O'REILLY®

Helion

Data science

Wyzwania i rozwiązania

Jak zostać ekspertem analizy danych

Daniel Vaughan

Tytuł oryginału: Data Science: The Hard Parts: Techniques for Excelling at Data Science

Tłumaczenie: Tomasz Walczak

ISBN: 978-83-289-1294-6

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

Pliki z przykładami omawianymi w książce można znaleźć pod adresem:
https://ftp.helion.pl/przyklady/dasctr.zip

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

Przedmowa **11**

Część I. Techniki analityki danych 17

1. I co z tego? Generowanie wartości dzięki danologii **19**
Czym jest wartość? 19
„Co?”, czyli zrozumieć biznes 20
„Co z tego?”, czyli istota generowania wartości dzięki danologii 22
„Co teraz?”, czyli bądź przebojowy 23
Pomiar wartości 23
Najważniejsze wnioski 25
Dalsza lektura 25

2. Projektowanie wskaźników **27**
Pożądane właściwości wskaźników 27
Mierzalność 27
Możliwość podejmowania działań 28
Trafność 28
Aktualność 28
Dekompozycja wskaźników 29
Lejek analityczny 29
Dekompozycje przepływów i zapasów 30
Dekompozycje typu P×Q 30
Przykład: inny sposób dekompozycji przychodów 31
Przykład: platformy sprzedażowe 31
Najważniejsze wnioski 32
Dalsza lektura 32

3. Dekompozycje wzrostu — zrozumienie przeszkód i sprzyjających czynników 34
Dlaczego dekompozycje wzrostu? 34
Dekompozycja addytywna 34
Przykład 35
Interpretacja i przypadki użycia 36
Dekompozycja multiplikatywna 36
Przykład 37
Interpretacja 37
Dekompozycja zmian wag i wartości 38
Przykład 39
Interpretacja 40
Wyprowadzanie równań matematycznych 40
Dekompozycja addytywna 41
Dekompozycja multiplikatywna 41
Dekompozycja mix-rate 41
Najważniejsze wnioski 42
Dalsza lektura 42

4. Projekty 2×2 43
Argumenty za upraszczaniem 43
Czym jest projekt 2×2? 44
Przykład: testowanie modelu i nowej cechy 45
Przykład: zrozumienie zachowań użytkownika 47
Przykład: udzielanie i akceptacja ofert kredytów 49
Przykład: ustalanie priorytetów w procesie pracy 50
Najważniejsze wnioski 51
Dalsza lektura 51

5. Tworzenie uzasadnienia biznesowego 53
Wybrane zasady tworzenia uzasadnień biznesowych 53
Przykład: proaktywna strategia zatrzymywania klientów 54
Zapobieganie oszustwom 55
Zakup zewnętrznych zbiorów danych 56
Praca nad projektem z obszaru danologii 57
Najważniejsze wnioski 57
Dalsza lektura 58

6. Czym jest wskaźnik przyrostu? 59
Definicja wskaźnika przyrostu 59
Przykład: model klasyfikatora 60
Błędy wynikające z samoselekcji i przeżywalności 60

Inne zastosowania wskaźników przyrostu 62
Najważniejsze wnioski 62
Dalsza lektura 63

7. Narracje 64
Co kryje się w narracji? Opowiadanie historii za pomocą danych 64
Jasna i rzeczowa 65
Wiarygodność 66
Zapadająca w pamięć 67
Możliwość podejmowania działań 68
Tworzenie narracji 68
Nauka jako opowiadanie historii 68
„Co?", „co z tego?" i „co teraz?" 69
Ostatnia prosta 71
Streszczenia TL;DR 71
Wskazówki dotyczące pisania zapadających w pamięć streszczeń TL;DR 71
Przykład: pisanie streszczenia TL;DR tego rozdziału 72
Skuteczne krótkie prezentacje 74
Prezentowanie narracji 74
Najważniejsze wnioski 75
Dalsza lektura 76

8. Wizualizacje danych — wybór właściwego wykresu do przekazania komunikatu 77
Kilka przydatnych i rzadko używanych wizualizacji danych 77
Wykres słupkowy a wykres liniowy 77
Wykres nachylenia 79
Wykres kaskadowy 79
Funkcje wygładzania dla wykresów punktowych 81
Prezentowanie rozkładów na wykresie 82
Ogólne zalecenia 83
Dobierz odpowiednią wizualizację dla przekazu 83
Mądrze dobieraj kolory 84
Różne wymiary na wykresie 85
Staraj się uzyskać odpowiednio wysoki współczynnik dane/atrament 85
Personalizacja a półautomatyzacja 86
Na samym początku dobierz odpowiedni rozmiar czcionki 86
Interaktywne czy nie? 86
Zachowaj prostotę 87
Zacznij od wyjaśnienia wykresu 87
Najważniejsze wnioski 87
Dalsza lektura 88

Część II. Uczenie maszynowe **89**

9. Symulacje i bootstrapping ... 91
Podstawy symulacji 92
Symulacja modelu liniowego i regresji liniowej 94
Czym są wykresy zależności częściowych? 96
Błąd systematyczny z powodu pominięcia zmiennej 100
Symulacja problemu klasyfikacji 103
Modele zmiennych ukrytych 103
Porównanie różnych algorytmów 104
Bootstrapping 106
Najważniejsze wnioski 108
Dalsza lektura 108

10. Regresja liniowa — powrót do podstaw ... 110
Co kryje się za współczynnikiem? 110
Twierdzenie Frischa-Waugha-Lovella 113
Dlaczego twierdzenie FWL jest ważne? 115
Czynniki zakłócające 116
Dodatkowe zmienne 118
Centralna rola wariancji w uczeniu maszynowym 120
Najważniejsze wnioski 123
Dalsza lektura 124

11. Wyciekanie danych ... 125
Czym jest wyciekanie danych? 125
Wynik również jest cechą 125
Funkcja wyniku sama też jest cechą 126
Złe zmienne kontrolne 126
Niewłaściwe oznaczenie znacznika czasu 126
Wiele zbiorów danych z nieprecyzyjnymi agregacjami czasowymi 127
Wyciekanie innych informacji 127
Wykrywanie wyciekania danych 128
Całkowita separacja 130
Metoda okien 132
Wybór długości okien 133
Etap treningu odzwierciedla etap oceny punktowej 134
Wdrażanie metody okien 135
Mam wyciek. Co teraz? 136
Najważniejsze wnioski 136
Dalsza lektura 137

12. Stosowanie modeli w środowisku produkcyjnym 138
Co oznacza „gotowość produkcyjna”? 138
Wsadowa ocena punktowa (w trybie offline) 138
Obiekty modeli czasu rzeczywistego 140
Dryf danych i modelu 141
Etapy niezbędne w każdym potoku produkcyjnym 142
Pobieranie i przekształcanie danych 143
Sprawdzanie poprawności danych 144
Etapy treningu i oceny punktowej 145
Sprawdzanie poprawności modelu i ocen punktowych 146
Zapisywanie modelu i ocen punktowych 146
Najważniejsze wnioski 146
Dalsza lektura 147

13. Opowiadanie historii w uczeniu maszynowym 149
Holistyczne spojrzenie na opowiadanie historii w uczeniu maszynowym 149
Opowiadanie historii przed opracowaniem modelu i w trakcie tego procesu 150
Tworzenie hipotez 151
Inżynieria cech 153
Opowiadanie historii po opracowaniu modelu: otwieranie czarnej skrzynki 156
Kompromis między interpretowalnością a skutecznością 156
Regresja liniowa: ustalenie punktu odniesienia 158
Znaczenie cech 160
Mapa cieplna 161
Wykresy zależności częściowych 163
Skumulowane efekty lokalne 165
Najważniejsze wnioski 166
Dalsza lektura 167

14. Od predykcji do decyzji 168
Analiza procesu podejmowania decyzji 168
Proste reguły decyzyjne oparte na inteligentnym wyznaczaniu wartości progowych 170
Precyzja i czułość 171
Przykład: pozyskiwanie list kontaktów 172
Optymalizacja macierzy błędów 174
Najważniejsze wnioski 176
Dalsza lektura 176

15. Zmiany dodatkowe — Święty Graal danologii? 177
Definiowanie zmian dodatkowych 177
Wnioskowanie przyczynowe w celu poprawy predykcji 178
Wnioskowanie przyczynowe jako wyróżnik 178
Usprawnione podejmowanie decyzji 179

Czynniki zakłócające i kolidery 179
Błąd doboru 182
Założenie o braku zmiennych zakłócających 186
Radzenie sobie z błędem doboru — randomizacja 187
Dopasowywanie 188
Uczenie maszynowe i wnioskowanie przyczynowe 191
Kod otwartoźródłowy 191
Podwójne uczenie maszynowe 192
Najważniejsze wnioski 194
Dalsza lektura 195

16. Testy A/B ... 198
Czym są testy A/B? 198
Kryterium decyzyjne 199
Minimalne wykrywalne efekty 202
Ustalanie mocy statystycznej, poziomu istotności i wartości P 205
Szacowanie wariancji wyniku 205
Symulacje 206
Przykład: współczynniki konwersji 207
Określanie wartości MWE 208
Lista hipotez do zbadania 209
Wskaźnik 209
Hipoteza 210
Uszeregowanie 210
Zarządzanie eksperymentami 210
Najważniejsze wnioski 211
Dalsza lektura 212

17. Modele LLM i praktyka danologii ... 213
Obecny stan sztucznej inteligencji 213
Czym zajmują się danologowie? 215
Ewolucja opisu stanowiska danologa 217
Studium przypadku: testy A/B 218
Studium przypadku: oczyszczanie danych 219
Studium przypadku: uczenie maszynowe 219
Modele LLM a ta książka 220
Najważniejsze wnioski 221
Dalsza lektura 222

Skorowidz ... 223

ROZDZIAŁ 13.

