Оглавление

Об авторе

Прадипта Мишра (Pradeepta Mishra) – руководитель отдела искусственного интеллекта в компании L&T Infotech (LTI), возглавляет группу специалистов по анализу данных, вычислительной лингвистике, машинному и глубокому обучению, участвующих в создании функций искусственного интеллекта для обработки данных. Он был два года подряд (2019, 2020) награжден премией «40 лучших специалистов по обработке данных Индии в возрасте до 40 лет», по версии журнала Analytics India. Как изобретатель, подал пять патентов, которые в настоящее время находятся на рассмотрении в разных странах мира. Является автором четырех книг, опубликованных на разных языках, включая английский, китайский и испанский. Его первая книга была рекомендована центром HSLS при Питтсбургском университете, штат Пенсильвания, США. Последняя его книга «PyTorch. Рецепты» была опубликована издательством Apress. Он выступил с основным докладом на конференции 2018 Global Data Science Conference, Калифорния, США. Выступил с более чем 500 докладами по анализу данных, машинному обучению, глубокому обучению, обработке естественного языка и искусственному интеллекту в различных университетах, на встречах, в технических институтах и на форумах, организованных сообществом. Является приглашенным преподавателем по курсам искусственного интеллекта, машинного обучения и кибербезопасности в магистратуре университета Рева, Бангалор, Индия, а также в других университетах. За последние девять лет обучал более 2 000 специалистов по анализу данных и инженеров по искусственному интеллекту.

О рецензентах

Абхишек Виджайваргия (Abhishek Vijayvargia) работает специалистом по данным и прикладным наукам в компании Microsoft. Ранее он работал во многих стартапах, связанных с машинным обучением и интернетом вещей. Разрабатывал алгоритмы искусственного интеллекта для решения различных проблем в области кибербезопасности, интернета вещей, производства, оптимизации перевозок и логистики. Он также является экспертом и активным исследователем в области объяснимого искусственного интеллекта и проектирования систем машинного обучения. Является автором работ по анализу данных, автором и рецензентом научных статей, представил свои идеи на многих конференциях по искусственному интеллекту.

Бхаратх (Bharath) имеет более чем десятилетний опыт работы, в настоящее время он старший инженер-консультант по науке о данных в компании Verizon, Бенгалуру. Имеет диплом аспиранта по анализу данных от бизнес-школы Praxis и степень магистра наук о жизни от Университета штата Миссисипи, США. Работал исследователем данных в университете Джорджии, Университете Эмори и компании Eurofins LLC. В компании Happiest Minds он работал над продуктами цифрового маркетинга на основе искусственного интеллекта и решениями на основе обработки естественного языка в сфере образования. Наряду с повседневными обязанностями является наставником и активным исследователем. Его особенно интересуют архитектуры самостоятельного, полусамостоятельного и эффективного глубокого обучения в обработке естественного языка и компьютерном зрении.

Благодарности

Эта книга основана на исследовании объяснимости моделей искусственного интеллекта на основе «черного ящика» и преобразовании решений, принимаемых моделями искусственного интеллекта, в прозрачные. Я благодарен моим друзьям и семье за то, что они побудили меня начать эту работу, упорно продолжать ее и дойти до последнего шага – опубликовать ее.

Я благодарю свою жену Праджну (Prajna) за ее постоянное ободрение и поддержку в завершении работы над книгой, помощь в расстановке приоритетов между книгой и отпуском, заботу о детях и за то, что давала мне достаточно времени для завершения книги.

Я благодарю моего редактора Дивию (Divya), которая оказывала мне постоянную поддержку на протяжении всей работы над книгой, проявляя гибкость в отношении сроков и давая мне достаточно времени, чтобы завершить начатое.

Я благодарю редакционную коллегию и Суреша (Suresh) за веру в то, что смогу написать о сложной теме объяснимости моделей, и предоставление возможности написать на эту тему.

Наконец, я благодарю своих дочерей Аарью (Aarya) и Аадию (Aadya) за то, что они любят меня и поддерживали в завершении работы над этой книгой.

Введение

Объяснимый искусственный интеллект (explainable artificial intelligent – XAI) является актуальной потребностью, поскольку все больше и больше моделей искусственного интеллекта (artificial intelligent – AI) используется для выработки бизнес-решений. Таким образом, эти решения также воздействуют на многих пользователей. При этом каждый пользователь может получить благоприятное или неблагоприятное воздействие. Поэтому важно знать ключевые особенности, приводящие к принятию этих решений. Часто утверждают, что модели AI являются по своей природе «черным ящиком», поскольку решения модели невозможно объяснить. Поэтому в промышленности они внедряются довольно медленно. Эта книга представляет собой попытку раскрыть так называемые модели «черного ящика», чтобы повысить адаптивность, интерпретируемость и объяснимость решений, принимаемых алгоритмами AI с использованием таких фреймворков, как библиотеки Python XAI, TensorFlow 2.0+, Keras, а также пользовательских фреймворков с использованием декораторов Python (Python Wrappers). Цель этой книги – объяснить модели AI на простом языке с использованием вышеупомянутых фреймворков.

Интерпретируемость и объяснимость модели – ключевые моменты этой книги. Существуют математические формулы и методы, которые обычно используются для объяснения решения, принятого моделью AI. Вам будут предоставлены методы, классы, фреймворки и функции программных библиотек, а также их использование для объяснения модели, прозрачности, надежности, этики, предвзятости и интерпретируемости. Если человек может понять причины решения, принятого моделью AI, это даст пользователю гораздо больше возможностей для внесения поправок и рекомендаций. Существует два различных типа пользователей – бизнес-пользователи и практикующие специалисты. Бизнес-пользователь хочет получить объяснение на простом языке без каких-либо статистических или математических терминов. Практик хочет знать объяснимость с точки зрения вычислений. Эта книга – попытка сбалансировать обе потребности при создании объяснений.

Эта книга начинается с введения в основы объяснимости и интерпретируемости модели, этических соображений при применении AI и предвзятости прогнозов, генерируемых моделями AI. Затем вы узнаете о надежности моделей искусственного интеллекта при создании прогнозов в различных случаях использования, изучите методы и системы интерпретации линейных моделей, нелинейных моделей и моделей временных рядов, используемых в AI. Далее узнаете о наиболее сложных ансамблевых моделях, объяснимости и интерпретируемости с использованием таких фреймворков, как Lime, SHAP, Skater, ELI5 и Alibi. Затем вы узнаете об объяснимости моделей для неструктурированных данных и обработки естественного языка, связанной с задачами классификации текстов. Изучение справедливости моделей также требует моделирова-

ния сценариев «что, если» с использованием результатов прогнозирования. Об этом вы узнаете далее. Затем вы прочитаете о контрфактных и контрастных объяснениях для моделей AI. Вы изучите объяснимость моделей для глубокого обучения, экспертных систем на основе правил, а также объяснения, не зависящие от модели, для инвариантности предсказаний и для задач компьютерного зрения, использующих различные фреймворки XAI.

