Титульный лист материалов по дисциплине

(заполняется по каждому виду учебного материала)

ДИСЦИПЛИНА	**Технологии извлечения знаний из больших данных** (полное наименование дисциплины без сокращений)
ИНСТИТУТ	ИКБ
КАФЕДРА	**КБ-4 «Интеллектуальные системы информационной безопасности»** полное наименование кафедры)
ВИД УЧЕБНОГО МАТЕРИАЛА	**Лекция** (в соответствии с пп.1-11)
ПРЕПОДАВАТЕЛЬ	**Никонов В.В.** (фамилия, имя, отчество)
СЕМЕСТР	**3 семестр 2023/2024 уч. года** (указать семестр обучения, учебный год)

Извлечение знаний и анализ данных

Задачи, решаемые в анализе данных, их сходства и отличия

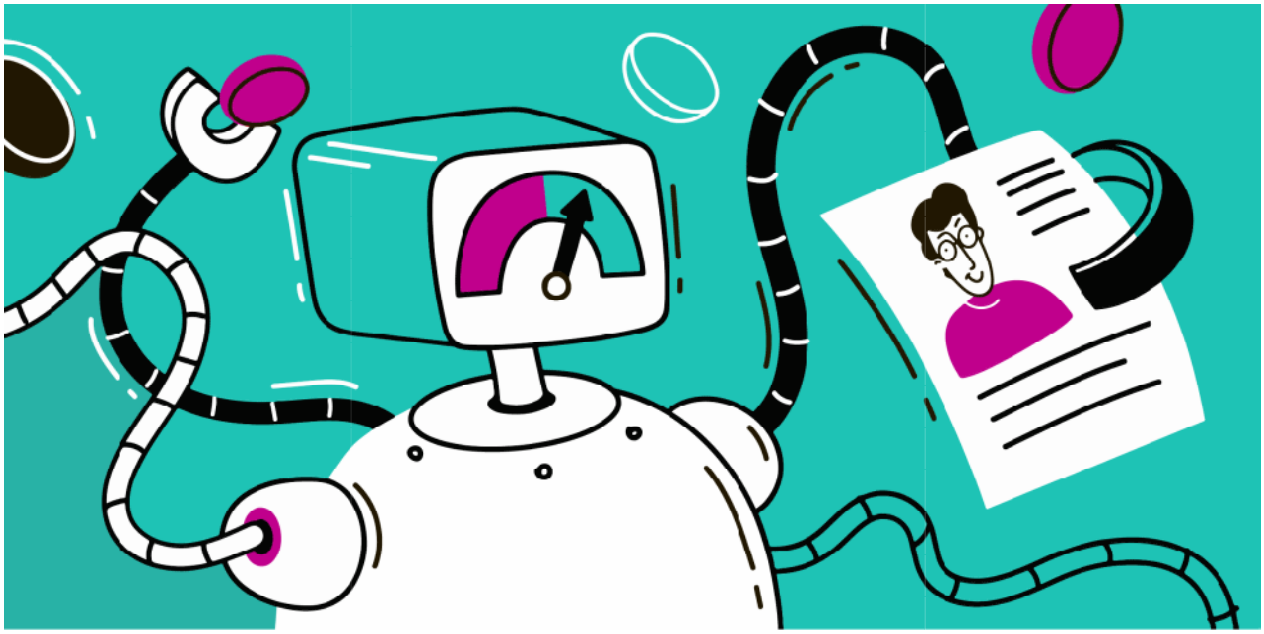
Анализ данных в современном мире чем-то напоминает добычу золота во времена золотой лихорадки. Данных очень много, и на первый взгляд их трудно структурировать, а тем более извлечь из них что-то полезное. Благодаря развитию вычислительной техники у нас есть возможность структурировать и исследовать данные, находить неожиданные зависимости в них и извлекать новые знания.

В этом лонгриде речь пойдет о задачах, решаемых с помощью современных методов анализа данных. Давайте для начала рассмотрим несколько примеров.

Анализ данных играет большую роль в банковской сфере. Давайте представим ситуацию: клиент пришел за кредитом на открытие своего дела. Но как нам оценить клиента, как понять, выдавать ему кредит или нет? Для того чтобы принять решение о выдаче кредита, мы должны предсказать, сможет ли клиент выплатить деньги банку. Ситуация, когда клиент не может заплатить по кредиту, называется дефолтом. И именно вероятность этого события является ключевым показателем при принятии решения банком.

