

«Искусственный интеллект – это новое электричество».

– *Эндрю Ин* (Andrew Ng), сооснователь Coursera,  
профессор и внештатный преподаватель  
Стэнфордского университета

# Содержание

От издательства .....	13
Кто мы? .....	14
О переводчике .....	16
Введение .....	17
<b>Часть I. ПОДГОТОВКА К РАСШИРЕННОМУ АНАЛИЗУ ДАННЫХ</b> .....	21
<b>Глава 1. Введение. Основы и не только</b> .....	22
1.1. Введение в промежуточный анализ данных .....	23
1.1.1. Ключевые концепции промежуточного анализа данных .....	26
1.1.2. Пример: промежуточный анализ данных с помощью Pandas и NumPy .....	32
1.1.3. Заполнение пропущенных значений .....	36
1.1.4. Вычисление скользящих средних .....	40
1.1.5. Оптимизация типов данных .....	42
1.1.6. Ключевые выводы .....	46
1.2. Путь от простого к сложному .....	47
1.2.1. От простых техник манипулирования данными к более сложным .....	47
1.2.2. Промежуточный уровень манипулирования данными .....	49
1.2.3. Построение эффективных рабочих процессов .....	50
1.2.4. Использование библиотеки NumPy для повышения производительности .....	53
1.3. Pandas, NumPy и Scikit-learn в действии .....	55
1.3.1. Pandas: манипулирование табличными данными на экспертном уровне .....	56
1.3.2. NumPy: высокоэффективные числовые вычисления .....	60
1.3.3. Использование инструментов NumPy для преобразований .....	64
1.3.4. Scikit-learn: эксперт в области машинного обучения .....	66
1.3.5. Почему Scikit-learn? .....	68
1.3.6. Собираем все вместе: полный рабочий процесс .....	69
1.3.7. Ключевые выводы .....	71
1.4. Практические упражнения .....	72
1.5. Возможные проблемы .....	76
1.5.1. Неэффективное манипулирование данными в Pandas .....	76

1.5.2. Неправильная обработка пропущенных значений.....	77
1.5.3. Неправильное применение масштабирования и преобразования признаков .....	78
1.5.4. Неправильное использование конвейеров Scikit-learn .....	78
1.5.5. Неправильная интерпретация результатов модели в Scikit-learn .....	79
1.5.6. Узкие места в операциях NumPy .....	79
1.5.7. Избыточное конструирование признаков.....	80
Заключение .....	80

## **Глава 2. Оптимизация потоков данных .....**

2.1. Расширенное манипулирование данными с Pandas .....	82
2.1.1. Сложная фильтрация и извлечение подмножеств .....	85
2.1.2. Многоуровневая группировка с агрегацией .....	93
2.1.3. Сводные таблицы и изменение структуры данных .....	96
2.1.4. Эффективный анализ временных рядов.....	101
2.1.5. Оптимизация производительности и использования памяти.....	105
2.2. Повышение производительности при помощи массивов NumPy.....	109
2.2.1. Работа с массивами в NumPy .....	109
2.2.2. Векторизованные операции: скорость и простота.....	113
2.2.3. Транслирование: гибкие операции с массивами .....	114
2.2.4. Работа с памятью: типы данных в NumPy .....	116
2.2.5. Многомерные массивы: работа со сложными структурами данных .....	118
2.3. Комбинирование инструментов для выполнения эффективного анализа данных.....	121
2.3.1. Шаг 1: предварительная обработка данных с помощью Pandas и NumPy .....	121
2.3.2. Шаг 2: конструирование признаков с помощью Pandas и NumPy .....	125
2.3.3. Шаг 3: построение модели машинного обучения с помощью Scikit-learn.....	127
2.3.4. Шаг 4: оптимизация рабочих процессов с помощью конвейеров Scikit-learn.....	129
2.4. Практические упражнения.....	132
2.5. Возможные проблемы .....	137
2.5.1. Большие накладные расходы при работе с большими наборами данных в Pandas .....	137
2.5.2. Игнорирование или неправильное использование векторизации в NumPy.....	138
2.5.3. Утечка информации в конвейерах Scikit-learn .....	138
2.5.4. Чрезмерная надежда на значения гиперпараметров модели по умолчанию .....	139
2.5.5. Излишняя сложность конвейеров.....	139
Заключение .....	140

