

УДК 004.85Н2О  
ББК 32.813  
К89

**Кук Д.**

К89 Машинное обучение с использованием библиотеки H2O / пер. с англ. А. Б. Огурцова. – М.: ДМК Пресс, 2018. – 250 с.: ил.

**ISBN 978-5-97060-508-0**

H2O – простая в использовании и открытая библиотека, которая поддерживает большое количество операционных систем и языков программирования, а также масштабируется для обработки больших данных. Эта книга научит вас использовать алгоритмы машинного обучения, реализованные в H2O, с упором на наиболее важные для продуктивной работы аспекты. Рассмотрены глубокое обучение, случайный лес, обучение на неразмеченных данных и ансамбли моделей.

В российское издание добавлены дополнительно два приложения, описывающих новейшие модули H2O – Deep Water и Stacked Ensemble. Их также можно найти в репозитории [https://github.com/statist-bhfh/h2o\\_book\\_translate](https://github.com/statist-bhfh/h2o_book_translate).

Издание предназначено для специалистов по анализу данных, желающих изучить и применять на практике относительно новый, но многообещающий инструмент – библиотеку H2O.

УДК 004.85Н2О  
ББК 32.813

Authorized Russian translation of the English edition of Practical Machine Learning with H2O, ISBN 9781491964606 © 2017 Darren Cook.

This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

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 978-1-491-96460-6 (анг.)  
ISBN 978-5-97060-508-0 (рус.)

Copyright © 2017 Darren Cook  
© Оформление, издание, перевод, ДМК Пресс, 2018

# Содержание

<b>Предисловие .....</b>	<b>10</b>
<b>Глава 1. Установка и начало работы .....</b>	<b>13</b>
Подготовка к установке.....	13
Установка R .....	13
Установка Python .....	14
Конфиденциальность .....	14
Установка Java .....	15
Установка H2O при помощи R (CRAN) .....	15
Установка H2O при помощи Python (pip).....	16
Наша первая задача машинного обучения.....	17
Обучение и предсказания в Python.....	21
Обучение и предсказания в R .....	23
Производительность и предсказания .....	25
Если вам не повезло .....	26
Веб-интерфейс Flow .....	27
Данные .....	28
Модели.....	29
Предсказания .....	31
Дополнительные сведения об интерфейсе Flow .....	32
Резюме.....	32
<b>Глава 2. Импортирование и экспортирование данных .....</b>	<b>33</b>
Требования к памяти .....	33
Подготовка данных.....	34
Загрузка данных в H2O.....	35
Загрузка файлов в формате CSV.....	35
Загрузка файлов в других форматах .....	37
Загрузка данных из R .....	38
Загрузка данных из Python .....	39
Операции с данными .....	40
«Ленивость», присвоение имен и удаление .....	40
Итоговые статистики .....	42
Операции со столбцами .....	42
Агрегирование строк.....	43
Индексация .....	44
Разделение данных в кластере H2O .....	46
Строки и столбцы .....	49
Выгрузка данных из H2O .....	52
Экспорт таблиц .....	52
Формат ROJO .....	53
Файлы моделей.....	54
Сохранение всех моделей .....	54
Резюме.....	55

<b>Глава 3. Наборы данных</b> .....	56
Набор данных об энергетической эффективности .....	56
Настройка и загрузка .....	57
Переменные .....	58
Разделение данных.....	59
Изучение данных.....	60
О наборе данных.....	64
Набор данных: рукописные цифры.....	64
Настройка и загрузка .....	65
Изучение данных.....	66
Как можно «помочь» модели .....	68
О наборе данных.....	70
Набор данных: результаты футбольных матчей .....	70
Корреляции .....	73
Пропущенные данные.....	76
Как обучать и тестировать? .....	77
Настройка и загрузка .....	77
Третий источник данных .....	78
Снова про пропущенные данные .....	80
Настройка и загрузка (снова).....	80
О наборе данных.....	83
Резюме.....	83
<b>Глава 4. Общие параметры моделей</b> .....	84
Поддерживаемые метрики .....	84
Метрики для регрессии.....	85
Метрики для классификации.....	85
Бинарная классификация .....	86
Основы .....	88
Объем выполняемой работы .....	89
Оценка и проверка .....	90
Ранняя остановка.....	90
Контрольные точки .....	92
Перекрестная проверка.....	94
Взвешивание наблюдений .....	95
Выборки и обобщающая способность.....	98
Регрессия .....	99
Контроль вывода результатов.....	100
Резюме .....	100
<b>Глава 5. Случайный лес</b> .....	101
Решающие деревья .....	101
Случайный лес .....	103
Параметры .....	103
Энергоэффективность зданий: случайный лес с настройками по умолчанию .....	105