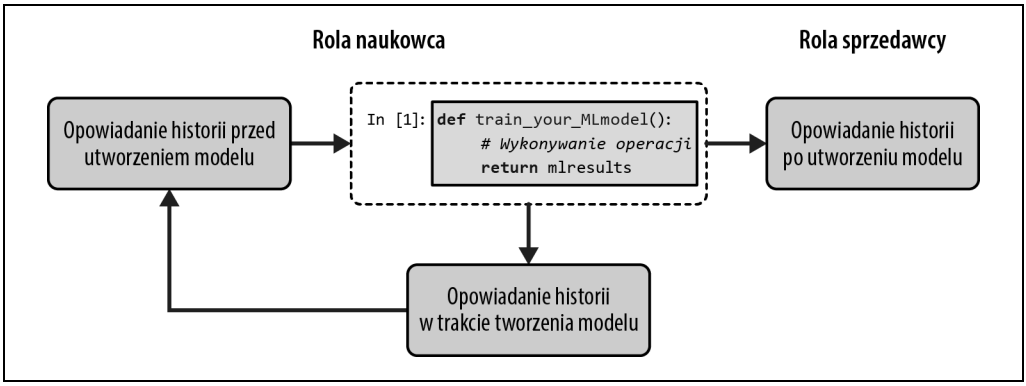
Opowiadanie historii w uczeniu maszynowym

W rozdziale 7. argumentowałem, że danologowie powinni uczyć się opowiadać historie. Jest to prawdą na ogólnym poziomie, ale nabiera szczególnego znaczenia w kontekście uczenia maszynowego.

Niniejszy rozdział przeprowadzi Cię przez główne aspekty opowiadania historii w uczeniu maszynowym, począwszy od inżynierii cech, a kończąc na problemie interpretowalności.

Holistyczne spojrzenie na opowiadanie historii w uczeniu maszynowym

Opowiadanie historii odgrywa dwie powiązane, ale różne role w uczeniu maszynowym (rysunek 13.1). Bardziej znana jest rola sprzedawcy, który musi nawiązać kontakt z odbiorcami, aby uzyskać lub utrzymać poparcie interesariuszy, co zwykle ma miejsce po opracowaniu modelu. Mniej znana jest rola naukowca, który musi znaleźć hipotezy, które poprowadzą go przez cały proces opracowywania modelu.

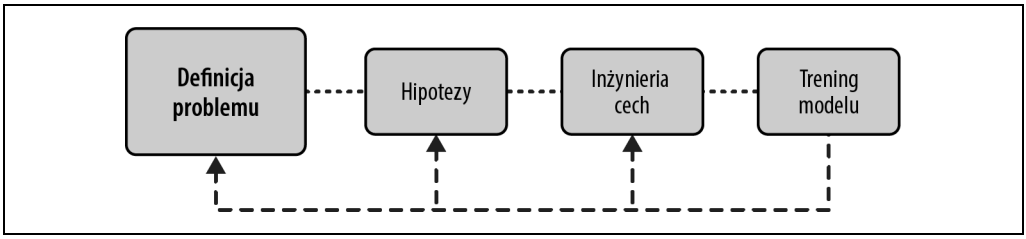


Rysunek 13.1. Opowiadanie historii w uczeniu maszynowym

Ponieważ pierwsze z wymienionych zadań ma miejsce *po* opracowaniu modelu, nazywam je *opowiadaniem historii po utworzeniu modelu*. Rola naukowca jest najczęściej przywoływana przed procesem trenowania modelu i w jego trakcie.

Opowiadanie historii przed opracowaniem modelu i w trakcie tego procesu

Opowiadanie historii przed opracowaniem modelu składa się z czterech głównych etapów: definiowania problemu, tworzenia hipotez, inżynierii cech i trenowania modelu (rysunek 13.2). Choć zazwyczaj są one realizowane w tej właśnie kolejności, między wszystkimi etapami występuje pętla sprzężenia zwrotnego, więc nierzadko po wytrenowaniu pierwszego modelu następuje iteracyjne modyfikowanie cech, hipotez, a nawet samego problemu.



Rysunek 13.2. Opowiadanie historii przed opracowaniem modelu

Pierwszym krokiem zawsze jest zdefiniowanie problemu: *co* chcesz prognozować i *dlaczego*. Lepiej jest to zrobić wcześnie i we współpracy z interesariuszami, aby mieć pewność, że masz ich poparcie, ponieważ wiele obiecujących projektów z dziedziny uczenia maszynowego kończy się niepowodzeniem z powodu jego braku.

Jak zapewne pamiętasz z rozdziału 12., model jest przydatny tylko wtedy, gdy został wdrożony w środowisku produkcyjnym. Takie wdrożenie jest kosztownym przedsięwzięciem nie tylko ze względu na czas i wysiłek, ale także z uwagi na inne projekty, które można było realizować (koszt utraconych korzyści). Z tego powodu zawsze warto zadać sobie pytanie: *czy naprawdę potrzebuję implementacji opartej na uczeniu maszynowym w tym projekcie?* Nie wpadaj w pułapkę stosowania uczenia maszynowego tylko dlatego, że jest to modne lub interesujące. Twoim celem zawsze powinno być wygenerowanie maksymalnej wartości, a uczenie maszynowe to tylko jeszcze jedno narzędzie w przyborniku.

W definicji problemu nie zapomnij też o dobrych odpowiedziach na następujące pytania:

- Jak model będzie używany?
- Jakie są dźwignie, które można wykorzystać dzięki prognozom z modelu?
- W jaki sposób poprawia to możliwości podejmowania decyzji w firmie?

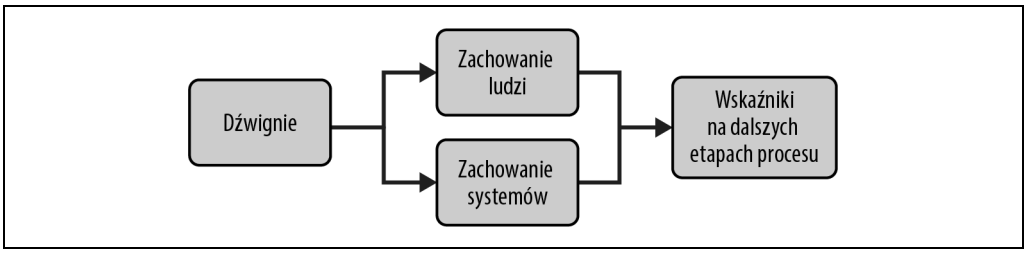
Posiadanie rzetelnych odpowiedzi na te pytania pomoże uzasadnić biznesowo opracowanie modelu opartego na uczeniu maszynowym i zwiększyć tym samym prawdopodobieństwo odniesienia sukcesu.

Zgodnie z ogólnym zaleceniem im szybciej zaangażujesz interesariuszy w definiowanie problemu, tym lepiej. Pomaga to w uzyskaniu akceptacji interesariuszy od samego początku. Upewnij się również, że uczenie maszynowe jest odpowiednim narzędziem dla danego problemu. Wdrażanie, monitorowanie i utrzymywanie modelu jest kosztowne, więc trzeba mieć dla niego dobre uzasadnienie biznesowe.

Tworzenie hipotez

Gdy masz już dobrze zdefiniowany problem, możesz wcielić się w rolę naukowca i zacząć tworzyć hipotezy dla danego problemu. Każda z tych hipotez to opowieść o czynnikach wpływających na prognozy. W tym konkretnym sensie naukowcy są również opowiadaczami.

W tym momencie kluczowe pytania brzmią: *co przewiduję i jakie czynniki wpływają na predykcje?* Rysunek 13.3 przedstawia ogólny przegląd rodzajów problemów związanych z przewidywaniem i ich związek z dostępnymi dźwigniami. Zrozumienie dźwigni ma kluczowe znaczenie dla zapewnienia, że model oparty na uczeniu maszynowym będzie generować wartość (zobacz rozdział 1.).



Rysunek 13.3. Proces obejmujący dźwignie, zachowania i wskaźniki

Z tego wynika, że większość problemów związanych z przewidywaniem należy do którejś z poniższych kategorii:

Wskaźniki, które wynikają z ludzkich zachowań

Często wartość wskaźnika, który Cię interesuje, zależy od tego, czy klienci zachowują się w określony sposób. Na przykład czy użytkownik kliknie baner? Czy kupi produkt po danej cenie? Czy zrezygnuje w przyszłym miesiącu? Ile czasu spędzi w sklepie?

Wskaźniki, które wynikają z zachowania systemów

Wskaźniki zależą również od działania systemów. Jednym z najbardziej znanych przykładów jest optymalizacja centrów danych, a przede wszystkim rozwiązanie problemu chłodzenia powietrzem (*https://oreil.ly/5guWh*). Innym jest przewidywanie czasu wczytywania strony internetowej, który, jak stwierdzono (*https://oreil.ly/xXtbS*), ma bezpośredni wpływ na wskaźniki rezygnacji.

Wskaźniki końcowe

Często zależy nam tylko na zagregowanych wskaźnikach końcowych, na przykład na przychodach. Najczęściej dotyczy to danologów pracujących bezpośrednio w dziedzinie planowania i analizy finansowej.

Wielu danologów ma trudności z tworzeniem i projektowaniem predykcyjnych cech. Ogólnym zaleceniem jest, aby zawsze zaczynać od spisania i przedyskutowania z innymi listy hipotez dotyczących problemu predykcji. Dopiero potem należy przejść do procesu inżynierii cech. Nie zapomnij zapisać powodów, dla których uważasz, że jakaś hipoteza może być słuszna. Tylko z takim uzasadnieniem będziesz w stanie zakwestionować swoją logikę i ulepszyć daną historię.

Oto kilka ogólnych wskazówek pomocnych przy wymyślaniu hipotez dotyczących problemu:

Naprawdę dobrze zapoznaj się z problemem

Niezbyt tajnym sekretem budowania świetnych modeli opartych na uczeniu maszynowym jest posiadanie dużej wiedzy z dziedziny.

Bądź dociekliwy

Jest to jedna z cech definicyjnych, która sprawia, że danolog jest naukowcem.

Kwestionuj status quo

Nie bój się podważać obecnego stanu rzeczy. Obejmuje to kwestionowanie własnych hipotez i iteracyjne wprowadzanie zmian, gdy jest to konieczne (miej świadomość wszelkich oznak efektu potwierdzenia po Twojej stronie).

Po tym wprowadzeniu przejdę do bardziej szczegółowych zaleceń dotyczących sposobu postępowania przy odkrywaniu i formułowaniu hipotez.

Przewidywanie ludzkich zachowań

Gdy przewidujesz ludzkie zachowania, warto zawsze pamiętać, że ludzie robią to, co *chcą* i *mogą*. Możesz chcieć pojechać do Włoch, ale jeśli nie możesz sobie na to pozwolić (finansowo lub czasowo), nie zrobisz tego. Gusta i dostępność zasobów mają podstawowe znaczenie, gdy chcesz przewidywać ludzkie zachowania, a to może zaprowadzić Cię daleko w procesie wymyślania hipotez dla badanego problemu.