<b>Контрольный опрос. Часть I. Подготовка данных для дальнейшего анализа.....</b>	<b>141</b>
---	------------

## **Часть II. КОНСТРУИРОВАНИЕ ПРИЗНАКОВ ДЛЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ..... 145**

### **Проект 1. Предсказание стоимости домов с помощью конструирования признаков ..... 146**

Исследование переменных и очистка данных .....	146
Конструирование признаков .....	151
Построение и оценка предсказательной модели .....	157
Итоги проекта .....	161
Дальнейшие улучшения .....	162

### **Глава 3. Роль конструирования признаков в машинном обучении ... 163**

3.1. Почему так важно конструировать признаки? .....	163
3.1.1. Области влияния признаков на качество моделей .....	164
3.2. Примеры эффективного конструирования признаков .....	172
3.2.1. Создание переменных взаимодействия .....	173
3.2.2. Создание признаков на основе временных рядов .....	176
3.2.3. Разбиение числовых переменных на интервалы .....	185
3.2.4. Кодирование на основе целевой переменной .....	189
3.3. Практические упражнения .....	193
3.4. Возможные проблемы .....	197
3.4.1. Переобучение из-за слишком большого количества признаков.....	197
3.4.2. Мультиколлинеарность .....	198
3.4.3. Утечка информации.....	198
3.4.4. Неправильная интерпретация признаков на основе времени .....	199
3.4.5. Неподобающее масштабирование признаков.....	199
3.4.6. Недооценка знаний о предметной области .....	200
Заключение .....	201

### **Глава 4. Сложные техники заполнения пропусков в данных ..... 202**

4.1. Использование продвинутых техник заполнения пропущенных значений..	202
4.1.1. Подстановка с помощью метода k-ближайших соседей .....	203
4.1.2. Метод множественной подстановки с помощью цепных уравнений (MICE) .....	208
4.1.3. Использование моделей машинного обучения для подстановки.....	212
4.2. Обработка пропущенных значений в больших наборах данных .....	216
4.2.1. Оптимизация техник подстановки с целью обеспечения масштабируемости.....	217
4.2.2. Обработка столбцов с большим количеством пропущенных значений ....	224
4.2.3. Использование распределенных вычислительных систем для заполнения пропусков .....	229
4.2.4. Ключевые выводы.....	238
4.3. Практические упражнения.....	240
4.4. Возможные проблемы .....	245
4.4.1. Погрешность модели при неправильной подстановке пропусков .....	245
4.4.2. Переобучение модели вследствие замены пропусков в тестовой выборке .....	246

4.4.3. Удаление слишком большого количества данных.....	246
4.4.4. Неправильное интерпретирование данных о временных рядах.....	247
4.4.5. Вычислительная сложность при работе с большими наборами данных ....	247
4.4.6. Сложности с нахождением шаблонов в пропущенных значениях .....	248
Заключение .....	248

## **Глава 5. Преобразование и масштабирование признаков**..... 250

5.1. Масштабирование и нормализация: оптимальное применение .....	250
5.1.1. Почему так важны масштабирование и нормализация.....	251
5.1.2. Масштабирование и нормализация: в чем разница? .....	252
5.1.3. Минимаксное масштабирование (нормализация).....	253
5.1.4. Стандартизация (z-нормализация).....	256
5.1.5. Когда использовать минимаксное масштабирование, а когда стандартизацию .....	259
5.1.6. Робастное масштабирование, устойчивое к выбросам .....	260
5.1.7. Винсоризация .....	264
5.2. Логарифм, квадратный корень и другие нелинейные преобразования признаков.....	267
5.2.1. Логарифмическое преобразование .....	268
5.2.2. Преобразование квадратного корня .....	270
5.2.3. Преобразование кубического корня.....	273
5.2.4. Преобразования Бокса–Кокса и Йео–Джонсона .....	276
5.3. Практические упражнения.....	282
5.4. Возможные проблемы .....	286
5.4.1. Неправильный выбор метода преобразования .....	286
5.4.2. Неправильное масштабирование тестовых данных .....	287
5.4.3. Излишнее преобразование признаков.....	288
5.4.4. Неправильная интерпретация результатов логарифмического преобразования.....	288
5.4.5. Игнорирование природы нелинейных зависимостей .....	289
5.4.6. Неправильное обращение с выбросами.....	289
Заключение .....	290