Если бы у нас не было машинного обучения, то предсказанием дефолта клиента занимался бы человек, оценивая клиента по различным признакам. Человеку для этого требуется время: нужно обратиться в бюро кредитных

историй, узнать о доходах клиента, судимости и еще множество показателей. Следовательно, банк может просто не успеть обработать все поступающие запросы.



Но мы живем в мире, где искусственный интеллект проник в банки и позволил автоматизировать многие процессы. Вместо того чтобы поручать человеку всю рутинную работу, связанную с оценкой клиента и предсказанием дефолта, мы создаем алгоритм, который способен на основании исторических данных предсказать возможный дефолт конкретного клиента. Такой алгоритм освобождает человека от рутинной работы и позволяет заниматься более сложными и интересными задачами. Также он позволяет сократить время на обработку запроса одного клиента, а значит, теперь мы способны обработать большее количество запросов, выдать больше кредитов и больше заработать.

Рекомендации фильмов, музыки, товаров

Наверняка вы, заходя на сервисы поиска товаров, музыки или фильмов, замечали, что кроме сортировки контента по тематикам есть отдельный блок с рекомендациями. Этот блок сформирован на основе предпочтений пользователя, т. е. на основе статистики просмотров того или иного контента.

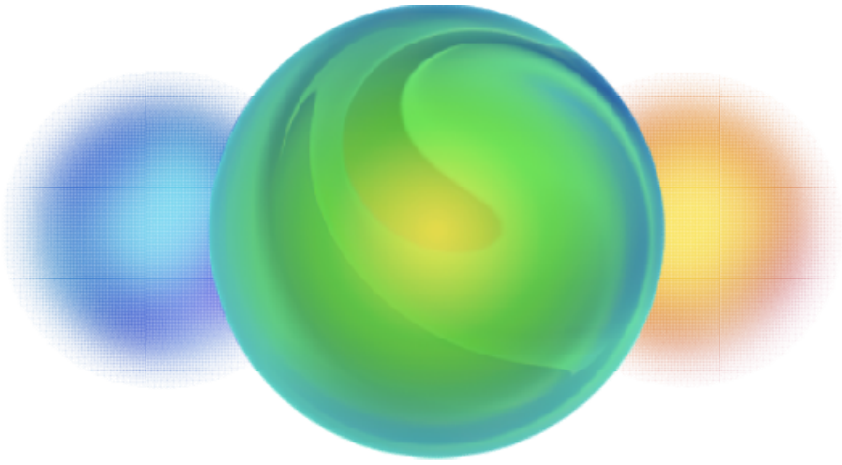
И эти рекомендации формируются не вручную, а с помощью методов анализа данных. Все действия, которые пользователь совершает на сайте, записываются и анализируются в автоматическом режиме. Благодаря этому есть возможность подбирать контент, релевантный для конкретного человека, и упрощать поиск контента. По сути, мы снова экономим время человека, находя за него то, что ему нужно.

Распознавание лиц

Кроме банковской сферы и сферы развлечений анализ данных применим и в сфере безопасности. Сейчас повсеместно используются системы видеонаблюдения, что позволяет фиксировать преступления и быстро на них реагировать. С развитием таких систем появилась возможность идентифицировать лица преступников. Проблема в том, что видеоматериалов очень много, и просматривать все вручную займет много времени. Поэтому здесь к нам на помощь вновь спешит анализ данных.

Оказывается, что обрабатывать изображения тоже можно в автоматическом режиме. Существуют методы, которые позволяют детектировать отдельные объекты на изображениях с точностью, сравнимой с человеком. В число этих объектов входят и лица. Методы распознавания лиц способны с большой скоростью и достаточно точно определить, кто конкретно находится в кадре, что значительно повышает скорость реагирования правоохранительных органов на возможные угрозы.

Голосовые помощники



Все слышали о Siri, Алисе и семействе голосовых помощников от «Сбера» (Салют, Афина, Джой). Наверное, никого не удивит, если я скажу, что в основе их работы лежит все тот же анализ данных. Голосовые помощники стали удобным инструментом при решении повседневных задач.

Например, можно голосом включить музыку, узнать погоду, получить информацию о пробках и многое другое. С голосовыми помощниками можно и просто поговорить, правда, иногда их ответы могут показаться странными. Алгоритмы, используемые здесь «под капотом», позволяют с высокой точностью распознавать человеческую речь и генерировать релевантный ответ.