Поиск по сетке .....	107
Полный перебор .....	108
Случайный поиск .....	110
Общая стратегия .....	112
Энергоэффективность зданий: настроенный случайный лес .....	113
MNIST: случайный лес с настройками по умолчанию .....	114
MNIST: настроенный случайный лес .....	116
Дополненные данные .....	119
Футбол: случайный лес с настройками по умолчанию .....	120
Футбол: настроенный случайный лес .....	122
Резюме .....	124
<b>Глава 6. Градиентный бустинг .....</b>	<b>125</b>
Бустинг .....	125
Хорошее, плохое... и непонятное .....	126
Параметры .....	127
Энергоэффективность зданий: градиентный бустинг с настройками по умолчанию .....	128
Энергоэффективность зданий: настроенный градиентный бустинг .....	130
MNIST: градиентный бустинг с настройками по умолчанию .....	133
MNIST: настроенный градиентный бустинг .....	134
Футбол: градиентный бустинг с настройками по умолчанию .....	137
Футбол: настроенный градиентный бустинг .....	138
Резюме .....	140
<b>Глава 7. Линейные модели .....</b>	<b>141</b>
Параметры GLM .....	141
Данные об энергоэффективности: GLM с настройками по умолчанию .....	145
Данные об энергоэффективности: настроенная GLM .....	147
MNIST: GLM с настройками по умолчанию .....	151
MNIST: настроенная GLM .....	153
Футбол: GLM с настройками по умолчанию .....	155
Футбол: настроенная GLM .....	156
Резюме .....	157
<b>Глава 8. Глубокое обучение (нейронные сети) .....</b>	<b>158</b>
Что такое нейронные сети? .....	159
Количественные и категориальные переменные .....	160
Слои нейронной сети .....	161
Функции активации .....	163
Параметры .....	164
Регуляризация .....	164
Оценка качества .....	165
Энергоэффективность зданий: модель глубокого обучения с настройками по умолчанию .....	168
Энергоэффективность зданий: настроенная модель глубокого обучения .....	168

MNIST: модель глубокого обучения с настройками по умолчанию.....	174
MNIST: настроенная модель глубокого обучения .....	175
Футбол: модель глубокого обучения с настройками по умолчанию .....	179
Футбол: настроенная модель глубокого обучения .....	180
Резюме .....	185
Приложение: дополнительные параметры .....	185
<b>Глава 9. Обучение на неразмеченных данных.....</b>	<b>187</b>
Кластеризация методом k-средних.....	188
Автокодировщики .....	191
Вложенные автокодировщики.....	193
Метод главных компонент.....	194
GLRM.....	196
Пропущенные данные.....	196
GLRM.....	200
Избавляемся от R .....	200
Резюме .....	203
<b>Глава 10. Все остальное.....</b>	<b>204</b>
Документация.....	204
Установка актуальной версии.....	204
Сборка из исходных кодов .....	204
Запуск из командной строки .....	205
Кластеры.....	205
EC2 .....	206
Другие облачные провайдеры .....	206
Hadoop .....	207
Spark / Sparkling Water .....	207
Наивный байесовский классификатор.....	207
Ансамбли.....	208
Стекинг: h2o.ensemble.....	208
Ансамбли для классификации.....	210
Резюме.....	210
<b>Глава 11. Эпилог .....</b>	<b>211</b>
Результаты для данных об энергоэффективности .....	211
Результаты для набора данных MNIST.....	213
Результаты для данных о футбольных матчах .....	214
Как далеко вы готовы зайти.....	216
Чем больше, тем лучше .....	217
Еще больше данных.....	218
Отбор сложных примеров.....	219
Автокодировщик .....	219
Сверточные сети .....	220
Ансамбли.....	221
Результаты.....	222
Резюме.....	223

<b>Приложение 1. Deep Water</b> .....	224
Установка.....	224
Сборка из исходных кодов .....	224
Amazon Machine Image .....	224
Образ Docker .....	224
Примеры данных .....	224
Обзор библиотеки Deep Water .....	224
Глубокое обучение в библиотеке H2O .....	225
Современные тенденции в глубоком обучении.....	225
Почему нужно использовать Deep Water .....	225
Начало работы: набор данных MNIST.....	226
Бекенды .....	227
CPU и GPU.....	227
Классификация изображений.....	229
Данные .....	229
Параметры изображений.....	229
Предварительно созданные архитектуры .....	229
Архитектуры, создаваемые пользователем.....	230
Предварительно обученные нейросети .....	230
Веб-интерфейс Flow .....	230
Поиск по сетке .....	233
Полный перебор .....	233
Случайный поиск.....	234
Контрольные точки .....	235
Ансамбли.....	237
Признаки скрытых слоев и меры сходства .....	238
Поддержка нескольких GPU.....	239
Развертывание моделей.....	240
MOJO.....	240
Prediction Service Builder .....	240
 <b>Приложение 2. Ансамбли (стекинг моделей).....</b>	 241
Вступление .....	241
Стекинг / Super Learner .....	241
Алгоритм .....	242
Вложенные ансамбли в библиотеке H2O.....	242
Пример .....	243
На языке R.....	243
На языке Python.....	245
Вопросы и ответы .....	247
Дополнительная информация .....	248
Список литературы.....	248
 <b>Краткий предметный указатель.....</b>	 249