Myślenie o motywacji zmusi Cię również do zastanowienia się nad produktem. Na przykład dlaczego ktoś chciałby go kupić? Jaka jest propozycja wartości? Którzy klienci byliby skłonni za to zapłacić?

Podważenie status quo: lekcje z linii frontu

Nie tak dawno temu danolog z jednego z moich zespołów pracowała nad modelem predykcyjnym do sprzedaży krzyżowej stosunkowo nowego produktu. Firma miała trudności ze zbudowaniem jego popularności i uzyskaniem odpowiedniej skali. Był to produkt z relatywnie słabą propozycją wartości, więc zrozumienie, którzy klienci byliby skłonni go używać i za niego zapłacić, było bardzo trudne (i tak samo było z budowaniem modelu do prognozowania tego!).

Pracowałem z nią i po włożeniu dużego wysiłku w zrozumienie produktu, propozycji wartości i klientów wróciliśmy z zaleceniem, że jeśli nie nastąpi poważne przeprojektowanie produktu, nie uzyskamy dopasowania go do rynku. Prawie rok zajęło nam przekonanie interesariuszy, że rzeczywiście tak jest, a w niektórych momentach kilku z nich nie było z nas zadowolonych.

Inną sztuczką jest wykorzystanie umiejętności empatycznego zrozumienia klientów. Zadaj sobie pytanie, co *Ty sam* byś zrobił, gdybyś był na ich miejscu. Oczywiście im łatwiej jest postawić się w ich sytuacji, tym lepiej (mnie naprawdę trudno byłoby wyobrazić sobie siebie jako influencera lub zawodowego boksera). Ta sztuczka może dać Ci dużo korzyści, ale pamiętaj, że być może nie jesteś typowym klientem dla danego produktu. To prowadzi do następnej wskazówki.

Przynajmniej na początku staraj się zrozumieć i modelować *przeciętnego* klienta. Musisz przede wszystkim poprawnie uchwycić efekty pierwszego rzędu, co oznacza, że modelowanie przeciętnej jednostki z analiz zapewnia akceptowalną skuteczność predykcji. Widziałem, jak wielu danologów zaczyna stawiać hipotezy na temat brzegowych lub skrajnych przypadków, które z definicji będą miały znikomy wpływ na ogólną skuteczność predykcji. Przypadki brzegowe są interesujące i ważne, ale na potrzeby prognozowania prawie zawsze lepiej jest zacząć od przypadków przeciętnych.

Przewidywanie zachowania systemu

Niektóre z poprzednich uwag mają zastosowanie również do przewidywania zachowania systemu. Główna różnica polega na tym, że ponieważ systemom brakuje celu lub świadomości, można ograniczyć się do zrozumienia technicznych wąskich gardeł.

Oczywiście musisz opanować szczegóły techniczne systemu, a im większą wiedzę zdobędziesz na temat fizycznych ograniczeń, tym łatwiej będzie Ci wymyślać hipotezy.

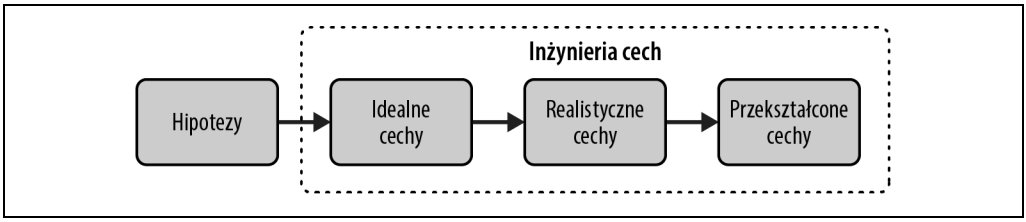
Przewidywanie wskaźników z dalszych etapów

Przewidywanie wskaźników z dalszych etapów jest zarazem trudniejsze i łatwiejsze niż prognozowanie indywidualnych wskaźników wynikających z zachowania człowieka lub systemu. Jest to trudniejsze, ponieważ im bardziej wskaźnik jest oderwany od podstawowych czynników, tym słabsze i mniej precyzyjne stają się hipotezy. Co więcej, przekłada się to na trudność wymyślania historii na temat tych czynników. Niektóre z tych problemów mogą się kumulować i prowadzić do komplikacji na wyższym poziomie.

Mimo to nieraz można zignorować niektóre z tych problemów i wykorzystać korelacje czasowe i przestrzenne do utworzenia cech. W pewnym sensie akceptujesz, że wszelkie historie, które wymyślisz, dadzą gorsze wyniki niż prosta struktura autoregresyjna powszechna w szeregach czasowych i w przestrzennych modelach autoregresyjnych.

Inżynieria cech

Na ogólnym poziomie proces inżynierii cech obejmuje przekształcanie hipotez w mierzalne zmienne, które zapewniają wystarczająco mocny sygnał, aby pomóc algorytmowi nauczyć się procesu generowania danych. Dobrą praktyką jest podzielenie tego procesu na kilka etapów (rysunek 13.4).



Rysunek 13.4. Proces inżynierii cech

Etapy inżynierii cech są następujące:

Utworzenie zestawu idealnych cech

Pierwszy krok wymaga przekształcenia hipotez na *idealne* cechy, które zastosowałbyś, gdyby możliwe było dokładne zmierzenie wszystkich czynników. Ten krok jest ważny, ponieważ pozwala wyznaczyć punkt odniesienia dla drugiego etapu.

Przykładem jest rola, jaką *intencjonalność* odgrywa we *wczesnej rezygnacji* definiowanej jako współczynnik klientów, którzy raz wypróbowują produkt i rezygnują. Jedna z hipotez dotyczy tego, że tacy klienci w rzeczywistości nie zamierzali korzystać z produktu (ponieważ chcieli go tylko wypróbować, zostali nakłonieni do zakupu (*https://oreil.ly/HDGj-*), miało miejsce oszustwo itp.). Czy nie byłoby wspaniale, gdybyś mógł ich o to zapytać, a oni odpowiedzieliby zgodnie z prawdą? Niestety nie jest to ani praktyczne, ani osiągalne.

Przybliżenie idealnych cech za pomocą realistycznych cech

Jeśli zdasz sobie sprawę, że idealny zestaw cech jest niedostępny, musisz znaleźć dobre cechy zastępcze, czyli takie, które są skorelowane z idealnymi. Często stopień korelacji może być bardzo niski i trzeba zadowolić się włączeniem zmiennych kontrolnych o bardzo słabej zgodności z pierwotną hipotezą.

Przykładem tej ostatniej sytuacji jest to, jak kultura wpływa na gusta, a tym samym na prawdopodobieństwo zakupu produktu. Na przykład mogą istnieć różnice kulturowe, które wyjaśniają, dlaczego użytkownicy w poszczególnych krajach decydują się akceptować lub odrzucać pliki cookie w przeglądarce (osoby z niektórych krajów mogą być bardziej niechętne do dzielenia się tymi informacjami). Nie trzeba dodawać, że pomiar kultury jest trudny. Ale jeśli podejrzewasz, że zmienność na poziomie kraju pozwoli uchwycić dużą część zmienności związanej z hipotezą kulturową, musisz jedynie włączyć zmienne zastępcze reprezentujące kraj. Jest to stosunkowo słaby zestaw cech, ponieważ będą one reprezentować dowolną cechę na tym poziomie, a nie tylko kulturę (na przykład różnice w przepisach prawnych).

Przekształcanie cech

Jest to proces wydobywania maksymalnej ilości sygnału z cech przez zastosowanie do nich zestawu przekształceń. Zauważ, że odchodzę w tym miejscu nieco od literatury, ponieważ większość podręczników na temat inżynierii cech dotyczy wyłącznie tego etapu.

Etap ten obejmuje transformacje takie jak skalowanie (*https://oreil.ly/Hak0v*), binaryzacja i kodowanie 1 z n (*https://oreil.ly/ralbT*), uzupełnianie brakujących wartości (*https://oreil.ly/MhGuK*), interakcje cech (*https://oreil.ly/bT-1q*) i tym podobne. Na końcu tego rozdziału podaję kilka źródeł, w których można zapoznać się z szeroką gamą dostępnych transformacji.

Co ważne, transformacje zależą od danych *i* wybranego algorytmu. Na przykład drzewa klasyfikacyjne i regresyjne mogą nie wymagać samodzielnego zajmowania się wartościami odstającymi, ponieważ algorytm zrobi to automatycznie. Podobnie w algorytmach nieliniowych, takich jak drzewa i zespoły oparte na drzewach, nie trzeba uwzględniać interakcji multiplikatywnych.

Przykład: przewidywanie sprzedaży

Załóżmy, że chcesz przewidywać sprzedaż w obszarach geograficznych (g). Typowym zastosowaniem takiego modelu jest kierowanie zasobów sprzedażowych do lokalizacji o najwyższym potencjale sprzedaży, zgodnie z przewidywaniami modelu.

Użyję sztuczki z rozdziału 2., aby uzyskać bardziej precyzyjne historie:

$$\text{sprzedaż}_g = \text{TAM}_g \times \frac{\text{sprzedaż}_g}{\text{TAM}_g} = \text{TAM}_g \times \text{Prawd. (sprz. jednostkowa w } g)$$

Ten wzór oznacza, że sprzedaż całkowita w obszarze g jest równa całkowitej wielkości rynku docelowego (ang. *total addressable market* — TAM) pomnożonej przez prawdopodobieństwo sprzedaży w tym obszarze.

W ten sposób, zamiast wymyślać hipotezy dotyczące liczby transakcji sprzedaży w poszczególnych lokalizacjach, mogę teraz skupić się na historiach, które pomogą mi przewidzieć TAM, *a także* historiach wyjaśniających, dlaczego firma dokonuje sprzedaży. To drugie wiąże się z ludzkim zachowaniem, a pierwsze jest zagregowanym wskaźnikiem.