Сходства задач, решаемых с помощью анализа данных

На первый взгляд может показаться, что анализ данных способен решать очень широкий спектр задач, и между ними сложно найти что-то общее. Но все же есть одна существенная деталь. Как можно увидеть из примеров, все задачи требуют наличия некоторого объема информации или примеров, на которых можно «обучить» алгоритм.

В случае с кредитным скорингом для того, чтобы сделать прогноз относительно дефолта конкретного клиента, алгоритм должен знать, какие клиенты уже попали в такую ситуацию, а какие нет. Получается, алгоритм должен обучиться на исторических данных, чтобы принять наиболее объективное решение в настоящем. Исторические данные в этом случае представляют собой таблицы, где для каждого клиента указаны его характеристики и факт дефолта. Алгоритм, обучаясь на этой информации, пытается построить зависимость между характеристиками клиента и фактом наступления дефолта.

Вообще, почти любая задача анализа данных сводится к нахождению неизвестной зависимости между двумя множествами. Одно из множеств — это наблюдения (в случае кредитного скоринга — характеристики клиентов), второе — целевое значение (факт дефолта). Решение обычно состоит из следующих шагов:

- **Бизнес-понимание задачи**

  Формулировка задачи совместно с представителями бизнеса для решения конкретной проблемы, а также подготовка плана решения.
- **Понимание данных**

Определение, в каких данных мы нуждаемся для решения задачи, поставленной в предыдущем пункте.

- **Подготовка данных**

  Приведение данных к такому виду, который будет пригоден для построения модели. Для разных моделей подготовка данных может происходить по-разному, поэтому через этот этап можно проходить несколько раз.

- **Моделирование**

  Выбор модели и ее обучение на имеющихся данных. Часто строится несколько моделей, и для каждой данные могут подготавливаться отдельно.

- **Валидация**

  Проверка качества лучшей построенной модели, построение альтернативных моделей. Часто проверяется не только точность модели, но и ее интерпретируемость, так как это важно для бизнеса.

- **Внедрение**

  Встраивание модели в бизнес-процесс и поддержка в течение ее работы.

В случае с рекомендациями алгоритм также ориентируется на историю. Здесь анализируются статистические данные, и на их основе принимаются решения о рекомендации.

Если говорить про распознавание лиц, то алгоритм обучается на «размеченных» изображениях, т. е. на таких изображениях, на которых вручную выделены и идентифицированы лица. Разметкой занимается человек, а алгоритм пытается подстроиться под поведение человека.

То же самое происходит и в других задачах, решаемых с помощью анализа данных. Алгоритм, зная о решениях, принятых в прошлом, находит закономерности, которые позволяют ему принимать наиболее релевантные решения в настоящем. Методы анализа данных не способны дать рекомендацию по контенту, если нет информации о предпочтениях

пользователя. Не получится идентифицировать лицо на изображении, если алгоритм никогда лиц «не видел», и т. д. Поэтому все, что нам нужно, — это больше данных.

Подумайте, как может использоваться анализ данных для выявления клиентов, которые скоро уйдут из банка, или для предсказания динамики толщины ледяного покрова в Арктике.

Различия задач, решаемых с помощью анализа данных

Несмотря на то, что у всех таких задач общая суть, нам нужны данные, чтобы найти закономерности и принять решение. Методы, которыми решаются эти задачи, могут быть совершенно разными. И выбор метода тоже зависит от данных.

Данные могут быть представлены в совершенно разных видах, например:

- Табличные данные по строкам содержат разные наблюдения (например, клиентов) и по столбцам характеристики этих наблюдений. Один из столбцов — это целевая переменная (например, доход клиента), которую мы хотим предсказать по всем остальным столбцам (другим характеристикам клиента)
- Изображения обычно представлены в виде трехмерных тензоров с числами, отражающими яркость пикселей по каналам RGB
- Текст состоит из набора слов, который преобразуется в матричный вид
- Звук представлен в виде последовательности амплитуд звукового сигнала во времени, которую часто преобразуют в формат спектрограммы

Например, в задаче кредитного скоринга мы используем табличные данные, и на них хорошо работают классические методы машинного обучения, такие как логистическая регрессия и градиентный бустинг над решающими деревьями.

А вот когда нужно идентифицировать человека на видео, понадобятся сверточные нейронные сети, чтобы извлечь признаки из изображения и с их помощью решить задачу.

Для голосовых помощников, помимо всего прочего, используются методы обработки естественного языка. В качестве данных мы используем обычный текст и работаем с ним уже с помощью специальных архитектур нейронных сетей: рекуррентных и трансформеров.