Сегодня у нас есть инженеры по AI и специалисты по анализу данных, которые обучают или создают эти модели; разработчики программного обеспечения, которые вводят эти модели в производство и в эксплуатацию; бизнес-пользователи, которые потребляют конечный результат или результат, созданный с помощью моделей; и лица, принимающие решения, которые обдумывают решения, принятые с помощью моделей. Руководители, занимающиеся продвижением проектов/продуктов AI, думают: «Есть ли способ добиться ясности вокруг моделей и разработчиков прогнозирующих моделей?» Био-статистики, конечно, думают, как объяснить предсказания модели и т. д. Ожидается, что будет разработана система объяснения, которая отвечает потребностям всех заинтересованных сторон, вовлеченных в процесс внедрения AI в реальную жизнь. В этой книге соблюден баланс между несколькими заинтересованными сторонами. Предпочтение предоставляется специалистам по анализу данных (data scientists), поскольку, если они убеждены в объяснимости моделей, то смогут разъяснить далее заинтересованным сторонам бизнеса.

Чтобы сделать модели AI понятными для бизнес-пользователей на простом, доступном языке, потребуется некоторое время. Возможно, для решения этой задачи появится новая система. На данный момент проблема заключается в том, что специалист по анализу данных, построивший модель, не имеет полной ясности о поведении модели, и не хватает ясности в ее объяснении. Новоиспеченные специалисты по анализу данных или выпускники вузов получат огромную пользу от этой книги. Аналогичным образом эта книга будет полезна и другим инженерам по AI. Это развивающаяся область, объяснения в данной книге были актуальны на июль 2021 года.

ГЛАВА 1

Объяснимость и интерпретируемость модели

Мы начнем эту книгу с введения в основы объяснимости и интерпретируемости моделей, этических аспектов применения AI и предвзятости прогнозов, генерируемых моделями AI, рассмотрим надежность моделей AI при создании прогнозов в различных случаях использования. Затем изучим методы и системы для интерпретации линейных, нелинейных моделей и моделей временных рядов, используемых в AI. Далее разберем наиболее сложные ансамблевые модели, объяснимость и интерпретируемость с использованием таких фреймворков, как Lime, SHAP, Skater, ELI5 и т. д. Затем обсудим объяснимость моделей для неструктурированных данных и задач, связанных с обработкой естественного языка.

Создание основ

За последние несколько лет был достигнут огромный прогресс в области машинного обучения и глубокого обучения при создании решений на основе искусственного интеллекта (AI) в различных областях, в том числе в розничной торговле, банковском деле, финансовых услугах, страховании, здравоохранении, производстве и отраслях, основанных на интернете вещей. AI является основой многих продуктов и решений, которые появляются в связи с быстрой цифровизацией различных бизнес-функций. Причина того, что AI лежит в основе этих продуктов и решений, заключена в том, что интеллектуальные машины в настоящее время обладают способностями к обучению, рассуждению и адаптации. Опыта мало. Если мы сможем использовать богатый опыт, накопленный умными людьми, и отразить его с помощью применения обучения и рассуждений в компьютерах, это может значительно повысить эффективность обучения. В силу этих возможностей современные модели машинного обучения и глубокого обучения способны достичь беспрецедентных уровней производительности при решении сложных бизнес-задач, тем самым повышая эффективность бизнеса.

За последние два года появилось множество инструментов автоматического машинного обучения (AutoML – Automatic Machine Learning), фреймворков, инструментов с низким содержанием кода и без кода (с минимальным вмешательством человека), что является еще одним уровнем сложности, ко-

торого достигли системы, поддерживаемые искусственным интеллектом. Это воплощение практически нулевого вмешательства человека, необходимого с точки зрения разработки, поставки и развертывания решений. Когда решения полностью принимаются машинами, а люди всегда находятся на стороне получателя, возникает острая необходимость понять, как машины пришли к этим решениям. Модели, на которых основаны системы AI, часто называют моделями «черного ящика». Следовательно, существует необходимость в объяснимости и интерпретируемости моделей для того, чтобы объяснить сделанные ими прогнозы.

Искусственный интеллект

Искусственный интеллект означает систему в виде компьютерной программы, которая автоматически принимает решения от имени человека в отношении некоторой задачи, без явного программирования. Рисунок 1.1 объясняет взаимосвязь между машинным обучением (machine learning – ML), глубоким обучением (deep learning – DL) и искусственным интеллектом.

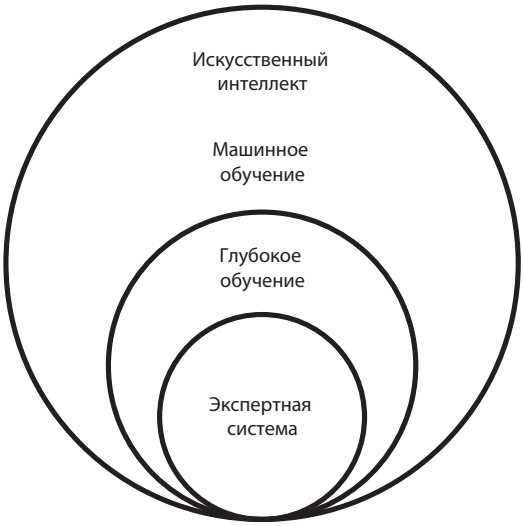

Рис. 1.1. Взаимосвязь между ML, DL и AI

Искусственный интеллект – это система, созданная с помощью компьютерной программы, в которой интеллектуальные выводы могут быть сделаны в отношении постановки задачи. В процессе получения вывода могут быть использованы алгоритмы машинного обучения или глубокого обучения. Алгоритмы машинного обучения – это простые математические функции, используемые в процессе оптимизации с использованием комбинаций входных и выходных данных. Кроме того, эти функции могут быть использованы для предсказания неизвестного выхода с использованием новых входных данных. Для структурированных данных мы можем использовать алгоритмы машинного обучения, но, когда размеры и объем данных, таких как изображения, аудиоданные, текстовые и видеоданные, увеличиваются, модели машинного

обучения не могут хорошо работать, поэтому требуется модель глубокого обучения. Экспертная система разработана как система, основанная на правилах, которая помогает в получении выводов. Это требуется, когда нет достаточного количества обучающих данных для обучения моделей машинного или глубокого обучения. В целом создание системы искусственного интеллекта требует для создания выводов сочетания экспертных систем, алгоритмов машинного обучения и алгоритмов глубокого обучения.

