

# Оглавление

<b>Предисловие от издательства</b> .....	13
<b>Предисловие</b> .....	14
<b>Введение</b> .....	17
Для кого предназначена эта книга.....	18
Почему мы используем TensorFlow и TensorFlow Extended.....	19
Обзор глав .....	19
Условные обозначения, используемые в этой книге .....	21
Использование примеров кода.....	22
Онлайн-обучение O'Reilly.....	22
Как с нами связаться .....	23
Благодарности .....	23
<b>Глава 1. Введение</b> .....	26
Почему и где используются конвейеры машинного обучения .....	26
Когда следует подумать о конвейере машинного обучения?.....	28
Обзор этапов конвейера машинного обучения .....	28
Этап загрузки данных и управление версиями данных.....	29
Проверка данных.....	29
Предварительная обработка данных .....	30
Обучение и настройка модели .....	31
Анализ модели.....	31
Управление версиями модели.....	32
Развертывание модели .....	32
Петли обратной связи .....	32
Приватность данных .....	33
Оркестровка конвейера.....	33
Для чего нужна оркестровка конвейера .....	33
Направленные ациклические графы .....	34
Наш демонстрационный проект машинного обучения .....	35
Структура проекта.....	36
Наша модель машинного обучения .....	36
Цель демонстрационного проекта .....	37
Резюме.....	37

<b>Глава 2. Введение в TensorFlow Extended</b> .....	38
Что такое TFX? .....	39
Установка TFX .....	40
Обзор компонентов TFX .....	41
Что такое метаданные ML Metadata? .....	42
Альтернативы TFX.....	45
Знакомство с Apache Beam.....	46
Установка .....	46
Базовый конвейер .....	47
Запуск элементарного конвейера .....	50
Резюме.....	50
<b>Глава 3. Загрузка данных</b> .....	51
Концепции загрузки данных .....	51
Загрузка локальных файлов данных .....	53
Загрузка удаленных файлов данных.....	57
Загрузка данных напрямую из баз данных .....	58
Подготовка данных .....	60
Разбиение наборов данных .....	60
Связующие наборы данных.....	62
Управление версиями наборов данных .....	63
Стратегии загрузки данных .....	64
Структурированные данные.....	64
Текстовые данные для задач обработки естественного языка .....	64
Графические данные для задач компьютерного зрения .....	64
Резюме.....	65
<b>Глава 4. Проверка данных</b> .....	66
Для чего нужна проверка данных?.....	67
TFDV.....	68
Установка .....	69
Генерация статистических показателей для набора данных .....	69
Генерация схемы на основе данных.....	71
Распознавание ошибок в данных.....	72
Сравнение наборов данных.....	73
Обновление схемы .....	75
Отклонения и дрейф данных .....	76
Наборы данных с систематической ошибкой выборки.....	77
Получение среза данных в TFDV .....	78
Обработка больших наборов данных с помощью Google Cloud Platform.....	80
Интеграция TFDV в конвейер машинного обучения .....	83
Резюме.....	84

<b>Глава 5. Предварительная обработка данных</b> .....	85
Для чего нужна предварительная обработка данных .....	86
Предварительная обработка данных в контексте всего набора данных .....	86
Масштабирование шагов предварительной обработки .....	86
Как избежать отклонения при обучении и работе модели .....	86
Развертывание шагов предварительной обработки и модели машинного обучения как единого артефакта .....	88
Проверка результатов предварительной обработки в конвейере .....	88
Предварительная обработка данных с помощью TFT .....	89
Установка .....	90
Стратегии предварительной обработки .....	90
Лучшие практики .....	92
Функции TFT .....	93
Автономная работа TFT .....	95
Интеграция TFT в конвейер машинного обучения .....	97
Резюме .....	101
<b>Глава 6. Обучение модели</b> .....	102
Определение модели для нашего демонстрационного проекта .....	103
Компонент TFX Trainer .....	106
Функция <code>run_fn()</code> .....	106
Запуск компонента <code>Trainer</code> .....	110
Другие соображения относительно компонента <code>Trainer</code> .....	112
Использование TensorBoard в интерактивном конвейере .....	113
Стратегии распределения .....	115
Настройка модели .....	118
Стратегии настройки гиперпараметров .....	118
Настройка гиперпараметров в конвейерах TFX .....	119
Резюме .....	119
<b>Глава 7. Анализ и проверка модели</b> .....	120
Как проанализировать модель .....	121
Метрики классификации .....	121
Метрики регрессии .....	124
Анализ модели TensorFlow .....	125
Анализ одной модели в TFMA .....	126
Анализ нескольких моделей в TFMA .....	129
Анализ достоверности модели .....	130
Формирование срезов для прогнозов модели в TFMA .....	132
Проверка пороговых значений решений с использованием метрик справедливости .....	134
Проведение более детального анализа с помощью инструмента анализа альтернатив (What-If Tool) .....	136

Объяснение модели.....	140
Генерация объяснений с помощью WIT .....	142
Другие методы объяснения .....	143
Анализ и проверка модели в TFX.....	145
ResolverNode .....	145
Компонент Evaluator .....	146
Проверка при помощи компонента Evaluator.....	147
Компонент TFX Pusher .....	148
Резюме.....	148