Для каждого типа данных придуманы мириады алгоритмов, каждый со своей спецификой. Их развитие и изобретение новых активно продолжается и сейчас.

Необходимость применения машинного обучения

В этом лонгриде мы рассмотрим, какую пользу для бизнеса приносит машинное обучение и какие есть особенности проектов по ML. Для лучшего понимания возьмем банк и команду кредитного скоринга. Команда занимается созданием алгоритма, который автоматически решает, кому выдать кредит, а кому нет. Цель команды — снизить риски банка и увеличить прибыль от выдачи кредитов.

Часто новички в анализе данных считают, что машинное обучение стоит внедрять по умолчанию, так как думают, что это новая технология. Однако новое не всегда значит лучшее и подходящее конкретно под ваши задачи. Некоторые компании хотят внедрить машинное обучение, чтобы показать своим клиентам стремление к инновациям, хотя для решения их задач машинное обучение не является чем-то необходимым. На самом деле задача машинного обучения — упростить и ускорить процессы в компании, автоматизировать рутинную работу, удешевить бизнес-процессы или создать новые и полезные.

Например, для принятия решения о выдаче кредита раньше нанималось большое количество сотрудников, которые работали в разных отделениях банков. Заявка о выдаче кредита могла рассматриваться до 2 недель. Однако с внедрением машинного обучения процесс ускорился до нескольких секунд (рис. 1). Алгоритм, в отличие от сотрудника, не может совершить ошибку по причине усталости, а также не способен мошенничать, например, одобрив кредит для друга.

Рис. 1. Процесс выдачи кредита в 90-х гг. и в наше время

Также ML может быть использовано для решения других задач:

- Следить за тем, какие сотрудники хотят покинуть компанию, чтобы предложить им более выгодные условия
- Прогнозировать экономические показатели, например инфляцию, чтобы лучше подбирать ставки по вкладам

Все эти действия прямо или косвенно помогают бизнесу увеличить прибыль и сократить расходы. Однако машинное обучение применяется не только для финансовых целей бизнеса: алгоритмы компьютерного зрения используются для распознавание номеров автомобилей (рис. 2) или мониторинга производственных систем, чтобы сохранить жизнь рабочих на опасном производстве.