## **Глава 6. Кодирование категориальных переменных**..... 291

6.1. Кодирование с одним активным состоянием: углубленное изучение.....	291
6.1.1. Совет 1: избегайте ловушки, связанной с фиктивными переменными ....	294
6.1.2. Совет 2: правильно кодируйте значения в столбцах с высокой кардинальностью .....	296
6.1.3. Совет 3: используйте разреженные матрицы для повышения эффективности .....	305
6.1.4. Выводы и рекомендации.....	308
6.2. Более сложные примеры применения кодирования на основе целевой переменной, частоты и порядкового кодирования .....	309
6.2.1. Кодирование на основе целевой переменной с регуляризацией и без ...	310
6.2.2. Пример использования кодирования на основе частоты.....	315
6.2.3. Порядковое кодирование .....	318
6.2.4. Выводы и рекомендации.....	322

6.3. Практические упражнения .....	323
6.4. Возможные проблемы .....	326
6.4.1. Переобучение при использовании кодирования на основе целевой переменной .....	327
6.4.2. Неправильное использование порядкового кодирования .....	327
6.4.3. Использование кодирования с одним активным состоянием для столбцов с высокой кардинальностью .....	328
6.4.4. Пренебрежение разреженностью матрицы при кодировании с одним активным состоянием .....	328
6.4.5. Утечка информации при использовании кодирования на основе целевой переменной .....	329
6.4.6. Ошибочная интерпретация результатов кодирования на основе частоты .....	329
Заключение .....	330

<b>Глава 7. Конструирование признаков и переменных взаимодействия .....</b>	<b>331</b>
7.1. Создание признаков на основе существующих переменных .....	331
7.1.1. Математические преобразования переменных .....	331
7.1.2. Извлечение компонентов из дат .....	337
7.1.3. Комбинирование признаков .....	340
7.2. Переменные взаимодействия и значимость признаков для моделей .....	342
7.2.1. Полиномиальные признаки .....	343
7.2.2. Перекрестные признаки .....	345
7.2.3. Переменные взаимодействия и нелинейные зависимости .....	351
7.2.4. Комбинирование полиномиальных и перекрестных признаков .....	354
7.3. Практические упражнения .....	358
7.4. Возможные проблемы .....	362
7.4.1. Переобучение модели при использовании избыточного количества признаков .....	362
7.4.2. Возникновение мультиколлинеарности .....	362
7.4.3. Добавление в модель избыточных признаков .....	363
7.4.4. Ошибки при интерпретации перекрестных признаков .....	364
7.4.5. Проблемы с производительностью при использовании полиномиальных признаков в больших наборах данных .....	364
Заключение .....	365

<b>Контрольный опрос. Часть II. Конструирование признаков для сложных моделей .....</b>	<b>366</b>
---	------------

<b>Часть III. Очистка и предварительная обработка данных .....</b>	<b>371</b>
--	------------

<b>Проект 2. Прогнозирование временных рядов с конструированием признаков .....</b>	<b>372</b>
Введение в прогнозирование временных рядов с использованием конструирования признаков .....	373

