

# Оглавление

<b>Предисловие .....</b>	<b>18</b>
От издательства .....	20
<b>Глава 1. Введение .....</b>	<b>21</b>
1.1. Прогнозирование и интерпретация .....	24
1.2. Ключевые ингредиенты предиктивных моделей.....	25
1.3. Терминология .....	27
1.4. Примеры наборов данных и типичные сценарии данных.....	29
Музыкальный жанр .....	29
Заявки на получение грантов .....	30
Поражение печени.....	31
Проницаемость .....	32
Производство химикатов.....	32
Мошенничество в финансовых отчетах .....	33
Сравнения наборов данных.....	34
1.5. Структура книги.....	36
1.6. Условные обозначения .....	38

## **ЧАСТЬ I ОБЩИЕ СТРАТЕГИИ**

<b>Глава 2. Краткий обзор процесса предиктивного моделирования.....</b>	<b>42</b>
2.1. Пример прогнозирования экономии топлива.....	42
2.2. Аспекты, заслуживающие отдельного рассмотрения .....	48

Разделение данных .....	48
Данные предикторов.....	48
Оценка эффективности .....	49
Оценка нескольких моделей.....	49
Выбор модели .....	49
2.3. Итоги .....	50
<b>Глава 3.</b> Предварительная обработка данных.....	51
3.1. Практический пример: сегментация клеток в высокопроизводительном скрининге .....	53
3.2. Преобразования данных для отдельных предикторов .....	55
Центрирование и масштабирование .....	55
Преобразования для устранения смещения.....	55
3.3. Преобразования данных с несколькими предикторами .....	58
Преобразования для решения проблемы выбросов .....	58
Прореживание данных и выделение признаков.....	60
3.4. Отсутствующие значения .....	66
3.5. Удаление предикторов.....	69
Корреляции между предикторами .....	71
3.6. Добавление предикторов.....	73
3.7. Группировка предикторов .....	75
3.8. Вычисления .....	77
Преобразования.....	79
Фильтрация.....	81
Создание фиктивных переменных .....	83
Упражнения.....	85
<b>Глава 4.</b> Переобучение и настройка модели .....	87
4.1. Проблема переобучения .....	88
4.2. Настройка модели.....	90
4.3. Разделение данных.....	93
4.4. Методы повторной выборки .....	96
К-кратная перекрестная проверка .....	96

Обобщенная перекрестная проверка.....	98
Повторное разделение тренировочного/тестового набора .....	98
Бутстрэп .....	99
4.5. Практикум: оценка кредитоспособности .....	101
4.6. Выбор итоговых параметров настройки .....	101
4.7. Рекомендации по разделению данных.....	105
4.8. Выбор между моделями .....	106
4.9. Вычисления .....	108
Разделение данных .....	109
Повторная выборка.....	110
Базовый процесс построения модели в R .....	111
Определение параметров настройки.....	112
Сравнение моделей.....	116
Упражнения.....	118

## **ЧАСТЬ II РЕГРЕССИОННЫЕ МОДЕЛИ**

<b>Глава 5.</b> Измерение эффективности регрессионных моделей .....	122
5.1. Количественные показатели эффективности .....	122
5.2. Обратное отношение между смещением и дисперсией.....	124
5.3. Вычисления .....	126
<b>Глава 6.</b> Модели с признаками линейной регрессии .....	128
6.1. Практикум: моделирование количественного соотношения «структура— активность».....	129
6.2. Линейная регрессия.....	135
Линейная регрессия для данных растворимости .....	139
6.3. Частные наименьшие квадраты.....	140
Применение методов PCR и PLSR для прогнозирования данных растворимости .....	144
Алгоритмические разновидности PLS .....	148
6.4. Штрафные модели.....	150

