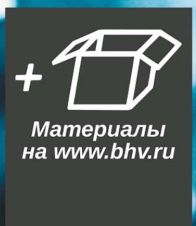
Серг Масис

ИНТЕРПРЕТИРУЕМОЕ МАШИННОЕ ОБУЧЕНИЕ НА PYTHON

Научитесь создавать интерпретируемые высокопроизводительные модели на практических примерах из реальной жизни



УДК 004.43
ББК 32.973-018.1
М31

Масис С.

М31 Интерпретируемое машинное обучение на Python: Пер. с англ. — СПб.: БХВ-Петербург, 2023. — 640 с.: ил.

ISBN 978-5-9775-1735-5

Книга поможет осознанно и эффективно работать с моделями машинного обучения. Дано введение в интерпретацию машинного обучения: раскрыты важность темы, ее ключевые понятия и проблемы. Рассмотрены методы интерпретации: модельно-агностические, якорные и контрфактические, для многопеременного прогнозирования, а также визуализации сверточных нейронных сетей. Раскрыты вопросы настройки на интерпретируемость: отбор и конструирование признаков, ослабление систематического смещения и причинно-следственный вывод, монотонные ограничения, настройка моделей и устойчивость к антагонизму. Показаны перспективы развития интерпретируемых моделей машинного обучения. Каждая глава книги включает подробные примеры исходного кода на языке Python.

На сайте издательства размещен архив с цветными иллюстрациями.

Для программистов в области машинного обучения

УДК 004.43
ББК 32.973-018.1

Научный редактор:
Инженер-разработчик компании КРОК *Анвар Хафизов*

Группа подготовки издания:

Руководитель проекта	*Евгений Рыбаков*
Зав. редакцией	*Людмила Гауль*
Перевод с английского	*Андрея Логунова*
Редактор	*Анна Кузьмина*
Компьютерная верстка	*Натальи Смирновой*
Корректор	*Светлана Крутоярова*
Оформление обложки	*Зои Канторович*

"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

ISBN 978-1-80020-390-7 (англ.)
ISBN 978-5-9775-1735-5 (рус.)

Оглавление

Об авторе **15**

О рецензентах **17**

Предисловие **19**

Для кого эта книга предназначена 20
Что эта книга охватывает 21
Как получить максимальную отдачу от этой книги 23
Загрузка файлов с исходным кодом 24
Загрузка цветных изображений 25
Используемые условные обозначения 25

Часть I. Введение в интерпретацию машинного обучения **27**

Глава 1. Интерпретация, интерпретируемость и объяснимость: почему всё это важно? **29**

Технические требования 30
Что такое интерпретация машинного обучения? 30
    Изучение простой модели предсказания веса 31
Понимание разницы между интерпретируемостью и объяснимостью 37
    Что такое интерпретируемость? 37
    Что такое объяснимость? 40
Деловое обоснование интерпретируемости 42
    Более качественные решения 42
    Более надежные бренды 44
    Более высокий уровень этичности 46
    Более высокая прибыльность 49
Резюме 50
Источники изображений 50
Справочные материалы 50

Глава 2. Ключевые понятия интерпретируемости **52**

Технические требования 52
Миссия 52
    Подробности о сердечно-сосудистых заболеваниях 53

Подход ... 54
Подготовительные работы ... 54
Загрузка библиотек ... 54
Изучение проблемы и подготовка данных ... 54
Ознакомление с типами методов интерпретации и диапазонами интерпретируемости ... 57
Типы методов модельной интерпретации ... 61
Диапазоны модельной интерпретируемости ... 61
Интерпретирование отдельных предсказаний с помощью логистической регрессии ... 62
Оценивание препятствий, мешающих интерпретируемости результатов машинного обучения ... 67
Нелинейность ... 69
Интерактивность ... 72
Немонотонность ... 72
Миссия выполнена ... 74
Резюме ... 75
Справочные материалы ... 75

Глава 3. Трудности интерпретации ... 76

Технические требования ... 76
Миссия ... 76
Подход ... 78
Подготовительные работы ... 78
Загрузка библиотек ... 78
Изучение проблемы и подготовка данных ... 79
Обзор традиционных методов модельной интерпретации ... 84
Предсказывание минут задержки с помощью различных регрессионных методов ... 84
Классифицирование рейсов как задержанных либо незадержанных с использованием различных классификационных методов ... 89
Визуализация задержанных рейсов с помощью методов понижения размерности ... 96
Ограничения традиционных методов модельной интерпретации ... 102
Изучение имманентно интерпретируемых моделей (типа белого ящика) ... 103
Обобщенные линейные модели ... 103
Деревья решений ... 118
RuleFit ... 123
Метод ближайших соседей ... 125
Наивный Байес ... 127
Распознавание компромисса между результативностью и интерпретируемостью ... 130
Особые модельные свойства ... 130
Диагностика результативности ... 131

Обнаружение более новых интерпретируемых (аквариумных) моделей 134
Объяснимая бустинговая машина 134
Skoped-Rules 138
Миссия выполнена 140
Резюме 141
Источник набора данных 141
Справочные материалы 142

Часть II. Освоение методов интерпретации 143

Глава 4. Основы важности признаков и их влияние 145

Технические требования 145
Миссия 146
Личность и очередность рождения 146
Подход 147
Подготовительные работы 147
Загрузка библиотек 147
Изучение проблемы и подготовка данных 148
Как измерить влияние признака на исход 150
Важность признаков в древовидных моделях 154
Важность признаков в логистической регрессии 156
Важность признаков в линейном дискриминантном анализе 159
Важность признаков в многослойном персептроне 161
Применение перестановочной важности признаков на практике 162
Недостатки метода перестановочной важности признаков 165
Интерпретирование графиков частичной зависимости 166
Интеракционные графики частичной зависимости 171
Недостатки графиков частичной зависимости 174
Объяснение графиков индивидуального условного ожидания 174
Недостатки графиков индивидуального условного ожидания 179
Миссия выполнена 179
Резюме 180
Источник набора данных 180
Справочные материалы 180

Глава 5. Модельно-агностические методы глобальной интерпретации 182

Технические требования 182
Миссия 183
Подход 184
Подготовительные работы 185
Загрузка библиотек 185
Изучение проблемы и подготовка данных 186
Значения Шепли 196

Интерпретирование сводки SHAP и графиков зависимости ... 198
Генерирование сводных графиков SHAP ... 202
Изучение взаимодействий ... 204
Графики зависимости SHAP ... 207
Силовые графики SHAP ... 215
Графики накопленных локальных эффектов ... 217
Глобальные суррогаты ... 221
Подгонка суррогатов ... 221
Оценивание суррогатов ... 222
Интерпретирование суррогатов ... 223
Миссия выполнена ... 225
Резюме ... 225
Справочные материалы ... 226

Глава 6. Модельно-агностические методы локальной интерпретации ... 227

Технические требования ... 227
Миссия ... 227
Подход ... 228
Подготовительные работы ... 229
Загрузка библиотек ... 229
Изучение проблемы и подготовка данных ... 230
Задействование ядерного объяснителя SHAP для локальных интерпретаций со значениями SHAP ... 236
Обучение модели C-SVC ... 237
Вычисление значений SHAP с помощью ядерного объяснителя ... 239
Локальная интерпретация для группы предсказаний с использованием графиков решений ... 241
Локальная интерпретация по одному предсказанию за раз с использованием силового графика ... 244
Применение локально интерпретируемых модельно-агностических объяснений ... 247
Что такое LIME? ... 247
Локальная интерпретация по одному предсказанию за раз с использованием табличного объяснителя на основе LIME ... 249
Использование метода LIME для NLP ... 251
Обучение модели LightGBM ... 253
Локальная интерпретация по одному предсказанию за раз с использованием текстового объяснителя на основе LIME ... 254
Опробование SHAP в обработке естественного языка ... 257
Сравнение SHAP с LIME ... 260
Миссия выполнена ... 261
Резюме ... 262
Источник набора данных ... 262
Справочные материалы ... 262

