcamy do zapoznania się z dyskusją w pracy: Shmueli, 2010.

jest, aby uświadomić sobie, że istnieją tutaj znaczne podobieństwa w zakresie stosowanych *technik*, ale wnioski z modelowania objaśniającego nie zawsze mają zastosowanie do modelowania predykcyjnego. Czytelnik mający pewne przygotowanie w zakresie analizy regresji może więc stanąć w obliczu nowych i nawet pozornie sprzecznych wniosków[6].

Uczenie maszynowe i eksploracja danych

Zbiór metod wydobywania modeli (predykcyjnych) z danych, znanych obecnie jako metody uczenia maszynowego, został opracowany w kilku dziedzinach jednocześnie, w szczególności w ramach uczenia maszynowego, statystyki stosowanej i rozpoznawania wzorców. Uczenie maszynowe jako dziedzina badań powstało jako obszar funkcjonujący w ramach sztucznej inteligencji, której badania koncentrowały się na metodach doskonalenia wiedzy i wydajności inteligentnego agenta w funkcji czasu, w reakcji na zdobywanie przez agenta doświadczenia w świecie. To doskonalenie często wiąże się z analizowaniem danych z otoczenia i dokonywaniem predykcji odnośnie nieznanych wielkości. Z biegiem czasu ten aspekt uczenia maszynowego związany z analizowaniem danych zaczął odgrywać bardzo znaczącą rolę w tej dziedzinie. Gdy metody uczenia maszynowego rozpowszechniły się, dyscypliny naukowe uczenia maszynowego, statystyki stosowanej i rozpoznawania wzorców wytworzyły ścisłe powiązania między sobą i podziały pomiędzy nimi uległy zatarciu.

Eksploracja danych (albo odkrywanie wiedzy i eksploracja danych — *Knowledge Discovery and Data Mining* [KDD]) jako dziedzina nauki powstała jako gałąź obszaru uczenia maszynowego i obie pozostają ze sobą ściśle powiązane. Obie dziedziny zajmują się analizą danych w celu wyszukiwania przydatnych lub informatywnych wzorców. Techniki i algorytmy są jednakowe; faktycznie oba obszary są ze sobą tak ściśle powiązane, że badacze często funkcjonują w obu społecznościach i bez problemu przemieszczają się pomiędzy nimi. Mimo wszystko warto jednak wskazać niektóre różnice, aby określić punkt widzenia.

Mówiąc ogólnie, dziedzina uczenia maszynowego zajmuje się wieloma rodzajami działań służących poprawie skuteczności i w związku z tym zawiera podpola, takie jak robotyka i wizja komputerowa, które nie są elementem KDD. Zajmuje się także kwestiami **sprawczości** i **kognicji** — czyli tego, w jaki sposób inteligentny agent będzie wykorzystywał odkrytą wiedzę do rozumowania i działania w swoim środowisku — czym nie zajmuje się eksploracja danych.

Historycznie KDD została wydzielona z dziedziny uczenia maszynowego jako obszar badań zorientowany na kwestie związane z rzeczywistymi zastosowaniami i półtorej dekady później społeczność KDD pozostaje bardziej związana z zastosowaniami niż dziedzina uczenia maszynowego. Wobec tego badania skoncentrowane na zastosowaniach komercyjnych i zagadnienia biznesowe analizy danych mają tendencję do kierowania się raczej w stronę społeczności eksploracji danych niż uczenia maszynowego. Eksploracja danych wydaje się być także bardziej związana z całym procesem analityki danych: przygotowywaniem danych, uczeniem modeli, ewaluacją itd.

[6] Osoby badające tę kwestię dokładniej zrozumieją tę pozorną sprzeczność. Takie szczegółowe badania nie są jednak konieczne, aby zrozumieć podstawowe zasady.

Odpowiadanie na pytania biznesowe z wykorzystaniem tych technik

W celu pokazania, w jaki sposób techniki te są wykorzystywane w analityce biznesowej, należy zastanowić się nad zestawem pytań, które mogą się pojawić, oraz technologii, które byłyby odpowiednie, aby na nie odpowiedzieć. Pytania te są powiązane ze sobą, ale każde z nich jest nieco inne. Ważne jest zrozumienie tych różnic, aby z kolei zrozumieć, jakie technologie należy zastosować i z kim być może należałoby się skonsultować.

