

Analityk danych : przewodnik po Data Science, statystyce i uczeniu maszynowym / Alex J. Gutman, Jordan Goldmeier. – Gliwice, copyright 2023

Spis treści

O redaktorach technicznych	ix
Podziękowania	xi
Przedmowa	xxi
Wprowadzenie	xxv
CZĘŚĆ I	
Myśl jak spec od danych	
ROZDZIAŁ 1	
Na czym polega problem?	3
Pytania, które powinien zadawać spec od danych	4
Dlaczego problem jest ważny?	4
Na kogo wpływa ten problem?	6
Co, jeśli nie mamy właściwych danych?	7
Kiedy projekt się zakończy?	7
Co, jeśli nie spodobają nam się rezultaty?	7
Dlaczego projekty związane z danymi kończą się niepowodzeniem?	8
Wrażenia klientów	8
Omówienie	10
Praca nad problemami, które mają znaczenie	11
Podsumowanie rozdziału	11
ROZDZIAŁ 2	
Czym są dane?	13
Dane a informacje	13
Przykładowy zbiór danych	14
Typy danych	15
Jak gromadzi się dane i jaką mają strukturę?	16
Dane obserwacyjne i eksperymentalne	16
Dane ustrukturyzowane i nieustrukturyzowane	17
Podstawowe statystyki zbiorcze	18
Podsumowanie rozdziału	19
ROZDZIAŁ 3	
Przygotowanie do myślenia statystycznego	21
Zadawaj pytania	22
Wszystko jest zmienne	23

Scenariusz: wrażenia klientów (kontynuacja)	24
Studium przypadku: zachorowalność na raka nerki	26
Prawdopodobieństwo i statystyka	28
Prawdopodobieństwo a intuicja	29
Odkrywanie informacji za pomocą statystyki	31
Podsumowanie rozdziału	33

CZEŚĆ II

Mów jak spec od danych

ROZDZIAŁ 4

Polemizuj z danymi	37
Co byś zrobił(a)?	38
Katastrofa spowodowana brakiem danych	39
Jaka jest historia pochodzenia danych?	43
Kto zebrał dane?	44
Jak zebrano dane?	44
Czy dane są reprezentatywne?	45
Czy poprawnie dobrano próbę?	46
Co zrobiono z wartościami odstającymi?	46
Jakich danych nie widzisz?	47
Jak rozwiązano problem brakujących wartości?	47
Czy dane mogą zmierzyć to, co ma być mierzone?	48
Polemizuj z danymi każdej wielkości	48
Podsumowanie rozdziału	49

ROZDZIAŁ 5

Eksploruj dane	51
Eksploracyjna analiza danych i Ty	52
Przyjmij nastawienie eksploracyjne	52
Pytania naprowadzające	53
Scenariusz	53
Czy dane mogą odpowiedzieć na pytanie?	54
Określ oczekiwania i użyj zdrowego rozsądku	54
Czy wartości mają intuicyjny sens?	54
Uważaj! Wartości odstające i brakujące	58
Czy odkryliście jakieś związki?	58
Korelacja	59
Uważaj! Błędne interpretowanie korelacji	60
Uważaj! Korelacja nie implikuje przyczynowości	62
Czy znaleźliście w danych nowe możliwości?	63
Podsumowanie rozdziału	63

ROZDZIAŁ 6

Badaj prawdopodobieństwa	65
Zgadnij odpowiedź	66
Reguły gry	67
Notacja	67

Prawdopodobieństwo warunkowe i zdarzenia niezależne	69
Prawdopodobieństwo wielu zdarzeń	69
Dwie rzeczy, które zdarzają się razem	69
Jedno albo drugie	71
Ćwiczenie myślowe z zakresu prawdopodobieństwa	72
Następne kroki	73
Uważaj z zakładaniem niezależności	74
Nie popełniaj błędu hazardzisty	75
Wszystkie prawdopodobieństwa są warunkowe	75
Nie przedstawiaj zależności	76
Twierdzenie Bayesa	77
Upewnij się, że prawdopodobieństwa mają znaczenie	80
Kalibracja	80
Rzadkie zdarzenia mogą się zdarzać i się zdarzają	80
Podsumowanie rozdziału	81

ROZDZIAŁ 7

Kwestionuj statystyki 83

Krótkie lekcje o wnioskowaniu	83
Zostaw sobie trochę przestrzeni	84
Więcej danych, więcej dowodów	84
Kwestionuj status quo	85
Dowody na twierdzenie przeciwne	86
Równoważenie błędów decyzyjnych	88
Proces wnioskowania statystycznego	89
Pytania, które pomogą Ci kwestionować statystyki	90
Jaki jest kontekst tych statystyk?	90
Jaki jest rozmiar próby?	91
Co testujecie?	91
Jaka jest hipoteza zerowa?	92
Zakładanie równoważności	93
Jaki jest poziom istotności?	93
Ile przeprowadzacie testów?	94
Czy mogę zobaczyć przedziały ufności?	94
Czy jest to praktycznie istotne?	96
Czy zakładacie przyczynowość?	96
Podsumowanie rozdziału	97

