

А. В. Протодьяконов  
П. А. Пылов  
В. Е. Садовников

# АЛГОРИТМЫ DATA SCIENCE

И ИХ ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ  
НА PYTHON

**А. В. Протодняконов  
П. А. Пылов  
В. Е. Садовников**

# **АЛГОРИТМЫ DATA SCIENCE**

**И ИХ ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ  
НА PYTHON**

Учебное пособие

**Москва Вологда  
«Инфра-Инженерия»  
2022**

УДК 004.89  
ББК 32.813  
П83

**Рецензенты:**

кандидат технических наук, доцент, заведующий кафедрой  
информационных и автоматизированных производственных систем  
ФГБОУ ВО «Кузбасский государственный технический университет»  
*И. В. Чичерин;*

доктор технических наук, профессор РАН, главный научный сотрудник  
ФГБНУ «Федеральный исследовательский центр угля и углехимии  
Сибирского отделения Российской академии наук»  
*А. Е. Майоров*

**Протодияконов, А. В.**

**П83** Алгоритмы Data Science и их практическая реализация на Python : учебное пособие / А. В. Протодияконов, П. А. Пылов, В. Е. Садовников. – Москва ; Вологда : Инфра-Инженерия, 2022. – 392 с. : ил., табл.  
ISBN 978-5-9729-1006-9

Рассмотрен полный каскад разработки моделей искусственного интеллекта. Проанализирована область Data Science, из которой выделены все необходимые для прикладной сферы алгоритмы машинного обучения, расположенные по уровню возрастания сложности работы с ними.

Для студентов, изучающих информационные технологии. Может быть полезно как начинающим программистам, так и специалистам высокого уровня.

УДК 004.89  
ББК 32.813

ISBN 978-5-9729-1006-9

© Протодияконов А. В., Пылов П. А., Садовников В. Е., 2022  
© Издательство «Инфра-Инженерия», 2022  
© Оформление. Издательство «Инфра-Инженерия», 2022

# ОГЛАВЛЕНИЕ

<b>Предисловие</b> .....	7
<b>Часть 1. Процесс машинного обучения</b> .....	8
Задачи машинного обучения.....	8
Модель и процесс машинного обучения.....	10
Понятие ETL.....	12
Понятие EDA.....	13
Подготовка данных.....	15
Разбиение выборки.....	18
Оптимизация гиперпараметров.....	21
Недообучение и переобучение.....	23
Смещение, разброс и ошибка данных.....	27
Использование HDF.....	30
<b>Часть 2. Метрики и модели общие</b> .....	33
Метод максимального правдоподобия.....	33
Метод наименьших квадратов.....	36
Аппроксимация пропусков в данных.....	38
Среднеквадратичная ошибка.....	40
Метрики и расстояния.....	42
<b>Часть практических навыков к 1-2</b> .....	45
Процесс ETL.....	45
Интерполяция и экстраполяция.....	50
Оценка модели.....	52
Линейная регрессия.....	55
Оптимизация потребления памяти.....	57
EDA и исследование зависимостей в данных.....	61
Заполнение пропусков в данных.....	69
<b>Часть 3. Модели линейной регрессии</b> .....	73
Линейная регрессия и L1_L2-регуляризация.....	73
Изотоническая регрессия.....	76
BIC и AIC.....	78
Полиномиальная регрессия.....	79
Линеаризация регрессии.....	81
<b>Часть практических навыков к 3</b> .....	84
Обогащение данных.....	84
Иерархия моделей.....	94
Оптимизация регрессии.....	101
Экспорт и импорт данных.....	105
Ансамбль регрессионных моделей.....	114
Расчет результатов.....	120