Машинное обучение можно определить как систему, в которой алгоритм обучается на примерах в отношении некоторой задачи, определенной ранее, и эффективность обучения увеличивается по мере ввода в систему все большего количества данных. Задачи могут быть определены как контролируемые, где выход/результат известен заранее; неконтролируемые, где выход/результат не известен заранее; и с подкреплением, где действия/результаты всегда определяются средой обратной связи, а обратная связь может быть вознаграждением или штрафом. Что касается алгоритмов обучения, то их можно разделить на следующие категории – линейные, детерминированные, аддитивные и мультипликативные, алгоритмы на основе деревьев, ансамблевые алгоритмы и на основе графов. Критерии эффективности могут быть определены в соответствии с выбором алгоритма. Объяснение решений модели AI называется объяснимым AI (explainable AI – XAI).

Необходимость XAI

Рассмотрим причину, по которой модели AI называют моделями «черного ящика». Рисунок 1.2 объясняет классический сценарий моделирования, когда набор независимых переменных передается через функцию, которая предопределена для получения выхода. Полученный результат сравнивается с истинным результатом, чтобы оценить, соответствует ли функция данным или нет. Если функция плохо подходит, то необходимо либо преобразовать данные, либо рассмотреть возможность использования другой функции, чтобы она подходила к данным. Но эксперимент ручной, и каждый раз, когда происходит обновление данных, статистикам или специалистам по моделированию приходится заново калибровать модели и снова проверять, соответствуют ли данные модели. Вот почему классический способ создания прогностических моделей для обработки выводов зависит от человека и всегда подлежит более чем одной интерпретации. Временами заинтересованным сторонам трудно доверять модели, так как все варианты, предложенные многими экспертами, могут звучать в определенном смысле хорошо, но нет обобщения. Таким образом, классическую систему разработки моделей, показанную на рис. 1.2, сложно реализовать в мире систем искусственного интеллекта, где данные постоянно меняются. Зависимость калибровки от человека является узким местом, поэтому существует необходимость в современной системе генерации выводов с использованием динамических алгоритмов.

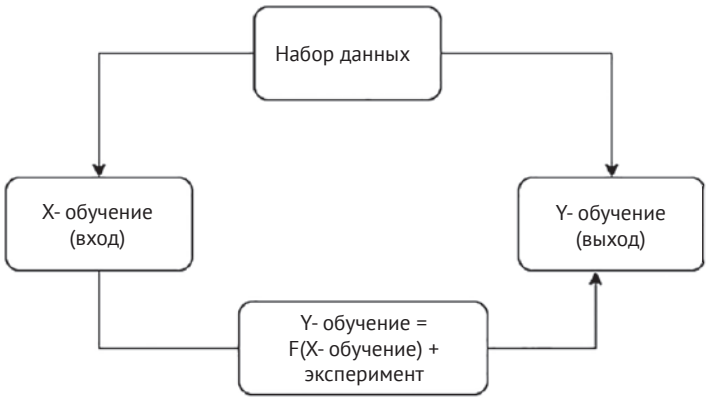

Рис. 1.2. Классическая система обучения модели

На рис. 1.2 показан классический сценарий разработки модели, когда модель может быть представлена через уравнение и это уравнение легко интерпретировать и просто объяснить кому угодно, но существование интерпретации на основе формул не всегда возможно в мире AI. На рис. 1.3 показана структура поиска наилучших возможных функций, которые производят выход, используя входные данные. Здесь нет ограничений модели конкретной функцией, например линейной или нелинейной. В этой структуре обучение происходит через множество итераций, и с помощью перекрестной проверки определяется лучшая модель. Проблема с моделями AI заключается в интерпретируемости и объяснимости, так как многие алгоритмы сложны, и поэтому нелегко объяснить прогнозы каждому.

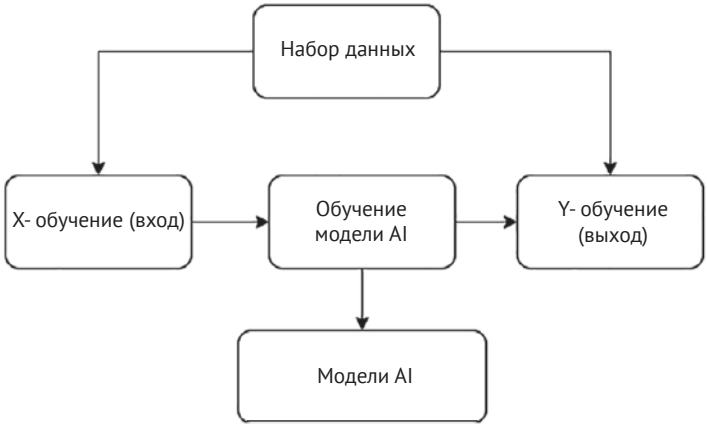

Рис. 1.3. Процесс обучения модели AI

С развитием компьютерных программ и алгоритмов разработчикам стало очень трудно искать различные виды линейных и нелинейных функций, и оценка таких функций также стала чрезвычайно сложной. Модели машин-

ного или глубокого обучения берут на себя поиск функций, которые хорошо подходят к обучающим данным. Рисунок 1.3 поясняет, что машина определяет окончательную модель, обеспечивающую лучшую эффективность не только с точки зрения точности, но также стабильности и надежности при генерировании прогнозов. Когда функциональная связь между входом и выходом четко определена, возникает меньше двусмысленности и прогнозы становятся прозрачными. Однако, когда модели AI делают выбор сложной функциональной связи, это очень трудно понять конечному пользователю. Поэтому модели AI считаются «черным ящиком». В этой книге мы хотим сделать модель «черного ящика» интерпретируемой, чтобы решения AI становились все более развертываемыми и адаптируемыми.

Повседневное использование моделей AI для принятия решений требует прозрачности, непредвзятости и этики. Существуют различные сценарии, в которых в настоящее время не хватает объяснимости:

- кто-то подает заявку на получение кредитной карты, и модель AI отклоняет его заявку. Важно объяснить, почему заявка была отклонена и какие корректирующие действия может предпринять заявитель, чтобы изменить свое поведение;
- в медицинской диагностике, основанной на образе жизни и жизненно важных параметрах, модель AI предсказывает, будет ли у человека диабет, или нет. Здесь если модель предсказывает, что у человека может развиться диабет, то она также должна объяснить, почему и каковы факторы, способствующие развитию заболевания в будущем;
- автономные транспортные средства идентифицируют объекты на дороге и принимают четкие решения. В этом случае также необходимо четкое объяснение того, почему они принимают эти решения.