## Глава 8. Развертывание модели с помощью

<b>TensorFlow Serving</b> .....	149
Простой сервер моделей .....	150
Недостатки развертывания моделей с помощью API на основе Python .....	151
Отсутствие разделения кода.....	151
Отсутствие контроля версий модели.....	152
Неэффективный вывод модели.....	152
TensorFlow Serving .....	152
Обзор архитектуры TensorFlow .....	153
Экспорт моделей для TensorFlow Serving.....	153
Сигнатуры моделей.....	155
Методы сигнатуры.....	155
Проверка экспортированных моделей .....	157
Проверка модели.....	158
Тестирование модели.....	159
Установка TensorFlow Serving .....	160
Установка Docker .....	160
Установка на Ubuntu.....	160
Сборка TensorFlow Serving из исходного кода.....	161
Настройка сервера TensorFlow.....	161
Конфигурация при работе с одной моделью .....	162
Конфигурация при работе с несколькими моделями.....	164
REST или gRPC .....	166
REST.....	166
gRPC.....	166
Выполнение прогнозов на сервере моделей .....	167
Получение прогнозов модели с использованием REST.....	167
Работа с TensorFlow Serving через gRPC.....	169
А/В-тестирование модели с использованием TensorFlow Serving .....	172
Запрос метаданных модели с сервера моделей .....	173
REST-запросы метаданных модели.....	173
Запросы gRPC для метаданных модели .....	174

Пакетные запросы на вывод прогнозов модели .....	175
Настройка использования пакетного режима в прогнозировании .....	177
Другие функции оптимизации TensorFlow Serving .....	179
Альтернативы TensorFlow Serving .....	180
BentoML .....	180
Seldon .....	180
GraphPipe .....	181
Simple TensorFlow Serving .....	181
MLflow .....	181
Ray Serve .....	181
Развертывание моделей с использованием услуг поставщиков облачных решений .....	182
Сценарии использования .....	182
Пример развертывания с помощью облачных платформ Google .....	182
Развертывание модели с помощью конвейеров TFX .....	188
Резюме .....	189

## **Глава 9. Расширенные концепции развертывания моделей с помощью TensorFlow Serving** .....

Разделение зон ответственности в процессе развертывания .....	190
Обзор рабочего процесса .....	191
Оптимизация загрузки удаленной модели .....	193
Оптимизация модели для развертываний .....	194
Квантование .....	194
Сокращение .....	195
Дистилляция .....	196
Использование TensorRT совместно с TensorFlow Serving .....	196
TFLite .....	197
Шаги по оптимизации моделей машинного обучения с помощью TFLite .....	197
Развертывание моделей TFLite с помощью TensorFlow Serving .....	199
Мониторинг экземпляров TensorFlow Serving .....	200
Установка Prometheus .....	200
Конфигурация TensorFlow Serving .....	202
Простое масштабирование с помощью TensorFlow Serving и Kubernetes .....	204
Дополнительная литература о Kubernetes и Kubeflow .....	205
Резюме .....	206

## **Глава 10. Расширенные концепции TensorFlow Extended** .....

Расширенные концепции конвейеров машинного обучения .....	207
Одновременное обучение нескольких моделей .....	208

Экспорт моделей TFLite .....	209
Ограничения TFLite.....	210
Обучение модели с «теплым» запуском .....	212
Участие человека в конвейере машинного обучения.....	212
Настройка компонента Slack .....	214
Как использовать компонент Slack .....	214
Пользовательские компоненты TFX .....	215
Сценарии использования пользовательских компонентов .....	216
Создание пользовательского компонента с нуля.....	216
Повторное использование существующих компонентов .....	225
Резюме.....	228
<b>Глава 11. Конвейеры, часть 1: Apache Beam и Apache Airflow....</b>	<b>230</b>
Какой инструмент оркестрации выбрать?.....	231
Apache Beam.....	231
Apache Airflow .....	231
Kubeflow Pipelines .....	231
Kubeflow Pipelines на платформе AI.....	232
Преобразование вашего интерактивного конвейера TFX в производственный конвейер.....	232
Преобразование элементарного интерактивного конвейера для Beam и Airflow.....	234
Введение в Apache Beam .....	235
Оркестрация конвейеров TFX с помощью Apache Beam .....	235
Введение в Apache Airflow.....	237
Установка и начальная настройка.....	237
Элементарный пример использования Airflow.....	239
Оркестрация конвейеров TFX с помощью Apache Airflow.....	242
Настройка конвейера .....	242
Запуск конвейера.....	244
Резюме.....	245
<b>Глава 12. Конвейеры, часть 2:</b>	
<b>Kubeflow Pipelines .....</b>	<b>246</b>
Введение в Kubeflow Pipelines.....	247
Установка и начальная настройка.....	249
Доступ к установленному экземпляру Kubeflow Pipelines.....	251
Оркестрация конвейеров TFX с помощью Kubeflow Pipelines.....	252
Настройка конвейера .....	254
Запуск конвейера.....	258
Полезные функции Kubeflow Pipelines.....	264

