

Я посвящаю эту книгу моему покойному отцу, который всегда вдохновлял меня стремиться к следующему уровню, никогда не останавливаясь на достигнутом и продолжать двигаться вперед.

Люблю тебя, папа, ты бы гордился этой книгой.

Трем женщинам в моей жизни – моей жене Праджне и моим дочерям, Аарши и Аади, которые всегда поддерживали и любили меня. Завершение этой книги было бы невозможно без их поддержки.

Оглавление

Об авторе	10
О рецензентах	11
Благодарности	12
Введение	13
Глава 1. Объяснимость и интерпретируемость модели.....	15
Создание основ	15
Искусственный интеллект	16
Необходимость ХАИ.....	17
Сравнение объяснимости с интерпретируемостью.....	20
Типы объяснимости	22
Инструменты для объяснимости моделей.....	22
SHAP	23
LIME	23
ELI5	24
Skater	25
Skope_rules.....	26
Методы ХАИ для МЛ.....	27
Совместимые с ХАИ модели	28
Объяснимый ИИ удовлетворяет требованиям ответственного ИИ.....	29
Оценка ХАИ	31
Заключение	33
Глава 2. Этика, предвзятость и надежность ИИ	34
Основы этики ИИ	34
Предвзятость в ИИ.....	37
Предвзятость данных	37
Алгоритмическая предвзятость	37
Процесс снижения предвзятости	38
Предвзятость интерпретации.....	38
Предвзятость при обучении	39
Надежность в ИИ	42
Заключение	44
ГЛАВА 3. Объяснимость для линейных моделей.....	45
Линейные модели.....	45
Линейная регрессия	45

VIF и проблемы, которые он может породить.....	53
Окончательная модель.....	57
Объяснимость модели.....	57
Доверие к модели ML: SHAP.....	59
Локальное объяснение и индивидуальные прогнозы в модели ML.....	62
Глобальное объяснение и общие прогнозы в модели ML.....	65
Объяснение LIME и модель ML.....	69
Объяснение Skater и модель ML.....	73
Объяснение ELI5 и модель ML.....	75
Логистическая регрессия.....	76
Интерпретация.....	85
Вывод LIME.....	86
Заключение.....	92

ГЛАВА 4. Объяснимость для нелинейных моделей 93

Нелинейные модели.....	93
Объяснение дерева решений.....	95
Подготовка данных для модели дерева решений.....	97
Создание модели.....	99
Дерево решений – SHAP.....	106
График частичной зависимости.....	106
PDP с использованием Scikit-Learn.....	115
Объяснение нелинейной модели – LIME.....	118
Нелинейное объяснение – Skope-Rules.....	121
Заключение.....	123

ГЛАВА 5. Объяснимость для ансамблевых моделей 124

Ансамблевые модели.....	124
Типы ансамблевых моделей.....	125
Почему ансамблевые модели?.....	125
Использование SHAP для ансамблевых моделей.....	128
Использование интерпретации, объясняющей модель повышения.....	133
Модель классификации ансамблей: SHAP.....	139
Использование SHAP для объяснения категориальных моделей повышения.....	146
Использование многоклассовой категориальной модели повышения SHAP.....	152
Использование SHAP для объяснения модели легкой GBM.....	154
Заключение.....	157

ГЛАВА 6. Объяснимость для моделей временных рядов 159

Модели временных рядов.....	159
Выбор подходящей модели.....	161
Стратегия прогнозирования.....	162
Доверительный интервал прогнозов.....	162
Что происходит с доверием?.....	163

Временные ряды: LIME	175
Заключение	178
ГЛАВА 7. Объяснимость для NLP.....	179
Задачи обработки естественного языка	179
Объяснимость для классификации текстов	180
Набор данных для классификации текста	180
Объяснение с помощью ELI5	182
Вычисление весов характеристик для локального объяснения	183
Локальное объяснение. Пример 1	184
Локальное объяснение. Пример 2	184
Локальное объяснение. Пример 3	185
Объяснение после удаления стоп-слов	185
Классификация текста на основе N-грамм.....	189
Объяснимость многоклассовой классификации текста по меткам	193
Локальное объяснение. Пример 1	198
Локальное объяснение. Пример 2	199
Локальное объяснение. Пример 3	201
Заключение	209
ГЛАВА 8. Справедливость модели ИИ, использующей сценарий «что, если»	210
Что такое WIT?	210
Установка WIT	211
Метрика оценки.....	220
Заключение	221
ГЛАВА 9. Объяснимость для моделей глубокого обучения	222
Объяснение моделей DL.....	222
Использование SHAP с DL.....	225
Использование Deep SHAP.....	225
Использование Alibi	225
Объяснитель SHAP для глубокого обучения	229
Еще один пример классификации изображений	231
Использование SHAP	234
Deep Explainer для табличных данных.....	237
Заключение	239
ГЛАВА 10. Контрфактуальные объяснения для моделей XAI	240
Что такое CFE?	240
Применение CFE	240
CFE с помощью Alibi	241
Контрфактуал для задач регрессии	248
Заключение	251