Существует множество других примеров использования, где объяснения, подтверждающие вывод модели, имеют решающее значение. Человеку свойственно не принимать то, что он не может интерпретировать или понять. Поэтому снижается фактор доверия к прогнозам модели AI. Мы используем модели AI для устранения предвзятости при принятии решений человеком. Однако решения будут опасными, если результат не будет оправданным, правомерным и прозрачным. С другой стороны, можно утверждать, что если мы не можем интерпретировать и объяснить решения моделей AI, то зачем их использовать. Причинами их использования являются точность и эффективность моделей. Всегда будет существовать компромисс между эффективностью модели и ее объяснимостью. Рисунок 1.4 объясняет компромисс между этими двумя понятиями.

На рис. 1.4 горизонтальная ось показывает производительность или точность модели, а вертикальная – интерпретацию и объяснение модели. Система, основанная на правилах, находится в позиции, где эффективность не является оптимальной, однако интерпретируемость хороша. Напротив, модели на основе глубокого обучения обеспечивают превосходную эффективность и хорошую точность при меньшей интерпретируемости и объяснимости.

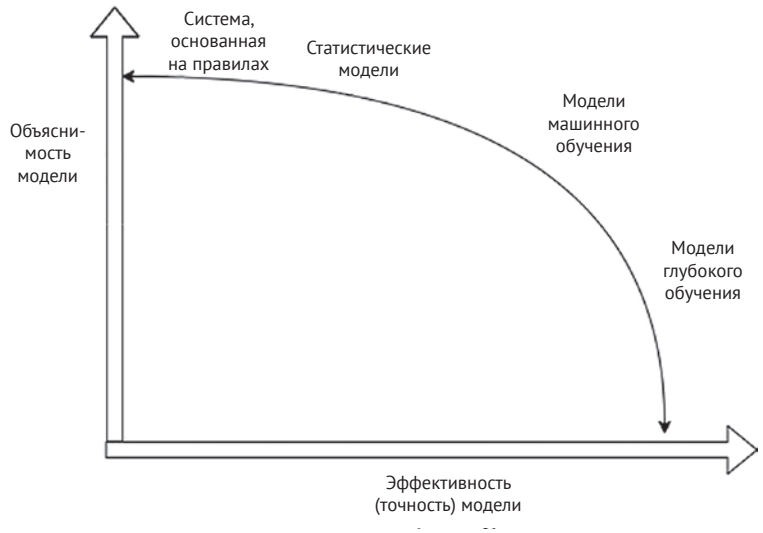

Рис. 1.4. Компромисс между объяснимостью и эффективностью (точностью) модели

Сравнение объяснимости с интерпретируемостью

Существует разница между интерпретируемостью и объяснимостью модели. Интерпретация – это смысл прогнозов. Объяснимость – это то, почему модель предсказывает что-либо и почему кто-то должен доверять модели. Чтобы лучше понять разницу, давайте рассмотрим реальный пример прогноза продаж, где факторами, которые помогают в прогнозировании, являются расходы на рекламу, качество продукции, источники для создания рекламы, размер рекламы и т. д. Каждый фактор имеет коэффициент после выполнения регрессионного моделирования. Коэффициенты могут быть интерпретированы как приращение продаж в результате дельта-изменения одного фактора, например расходов на рекламу. Однако, если вы прогнозируете, что продажи в следующем месяце составят 20 000 долл., когда среднемесячные продажи исторически были меньше или равны 15 000 долл., это требует объяснения. В рамках объяснимости модели нам необходимо следующее:

- интерпретируемость модели должна обеспечить естественность принятия решений и отсутствие предвзятости в прогнозе;
- дифференциация ложной причинности от истинной причинности, которая помогает сделать прогнозы прозрачными;
- необходимость создания объяснимых моделей без ущерба для высокой производительности опыта обучения и итераций;
- возможность лицам, принимающим решения, доверять моделям AI.

«Приложения и продукты XAI станут новым трендом в 2021 году, поскольку спрос на интерпретируемость, доверие и этику в AI громкий и ясный». – Прадипта Мишра (Pradeepta Mishra; источник – различные научно-исследовательские отчеты).

Вокруг XAI постоянно ведутся исследования, которые делают все возможное, чтобы объяснить конечным пользователям модели AI и их поведение с целью повысить уровень принятия моделей. Теперь возникает вопрос – кто является конечными пользователями XAI? Ими являются:

- кредитные инспекторы, которые оценивают заявки на получение кредитов, кредитные запросы и многое другое. Если они понимают, какие решения принимаются, то могут помочь обучить клиентов корректировать свое поведение;
- специалисты по анализу данных, оценивающие собственные решения и обеспечивают введение улучшений в модели. Является ли это лучшей моделью, которую можно сделать, используя имеющийся набор данных?
- старшие менеджеры, которым необходимо соответствовать нормативным требованиям на высоком уровне;
- руководители предприятий, кому необходимо доверять решениям «черного ящика» AI и кто ищет любые исторические свидетельства, на которые они могут положиться;
- руководители службы поддержки клиентов, кому необходимо отвечать на жалобы и объяснять решения;
- внутренние аудиторы и регуляторы, обязанные обеспечить прозрачность процесса, основанного на данных.

Целью XAI является достижение следующих показателей:

- **доверия:** точность прогноза является четкой функцией качества данных, истинности причинно-следственной связи и выбора подходящего алгоритма. Однако модели могут генерировать ложные срабатывания в процессе прогнозирования. Если модели генерируют много ложных срабатываний, то конечный пользователь потеряет доверие к модели. Таким образом, важно передать доверие к модели конечному пользователю;
- **ассоциаций:** модели ML или DL учатся делать прогнозы на основе ассоциаций между различными признаками. Ассоциации могут быть корреляциями или просто ассоциациями. Необъяснимые корреляции являются ложными корреляциями, которые делают модель невозможной для интерпретации. Следовательно, важно уловить истинные корреляции;
- **надежности:** уверенность в модели, стабильность модели в прогнозах, а также устойчивость модели также очень важны. Это необходимо для того, чтобы модели AI вызывали больше доверия, и для того, чтобы конечный пользователь был достаточно уверен в прогнозах модели. Если этого нет, то ни один пользователь не будет доверять моделям;
- **справедливости:** модели AI должны быть справедливыми и соответствовать этическим нормам. Они не должны дискриминировать религию, пол, класс и расу при генерировании прогнозов;
- **идентичности:** модели AI должны быть способны сохранять соображения конфиденциальности без раскрытия личности человека. Конфиденциальность и управление идентификацией при создании XAI очень важны.