Конвейеры, работающие на Google Cloud AI Platform .....	269
Настройка конвейера .....	269
Настройка конвейера TFX.....	273
Запуск и работа конвейера .....	276
Резюме.....	277
<b>Глава 13. Петли обратной связи .....</b>	<b>279</b>
Явная и неявная обратная связь.....	280
Маховик данных .....	281
Петли обратной связи в реальном мире .....	282
Конструктивные шаблоны для сбора отзывов .....	284
Пользователи предпринимают определенные действия в результате прогноза .....	284
Пользователи оценивают качество прогноза.....	285
Пользователи исправляют прогноз.....	285
Краудсорсинг аннотаций .....	286
Экспертные аннотации .....	287
Обратная связь автоматически предоставляется системой .....	287
Как отслеживать петли обратной связи.....	287
Отслеживание явной обратной связи .....	288
Отслеживание неявной обратной связи .....	289
Резюме.....	289
<b>Глава 14. Приватность данных, используемых для машинного обучения .....</b>	<b>290</b>
Введение в приватность данных .....	290
Почему мы заботимся о приватности данных? .....	291
Самый простой способ повысить приватность данных .....	291
Какие данные должны быть приватными? .....	292
Дифференцированная приватность.....	293
Локальная и глобальная дифференцированная приватность.....	294
Эпсилон-дельта и бюджет приватности .....	295
Дифференцированная приватность в машинном обучении .....	296
Введение в TensorFlow Privacy .....	296
Обучение с оптимизатором, использующим подход дифференцированной приватности .....	297
Расчет параметра $\epsilon$ .....	298
Введение в федеративное обучение.....	299
Федеративное обучение в TensorFlow.....	301
Зашифрованное машинное обучение.....	302
Зашифрованное обучение модели .....	303
Преобразование обученной модели для обслуживания зашифрованных прогнозов .....	304

---

Другие методы обеспечения приватности данных.....	305
Резюме.....	305
<b>Глава 15. Будущее конвейеров машинного обучения и следующие шаги.....</b>	<b>307</b>
Отслеживание экспериментов с моделью .....	307
Предложения в области управления релизами модели.....	308
Будущие возможности конвейеров.....	309
Использование TFX с другими фреймворками машинного обучения.....	310
Тестирование моделей машинного обучения .....	310
Системы непрерывной интеграции и развертывания для машинного обучения.....	311
Сообщество инженеров машинного обучения.....	311
Резюме.....	311
<b>Приложение А. Введение в инфраструктуру машинного обучения.....</b>	<b>313</b>
<b>Приложение В. Настройка кластера Kubernetes в Google Cloud .....</b>	<b>326</b>
<b>Приложение С. Советы по работе с Kubeflow Pipelines .....</b>	<b>332</b>
<b>Предметный указатель .....</b>	<b>340</b>

# Предисловие от издательства

## Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте [www.dmkpress.com](http://www.dmkpress.com), зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com); при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу [http://dmkpress.com/authors/publish\\_book/](http://dmkpress.com/authors/publish_book/) или напишите в издательство по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com).

## Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в основном тексте или программном коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com), и мы исправим это в следующих тиражах.

## Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу [dmkpress@gmail.com](mailto:dmkpress@gmail.com).

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

# Предисловие

Когда в 1913 году компания Генри Форда построила свой первый сборочный конвейер для производства своей легендарной Model T, время, необходимое для сборки каждой машины, сократилось с 12 до 3 часов. Затраты на производство резко снизились, что позволило Model T стать первым доступным автомобилем в истории. Это также сделало возможным массовое производство: вскоре Model T стала королевой автомобильных дорог.

Поскольку производственный процесс теперь представлял собой четкую последовательность четко определенных шагов (он же конвейер), стало возможно автоматизировать некоторые из этих шагов, сэкономив еще больше времени и денег. Сегодня производство автомобилей невозможно без автоматизации.

Но дело не только во времени и деньгах. При выполнении многих повторяющихся задач машина будет производить гораздо более стабильный результат, чем люди, в результате чего конечный продукт станет более предсказуемым, последовательным и надежным. Наконец, избавляя людей от тяжелого физического труда, автоматика позволяет значительно повысить безопасность. Многие рабочие от монотонной физической работы перешли к работе, требующей более высокого уровня квалификации (хотя, честно говоря, многие просто потеряли работу).

Если посмотреть на автоматизацию с другой стороны, установка автоматизированной сборочной линии может занять много времени и стать дорогостоящим проектом. Кроме того, сборочный конвейер – не идеальное решение, если вы хотите производить небольшие партии или продукцию по индивидуальному заказу. Форд сказал: «Цвет автомобиля может быть любым, при условии что он будет черным».