CZEŚĆ III

Przybornik specjalisty data science

ROZDZIAŁ 8

W poszukiwaniu ukrytych grup 101

Uczenie nienadzorowane	102
Redukcja wymiarowości	102
Tworzenie cech złożonych	103
Analiza składowych głównych	105
Składowe główne zdolności sportowych	105

Podsumowanie PCA	108
Potencjalne pułapki	109
Klasteryzacja	109
Klasteryzacja metodą k-średnich	111
Klasteryzacja sklepów detalicznych	111
Potencjalne pułapki	113
Podsumowanie rozdziału	114

ROZDZIAŁ 9

Model regresji	117
Uczenie nadzorowane	117
Jak działa regresja liniowa?	119
Regresja metodą najmniejszych kwadratów: nie tylko pomysłowa nazwa	120
Regresja liniowa: co Ci daje?	123
Rozszerzanie modelu na wiele cech	124
Regresja liniowa: jakie powoduje nieporozumienia?	125
Pominięte zmienne	126
Współliniowość	126
Przeciek danych	127
Błędy ekstrapolacji	128
Relacje nie zawsze są liniowe	128
Wyjaśniasz czy przewidujesz?	128
Skuteczność regresji	130
Inne modele regresji	131
Podsumowanie rozdziału	131

ROZDZIAŁ 10

Model klasyfikacji	133
Wprowadzenie do klasyfikacji	133
Czego się nauczysz?	134
Przykładowy problem klasyfikacji	135
Regresja logistyczna	135
Regresja logistyczna — i co z tego?	138
Drzewa decyzyjne	139
Metody zespołowe	142
Lasy losowe	143
Drzewa wzmacniane gradientowo	143
Interpretowalność modeli zespołowych	145
Strzeż się pułapek	145
Złe podejście do problemu	146
Przeciek danych	146
Brak podziału danych	146
Wybór odpowiedniego progu decyzyjnego	147
Błędne rozumienie dokładności	147
Macierze błędów	148
Podsumowanie rozdziału	150

ROZDZIAŁ 11

Analiza tekstu	151
Oczekiwanie wobec analizy tekstu	151
Jak tekst staje się liczbami	153
Wielki worek słów	153
N-gramy	157
Osadzenia słów	158
Modelowanie tematyczne	160
Klasyfikacja tekstu	162
Naiwny klasyfikator bayesowski	164
Analiza odczuć	166
Kwestie praktyczne podczas pracy z tekstem	167
Giganci technologiczni mają przewagę	168
Podsumowanie rozdziału	169

ROZDZIAŁ 12

Uczenie głębokie	171
Sieci neuronowe	172
Pod jakimi względami sieci neuronowe przypominają ludzki mózg?	172
Prosta sieć neuronowa	173
Jak uczy się sieć neuronowa?	174
Nieco bardziej złożona sieć neuronowa	175
Zastosowania uczenia głębokiego	178
Korzyści z uczenia głębokiego	179
Jak komputery „widzą” obrazy?	180
Konwolucyjne sieci neuronowe	181
Uczenie głębokie w języku i sekwencjach	183
Uczenie głębokie w praktyce	184
Czy masz dane?	184
Czy Twoje dane są ustrukturyzowane?	185
Jak będzie wyglądać sieć?	186
Sztuczna inteligencja i Ty	187
Giganci technologiczni mają przewagę	188
Etyka w uczeniu głębokim	188
Podsumowanie rozdziału	189

CZEŚĆ IV

Droga do sukcesu

ROZDZIAŁ 13

Strzeż się pułapek	193
Tendycyjność i dziwne zjawiska w danych	194
Błąd przeżywalności	194
Regresja do średniej	195
Paradoks Simpsona	195
Błąd potwierdzenia	197
Błąd utopionych kosztów	197
Dyskryminacja algorytmiczna	197

Nieskategoryzowane przejawy tendencyjności	198
Wielka lista pułapek	199
Pułapki związane ze statystyką i uczeniem maszynowym	199
Pułapki związane z projektem	200
Podsumowanie rozdziału	202

ROZDZIAŁ 14

Znaj ludzi i osobowości	203
Siedem scenariuszy fiaska komunikacyjnego	204
Post mortem	204
Wieczorynka	205
Głuchy telefon	206
W gąszczu szczegółów	206
Konfrontacja z rzeczywistością	207
Wrogie przejęcie	207
Egocentryk	207
Osobowości w świecie danych	208
Entuzjasta	208
Cynik	209
Spec od danych	209
Podsumowanie rozdziału	209

ROZDZIAŁ 15

Co dalej?	211
------------------	------------