Глава 7. Якорные и контрфактические объяснения ... 264

Технические требования ... 264
Миссия ... 264
Необъективная смещенность в диагностиках риска рецидивизма ... 266
Подход ... 267
Подготовительные работы ... 267
Загрузка библиотек ... 267
Изучение проблемы и подготовка данных ... 268
Якорные объяснения ... 278
Подготовительные работы для якорных и контрафактических объяснений с помощью библиотеки *alibi* ... 279
Локальные интерпретации якорных объяснений ... 281
Анализ контрфактических объяснений ... 284
Контрфактические объяснения под руководством прототипов ... 285
Получение контрфактических экземпляров и многого другого с помощью инструмента What-If Tool (WIT) ... 289
Сравнение с помощью метода контрастивного объяснения ... 299
Миссия выполнена ... 303
Резюме ... 304
Источник набора данных ... 304
Справочные материалы ... 304

Глава 8. Визуализация сверточных нейронных сетей ... 306

Технические требования ... 306
Миссия ... 307
Подход ... 308
Подготовительные работы ... 309
Загрузка библиотек ... 309
Изучение проблемы и подготовка данных ... 310
Диагностика CNN-классификатора традиционными методами интерпретации ... 315
Визуализирование процесса усвоения с помощью активационных методов ... 323
Промежуточные активации ... 325
Максимизация активации ... 328
Оценивание ошибочных классификаций с помощью градиентных методов атрибуции ... 332
Карты значимости ... 333
Метод градиентных карт активаций классов Grad-CAM ... 336
Интегрированные градиенты ... 338
Окончательная сборка ... 341
Объяснение классификаций с помощью пертурбационных методов атрибуции ... 344
Окклюзивная чувствительность ... 344
Объяснитель изображений методом LIME ... 347
Метод контрастивных объяснений ... 349

Окончательная сборка ... 354
Бонусный метод: глубокий объяснитель SHAP ... 357
Миссия выполнена ... 358
Резюме ... 359
Источники данных и изображений ... 359
Справочные материалы ... 360

Глава 9. Методы интерпретации для многопеременного прогнозирования и анализа чувствительности ... 361

Технические требования ... 362
Миссия ... 362
Подход ... 364
Подготовительные работы ... 365
Загрузка библиотек ... 365
Изучение проблемы и подготовка данных ... 366
Диагностика моделей временного ряда с использованием традиционных методов интерпретации ... 375
Использование стандартных регрессионных метрик ... 376
Агрегации предсказательных ошибок ... 378
Оценивание как классификационная задача ... 380
Генерирование LSTM-атрибуций с помощью интегрированных градиентов ... 381
Вычисление глобальных и локальных атрибуций с помощью ядерного объяснителя SHAP ... 387
Зачем использовать ядерный объяснитель? ... 387
Определение стратегии, позволяющей работать с моделью многопеременного временного ряда ... 388
Заложение основы для стратегии аппроксимации перестановок ... 389
Выявление влияющих признаков с помощью факторной приоритизации ... 394
Вычисление индексов чувствительности Морриса ... 395
Анализирование элементарных эффектов ... 398
Квантифицирование неопределенности и стоимостной чувствительности с помощью фиксирования факторов ... 401
Генерирование и предсказывание на образцах Сальтелли ... 402
Выполнение анализа чувствительности по методу Соболя ... 403
Встраивание реалистичной функции стоимости ... 405
Миссия выполнена ... 409
Резюме ... 410
Источники данных и изображений ... 411
Справочные материалы ... 411

Часть III. Настройка на интерпретируемость ... 413

Глава 10. Отбор и конструирование признаков для обеспечения интерпретируемости ... 415

Технические требования ... 416

Миссия ... 416
Подход ... 417
Подготовительные работы ... 418
Загрузка библиотек ... 418
Изучение проблемы и подготовка данных ... 419
Изучение эффекта нерелевантных признаков ... 420
Построение базовой модели ... 421
Оценивание модели ... 422
Обучение базовой модели на разных максимальных глубинах ... 425
Обзор фильтрационных методов отбора признаков ... 427
Базовые фильтрационные методы ... 428
Корреляционные фильтрационные методы ... 430
Ранжирующие фильтрационные методы ... 432
Сравнение фильтрационных методов ... 434
Анализ встроенных методов отбора признаков ... 435
Раскрытие потенциала оберточных, гибридных и продвинутых методов отбора признаков ... 439
Оберточные методы ... 439
Гибридные методы ... 441
Продвинутые методы ... 443
Оценивание всех моделей, построенных с применением отбора признаков ... 445
Обзор конструирования признаков ... 447
Миссия выполнена ... 455
Резюме ... 457
Источники наборов данных ... 457
Справочные материалы ... 457

Глава 11. Ослабление систематического смещения и причинно-следственный вывод ... 459

Технические требования ... 460
Миссия ... 460
Подход ... 461
Подготовительные работы ... 462
Загрузка библиотек ... 462
Изучение проблемы и подготовка данных ... 463
Обнаружение систематического смещения ... 467
Визуализирование систематического смещения набора данных ... 469
Квантифицирование систематического смещения набора данных ... 472
Квантифицирование систематического смещения модели ... 476
Ослабление систематического смещения ... 479
Методы ослабления систематического смещения стадии предварительной обработки ... 480

Методы ослабления систематического смещения стадии промежуточной обработки ... 487
Методы ослабления систематического смещения стадии последующей обработки ... 490
Окончательная сборка ... 493
Построение причинно-следственной модели ... 495
Изучение результатов эксперимента ... 497
Изучение причинно-следственных моделей ... 500
Инициализация линейного дважды устойчивого ученика ... 502
Обучение причинно-следственной модели ... 502
Гетерогенные эффекты экспериментальной процедуры ... 503
Выбор политики ... 507
Проверка устойчивости оценки ... 510
Добавление случайной общей причины ... 510
Замена экспериментальной процедуры случайной переменной ... 511
Миссия выполнена ... 512
Резюме ... 513
Источник набора данных ... 513
Справочные материалы ... 513

Глава 12. Монотонные ограничения и настройка моделей на интерпретируемость ... 515

Технические требования ... 516
Миссия ... 516
Подход ... 518
Подготовительные работы ... 518
Загрузка библиотек ... 519
Изучение проблемы и подготовка данных ... 519
Установка ограничений с помощью конструирования признаков ... 522
Упорядочение ... 523
Дискретизация ... 525
Члены взаимодействия и нелинейные преобразования ... 526
Категориальное кодирование ... 530
Другие подготовительные работы ... 531
Настройка моделей на интерпретируемость ... 532
Настройка нейронной сети Keras ... 533
Настройка других популярных модельных классов ... 536
Оптимизация под объективность с помощью байесовой гиперпараметрической настройки и прикладных метрик ... 544
Имплементирование модельных ограничений ... 550
Ограничения в XGBoost ... 551
Ограничения в TensorFlow Lattice ... 556
Миссия выполнена ... 563

Резюме ... 564
Источник набора данных ... 565
Справочные материалы ... 565

Глава 13. Устойчивость к антагонизму ... 566

Технические требования ... 567
Миссия ... 567
Подход ... 569
Подготовительные работы ... 569
Загрузка библиотек ... 569
Изучение проблемы и подготовка данных ... 570
Загрузка базовой модели CNN ... 573
Диагностика базового классификатора CNN ... 575
Эвазивные атаки ... 576
Атака быстрым методом на основе знака градиента ... 578
Атака методом инфинитной нормы Карлини и Вагнера ... 581
Целенаправленная атака методом антагонистических заплат ... 583
Защита от целенаправленных атак с помощью предобработки ... 585
Защита от любой эвазивной атаки с помощью антагонистического обучения устойчивого классификатора ... 590
Оценивание и сертифицирование устойчивости к антагонизму ... 595
Сравнение устойчивости модели с силой атаки ... 595
Сертифицирование устойчивости с помощью рандомизированного сглаживания ... 597
Миссия выполнена ... 604
Резюме ... 605
Источники наборов данных ... 605
Справочные материалы ... 606

Глава 14. Интерпретируемость машинного обучения: что дальше? ... 607

Современное состояние интерпретируемости машинного обучения ... 607
Связываем всё воедино! ... 607
Текущие тренды ... 612
Размышления о будущем интерпретируемости машинного обучения ... 614
Новое видение машинного обучения ... 615
Междисциплинарный подход ... 616
Соответствующая требованиям стандартизация ... 616
Исполнение регуляторных предписаний ... 616
Бесшовная автоматизация машинного обучения со встроенной интерпретацией ... 617
Более тесная интеграция с инженерами MLOps ... 617
Справочные материалы ... 618

Предметный указатель ... 619

1 Интерпретация, интерпретируемость и объяснимость: почему всё это важно?

