

УДК 004.438Python
ББК 32.973.22
М71

Прадипта Мишра

М71 Объяснимые модели искусственного интеллекта на Python. Модель искусственного интеллекта. Объяснения с использованием библиотек, расширений и фреймворков на основе языка Python / пер. с англ. С. В. Минца. – М.: ДМК Пресс, 2022. – 298 с.: ил.

ISBN 978-5-93700-124-5

В книге рассматриваются так называемые модели «черного ящика» для повышения адаптивности, интерпретируемости и объяснимости решений, принимаемых алгоритмами искусственного интеллекта (ИИ) с использованием библиотек Python XAI, TensorFlow 2.0+, Keras, а также пользовательских фреймворков с использованием Python Wrappers.

Издание предназначено специалистам по анализу данных, инженерам по внедрению моделей ИИ, а также может быть полезно бизнес-пользователям и руководителям проектов, использующих результаты работы решений ИИ в своей деятельности.

УДК 004.438Python
ББК 32.973.22

First published in English under the title Practical Explainable ИИ Using Python
This edition has been translated and published under licence from APress Media, LLC, part of Springer Nature.

APress Media, LLC, part of Springer Nature takes no responsibility and shall not be made liable for the accuracy of the translation.

Все права защищены. Любая часть этой книги не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами без письменного разрешения владельцев авторских прав.

ISBN (англ.) 978-1-4842-7157-5
ISBN (рус.) 978-5-93700-124-5

Copyright © 2022 by Pradeepta Mishra
© Оформление, издание, перевод, ДМК Пресс, 2022

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 ML.....	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 СЕ для ML?.....	252
СЕМ, использующие Alibi.....	253
Сравнение оригинального изображения и изображения,	
сгенерированного автокодировщиком.....	258
Объяснения СЕМ 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