

# Содержание

<b>Введение</b>	15
О чем эта книга	15
Для кого предназначена эта книга	16
Используемые пиктограммы	16
Дополнительный материал	17
Что дальше	17
<b>Часть 1. Введение в аналитическое прогнозирование</b>	19
<b>Глава 1. Выход на арену</b>	21
Что такое аналитическое прогнозирование	21
Интеллектуальный анализ данных	22
Создание модели	23
Добавление бизнес-ценности	25
Бесконечные возможности	25
Расширение возможностей организации	26
Начало проекта по аналитическому прогнозированию	28
Знания о бизнесе	28
Группа специалистов по анализу данных и информационным технологиям	29
Данные	30
Аналитическое прогнозирование	31
Формирование группы аналитического прогнозирования	33
Привлечение опытных практиков	33
Инициативность и любознательность	34
Исследование рынка	34
Обработка больших данных	35
Работа с большими данными	35
<b>Глава 2. Аналитическое прогнозирование на практике</b>	39
Интернет-маркетинг и розничная торговля	42
Рекомендательные системы	42
Персонализированные покупки в Интернете	44
Реализация рекомендательной системы	44
Совместная фильтрация	45
Контентная фильтрация	53
Гибридные рекомендательные системы	58
Целевой маркетинг	59

Целевой маркетинг с использованием прогнозного моделирования	60
Моделирование воздействия	62
Персонализация	65
Поведение клиентов в Интернете	65
Перенацеливание	65
Реализация	66
Оптимизация с использованием персонализации	67
Сходство персонализации и рекомендации	68
Контент и анализ текстов	69
<b>Глава 3. Методы исследования данных</b>	<b>71</b>
Распознавание типов данных	72
Структурированные и неструктурированные данные	72
Статические и потоковые данные	77
Определение категорий данных	78
Оценочные данные	80
Поведенческие данные	81
Демографические данные	81
Генерирование моделей аналитического прогнозирования	82
Анализ, ориентированный на данные	83
Анализ, ориентированный на пользователей	85
Связь со смежными дисциплинами	86
Статистика	87
Интеллектуальный анализ данных	88
Машинное обучение	88
<b>Глава 4. Сложность данных</b>	<b>91</b>
Поиск ценности в ваших данных	92
Погружение в данные	93
Разнообразие данных	94
Постоянно меняющиеся данные	95
Скорость передачи данных	95
Большой объем данных	95
Сложности в поиске данных	96
Поиск по ключевым словам	96
Семантический поиск	97
Контекстный поиск	99
Отличие бизнес-аналитики от анализа больших данных	103
Исследование необработанных данных	104
Идентификация атрибутов данных	104
Изучение типичных визуализаций данных	105

Табличные визуализации	105
Облака слов	107
“Подобное притягивает подобное” как принцип представления новых данных	107
Графы	109
Типичные средства визуализации	111
<b>Часть 2. Включение алгоритмов в модели</b>	<b>113</b>
<b>Глава 5. Применение моделей</b>	<b>115</b>
Данные для моделирования	116
Модели и моделирование	117
Классификация моделей	119
Описание и обобщение данных	121
Принятие более эффективных бизнес-решений	121
Примеры аналитики в сфере здравоохранения	122
Проект Google Flu Trends	122
Предикторы выживаемости при раке	124
Социальная и маркетинговая аналитика	126
Сеть магазинов Target предсказывает беременность	126
Предсказание землетрясений на базе социальной сети Twitter	127
Предикторы результатов политической кампании, основанные на Твиттере	129
Твиты как предикторы для фондового рынка	131
Прогнозирование колебаний цен на акции по новостным статьям	132
Анализ использования велосипедов в Нью-Йорке	133
Предсказания и ответы	136
Сжатие данных	137
Прогностика и ее связь с аналитическим прогнозированием	138
Прогностика и обеспечение надежности оборудования	138
Рост использования открытых данных	139
<b>Глава 6. Выявление сходства в данных</b>	<b>141</b>
Объяснение кластеризации данных	142
Обоснование	144
Преобразование необработанных данных в матрицу	146
Создание матрицы терминов в документах	146
Выбор термина	148
Идентификация групп в данных	148
Алгоритм кластеризации K-средних	149
Кластеризация методом ближайших соседей	153
Алгоритмы на основе плотности	156

