
Nowoczesne analizy biznesowe

*Zwiększanie wartości danych
przy użyciu Pythona i R*

Deanne Larson

przekład: Marek Włodarz

Spis treści

Przedmowa	xi
1. Rola analityka biznesowego i analiz.	1
Czym jest rola analityka biznesowego?	1
Umiejętności	2
Obowiązki	3
Typy analityków	4
Dlaczego analityk biznesowy musi znać analitykę?	8
Eksploracja danych	8
Kontekst biznesowy	9
Analityka	9
Wkład analityka biznesowego w wartość analityczną	11
Problemy biznesowe rozwiązywane przez analitykę	12
Współpraca z innymi zespołami	13
Umiejętności wykorzystywane w analityce	13
Python i R	13
Cykl życia projektu analitycznego	14
Podsumowanie	14
2. Metodologie projektów analitycznych	15
Zrozumienie biznesu	16
Ustalanie zamierzeń biznesowych	16
Ocena sytuacji	17
Ustalanie celów	17
Ustalenie podejścia i planu	18
Ocena narzędzi i technik	18
Poznanie i przygotowywanie danych	18
Ocena zawartości i jakości danych	18
Wybieranie i czyszczenie danych	19

Konstruowanie i integrowanie zestawu danych	19
Tworzenie zestawu danych dla budowania modelu	19
Modelowanie i ocena	20
Wybór techniki analitycznej	20
Budowanie i ocena modelu	20
Wdrożenie	20
Ocena sprawności modelu	21
Ustalanie odstępów czasu pomiędzy ocenami	21
Operacje modelu	21
Monitorowanie modeli	21
Życie modelu	22
Ponowne trenowanie	22
Podsumowanie	22
3. Wprowadzenie do R i Pythona	23
Instalowanie R i Pythona oraz opcje konfiguracji	24
Dlaczego uczyć się R i Pythona?	24
Uczyć się obu na raz, czy każdego kolejno?	24
Zalety i wady różnych strategii nauki	25
Instalacja R	25
Instalacja Pythona	26
Pisanie skryptów w R i Pythonie	27
Skrypty języka R	27
Skrypty języka Python	29
Koncepcje obiektowe	32
Struktura OOP	32
Zasady OOP	34
Typy danych w R i Pythonie	40
Typy danych w R	41
Struktury danych w R	42
Typy danych w Pythonie	43
Struktury danych Pythona	46
Interakcja z relacyjnymi bazami danych	53
Dlaczego relacyjne bazy danych?	54
Połączenie R z relacyjną bazą danych	54
Łączenie się z relacyjnymi bazami danych z poziomu R	55
SQLite	56
Łączenie się z relacyjnymi bazami danych z poziomu Pythona	57
Połączenie Pythona z relacyjną bazą danych	57
Podsumowanie	57

4. Analizy statystyczne	59
Przykładowy projekt analityczny	59
Rotacja klientów w firmie telekomunikacyjnej	60
Test A/B	61
Kampanie marketingowe	62
Prognozowanie finansowe	63
Diagnostyka zdrowia	64
Zacznijmy od sformułowania problemu	65
Rozpoznawanie problemu analitycznego	66
Klasyfikacja	69
Regresja	70
Co chcemy mierzyć?	70
Podejścia analityczne	70
EDA	71
Uczenie nienadzorowane	71
Statystyczna analiza dla regresji	72
Analizy dla klasyfikacji	74
Rola testowania hipotez	75
Wizualizacje	76
Wizualizacje wspomagające EDA w R i Pythonie	76
Wizualizacje dla regresji	78
Wizualizacje dla klasyfikacji	85
Podsumowanie	95
5. Eksploracyjna analiza danych w R i Pythonie	97
Jakość danych	97
Charakterystyki jakości danych	97
Profilowanie danych	99
Klasteryzacja i uczenie nienadzorowane	106
Cele uczenia nienadzorowanego	107
Przykład, jaki wpływ na nadzorowane uczenie może mieć klasteryzacja	108
Metoda centroidów	110
Klasteryzacja hierarchiczna	114
Inne metody nienadzorowane używane w fazie EDA	117
Identyfikowanie wartości odstających	117
Wartości odstające w regresji	118
Wartości odstające w klasyfikacji	119
Przygotowanie danych do modelowania	122
Próbkowanie	122
Uczenie się i testowanie	124
Transformacje danych	125
Przekształcanie danych	133