Признаки на основе временного лага в прогнозировании временных рядов .....	373
Признаки на основе скользящего окна для обнаружения трендов и сезонности .....	377
Циклические признаки на основе гармонических функций .....	381
Часовые пояса и пропущенные значения во временных рядах .....	382
Детрендирование и работа с сезонностью во временных рядах .....	386
Что такое детрендирование? .....	386
Методы детрендирования временных рядов .....	387
Работа с сезонностью во временных рядах .....	394
Как детрендирование и выделение сезонности влияют на качество моделей .....	397
Применение методов из семейства ARIMA и алгоритмов машинного обучения для прогнозирования временных рядов .....	397
Шаг 1. Подготовка данных для алгоритма машинного обучения .....	398
Шаг 2. Применение методов прогнозирования временных рядов .....	403
Шаг 2(б). Применение методов машинного обучения для прогнозирования временных рядов .....	424
Подбор гиперпараметров для методов машинного обучения .....	434
Что такое гиперпараметры? .....	435
Использование поиска по сетке для подбора гиперпараметров .....	436
Использование случайного поиска для подбора гиперпараметров .....	438
Итоги проекта .....	440
Особенности развертывания моделей прогнозирования временных рядов .....	441
Практические упражнения .....	442
Возможные проблемы .....	444
Утечка информации в результате неправильно созданных признаков .....	444
Неправильно выбранный размер окна при создании скользящих признаков .....	445
Пропуски, возникающие в результате создания новых признаков .....	445
Неправильная интерпретация циклических переменных .....	446
Разреженность данных при создании скользящих признаков .....	446
Неправильный учет часовых поясов в данных .....	447
<b>Глава 8. Корректировка аномалий в данных при помощи Pandas</b> ....	448
8.1. Обработка некорректных форматов данных .....	449
8.2. Поиск и удаление дубликатов .....	455
8.3. Исправление неконсистентных категориальных данных .....	459
8.4. Обработка значений, выходящих за допустимые границы .....	463
8.5. Обработка пропущенных значений, образовавшихся в результате коррекции аномалий .....	467
8.6. Практические упражнения .....	469
8.7. Возможные проблемы .....	473
8.7.1. Удаление важных наблюдений вместе с выбросами .....	473
8.7.2. Чрезмерная стандартизация категориальных данных .....	474
8.7.3. Ошибочное интерпретирование дубликатов .....	474

8.7.4. Ошибочное удаление значений, выходящих за границы диапазона.....	475
8.7.5. Ошибки, появляющиеся в результате автоматической стандартизации .....	475
8.7.6. Ошибки в результате подстановки пропущенных значений .....	476
Заключение .....	476

## **Глава 9. Методы снижения размерности .....**

9.1. Анализ главных компонент (PCA).....	477
9.1.1. Суть анализа главных компонент .....	479
9.1.2. Реализация анализа главных компонент при помощи Scikit-learn .....	483
9.1.3. Объясненная дисперсия и анализ главных компонент .....	487
9.1.4. Когда стоит применять анализ главных компонент .....	491
9.1.5. Ключевые выводы об анализе главных компонент .....	492
9.2. Техники отбора признаков.....	493
9.2.1. Методы фильтрации .....	494
9.2.2. Оберточные методы .....	501
9.2.3. Встроенные методы.....	506
9.2.4. Ключевые выводы о техниках отбора признаков.....	511
9.3. Практические упражнения.....	512
9.4. Возможные проблемы .....	516
9.4.1. Удаление слишком большого количества признаков.....	516
9.4.2. Опасности использования только методов фильтрации .....	516
9.4.3. Утечка информации при использовании оберточных методов .....	517
9.4.4. Чрезмерные штрафы при использовании встроенных методов .....	517
9.4.5. Ошибки при интерпретации главных компонент .....	518
9.4.6. Избыточность данных при использовании техник отбора признаков....	519
Заключение .....	519

## **Контрольный опрос. Часть III. Очистка и предобработка данных .....**

### **Заключение .....**

### **Предметный указатель.....**



# От издательства

## ***Отзывы и пожелания***

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте [www.dmkpress.com](http://www.dmkpress.com), зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com); при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу [http://dmkpress.com/authors/publish\\_book/](http://dmkpress.com/authors/publish_book/) или напишите в издательство по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com).

## ***Список опечаток***

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com). Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

## ***Нарушение авторских прав***

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты [dmkpress@gmail.com](mailto:dmkpress@gmail.com).

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

# Кто мы?

Вы держите в руках книгу, ставшую плодом совместных усилий разработчиков из компании Quantum Technologies. Мы – команда людей, преданных идее создания программного обеспечения, наделенного творческим началом и помогающего решать реальные задачи. Мы стремимся создавать высококачественные и дружелюбные веб-приложения, отвечающие всем требованиям наших заказчиков.

