Оглавление

Предисловие 12
От издательства 16

ЧАСТЬ I
ОБЗОР ПРОЦЕССА МАШИННОГО ОБУЧЕНИЯ И РЕАЛЬНЫХ ЗАДАЧ, РЕШАЕМЫХ В КНИГЕ

Глава 1. Основные этапы процесса машинного обучения 18

Глава 2. Примеры данных и практических задач, рассмотренные в книге 23
2.1. Обнаружение мошенничества 23
2.2. Прогнозирование временных рядов и выявление аномалий 25

ЧАСТЬ II
ПОДГОТОВКА К ПРОГРАММИРОВАНИЮ НА PYTHON

Глава 3. Настройка рабочего окружения 28
3.1. Установка Python 28
3.2. Установка интегрированной среды разработки 29
3.3. Организация структуры папок проекта 30

ЧАСТЬ III
ОТ СБОРА ДАННЫХ К ИХ ПРОВЕРКЕ И ОЦЕНКЕ

Глава 4. Сбор данных: от файлов до баз данных 36
4.1. Создание искусственных данных с помощью функций из библиотеки scikit-learn 36
4.1.1. Установка библиотеки 36
4.1.2. Генерация данных 37

4.2. Загрузка данных с помощью библиотеки pandas 40
4.2.1. Установка библиотеки 41
4.2.2. Загрузка данных 41
4.3. Загрузка данных с помощью библиотеки Dask 48
4.3.1. Установка библиотеки 50
4.3.2. Генерация данных 50
4.3.3. Загрузка данных 51
4.4. Сбор данных с помощью других библиотек 55
4.4.1. Загрузка данных из интернета 55
4.4.2. Загрузка данных из реляционных баз данных 59

Глава 5. Визуализация данных 62
5.1. Основы визуализации с помощью библиотеки matplotlib 62
5.1.1. Установка библиотеки 62
5.1.2. Возможности библиотеки 63
5.1.3. Примеры использования библиотеки 65
5.2. Основы визуализации с помощью библиотеки seaborn 72
5.2.1. Установка библиотеки 72
5.2.2. Возможности библиотеки 73
5.2.3. Примеры использования библиотеки 73
5.3. Основы визуализации с помощью библиотеки pandas 82
5.4. Основы визуализации с помощью библиотеки Bokeh 88
5.4.1. Установка библиотеки 88
5.4.2. Возможности библиотеки 89
5.4.3. Примеры использования библиотеки 90
5.5. Сравнение библиотек визуализации 96

Глава 6. Проверка и профилирование данных 98
6.1. Профилирование с помощью библиотеки ydata-profiling 99
6.1.1. Установка библиотеки 100
6.1.2. Возможности библиотеки 100
6.1.3. Примеры использования библиотеки 103
6.2. Профилирование с помощью библиотеки Great Expectations 108
6.2.1. Установка библиотеки 109
6.2.2. Возможности библиотеки 109
6.2.3. Примеры использования библиотеки 112
6.2.4. Сценарии применения утверждений для практических задач 121

ЧАСТЬ IV
ОТ ПРЕДОБРАБОТКИ ДАННЫХ К ВЫБОРУ МЕТРИКИ ОЦЕНКИ ЭФФЕКТИВНОСТИ МОДЕЛИ

Глава 7. Предобработка данных: устранение дефектов и подготовка данных к моделированию ... 125
7.1. Генерация искусственных данных ... 126
7.2. Предобработка данных с помощью библиотеки scikit-learn ... 132
7.2.1. Удаление строк-дубликатов ... 133
7.2.2. Приведение признаков к правильному типу представления ... 133
7.2.3. Заполнение пропусков значений ... 134
7.2.4. Преобразования категориальных признаков в числовое представление ... 137
7.2.5. Поиск и обработка аномальных (нетипичных) значений ... 141
7.2.6. Масштабирование значений числовых признаков ... 142
7.2.7. Создание новых признаков, если это необходимо ... 143
7.2.8. Сокращение количества признаков ... 143
7.2.9. Создание конвейера, объединяющего все операции предобработки ... 144
7.3. Предобработка данных с помощью библиотеки feature-engine ... 151
7.3.1. Установка библиотеки ... 151
7.3.2. Удаление строк-дубликатов ... 151
7.3.3. Приведение признаков к правильному типу представления ... 152
7.3.4. Заполнение пропусков значений ... 152
7.3.5. Преобразования категориальных признаков в числовое представление ... 156
7.3.6. Поиск и обработка аномальных (нетипичных) значений ... 158
7.3.7. Масштабирование значений числовых признаков ... 159
7.3.8. Создание новых признаков, если это необходимо ... 160
7.3.9. Сокращение количества признаков ... 161
7.3.10. Создание конвейера, объединяющего все операции предобработки ... 162
7.4. Предобработка данных с помощью библиотеки Dask ... 166
7.4.1. Удаление строк-дубликатов ... 169
7.4.2. Приведение признаков к правильному типу представления ... 170
7.4.3. Заполнение пропусков значений ... 170
7.4.4. Преобразования категориальных признаков в числовое представление ... 171

7.4.5. Поиск и обработка аномальных (нетипичных) значений ... 171
7.4.6. Масштабирование значений числовых признаков ... 171
7.4.7. Создание новых признаков, если это необходимо ... 172
7.4.8. Сокращение количества признаков ... 172
7.4.9. Создание конвейера, объединяющего все операции предобработки ... 173
7.5. Несколько слов о библиотеке polars ... 178
7.6. Общие рекомендации ... 179

