

ОТ ПРОФЕССИОНАЛОВ
ДЛЯ ПРОФЕССИОНАЛОВ

kaggle™ 

Книга Kaggle

Машинное обучение
и анализ данных



Конрад Банахевич,
Лука Массарон

Packt >

Конрад Банахевич,
Лука Массарон

Книга Kaggle

Машинное обучение
и анализ данных

Санкт-Петербург
«БХВ-Петербург»
2024

УДК 004.43
ББК 32.973.26-018.1
Б23

Банахевич, К.

Б23 Книга Kaggle. Машинное обучение и анализ данных / К. Банахевич, Л. Массарон: Пер. с англ. — СПб.: БХВ-Петербург, 2024. — 448 с.: ил.

ISBN 978-5-9775-1903-8

Книга рассказывает о продвинутых приёмах машинного обучения и науки о данных (data science) на основе задач, решаемых на всемирно известной платформе Kaggle. Показано (в том числе на примере увлекательных интервью с Kaggle-гроссмейстерами), как устроена платформа Kaggle и проводимые на ней соревнования. Изложенный материал позволяет развить необходимые навыки и собрать портфолио по машинному обучению, анализу данных, обработке естественного языка, работе с множествами. Подобран уникальный пул задач, охватывающих различные классификационные и оценочные метрики, методы обучения нейронных сетей, схемы валидации, выявление паттернов и трендов в материале любой сложности.

Для специалистов по анализу данных и машинному обучению

УДК 004.43
ББК 32.973.26-018.1

Группа подготовки издания:

Руководитель проекта	<i>Олег Сивченко</i>
Зав. редакцией	<i>Людмила Гауль</i>
Редактор	<i>Анна Кузьмина</i>
Компьютерная верстка	<i>Натальи Смирновой</i>
Оформление обложки	<i>Зои Канторович</i>

© Packt Publishing 2022. First published in the English language under the title 'The Kaggle Book – (9781801817479)'
Впервые опубликовано на английском языке под названием 'The Kaggle Book – (9781801817479)'

Подписано в печать 08.11.23.
Формат 70×100^{1/16}. Печать офсетная. Усл. печ. л. 36,12.
Тираж 1300 экз. Заказ № 8042.
"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.
Отпечатано с готового оригинал-макета
ООО "Принт-М", 142300, М.О., г. Чехов, ул. Полиграфистов, д. 1

ISBN 978-1-80181-747-9 (англ.)
ISBN 978-5-9775-1903-8 (рус.)

© Packt Publishing, 2022
© Перевод на русский язык, оформление.
ООО "БХВ-Петербург", ООО "БХВ", 2024

Содержание

Предисловие	11
Составители	13
Об авторах.....	13
О рецензентах	14
О респондентах.....	15
Введение	18
Для кого эта книга	19
О чем эта книга	19
Часть I. Знакомство с соревнованиями	19
Часть II. Оттачивание соревновательных навыков	20
Часть III. Использование соревнований в своей карьере	21
Как получить максимальную отдачу от этой книги.....	21
Загрузите файлы с примерами кода.....	21
Загрузите цветные изображения	22
Условные обозначения и соглашения	22
Часть I. Знакомство с соревнованиями Kaggle	23
Глава 1. Знакомство с Kaggle и другими соревнованиями по науке о данных	25
Появление и рост соревновательных платформ	26
Соревновательная платформа Kaggle.....	28
История Kaggle.....	28
Другие конкурсные платформы.....	31
Знакомство с Kaggle.....	33
Стадии соревнования	33
Типы соревнований и примеры	36

Отправка решения и таблица результатов	41
Парадигма каркаса для общих задач	41
Что может пойти не так	42
Вычислительные ресурсы	44
Kaggle Notebooks	45
Создание команд и нетворкинг	46
Уровни и рейтинг	49
Критика и возможности	50
Резюме	51
Глава 2. Организация данных	53
Создание датасета	53
Сбор данных	57
Работа с датасетами	62
Kaggle Datasets и Google Colab	63
Юридические вопросы	65
Резюме	66
Глава 3. Работаем и учимся с Kaggle Notebooks	67
Создание блокнота	68
Запуск блокнота	71
Сохранение блокнотов на GitHub	73
Как получить максимум от Kaggle Notebooks	75
Переход на Google Cloud Platform	76
На шаг дальше	77
Курсы Kaggle Learn	82
Резюме	86
Глава 4. Используем форумы	87
Как работают форумы	87
Примеры обсуждений	92
Сетевой этикет	97
Резюме	97
Часть II. Оттачивание соревновательных навыков	99
Глава 5. Задачи и метрики на соревнованиях	101
Метрики оценивания и целевые функции	102
Основные типы задач	103
Регрессия	104
Классификация	104
Задачи ранжирования	105