Мы живем в мире, правила и процедуры которого регулируются данными и алгоритмами.

Например, существуют правила разрешения на предоставление кредита или освобождения под залог, правила цензуры сообщений в социальных сетях. Существуют также процедуры, позволяющие определять, какая маркетинговая тактика является наиболее эффективной и какие рентгеновские снимки грудной клетки могут доказывать наличие пневмонии.

Вы этого ожидаете, потому что в этом нет ничего нового!

Еще совсем недавно подобные правила и процедуры было принято формализировать в программах, учебниках и на бумажных бланках, хотя последнее слово при принятии решения оставалось за людьми. Зачастую это решение полностью зависело от людей, потому что правила и процедуры были жесткими и, следовательно, не всегда применимыми. *Всегда* были исключения, поэтому для их принятия требовался человек.

Например, если вы подадите заявку на ипотеку, ее одобрение будет зависеть от приемлемой и достаточно длительной кредитной истории. Эти данные позволят рассчитать кредитный балл с использованием алгоритма начисления баллов. В свою очередь, у банка есть правила, определяющие величину балла, который является достаточным для желаемой вами ипотеки. Ваш кредитный инспектор может дать ей ход или же отклонить ее.

В наши дни финансовые учреждения тренируют нейросетевые модели на тысячах ипотечных результатов с десятками переменных. Эти модели можно использовать для определения вероятной возможности дефолта по ипотеке с допускаемой высокой точностью. Если есть кредитный инспектор, который визирует одобрение или отказ, то это уже не просто руководящий принцип, а алгоритмическое решение. Разве оно может быть неверным? А правильным?

Стоит попридержать эту мысль, потому что на протяжении всей книги мы будем изучать ответы на эти и многие другие вопросы!

Интерпретировать решения, принимаемые моделью машинного обучения, значит отыскивать в решении смысл, но, кроме того, можно прослеживать его источник и процесс, который его преобразовал. В этой главе рассматриваются интерпретация машинного обучения и связанные с ней понятия, такие как интерпретируемость, объяснимость, модели типа черного ящика и прозрачность. Данная глава предоставляет определения этих терминов, чтобы можно было избежать двусмысленности, и подкрепляет их объяснением ценности интерпретируемости машинного обучения. Вот главные темы, которые мы собираемся осветить:

- что такое интерпретация машинного обучения;
- понимание разницы между интерпретацией и объяснимостью;
- бизнес-аспекты интерпретируемости.

Давайте начнем!

Технические требования

В целях сверки с примерами по ходу чтения этой главы вам понадобится язык Python 3, работающий в среде Jupyter Notebook либо в вашей любимой интегрированной среде разработки (integrated development environment, IDE), такой как PyCharm, Atom, VSCode, PyDev или Idle. Для запуска примеров также потребуются установленные Python-библиотеки `requests`, `BS4`, `pandas`, `sklearn`, `matplotlib` и `scipy`. Исходный код этой главы находится по адресу: **https://github.com/PacktPublishing/Interpretable-Machine-Learning-with-Python/tree/master/Chapter01**.

Что такое интерпретация машинного обучения?

Интерпретировать что-то — значит *объяснять его смысл*. В контексте машинного обучения это "что-то" является алгоритмом. Если конкретнее, то это математический алгоритм, который принимает данные на входе и выдает данные на выходе, как и в любой формуле.

Давайте рассмотрим самую базовую из моделей — простую линейную регрессию, модель которой проиллюстрирована в следующем уравнении:

$$\hat{y} = \beta_0 + \beta_1 x_1 .$$

После аппроксимации данных этой моделью ее смысл заключается в том, что предсказания $\hat{y}$ представляют собой взвешенную сумму признаков x с коэффициентами β. В данном случае существует только один **признак** x, или **предсказательная** переменная, а переменная $\hat{y}$ обычно называется **откликом**, или **целевой** переменной. Простая линейно-регрессионная модель объясняет преобразование, которое выполняется на данных x_1 на входе для порождения результата $\hat{y}$ на выходе. Следующий ниже пример может проиллюстрировать эту концепцию подробнее.

Изучение простой модели предсказания веса

Если вы зайдете на веб-страницу, поддерживаемую Калифорнийским университетом, **http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights**, то сможете найти ссылку на скачивание набора данных из 25 000 синтетических записей о весе и росте 18-летних юношей и девушек. Мы не будем использовать весь набор данных целиком, а возьмем только образцовую таблицу на самой веб-странице с 200 записями. Мы извлекаем эту таблицу с веб-страницы и выполняем подгонку линейно-регрессионной модели под эти данные. В указанной модели рост используется для предсказания веса индивида.

Другими словами, x_1 — рост, а $\hat{y}$ — вес (вернее, масса тела), поэтому формула линейно-регрессионной модели будет следующей:

$$\text{вес} = \beta_0 + \beta_1 \cdot \text{рост} .$$

Исходный код этого примера находится по адресу:
https://github.com/PacktPublishing/Interpretable-Machine-Learning-with-Python/blob/master/Chapter01/WeightPrediction.ipynb.

В целях выполнения этого примера вам необходимо установить следующие библиотеки:

- `requests` для получения запроса содержания веб-страницы;
- `bs4` (Beautiful Soup) для парсинга таблицы с веб-страницы;
- `pandas` для загрузки таблицы в кадр данных;
- `sklearn` (scikit-learn) для подгонки линейно-регрессионной модели и расчета ее ошибки;
- `matplotlib` для визуализации модели;
- `scipy` для проверки корреляции.

Сначала необходимо импортировать все библиотеки следующим образом:

```
import math
import requests
from bs4 import BeautifulSoup
import pandas as pd
from sklearn import linear_model
from sklearn.metrics import mean_absolute_error
import matplotlib.pyplot as plt
from scipy.stats import pearsonr
```

Библиотека `requests` используется для получения содержимого веб-страницы следующим образом:

```
url = \
'http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights'
page = requests.get(url)
```

Затем надо взять это содержимое и извлечь только таблицу с помощью библиотеки `BeautifulSoup` следующим образом:

```
soup = BeautifulSoup(page.content, 'html.parser')
tbl = soup.find("table",{"class":"wikitable"})
```

Библиотека `pandas` может превратить содержимое таблицы из исходного источника на **языке гипертекстовой разметки** (HyperText Markup Language, HTML) в кадр данных, как показано ниже:

```
height_weight_df = pd.read_html(str(tbl))[0][['Height(Inches)', 'Weight(Pounds)']]
```

И вуаля! Теперь у нас есть кадр данных с ростом в дюймах (`Heights(Inches)`) в одном столбце и весом в фунтах (`Weights(Pounds)`) в другом. В качестве проверки исправности можно подсчитать число записей. Оно должно равняться 200. Исходный код показан ниже:

```
num_records = height_weight_df.shape[0]
print(num_records)
```