История автомобилестроения повторилась в индустрии программного обеспечения за последние пару десятилетий: каждая значительная часть программного обеспечения в настоящее время создается, тестируется и развертывается с использованием таких инструментов автоматизации, как Jenkins или Travis. Однако метафоры Model T уже недостаточно. Программное обеспечение не просто развертывается и работает как есть; его необходимо регулярно контролировать, поддерживать и обновлять. Программные конвейеры теперь больше похожи на динамические циклы, чем на статические производственные линии. Крайне важно иметь возможность быстро обновлять программное обеспечение (или сам конвейер), не нарушая его целостности. А программное обеспечение гораздо более вариативно, чем когда-либо была Model T: программное обеспечение можно раскрасить в любой цвет (попробуйте, например, подсчитать количество существующих вариантов для MS Office).

К сожалению, «классические» инструменты автоматизации не подходят для создания полноценного конвейера машинного обучения. Действительно, модель машинного обучения не является обычным программным обеспечением.

Во-первых, большая часть его поведения определяется данными, на которых он обучается. Следовательно, сами обучающие данные должны рассматриваться как код (и, соответственно, иметь версии). Это довольно сложная проблема, потому что новые данные появляются каждый день (часто в больших количествах), изменяются и дрейфуют с течением времени, часто включают персональные данные; также новые данные должны быть размечены, прежде чем вы сможете передать их в работу, которую выполняют алгоритмы машинного обучения.

Во-вторых, поведение модели нередко бывает довольно непрозрачным: она может пройти все тесты для одних данных, но полностью потерпеть неудачу для других. Таким образом, вы должны убедиться, что ваши тесты охватывают все области данных, на которых ваша модель будет использоваться в производстве. В частности, вы должны убедиться, что она не проявляет дискриминацию в отношении какой-либо группы ваших пользователей.

По этим (и другим) причинам специалисты по обработке данных и инженеры-программисты сначала начали создавать и обучать модели машинного обучения вручную, так сказать, «в своем гараже», и многие из них до сих пор это делают. Но за последние несколько лет были разработаны новые инструменты автоматизации, которые решают задачи конвейеров машинного обучения, такие как TensorFlow Extended (TFX) и Kubeflow. Все больше и больше организаций начинают использовать эти инструменты для создания конвейеров машинного обучения, которые автоматизируют большую часть (или все) этапов построения и обучения моделей машинного обучения. Преимущества этой автоматизации в основном те же, что и для автомобильной промышленности: экономия времени и денег; возможность создавать более качественные, надежные и безопасные модели и тратить больше времени на выполнение более полезных задач, чем на копирование данных или изучение кривых обучения. Однако построить конвейер машинного обучения непросто. Так с чего же начать?

Начать с этой книги!

В этой книге Ханнес и Кэтрин дают четкое и понятное руководство по автоматизации конвейеров машинного обучения. Как твердому стороннику практического подхода, особенно для такой технической темы, мне особенно понравилось то, как эта книга шаг за шагом проведет вас через конкретный пример проекта от начала до конца. Благодаря множеству примеров кода и ясным, лаконичным объяснениям вы сможете создать свой собственный конвейер машинного обучения и запустить его в кратчайшие сроки, а также все концептуальные инструменты, необходимые для адаптации этих конвейеров машинного обучения к вашим собственным вариантам использования. Я настоятельно рекомендую вам взять свой ноутбук и попробовать что-то во время чтения: так вы научитесь намного быстрее.

Я впервые встретился с Ханнесом и Кэтрин в октябре 2019 года на конференции TensorFlow World в Санта-Кларе, Калифорния, где я делал доклад о создании конвейеров машинного обучения с использованием TFX. Они работали над этой книгой по той же теме, и у нас был один редактор, так что, естественно, нам было о чем поговорить. Некоторые слушатели задавали технические вопросы о TensorFlow Serving (который является частью TFX), и у Ханнеса и Кэтрин были все ответы, которые я искал. Ханнес даже любезно принял мое при-

глашение выступить с докладом о расширенных функциях TensorFlow Serving в конце моего курса в очень короткие сроки. Его выступление было сокровищницей идей и полезных советов, которые вы найдете в этой книге, а также во многих, многих других.

Пришло время приступить к созданию профессиональных конвейеров машинного обучения!

– *Орелиен Жерон*,  
бывший руководитель группы классификации видео YouTube,  
автор книги «Практическое машинное обучение  
с использованием Scikit-Learn, Keras и TensorFlow» (O'Reilly)  
*Окленд, Новая Зеландия, 18 июня 2020 г.*

# Введение

Все говорят о машинном обучении. Из академической дисциплины оно превратилось в одну из самых удивительных технологий. Машинное обучение используется везде – от обработки видеопотока, регистрируемого в автомобилях с автоматическим управлением, до персонализации назначений лекарств. Оно становится важным элементом в каждой отрасли. В то время как моделям архитектуры и концепциям уделялось большое внимание, машинное обучение еще не прошло стадию стандартизации процессов, появившихся в отрасли программного обеспечения в последнее десятилетие. В этой книге мы хотели бы показать вам, как создать стандартизированную систему машинного обучения, которая была бы автоматизированной и воспроизводимой.