Wybieranie i redukowanie cech	137
Wybieranie cech	138
Techniki redukowania cech	146
Podsumowanie	149
6. Stosowanie i ocena modelowania w R i Pythonie	151
Kroki modelowania	151
Wybór i trenowanie modelu	151
Ocena modelu	153
Optymalizacja modelu	153
Wdrażanie modelu	154
Monitorowanie modelu i konserwacja	155
Wybieranie właściwego algorytmu	155
Regresja	157
Typowe przypadki użycia	157
Równanie regresji liniowej	158
Regresja liniowa w R	159
Regresja liniowa w Pythonie	161
Wielokrotna regresja liniowa	166
Inne rodzaje regresji	169
Wyzwania związane z modelami regresji	171
Inne algorytmy dla regresji	172
Drzewa decyzyjne w regresji	175
Ewaluacja regresji liniowej	177
Klasyfikacja	183
Typowe przypadki użycia	184
Algorytmy klasyfikacyjne	186
Klasyfikacja w R	188
Klasyfikacja w Pythonie	190
Przypadek użycia: rotacja klientów firmy telekomunikacyjnej	191
Ewaluacja klasyfikacji	205
Ewaluacja klasyfikacji: przypadek użycia	215
Podsumowanie	224
7. Modelowanie i wybór algorytmu	225
Algorytmy	225
Kryteria wyboru algorytmu	226
Typ problemu	226
Modele interpretowalne	228
Dokładność prognozowania	238
Szybkość uczenia się	239
Szybkość przewidywania	240
Dostrajanie hiperparametrów	241

Praca z małymi zbiorami danych.	246
Praca z wielkimi zbiorami danych.	247
Interakcja cech	248
Charakterystyka danych	252
Przykład: Wybieranie odpowiedniego algorytmu.	254
Wybieranie właściwego algorytmu dla przewidywania sprzedaży	254
Ocena kryteriów.	256
Decyzja i implementacja.	257
Podsumowanie	257
8. ModelOps	259
Przegląd operacji modelu	259
Procesy operacji modelu.	261
Ocenianie modelu	261
Ocenianie modelu w R: wykorzystanie aplikacji Shiny do punktowania w czasie rzeczywistym	262
Ocenianie modelu w Pythonie: wdrażanie modelu przy użyciu Streamlit	267
Monitorowanie modelu.	269
Kluczowe metryki i wskaźniki monitorowania wydajności modelu	269
Techniki automatycznego monitorowania modelu.	270
Ponowne trenowanie modelu.	276
Zdarzenia wyzwające ponowne trenowanie modelu.	277
Zautomatyzowane techniki ponownego trenowania	277
Implementacja w R: wykorzystanie zadań cron do okresowego ponownego trenowania	278
Implementacja w Pythonie: wykorzystanie narzędzi typu Airflow do zarządzania przepływem pracy	278
Generowanie raportów	279
Zawartość i struktura raportu końcowego	280
Techniki zautomatyzowanego generowania raportów	283
Implementacja w R: generowanie raportów w R za pomocą Markdown i knitr.	284
Implementacja w Pythonie: tworzenie raportów za pomocą notatników Jupyter i nbconvert	286
Kontrola wersji i odtwarzalność modelu	289
Praktyki współdziałania i dokumentowania	289
Przypadki użycia ModelOps.	290
Prognozowanie sprzedaży detalicznych: automatyzacja ocen i monitorowania	290
Wykrywanie oszustw: dynamiczne ponowne trenowanie modelu i raportowanie	290