Aby modelować TAM, muszę najpierw zrozumieć, kto jest moim docelowym klientem, a następnie znaleźć historie o tym, co sprawia, że liczba takich osób jest większa w określonych lokalizacjach. Na przykład aby przewidzieć TAM dla tej książki, chcę oszacować liczbę danologów w danym regionie. Jedną z prawdopodobnych historii jest to, że danologowie są tam, gdzie firmy ich potrzebują. Mogę dalej doprecyzować tę historię przez przyjęcie różnych założeń: że wielkość firmy ma znaczenie (*z powodu* ilości danych potrzebnych do uzasadnienia zatrudnienia danologa, ale także dlatego, że danologowie są stosunkowo drodzy i tylko wystarczająco duże firmy mogą ich zatrudnić), że branże, w jakich działają te firmy, mają znaczenie (*ponieważ* bardziej kapitałochłonne sektory mogą mieć więcej zautomatyzowanych danych generowanych przez systemy niż branże z większą intensywnością pracy z ręcznymi procesami, *z powodu* nacisków regulacyjnych, czy to *z powodu* różnic w koncentracji rynku), a także że wielkość populacji i rozkład wieku mają znaczenie (*ponieważ* dziedzina ta jest stosunkowo nowa, a młodsi ludzie, ale nie zbyt młodzi, są bardziej skłonni do inwestowania w naukę trudnego przedmiotu technicznego, takiego jak danologia). Te hipotezy wskazują, jakiego rodzaju danych muszę szukać, aby rozwiązać omawiany problem predykcji.

Gdy modelujesz prawdopodobieństwo sprzedaży, trzeba pamiętać, że muszą istnieć ludzie, którzy chcą i mogą sobie pozwolić na produkt (popyt), a towar musi być dostępny dla nich w odpowiednich lokalizacjach (podaż). Idealnymi cechami do modelowania popytu są preferencje konsumentów dotyczące produktu, a także dochód gospodarstwa domowego. Preferencje są zazwyczaj trudne do uzyskania, ale można je przybliżyć za pomocą wcześniejszej sprzedaży firmy w danej lokalizacji lub na podstawie zachowań związanych z wyszukiwaniem w internecie (odzwierciedlanych na przykład w trendach Google lub w dostępnych danych od podobnych sprzedawców). Dane dotyczące podaży są łatwiejsze do uzyskania, ponieważ zwykle wiadomo, czy firma i jej konkurenci są obecni w różnych lokalizacjach.

Opowiadanie historii po opracowaniu modelu: otwieranie czarnej skrzynki

Problem z opowiadaniem historii po opracowaniu modelu związany jest głównie ze zrozumieniem, dlaczego model tworzy prognozy w taki sposób, w jaki to robi, które cechy są najbardziej predykcyjne i w jaki sposób są one skorelowane z prognozami. Dwa główne punkty, które chcesz przekazać swoim odbiorcom, to:

- Model zwiększa jakość predykcji, co oznacza, że błąd predykcji jest niższy niż dla podstawowego podejścia.
- Model *ma sens*. Dobrą praktyką jest rozpoczęcie od omówienia hipotez, sposobu ich modelowania i ich zgodności z wynikami.

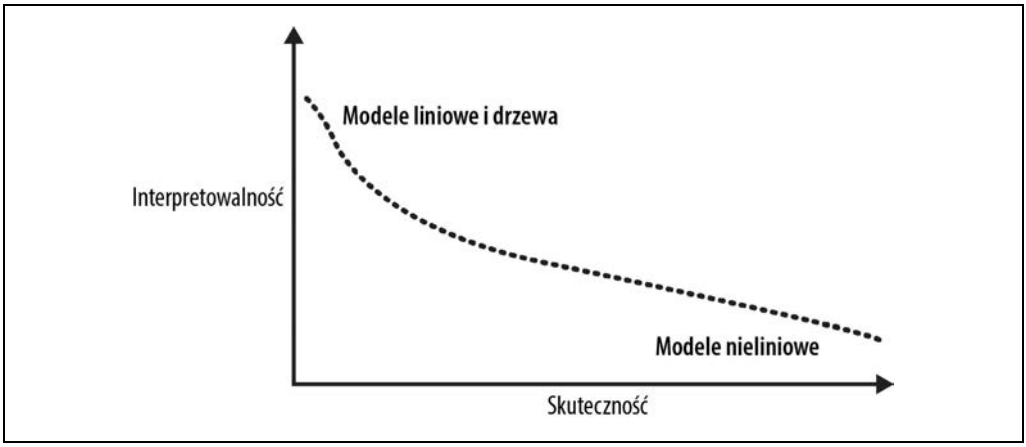
Na ogólnym poziomie model jest *interpretowalny*, jeśli można zrozumieć, z czego wynikają predykcje. *Lokalna* interpretowalność ma na celu zrozumienie konkretnych przewidywań, na przykład dlaczego uważa się za wysoce prawdopodobne, że klient nie spłaci kredytu. *Globalna* interpretowalność ma na celu zapewnienie ogólnego zrozumienia, w jaki sposób cechy wpływają na wynik. Temat ten zasługuje na odrębną książkę, ale w tym rozdziale mogę przyjrzeć się wyłącznie praktycznym kwestiom, a konkretnie omówię tylko metody osiągania globalnej interpretowalności, ponieważ uważam je za najbardziej przydatne do celów opowiadania historii.

Przed otwarciem czarnej skrzynki upewnij się, że model ma wystarczającą skuteczność predykcji i że nie występuje wyciekanie danych. Opowiadanie historii po opracowaniu modelu wymaga dużo czasu i wysiłku, więc lepiej zacząć od dobrego modelu predykcyjnego.

Ponadto gdy prezentujesz wskaźniki skuteczności, staraj się, aby były one jak najbardziej zrozumiałe dla odbiorców. Typowe wskaźniki, takie jak błąd średniokwadratowy (RMSE) lub obszar pod krzywą (AUC), mogą być zagadkowe dla interesariuszy biznesowych. Zwykle warto włożyć dodatkową pracę w przetłumaczenie ich na precyzyjne skutki biznesowe. Na przykład w jaki sposób spadek błędu średniokwadratowego o 5% przełoży się na poprawę działalności biznesowej?

Kompromis między interpretowalnością a skutecznością

Można argumentować, że idealny algorytm uczenia maszynowego powinien być zarówno wydajny, jak i interpretowalny. Niestety zwykle niezbędny jest kompromis między interpretowalnością a skutecznością predykcji, więc musisz częściowo zrezygnować ze zrozumienia tego, co dzieje się wewnątrz algorytmu, jeśli chcesz osiągnąć niższy błąd predykcji (rysunek 13.5).



Rysunek 13.5. Kompromis między interpretowalnością a skutecznością

Po jednej stronie spektrum znajdują się modele liniowe, które są ogólnie uważane za wysoce interpretowalne, ale mają niższą skuteczność predykcji. Zestaw ten obejmuje regresję liniową i logistyczną, a także nieliniowe algorytmy uczenia się, takie jak drzewa klasyfikacyjne i regresyjne. Po drugiej stronie spektrum znajdują się bardziej elastyczne i zazwyczaj wysoce nieliniowe modele, takie jak głębokie sieci neuronowe, zespoły oparte na drzewach i maszyny wektorów nośnych. Algorytmy te są ogólnie znane jako *czarnoskrzynkowe* mechanizmy uczenia się. Celem jest otwarcie czarnej skrzynki i lepsze zrozumienie tego, co się w niej dzieje.

Zanim przejdziemy dalej, warto wspomnieć, że nie jest oczywiste, iż trzeba interpretować wyniki, więc krótko omówię, dlaczego warto to robić:

Wdrażanie i pozyskanie poparcia

Wiele osób musi najpierw zrozumieć, dlaczego generowane są dane prognozy, aby zaakceptować je jako ważne i przyjąć. Najczęściej zdarza się to w organizacjach, które nie są przyzwyczajone do podejścia opartego na uczeniu maszynowym. W takich organizacjach decyzje są zwykle podejmowane z wykorzystaniem podejścia pozornie opartego na danych, a w rzeczywistości często na podstawie intuicji. Otwarcie czarnej skrzynki może ułatwić interesariuszom zaakceptowanie wyników i przekonać ich do sfinansowania projektu.

Niska skuteczność predykcji w praktyce

Otwarcie czarnej skrzynki jest jednym z najskuteczniejszych sposobów wykrywania i korygowania problemów, takich jak wyciekanie danych (zobacz rozdział 11.).

Etyka i wymogi regulacyjne

W niektórych branżach wymagane jest, aby firmy wyjaśniały, dlaczego podjęto określone decyzje. Na przykład w Stanach Zjednoczonych (*https://oreil.ly/5zj9j*) ustawa o zapobieganiu dyskryminacji (ang. *Equal Opportunity Act*) uprawnia każdego do zapytania o powody odmowy udzielenia kredytu. Podobnie jest z europejskim ogólnym rozporządzeniem o ochronie danych osobowych (RODO). Nawet jeśli nie masz tego rodzaju obowiązku informacyjnego, możesz chcieć sprawdzić, czy prognozy i późniejsze decyzje są zgodne z minimalnymi standardami etycznymi, i w tym celu otworzyć czarną skrzynkę.

Regresja liniowa: ustalenie punktu odniesienia

Regresja liniowa stanowi przydatny punkt odniesienia do zrozumienia interpretacji (zobacz także rozdział 10.). Rozważ następujący prosty model:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \epsilon$$

Jeśli przyjmiesz silne założenia o liniowości podstawowego procesu generowania danych, natychmiast uzyskasz:

Kierunkowość efektu

Znak każdego współczynnika informuje, czy cecha jest dodatnio, czy ujemnie powiązana z wynikiem (przy kontrolowaniu wszystkich innych cech).

Wielkość efektu

Każdy współczynnik jest interpretowany jako zmiana wyniku związana ze zmianą każdej cechy o jedną jednostkę przy utrzymaniu innych cech na stałym poziomie. Co ważne, bez przyjęcia dalszych założeń nie można podać interpretacji przyczynowej.

Interpretowalność lokalna

Na podstawie dwóch pierwszych elementów można stwierdzić, dlaczego wygenerowano konkretną prognozę.

Niektórzy danologowie popełniają błąd, ponieważ nadają bezwzględnej wielkości współczynników interpretację względnego *znaczenia*. Aby zobaczyć, dlaczego jest to niewłaściwe, przyjrzyj się następującemu modelowi, w którym przychody są wyrażone jako funkcja wielkości zespołu sprzedażowego i płatnych wydatków na marketing (a konkretnie na marketing w wyszukiwarkach):

$$\text{przychody} = 100 + 1000 \cdot \text{kierownicy sprzedaży} + 0{,}5 \cdot \text{wydatki na marketing w wyszukiwarkach}$$

Oznacza to, że średnio i przy utrzymaniu innych czynników na stałym poziomie każdy dodatkowy:

- kierownik sprzedaży daje wzrost przychodów o 1000 dolarów;
- dolar wydany na marketing w wyszukiwarkach (na przykład na oferty na powierzchnię reklamową w witrynach Google, Bing lub Facebook) daje wzrost przychodów o 50 centów.