---

6.5. Вычисления .....	156
Обычная линейная регрессия.....	157
Частные наименьшие квадраты.....	162
Штрафные регрессионные модели .....	163
Упражнения.....	166
<b>Глава 7. Нелинейные регрессионные модели .....</b>	<b>169</b>
7.1. Нейросети.....	169
7.2. Многомерные адаптивные регрессионные сплайны.....	174
7.3. SVM, или метод опорных векторов .....	180
7.4. Метод k ближайших соседей .....	188
7.5. Вычисления .....	191
Нейросети.....	191
Многомерные адаптивные регрессионные сплайны.....	193
SVM, метод опорных векторов.....	196
Метод KNN .....	198
Упражнения.....	198
<b>Глава 8. Древоподобные модели. Модели на базе правил.....</b>	<b>202</b>
8.1. Базовые деревья регрессии .....	204
8.2. Деревья регрессионных моделей .....	214
8.3. Модели на базе правил.....	221
8.4. Бэггинг-деревья.....	224
8.5. Случайные леса .....	230
8.6. Усиление .....	235
8.7. Модель Cubist.....	241
8.8. Вычисления .....	246
Простые деревья.....	246
Деревья моделей .....	247
Деревья бэггинга .....	248
Случайные леса .....	248

---

Усиленные деревья .....	249
Модель Cubist.....	250
Упражнения.....	250
<b>Глава 9.</b> Обзор моделей растворимости.....	254
<b>Глава 10.</b> Практический пример: сопротивление сжатию бетонных смесей .....	257
10.1. Стратегия построения модели.....	261
10.2. Эффективность моделей .....	262
10.3. Оптимизация сопротивления сжатию .....	265
10.4. Вычисления .....	269

### **ЧАСТЬ III КЛАССИФИКАЦИОННЫЕ МОДЕЛИ**

<b>Глава 11.</b> Определение эффективности в классификационных моделях.....	278
11.1. Прогнозы классов .....	278
Хорошо откалиброванные вероятности .....	280
Представление вероятностей классов .....	282
Неоднозначные зоны .....	284
11.2. Оценка прогнозируемых классов.....	286
Задача двух классов.....	288
Критерии, не основанные на точности .....	292
11.3. Оценка вероятностей классов .....	294
ROC-кривые .....	295
Диаграммы точности прогнозов .....	297
11.4. Вычисления .....	299
Чувствительность и специфичность .....	301
Матрица несоответствий .....	301
ROC-кривые .....	302
Диаграммы точности прогнозов .....	303
Калибровка вероятностей .....	304

---

<b>Глава 12. Дискриминантный анализ и другие линейные классификационные модели</b> .....	307
12.1. Практикум: прогнозирование успешных заявок на получение грантов .....	307
12.2. Логистическая регрессия .....	315
12.3. Линейный дискриминантный анализ (LDA).....	320
12.4. Дискриминантный анализ методом частных наименьших квадратов.....	331
12.5. Штрафные модели .....	337
12.6. Ближайшие сжатые центроиды .....	341
12.7. Вычисления.....	344
Логистическая регрессия .....	347
Линейный дискриминантный анализ .....	353
Дискриминантный анализ методом частных наименьших квадратов.....	355
Штрафные модели .....	357
Метод ближайших сжатых центроидов.....	359
Упражнения.....	362
<b>Глава 13. Нелинейные классификационные модели</b> .....	365
13.1. Нелинейный дискриминантный анализ.....	365
Квадратичный и регуляризованный дискриминантный анализ.....	365
Смешанный дискриминантный анализ.....	367
13.2. Нейросети.....	369
13.3. Гибкий дискриминантный анализ .....	374
13.4. SVM, метод опорных векторов.....	380
13.5. Метод KNN .....	389
13.6. Наивный байесовский классификатор .....	391
13.7. Вычисления.....	396
Нелинейный дискриминантный анализ.....	397
Нейросети.....	398
FDA.....	400
Модель SVM .....	401
Модель KNN (к ближайших соседей) .....	403

