

БНТУ

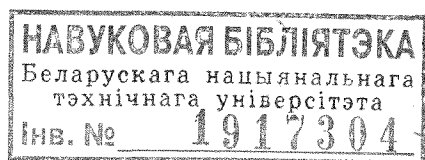
Научная библиотека



Машинное обучение на табличных данных

XGBoost, глубокое обучение и ИИ

МАРК РАЙАН, ЛУКА МАССАРОН
ПРЕДИСЛОВИЕ АНТОНИО ГУАЛИ



305/2)

 ПИТЕР®

Санкт-Петербург • Москва • Минск

2026

Краткое содержание

Часть 1

Понятие о машинном обучении для табличных данных

Глава 1. Табличные данные	32
Глава 2. Табличные датасеты	45
Глава 3. Машинное и глубокое обучение	99

Часть 2

Машинное обучение и градиентный бустинг для табличных данных

Глава 4. Классические алгоритмы для табличных данных	130
Глава 5. Деревья решений и градиентный бустинг	178
Глава 6. Продвинутое методы обработки признаков	245
Глава 7. Полный пример применения XGBoost	302

Часть 3

Глубокое обучение для анализа табличных данных

Глава 8. С чего начать?	365
Глава 9. Лучшие практики глубокого обучения	395

6

Глава 10.	Развертывание модели.....	431
Глава 11.	Пайплайн машинного обучения.....	473
Глава 12.	Комбинация градиентного бустинга и глубокого обучения.....	524
Приложение А.	Гиперпараметры для классических моделей машинного обучения.....	543
Приложение Б.	Методы k ближайших соседей и опорных векторов.....	547

Оглавление

От издательства	18
О научном редакторе русского издания	18
Предисловие.....	19
Введение	21
Благодарности	22
Об этой книге	24
Для кого эта книга	24
Структура книги	25
О коде в книге	26
Форум liveBook	28
Об авторах	29
Иллюстрация на обложке	30

Часть 1

Понятие о машинном обучении для табличных данных

Глава 1. Табличные данные	32
1.1. Что такое табличные данные?	33
1.2. Весь мир работает на табличных данных	35

8 Оглавление

1.3. Сравнение машинного и глубокого обучения.....	36
1.4. В чем особенность табличных данных?	38
1.5. Генеративный ИИ и табличные данные.....	41
Итоги	44
Глава 2. Табличные датасеты.....	45
2.1. Свойства строк и столбцов	46
2.1.1. Как должна выглядеть идеальная строка таблицы	47
2.1.2. Как должен выглядеть идеальный столбец таблицы	52
2.1.3. Представление строк и столбцов.....	55
2.2. Возможные проблемы и способы их устранения.....	57
2.2.1. Столбцы с нулевой или почти нулевой дисперсией	59
2.2.2. Дублирующиеся и высококоллинеарные признаки	59
2.2.3. Нерелевантные признаки.....	62
2.2.4. Пропуски.....	62
2.2.5. Редкие категории	63
2.2.6. Ошибки в данных.....	64
2.2.7. Утечка данных	64
2.3. Поиск данных вовне и внутри организации	65
2.3.1. Использование pandas для доступа к хранилищам данных.....	66
2.3.2. Данные из интернета.....	71
2.3.3. Синтетические данные	74
2.4. Разведочный анализ данных.....	78
2.4.1. Загрузка датасета Auto MPG.....	80
2.4.2. Изучение меток, значений, распределений.....	82
2.4.3. Изучение двумерных и многомерных взаимосвязей.....	92
Итоги.....	97
Глава 3. Машинное и глубокое обучение.....	99
3.1. Прогнозирование цен на Airbnb в Нью-Йорке.....	100

3.1.1. Датасет Airbnb в Нью-Йорке	100
3.1.2. Знакомство с кодом	103
3.1.3. Решение для глубокого обучения на Keras	105
3.1.4. Обучение моделей	105
3.1.5. Сравнение градиентного бустинга и глубокого обучения	107
3.1.6. Выводы	115
3.2. Прозрачность	115
3.2.1. Объяснимость	116
3.2.2. Важность признаков	119
3.2.3. Выводы	121
3.3. Эффективность	121
3.3.1. Оценка производительности	122
3.4. Если копнуть глубже	123
Итоги	128

Часть 2

Машинное обучение и градиентный бустинг для табличных данных

Глава 4. Классические алгоритмы для табличных данных	130
4.1. Знакомство с scikit-learn	131
4.1.1. Общие особенности пакетов scikit-learn	132
4.1.2. Общий интерфейс scikit-learn	134
4.1.3. Введение в пайплайны scikit-learn	137
4.2. Исследование и обработка признаков датасета Airbnb NYC	139
4.2.1. Исследование датасета	139
4.2.2. Подготовка пайплайнов	147
4.3. Классическое машинное обучение	149
4.3.1. Линейная и логистическая регрессии	152
4.3.2. Методы регуляризации	157