Można by pokusić się o stwierdzenie, że zwiększenie liczby kierowników sprzedaży jest *ważniejsze* dla przychodów w porównaniu z wydatkami na marketing. Niestety jest to porównanie jabłek do pomarańczy, ponieważ każda cecha jest mierzona w innych jednostkach. Sztuczka polegająca na mierzeniu wszystkiego w tych samych jednostkach wymaga przeprowadzenia regresji na znormalizowanych cechach:

$$y = \beta_0 + \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \eta$$

gdzie dla każdej zmiennej z mamy $\tilde{z} = \frac{z - \text{średnia}\ (z)}{\text{odch.stand.}\ (z)}$.

Należy pamiętać, że współczynniki regresji dla zmiennych standaryzowanych są zazwyczaj różne od tych z oryginalnego modelu (stąd inne greckie litery), a zatem mają inną interpretację. Gdy stosujesz standaryzację wszystkich cech, mierzysz wszystko w jednostkach odchyleń standardowych (lepszym terminem byłoby *bezjednostkowo*), dzięki czemu porównujesz jabłka z jabłkami.

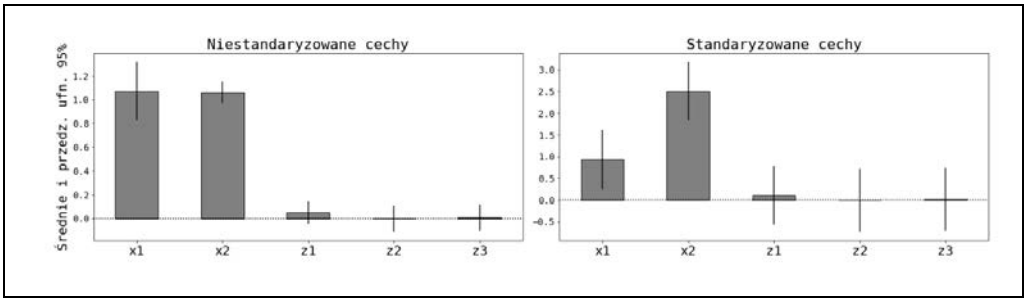
Można więc powiedzieć, że x_1 *jest ważniejsze niż* x_2, *ponieważ dodatkowe odchylenie standardowe* x_1 *zwiększa przychody o więcej niż analogiczny wzrost* x_2.

Sztuczka polega na znalezieniu sposobu na konwersję oryginalnych jednostek na wspólną jednostkę. W tym przykładzie używam odchylenia standardowego, ale odpowiednia może być też każda inna wspólna jednostka. Wyobraźmy sobie na przykład, że każdy dodatkowy kierownik sprzedaży kosztuje średnio 5000 dolarów miesięcznie. Ponieważ wydatki na marketing są już wyrażone w dolarach, można powiedzieć, że średnio każdy dodatkowy dolar wydany na:

- kierowników sprzedaży wiąże się ze wzrostem przychodów o 20 centów;
- płatny marketing wiąże się ze wzrostem przychodów o 50 centów.

Chociaż ta ostatnia metoda również działa, standaryzacja jest znacznie bardziej powszechną metodą określania wspólnej jednostki dla wszystkich cech. Ważną rzeczą do zapamiętania jest to, że jesteś teraz w stanie *uszeregować* cechy w sensowny sposób.

Rysunek 13.6 przedstawia wykres oszacowanych współczynników wraz z przedziałami ufności 95% dla symulowanego modelu liniowego z dwiema cechami o zerowej średniej i rozkładzie normalnym (x_1, x_2), tak jak w poprzednich równaniach. Cechy z_1, z_2, z_3 są dodatkowymi zmiennymi skorelowanymi z x_2, ale poza tym nie są powiązane z wynikiem. Co ważne, ustawiłem prawdziwe parametry na $\alpha_1 = \alpha_2 = 1$ i $\text{War}(x_1) = 1$, $\text{War}(x_2) = 5$.



Rysunek 13.6. Regresja liniowa dla niestandaryzowanych i standaryzowanych cech

Ma to dwa skutki:

- zwiększa stosunek sygnału do szumu dla drugiej cechy, przez co staje się ona bardziej informatywna;
- zwiększa prawdziwy współczynnik[1]: $\beta_2 = \sqrt{5}\alpha_2$.

W wyniku standaryzacji obu cech okazuje się, że druga cecha jest ważniejsza od pierwszej, co zdefiniowałem wcześniej. Dzięki przedziałom ufności można również stwierdzić, że ostatnie trzy cechy są nieinformatywne. Zamiast podejścia statystycznego można też zastosować *regularyzację*, tak jak w regresji lasso.

[1] Łatwo jest wykazać, że w regresji liniowej przeskalowanie cechy x do kx zmienia prawdziwy współczynnik z α na α/k.

Znaczenie cech

Często cechy są szeregowane według jakiejś obiektywnej miary ważności. Jest to przydatne do celów opowiadania historii przed opracowaniem modelu i po jego otrzymaniu. Jeśli chodzi o opowiadanie historii po uzyskaniu modelu, można powiedzieć na przykład, że *odkryliśmy, iż czas transakcji jest najważniejszym predyktorem oszustwa*. Może to pomóc w wykazaniu przydatności wyników, a także doprowadzić do potencjalnie cennych momentów olśnienia zarówno dla Ciebie, jak i dla odbiorców (zobacz także rozdział 7.).

Z kolei przed opracowaniem modelu posiadanie sposobu na szeregowanie cech według ich ważności może pomóc w iteracyjnym rozwijaniu hipotez, w inżynierii cech lub też w lepszym zrozumieniu problemu. Jeżeli masz dobrze przemyślane hipotezy, a wyniki wyglądają podejrzanie, bardziej prawdopodobne jest, że popełniłeś błąd programistyczny na etapie inżynierii cech lub że nastąpiło wyciekanie danych.

Wcześniej użyłem ustandaryzowanych cech w regresji liniowej, aby uzyskać jedno z możliwych uporządkowań według ważności:

Ważność standaryzowanych cech w regresji liniowej

> Cecha *x* jest ważniejsza niż cecha *z*, jeśli wzrost *x* o jedno odchylenie standardowe jest związany z większą zmianą wyniku (mierzoną wartością bezwzględnej).

Ważność można też zdefiniować w kategoriach ilości informacji, jaką każda cecha wnosi w danym problemie predykcji. Intuicyjnie im wyższa zawartość informacyjna cechy (dla danego wyniku), tym niższy błąd predykcji, jeśli ta cecha zostanie uwzględniona. Istnieją dwa powszechnie stosowane wskaźniki tego rodzaju:

Znaczenie cechy ze względu na niejednorodność

> Cecha *x* jest ważniejsza niż cecha *z* ze względu na niejednorodność węzła, jeśli względny spadek błędu predykcji dla węzłów, w których *x* została wybrana jako zmienna dzieląca, jest większy niż analogiczny spadek dla *z*.

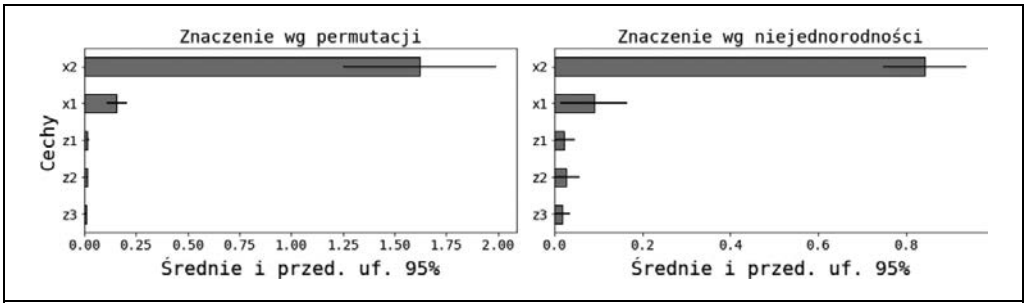
Znaczenie cechy ze względu na permutacje

> Cecha *x* jest ważniejsza niż cecha *z* ze względu na permutacje, jeśli względna utrata skuteczności, gdy wartości *x* są poddawane permutacji, jest większa niż dla *z*.

Należy pamiętać, że ważność cech ze względu na niejednorodność (*https://oreil.ly/acJDH*) działa tylko dla algorytmów uczenia maszynowego opartych na drzewach. Za każdym razem gdy węzeł jest dzielony na podstawie danej cechy, zapisywana jest poprawa skuteczności. Dzięki temu na końcu można obliczyć udział poszczególnych cech w poprawie w stosunku do całkowitej poprawy.

Z kolei ważność ze względu na permutacje (*https://oreil.ly/84XXY*) działa w każdym algorytmie uczenia maszynowego, ponieważ wystarczy przestawić wartości każdej cechy (kilka razy, jak w procedurze bootstrappingu) i obliczyć spadek skuteczności. Intuicja podpowiada, że kolejność wartości ma większe znaczenie dla *ważnych* cech, więc permutacja tych wartości powinna powodować dla nich większy spadek skuteczności.

Rysunek 13.7 przedstawia znaczenie cech ze względu na permutacje i niejednorodność dla tego samego symulowanego zbioru danych co poprzednio, wytrenowanego za pomocą regresji ze wzmacnianiem gradientowym (bez optymalizacji metaparametrów), wraz z przedziałami ufności 95%. Przedziały ufności dla znaczenia ze względu na permutacje są ustalane parametrycznie (przy założeniu rozkładu normalnego) z użyciem średnich i odchyleń standardowych obliczanych za pomocą pakietu `scikit-learn`. Analogiczne przedziały dla znaczenia ze względu na niejednorodność uzyskuję za pomocą bootstrappingu (zobacz rozdział 9.).