За последние несколько лет разработки в области машинного обучения достигли впечатляющих результатов. Благодаря широкой доступности графических процессоров (Graphical Processing Units, GPU) и разработке новых концепций глубокого обучения, таких как Transformers (например, BERT) или Generative Adversarial Networks (например, DCGAN), количество проектов в области искусственного интеллекта, или ИИ (Artificial Intelligence, AI), резко возросло. Количество стартапов в сфере ИИ огромно. Корпорации применяют новейшие технологии машинного обучения для решения всех видов бизнес-задач. В этом стремлении к наиболее эффективному решению для задач машинного обучения мы наблюдали несколько вещей, которым уделялось меньше внимания. Мы увидели, что специалистам в области ИИ – ученым, специализирующимся в области машинного обучения и искусственного интеллекта, и инженерам по машинному обучению – не хватает хороших источников информации для создания концепций и инструментов для ускорения, повторного использования, управления и развертывания своих разработок. Необходима стандартизация конвейеров машинного обучения.

Конвейеры машинного обучения – это процессы для ускорения, повторного использования, управления и развертывания моделей машинного обучения. Примерно десять лет назад разработка программного обеспечения претерпела такие же изменения благодаря внедрению непрерывной интеграции (Continuous Integration, CI) и непрерывного развертывания (Continuous Deployment, CD). Когда-то это был длительный процесс тестирования и развертывания веб-приложения. В наши дни эти процессы были значительно упрощены с помощью нескольких инструментов и концепций. Ранее для развертывания веб-приложений требовалось сотрудничество между инженером DevOps и разработчиком программного обеспечения. Сегодня приложение можно надежно протестировать и развернуть за считанные минуты. Специалисты по обработке данных и инженеры машинного обучения могут заимствовать концепции рабочих процессов из программной инженерии.

Исходя из нашего личного опыта, большинство проектов в области науки о данных, нацеленных на внедрение моделей в производство, не могут позволить себе роскошь создавать большие команды, что затрудняет построение всего конвейера собственными силами с нуля. Это может означать, что проекты машинного обучения превращаются в одиночные попытки построения моделей, и их результативность ухудшается со временем; специалист тратит большую часть своего времени на исправление ошибок, когда меняются данные, лежащие в основе решения, или модель не находит широкого применения. Автоматизированный конвейер, позволяющий построить воспроизводимый сквозной рабочий процесс, уменьшает усилия, необходимые для развертывания модели. Конвейер машинного обучения должен включать процессы, которые:

- эффективно отслеживают изменение версий исходных данных и запускают новое обучение моделей;
- эффективно выполняют предварительную обработку данных для обучения и проверки модели;
- следят за версиями контрольных точек модели во время обучения;
- отслеживают ваши эксперименты по обучению моделей;
- анализируют и проверяют обученные и настроенные модели;
- выполняют развертывание проверенных моделей;
- масштабируют развернутую модель;
- собирают новые данные для обучения и регистрируют показатели точности модели, используя петли обратной связи.

В этом списке пропущен один важный момент: обучение и настройка модели. Мы предполагаем, что вы уже обладаете достаточными знаниями и опытом для выполнения этого шага. Если же вы только начинаете погружаться в тему машинного или глубокого обучения, следующие книги, опубликованные O'Reilly, станут отличной отправной точкой для знакомства с машинным обучением:

- *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms 1st Edition* by Nikhil Buduma, Nicholas Locascio (*Нихил Будума, Николас Локасио*. Основы глубокого обучения: разработка алгоритмов искусственного интеллекта следующего поколения. 1-е изд.);
- *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition* by Aurélien Géron (*Аурелиен Жерон*. Практическое машинное обучение с использованием Scikit-Learn, Keras и TensorFlow: концепции, инструменты и методы для построения интеллектуальных систем. 2-е изд.).

## Для кого предназначена эта книга

Основная аудитория этой книги – ученые, специализирующиеся в области машинного обучения и искусственного интеллекта и инженеры по машинному обучению, которые хотят выйти за рамки обучения единичной модели машинного обучения и успешно реализовать свои проекты в области науки о данных. Вы должны быть знакомы с основными концепциями машинного обучения и хотя бы с одним из фреймворков, используемых в машинном

обучении (например, PyTorch, TensorFlow, Keras). Примеры в этой книге основаны на TensorFlow и Keras, но основные концепции могут быть применены к любой среде.

Также эта книга может быть полезна менеджерам проектов в области науки о данных, разработчикам программного обеспечения или инженерам DevOps, которые хотят, чтобы их организация ускорила свои проекты, использующие технологии машинного обучения и искусственного интеллекта. Если вы заинтересованы в лучшем понимании жизненного цикла разработки систем на основе автоматизированного машинного обучения и хотите понять, как это может принести пользу вашей организации, то в следующих главах будет представлен набор инструментов для достижения данной цели.

## ПОЧЕМУ МЫ ИСПОЛЬЗУЕМ TENSORFLOW И TENSORFLOW EXTENDED

В этой книге во всех наших примерах конвейера будут использоваться инструменты из экосистемы TensorFlow, в частности TensorFlow Extended (TFX). Мы выбрали этот фреймворк по ряду причин:

- экосистема TensorFlow является наиболее доступной для машинного обучения на момент написания статьи. Она включает в себя несколько полезных проектов и библиотек поддержки, таких как TensorFlow Privacy и TensorFlow Probability;
- она популярна и широко используется как в малых, так и на крупных производственных предприятиях, и существует активное сообщество заинтересованных пользователей;
- поддерживаемые варианты использования простираются от академических исследований до машинного обучения в производственной среде. TFX тесно интегрирован с базовой платформой TensorFlow для использования в производственных процессах;
- и TensorFlow, и TFX являются инструментами с открытым исходным кодом, и нет никаких ограничений на их использование.