4.3.3. Логистическая регрессия	162
4.3.4. Обобщенные линейные методы	166
4.3.5. Обработка больших датасетов с помощью стохастического градиентного спуска.....	169
4.3.6. Выбор алгоритма	174
Итоги.....	175
Глава 5. Деревья решений и градиентный бустинг.....	178
5.1. Знакомство с алгоритмами на основе деревьев.....	179
5.1.1. Бэггинг и семплирование.....	186
5.1.2. Прогнозирование с помощью случайных лесов.....	191
5.1.3. Использование особо случайных деревьев	194
5.2. Градиентный бустинг	196
5.2.1. Как работает градиентный бустинг	199
5.2.2. Экстраполяция с градиентным бустингом.....	202
5.2.3. Почему градиентный бустинг эффективен?	207
5.3. Бустинг в scikit-learn.....	210
5.3.1. Применение ранней остановки для предотвращения переобучения	212
5.4. Использование XGBoost.....	216
5.4.1. Основные параметры XGBoost	218
5.4.2. Как работает XGBoost	223
5.4.3. Разбиение на основе гистограмм	226
5.4.4. Применение ранней остановки в XGBoost.....	229
5.5. Введение в LightGBM	231
5.5.1. Как LightGBM строит деревья.....	234
5.5.2. Увеличение скорости с помощью эксклюзивного объединения признаков и одностороннего семплирования на основе градиента	236
5.5.3. Применение ранней остановки в LightGBM.....	238

5.5.4. Заставляем XGBoost работать, как LightGBM	240
5.5.5. Как LightGBM повлиял на scikit-learn	240
Итоги	242
Глава 6. Продвинутые методы обработки признаков	245
6.1. Особенности обработки признаков	246
6.1.1. Многомерная импутация пропущенных значений	248
6.1.2. Обработка пропущенных данных с помощью GBDT	252
6.1.3. Кодирование по целевой переменной	253
6.1.4. Преобразование числовых данных	259
6.2. Отбор признаков	267
6.2.1. Стабильный отбор предикторов для линейных моделей	268
6.2.2. Фоновые признаки и Boruta	270
6.2.3. Прямое включение и обратное исключение	273
6.3. Оптимизация гиперпараметров	276
6.3.1. Системный поиск	277
6.3.2. Случайный поиск	279
6.3.3. Снижение вычислительной нагрузки	281
6.3.4. Расширение поиска с помощью байесовских методов	283
6.3.5. Ручная настройка гиперпараметров	288
6.4. Освоение градиентного бустинга	290
6.4.1. Выбор между XGBoost и LightGBM	290
6.4.2. Исследование структуры деревьев	291
6.4.3. Ускорение с помощью GBDT и компиляции	296
Итоги	299
Глава 7. Полный пример применения XGBoost	302
7.1. Подготовка и исследование данных	303
7.1.1. Использование генеративного ИИ для подготовки данных	303

7.1.2. Получение и подготовка данных	304
7.1.3. Генерация более сложных признаков.....	313
7.1.4. Завершение обработки данных	321
7.1.5. Исследование и исправление данных.....	322
7.1.6. Изучение целевой переменной.....	327
7.2. Создание и оптимизация модели	329
7.2.1. Подготовка стратегии кросс-валидации	330
7.2.2. Подготовка пайплайна	332
7.2.3. Построение базовой модели	336
7.2.4. Создание первой предварительной модели	342
7.2.5. Оптимизация модели.....	344
7.2.6. Обучение финальной модели	348
7.3. Объяснение модели с помощью SHAP	351
Итоги.....	361

Часть 3

Глубокое обучение для анализа табличных данных

Глава 8. С чего начать?	365
8.1. Стек технологий для глубокого обучения на табличных данных.....	366
8.2. PyTorch с fastai	371
8.2.1. Ключевые особенности кода решения на fastai	371
8.2.2. Сравнение решений на fastai и Keras	379
8.3. PyTorch с TabNet.....	380
8.3.1. Ключевые особенности кода решения на TabNet	380
8.3.2. Сравнение решений на TabNet и Keras.....	383
8.4. PyTorch с Lightning Flash	384
8.4.1. Ключевые особенности решения на Lightning Flash	385
8.4.2. Сравнение решений на Lightning Flash и Keras	388
8.5. Сравнение всех стеков.....	389