Rysunek 13.7. Znaczenie cech w symulowanym modelu opartym na regresji ze wzmacnianiem gradientowym

Mapa cieplna

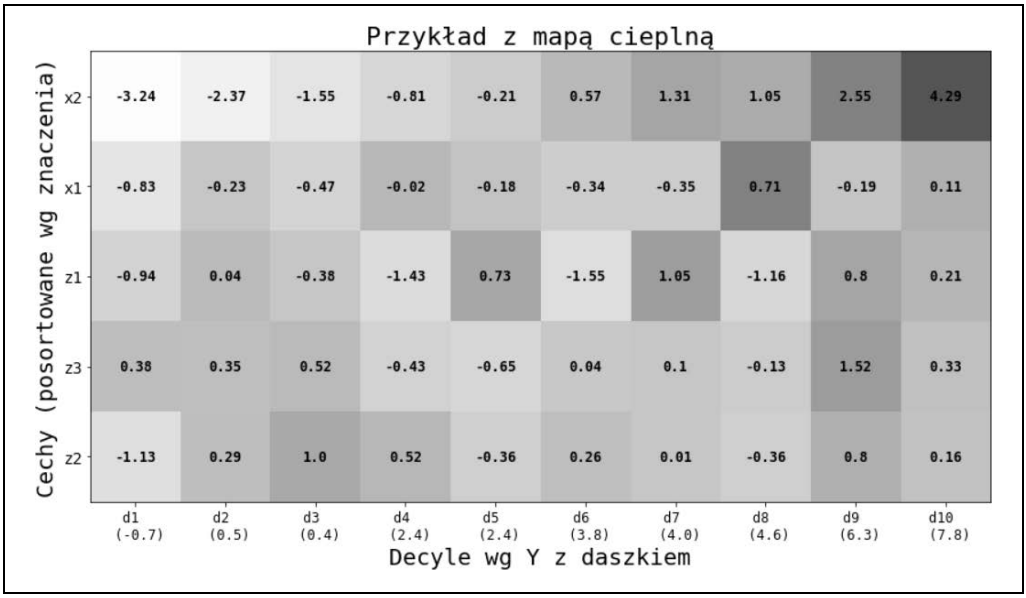
Mapy cieplne są bardzo łatwe do obliczenia i zwykle dobrze się sprawdzają do wizualnego przedstawiania korelacji między każdą cechą a przewidywanym wynikiem. Jest to bardzo przydatne do przekazywania informacji typu: *gdy x rośnie, y spada*. Wiele hipotez ma charakter kierunkowy, więc szybki pierwszy test sprawdzający, czy dana hipoteza znajduje potwierdzenie w praktyce, jest całkiem przydatny. Proces generowania takich map wygląda następująco:

1. Podziel przewidywane wyniki (za pomocą regresji) lub wartości prawdopodobieństwa (za pomocą klasyfikacji) na decyle lub dowolne inne kwantyle.
2. Dla każdej cechy x_j i decyla d oblicz średnią dla wszystkich jednostek z danego kubełka: $\bar{x}_{j,d}$.

Otrzymane wartości można umieścić w tabeli z decylami w kolumnach i cechami w wierszach. Zwykle dobrze jest uporządkować cechy za pomocą jakiejś miary ważności, aby najpierw skupić się na najbardziej istotnych cechach.

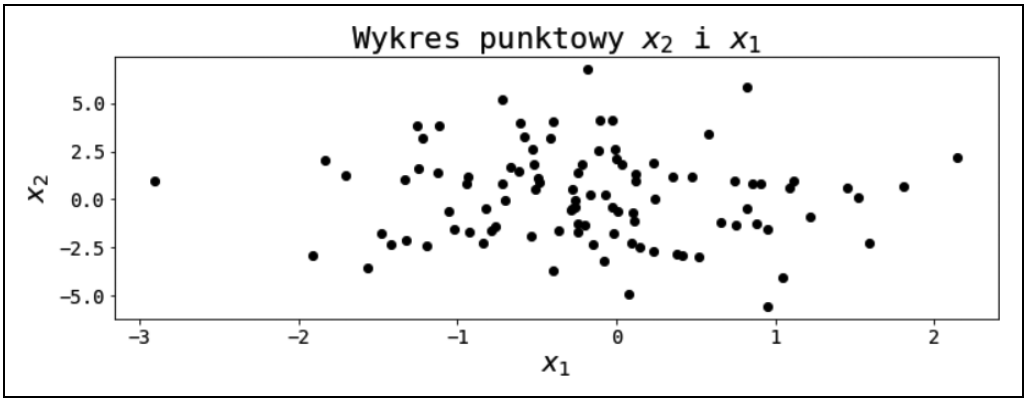
Rysunek 13.8 przedstawia mapę cieplną dla regresji liniowej wytrenowanej na opisanych wcześniej zasymulowanych danych, w których cechy zostały już posortowane według ważności. Samo sprawdzenie względnych odcieni dla każdej cechy (wiersza) pozwala łatwo zidentyfikować wzorce lub ich brak.

Na przykład cecha x_2 jest dodatnio skorelowana z wynikiem, co jest zgodne z oczekiwaniami, ponieważ prawdziwy współczynnik w symulacji jest równy 1. W dolnym decylu średnia wartość to –3,58 jednostki i rośnie ona monotonicznie do średnio 4,23 jednostki w górnym decylu.



Rysunek 13.8. Mapa cieplna cech dla wcześniejszego przykładu z symulacją

Sprawdzenie wiersza z x_1 ujawnia główny problem map cieplnych: przedstawiają one tylko korelacje dwuwymiarowe. Prawdziwa korelacja jest dodatnia ($\alpha_1 = 1$), ale mapa cieplna nie obrazuje tej monotonicznej zależności. Aby zrozumieć, dlaczego tak się dzieje, należy zauważyć, że x_1 i x_2 są *ujemnie* skorelowane (rysunek 13.9). Jednak większa wariancja drugiej cechy daje jej większą moc predykcyjną, a tym samym większą wagę w ostatecznym uporządkowaniu cech ze względu na prognozowanie wyników (i decyle). Te dwa fakty naruszają zależność monotoniczną, której oczekiwano dla drugiej cechy



Rysunek 13.9. Cechy x_2 i x_1 są ujemnie skorelowane

Wykresy zależności częściowych

Za pomocą wykresów zależności częściowych można przewidzieć wynik lub prawdopodobieństwo przez zmianę tylko jednej cechy w danym momencie przy zachowaniu stałych wszystkich pozostałych wartości. Jest to dość atrakcyjne podejście, ponieważ jest podobne do tego, co można uzyskać za pomocą pochodnych cząstkowych w regresji liniowej.

W rozdziale 9. zastosowałem następującą metodę generowania wykresów zależności częściowych, która bardzo dobrze oddaje to intuicyjne podejście. Najpierw należy obliczyć średnie dla wszystkich cech, a następnie utworzyć siatkę liniową o rozmiarze *G* dla cechy, którą chcesz zasymulować. Wszystkie te dane należy zapisać w macierzy o następującej postaci:

$$\overline{\mathbf{X}}_{\mathbf{j}} = \begin{pmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & x_{0j} & \cdots & \bar{x}_K \\ \bar{x}_1 & \bar{x}_2 & \cdots & x_{1j} & \cdots & \bar{x}_K \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ \bar{x}_1 & \bar{x}_2 & \cdots & x_{Gj} & \cdots & \bar{x}_K \end{pmatrix}_{G \times K}$$

Następnie należy użyć tej macierzy do utworzenia prognozy za pomocą wytrenowanego modelu:

$$\text{wykre}\square\text{zależności częściowych}^{(1)}\left(x_j\right) = \hat{f}\left(\overline{\mathbf{X}}_j\right)$$

Metoda ta jest szybka i intuicyjnie atrakcyjna, a także pozwala sprawnie zasymulować wpływ interakcji między cechami. Jednak ze statystycznego punktu widzenia nie jest ona do końca poprawna, ponieważ średnia funkcji różni się od funkcji wykonywanej dla średnich danych wejściowych (chyba że model jest liniowy). Główną zaletą tego podejścia jest to, że wymaga tylko jednej oceny wytrenowanego modelu.

Prawidłowy sposób — i jest to metoda używana w pakiecie `scikit-learn` (*https://oreil.ly/waddK*) do generowania wykresów zależności częściowych — wymaga *N* (wielkość próby) ocen wytrenowanego modelu dla każdej wartości *g* w siatce. Są one następnie uśredniane w celu uzyskania:

$$\text{wykre}\square\text{zależności częściowych}^{(2)}\left(x_j = g\right) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}\left(x_{1,i}, \cdots, x_{j-1,i}, g, x_{j+1,i}, \cdots, x_{K,i}\right)$$

Interakcje można łatwo zasymulować przez zmianę kilku cech jednocześnie. W praktyce obie metody często dają podobne wyniki, ale tak naprawdę zależy to od rozkładu cech i rzeczywistego nieobserwowanego procesu generowania danych.

Zanim przejdę dalej, warto zauważyć, że w tych ostatnich obliczeniach należy uzyskać prognozę dla każdego wiersza ze zbioru danych. Na *wykresach ICE* (ang. *individual conditional expectation*) można wizualnie wyświetlić te efekty dla różnych jednostek, co sprawia, że jest to metoda lokalnej interpretacji (w odróżnieniu od wykresów zależności częściowych)[2].

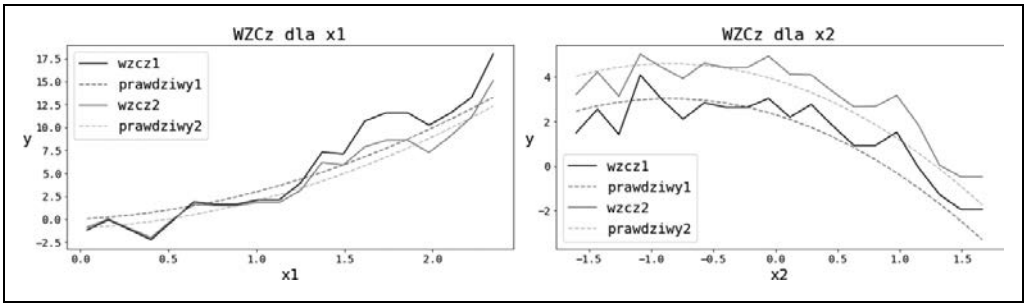
Przeprowadzę teraz symulację dla modelu nieliniowego, aby zobaczyć działanie obu opisanych podejść w praktyce. Wykorzystam następujący proces generowania danych:

[2] W implementacji z repozytorium kodu (*https://oreil.ly/dshp-repo*) generowane są zarówno wykresy ICE, jak i wykresy zależności częściowych.

$$y = x_1 + 2x_1^2 - 2x_1x_2 - x_2^2 + \epsilon$$
$$x_1 \sim gamma\ (\text{shape } = 1, \text{scale} = 1)$$
$$x_2 \sim N(0,1)$$
$$\epsilon \sim N(0,5)$$

Dla pierwszej cechy używam rozkładu gamma, aby podkreślić wpływ, jaki mogą mieć wartości odstające, gdy używasz którejś z omawianych metod.