Однако все принципы, которые мы описываем в этой книге, применимы и к другим инструментам и фреймворкам.

## ОБЗОР ГЛАВ

В следующих главах мы представим конкретные шаги для построения конвейеров машинного обучения и продемонстрируем, как они работают, на примере демонстрационного проекта.

*Глава 1 «Введение».* В этой главе представлен обзор конвейеров машинного обучения, обсуждается, когда их следует использовать, и описываются все этапы, входящие в состав конвейера. Также мы представляем здесь пример проекта, который будем использовать на протяжении всей книги.

*Глава 2 «Введение в TensorFlow Extended»* знакомит читателей с экосистемой TFX, объясняет, как задачи взаимодействуют друг с другом, и описывает внутреннюю работу компонентов TFX. Мы также рассмотрим ML MetadataStore,

покажем, как он используется в контексте TFX, и как Apache Beam запускает компоненты TFX без вмешательства пользователя.

*Глава 3 «Загрузка данных»* посвящена процессу систематической загрузки данных в наши конвейеры, а также в ней обсуждается концепция управления версиями данных.

*Глава 4 «Проверка данных»* объясняет, каким образом организовать эффективную проверку данных, поступающих в ваш конвейер, с помощью инструмента TensorFlow Data Validation. Система предупредит вас, если новые данные существенно изменятся по сравнению с предыдущими данными, что может повлиять на точность вашей модели.

*Глава 5 «Предварительная обработка данных»* посвящена подготовке данных (конструированию признаков) с использованием TensorFlow Transform для преобразования необработанных данных в признаки, подходящие для обучения модели.

*Глава 6 «Обучение модели»* объясняет, как обучать модели в рамках задачи машинного обучения. В этой главе мы также объясним концепцию настройки модели.

*Глава 7 «Анализ и проверка модели»* содержит полезные метрики для понимания, как работает ваша модель в производственной среде. В число этих метрик входят и те, которые могут позволить вам выявить ошибки в предсказаниях модели. В разделе «Анализ и проверка модели в TFX» объясняется, как управлять версиями вашей модели при улучшении одного из показателей в новой версии. Модель в конвейере может быть автоматически обновлена до новой версии.

*Глава 8 «Развертывание моделей с помощью TensorFlow Serving»* посвящена тому, как эффективно развернуть модель машинного обучения. Начиная с простой реализации Flask мы подчеркиваем ограничения в использовании таких пользовательских моделей в приложениях. Мы расскажем о TensorFlow Serving и о том, как настроить ваши исполнительные экземпляры. Мы также обсуждаем его пакетный режим работы и демонстрируем настройку клиентов для запроса результатов прогнозирования на основе модели.

*Глава 9 «Расширенные варианты развертывания моделей с помощью TensorFlow Serving»* показывает, как оптимизировать ваши варианты развертывания модели и как их контролировать. Мы обсуждаем стратегии оптимизации ваших моделей TensorFlow для повышения производительности. Мы также продемонстрируем базовый сценарий настройки развертывания с использованием Kubernetes.

*Глава 10 «Расширенные концепции TensorFlow Extended»* представляет концепцию специальных компонентов для ваших конвейеров машинного обучения, так что вы не ограничены стандартными компонентами в TFX. Если вы хотите добавить дополнительные этапы загрузки данных или преобразовать экспортированные модели в TensorFlow Lite (TFLite), мы проведем вас через все этапы создания таких компонентов.

*Глава 11 «Конвейеры, часть 1: Apache Beam и Apache Airflow»* собирает воедино все, что обсуждалось в предыдущих главах. Мы обсудим, как превратить ваши компоненты в конвейеры и как их настроить для выбранной платформы оркестровки. Мы также проведем вас от начала до конца, через все шаги конвейера, работающего на Apache Beam и Apache Airflow.

*Глава 12 «Конвейеры, часть 2: Конвейеры Kubeflow Pipelines»* является продолжением предыдущей главы. В ней рассматривается построение конвейера с использованием Kubeflow и платформы искусственного интеллекта Google.

*Глава 13 «Петли обратной связи»* посвящена тому, как превратить ваш модельный конвейер в цикл, позволяющий вносить улучшения на основе отзывов пользователей конечного продукта. Мы обсудим, какие типы данных необходимо собирать, чтобы улучшить модель в будущих версиях, и как передать эти данные обратно в конвейер.

*Глава 14: «Приватность данных, используемых для машинного обучения»* представляет концепцию быстро меняющейся области машинного обучения с сохранением приватности данных. В ней обсуждается три важных элемента этой концепции: дифференцированная приватность, федеративное обучение и зашифрованное машинное обучение.