1. *Którzy klienci są najbardziej rentowni?*

   Jeśli „rentowność" może zostać jasno zdefiniowana na podstawie istniejących danych, to mamy do czynienia z prostym zapytaniem do bazy danych. Standardowe narzędzie obsługi zapytań można wykorzystać do pobrania zbioru rekordów klientów z bazy danych. Wyniki mogą zostać posortowane według łącznej kwoty transakcji lub jakiegoś innego operacyjnego wskaźnika rentowności.

2. *Czy rzeczywiście istnieje różnica między klientami rentownymi i przeciętnymi?*

   To pytanie o domysły, czyli hipoteza (w tym przypadku: „Dla firmy istnieje różnica wartości między klientem rentownym a klientem przeciętnym"), i do potwierdzenia lub zaprzeczenia jej należy wykorzystać testowanie hipotez statystycznych. Analiza statystyczna może również określić prawdopodobieństwo lub stopień pewności, że ta różnica jest prawdziwa. Zazwyczaj wynik będzie wyglądał następująco: „Wartość rentownych klientów jest znacząco inna niż przeciętnego klienta, z prawdopodobieństwem <5%, że jest to spowodowane przypadkiem".

3. *Ale kim tak naprawdę są ci klienci? Czy mogę ich scharakteryzować?*

   Często chcielibyśmy zrobić więcej, niż tylko stworzyć listę rentownych klientów. Chcielibyśmy opisać ich wspólne cechy. Cechy poszczególnych klientów można wyodrębnić z bazy danych, stosując takie techniki jak zapytania do baz danych, które można również wykorzystać do generowania miar rozkładu. Głębsza analiza powinna dotyczyć ustalenia, jakie cechy *różnicują* klientów rentownych i pozostałych. To domena nauki o danych, wykorzystującej techniki eksploracji danych do zautomatyzowanego wyszukiwania wzorców — co będziemy omawiali szczegółowo w kolejnych rozdziałach.

4. *Czy określony nowy klient będzie rentowny? Jaki oczekiwany przychód jest w stanie wygenerować?*

   Takie pytania mogłyby pochodzić od technik eksploracji danych, które badają historyczne rekordy klientów i tworzą predykcyjne modele rentowności. Takie techniki generują z danych historycznych modele, które mogą następnie zostać zastosowane w celu tworzenia prognoz dla nowych klientów. To również zostanie omówione w kolejnych rozdziałach.

Należy zwrócić uwagę na fakt, że ostatnia para pytań to subtelnie różniące się pytania z zakresu eksploracji danych. Pierwsze pytanie, klasyfikacyjne, może zostać sformułowane jako predykcja, czy dany nowy klient będzie rentowny (tak/nie lub stopień prawdopodobieństwa). Drugie można sformułować jako predykcję wartości (liczbowej), którą klient wniesie do firmy. Więcej na ten temat w kolejnych rozdziałach.

Podsumowanie

Eksploracja danych to rzemiosło. Jak w przypadku wielu rzemiosł, mamy tu do czynienia z jednoznacznie zdefiniowanym procesem, który może przyczynić się do zwiększenia prawdopodobieństwa osiągnięcia pomyślnego wyniku. Proces ten jest podstawowym narzędziem koncepcyjnym w rozważaniach o projektach z zakresu nauki o danych. W całej tej książce będziemy wielokrotnie wracać do procesu eksploracji danych, ukazując, w jaki sposób każde podstawowe pojęcie do niego pasuje. Z kolei zrozumienie podstaw nauki o danych znacznie zwiększa szanse na sukces, gdy firma uruchamia proces eksploracji danych.

