

Spis treści

PODZIĘKOWANIA	13
WPROWADZENIE	15
1	
EKSPLORACYJNA ANALIZA DANYCH	23
Twój pierwszy dzień na stanowisku prezesa	24
Znajdowanie wzorców w zbiorach danych	24
Używanie plików .csv do przeglądania i przechowywania danych	26
Wyświetlanie danych w Pythonie	27
Obliczanie statystyk zbiorczych	29
Analizowanie podzbiorów danych	32
Dane nocne	32
Dane sezonowe	33
Wizualizacja danych za pomocą Matplotlib	35
Rysowanie i wyświetlanie prostego wykresu	35
Tytuły i etykiety na wykresach	36
Kreślenie podzbiorów danych	37
Testowanie różnych typów wykresów	38
Eksplorowanie korelacji	44
Obliczanie korelacji	44
Silna i słaba korelacja	45
Znajdowanie korelacji między zmiennymi	48
Tworzenie map cieplnych	49
Dalsza eksploracja	52
Podsumowanie	52

2		
PROGNOZOWANIE		53
Przewidywanie popytu		53
Oczyszczanie błędnych danych		54
Kreślenie danych w celu odkrywania trendów		57
Przeprowadzanie regresji liniowej		58
Algebraiczne podstawy linii regresji		60
Obliczanie miar błędu		62
Używanie regresji do prognozowania przyszłych trendów		66
Wypróbowywanie innych modeli regresji		68
Przewidywanie sprzedaży z wykorzystaniem multiwariantnej regresji liniowej		68
Oddawanie zmienności z wykorzystaniem trygonometrii		71
Wybieranie najlepszej regresji na potrzeby prognozowania		74
Dalsza eksploracja		79
Podsumowanie		80
3		
PORÓWNYWANIE GRUP		81
Dane dotyczące populacji		81
Statystyki zbiorcze		82
Próby losowe		83
Różnice między próbami		86
Testowanie hipotez		90
Test t		91
Niuanse testowania hipotez		92
Porównywanie grup w kontekście praktycznym		94
Podsumowanie		98
4		
TESTY A/B		99
Potrzeba eksperymentowania		100
Przeprowadzanie eksperymentów w celu testowania nowych hipotez		101
Matematyka testów A/B		105
Przekładanie matematyki na praktykę		106
Optymalizacja z wykorzystaniem modelu mistrz – pretendent		108
Zapobieganie pomyłkom z wykorzystaniem prawa Twymana i testów A/A		109
Wielkości efektu		110
Obliczanie istotności danych		113
Zastosowania i kwestie zaawansowane		115
Etyka testów A/B		116
Podsumowanie		118

5	
KLASYFIKACJA BINARNA	119
Minimalizowanie odpływu klientów	120
Używanie liniowych modeli prawdopodobieństwa do znajdowania klientów wysokiego ryzyka	121
Wykres ryzyka odpływu	122
Potwierdzanie zależności za pomocą regresji liniowej	124
Przewidywanie przyszłości	126
Dokonywanie rekomendacji biznesowych	128
Mierzenie dokładności prognoz	129
Multiwariantne modele LPM	131
Tworzenie nowych miar	133
Wady modeli LPM	135
Przewidywanie binarnych wyników za pomocą regresji logistycznej	135
Rysowanie krzywych logistycznych	136
Dopasowywanie krzywej logistycznej do danych	138
Zastosowania klasyfikacji binarnej	140
Podsumowanie	140
6	
UCZENIE NADZOROWANE	141
Przewidywanie ruchu w witrynie internetowej	142
Wczytywanie i wykreślanie danych dotyczących artykułów	142
Używanie regresji liniowej jako metody przewidywania	145
Uczenie nadzorowane	147
Metoda k najbliższych sąsiadów	149
Implementowanie metody k-NN	150
Analiza k-NN z wykorzystaniem pakietu sklearn	152
Używanie innych algorytmów uczenia nadzorowanego	153
Drzewa decyzyjne	154
Lasy losowe	156
Sieci neuronowe	157
Mierzenie dokładności prognoz	159
Praca z modelami multiwariantnymi	161
Używanie klasyfikacji zamiast regresji	162
Podsumowanie	164

UCZENIE NIENADZOROWANE	165
Uczenie nadzorowane a uczenie nienadzorowane	165
Generowanie i eksplorowanie danych	167
Rzucanie kością	167
Używanie innego typu kości	171
Określanie źródła obserwacji przez klasteryzację	174
Klasteryzacja w zastosowaniach biznesowych	177
Analizowanie wielu wymiarów	179
Klasteryzacja E-M	181
Etap zgadywania	183
Etap oczekiwania	184
Etap maksymalizacji	186
Etap konwergencji	189
Inne metody klasteryzacji	191
Inne metody uczenia nienadzorowanego	194
Podsumowanie	195

WEB SCRAPING	197
Jak działają witryny internetowe?	198
Twój pierwszy scraper	199
Parsowanie kodu HTML	201
Scraping adresów e-mail	201
Bezpośrednie wyszukiwanie adresów	203
Wyrażenia regularne	204
Używanie metaznaków do elastycznych wyszukiwań	206
Dostrajanie wyszukiwań z wykorzystaniem sekwencji unikowych	207
Łączenie metaznaków w wyszukiwaniach zaawansowanych	209
Używanie wyrażeń regularnych do wyszukiwania adresów e-mail	210
Przekształcanie wyników w użyteczne dane	211
Używanie biblioteki BeautifulSoup	213
Parsowanie elementów etykiety HTML	214
Scraping i parsowanie tabel HTML	215
Zaawansowany scraping	217
Podsumowanie	218

9	
SYSTEMY REKOMENDACYJNE	219
Rekomendacje oparte na popularności	220
Filtrowanie kolaboratywne oparte na artykułach	222
Mierzenie podobieństwa wektorów	223
Obliczanie podobieństwa kosinusowego	226
Implementowanie filtrowania kolaboratywnego opartego na artykułach	227
Filtrowanie kolaboratywne oparte na użytkownikach	229
Studium przypadku: rekomendacje muzyczne	232
Generowanie rekomendacji za pomocą zaawansowanych systemów	234
Podsumowanie	236
10	
PRZETWARZANIE JĘZYKA NATURALNEGO	237
Używanie NLP do wykrywania plagiatów	238
Model NLP word2vec	239
Ocenianie podobieństw między słowami	239
Tworzenie układu równań	241
Analizowanie wektorów liczbowych w word2vec	245
Manipulowanie wektorami poprzez obliczenia matematyczne	248
Wykrywanie plagiatów z wykorzystaniem word2vec	249
Model skip-thoughts	250
Modelowanie tematyczne	253
Inne zastosowania NLP	255
Podsumowanie	256
11	
DATA SCIENCE W INNYCH JĘZYKACH	257
Wygrywanie meczów piłkarskich z pomocą SQL-a	258
Wczytywanie i analizowanie danych	258
Podstawy SQL-a	260
Tworzenie bazy danych SQL	260
Wykonywanie kwerend SQL	261
Łączenie danych poprzez łączenie tabel	264
Wygrywanie meczów piłkarskich z pomocą języka R	267
Podstawy języka R	267
Stosowanie regresji liniowej w języku R	269
Kreślenie danych w R	270
Inne przydatne umiejętności	272
Podsumowanie	274
SKOROWIDZ	275