Поиск ассоциаций в элементах данных	158
Алгоритм Apriori	159
Применение биологически вдохновленных методов кластеризации	162
Стая птиц: алгоритм Flock by Leader	163
Муравьиные колонии	168
<b>Глава 7. Прогнозирование на основе классификации данных</b>	<b>173</b>
Введение в классификацию данных	175
Кредитование	175
Маркетинг	176
Здравоохранение	177
Что дальше?	178
Использование классификации данных в бизнесе	178
Изучение процесса классификации данных	181
Использование классификации данных для прогнозирования будущего	182
Деревья решений	183
Алгоритмы генерации деревьев решений	185
Метод опорных векторов	190
Ансамблевые методы для повышения точности прогноза	192
Наивный байесовский алгоритм классификации	193
Основы наивного байесовского классификатора	194
Марковская модель	198
Линейная регрессия	204
Нейронные сети	204
Глубокое обучение	207
Ренессанс нейронных сетей	207
Введение в глубокое обучение	208
<b>Часть 3. Планирование</b>	<b>213</b>
<b>Глава 8. Как убедить руководство одобрить проект по аналитическому прогнозированию</b>	<b>215</b>
Разработка бизнес-сценария	217
Выгоды для бизнеса	217
Получение поддержки от заинтересованных сторон	225
Работа со спонсорами	226
Одобрение проекта со стороны бизнес-руководства и администрации	228
Одобрение проекта со стороны IT-менеджеров	230
Быстрое создание прототипов	235
Презентация предложения	236

<b>Глава 9. Подготовка данных</b>	239
Перечисление бизнес-целей	240
Определение связанных целей	241
Сбор требований пользователей	242
Обработка данных	242
Идентификация данных	242
Очистка данных	244
Генерация любых производных данных	245
Уменьшение размерности данных	246
Применение анализа главных компонент	247
Использование сингулярного разложения	249
Работа с признаками	251
Выбор признаков	252
Извлечение признаков	254
Ранжирование признаков	255
Структурирование данных	256
Извлечение, преобразование и загрузка данных	256
Поддержание данных в актуальном состоянии	257
Планирование тестирования и организация тестовых данных	258
<b>Глава 10. Создание прогностической модели</b>	261
Начало	262
Определение бизнес-целей	264
Подготовка данных	265
Выбор алгоритма	268
Разработка и тестирование модели	270
Разработка модели	270
Тестирование модели	271
Оценка модели	274
Дальнейшая работа с моделью	275
Развертывание модели	275
Мониторинг и поддержка модели	276
<b>Глава 11. Визуализация аналитических данных</b>	277
Визуализация как инструмент прогнозирования	278
Чем важна визуализация	278
Получение выгоды от визуализации	280
Устранение сложностей	281
Оценка визуализации	282
Насколько релевантная эта картина?	282
Насколько интерпретируема картина?	282

## 10 Содержание

Достаточно ли проста картина?	283
Приводит ли картина к новым плодотворным идеям?	283
Визуализация аналитических результатов моделирования	284
Визуализация скрытых группировок в данных	284
Визуализация результатов классификации данных	284
Визуализация выбросов в данных	286
Визуализация деревьев решений	287
Визуализация прогнозов	288
Новые средства визуализации в аналитическом прогнозировании	290
Алгоритм Flock by Leader для визуализации данных	291
Инструменты визуализации больших данных	295
Tableau	295
Google Charts	296
Plotly	296
Infogram	296