Мы в нашей команде верим, что программирование не ограничивается исключительно написанием кода. Задача программирования состоит в решении насущных проблем и облегчении жизни людей. С завидным постоянством исследуем новые технологии и науки, чтобы оставаться в авангарде индустрии, и с удовольствием делимся накопленными знаниями с вами, наши читатели.

Наш подход к созданию приложений базируется на совместной работе и творческом начале. Мы работаем в тесном контакте с заказчиками, чтобы досконально понять их нужды и создать продукт, отвечающий всем их требованиям. Верим в то, что выходящие из-под пера программиста приложения должны быть интуитивно понятными, легкими в обращении и привлекательными визуально, и делаем все от нас зависящее для воплощения этой веры.

В этой книге мы постарались изложить практические рекомендации по подготовке данных для их интеллектуального анализа. Делаете ли вы свои первые шаги в программировании или уже имеете определенный опыт в написании приложений, эта книга поможет вам освоить приемы работы с данными, которые позволят вам в будущем погрузиться в мир машинного и глубокого обучения.

## *Наша философия*

В Quantum Technologies мы свято верим, что учиться программировать и развивать свои навыки в написании приложений можно на протяжении всей жизни. В связи с этим мы всячески поощряем стремление наших разработчиков осваивать новые технологии и техники и обеспечиваем их всеми необходимыми инструментами, для того чтобы не терять позиций в быстро развивающейся отрасли. Кроме того, мы полагаем, что программирование должно приносить радость и удовлетворение, и стараемся создавать рабочее окружение, поощряющее творческий подход и внедрение инноваций.

## ***Наш опыт***

Компания Quantum Technologies специализируется на создании веб-приложений, помогающих решать заказчикам их насущные проблемы творчески и эффективно. Наши разработчики обладают богатым опытом написания приложений на разных языках программирования и использования различных фреймворков и технологий, включая Python, AI, ChatGPT, Django, React, Three.js, Vue.js и пр. Мы постоянно осваиваем новые технологии и стараемся реализовывать на практике все пожелания клиентов.

Также наши разработчики имеют все необходимые навыки в области анализа данных, визуализации, машинного обучения и искусственного интеллекта. Мы убеждены, что именно за этими технологиями будущее индустрии, и делаем все возможное, чтобы оставаться на переднем крае этой революции.

# О переводчике



**Александр Гинько**, обладающий богатым опытом работы в сфере ИТ и более десяти лет посвятивший переводам книг и статей на самые разные темы, в последние годы специализируется на переводе книг в области бизнес-аналитики и программирования для издательства «ДМК Пресс» по направлениям Python, SQL, Power BI, DAX, Excel, Power Query, Tableau, R... На данный момент в активе Александра уже более 25 книг, включая одну авторскую, и он продолжает плодотворно работать над переводом и написанием новых книг.

Возможно, вам также будут интересны книги «Сверхбыстрый Python» (<https://dmkpress.com/catalog/computer/programming/python/978-5-93700-226-6>), «Python: практическое руководство по Pandas (200 упражнений)» (<https://dmkpress.com/catalog/computer/programming/python/978-5-93700-227-3>) и «Введение в статистическое обучение с примерами на Python» (<https://dmkpress.com/catalog/computer/statistics/978-5-93700-217-4>) в переводе Александра.

Помимо перевода книг, Александр ведет свой канал в Telegram ([https://t.me/alexanderginko\\_books](https://t.me/alexanderginko_books)), на котором вы можете из первых уст получить ответы на все интересующие вас вопросы об уже переведенных книгах, находящихся в работе и запланированных на будущее. Также на канале можно найти промокоды на все книги Александра для покупки книг на сайте издательства «ДМК Пресс» с большими скидками. Купить книги Александра и следить за переводом новых книг в режиме реального времени можно также с помощью его бота в Telegram по адресу [https://t.me/alexanderginko\\_books\\_bot](https://t.me/alexanderginko_books_bot).

# Введение

Данные – это один из наиболее ценных активов в современном цифровом мире, и они лежат в основе всего: от принятия бизнес-решений до осуществления технического прогресса. В то же время сырые необработанные данные зачастую содержатся в разрозненном, беспорядочном и неструктурированном виде. Истинная ценность данных кроется в их сути, добраться до которой можно только путем их преобразований. Но для выполнения этих преобразований недостаточно просто знать нужные алгоритмы. Нужно хорошо понимать приемы и способы эффективной очистки, подготовки и манипулирования данными. В этой книге мы поговорим о **ключевых концепциях и техниках**, связанных с анализом данных, а также с конструированием и отбором признаков, лежащих в основе машинного и глубокого обучения.