Теперь, когда мы подтвердили, что у нас есть данные, необходимо преобразовать их так, чтобы они сочетались со спецификациями модели. Библиотека `sklearn` принимает их как в NumPy-массиве размерностью (200, 1), поэтому для начала необходимо извлечь числовые ряды `Heights(Inches)` и `Weights(Pounds)` из `pandas`. Следующий шаг — превратить их в NumPy-массив и, наконец, преобразовать их размерность для соответствия (200, 1). Следующий код выполняет все необходимые операции преобразования:

```
x = height_weight_df['Height(Inches)'].values.reshape(num_records, 1)
y = height_weight_df['Weight(Pounds)'].values.reshape(num_records, 1)
```

Затем надо инициализировать линейно-регрессионную модель (`LinearRegression`) библиотеки `scikit-learn` и выполнить ее подгонку под тренировочные данные:

```
model = linear_model.LinearRegression()
_ = model.fit(x,y)
```

Работая в библиотеке `scikit-learn`, для вывода на экран формулы линейно-регрессионной модели необходимо извлечь коэффициент β_0 вертикального сдвига (т. е. *y*-координату точки пересечения линии регрессии с вертикальной осью системы координат) и коэффициент наклона линии β_1. Вот **формула**, которая объясняет, как она делает предсказания:

```
print("ŷ =" + str(model.intercept_[0]) + " + " + str(model.coef_.T[0][0]) + " x1 "
```

Результат:

```
ŷ = -106.02770644878132 + 3.432676129271629 x1
```

Он говорит о том, что в среднем на каждый дополнительный фунт приходится 3,4 дюйма роста.

Однако *объяснение алгоритма работы модели* — это всего лишь один из способов объяснения данной линейно-регрессионной модели, и это только одна сторона ис-

тории. Модель не является идеальной, потому что фактические результаты, приведенные в таблице, и предсказанные моделью результаты для тренировочных данных не эквивалентны. Разница между ними называется **ошибкой**, или **остатками**.

Понимать ошибку в модели можно совершенно по-разному. К примеру, можно применить функцию ошибки, такую как функция средней абсолютной ошибки (`mean_absolute_error`), используемую для измерения отклонения по модулю между предсказанными и фактическими значениями, как показано в следующем фрагменте исходного кода:

```
y_pred = model.predict(x)
mae = mean_absolute_error(y, y_pred)
print(mae)
```

Результат:

```
7.7587373803882205
```

Средняя абсолютная ошибка 7,8 означает, что в среднем предсказание отстоит на 7,8 фунта от фактической величины, но эта величина ошибки, возможно, не будет интуитивно понятной или в некоторых случаях даже не будет информативной. Визуализация линейно-регрессионной модели может пролить некоторый свет на то, насколько эти предсказания точны на самом деле.

Это можно сделать с помощью точечной диаграммы библиотеки `matplotlib` и наложения линейной модели (синим цветом) и *средней абсолютной ошибки* (в виде двух параллельных полос серым цветом):

```
plt.scatter(x, y, color='black')
plt.plot(x, y_pred, color='blue', linewidth=3)
plt.plot(x, y_pred + mae, color='lightgray')
plt.plot(x, y_pred - mae, color='lightgray')
plt.xlabel('Рост, дюймы')
plt.ylabel('Вес, фунты')
```

Если выполнить приведенный выше фрагмент кода, то на выходе будет получен график, показанный на рис. 1.1.

Как следует из графика на рис. 1.1, во многих случаях фактические точки находятся на расстоянии 20–25 фунтов от предсказания. Тем не менее средняя абсолютная ошибка может вас дезинформировать, заставив думать, что ошибка всегда ближе к 8. Поэтому крайне важно выполнять визуализацию ошибки модели, чтобы понимать ее распределение. По этому графику можно сказать, что в нашем распределении нет красных флажков, которые выделялись бы, например, как в случае большей разбросанности остатков для одного диапазона роста, чем для других. Поскольку точки на графике распределены более-менее равномерно относительно синей прямой, мы говорим, что распределение величины **гомоскедастично**. В случае линейной регрессии гомоскедастичность — одно из многих модельных допущений, которые вы должны проверять наряду с *линейностью*, *нормальностью*, *не-*

зависимостью и отсутствием *мультиколлинеарности* (если есть более одного признака). Эти допущения гарантируют, что вы используете модель, которая подходит для выполняемой работы. Другими словами, рост и вес *можно объяснить* линейной связью, и это объяснение будет неплохим с точки зрения статистики.

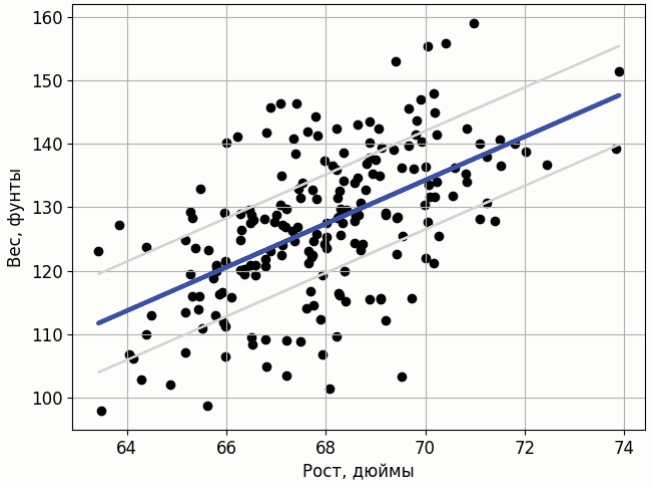


Рис. 1.1. Линейно-регрессионная модель предсказания веса на основе роста

С помощью этой модели мы пытаемся установить факт линейной связи между ростом и весом. Эта ассоциация называется **линейной корреляцией**. Измерить ее силу можно с помощью **коэффициента корреляции Пирсона**. Указанный статистический метод измеряет связь между двумя случайными величинами, используя их ковариацию, деленную на их среднеквадратичные отклонения. Ковариация — это число между −1 и 1, причем чем указанное число ближе к нулю, тем слабее связь между величинами. Если это число является положительным, существует положительная связь, а если оно является отрицательным — отрицательная связь. В Python коэффициент корреляции Пирсона можно вычислить с помощью функции `pearsonr` из библиотеки `scipy`:

```
corr, pval = pearsonr(x[:,0], y[:,0])
print(corr)
```

Результат:

```
0.5568647346122992
```

Данное число является положительным, что неудивительно, т. к. по мере увеличения роста вес тоже имеет тенденцию к увеличению, но оно также ближе к 1, чем к 0, а значит, вес сильно коррелирован. Второе число, произведенное функцией `pearsonr`, является *p*-значением, служащим для проверки некорреляции. Если протестировать уровень ошибки на значение менее 5%, можно уверенно сказать, что для этой корреляции существует достаточно оснований:

```
print(pval < 0.05)
```

Результат:

```
True
```

Понимание поведения модели и соответствующих обстоятельств ее поведения помогает нам **объяснять, почему, собственно говоря, она делает определенные предсказания**, и когда она на них не способна. Давайте вообразим, что нас просят объяснить, почему кто-то с ростом 71 дюйм должен весить 134 фунта, но вместо этого весил на 18 фунтов больше. Судя по тому, что мы знаем о модели, это погрешность ошибки, т. к. существует целый ряд обстоятельств, при которых мы не можем ожидать надежности от модели. Что делать, если нас попросят использовать эту модель для предскания веса человека ростом 56 дюймов? Сможем ли мы обеспечить такой же уровень точности? Определенно нет, потому что мы выполняем подгонку модели под данные испытуемых не ниже 63 дюймов. То же самое было бы, если бы нас попросили предсказать вес 9-летнего ребенка, потому что тренировочные данные были предназначены для 18-летних.