---

Наивный байесовский классификатор .....	403
Упражнения.....	405
<b>Глава 14.</b> Деревья классификации и модели на базе правил.....	407
14.1. Базовые деревья классификации .....	408
14.2. Модели на базе правил .....	423
Модель C4.5Rules .....	423
Модель PART .....	424
14.3. Бэггинг деревьев.....	425
14.4. Случайные леса .....	427
14.5. Бустинг .....	429
Алгоритм AdaBoost.....	429
Стохастический градиентный бустинг .....	431
14.6. Модель C5.0.....	434
Классификационные деревья .....	435
Правила классификации .....	436
Усиление.....	437
Другие аспекты модели.....	438
Данные грантов .....	440
14.7. Сравнение двух кодировок категориальных предикторов .....	442
14.8. Вычисления .....	443
Классификационные деревья .....	443
Правила .....	447
Деревья с бэггингом.....	449
Случайный лес.....	450
Деревья с усилением .....	451
Упражнения.....	453
<b>Глава 15.</b> Сравнительный анализ моделей для заявок на получение грантов.....	456
<b>Глава 16.</b> Решение проблемы дисбаланса классов .....	460
16.1. Практикум: прогнозирование политики страхования.....	461

16.2. Эффект дисбаланса классов.....	462
16.3. Настройка модели.....	465
16.4. Альтернативные пороги отсечения.....	465
16.5. Корректировка априорных вероятностей .....	468
16.6. Неравные веса .....	469
16.7. Методы выборки .....	469
16.8. Тренировка с учетом стоимости .....	473
16.9. Вычисления .....	478
Альтернативные пороги отсечения.....	482
Методы выборки .....	482
Тренировка с учетом стоимости .....	483
Упражнения.....	486
<b>Глава 17. Практикум: планирование заданий.....</b>	<b>488</b>
17.1. Разделение данных и стратегия модели .....	496
17.2. Результаты.....	498
17.3. Вычисления.....	501

## **ЧАСТЬ IV ПРОЧИЕ ВОПРОСЫ ПРЕДИКТИВНОГО МОДЕЛИРОВАНИЯ**

<b>Глава 18. Определение важности предикторов .....</b>	<b>506</b>
18.1. Числовые результаты.....	507
18.2. Категорийные результаты.....	511
18.3. Прочие методы .....	516
18.4. Вычисления .....	522
Числовые результаты.....	522
Категорийные результаты .....	525
Показатели важности для разных моделей.....	528
Упражнения.....	529
<b>Глава 19. Выбор признаков.....</b>	<b>531</b>
19.1. Последствия использования неинформативных предикторов .....	532



---

19.2. Методы сокращения количества предикторов .....	534
19.3. Методы-обертки.....	535
Прямой, обратный и пошаговый выбор .....	539
Имитация отжига .....	540
Генетические алгоритмы.....	541
19.4. Методы-фильтры .....	544
19.5. Смещение выбора .....	545
19.6. Практикум: прогнозирование когнитивного расстройства .....	548
19.7. Вычисления.....	557
Прямой, обратный и пошаговый выбор .....	558
Рекурсивное исключение признаков .....	560
Методы-фильтры .....	563
Упражнения.....	565
<b>Глава 20. Факторы, влияющие на эффективность модели.....</b>	<b>567</b>
20.1. Ошибки III типа .....	568
20.2. Ошибка измерения результата .....	571
20.3. Погрешность измерений в предикторах .....	572
Практикум: прогнозирование нежелательных побочных эффектов .....	576
20.4. Дискретизация непрерывных результатов.....	578
20.5. Когда следует доверять прогнозу вашей модели? .....	582
20.6. Влияние большой выборки.....	586
20.7. Вычисления.....	588
Упражнения.....	590

## ПРИЛОЖЕНИЯ

<b>Приложение А. Краткая сводка различных моделей .....</b>	<b>596</b>
<b>Приложение Б. Введение в R.....</b>	<b>599</b>
Б.1. Запуск и вывод справочной информации.....	599
Б.2. Пакеты .....	600
Б.3. Создание объектов .....	602

Б.4. Типы данных и базовые структуры .....	602
Б.5. Работа с прямоугольными наборами данных.....	606
Б.6. Объекты и классы.....	609
Б.7. Функции R.....	610
Б.8. Три грани = .....	611
Б.9. Пакет AppliedPredictiveModeling.....	611
Б.10. Пакет caret.....	612
Б.11. Пакеты, используемые в книге.....	615
<b>Приложение В. Рекомендуемые веб-сайты.....</b>	<b>616</b>
<b>Список источников .....</b>	<b>619</b>