*Глава 15 «Будущее конвейеров машинного обучения и следующие шаги»* содержит обзор технологий, которые повлияют на будущие конвейеры машинного обучения и на то, как мы будем смотреть на машинное обучение в ближайшие годы.

*Приложение А «Введение в инфраструктуру машинного обучения»* содержит краткое введение в Docker и Kubernetes.

*Приложение В «Настройка кластера Kubernetes в Google Cloud»* содержит дополнительные материалы по настройке Kubernetes в Google Cloud.

*Приложение С «Советы по работе с конвейерами Kubeflow»* содержит несколько полезных советов по работе с настройками конвейеров Kubeflow, включая обзор интерфейса командной строки TFX.

## УСЛОВНЫЕ ОБОЗНАЧЕНИЯ, ИСПОЛЬЗУЕМЫЕ В ЭТОЙ КНИГЕ

В этой книге используются следующие условные обозначения.

### *Курсив*

Обозначает новые термины, URL-адреса, адреса электронной почты, имена файлов и расширения файлов.

### Моноширинный шрифт

Используется для оформления листингов программ, а также в тексте для обозначения фрагментов программы, таких как имена переменных или функций, базы данных, типы данных, переменные среды, операторы и ключевые слова.

### **Жирный моноширинный шрифт**

Показывает команды или другой текст, который пользователь должен набирать точь-в-точь так, как показано в книге.

### *Моноширинный шрифт, выделенный курсивом*

Показывает текст, который следует заменить пользовательскими значениями или значениями, определяемыми контекстом.



Данный элемент означает совет или предложение.



Данный элемент обозначает общее примечание.



Так обозначаются предупреждения и предостережения.

## ИСПОЛЬЗОВАНИЕ ПРИМЕРОВ КОДА

Дополнительные материалы (примеры кода и т. п.) доступны для загрузки по ссылке <https://oreil.ly/bmlp-git>.

Если у вас есть технический вопрос или проблема с использованием примеров кода, отправьте электронное письмо на адрес [bookquestions@oreilly.com](mailto:bookquestions@oreilly.com) и [buildingmlpipelines@gmail.com](mailto:buildingmlpipelines@gmail.com).

Эта книга предназначена для того, чтобы помочь вам в работе над проектами. Вы можете использовать приведенные в этой книге примеры кода в своих программах и документации. Вам не нужно связываться с нами для получения разрешения на использование этих примеров, если вы используете незначительную часть кода. Например, для написания программы, использующей несколько фрагментов кода из этой книги, не требуется разрешения. Однако для коммерческого использования или распространения примеров из книг O'Reilly потребуется разрешение. Для ответа на вопрос с использованием цитат из этой книги и примеров кода разрешение не требуется. Но для включения значительного количества примеров кода из этой книги в документацию к вашему программному продукту потребуется разрешение.

Мы ценим, но не требуем указания авторства. Обычно при цитировании указывается название, автор, издатель и ISBN, например: «Создание конвейеров машинного обучения», Ханнес Хапке и Кэтрин Нельсон (O'Reilly). © 2020 Ханнес Хапке и Кэтрин Нельсон, 978-1-492-05319-4».

Если вы считаете, что ваше использование примеров кода выходит за рамки условий добросовестного использования или разрешений, приведенных выше, обращайтесь к нам по адресу электронной почты [permissions@oreilly.com](mailto:permissions@oreilly.com).

## ОНЛАЙН-ОБУЧЕНИЕ O'REILLY

На протяжении более 40 лет O'Reilly Media разрабатывает технологические и бизнес-тренинги, продвигая знания и развивая идеи, чтобы помочь компаниям добиться успеха.

Наша уникальная сеть экспертов и новаторов делится своими знаниями и опытом через книги, статьи и нашу платформу онлайн-обучения. Платформа онлайн-обучения O'Reilly предоставляет по запросу доступ к курсам обучения в режиме реального времени, схемам углубленного обучения, интерактивным средам программирования и обширной коллекции текстов и видео от O'Reilly и более 200 других издателей. Для получения дополнительной информации перейдите по ссылке <http://oreilly.com>.

## КАК С НАМИ СВЯЗАТЬСЯ

Оба автора этой книги хотели бы поблагодарить вас за уделенное ей внимание. Если вы хотите связаться с ними, вы можете сделать это через их веб-сайт [www.buildingmlpipelines.com](http://www.buildingmlpipelines.com) или по электронной почте [buildingmlpipelines@gmail.com](mailto:buildingmlpipelines@gmail.com). Авторы желают вам всяческих успехов в создании собственных конвейеров машинного обучения.

Комментарии и вопросы, касающиеся этой книги, просьба направлять издателю:

O'Reilly Media, Inc.

1005 Gravenstein Highway Северный Севастополь, CA 95472

800-998-9938 (в США или Канаде) 707-829-0515 (международный или местный)

707-829-0104 (факс)

Для этой книги разработана веб-страница, где мы публикуем исправления, примеры, а также другую дополнительную информацию. Для того чтобы ознакомиться с материалами, размещенными на этой странице, перейдите по ссылке: <https://oreil.ly/build-ml-pipelines>.