Глава 8. Оценка эффективности модели: как отличить «хорошую» модель от «плохой» ... 181
8.1. Оценка эффективности модели с помощью библиотеки scikit-learn ... 183
8.1.1. Базовые метрики для классификации ... 183
8.1.2. Метрики для классификации при дисбалансе классов ... 198
8.1.3. Метрики для регрессии ... 202
8.1.4. Метрики для кластеризации ... 208
8.1.5. Калибровка вероятностей ... 214
8.1.6. Статистическое сравнение моделей ... 226
8.2. Оценка эффективности модели с помощью библиотеки Yellowbrick ... 241
8.2.1. Визуализаторы для классификации ... 242
8.2.2. Визуализаторы для регрессии ... 249
8.2.3. Визуализаторы для кластеризации ... 252
8.2.4. Визуализаторы для анализа целевой переменной ... 257
8.2.5. Выводы и перспективы ... 258
8.3. Конформное предсказание ... 258
8.3.1. Конформное предсказание для классификации ... 260
8.3.2. Конформное предсказание для регрессии ... 264

ЧАСТЬ V
ОТ ОБУЧЕНИЯ ДО ОПТИМИЗАЦИИ МОДЕЛИ

Глава 9. Создание, обучение и применение модели ... 272
9.1. Создание, обучение и применение моделей классификации с помощью библиотеки scikit-learn ... 273
9.1.1. Модели классификации ... 273
9.1.2. Модели классификации по многим меткам ... 291
9.1.3. Модели многозадачной классификации ... 293
9.1.4. Модели регрессии ... 294
9.1.5. Модели кластеризации ... 297
9.1.6. Модели для обнаружения аномалий ... 301

9.2. Создание, обучение и применение моделей с помощью библиотеки XGBoost ... 307
9.2.1. Модели классификации ... 308
9.2.2. Модели регрессии ... 315
9.3. Другие возможности XGBoost и альтернативные библиотеки ... 320
9.4. Глубокое обучение и архитектура модели TabNet ... 321
9.5. Создание, обучение и применение моделей с помощью библиотеки TabNet ... 326
9.5.1. Модели классификации ... 326
9.5.2. Модели регрессии ... 331

Глава 10. Оптимизация гиперпараметров модели ... 335
10.1. Классическая оптимизация с помощью библиотеки scikit-learn ... 335
10.1.1. Перекрестная проверка ... 336
10.1.2. Настройка гиперпараметров ... 341
10.1.3. Анализ модели с помощью проверочных и обучающих кривых ... 348
10.2. Рекомендации по настройке некоторых параметров моделей ... 351
10.3. Байесовская оптимизация с помощью сторонних библиотек ... 352
10.3.1. Оптимизация с помощью Optuna ... 352
10.3.2. Оптимизация в Hyperopt ... 357

ЧАСТЬ VI
ОТ РАЗВЕРТЫВАНИЯ ДО МОНИТОРИНГА МОДЕЛИ

Глава 11. Развертывание модели на конечных точках с помощью FastAPI ... 363
11.1. Почему FastAPI? Введение и преимущества ... 364
11.2. Первые шаги: от Hello, World! до работающего API ... 364
11.3. Подготовка модели и данных для развертывания ... 369
11.4. Создание и тестирование API для предсказаний ... 370
11.5. Расширение API: загрузка файлов и пакетная обработка ... 374
11.6. Темы для дальнейшего изучения ... 377

Глава 12. Мониторинг эффективности модели на конечной точке ... 378
12.1. Основные понятия и типы дрейфа ... 379
12.1.1. Дрейф данных ... 380
12.1.2. Сдвиг целевой переменной ... 381
12.1.3. Дрейф концепции ... 382
12.1.4. Сдвиг совместного распределения или области ... 382
12.2. Способы обнаружения дрейфа ... 383
12.2.1. Выявление дрейфа данных и сдвига целевой переменной ... 383
12.2.2. Способы обнаружения дрейфа концепции ... 386

12.3. Практика работы с дрейфом с помощью библиотеки Evidently ... 390
12.3.1. Обзор библиотеки Evidently ... 390
12.3.2. Обнаружение дрейфа данных ... 390
12.3.3. Мониторинг дрейфа ... 401

ЧАСТЬ VII
ОБЪЯСНЕНИЕ ЛОГИКИ ОБУЧЕННОЙ МОДЕЛИ

Глава 13. Интерпретация принятия решений моделью ... 409
13.1. Важность признаков как первый инструмент интерпретируемости моделей ... 410
13.2. Использование библиотек LIME и SHAP для интерпретации предсказаний моделей ... 415
13.2.1. Применение LIME для локальной интерпретации ... 415
13.2.2. Использование SHAP для локальной и глобальной интерпретации ... 421
13.3. Краткое обобщение: случай регрессии ... 429

ЧАСТЬ VIII
«ОБЕРТКА» ДЛЯ ВСЕГО, ЧТО ВЫ ИЗУЧИЛИ В КНИГЕ

Глава 14. Управление жизненным циклом моделей с помощью MLflow ... 433
14.1. Применение MLflow на практике ... 434
14.1.1. Подготовка данных для модели ... 434
14.1.2. Обучение, оптимизация, журналирование модели ... 437
14.1.3. Оценка производительности, интерпретация и мониторинг ... 439
14.1.4. Итоговая структура проекта: папки ... 440
14.2. Дополнительные возможности MLflow ... 442

Послесловие ... 447

ПРИЛОЖЕНИЯ

Приложение А. Источники данных ... 450
Приложение Б. Основные типы данных ... 452
Приложение В. Различные представления данных ... 455
Приложение Г. Систематика задач машинного обучения ... 456

Список рекомендуемых источников ... 461