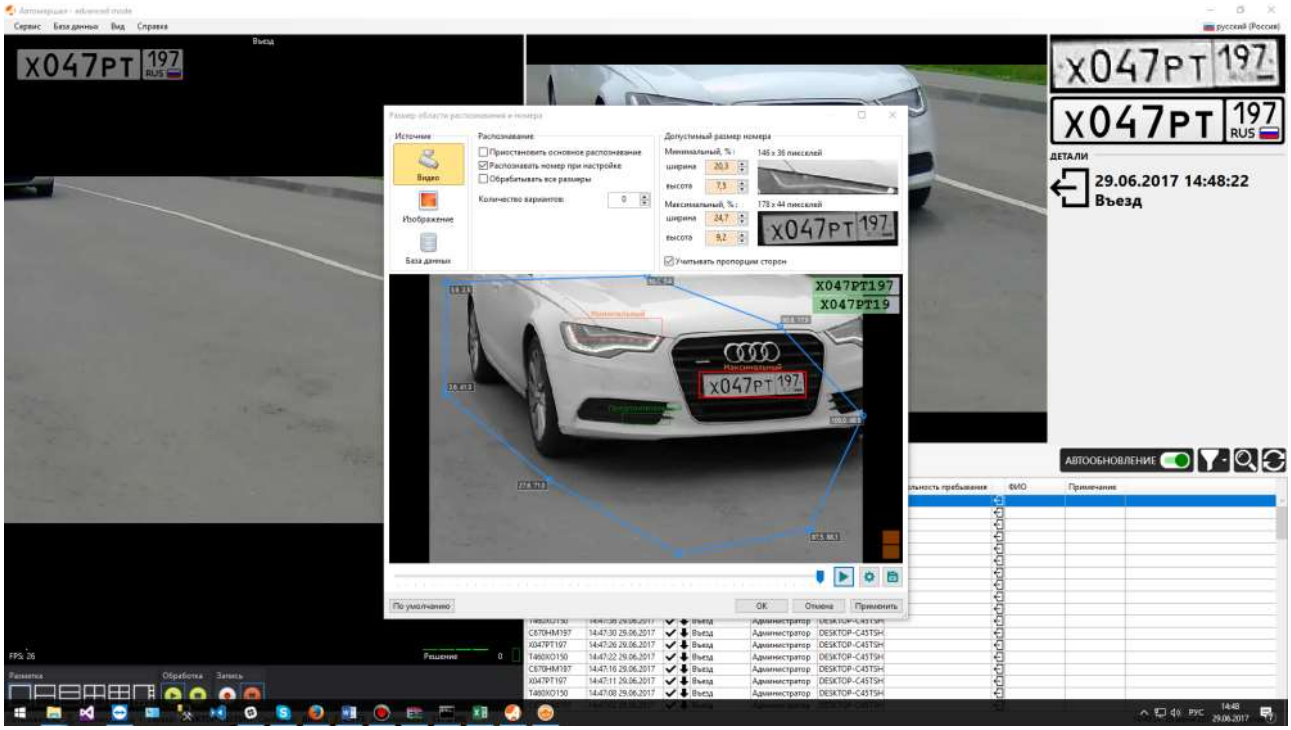


Рис. 2. Система распознавания номеров автомобилей «АвтоУраган» / © recognize.ru

Раньше задачи машинного обучения решались в ручном режиме или при помощи других, менее эффективных методов. Причины взрывной популярности машинного обучения схожи с ростом популярности ИТ 20 лет назад. Информационные технологии помогли упростить и ускорить операции, которые делались людьми вручную. Так, раньше бухгалтерия велась на бумаге, а сейчас бухгалтеры пользуются автоматизированными системами и тратят намного меньше времени на рутинные расчеты. С машинным обучением похожая история: если раньше решение о выдаче кредита принимал сотрудник банка, то сейчас мы разрабатываем системы, которые могут делать это непредвзято и в течение доли секунды.

Как использовать машинное обучение в бизнесе

В больших компаниях инновации происходят постепенно, в проектной форме. Для того чтобы сделать переход быстрым, результативным и комфортным для сотрудников, есть отдельные стандарты по управлению проектами, в частности в сфере разработки ПО. Так, **PMBOK** является ключевым стандартом в области управления проектами. Однако есть **ряд**

отличий проектов по машинному обучению по сравнению с проектами по разработке ПО:

1. Успех проектов по машинному обучению зависит от разных факторов, в частности качества данных. Даже самые точные модели машинного обучения будут работать плохо, если они обучены на некачественных данных с большим количеством ошибок и пропусков. Поиск качественных данных — важная часть проекта по машинному обучению

2. При реализации проекта тестируется большое количество гипотез. Если в ИТ-разработке заранее можно понять, какие программные библиотеки или платформы нужны для решения задачи, то в DataScience приходится работать с моделями. Перед внедрением в production (запуском модели для решения задачи бизнеса) нужно протестировать большое количество разных моделей, способов предобработки данных, подобрать параметры моделей. Иначе говоря, нужно проверить много гипотез и провести много экспериментов. Собственно, поэтому DataScience — это именно наука о данных

3. Изначально крайне сложно оценить ресурсы, необходимые для реализации проектов. Неясно, какой сложности должна быть изначальная модель или модель, которая должна улучшить результат. Иногда лучший результат показывают вычислительно простые модели, а иногда нужны большие нейронные сети, которые во время работы потребляют много ресурсов. В начале проекта нельзя сказать наверняка, какая модель заработает лучше

Таким образом, помимо нюансов ИТ-проектов, в ML-проектах возникают трудности, связанные с дополнительной неопределенностью, присущей исследовательским проектам. Поэтому для проектов по машинному обучению существуют отдельные стандарты.

Стандарт CRISP-DM: решение задач анализа данных

Стандарт CRISP-DM для работы с данными описывает, как происходит создание и внедрение моделей в индустрии. Также мы узнаем, чем аналитик данных занимается на самом деле.

В разработке ПО существуют стандарты по управлению проектами. Эти стандарты плохо подходят для проектов по анализу данных, поэтому в анализе данных появились свои. Их порядка 10, и они похожи между собой. Часто крупные компании, напримерMicrosoft, стараются делать собственные внутренние стандарты. Так, стандарт Microsoft включает в себя 4 этапа работы с моделью: постановка бизнес-задачи, изучение данных, моделирование и внедрение. Мы же остановимся на стандарте CRISP-DM. Несмотря на то, что стандарт был создан в 1999 г., 42 % компаний все еще используют именно его. CRISP-DM хорошо отражает основные этапы проектов по анализу данных. К тому же этот стандарт межотраслевой и не включает специфику отдельных компании.