Типы объяснимости

Интерпретируемость машинного обучения является неотъемлемой частью объяснимости модели. Существуют различные классификации интерпретаций моделей:

- **внутреннее объяснение:** в эту категорию попадают простые модели, такие как простые линейные регрессионные модели и модели на основе дерева решений, где простое условие «если/иначе» (if/else) может объяснить предсказания. Это означает, что XAI присущ самой модели, и нет необходимости проводить какой-либо постанализ;
- **объяснение post-hoc:** сложные модели, такие как нелинейные, ансамблевые древовидные, стохастическая древовидная модель с усилением градиента и стековые модели, где необходимо уделить больше внимания созданию объяснимости;
- **конкретная модель:** существует набор объяснений, которые могут быть получены из конкретного типа модели, не более того. Например, модель линейной регрессии не обеспечивает значимость признаков. Однако коэффициенты линейной регрессионной модели кто-то может использовать в качестве косвенного показателя;
- **не зависящие от модели:** эти объяснения кто-то может интерпретировать, глядя на пару комбинаций обучающих входных данных и обучающих выходных данных. В этой книге мы рассмотрим объяснения, не зависящие от модели, в последующих главах;
- **локальная интерпретация:** дает представление об отдельных предсказаниях, что является интерпретацией одной точки данных. Например, если заемщик предсказан моделью как склонный к дефолту, то почему это так? Это локальная интерпретация;
- **глобальная интерпретация:** дает представление о глобальном понимании прогнозов для всех точек данных, общем поведении модели и многом другом;
- **сублокальная интерпретация:** объясняет локальные интерпретации для группы точек данных, а не всех точек данных. Это отличается от локальной интерпретации;
- **текстовые объяснения:** включают числовую часть, а также язык для передачи значения определенных параметров модели;
- **визуальные объяснения:** визуальные объяснения хороши, но иногда они недостаточно интуитивны для объяснения прогнозов, поэтому очень необходимы визуальные объяснения, сопровождаемые текстовой интерпретацией.

Инструменты для объяснимости моделей

Существуют различные инструменты и механизмы для создания объяснимости из ML- и DL-моделей. Библиотеки Python с открытым исходным кодом имеют некоторые преимущества и недостатки. В примерах на протяжении всей этой

книги мы будем использовать сочетание библиотек Python с открытым исходным кодом и общедоступных наборов данных с различных веб-сайтов. Далее перечислены инструменты, которые необходимо установить, и среда, которую необходимо настроить.

SHAP

Библиотека SHAP (SHapley Additive exPlanations) – это унифицированный подход на базе Python для объяснения результатов любой модели машинного обучения. Библиотека SHAP Python основана на теории игр с локальными объяснениями. Подход теории игр – это способ получить прогнозы при наличии одного фактора по сравнению с его отсутствием. Если происходит значительное изменение в ожидаемом результате, значит, фактор очень важен для целевой переменной. Этот метод объединяет несколько предыдущих методов для объяснения результатов, генерируемых моделями машинного обучения. Фреймворк SHAP может быть использован для различных типов моделей, за исключением моделей на основе временных рядов (см. рис. 1.5). Библиотека SHAP может быть использована для осмысления моделей.

Для установки SHAP можно использовать следующие методы:

```
! Pip install shap                    (из ноутбука Jupyter)
conda install -c conda-forge shap     (с помощью терминала)
   !pip3 install shap
```

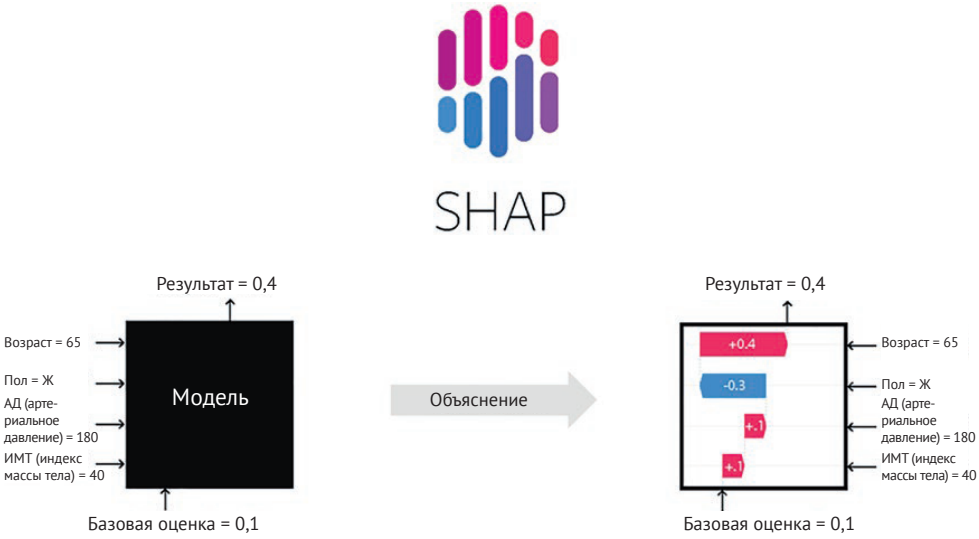

Рис. 1.5. Значения Шепли и пример объяснения изображения

LIME

LIME расшифровывается как локальные интерпретирумые не зависящие от модели объяснения (Local Interpretable Model-Agnostic Explanations). Локальное относится к объяснению локальности класса, который был предсказан мо-

делью. Поведение классификатора при локальности дает хорошее понимание прогноза. Интерпретируемость означает, что если предсказание не может быть интерпретировано человеком, то в нем нет смысла. Следовательно, предсказания классов должны быть интерпретируемыми. Независимость от модели подразумевает, что вместо понимания конкретного типа модели система и метод должны быть способны генерировать интерпретации.

Рис. 1.6. Пример классификации грибов

Проблема классификации текста (например, анализ настроения) – это когда на вход подаются предложения в виде документов, а на выходе получается класс (см. рис. 1.6). Когда модель предсказывает положительное настроение для предложения, нам необходимо знать, какие слова заставили модель предсказать класс как положительный. Эти векторы слов иногда очень просты, например отдельные слова. Иногда они сложны (например, вкрапления слов), и в этом случае нужно знать, как модель интерпретировала вкрапления слов и как это влияет на классификацию. В этих сценариях LIME чрезвычайно полезна для понимания смысла моделей машинного обучения и глубокого обучения. LIME – это библиотека на базе Python, которую можно использовать для демонстрации ее работы. Для ее установки необходимо выполнить следующее:

```
! pip install lime
```

ELI5

ELI5 – это библиотека на базе Python, предназначенная для создания объяснимого конвейера AI, который позволяет визуализировать и отлаживать различные модели машинного обучения с помощью унифицированного API. Она имеет встроенную поддержку нескольких ML-фреймворков и предоставляет возможность объяснять «черные ящики» моделей. Цель библиотеки – сделать объяснения простыми для всех видов моделей «черного ящика» (см. рис. 1.7).