Несмотря на приемлемые результаты, этот пример модели предсказания веса не был реалистичным. Если вы хотите получать более точные результаты, но, что важнее, достоверно прогнозировать влияние разных факторов на вес человека, вам нужно будет добавить больше переменных. К примеру, можно добавить пол, возраст, рацион и уровень физической активности. Вот где ситуация станет интересней, потому что необходимо обеспечить **объективность их включения или не включения**. Например, если бы был включен пол, но большинство нашего набора данных состояло бы из мужчин, то как бы вы смогли обеспечить точность для женщин? Такая ситуация называется **систематическим смещением вследствие отбора данных**. А что, если вес больше связан с образом жизни и такими обстоятельствами, как бедность и беременность, нежели с полом? Если эти переменные не включены, это называется **систематическим смещением вследствие пропущенных переменных**. И тогда имеет ли вообще смысл включать чувствительную переменную пола с риском добавления в модель систематического смещения?

Если у вас есть несколько признаков, и вы их проверили на предмет объективности, можно узнать и *объяснить признаки, которые влияют на результативность модели*. Мы называем эту процедуру определением **важности признаков**. Однако, по мере добавления большего числа переменных, мы увеличиваем сложность модели. Парадоксально, но эта проблема касается интерпретации, и мы рассмотрим ее подробнее в следующих главах. А пока ключевым выводом должно быть утверждение, что интерпретация модели во многом связана с объяснением следующего:

1. Можно ли объяснить, что предсказания были сделаны объективно?
2. Можно ли надежно проследить предсказания назад до чего-то или кого-то?
3. Можно ли объяснить, как были сделаны предсказания? Можно ли объяснить ход работы модели?

И в конечном счете вопрос, на который мы пытаемся ответить, звучит так: *можно мы доверять модели*?

Три главных понятия интерпретируемого машинного обучения напрямую связаны с тремя приведенными выше вопросами и имеют аббревиатуру **FAT**, что означает fairness (**объективность**)[1], accountability (**подотчетность**) и transparency (**прозрачность**). Если можно объяснить, что предсказания были сделаны без заметного систематического смещения, тогда есть **объективность**. Если можно объяснить, почему, собственно, модель делает определенные предсказания, тогда есть **подотчетность**. И если можно объяснить, каким образом были сделаны предсказания и как модель работает, тогда есть **прозрачность**. С этими понятиями связано много вопросов, вызывающих этические соображения (рис. 1.2).

Рис. 1.2. Три главных понятия интерпретируемого машинного обучения

Некоторые исследователи и компании расширили показатель FAT, поместив его под более широким зонтиком этического **искусственного интеллекта** (ИИ), тем самым превратив FAT в FATE (где буква E в аббревиатуре соответствует английскому слову ethical). Этический ИИ является частью еще более широкого обсуждения темы алгоритмического регулирования и регулирования на основе данных. Тем не менее оба понятия очень сильно пересекаются, поскольку интерпретируемое машинное обучение связано с тем, как в машинном обучении имплементируются принципы FAT и этические вопросы. В этой книге мы будем обсуждать этику именно в таком контексте. Например, *глава 13* относится к надежности, безвредности и безопасности. В *главе 11* обсуждается тема объективности.

[1] Объективность (fairness), согласно Оксфордскому словарю английского языка, — это нейтральное и равное обращение с кем-либо/чем-либо или поведение без фаворитизма или дискриминации. В русском языке указанный термин может переводиться как "справедливость". В книге используется первый вариант, поскольку слово "справедливость" относится к поведению людей, а книга посвящена вычислительным моделям. К тому же термин justice, который тоже переводится как "справедливость", вступает в коллизию и вносит путаницу. — *Прим. перев.*

Понимание разницы между интерпретируемостью и объяснимостью

Читая первые несколько страниц этой книги, вы, вероятно, кое-что заметили, а именно, что глаголы *"интерпретировать"* и *"объяснять"*, а также существительные *"интерпретация"* и *"объяснение"* использовались взаимозаменяемо. Это неудивительно, учитывая, что интерпретировать — значит толковать, объяснять смысл чего-то. Несмотря на это, родственные термины *"интерпретируемость"* и *"объяснимость"* не следует использовать взаимозаменяемо, даже если их часто принимают за синонимы.

Что такое интерпретируемость?

Интерпретируемость позволяет уяснить, насколько люди, включая экспертов, не относящихся к обсуждаемому предмету, могут понимать причину и следствие, а также данные на входе в модель машинного обучения и на выходе из нее. Сказать, что модель обладает высоким уровнем интерпретируемости, означает, что можно описать ее вывод в доступном для толкования человеком ключе. Другими словами, почему данные на входе в модель дают определенный результат на выходе? Каковы требования и ограничения данных на выходе? Какими границами уверенности обладают предсказания? Или почему одна переменная оказывает более существенное влияние, чем другая? Говоря об интерпретируемости, следует учеть, что подробное описание хода работы модели имеет значение только в той степени, в которой она может объяснять свои предсказания и оправдывать свое применение.

В приведенном в этой главе примере вы могли бы заметить, что между ростом и весом человека существует линейная связь, поэтому имеет смысл использовать линейно-регрессионную модель, а не нелинейную. Это можно доказать статистически, потому что участвующие переменные не нарушают допущения линейной регрессии. Даже когда статистика на нашей стороне, все равно следует учесть знания из предметной области, связанной с вариантом использования. В данном конкретном случае наша уверенность опирается на биологию, потому что наши знания физиологии человека не противоречат взаимосвязи между ростом и весом.

Осторожно, сложность!

Многие модели машинного обучения изначально сложнее понимать просто из-за математики, связанной с внутренним механизмом модели или конкретной архитектурой модели. В дополнение к этому выбирается много вариантов решений, которые могут увеличивать сложность и делать модели менее интерпретируемыми, от выбора набора данных до селекции и конструирования признаков, до выбора вариантов обучения и настройки модели. Эта сложность делает задачу объяснения хода ее работы чрезвычайно непростой. Интерпретируемость машинного обучения является весьма активной областью исследований, поэтому по-прежнему существует много споров о ее точном определении. Указанная полемика включает вопрос о необходимости полной прозрачности, чтобы мы имели возможность квалифициро-

вать модель машинного обучения как достаточно интерпретируемую. В этой книге отдается предпочтение пониманию того, что определение понятия "интерпретируемость" не должно с неизбежностью исключать непрозрачные модели, которые по большей части являются сложными, если выбор не ставит под угрозу их надежность. Этот компромисс обычно называется **постфактумной интерпретируемостью** (post-hoc interpretability). В конце концов, невозможно объяснить точно, каким образом человеческий мозг, во многом подобный сложной модели машинного обучения, делает выбор, но мы часто доверяем его решению, потому что можем попросить человека рассказать о своем ходе рассуждений. Постфактумная интерпретация машинного обучения — это в точности то же самое, за исключением того, что здесь рассуждения объясняются человеком от имени модели. Использование конкретного понятия интерпретируемости выгодно тем, что можно интерпретировать непрозрачные модели и не жертвовать точностью наших предсказаний. Мы обсудим эту тему подробнее в *главе 3*.

Когда важна интерпретируемость?

Системы принятия решений не всегда требуют интерпретируемости. В исследованиях предлагаются два случая в качестве исключений.

- Когда неправильные результаты не имеют существенных последствий. Например, что делать, если модель машинного обучения, натренированная находить и читать почтовый индекс на бандероли, иногда неправильно его читает и отправляет бандероль по другому адресу? Вероятность дискриминационного смещения невелика, а стоимость ошибочной классификации относительно низкая. Такие ситуация встречаются не слишком часто, чтобы можно было увеличивать стоимость.
- Когда есть последствия, но они достаточно изучены и подтверждены в реальном мире, чтобы можно было принимать решения без участия человека. Так обстоит дело с **системой оповещения о воздушном движении и предупреждения столкновений в воздухе** (traffic-alert and collision-avoidance system, TCAS), которая предупреждает пилота другого самолета, который представляет угрозу столкновения в воздухе.

С другой стороны, интерпретируемость необходима для того, чтобы эти системы имели следующие атрибуты.