Полное название **CRISP-DM** — Cross-IndustryStandardProcessforDataMining (Межотраслевой стандартный процесс интеллектуального анализа данных). Стандарт включает в себя 6 этапов. Они представлены на схеме ниже (рис. 1).

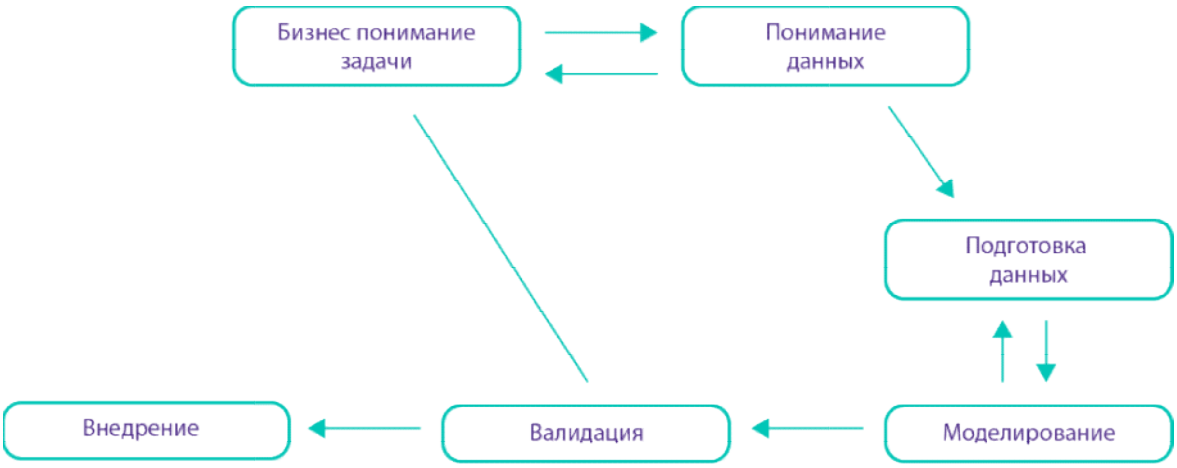


Рисунок 1. Цикл управления в проектах на основе анализа данных по CRISP-DM

Многие новички в машинном обучении думают, что на работе им придется работать только с моделями машинного обучения, а поиском данных и внедрением занимаются другие люди. Это не так. **Обычно аналитики данных участвуют во всех этапах цикла управления проектом.** В этом основное отличие DataScience в индустрии от учебных курсов в университете или интернете. Аналитики данных должны не только уметь строить модели и оценивать их, но и вместе с бизнесом определять, какую задачу нужно решить, искать данные внутри компании, а также внедрять модели. Рассмотрим роли и задачи аналитиков данных на всех этапах работы.

- **1. Бизнес-понимание задачи**

На этом этапе мы формулируем задачу и делаем план проекта. Проект стартует с того, что у бизнеса есть какая-то проблема, и он хочет решить ее при помощи анализа данных. Например, банк хочет выдавать кредиты более надежным заемщикам. Но бизнес не умеет формулировать задачу с точки зрения машинного обучения: менеджеры очень поверхностно знакомы с такими терминами, как классификация, регрессия, кластеризация, построение рекомендаций и т. д. Их задача — понять, вернет человек кредит или нет, оценить по пятибалльной шкале или повысить продажи товаров на маркетплейсе. Главная задача на этапе бизнес-понимания задачи — перевести проблему с языка бизнеса на язык машинного обучения. Также нам надо выбрать метрику, по которой мы поймем, решили мы проблему бизнеса или нет. Бизнес использует свои метрики: объем продаж, прибыль, средняя доходность с заемщика и т. д. В команде аналитиков данных мы используем метрики машинного обучения, например точность (рис. 2).