Вес	Характеристика
0.3717	взаимосвязь
0.1298	семейное положение
0.1247	длительность обучения
0.1108	прирост капитала
0.0611	потеря капитала
0.0362	возраст
0.0307	занятость
0.0298	пол
0.0289	количество рабочих часов в неделю
0.0188	рабочий класс
0.0161	родная страна
0.0160	раса
0.0132	вес выборки[1]
0.0123	образование

Рис. 1.7. Пример изображения в ELI5

Рисунок 1.7 из ELI5 показывает важность факторов в прогнозировании дохода класса в примере использования классификации доходов, который мы рассмотрим в последующих главах. Установка ELI5 на Python может быть выполнена с помощью следующего синтаксиса:

```
!pip install eli5
```

Это требует обновления многих библиотек на базе Python, и вам, возможно, придется подождать некоторое время, пока это произойдет.

Skater

Skater – это унифицированный фреймворк с открытым исходным кодом, позволяющий интерпретировать модели для всех форм моделей, чтобы помочь построить интерпретируемую систему машинного обучения, что часто необходимо для использования в реальном мире. Skater поддерживает алгоритмы для прояснения изученных структур модели «черного ящика» как глобально (вывод на основе полного набора данных), так и локально (вывод на основе индивидуального прогноза).

[1] Примерная оценка количества людей, которое представляет каждая строка данных

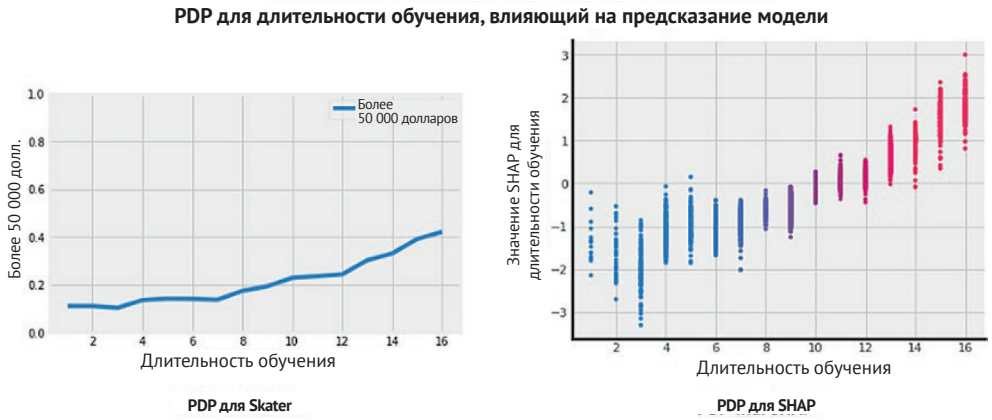

Рис. 1.8. Пример изображения, показывающего PDP[2] с использованием SHAP и Skater

Skater позволяет реализовать это видение, предоставляя по мере необходимости возможность выводить и отлаживать политики принятия решений модели, т. е. обеспечивая «участие человека в процессе» (см. рис. 1.8). Чтобы установить библиотеку Skater, можно использовать следующую команду:

```
!pip install skater
```

Skope_rules

Skope-rules нацелен на изучение логических, интерпретируемых правил для «обследования» целевого класса (т. е. обнаружения с высокой точностью экземпляров этого класса). Skope-rules – это компромисс между интерпретируемостью дерева решений и моделирующей способностью случайного леса (см. рис. 1.9).

Вышеупомянутые библиотеки на базе Python в основном имеют открытый исходный код и бесплатны для использования и интеграции в любое программное приложение. Однако существует множество корпоративных инструментов и фреймворков, таких как H2O.ai. Область XAI является относительно новой, так как еще ведутся исследования в направлении упрощения интерпретации моделей. Инструменты и фреймворки выходят на передний план, чтобы сделать этот процесс доступным для применения в промышленности.

[2] PDP (Partial Dependence Plot) – график частичной зависимости, показывает краевой эффект одного или двух признаков на прогнозируемый результат модели машинного обучения. – *Прим. перев.*

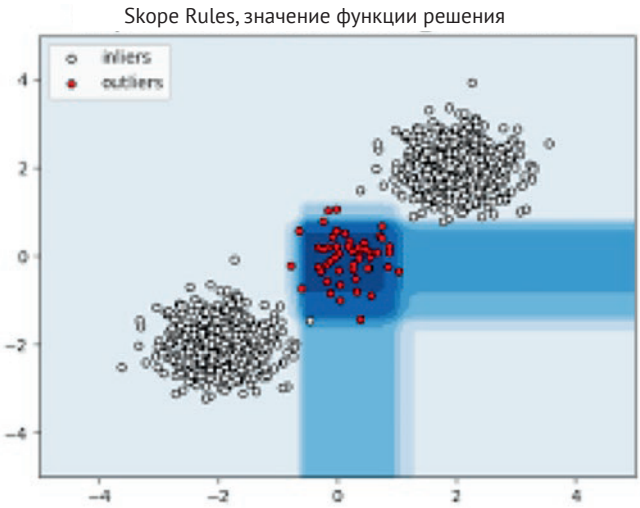

Рис. 1.9. Пример изображения Skope-rules

Методы XAI для ML

Уровни прозрачности моделей машинного обучения можно разделить на три группы – прозрачность алгоритмов, декомпозицию параметров и гиперпараметров и воспроизводимость одного и того же результата в аналогичных ситуациях. Некоторые модели ML по своей конструкции интерпретируемы, а некоторые требуют набора других программ, чтобы сделать их объяснимыми. Рисунок 1.10 объясняет методы объяснимости моделей.

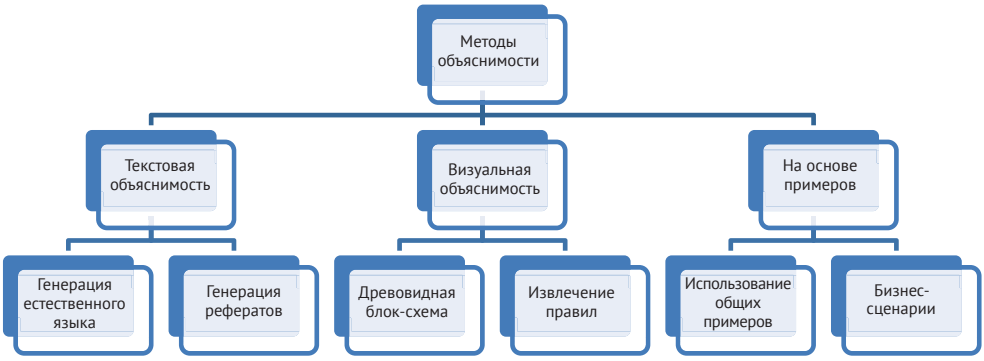

Рис. 1.10. Методы XAI

Существует три метода объяснения модели:

- текстовое объяснение требует уточнения значения математической формулы, параметра модели или метрик, определенных моделью. Интерпретации могут быть разработаны на основе определенного шаблона, где сюжетная линия должна быть подготовлена заранее и только параметры должны быть введены в шаблон. Существует два различных

подхода для достижения этой цели – использование методов генерации естественного языка (Natural Language Generation – NLG), для чего необходимо собрать текст и составить предложение, описывающее объект, и использование генерации реферата;

- визуальное объяснение может быть предоставлено с помощью пользовательских графиков и диаграмм. Древовидные графики довольно просты и понятны для конечных пользователей. Каждый древовидный метод опирается на набор правил. Если эти правила могут быть показаны пользователю в виде простых инструкций «если/иначе» (`if/else`), то это будет гораздо более мощным;
- метод, основанный на примерах, гарантирует, что мы можем взять обычные повседневные примеры для объяснения модели путем проведения параллелей. Также для объяснения моделей можно использовать обычный бизнес-сценарий.