Для отправки комментариев или технических вопросов по содержанию данной книги используйте адрес электронной почты [bookquestions@oreilly.com](mailto:bookquestions@oreilly.com).

Новости и информацию о наших книгах и курсах можно найти на сайте <http://oreilly.com>.

Наша страница в Facebook: <http://facebook.com/oreilly>.

Мы в Twitter: <http://twitter.com/oreillymedia>.

Наш канал YouTube: <http://www.youtube.com/oreillymedia>.

## БЛАГОДАРНОСТИ

В процессе написания этой книги нас поддерживали многие замечательные люди. Большое спасибо всем, кто помог воплотить в жизнь наши замыслы! Мы хотели бы выразить особую благодарность всем вам.

Со всей командой O'Reilly было здорово работать все время, пока готовилась эта книга. Спасибо нашим редакторам Мелиссе Поттер (Melissa Potter), Николь Таше (Nicole Taché) и Амелии Блевинс (Amelia Blevins) за постоянную поддержку и вдумчивые отзывы. Спасибо также Кэти Тозер (Katie Tozer) и Джонатану Хасселлу (Jonathan Hassell) за оказанную нам поддержку.

Спасибо Орелиену Жерону (Aurélien Géron), Роберту Кроу (Robert Crowe), Маргарет Мейнард-Рид (Margaret Maynard-Reid), Сергею Хоменко (Sergii Khomenko) и Викраму Тивари (Vikram Tiwari), которые внимательно просмотрели

рели всю книгу и дали множество полезных советов и ценных комментариев. Вы много сделали для того, чтобы ее окончательный вариант стал лучше. Спасибо за часы, потраченные вами на внимательное изучение материалов книги.

Спасибо Янну Дюпису (Yann Dupis), Джейсону Манкузо (Jason Mancuso) и Мортену Далю (Morten Dahl) за ваш тщательный и всесторонний обзор главы о конфиденциальности машинного обучения.

Мы получили фантастическую поддержку от многих замечательных людей в Google. Благодарим вас за помощь в поиске и исправлении ошибок, а также за то, что вы выпускаете инструменты машинного обучения в виде пакетов с открытым исходным кодом! Помимо сотрудников Google, мы хотим особо поблагодарить Эми Унру (Amy Unruh), Анушу Рамеш (Anusha Ramesh), Кристину Грир (Christina Greer), Клеменс Мевальд (Clemens Mewald), Дэвида Заца (David Zats), Эдда Уайлдера-Джеймса (Edd Wilder-James), Ирен Джаннумис (Irene Giannoumis), Ярека Вилькевича (Jarek Wilkiewicz), Джайи Чжао (Jiayi Zhao), Джери Симсу (Jiri Simsa), Константиноса Катсиаписана (Konstantinos Katsiapis), Лак Лакшана (Lak Lakshmanan), Майка Древеса (Mike Dreves), Пейджа Бейли (Paige Bailey), Педрама Пеймана (Pedram Pejman), Сару Робинсон (Sara Robinson), Сунсон Квон (Soonson Kwon), Тею Ламкин (Thea Lamkin), Триса Варкентина (Tris Warkentin), Варшу Наганатан (Varshaa Naganathan), Чжитао Ли (Zhitaο Li) и Зохара Яхава (Zohar Yahav).

Мы выражаем огромную благодарность сообществу TensorFlow и Google Developer Expert и его замечательным участникам. Спасибо за поддержку в этом начинании.

Спасибо другим участникам, которые помогли на разных этапах: Барбаре Фусинска (Barbara Fusinska), Хамелю Хусайну (Hamel Husain), Михалу Яцшембскому (Michał Jastrzębski) и Яну Хенселю (Ian Hensel).

Спасибо сотрудникам Concur Labs (прошлым и настоящим) и другим сотрудникам SAP Concur за плодотворные обсуждения и полезные идеи. В частности, спасибо Джону Дитцу (John Dietz) и Ричарду Пакетту (Richard Puckett) за неоценимую поддержку, оказанную авторам книги.

### *Ханнес*

Я хочу поблагодарить мою замечательную партнершу Уитни за ее огромную поддержку на протяжении всего времени работы над этой книгой. Спасибо за постоянную поддержку, за отклики и за терпение, когда я провожу долгие часы за написанием статей. Спасибо моей семье, особенно моим родителям, которые позволили мне следовать за моей мечтой по всему миру.

Эта книга не появилась бы, не будь у меня замечательных друзей. Спасибо, Коул Ховард (Cole Howard), за то, что вы прекрасный друг и учитель. Наше сотрудничество подтолкнуло меня к этой работе и размышлениям о конвейерах машинного обучения. Моим друзьям Тимо Метцгеру (Timo Metzger) и Аманде Райт (Amanda Wright): спасибо за то, что вы улучшили язык этой книги. И спасибо Еве и Килиану Рамбах (Eva Rambach, Kilian Rambach), а также Деб и Дэвиду Хаклеманам (Deb Hackleman, David Hackleman). Без вашей помощи я бы не добрался до Орегона.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

[e-Univers.ru](http://e-Univers.ru)