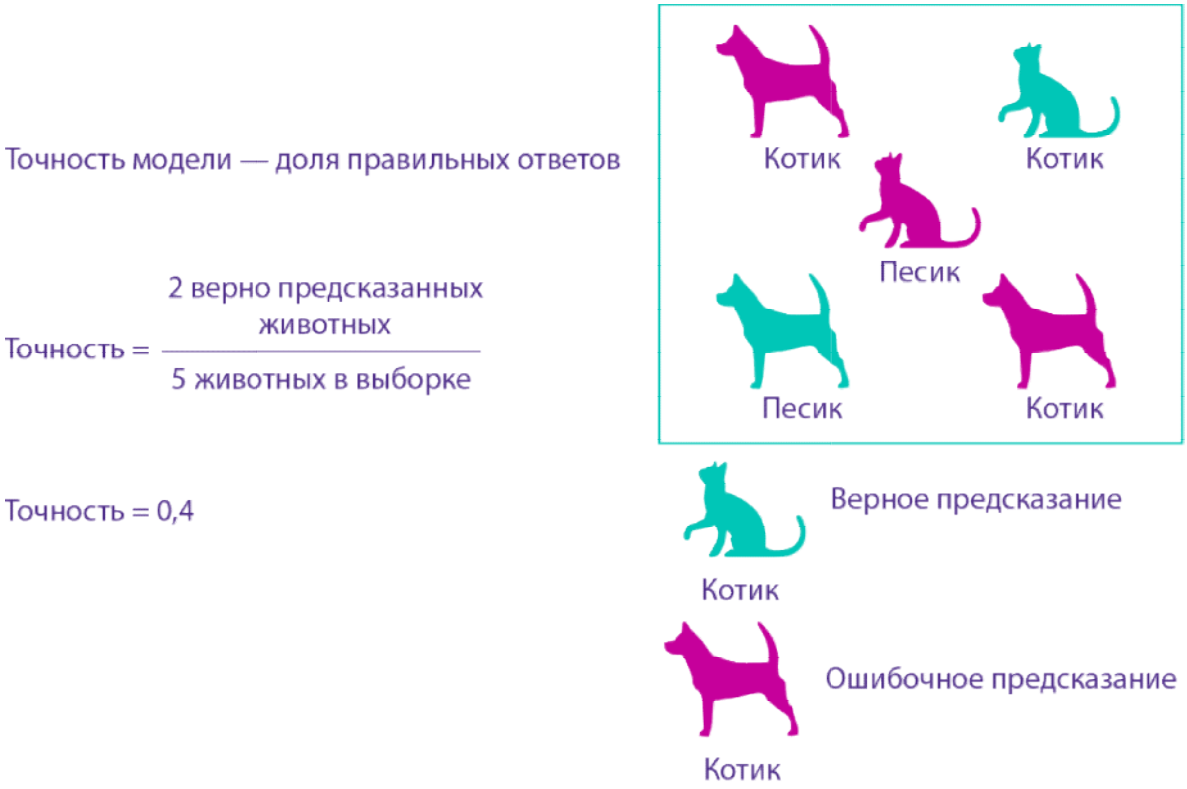


Рис. 2. Точность модели машинного обучения

Надо понять, какую бизнес-метрику необходимо улучшить, подобрать максимально похожую метрику машинного обучения и по ней потом оценивать модель.

- **Понимание данных**

  На этом этапе аналитик данных определяет, какие данные нужны для решения задачи. Половина успеха проектов по машинному обучению — качественные данные и правильная разметка.

  На этом этапе мы решаем:

  - Какие данные нужны
  - Какие данные у нас есть внутри компании
  - Какие данные придется закупать во внешних источниках
  - Нуждаются ли данные в разметке

  Другая проблема — найти данные внутри компании и получить к ним доступ. В больших компаниях могут храниться терабайты данных. Не всегда сразу понятно, кто отвечает за данные и где они лежат. Другой нюанс — безопасность. Некоторые данные персональные, и доступ к ним одобряет служба безопасности. Процесс поиска данных внутри

компании и получения доступа к ним иногда может занимать 2 месяца и более.

- **Подготовка данных**

  На этом этапе решается целый комплекс задач. Мы должны:

  - понять, какие типы данных у нас есть

  Данные могут быть представлены в виде таблицы, необработанных текстов, изображений или графов. С каждым из этих типов данных хорошо работают разные модели машинного обучения.

  - проверить качество данных

  Данные могут содержать ошибки, пропуски, аномальные значения. Необходимо хорошо изучить данные и решить, как мы будем чистить и трансформировать данные.

  - почистить данные

  Тут мы заполняем пропуски, исправляем ошибки, где это возможно, удаляем аномальные объекты.

  - преобразовать данные

  Не все модели умеют работать с любыми типами признаков, например с категориальными переменными. Нужно преобразовать такие переменные в формат, понятный модели.

  - Отобрать релевантные признаки

  Не всегда больше признаков — лучше. Иногда бывает, что в выборке более 3 тысяч признаков, но не все из них полезны. Нам надо оставить только самые актуальные. Обычно оставляют 20–30 самых подходящих. Однако этот этап нужен не всегда. Например, глубокие нейронные сети могут сами извлекать признаки из неструктурированных данных: текстов, картинок, транзакций.