## **Часть 4. Программирование методов аналитического прогнозирования** 297

### **Глава 12. Примеры создания типичных прогностических моделей** 299

Инсталляция программных пакетов	300
Инсталляция интерпретатора Python	300
Инсталляция модуля машинного обучения	303
Инсталляция зависимостей	307
Подготовка данных	311
Получение примера набора данных	311
Разметка данных	311
Прогнозирование с использованием алгоритмов классификации	313
Создание модели обучения с учителем с помощью метода SVM	313
Загрузка данных	314
Обучение модели	315
Создание модели обучения с учителем на основе логистической регрессии	321
Создание модели обучения с учителем на основе случайного леса	327
Сравнение моделей классификации	329

### **Глава 13. Примеры прогнозирования без учителя** 331

Получение примера набора данных	332
Использование алгоритмов кластеризации для прогнозирования	332
Сравнение моделей кластеризации	333
Создание модели обучения без учителя с помощью K-средних	334

Создание модели обучения без учителя с помощью алгоритма DBSCAN	345
Создание модели обучения без учителя с помощью алгоритма сдвига среднего значения	349
<b>Глава 14. Аналитическое прогнозирование на языке R</b>	<b>353</b>
Программирование на языке R	355
Инсталляция интерпретатора языка R	356
Инсталляция среды RStudio	356
Знакомство со средой	357
Немного о языке R	359
Вызов функции	363
Прогнозирование с помощью языка R	364
Прогнозирование с помощью регрессии	364
Использование классификации для прогнозирования	375
Классификация с помощью случайного леса	383
<b>Глава 15. Как избежать ловушек в процессе анализа данных</b>	<b>389</b>
Проблемы, связанные с данными	390
Ограничения, связанные с данными	392
Работа с экстремальными случаями (выбросами)	394
Сглаживание данных	398
Приближение кривой	402
Делайте как можно меньше предположений	405
Проблемы анализа	406
Анализ с учителем	407
Опираясь только на один анализ	407
Описание ограничений модели	408
Избегайте немасштабируемых моделей	410
Точная оценка прогнозов	411
<b>Часть 5. Большие данные</b>	<b>413</b>
<b>Глава 16. Ориентация на большие данные</b>	<b>415</b>
Основные технологические тенденции в аналитическом прогнозировании	416
Изучение аналитического прогнозирования как услуги	417
Агрегирование распределенных данных для анализа	417
Анализ, управляемый данными в реальном времени	419
Применение инструментов с открытым исходным кодом к большим данным	420
Платформа Apache Hadoop	421
Apache Yarn	423
Платформа Apache Spark	427
Основные компоненты платформы Spark	430

<b>Глава 17. Подготовка к анализу данных предприятия</b>	433
Корпоративная архитектура для анализа больших данных	434
Аналитика как услуга	437
Google Analytics	438
Microsoft Revolution R Enterprise	440
Подготовка прототипа по аналитическому прогнозированию	441
Создание прототипов для аналитического прогнозирования	441
Тестирование модели аналитического прогнозирования	445
<b>Часть 6. Великолепные десятки</b>	447
<b>Глава 18. Десять причин для внедрения аналитического прогнозирования</b>	449
Определение бизнес-целей	450
Изучение данных	451
Организация данных	452
Удовлетворение клиентов	453
Сокращение эксплуатационных расходов	454
Увеличение доходности инвестиций (ROI)	455
Получение быстрого доступа к информации	456
Принятие обоснованных решений	457
Получение конкурентного преимущества	458
Улучшение бизнеса	458
<b>Глава 19. Десять шагов к построению модели</b>	461
Создание группы аналитического прогнозирования	462
Получение бизнес-опыта	462
Привлечение IT-специалистов и математиков	462
Постановка бизнес-целей	463
Подготовка данных	464
Выборка данных	464
Избегайте ситуации “мусор на входе, мусор на выходе”	465
Простота — не глупость	465
Подготовка данных — важный фактор успеха	466
Достижение быстрых побед	466
Стимулирование изменений в организации	467
Создание развертываемых моделей	468
Оценка модели	469
Обновление модели	470
<b>Предметный указатель</b>	472