Rysunek 13.10 przedstawia szacowane i rzeczywiste wykresy zależności częściowych dla obu metod. Wykresy zależności częściowych dla pierwszej cechy dobrze oddają kształt prawdziwej zależności, ale dla większych wartości x_1 te dwie metody zaczynają się od siebie różnić. Jest to zgodne z oczekiwaniami, ponieważ średnia z próby jest wrażliwa na wartości odstające, więc w pierwszej metodzie używana jest uśredniona jednostka ze stosunkowo dużą wartością pierwszej cechy. W drugiej metodzie nie jest to aż tak odczuwalne, ponieważ uśredniane są indywidualne prognozy, a w tym konkretnym przykładzie postać funkcjonalna wygładza efekt wartości odstających.



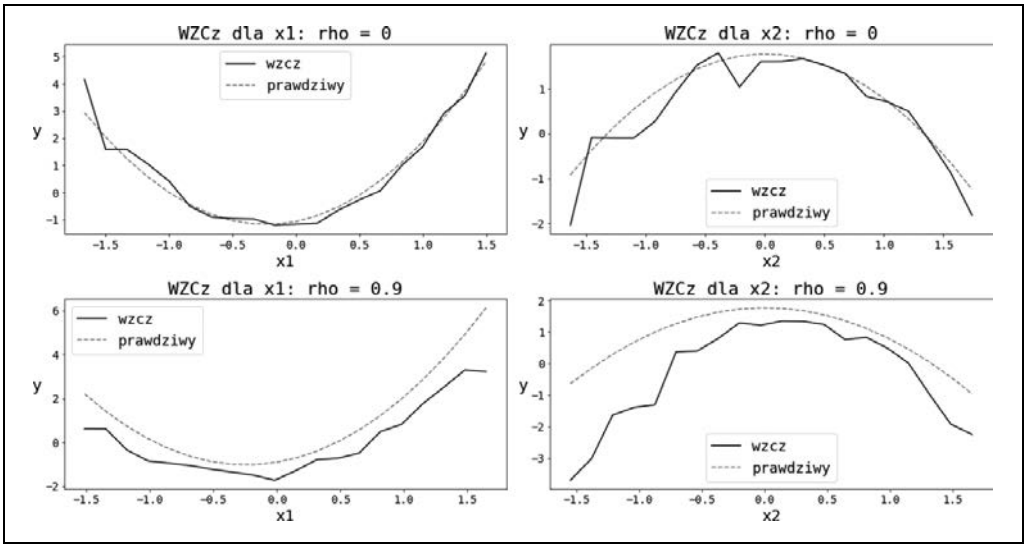
Rysunek 13.10. Wykresy zależności częściowych w obu metodach dla symulowanych danych

Chociaż wykresy zależności częściowych są świetne, są obciążone błędem systematycznym, gdy cechy są skorelowane. Na przykład jeśli x_1 i x_2 są dodatnio skorelowane, zwykle obie będą miały albo małe, albo duże wartości. Kiedy stosujesz wykresy zależności częściowych, możesz otrzymać nierealistyczne wyniki, jeśli zastosujesz małą wartość (z siatki) dla x_1, gdy powiązana wartość drugiej cechy jest duża.

Aby pokazać to w praktyce, użyłem w symulacji zmodyfikowanej wersji poprzedniego modelu nieliniowego:

$$y = x_1 + 2x_1^2 - 2x_1x_2 - x_2^2 + \epsilon$$
$$x_1, x_2 \sim N(\mathbf{0}, \Sigma(\rho))$$
$$\epsilon \sim N(0,5)$$

Cechy są teraz losowane z wielowymiarowego rozkładu normalnego z macierzą kowariancji indeksowaną parametrem korelacji. Rysunek 13.11 przedstawia szacowane i prawdziwe wykresy zależności częściowych dla nieskorelowanych ($\rho = 0$) i skorelowanych ($\rho = 0{,}9$) cech. Można łatwo zweryfikować, że gdy cechy są skorelowane, wykresy zależności częściowych są obciążone błędem systematycznym.



Rysunek 13.11. Wykresy zależności częściowych dla skorelowanych i nieskorelowanych cech

Skumulowane efekty lokalne

Skumulowane efekty lokalne (ang. *accumulated local effect* — ALE) to stosunkowo nowa metoda, która eliminuje wady wykresów zależności częściowych dla skorelowanych cech. Jest również mniej kosztowna obliczeniowo, ponieważ liczba wywołań wytrenowanej funkcji jest mniejsza[3].

Jak wspomniałem wcześniej, problem z wykresami zależności częściowych wynika z wymuszania nierealistycznych wartości cechy, jeśli wziąć pod uwagę jej korelację z innymi cechami. W efekcie szacunki są obciążone błędem systematycznym. Tak jak poprzednio należy rozpocząć od utworzenia siatki dla dowolnej analizowanej cechy k. W metodzie skumulowanych efektów lokalnie trzeba wykonać trzy czynności:

Skupić się na efektach lokalnych

Dla danej wartości w siatce g wybierz z danych tylko te jednostki (i), dla których wartość cechy spada w sąsiedztwie tego punktu ($\{i: g - \delta \leq x_{ik} \leq g + \delta\}$). Dla cech skorelowanych wszystkie te jednostki powinny mieć stosunkowo spójne wartości wszystkich innych zmiennych.

Obliczyć nachylenie funkcji

W tym sąsiedztwie oblicz nachylenie dla każdej jednostki, a następnie uśrednij otrzymane wartości.

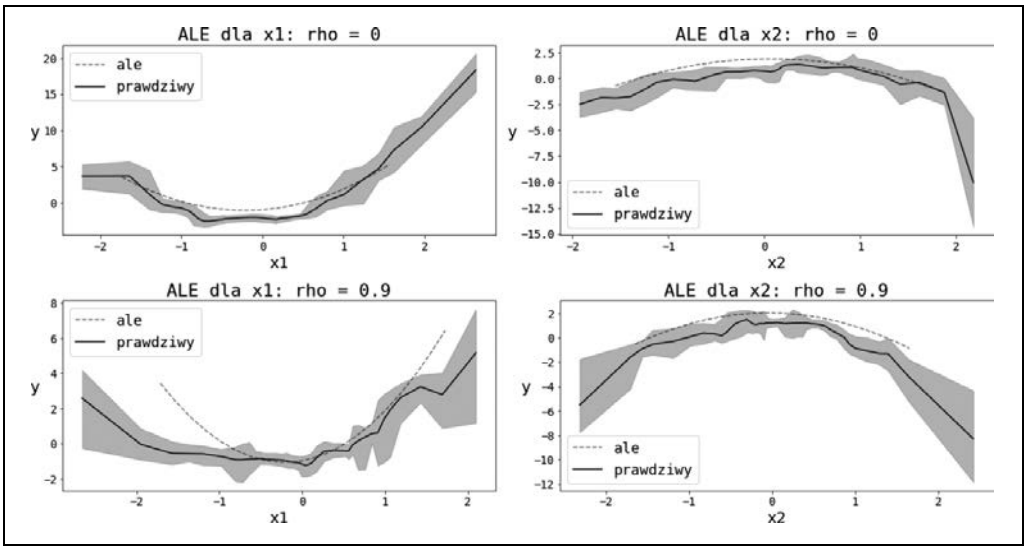
Obliczyć skumulowane wartości tych efektów

Na potrzeby wizualizacji wszystkie te efekty są kumulowane. Pozwala to przejść od lokalnego poziomu sąsiedztwa w siatce do globalnego zakresu funkcji.

[3] W czasie gdy powstaje ten tekst, dostępne są dwa pakiety Pythona, które obliczają skumulowane efekty lokalne: ALEPython (*https://oreil.ly/znDHe*) i alibi (*https://oreil.ly/QIZkS*). Moją implementację dla scenariusza z cechami ciągłymi i brakiem interakcji znajdziesz w repozytorium kodu (*https://oreil.ly/dshp-repo*).

Drugi krok jest istotny. Zamiast ograniczać się do obliczenia funkcji w jednym punkcie siatki, obliczasz nachylenie funkcji w określonym przedziale. W przeciwnym razie możesz pomylić efekt interesującej Cię cechy z efektem innych wysoce skorelowanych cech.

Rysunek 13.12 pokazuje skumulowany efekt lokalny dla tego samego symulowanego zestawu danych, który był używany wcześniej, z przedziałami ufności 90% otrzymanymi w wyniku bootstrappingu. Dla nieskorelowanych cech (pierwszy wiersz) metoda skumulowanych efektów lokalnych świetnie radzi sobie z obliczaniem prawdziwych efektów. Dla cech skorelowanych (drugi rząd) prawdziwy efekt drugiej cechy jest obliczany poprawnie, ale w niektórych punktach dla pierwszej cechy nadal występuje błąd systematyczny. Mimo to metoda skumulowanych efektów lokalnych nadal daje lepsze rezultaty niż wykresy zależności częściowych.



Rysunek 13.12. Skumulowane efekty lokalne dla tych samych symulowanych danych (z przedziałami ufności 90%)

Najważniejsze wnioski

Oto kluczowe wnioski z tego rozdziału:

Holistyczne opowiadanie historii w uczeniu maszynowym

Opowiadanie historii w uczeniu maszynowym najczęściej ma miejsce po opracowaniu modelu w trakcie spotkania z interesariuszami. Holistyczne podejście przedstawione w tym rozdziale jest zgodne z wizją, w której wcielasz się w rolę naukowca, aby tworzyć i iteracyjnie modyfikować historie pomagające opracować dobry model predykcyjny. Następnie możesz wcielić się w bardziej tradycyjną rolę sprzedawcy.

Opowiadanie historii przed opracowaniem modelu

Gdy chcesz opowiadać historie przed opracowaniem modelu, zacznij od stworzenia historii lub hipotez dotyczących tego, co wpływa wynik, który chcesz prognozować. Następnie na podstawie tych historii lub hipotez należy opracować cechy w wieloetapowym procesie inżynierii cech.

Opowiadanie historii po opracowaniu modelu

Opowiadanie historii po opracowaniu modelu pomaga zrozumieć i zinterpretować prognozy pochodzące z modelu. Techniki takie jak mapy cieplne, wykresy zależności częściowych i skumulowane efekty lokalne powinny pomóc w opowiedzeniu historii o wpływie, jaki poszczególne cechy mają na wynik. Znaczenie cech pozwala je uszeregować.