Цель этой книги – научить вас подготавливать и преобразовывать сырые данные, конструировать новые признаки и в целом придавать исходным данным форму, пригодную для будущего интеллектуального анализа при помощи методов машинного и глубокого обучения. Работаете ли вы с небольшими объемами данных или оперируете сложными наборами данных высокой размерности, эта книга познакомит вас с эффективными техниками и приемами предварительной обработки и подготовки данных к дальнейшему анализу. По большей части мы будем пользоваться богатыми средствами и инструментами наиболее популярных библиотек Python для работы с данными, таких как Pandas, NumPy и Scikit-learn.

## ***Почему так важно подготавливать данные и заниматься построением признаков?***

В машинном обучении часто можно услышать фразу «Данные – наше всё». Хотя выбор модели и тщательная настройка используемых алгоритмов играют важную роль, качество исходных данных оказывает гораздо большее влияние на эффективность итоговой модели. Этап подготовки данных и конструирования признаков зачастую занимает наиболее продолжительное время в проекте, и на то есть веские причины. Искусно подготовленные данные позволяют модели выявлять значимые шаблоны и делать точные предсказания, а также улучшают ее обобщающую способность при применении к новым данным.

*Конструирование признаков* (feature engineering) относится к области создания новых информативных признаков на основе сырых данных и играет важнейшую роль в процессе интеллектуального анализа данных. Именно *признакам* (feature) – преобразованным, комбинированным или созданным

на основе существующих данных – модели обязаны своим высоким предсказательным потенциалом. В процессе чтения книги вы увидите, что хорошо сконструированные признаки способны определять зависимости и шаблоны, которые невозможно выявить на основе одних только исходных данных. Такие тщательно отобранные признаки лежат в основе моделей, характеризующихся высоким качеством предсказаний, устойчивостью и интерпретируемостью результатов.

## ***Инструменты, которые мы будем использовать***

В последнее время практическим стандартом для специалистов по работе с данными стал язык программирования Python, и в этой книге мы будем в основном использовать три наиболее популярные библиотеки этого языка для преобразования данных: Pandas, NumPy и Scikit-learn.

*Pandas* – мощная библиотека для анализа и манипулирования данными. Pandas представляет собой интуитивно понятный фреймворк для управления исходными данными, представленными в виде строк и столбцов. Он особенно полезен в задачах первичной обработки данных, таких как очистка, фильтрация, агрегация и объединение наборов данных. Pandas поможет вам значительно облегчить процедуру обработки сырых данных и извлечения из них важных сведений.

*NumPy* – высокоэффективная библиотека для быстрой работы с массивами и применения математических операций к большим наборам данных. Преимущества NumPy в работе с данными делают эту библиотеку незаменимой при выполнении ресурсоемких операций с большими массивами информации, таких как масштабирование, нормализация и математические преобразования.

*Scikit-learn* – будучи одной из наиболее популярных библиотек в области машинного обучения, Scikit-learn не ограничивается одними лишь инструментами для построения моделей, но также предлагает богатый арсенал средств для манипулирования данными. В модулях фреймворка, связанных с предварительной обработкой данных, присутствуют средства для кодирования категориальных переменных, масштабирования непрерывных переменных, создания конвейеров обработки данных и многого другого. Библиотека Scikit-learn играет ключевую роль на этапе предварительной обработки данных, обеспечивая вашему анализу согласованность и воспроизводство.

В совокупности перечисленные библиотеки дают вам полный набор средств для управления процессом подготовки данных и конструирования признаков, включая очистку и преобразование данных, а также отбор и кодирование переменных.

## ***Чему вы научитесь***

Главы этой книги разбиты на три части, каждая из которых посвящена отдельному этапу подготовки данных и конструирования признаков.