- **Пригодность для получения научных знаний**: метеорологам есть чему поучиться на основе климатической модели, но только если ее легко интерпретировать.
- **Надежность и безвредность**: решения, принимаемые беспилотным транспортным средством, должны быть доступны для отлаживания, чтобы его разработчики могли понимать точки отказа.
- **Этичность**: в модели перевода с языка на язык могут использоваться гендерно-смещенные вложения слов, которые приводят к дискриминационным переводам, но вы должны уметь легко находить эти примеры и исправлять их. Однако сис-

тема должна быть спланирована так, чтобы она могла извещать вас о проблеме до момента ее выпуска в публичное пространство.

- **Убедительность и выверенность**: иногда модели машинного обучения могут иметь неполные и взаимоисключающие целевые задачи; например, система контроля холестерина может не учитывать степень вероятности того, что пациент действительно будет придерживаться диеты или режима приема лекарственных препаратов, либо между одной целевой задачей и другими, такими как безвредность и отсутствие дискриминации, может существовать компромисс.

Объясняя решения модели, можно устранять пробелы нашего понимания задачи — *ее неполноту*. Одна из наиболее важных проблем заключается в том, что, учитывая высокую точность наших технических решений в области машинного обучения, мы склонны повышать уровень нашей уверенности до такой степени, что считаем, будто полностью понимаем задачу. И тогда мы оказываемся дезориентированными, полагая, что наше решение охватывает *ее целиком*!

В начале этой книги мы затронули вопрос о том, что привлечение данных для порождения алгоритмических правил не является чем-то новым. Однако раньше мы судили об этих правилах задним числом, а теперь не можем. Следовательно, раньше был ответственен человек, а теперь алгоритм. В данном случае алгоритмом является модель машинного обучения, которая отвечает за все этические последствия, которые она влечет за собой. Указанный поворот во многом связан с точностью. Проблема в том, что, хотя модель машинного обучения может превосходить человеческую точность в целом, она еще должна интерпретировать свои результаты так, как это делал человек. Следовательно, она не судит о своих решениях постфактум, поэтому в качестве решения ей не хватает желаемого уровня полноты, и именно по этой причине нам нужно интерпретировать модели так, чтобы иметь возможность закрывать хотя бы часть этого пробела. Тогда почему интерпретация машинного обучения до сих пор еще не является стандартной частью конвейера? Дело в том, что в дополнение к нашему уклону в сторону сосредоточенности лишь на точности одним из самых больших препятствий является устращающая концепция моделей типа черного ящика.

Что такое модели типа черного ящика?

Это просто еще один термин для непрозрачных моделей. Черный ящик обозначает систему, в которой известны только входы и выходы и нет возможности увидеть, что именно преобразовывает входы в выходы. В случае машинного обучения модель типа черного ящика может быть открыта, но ее механизмы нелегко понять.

Что такое модели типа белого ящика?

Они противоположны моделям типа черного ящика (рис. 1.3). Их также называют прозрачными, поскольку они достигают полной или почти полной прозрачности интерпретации. В этой книге мы называем их **имманентно интерпретируемыми** и рассмотрим их подробнее в *главе 3*.

Взгляните на сравнение этих моделей.

Рис. 1.3. Визуальное сравнение моделей типа типа белого ящика и типа черного ящика

Что такое объяснимость?

Объяснимость охватывает все сущности, входящие в сферу интерпретируемости. Разница заключается в том, что она уходит глубже в требование прозрачности, чем интерпретируемость, т. к. требует удобных для человека объяснений внутреннего механизма работы модели и процесса обучения модели, а не только выведения модельного результата. В зависимости от приложения это требование может распространяться на различные степени модельной, конструктивной и алгоритмической прозрачности. Указанные три типа прозрачности описаны далее.

- **Модельная прозрачность** — возможность объяснять, как выполняется пошаговое обучение модели. В случае нашей простой модели предсказания веса можно объяснить, как метод оптимизации, именуемый **обычными наименьшими квадратами**, находит коэффициент β, который минимизирует ошибки в модели.
- **Конструктивная прозрачность** — возможность объяснять выбираемые варианты, такие как модельная архитектура и гиперпараметры. Например, мы могли бы обосновать эти варианты выбора на основе размера или природы тренировочных данных. Если бы мы прогнозирвали продажи и знали, что у них сезонность составляет 12 месяцев, то это могло бы быть правильным вариантом выбора параметров. В случае сомнений для отыскания правильной сезонности мы всегда могли бы использовать какой-нибудь хорошо зарекомендовавший себя статистический метод.
- **Алгоритмическая прозрачность** — возможность объяснять автоматизированные оптимизации, такие как поиск по сетке (grid search) гиперпараметров; но обратите внимание, что те, которые невозможно воспроизвести из-за их случайной природы, такие как случайный поиск гиперпараметрической оптимизации, ранняя остановка и стохастический градиентный спуск, делают алгоритм непрозрачным.

Непроницаемые модели называются *непроницаемыми* просто потому, что им не хватает *модельной прозрачности*, но для многих моделей это неизбежно, каким бы оправданным ни был выбор модели. Во многих сценариях, даже если вывести математику, участвующую, скажем, в обучении нейронной сети или случайного леса, она будет вызывать больше сомнений, чем доверие. Для этого есть по меньшей мере несколько причин, которые изложены ниже:

- **Статистическая необоснованность**: непроницаемый процесс обучения модели соотносит данные на входе с оптимальными данными на выходе, основываясь на произвольных значениях параметров. Эти параметры были найдены путем оптимизации функции стоимости, но не основаны на статистической теории.
- **Неопределенность и невоспроизводимость**: когда вы выполняете обучение прозрачной модели под одинаковые данные, вы всегда получаете одинаковые результаты. С другой стороны, непроницаемые модели не являются одинаково воспроизводимыми, потому что для инициализации своих весов или для регуляризации или оптимизации своих гиперпараметров в них используются случайные числа или задействуется стохастическая избирательность (так обстоит дело, например, с алгоритмом случайного леса).
- **Переобучение и проклятие размерности**: многие из этих моделей оперируют на высокоразмерном пространстве. Это не вызывает доверия, потому что труднее делать обобщения на более крупное число размерностей. В конце концов, имеется тем больше риска для переобучения, чем больше размерностей вы добавляете.
- **Человеческое познание и проклятие размерности**: прозрачные модели часто используются для малых наборов данных с меньшим числом размерностей, и даже если они не являются прозрачными, в них никогда не используется больше размерностей, чем необходимо. Они также, как правило, не усложняют взаимодействие между этими размерностями больше, чем это необходимо. Это отсутствие ненужной сложности облегчает визуализацию того, что модель делает, и ее исходов. Люди не особо справляются с комплексным учетом многочисленных факторов, поэтому использование прозрачных моделей, как правило, облегчает их понимание.
- **Бритва Оккама**: это правило называется принципом простоты, или бережливости. В нем говорится, что самое простое решение обычно является правильным. Правда это или нет, но люди тоже склонны к простоте, и прозрачные модели известны — если что — своей простотой.

Почему и когда объяснимость имеет важное значение?

Заслуживающее доверия и этичное принятие решений являются главными мотивами для интерпретируемости. Объяснимость имеет дополнительные мотивы, такие как причинно-следственная связь, переносимость и информативность. Поэтому существует много вариантов использования, в которых ценится полная или почти полная прозрачность, и это правильно. Некоторые из них описаны ниже:

- **Научные исследования**: воспроизводимость имеет важное значение для научного метода. Кроме того, использовать статистически обоснованные методы оп-

тимизации особенно желательно, когда необходимо доказать причинно-следственную связь.