Ustrukturyzuj proces opowiadania historii przez podział go na etapy

Przynajmniej na początku dobrze jest nadać zestawowi narzędzi do opowiadania historii jakąś strukturę, zarówno w podejściu przed opracowaniem modelu, jak i w podejściu po opracowaniu modelu.

Dalsza lektura

Efekty pierwszego i drugiego rzędu omawiam w książce pod tytułem *Umiejętności analityczne w pracy z danymi i sztuczną inteligencją*.

Książka Rolfa Dobelliego *The Art of Thinking Clearly* (wydawnictwo Harper) jest dobra, jeśli chcesz zdobyć wiedzę na temat rozmaitych błędów i heurystyk obecnych w ludzkich zachowaniach. Mogą one znacznie wzbogacić zestaw hipotez dotyczących badanego problemu.

Dostępnych jest kilka kompletnych pozycji dotyczących inżynierii cech w kontekście transformacji danych. Możesz zapoznać się z książkami *Feature Engineering for Machine Learning* Alice Zheng i Amandy Casari (wydawnictwo O'Reilly), *Feature Engineering Bookcamp* Sinana Ozdemira (wydawnictwo Manning), *Python Feature Engineering Cookbook* (wydanie drugie) Soledad Galli (wydawnictwo Packt Publishing) lub z serią artykułów *Feature Engineering for Machine Learning* na blogu Winga Poona (*https://oreil.ly/Zg3EI*).

Rysunek 13.5 zaadaptowałem z rysunku 2.7 z książki *An Introduction to Statistical Learning with Applications in R* (wydanie drugie) Garetha Jamesa i in. (wydawnictwo Springer), udostępnianej w internecie przez autorów (*https://oreil.ly/LZPDX*). Gorąco polecam tę pozycję, jeśli bardziej interesuje Cię intuicyjne zrozumienie różnych zagadnień niż opanowanie technicznych szczegółów.

Jeśli chodzi o interpretowalność uczenia maszynowego, gorąco polecam *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* Christopha Molnara (dostępna w internecie, *https://oreil.ly/FujJr*, opublikowana niezależnie, 2023). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (wydanie drugie) Trevora Hastiego i in. (wydawnictwo Springer) zawiera doskonałe omówienie znaczenia cech i możliwości interpretacji w różnych algorytmach (zobacz przede wszystkim punkty 10.13 i 15.13.2). Wreszcie Michael Munn i David Pitman przedstawiają bardzo kompleksowy i aktualny przegląd różnych technik w książce *Explainable AI for Practitioners: Designing and Implementing Explainable ML Solutions* (wydawnictwo O'Reilly).

Jeśli chodzi o skumulowane efekty lokalne, możesz zapoznać się z artykułem Daniela W. Apleya i Jingyu Zhu *Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models* (sierpień 2019, pobrane z serwisu arXiv; *https://oreil.ly/gbZlu*). Omówienie tego zagadnienia przez Molnara jest bardzo dobre, ale artykuł zawiera więcej szczegółów na temat tego dość mało intuicyjnego algorytmu.

Skorowidz

A

agregacje czasowe, 127

B

biznes, 20
  tworzenie uzasadnień, 53
  zapobieganie oszustwom, 55
  zatrzymywanie klientów, 54
błąd
  doboru, 182, 187, 194
  systematyczny, 100, 180
  średniokwadratowy, 129
bootstrapping, 106

C

cechy, 126, 160
czułość, 171
czynniki zakłócające, 116, 123, 179

D

danolog, 215, 217
dekompozycja
  addytywna, 34, 41
  mix-rate, 41
  multiplikatywna, 36, 41
  przepływów i zapasów, 30
  przychodów, 31
  typu P×Q, 30
  wzrostu, 34
  zmian wag i wartości, 38
dopasowywanie, 188, 195
dryf
  danych, 141, 147
  modelu, 143, 147

E

etap
  oceny punktowej, 132, 134
  treningu, 132, 134

F

funkcja
  jako usługa, FaaS, 141
  wyniku, 126

G

generowanie
  danych wyjściowych, 141
  ocen punktowych, 141
  wartości, 19
gotowość produkcyjna, 138, 147

H

hipoteza, 210
  zerowa o braku efektu, 200
histogram, 82

K

klasyfikacje, 103
klasyfikator, 60
kolidery, 179, 194
korelacja, 123
kryterium decyzyjne, 199
  etapy wprowadzania, 202

L

lejek analityczny, 29
LTV, customer lifetime value, 47

M

macierz błędów, 171, 174
mapa cieplna, 161
metoda najmniejszych kwadratów, 110
minimalny wykrywalny efekt, MWE, 202-203, 212
  określanie wartości, 208
moc statystyczna, 205
modele
  czasu rzeczywistego, 140
  LLM, 168, 213, 220
moment olśnienia, 67

N

narracje, 64
  klarowność, 65
  prezentowanie, 74
  przejrzystość, 66
  tworzenie, 68
  wiarygodność, 66

O

oceny punktowe, 138
oczyszczanie danych, 219
okno
  obserwacji, 133, 134
  predykcji, 133, 134

P

podejmowanie decyzji, 168, 176, 179
potoki produkcyjne, 143, 147
poziom istotności, 205
precyzja, 171
prezentowanie rozkładów, 82
prognozowanie, 176
projekt 2×2, 44
  testowanie modelu, 45
  testowanie nowej cechy, 45
przekształcanie danych, 143
przewidywanie sprzedaży, 155
przeżywalność, 60

R

randomizacja, 187, 195
regresja liniowa, 110
  błąd systematyczny, 100
  czynniki zakłócające, 116
  dodatkowe zmienne, 118
  twierdzenie FWL, 113
  ustalenie punktu odniesienia, 158
  współczynniki, 110
reguły decyzyjne, 170, 176
rozkłady, 82

S

samoselekcja, 60
separacja
  całkowita, 130
  quasi-całkowita, 130
skumulowane efekty lokalne, ALE, 165
sprawdzanie poprawności
  danych, 144
  modelu, 146
  ocen punktowych, 146
symulacja, 91, 206
  modelu liniowego, 94
  regresji liniowej, 94
sztuczna inteligencja, 213

Ś

środowisko produkcyjne, 138
  kategorie modeli, 139
  modele czasu rzeczywistego, 140

T

testowanie
  hipotez, 209
  modelu, 45
  nowej cechy, 45
testy A/B, 198, 211, 218
trenowanie klasyfikatora, 60
twierdzenie Frischa-Waugha-Lovella, 113, 115, 123

U

uczenie maszynowe, 219
  etapy, 142, 145
  generowanie danych wyjściowych, 141
  oceny punktowe, 140
  opowiadanie historii
    holistyczne, 149, 166
    po opracowaniu modelu, 156, 167
    przed opracowaniem modelu, 150, 166
  podwójne, 192
  zastosowania, 170
uszeregowanie hipotez, 210

W

wariancja
  oszacowań, 120
  wyniku, 205
wartość, 19
  generowanie, 22
  P, 205
  pomiar, 23
  życiowa klienta, LTV, 47
wiarygodność
  biznesowa, 66
  danych, 66
  techniczna, 66
wnioskowanie przyczynowe, 178, 191
wsadowa ocena punktowa, 138
wskaźnik przyrostu, 59
wskaźniki, 209
  dekompozycja, 29
  właściwości, 27
  zastosowania, 62
współczynniki konwersji, 207
wyciekanie danych, 125
  błąd średniokwadratowy, 129
  identyfikowanie, 136
  metoda okien, 132
  usuwanie, 136
  wykrywanie, 128
wykres
  kaskadowy, 79
  liniowy, 77
  nachylenia, 79
  punktowy, 81
  słupkowy, 77
wykresy zależności częściowych, 96, 163
wzmacnianie gradientowe, 121

Z

zarządzanie eksperymentami, 210
zmiany dodatkowe, 177, 194
zmienne
  kontrolne, 126
  sztuczne, 121, 124
  zakłócające, 186, 194
znacznik czasu, 126

Ten podręcznik zaczyna się tam, gdzie większość książek się kończy — od rzeczywistych procesów decyzyjnych opartych na wnioskach wynikających z danych.

Brett Holleman, niezależny danolog

Uczenie się i praktykowanie danologii nie należy do najłatwiejszych zadań. Edukacja w tej dziedzinie zazwyczaj dotyczy programowania i uczenia maszynowego, a przecież świetny analityk danych musi się znać na wielu innych zagadnieniach. Może się ich nauczyć w pracy, ale w tym celu konieczne jest znalezienie mentora. A to niestety nie zawsze jest możliwe.

Dzięki tej książce przyswoisz różne techniki, które pomogą Ci stać się bardziej produktywnym analitykiem danych. Najpierw zapoznasz się z tematami związanymi z rozumieniem danych i umiejętnościami miękkimi, które okazują się konieczne w pracy dobrego danologa. Dopiero potem skupisz się na kluczowych aspektach uczenia maszynowego. W ten sposób stopniowo przejdziesz ścieżkę od przeciętnego kandydata do wyjątkowego specjalisty data science. Umiejętności opisane w tym przewodniku przez wiele lat były rozpoznawane, katalogowane, analizowane i stosowane do generowania wartości i szkolenia danologów w różnych firmach i branżach.

Oto brakujący podręcznik pozwalający odnieść sukces komercyjny dzięki data science!

Adri Purkayastha, dyrektor do spraw zagrożeń związanych z AI, BNP Paribas

Z książki dowiesz się:

- jak sprawić, by procesy oparte na analizie danych generowały wartość
- jak zaprojektować przydatne wskaźniki
- jak zdobywać poparcie interesariuszy
- jak się upewnić, że algorytm uczenia maszynowego nadaje się do rozwiązania danego zadania
- jak zapanować nad wyciekami danych

Dr Daniel Vaughan od piętnastu lat zajmuje się rozwiązywaniem problemów przy użyciu metod predykcyjnych i normatywnych. Kierował zespołami danologów, a obecnie doradza kilku organizacjom z branży fintech. Jest autorem książki *Umiejętności analityczne w pracy z danymi i sztuczną inteligencją* (Helion, 2021).

Helion
helion.pl
HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl
KOD KORZYŚCI
Sięgnij po więcej!
ISBN 978-83-289-1294-6
9 788328 912946
Cena: 79,00 zł