Przewidywanie odejść klientów: planowe ponowne trenowanie modelu i generowanie raportów końcowych	290
Integracja z istniejącymi systemami i infrastrukturą	291
Kierunki rozwojowe MLOps	291
Podsumowanie	293
9. Zaawansowane wizualizacje	295
Zaawansowane wizualizacje w R Shiny	296
Czym jest R Shiny?	296
Kluczowe funkcjonalności i możliwości R Shiny	296
Konfigurowanie środowiska	300
Budowanie pierwszej aplikacji Shiny	300
Tworzenie zaawansowanych UI	308
Przykład: Tworzenie tablicy kontrolnej do monitorowania sprzedaży w czasie rzeczywistym	319
Wizualizacje w Pythonie	324
Przegląd możliwości wizualizacji w Pythonie	325
Powszechnie używane biblioteki: Matplotlib, Seaborn, Plotly oraz Dash	325
Matplotlib: podstawy wizualizacji w Pythonie	327
Dostosowywanie wykresów za pomocą stylów i kolorów	328
Wykresy statystyczne: punktowe, skrzypcowe, mapy ciepła	330
Interaktywne wykresy w Plotly	334
Wykresy 3D za pomocą Matplotlib i Plotly	336
Wizualizacja danych geoprzestrzennych	340
Tworzenie tablicy kontrolnej: używanie Plotly Dash	342
Przypadek użycia: wykorzystanie Pythona w zaawansowanym projekcie wizualizacji	344
Wizualizacje: R Shiny czy Python?	352
Podsumowanie	353
10. Nowoczesne typy danych w analityce	355
Dane półstrukturalne (JSON)	355
Używanie danych JSON w Pythonie	356
Używanie danych JSON w R	360
Dane mediów społecznościowych	364
Typy danych mediów społecznościowych	364
Używanie Pythona do analizy danych mediów społecznościowych	367
Używanie R do analizy danych mediów społecznościowych	389
Dane obrazów	399
Przetwarzanie obrazów w Pythonie	399
Przetwarzanie obrazów w R	401
Dane wideo	402
Używanie Pythona do danych wideo	403

Używanie R do danych wideo	408
Podsumowanie	411
11. Mierzenie wartości biznesowej uzyskiwanej z analityki i rola AI	413
Czym jest wartość biznesowa w analityce?	413
Wpływ strategiczny	414
Wydajność operacyjna	414
Satysfakcja i lojalność klientów	415
Metryki i kluczowe wskaźniki do mierzenia wartości biznesowej	415
Metryki finansowe	415
Metryki operacyjne	416
Metryki klienckie	416
Dopasowywanie metryk do celów organizacji	417
Wykorzystanie metryk do zademonstrowania wartości	417
Metryki i KPI w praktyce	417
Przykłady przypadków biznesowych dla wartości analityki	418
Krok 1: Definiowanie problemu i określenie mierzalnych wyników	418
Krok 2: Identyfikowanie metryk do zmierzenia sukcesu lub porażki	418
Krok 3: Implementowanie rozwiązań analitycznych	419
Krok 4: Mierzenie i demonstrowanie wartości	420
Krok 5: Raportowanie i ciągle ulepszanie	420
AI oraz generatywna AI w analityce biznesowej	421
Wprowadzenie do generatywnych AI	422
Zastosowania w projektowaniu produktów	422
Zastosowania w tworzeniu treści	422
Zastosowania w marketingu	422
Poprawa jakości obsługi klienta	423
Ulepszanie sprawności operacyjnej	423
Perspektywy i wyzwania	423
Przypadki użycia AI i generatywnej AI w analityce biznesowej	424
Przypadek 1: Spostrzeżenia i rekomendacje dla klientów oparte na sztucznej inteligencji	424
Przypadek 2: Wspomagana przez AI optymalizacja łańcucha dostaw	424
Przypadek 3: Poprawianie podejmowania decyzji	425
Przypadek 4: Sztuczna inteligencja w analityce medycznej	425
Przypadek 5: Użycie generatywnej AI do spersonalizowania doświadczeń klientów	425
Przypadek 6: AI w analityce branży detalicznej	425
Problemy niespójności z faktami i współpraca człowieka ze sztuczną inteligencją	426
Perspektywy na przyszłość	426
Wyzwania i uwarunkowania	427
Wyzwania dotyczące integracji i skalowalności	427

Łagodzenie obciążeń i zapewnienie uczciwości w decyzjach podejmowanych przy użyciu sztucznej inteligencji	427
Techniczne i organizacyjne wyzwania wdrożeń AI	428
Uwarunkowania związane z kosztami i zasobami	428
Zapewnianie trwałości inwestycji w AI	428
Podsumowanie	429
Indeks	431
O autorce	453