- **Клинические испытания**: они также должны давать воспроизводимые результаты и быть статистически обоснованными. В дополнение к этому, учитывая потенциальную тяжесть последствий переобучения, в них должны использоваться наименьшие возможные размерности и модели, которые их не усложняют.
- **Верификация безвредности потребительских товаров**: как и в случае клинических испытаний, когда речь идет о безвредности, касающейся жизни и смерти, простота по возможности бывает предпочтительнее.
- **Общественная политика и право**: это более узкая тема обсуждения в рамках того, что ученые в области правоведения называют **алгоритмическим регулированием**, и они проводят различие между **авквариумной прозрачностью** и **обоснованной прозрачностью**. Первая из них ближе к строгости, необходимой для верификации безвредности потребительских товаров, а вторая — там, где будет достаточно постфактумной интерпретируемости. Когда-нибудь правительство будет полностью управляться алгоритмами. Сейчас трудно сказать, какая политика будет соответствовать какой форме прозрачности, но существует много областей общественной политики, таких как уголовное правосудие, где необходима абсолютная прозрачность. Однако всякий раз, когда полная прозрачность противоречит целевым задачам обеспечения конфиденциальности или безопасности, должна применяться менее строгая форма прозрачности.
- **Уголовное расследование и аудит соответствия регуляторным требованиям.** Если что-то пойдет не так, например, произойдет авария на химическом заводе, вызванная неисправностью робота, или авария на автономном транспортном средстве, то следователю необходимо проследить **цепочку принятых решений**. Это делается для того, чтобы "облегчить возложение процедурной и юридической ответственности". Даже если никаких несчастных случаев не произошло, этот вид аудита может проводиться по требованию властей. Аудит соответствия требованиям применяется к регулируемым отраслям, таким как финансовые и коммунальные услуги, транспорт и здравоохранение. Во многих случаях "аквариумная" прозрачность является предпочтительной.

Деловое обоснование интерпретируемости

В этом разделе описывается несколько практических деловых выгод интерпретируемости машинного обучения, таких как более качественные решения, а также большее доверие, этичность и прибыльность.

Более качественные решения

Как правило, модели машинного обучения тренируются, а затем оцениваются в отношении желаемых метрик. Если они проходят контроль качества на тестовом наборе данных (т. е. на множестве данных, которые не использовались при обуче-

нии), они развертываются в производстве. Однако после проверки в реальных условиях всё принимает буйный характер, как в следующих ниже гипотетических сценариях.

- Алгоритм высокочастотной торговли может в одиночку разрушить фондовый рынок.
- Сотни устройств для умного дома могут необъяснимо разразиться необузданным хохотом, приводя в ужас своих пользователей.
- Системы распознавания номерных знаков могут начать считывать номерные знаки нового типа неправильно и штрафовать не тех водителей.
- Расово смещенная система наблюдения может неправильно обнаружить злоумышленника, и из-за этого охранники выстрелят в невинного офисного служащего.
- Самоходный автомобиль может принять снег за тротуар, врезаться в скалу и травмировать пассажиров.

Любая система подвержена ошибкам, поэтому нельзя сказать, что интерпретируемость есть средство от всех ошибок. Однако концентрация на простой оптимизации метрик может стать причиной катастрофы. В лабораторных условиях модель может и будет хорошо обобщать, но если вы не знаете, почему модель принимает те или иные решения, тогда существует вероятность упустить шанс ее улучшения. Например, знать, *что* именно самоходный автомобиль считает дорогой, недостаточно, но знания, *как* он это делает, помогут в совершенствовании модели. Если, скажем, одной из причин была светлая дорога, как снег, то это могло представлять опасность. Проверка допущений и выводов модели может приводить к улучшению модели путем введения в набор фотоснимков зимних дорог или потоковой подачи в модель реально-временны́х данных о погоде. Кроме того, если это не сработает, возможно, алгоритмическая отказоустойчивость способна помешать ей осуществлять действия по принятому решению, в котором она не совсем уверена.

Одна из главных причин, по которой акцент на интерпретируемости машинного обучения приводит к более качественному принятию решений, была упомянута ранее, когда мы говорили о полноте. Если вы считаете, что модель является полной, то какой смысл делать ее лучше? Более того, если вы не подвергаете рассуждения модели сомнению, тогда ваше понимание задачи должно быть полным. Если это так, то, возможно, вам вообще не следует использовать машинное обучение для решения задачи! Машинное обучение создает алгоритм, который в иной ситуации было бы слишком сложно запрограммировать с использованием инструкций `if-else`, именно для использования в тех случаях, когда наше понимание задачи является неполным!

Оказывается, когда мы что-то предсказываем или оцениваем, в особенности с высокой степенью точности, мы полагаем, что имеем над этим контроль. Это как раз и называется **систематическим смещением вследствие иллюзии контроля**. Нельзя недооценивать сложность задачи только потому, что в совокупности модель почти все время выполняет ее правильно. Разница между снегом и бетонным покрытием

бывает размытой и трудно объяснимой даже для человека. Откуда вообще начинать описывать это различие в таком ключе, чтобы она всегда была точной? Модель может усваивать эти различия, но это не делает ее менее сложной. Проверка модели на предмет точек отказа и постоянное наблюдение за выбросами требует другого взгляда, в соответствии с которым мы признаем, что не способны контролировать модель, но можем попытаться понять ее с помощью интерпретации.

Ниже приведены несколько дополнительных систематических смещений при принятии решений, способные влиять на модель негативно и служить причинами, по которым интерпретируемость может приводить к более качественному принятию решений.

- **Систематическое смещение вследствие консерватизма.** При получении новой информации мы не меняем своих прежних мнений. С таким систематическим смещением укоренившаяся ранее существовавшая информация одерживает верх над новой информацией, но модели должны эволюционировать. Следовательно, подход, который предпочитает ставить предыдущие допущения под сомнение, является здравым.
- **Систематическое смещение вследствие заметности.** Некоторые выступающие наружу или более заметные детали могут выделяться больше других, но, говоря статистически, им следует уделять равное внимание наряду с другими. Это систематическое смещение может влиять на наш выбор признаков, поэтому менталитет на основе интерпретируемости может расширять наше понимание задачи за счет включения в нее других, менее воспринимаемых, признаков.
- **Фундаментальная ошибка атрибуции.** Это систематическое смещение заставляет нас приписывать исходы поведению, а не обстоятельствам, характеру, а не ситуациям, природе, а не формированию/воспитанию. Интерпретируемость требует от нас проведения более глубокого исследования и поиска менее очевидных взаимосвязей между нашими переменными или теми, которые могут быть пропущены.

Одной из важнейших выгод от интерпретации моделей является локализация *выбросов*. Выбросы, как аномальные значения, бывают потенциальным новым источником активов или пассивов, которые могут происходить в любой момент. Знание этого факта помогает нам соответствующим образом подготавливаться и вырабатывать стратегию.

Более надежные бренды

Доверие определяется как вера в надежность, способность или авторитетность чего-либо или кого-либо. В контексте организаций доверие — это ее репутация; и в неумолимом суде общественного мнения требуется всего один-единственный неуспешный случай, разногласие или фиаско, чтобы значительное доверие общественности было утрачено. А это, в свою очередь, может приводить к ослабеванию доверия инвесторов.

Давайте посмотрим, что случилось с Boeing после крушения 737 MAX или с Facebook после скандала с президентскими выборами 2016 года. В обоих случаях были приняты недальновидные решения исключительно с целью оптимизации одной метрики, будь то прогнозные продажи самолетов или продажи цифровой рекламы. Они недооценили известные потенциальные точки отказа и совершенно упустили очень большие точки. После этого нередко ситуация ухудшается, когда организации прибегают к ложной аргументации, чтобы оправдывать свои рассуждения, дезориентировать общественность или отвлечь внимание СМИ на что-то другое. Такое поведение может приводить к дополнительным промахам в связях с общественностью. Компании не только теряют авторитетность в том, *что они делают* со своей первой ошибкой, но и пытаются обманывать людей, теряя авторитетность в том, *что говорят их представители*.