8.6. Стеки, которые остались за кадром	390
Итоги	393
Глава 9. Лучшие практики глубокого обучения	395
9.1. Знакомство с данными о недвижимости Куала-Лумпура	396
9.2. Обработка датасета	401
9.2.1. Обработка столбцов Bathrooms, Car Parks, Furnishing, Property Type и Location	401
9.2.2. Обработка столбца Price	403
9.2.3. Обработка столбца Rooms	404
9.2.4. Обработка столбца Size	407
9.3. Определение модели глубокого обучения	411
9.3.1. Сравнение подходов с пользовательскими слоями и слоями предварительной обработки Keras	412
9.3.2. Изучение кода для определения модели с использованием слоев предварительной обработки Keras	416
9.4. Обучение модели глубокого обучения	422
9.4.1. Кросс-валидация в процессе обучения	423
9.4.2. Регуляризация в процессе обучения	423
9.4.3. Нормализация в процессе обучения	424
9.5. Применение модели глубокого обучения	425
9.5.1. Тестирование обученной модели на новых данных	425
9.5.2. Тестирование обученной модели на новых точках данных	427
Итоги	429
Глава 10. Развертывание модели	431
10.1. Простое развертывание в веб-приложении	432
10.1.1. Общая схема веб-развертывания	432
10.1.2. Модуль сервера Flask	434
10.1.3. Страница home.html	437

10.1.4. Страница show-prediction.html	442
10.1.5. Веб-развертывание	443
10.2. Публичные облачные сервисы и операции машинного обучения	444
10.3. Начало работы с Google Cloud	445
10.3.1. Первый вход в Google Cloud	445
10.3.2. Создание проекта Google Cloud	447
10.3.3. Создание бакета Google Cloud Storage	448
10.4. Развертывание модели в Vertex AI	450
10.4.1. Загрузка модели в бакет Cloud Storage	450
10.4.2. Импорт модели в Vertex AI	452
10.4.3. Развертывание модели на эндпоинте	456
10.4.4. Начальное тестирование развертывания модели	458
10.5. Использование развертывания Vertex AI с Flask	461
10.5.1. Настройка Vertex AI SDK	462
10.5.2. Обновление модуля сервера Flask для вызова эндпоинта	462
10.5.3. Преимущества развертывания модели на эндпоинте	465
10.6. Gemini для Google Cloud: помощь генеративного ИИ в Google Cloud	467
10.6.1. Настройка Gemini для Google Cloud	468
10.6.2. Использование Gemini для ответов на вопросы о Google Cloud	468
Итоги	471
Глава 11. Пайплайн машинного обучения	473
11.1. Знакомство с пайплайнами машинного обучения	474
11.1.1. Три вида пайплайнов	474
11.1.2. Обзор ML-пайплайнов Vertex AI	476
11.2. Этапы подготовки ML-пайплайна	477
11.2.1. Создание сервисного аккаунта для ML-пайплайна	477
11.2.2. Создание ключа сервисного аккаунта	480

11.2.3. Предоставление сервисному аккаунту права доступа к базовому сервисному аккаунту Compute Engine.....	482
11.2.4. Знакомство с Cloud Shell	484
11.2.5. Загрузка ключа сервисного аккаунта	486
11.2.6. Загрузка очищенного датасета в бакет Google Cloud Storage	488
11.2.7. Создание управляемого набора данных Vertex AI	490
11.3. Определение ML-пайплайна	494
11.3.1. Что выбрать: локальную реализацию или ML-пайплайн	495
11.3.2. Знакомство с контейнерами.....	497
11.3.3. Преимущества контейнеризации в ML-пайплайне	498
11.3.4. Адаптация кода для запуска в контейнере.....	498
11.3.5. Изменение кода обучения для работы в контейнере	500
11.3.6. Скрипт пайплайна	502
11.3.7. Тестирование модели, обученной в пайплайне	505
11.4. Использование генеративного ИИ для создания ML-пайплайна	507
11.4.1. Использование Gemini для ответов на вопросы об ML-пайплайне	508
11.4.2. Использование Gemini для генерации кода ML-пайплайна	510
11.4.3. Использование Gemini для объяснения кода ML-пайплайна.....	513
11.4.4. Использование Gemini для анализа записей логов	514
11.4.5. Дообучение базовой модели в Vertex AI.....	519
Итоги	522
Глава 12. Комбинация градиентного бустинга и глубокого обучения.....	524
12.1. Обзор решения с градиентным бустингом из главы 7	525
12.2. Выбор решения на основе глубокого обучения	531
12.3. Решение задачи Airbnb Токио с помощью глубокого обучения.....	532

12.4. Сравнение решений на XGBoost и fastai для задачи Airbnb Токио	535
12.5. Ансамблирование двух решений задачи Airbnb Токио	538
12.6. Общее сравнение градиентного бустинга и глубокого обучения	541
Итоги	542
Приложение А. Гиперпараметры для классических моделей машинного обучения	543
Приложение Б. Методы k ближайших соседей и опорных векторов	547
Б.1. Алгоритм k ближайших соседей (k-NN)	548
Б.2. Метод опорных векторов (SVM)	552
Б.3. Использование графических процессоров для машинного обучения	555