ГЛАВА 11. Контрастные объяснения для машинного обучения.....	252
Что такое SE для ML?.....	252
SEM, использующие Alibi.....	253
Сравнение оригинального изображения и изображения, сгенерированного автокодировщиком.....	258
Объяснения SEM для табличных данных	262
Заключение	267
ГЛАВА 12. Модельно независимые объяснения путем определения инвариантности прогноза	268
Что такое независимость от модели?.....	268
Что такое якорь?	268
Объяснения якорей с помощью Alibi	269
Якорь текста для классификации текста.....	273
Якорь изображения для классификации изображений	277
Заключение	280
ГЛАВА 13. Объяснимость модели для экспертных систем, основанных на правилах.....	281
Что такое экспертная система?	281
Прямая и обратная цепочки	282
Извлечение правил с помощью Scikit-Learn	283
Потребность в системе, основанной на правилах.....	289
Проблемы экспертной системы	290
Заключение	290
ГЛАВА 14. Объяснимость моделей для компьютерного зрения.....	291
Почему объяснимость для данных изображений?.....	291
Якорь изображения с помощью Alibi	292
Метод интегрированных градиентов	292
Заключение	297

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 Harpiest 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 (ИИ) 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 ИИ 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 ИИ 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 ИИ 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 ИИ. 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 ИИ, 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 ИИ и предвзятости прогнозов, генерируемых моделями ИИ. 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 ИИ. 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 ИИ. 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 ХАИ.

Сегодня у нас есть инженеры по ИИ 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 ИИ, 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 ИИ 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 ИИ 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 получают огромную пользу от этой книги. Аналогичным образом эта книга будет полезна и другим инженерам по ИИ. 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 ИИ и предвзятости прогнозов, генерируемых моделями ИИ, рассмотрим надежность моделей ИИ при создании прогнозов в различных случаях использования. Затем изучим методы и системы для интерпретации линейных, нелинейных моделей и моделей временных рядов, используемых в ИИ. 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 (ИИ) в различных областях, в том числе в розничной торговле, банковском деле, финансовых услугах, страховании, здравоохранении, производстве и отраслях, основанных на интернете вещей. ИИ является основой многих продуктов и решений, которые появляются в связи с быстрой цифровизацией различных бизнес-функций. Причина того, что ИИ 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-

того 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 ИИ, 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 ИИ

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 ИИ называется объяснимым ИИ (explainable ИИ – ХАИ).

Необходимость ХАИ

Рассмотрим причину, по которой модели ИИ 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.

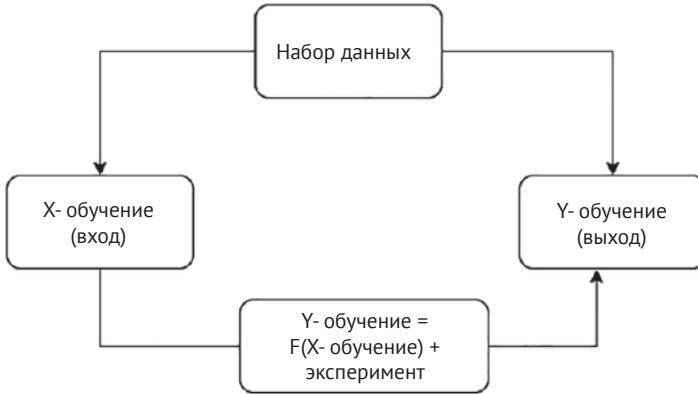


Рис. 1.2. Классическая система обучения модели

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 ИИ. 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 ИИ заключается в интерпретируемости и объяснимости, так как многие алгоритмы сложны, и поэтому нелегко объяснить прогнозы каждому.

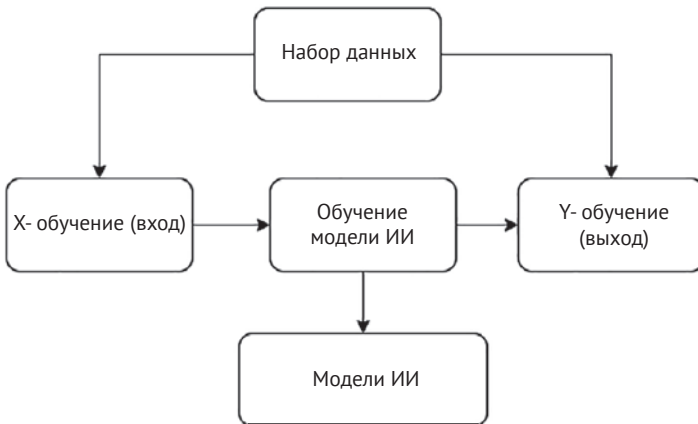


Рис. 1.3. Процесс обучения модели ИИ

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 ИИ делают выбор сложной функциональной связи, это очень трудно понять конечному пользователю. Поэтому модели ИИ считаются «черным ящиком». В этой книге мы хотим сделать модель «черного ящика» интерпретируемой, чтобы решения ИИ становились все более развертываемыми и адаптируемыми.

