

СЕРГ МАСИС

# ИНТЕРПРЕТИРУЕМОЕ МАШИННОЕ ОБУЧЕНИЕ НА PYTHON

НАУЧИТЕСЬ СОЗДАВАТЬ ИНТЕРПРЕТИРУЕМЫЕ  
ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ МОДЕЛИ  
НА ПРАКТИЧЕСКИХ ПРИМЕРАХ ИЗ РЕАЛЬНОЙ ЖИЗНИ



Материалы  
на [www.bhv.ru](http://www.bhv.ru)

**Packt**>

СЕРГ МАСИС

# ИНТЕРПРЕТИРУЕМОЕ МАШИННОЕ ОБУЧЕНИЕ НА PYTHON

НАУЧИТЕСЬ СОЗДАВАТЬ ИНТЕРПРЕТИРУЕМЫЕ  
ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ МОДЕЛИ  
НА ПРАКТИЧЕСКИХ ПРИМЕРАХ ИЗ РЕАЛЬНОЙ ЖИЗНИ

Санкт-Петербург

«БХВ-Петербург»

2023

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	<i>Евгений Рыбаков</i>
Зав. редакцией	<i>Людмила Гауль</i>
Перевод с английского	<i>Андрея Логанова</i>
Редактор	<i>Анна Кузьмина</i>
Компьютерная верстка	<i>Натальи Смирновой</i>
Корректор	<i>Светлана Крутярова</i>
Оформление обложки	<i>Зои Канторович</i>

© Packt Publishing 2021. First published in the English language under the title 'Interpretable Machine Learning with Python – (9781800203907)'

Впервые опубликовано на английском языке под названием 'Interpretable Machine Learning with Python – (9781800203907)'

Подписано в печать 02.12.22.  
Формат 70×100<sup>1/16</sup>. Печать офсетная. Усл. печ. л. 51,6.  
Тираж 1200 экз. Заказ № 5660.  
"БХВ-Петербург", 191036, Санкт-Петербург, Гончарная ул., 20.

Отпечатано с готового оригинал-макета  
ООО "Принт-М", 142300, М.О., г. Чехов, ул. Полиграфистов, д. 1

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

© Packt Publishing, 2021  
© Перевод на русский язык, оформление.  
ООО "БХВ-Петербург", ООО "БХВ", 2023

# Оглавление

---

Об авторе .....	15
О рецензентах.....	17
<b>Предисловие.....</b>	<b>19</b>
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
<b>Часть I. ВВЕДЕНИЕ В ИНТЕРПРЕТАЦИЮ МАШИННОГО ОБУЧЕНИЯ .....</b>	<b>27</b>
<b>Глава 1. Интерпретация, интерпретируемость и объяснимость: почему всё это важно? .....</b>	<b>29</b>
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
<b>Глава 2. Ключевые понятия интерпретируемости.....</b>	<b>52</b>
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
<b>Глава 3. Трудности интерпретации.....</b>	<b>76</b>
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 перцептроне .....	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
<b>Глава 6. Модельно-агностические методы локальной интерпретации .....</b>	<b>227</b>
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

<b>Глава 7. Якорные и контрфактические объяснения.....</b>	<b>264</b>
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 <i>alibi</i> .....	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
<b>Глава 8. Визуализация сверточных нейронных сетей.....</b>	<b>306</b>
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
<b>Глава 9. Методы интерпретации для многопеременного прогнозирования и анализа чувствительности .....</b>	<b>361</b>
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
<b>Часть III. НАСТРОЙКА НА ИНТЕРПРЕТИРУЕМОСТЬ .....</b>	<b>413</b>
<b>Глава 10. Отбор и конструирование признаков для обеспечения интерпретируемости .....</b>	<b>415</b>
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

<b>Глава 11. Ослабление систематического смещения и причинно-следственный вывод .....</b>	<b>459</b>
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**

<b>на интерпретируемость.....</b>	<b>515</b>
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
<b>Глава 13. Устойчивость к антагонизму .....</b>	<b>566</b>
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
Эвразивные 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 эвразивной 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
<b>Глава 14. Интерпретируемость машинного обучения: что дальше? .....</b>	<b>607</b>
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
<b>Предметный указатель .....</b>	<b>619</b>