Совместимые с XAI модели

Давайте посмотрим на текущее состояние моделей, их характер, насколько они совместимы с XAI, и нужны ли этим моделям для объяснимости дополнительные фреймворки:

- **линейные модели** – модели линейной регрессии или логистической регрессии легко интерпретировать, анализируя значение их коэффициента, который представляет собой число. Эти значения очень легко интерпретировать. Однако, если мы распространим это на регуляризованные регрессионные семейства, это становится очень сложным для объяснения. Обычно мы придаем большее значение отдельным признакам. Мы не учитываем особенности взаимодействия. Сложность модели возрастает, если мы включаем взаимодействия, такие как аддитивные, мультипликативные, полиномиальные взаимодействия второй или третьей степени. В этих сложных сценариях математический результат требует более простой интерпретации;
- **модели прогнозирования временных рядов** – это также очень простые модели, они следуют регрессионному типу сценария, который легко объяснить с помощью параметрического подхода;
- **модели на основе деревьев** более просты для анализа и также очень интуитивно понятны для интерпретации человеком. Однако эти модели часто не обеспечивают лучшей точности и производительности. Им также не хватает устойчивости, и они имеют присущие им проблемы предвзятости и чрезмерной подгонки. Поскольку недостатков так много, их интерпретация не имеет смысла для конечного пользователя;
- **ансамблевые модели** – существуют три различных типа ансамблевых моделей: упаковка, повышение и укладка. Все три типа не обладают достаточной объяснимостью. Необходимо упрощенное описание, чтобы передать результаты модели. Важность признаков также должна быть упрощена;

- **математические модели** – машины, поддерживающие векторную математику, используются для задач, основанных на регрессии и классификации. Эти модели довольно сложны для объяснения, поэтому очень важно упрощение модели;
- **модели глубокого обучения** – модели глубоких нейронных сетей (Deep Neural Network – DNN) обычно имеют более трех скрытых слоев. Помимо слоев модели глубокого обучения, существуют различные параметры настройки модели, такие как веса, типы регуляризации, сила регуляризации, типы функций активации для различных слоев, типы функций потерь, используемые в модели, и алгоритмы оптимизации, включающие скорость обучения и параметры импульса. Все это очень сложно по своей природе и требует упрощенной структуры для интерпретации;
- **конволюционная нейронная сеть** (Convolutional Neural Network – CNN) – это еще один тип нейросетевой модели, которая обычно применяется для обнаружения объектов и задач, связанных с классификацией изображений. Она рассматривается как полная модель «черного ящика». В ней есть слои свертки, максимальное или среднее число слоев объединения и многое другое. Если кто-то спросит, почему эта модель предсказала кошку как собаку, можем ли мы объяснить, что пошло не так? В настоящее время ответ отрицательный. Требуется большая работа по объяснению этой модели конечному пользователю;
- **рекуррентные нейронные сети** (Recurrent Neural Networks – RNNs) – рекуррентные нейронные модели обычно применяются для классификации текстов и предсказаний текста. Существуют различные варианты, такие как сеть с долговременной краткосрочной памятью (long short term memory – LSTM) и двунаправленные LSTM, очень сложные для объяснения. Есть постоянная потребность в более совершенных структурах и методах, которые можно использовать для объяснения таких моделей;
- **модели, основанные на правилах,** – это очень простые модели, поскольку нам нужны только условия `if/else` для создания таких моделей.

XAI удовлетворяет требованиям ответственного AI

Ответственный искусственный интеллект – это структура, в которой объяснимость, прозрачность, этика и подотчетность обеспечиваются в различных программных приложениях, цифровых решениях и продуктах. Развитие искусственного интеллекта стремительно создает множество возможностей в различных областях, где эти технологии затрагивают жизнь простых людей, поэтому искусственный интеллект должен быть ответственным, а решения – объяснимыми.

Семь основных принципов ответственного искусственного интеллекта являются критически важной частью объяснимости (рис. 1.11). Давайте рассмотрим каждый из них.

- **Справедливость**. Предсказания, генерируемые системами AI, не должны приводить к дискриминации людей по их касте, вероисповеданию, религии, полу, политическим убеждениям, этнической принадлежности и т. д., поэтому требуется большая степень справедливости.

Рис. 1.11. Основные принципы ответственного AI

- **Этика**. В стремлении к построению интеллектуальных систем мы не должны забывать об этике при сборе данных.
- **Прозрачность**. Прогнозы моделей и методы их генерации должны быть прозрачными.
- **Конфиденциальность**. При разработке систем AI должны быть защищены персональные данные (персонализированная идентифицируемая информация – personalized identifiable information – PII).
- **Информационная безопасность**. Интеллектуальные системы должны быть безопасными.
- **Ответственность**. В случае ошибочного прогноза модель AI должна быть способна взять на себя ответственность за устранение проблемы.
- **Безопасность**. Когда модели AI принимают решения по навигации самодвижущихся автомобилей, роботизированной стоматологической хирургии и медицинской диагностике, любой неверный прогноз может привести к опасным последствиям.

Многие организации находятся в процессе подготовки руководящих принципов и стандартов для использования AI в своих решениях, чтобы избежать непреднамеренных негативных последствий в будущем. Возьмем организацию A. Она использует AI для прогнозирования объема продаж. AI предсказывает, что объем продаж будет на 30 % выше среднего, поэтому предприятие делает запасы продукта и мобилизует рабочую силу для поддержки продаж. Но если фактические продажи окажутся на уровне среднего исторического уровня продаж, то создание этих запасов было напрасным. Здесь прогноз AI оказался неверным. С помощью объяснимости модели эта ситуация могла бы быть проанализирована, и, возможно, модель могла быть исправлена.