W ramach różnych dziedzin wiedzy związanych z nauką o danych opracowano zestaw kanonicznych typów zadań, takich jak klasyfikacja, regresja i klastrowanie. Każdy typ zadania służy innemu celowi i posiada przypisany zestaw technik umożliwiających jego rozwiązanie. Badacz danych zwykle atakuje nowy projekt, rozkładając go na czynniki w taki sposób, że pojawia się jedno lub kilka kanonicznych zadań. Badacz wybiera odpowiednią dla każdego z nich technikę, a następnie łączy rozwiązania. Przeprowadzenie tego procesu we właściwy sposób może wymagać znaczącego poziomu doświadczenia i umiejętności. Udany projekt z zakresu eksploracji danych to inteligentny kompromis pomiędzy tym, co dane mogą przynieść (tzn. co na ich podstawie można przewidzieć i na ile skutecznie), a celami projektu. Z tego powodu ważne jest, aby pamiętać, w jaki sposób wyniki eksploracji danych zostaną wykorzystane, i wykorzystać to z kolei do udoskonalenia samego procesu eksploracji danych.

Eksploracja danych różni się od istotnych technologii wspomagających, takich jak testowanie hipotez statystycznych i zapytania do baz danych (które są przedmiotem innych książek i zajęć), oraz jest w stosunku do nich komplementarna. Choć granice pomiędzy eksploracją danych i związanymi z nią technikami nie zawsze są ostre, istotne jest poznanie możliwości i mocnych stron tych innych technik, aby wiedzieć, kiedy należy je stosować.

Dla menedżera proces eksploracji danych jest użyteczny jako platforma do analizy projektu lub propozycji eksploracji danych. Proces ten umożliwia systematyczne ujęcie tego projektu lub propozycji, zapewniając między innymi zestaw pytań, które można zadać na ich temat, aby móc zrozumieć, czy są dobrze przemyślane czy całkowicie chybione. Wrócimy do tej kwestii po szczegółowym omówieniu szeregu podstawowych zasad jako takich. A tym właśnie zajmiemy się teraz.

Skorowidz

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

Ś

T

U

W

Z

Przeanalizuj posiadane dane i podejmij trafne decyzje!

Posiadanie zbiorów danych to połowa sukcesu. Druga połowa to umiejętność ich skutecznej analizy i wyciągania wniosków! Dopiero na tej podstawie będziesz w stanie właściwie ocenić kondycję Twojej firmy oraz podjąć słuszne decyzje. Wiedza zawarta w tej książce może zadecydować o sukcesie lub porażce Twojego biznesu. Nie ryzykuj i sięgnij po to doskonałe źródło wiedzy poświęcone nauce o danych.

To unikalny podręcznik, który pomoże Ci sprawnie opanować nawet najtrudniejsze zagadnienia związane z analizą danych. Dowiesz się, jak zbudowany jest proces eksploracji danych, z jakich narzędzi możesz skorzystać oraz jak stworzyć model predykcyjny i dopasować go do danych. W kolejnych rozdziałach przeczytasz o tym, czym grozi nadmierne dopasowanie modelu i jak tego unikać oraz jak wyciągać wnioski metodą najbliższych sąsiadów. Na koniec zaznajomisz się z możliwościami wizualizacji skuteczności modelu oraz odkryjesz związek pomiędzy nauką o danych a strategią biznesową. To obowiązkowa lektura dla wszystkich osób chcących podejmować świadome decyzje na podstawie posiadanych danych!

Dzięki tej książce:

- poznasz model predykcyjny
- dowiesz się, jak dopasować model do danych
- zwizualizujesz skuteczność zbudowanego modelu
- zwiększysz swoje szanse na osiągnięcie sukcesu w biznesie!

O'REILLY

helion.pl księgarnia internetowa

Nr katalogowy: 26131

Księgarnia internetowa: http://helion.pl

Zamówienia telefoniczne: 0 801 339900 0 601 339900

one press

Helion

Sprawdź najnowsze promocje: http://helion.pl/promocje
Książki najchętniej czytane: http://helion.pl/bestsellery
Zamów informacje o nowościach: http://helion.pl/nowosci

Helion SA
ul. Kościuszki 1c, 44-100 Gliwice
tel.: 32 230 98 63
e-mail: helion@helion.pl
http://helion.pl

Informatyka w najlepszym wydaniu

sięgnij po WIĘCEJ

KOD KORZYŚCI

ISBN 978-83-246-9610-9

cena: 59,00 zł

9 788324 696109