<b>Часть 4. Модели классификации и её метрики</b> .....	132
Точность и полнота .....	132
F-мера .....	134
ROC AUC и Gini .....	136
Оценка Каппа Коэна .....	139
Взвешенная квадратичная оценка Каппа Коэна .....	140
Логистическая функция потерь .....	142
Метод ближайших соседей .....	144
<b>Часть практических навыков к 4</b> .....	147
Страховой скоринг .....	147
F1 и Каппа оценки классификации .....	155
Метод ближайших соседей .....	161
Наивный Байес в задаче классификации скоринга и оптимизации потребления памяти .....	165
Логистическая регрессия .....	170
Иерархия логистической регрессии .....	174
Метод опорных векторов (Support-Vector Machine) .....	179
<b>Часть 5. Ансамблевые модели</b> .....	183
Ансамблевые модели .....	183
Бутстрэп .....	185
Бэггинг .....	186
Случайный лес .....	188
Out-of-Bag .....	190
Сверхслучайные деревья .....	192
Адаптивный бустинг .....	194
LogitBoost, BrownBoost и L2Boost .....	197
Градиентный спуск .....	200
Градиентный бустинг и XGBoost .....	203
Стохастический градиентный бустинг .....	205
<b>Часть практических навыков к 5</b> .....	208
Решающие деревья .....	208
Случайный лес .....	212
Бустинг с XGBoost .....	216
<b>Часть 6. Продвинутое ансамбли</b> .....	220
LightGBM .....	220
CatBoost .....	222
Ансамбль стекинга .....	224
<b>Часть практических навыков к 6</b> .....	228
LightGBM .....	228
CatBoost .....	232
Ансамбль классификации .....	238
Расчет результатов .....	243

<b>Часть 7. Искусственные нейронные сети</b> .....	247
Искусственные нейронные сети .....	247
Слой в нейросетях .....	250
Нейрон смещения .....	251
Функции активации .....	253
Обратное распространение ошибки .....	256
Многослойный перцептрон .....	258
<b>Часть практических навыков к 7</b> .....	261
Задача предсказания формы облаков .....	261
Предобработка изображений .....	266
Опорные векторы и коэффициент сходства .....	270
Двухслойный перцептрон .....	273
<b>Часть 8. Обучение нейросети</b> .....	278
Эпохи, пакеты, итерации .....	278
Оптимизация нейронной сети по Нестерову .....	279
Адаптивная оптимизация нейронной сети .....	281
RMSProp, Adadelta, Adam .....	282
Оптимизация нейронных сетей .....	283
Пакетная нормализация .....	285
Регуляризация обучения нейронных сетей .....	287
Методы инициализации весов в нейронных сетях .....	288
Дополнение данных .....	290
Свертка и подвыборка .....	292
Сверточные нейронные сети .....	294
<b>Часть практических навыков к 8</b> .....	296
Свертка и предвыборка .....	296
Активация и оптимизаторы .....	300
Нормализация и переобучение .....	305
Дополнение изображений .....	310
<b>Часть 9. Архитектуры сверточных нейросетей</b> .....	315
LeNet .....	315
AlexNet .....	317
VGG .....	320
GoogLeNet .....	323
Inception .....	325
ResNet .....	329
ResNetXt .....	331
SE-ResNet .....	333
EfficientNet .....	334
DenseNet .....	336
MobileNet .....	338
<b>Часть практических навыков к 9</b> .....	341
LeNet и AlexNet .....	341
VGG16 и VGG19 .....	345

GoogLeNet и Inception-BN .....	349
Inception V3 и V4.....	361
ResNet.....	365
Архитектура нейросети .....	369
MobileNet для различных предметных областей .....	375
<b>Библиографический список .....</b>	<b>380</b>
<b>Приложение 1. Варианты заданий для самостоятельной реализации алгоритмов машинного обучения.....</b>	<b>384</b>
<b>Приложение 2. Варианты заданий для исследовательских работ в области машинного обучения.....</b>	<b>387</b>
<b>Приложение 3. Варианты заданий, включающие в себя самостоятельный этап Data Mining, для построения End-To-End ррешний в области машинного обучения.....</b>	<b>388</b>