И это были примеры решений, принимаемых по большей части людьми. С решениями, принимаемыми исключительно моделями машинного обучения, все может ухудшиться, потому что легко бросить мяч и удерживать подотчетность на стороне модели. Например, если в вашей ленте Facebook стали появляться оскорбительные материалы, то Facebook мог бы сказать, что это потому, что его модель была натренирована на ваших данных, таких как ваши комментарии и лайки, поэтому социальная сеть на самом деле отражает то, что вы хотите увидеть. Это не ее вина, а ваша. Если полиция нацелилась на ваш район для проведения агрессивной охраны общественного порядка, так это потому, что она использует PredPol — алгоритм, который предсказывает, где и когда произойдут преступления, и представители охраны правопорядка могут обвинить в этом алгоритм. С другой стороны, создатели этого алгоритма могут обвинить полицию в том, что программно-информационное обеспечение было натренировано на их полицейских отчетах. Это генерирует потенциально тревожный и порочный цикл обратной связи, не говоря уже о разрыве в подотчетности. И если некоторые пранкеры или хакеры устранят разметку полос, то это может привести к тому, что беспилотный автомобиль Tesla свернет на неправильную полосу. Разве это вина Tesla в том, что разработчики не предвидели такой возможности, или хакеров, которые бросили гаечный ключ в эту модель? Это называется **антагонистической атакой**, и мы обсудим эту тему в *главе 13*.

Несомненно, одна из целей интерпретируемости машинного обучения как таковой состоит в том, чтобы делать модели качественнее при принятии решений. Но даже когда они отказывают, можно продемонстрировать, что усилия прилагались. Доверие теряется не полностью из-за самого отказа, а из-за отсутствия подотчетности, и даже в тех случаях, когда принимать всю вину на себя было бы необъективно, некоторая подотчетность (или ответственность) лучше, чем никакая. Например, в предыдущем наборе примеров Facebook мог бы поискать подсказки о том, почему оскорбительные материалы показываются чаще, а затем посвятить себя поиску способов делать это реже, даже если это принесет меньше денег. Пользователи алгоритма PredPol могли бы найти другие источники наборов данных о преступности, которые потенциально менее смещены, даже если они меньше. Они также могут использовать методы для устранения систематического смещения в существующих наборах данных (*см. главу 11*). И Tesla может провести аудит своих систем на

предмет антагонистических атак, даже если это задержит старт продаж автомобилей. Все эти решения касаются интерпретации. После того как это станет обычной практикой, они могут привести к повышению доверия не только общественности, например со стороны пользователей и клиентов, но и внутренних интересантов, таких как сотрудники и инвесторы.

На рис. 1.4 показано несколько промашек ИИ в связях с общественностью, которые произошли за последние пару лет.

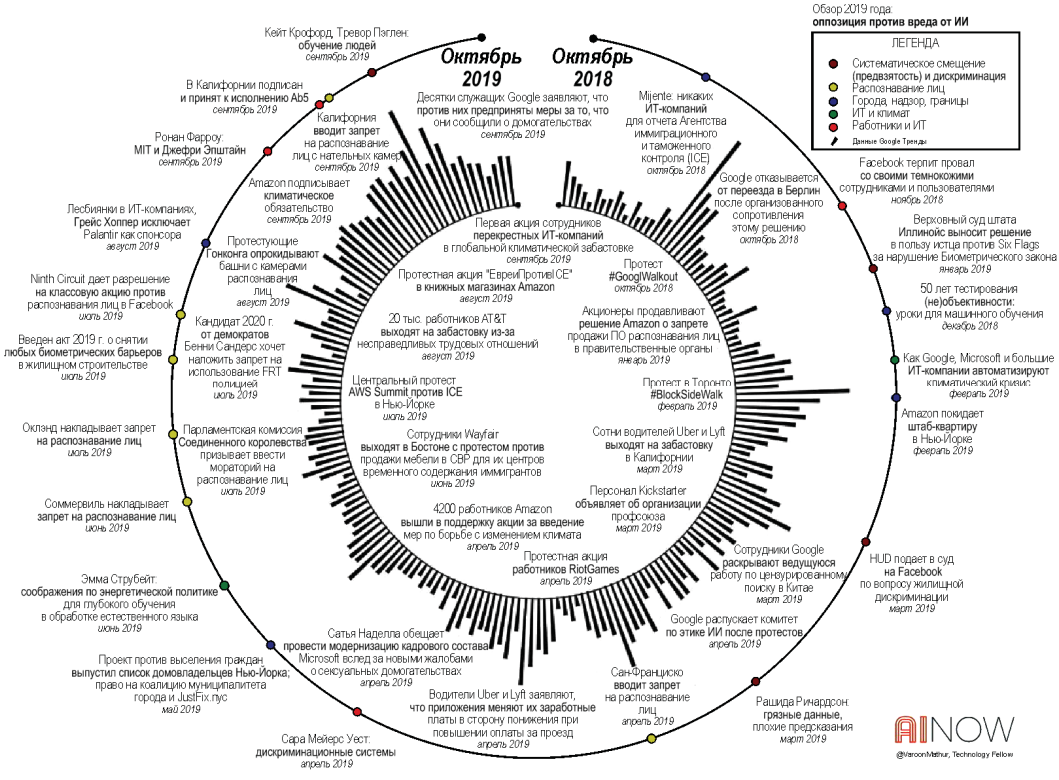


Рис. 1.4. Инфографика института AI Now с промашками искусственного интеллекта в связях с общественностью за 2019 год

Из-за проблем с доверием многие технологии, основанные на искусственном интеллекте, теряют общественную поддержку в ущерб как компаниям, монетизирующим ИИ, так и пользователям, которые могли бы извлекать из них выгоду (см. рис. 1.4). Это требует правовой базы отчасти на национальном или глобальном уровне, а также большей подотчетности в организационном плане для тех, кто внедряет эти технологии.

Более высокий уровень этичности

Существует три школы этики: утилитаристы сосредоточены на последствиях, деонтологи озабочены долгом, а телеологи больше интересуются совокупным моральным

ным характером. Таким образом, это означает, что существуют разные способы изучения этических проблем. Например, из них всех можно извлекать полезные уроки. Бывают случаи, когда хочется произвести наибольшее количество "добра" несмотря на то, что по ходу причиняется и некоторый вред. В отдельных случаях этические границы должны трактоваться как линии на песке (иными словами, жесткие границы), которые нельзя пересекать. В других ситуациях речь идет о развитии праведного нрава, к чему стремятся многие религии. Независимо от этических воззрений, которых мы придерживаемся, наше представление о них со временем эволюционирует, потому что оно отражает наши текущие ценности. На данный момент в западных культурах к этим ценностям относятся следующие:

- благосостояние человека;
- владение и право собственности;
- конфиденциальность;
- свобода от предвзятости;
- универсальное удобство;
- доверие;
- автономность;
- информированное согласие;
- подотчетность;
- вежливость;
- экологическая стабильность.

Этические нарушения — это случаи, когда вы пересекаете моральные границы, которые указанные ценности стремятся поддерживать, будь то дискриминация по отношению к кому-либо или загрязнение окружающей среды, независимо от того, противоречит это закону или нет. Этические дилеммы возникают, когда у вас есть выбор между вариантами, которые приводят к нарушениям, поэтому вам приходится выбирать между одним и другим.

Первая причина, по которой машинное обучение связано с этикой, заключается в том, что технологии и этические дилеммы имеют имманентно связанную историю.

Начиная с первого широко распространенного созданного людьми инструмента, технологии несли прогресс, но также наносили вред; примерами могут служить несчастные случаи, войны и потери рабочих мест. Это не значит, что технологии всегда плохи, но нам не хватает предусмотрительности измерять и контролировать их последствия во временно́й динамике. В случае с ИИ совершенно не ясно, каковы вредные долгосрочные последствия. Чего можно ожидать, так это серьезной потери рабочих мест и огромного спроса на энергию для питания наших центров обработки данных, что может нанести вред окружающей среде. Существуют соображения, что ИИ может создать "алгократическое" надзорное государство, управляемое алгоритмами, нарушая такие ценности, как конфиденциальность, автономия и право собственности.