**Часть I. Подготовка к расширенному анализу данных.** В первой части книги мы познакомимся с базовыми принципами промежуточного анализа данных и закроем пробелы в знаниях, касающихся обработки данных средствами Python. Мы научимся подходить к обработке данных системно и обеспечивать полную пригодность и структурированность данных перед их дальнейшим анализом. В процессе мы пройдемся по основным возможностям библиотек Pandas и NumPy и научимся выполнять требуемые операции над данными максимально эффективно. Таким образом, в этой части мы заложим основы для знакомства с более сложными техниками обработки данных, о которых будем говорить далее.

**Часть II. Конструирование и отбор признаков для повышения качества моделей.** Во второй части книги мы с головой погрузимся в область конструирования и отбора признаков. Мы познакомимся с эффективными способами управления пропущенными значениями, масштабирования и преобразования признаков, кодирования категориальных переменных и создания новых признаков. Область конструирования признаков требует творческого подхода и глубокого понимания поставленной задачи, и в этой части книги мы научимся выстраивать мыслительный процесс и использовать наиболее подходящие техники для повышения предсказательной способности будущей модели. Мы обсудим множество полезных методов для создания полиномиальных признаков, комбинирования существующих переменных с целью отслеживания эффектов их взаимодействий и кодирования категориальных переменных с использованием разных стратегий. Завершив чтение этой части книги, вы будете вооружены полным спектром техник и приемов для конструирования и отбора признаков, которые сможете полноценно использовать в собственных проектах.

**Часть III. Очистка и предварительная обработка данных.** В заключительной части книги мы сосредоточимся на критически важных задачах, связанных с очисткой данных и их предварительной обработкой. Здесь вы узнаете о передовых техниках обработки выбросов в данных, корректировки аномалий и подготовки данных для анализа временных рядов. Мы также обсудим способы снижения размерности данных, такие как метод главных компонент, незаменимые при работе с многомерными данными. Вы узнаете, как можно уменьшить сложность пространства признаков без существенного вреда для качества модели. Это позволит повысить эффективность итоговой модели и ее интерпретируемость. Освоив эти техники предварительной обработки данных, вы сможете строить хорошо структурированные наборы данных, что существенно повысит качество создаваемых моделей.

## ***Практические примеры и реальные задачи***

Каждая глава книги сопровождается примерами и упражнениями, которые позволят вам проверить полученные знания на практике в самых разных областях – от финансов до здравоохранения и розничных продаж. Столь разные примеры помогут вам понять, какие преобразования, способы кодирования

и масштабирования могут успешно применяться в тех или иных областях. Кроме того, эти примеры помогут вам критически взглянуть на конструирование признаков и научиться применять только нужные преобразования в зависимости от набора данных и итоговой модели.

Также в конце каждой главы вы встретите раздел с названием «Что может пойти не так?», в котором мы будем обсуждать распространенные ловушки и трудности, поджидающие вас в процессе конструирования и отбора признаков. Эти разделы призваны помочь вам развить критическое мышление и всегда находиться на шаг впереди, предвосхищая возможные трудности.

## ***Важность воспроизводимости***

В науке о данных воспроизводимость является одним из ключевых аспектов в построении надежных и качественных моделей. В книге мы часто будем обращаться к теме воспроизводимости, в частности при обсуждении использования конвейеров в Scikit-learn. Автоматизация шагов по преобразованию данных внутри конвейеров позволит вам применять одни и те же действия к обучающему и тестовому наборам данных и тем самым избежать утечки информации и повысить надежность результатов.

## ***Заключение***

Книга, которую вы держите в руках, является подробным руководством по освоению ключевых навыков, необходимых для подготовки данных и конструирования и отбора признаков. Эти аспекты составляют основу любого успешного проекта с применением машинного и глубокого обучения, и у вас есть прекрасная возможность овладеть ими в полной мере.

Делаете ли вы свои первые шаги в освоении науки о данных или являетесь практикующим специалистом в этой области, желающим улучшить свои навыки, эта книга поможет вам обрести уверенность при работе с большими объемами данных.

Что ж, давайте вместе отправимся в это увлекательное путешествие и научимся из сырых данных готовить восхитительные полуфабрикаты для построения успешных моделей машинного обучения!



Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

[e-Univers.ru](http://e-Univers.ru)