Датасет Meta Kaggle	105
Как быть с незнакомыми метриками	108
Метрики для задач регрессии	112
Средний квадрат и R-квадрат	112
Среднеквадратичная ошибка	113
Среднеквадратичная логарифмическая ошибка	114
Средняя абсолютная ошибка	115
Метрики для задач классификации	116
Доля правильных ответов	116
Точность и полнота	118
F-мера	120
Log Loss и ROC-AUC	120
Коэффициент корреляции Мэтьюса	122
Метрики для многоклассовой классификации	123
Метрики для задач детектирования объектов	129
Отношение площадей ограничивающих рамок	131
Коэффициент Дайса	132
Метрики для многоклассовой классификации и построение рекомендаций	133
MAP@{K}	133
Оптимизация метрики	134
Нестандартные метрики и целевые функции	135
Постобработка предсказаний	138
Предсказание вероятностей и их корректировка	139
Резюме	143
Глава 6. Построение схемы валидации	144
Подглядывание	144
Почему важна валидация	147
Смещение и разброс	150
Стратегии разделения данных	152
Контроль на отложенных данных	153
Вероятностные методы оценки качества	154
Контроль по k блокам	154
Случайные разбиения	162
Бутстрэп	162
Настройка системы валидации	166
Применение adversarial validation	169
Пример реализации	171
Различные распределения обучающих и тестовых данных	172
Работа с утечками в данных	176
Резюме	180

Глава 7. Моделирование для табличных данных	182
Tabular Playground Series	183
Начальное состояние случайного генератора и воспроизводимость	186
Разведочный анализ данных.....	188
Понижение размерности методами t-SNE и UMAP.....	190
Уменьшение размера данных.....	191
Преобразования признаков.....	193
Простые производные признаки.....	194
Метапризнаки на основе строк и столбцов.....	196
Целевое кодирование.....	197
Важность признаков и оценка качества.....	202
Псевдометки	205
Удаление шума с помощью автокодировщиков	207
Нейросети для табличных конкурсов.....	210
Резюме.....	216
Глава 8. Оптимизация гиперпараметров	218
Базовые методы оптимизации.....	219
Поиск по сетке.....	219
Случайный поиск	221
Поиск сокращением вдвое.....	222
Ключевые параметры и их использование.....	225
Линейные модели.....	225
Машины опорных векторов	225
Случайные леса и экстремально рандомизированные деревья	227
Градиентный бустинг над деревьями.....	228
LightGBM.....	228
XGBoost	230
CatBoost.....	231
HistGradientBoosting.....	232
Байесовская оптимизация.....	235
Использование Scikit-optimize	236
Настройки байесовской оптимизации.....	241
Обобщение байесовской оптимизации на параметры нейронных сетей	248
Создание моделей с KerasTuner.....	256
Подход TPE и Optuna.....	265
Резюме	270
Глава 9. Ансамбли: блендинг и стекинг	271
Краткое введение в ансамблевые алгоритмы	272
Усреднение	275
Голосование.....	277
Усреднение предсказаний	279

Взвешенные средние.....	280
Усреднение и кросс-валидация.....	281
Корректируем усреднение для оценок ROC-AUC.....	282
Блендинг и метамоделли.....	283
Блендинг: лучшие практики.....	284
Стекинг.....	289
Варианты стекинга.....	293
Сложные решения с блендингом и стекингом.....	294
Резюме.....	297
Глава 10. Моделирование в компьютерном зрении.....	299
Стратегии аугментации.....	299
Встроенные аугментации Keras.....	305
Подход на основе <i>ImageDataGenerator</i>	305
Слой предварительной обработки.....	308
Пакет <i>alumentations</i>	309
Классификация.....	312
Обнаружение объектов.....	319
Семантическая сегментация.....	333
Резюме.....	349
Глава 11. Моделирование для обработки естественного языка.....	350
Анализ тональности текста.....	350
Вопросы и ответы в открытом домене.....	359
Стратегии аугментации текста.....	374
Основные приемы.....	375
Пакет <i>npraug</i>	380
Резюме.....	383
Глава 12. Соревнования по моделированию и оптимизации.....	384
Игра Connect X.....	385
Игра "Камень, ножницы, бумага".....	390
Соревнование Santa 2020.....	393
Такие разные игры.....	397
Резюме.....	402
Часть III. Использование соревнований в своей карьере.....	403
Глава 13. Создание портфолио проектов и идей.....	405
Создание портфолио с помощью Kaggle.....	405
Использование блокнотов и обсуждений.....	410
Использование датасетов.....	413

Организация своего присутствия в Интернете за пределами Kaggle	417
Блоги и публикации	418
GitHub	421
Мониторинг обновлений и информационных бюллетеней о соревнованиях	423
Резюме	425
Глава 14. Поиск новых профессиональных возможностей	426
Налаживание связей с другими исследователями данных на соревнованиях	427
Участие в Kaggle Days и других встречах Kaggle	438
Привлечение к себе внимания и другие возможности трудоустройства	439
Методика STAR	440
Резюме (и несколько напутственных слов)	442
Предметный указатель	444