Оценка XAI

Не существует единого стандарта для оценки различных объяснений, генерируемых библиотеками на основе Python через интернет. Процесс XAI должен следовать следующим шагам при оценке объяснений:

- каждая страта должна иметь отдельное объяснение. Если у нас есть набор данных, который не может быть использован для обучения модели из-за большого объема, мы обычно делаем выборку из этого набора данных. Если мы используем стратифицированную выборку, то каждая страта должна иметь отдельное объяснение;

- **ограничения по времени**. Мы знаем, что реальные наборы данных достаточно велики. Даже если мы работаем с распределенными вычислительными системами, объяснения, генерируемые библиотеками XAI, как правило, не должны занимать много времени;
- **инвариантность экземпляра**. Если точки данных идентичны по своим атрибутам, они должны быть частью одной и той же группы и, следовательно, должны давать схожие интерпретации.

В различных проектах и инициативах в области AI, когда мы делаем прогнозы, часто возникает вопрос, почему кто-то должен доверять нашей модели. В предиктивной аналитике, машинном или глубоком обучении существует компромисс между тем, что было предсказано, и тем, почему это было предсказано. Если предсказание соответствует ожиданиям человека, то это хорошо. Если оно выходит за рамки человеческих ожиданий, то нам нужно знать, почему модель приняла такое решение. Прогнозирование и отклонение от ожиданий – это нормально, если речь идет о сценарии с низким уровнем риска, например о таргетировании клиентов, цифровом маркетинге или рекомендации контента. Однако в условиях высокого риска, таких как клинические испытания или система тестирования лекарств, незначительное расхождение между прогнозом и ожиданиями имеет большое значение, и возникнет множество вопросов о том, почему модель сделала такой прогноз. Как человеческие существа, мы считаем себя выше всех. Возникает любопытство, как модель пришла к такому прогнозу и почему этого не сделал человек. Система XAI – это отличный инструмент для выявления предвзятости, присущей процессу машинного обучения. XAI помогает нам выяснить, где именно появляется предвзятость.

Поскольку многие люди не в состоянии объяснить результаты работы модели машинного обучения, они не могут обосновать решения модели и поэтому не готовы к использованию моделей AI (рис. 1.12). Эта книга является попыткой популяризировать концепцию фреймворков XAI для отладки моделей машинного и глубокого обучения, чтобы повысить уровень внедрения AI в промышленности. В средах с высоким уровнем риска при использовании существуют нормативные требования и требования аудита для обоснования решения модели. Книга организована таким образом, чтобы обеспечить практическое выполнение фреймворков XAI для задач, связанных с контролируемой регрессией, классификацией, обучением без учителя, кластеризацией и сегментацией. Кроме того, некоторые модели прогнозирования временных рядов нуждаются в XAI для описания прогнозов. Далее, фреймворки XAI могут быть использованы для решения задач классификации неструктурированного текста. Один фреймворк XAI не подходит для всех типов моделей, поэтому мы собираемся обсудить различные типы библиотек Python с открытым исходным кодом и их использование для создания объяснений на основе XAI.

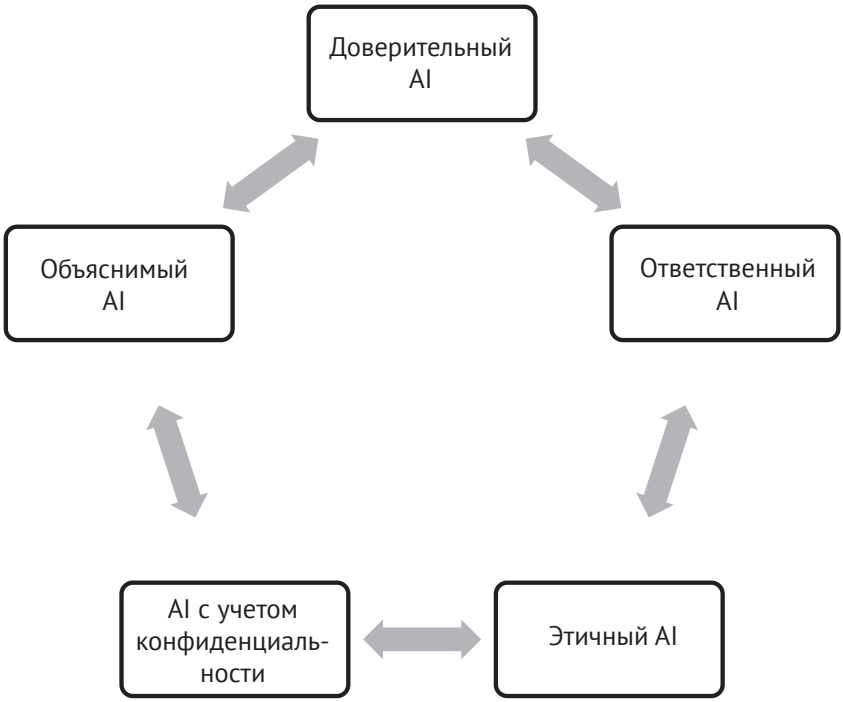

Рис. 1.12. Что необходимо для повышения уровня внедрения AI

Изучение справедливости модели также требует моделирования сценариев «что, если» (what-if) с использованием результатов прогнозов. Мы рассмотрим и это. Затем обсудим контрафактные и контрастные объяснения для моделей AI. Мы рассмотрим объяснимость для моделей глубокого обучения, экспертных систем, основанных на правилах, а также объяснения, не зависящие от модели, для инвариантности предсказаний и для задач компьютерного зрения, использующих различные фреймворкиXAI. Интерпретируемость и объяснимость моделей – ключевые темы этой книги. Существуют математические формулы и методы, которые обычно используются для объяснения решений, принимаемых моделями AI. Читателям предоставляются методы программных библиотек, классы, фреймворки и функции, а также методы их использования для объяснения моделей, прозрачности, надежности, этики, предвзятости и интерпретируемости. Если человек может понять причины решения, принятого моделью AI, это даст пользователю гораздо больше возможностей для внесения поправок и рекомендаций.

Заключение

Интерпретируемость и объяснимость модели необходимы для всех процессов, использующих AI для прогнозирования чего-либо, потому что нам нужно знать причины, стоящие за прогнозом. В этой главе вы узнали следующее:

- основы объяснимости и интерпретируемости моделей;

- этические соображения при применении AI и предвзятость прогнозов, генерируемых моделями AI;
- надежность моделей AI при создании прогнозов в различных случаях использования;
- методы и системы для интерпретации линейных моделей, которые используются в AI, нелинейные модели и модели временных рядов, используемые в AI;
- наиболее сложные ансамблевые модели, объяснимость и интерпретируемость с использованием таких фреймворков, как Lime, SHAP, Skater, ELI5 и др.;
- объяснимость моделей для неструктурированных данных и задач, связанных с обработкой естественного языка.