Повседневное использование моделей ИИ 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 ИИ 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 ИИ 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 ИИ. Мы используем модели ИИ 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 ИИ, 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.

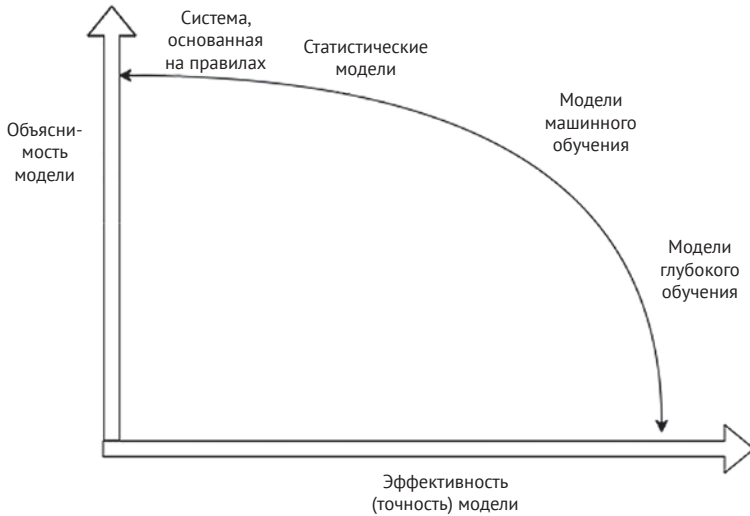


Рис. 1.4. Компромисс между объяснимостью и эффективностью (точностью) модели

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 ИИ.

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

Вокруг ХАИ постоянно ведутся исследования, которые делают все возможное, чтобы объяснить конечным пользователям модели ИИ и их поведение с целью повысить уровень принятия моделей. Теперь возникает вопрос – кто является конечными пользователями ХАИ? 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» ИИ 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 ХАИ 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 ИИ 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 ИИ должны быть справедливыми и соответствовать этическим нормам. Они не должны дискриминировать религию, пол, класс и расу при генерировании прогнозов;
- **идентичности:** модели ИИ должны быть способны сохранять соображения конфиденциальности без раскрытия личности человека. Конфиденциальность и управление идентификацией при создании ХАИ 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 ХАИ 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
```

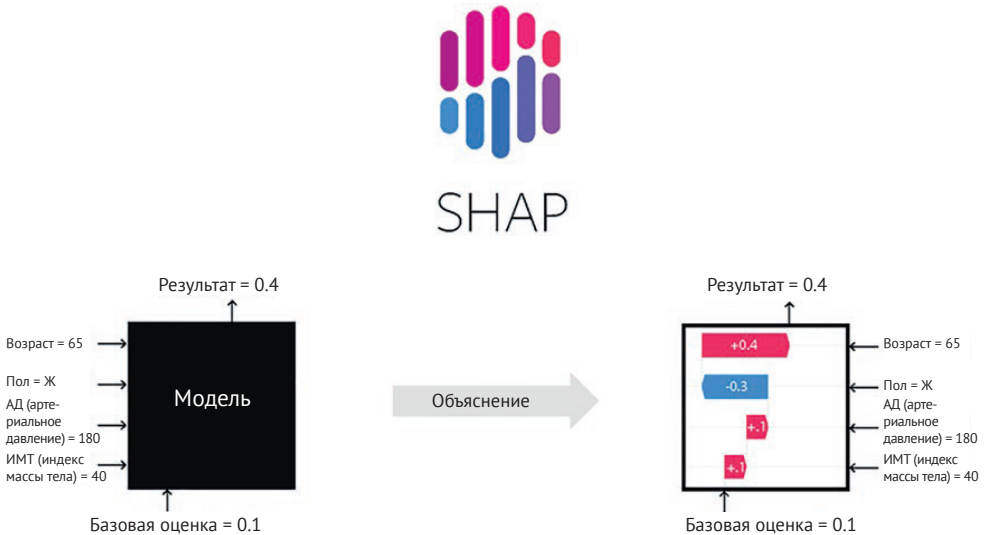


Рис. 1.5. Значения Шепли и пример объяснения изображения

LIME

LIME расшифровывается как локальные интерпретируемые 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 ИИ, 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).

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

e-Univers.ru