Сбор данных, очистка и подготовка данных занимают больше 50 % времени аналитика данных, в то время как моделирование может занимать только 20 %.

Хорошая подготовка данных может значительно увеличить качество модели. Поэтому процесс подготовки данных обычно происходит несколько раз.

- **Моделирование.**

  На этом этапе нам надо выбрать модель машинного обучения и обучить ее, проверить качество модели по выбранным метрикам, а также подобрать параметры модели. Обычно тестируют сначала самую простую модель, ее называют бейзлайн. Потом проверяют более сложные модели, которые могут дать более высокое качество. Причем некоторые модели требуют иного набора данных. Поэтому часто необходимо повторить фазу подготовки данных. Этап моделирования заканчивается тем, что аналитики данных сравнивают несколько моделей, выбирают лучшую и отдают на проверку бизнесу. При этом мы должны следовать от простого к сложному.

Сначала учим простые модели, так как это быстрее и проще внедрять, а потом уже начинаем обучать нейронные сети.

- **Валидация модели.**

  На этом этапе бизнес решает, подходит ли ему построенная модель. Существует несколько причин не принять модель:

  - модель не решает проблему, плохо работает на бизнес-метрике, хотя на метрике машинного обучения работала хорошо
  - модель дорого или сложно поддерживать
  - модель плохо работает на некоторой группе объектов, например на заемщиках с очень высоким доходом

  Основная цель этапа — проверить, существуют ли какие-то другие нюансы, связанные с моделью, которые мы не рассмотрели ранее. Например, модель может отказывать в кредите всем женщинам, что не этично. Если все хорошо, бизнес принимает решение о внедрении модели. Если модель отклонена, мы возвращаемся на более ранние этапы. При этом то, на какой этап мы возвращаемся, зависит от того, чем именно заказчик недоволен. Мы могли изначально взять не те

данные для обучения или неверно понять задачу. В таком случае необходимо вернуться на стадию бизнес-понимания задачи.

- **Внедрение.**

  В самом конце нужно внедрить модель и поддерживать ее. Внедрением модели обычно занимается бизнес, однако часто приходится помогать бизнесу на этапе внедрения. На данном этапе модель часто переписывается на более эффективный язык программирования, напримерScala, C++, Java. Также модель со временем может устаревать, так как данные меняются. Периодически необходимо переобучать модели на новых данных.

Иногда цель проекта — это не построить модель, а улучшить понимание данных для принятия более эффективных решений. Тогда на этапе внедрения полученные знания структурируются и представляются в виде отчета, презентации и т. д.

Роли в проектах по анализу данных

Мы рассмотрели, как проходят проекты по анализу данных. Теперь давайте поговорим о том, на какие профессии делятся аналитики данных и чем они отличаются. На самом деле «аналитик данных» не всегда является отдельной профессией: все зависит от размера команды по анализу данных. Если в отделе по анализу данных 2–3 человека, то они обычно занимаются всеми задачами, так или иначе связанными с DataScience. Однако с ростом бизнеса и количества задач роль аналитика данных делится на отдельные, более мелкие роли.

Ключевые профессии в анализе данных

- **Аналитики (дата-аналитики, dataanalytics).**

Это люди, которые помогают бизнесу принимать решения. Основной продукт их работы — отчет, презентация к совещанию или другой инструмент, который помогает менеджерам понять, куда вести бизнес. Обычно аналитики слабо касаются машинного обучения, в основном для

понимания влияния разных признаков друг на друга. По CRISP-DM аналитики вовлечены на этапе бизнес-понимания задачи и понимания данных, а также на этапе валидации.

Главные инструменты работы аналитика:

- SQL

Данные хранятся в базах, и аналитики должны уметь доставать их оттуда с помощью языка запросов.

- BI-системы

Это специальные системы для построения интерактивных диаграмм-отчетов для мониторинга состояния бизнеса. Такие диаграммы называются дашбордами.

- Python

Аналитики используют Python для аналитики, тестирования гипотез, проведения A/B-тестов.

От аналитика ждут хороших знаний математики, в первую очередь теории вероятностей и статистики, а также понимания бизнес-области. Знание ИТ-инструментов ограничивается Python, SQL.

- **Инженеры данных (dataengineers).**

Их задача — собрать все данные внутри фирмы и структурно хранить их в специальных системах. Также инженеры данных следят за целостностью и качеством данных, контролируют доступы к данным, разворачивают системы высокоэффективной обработки данных. В проектах по предсказательной аналитике они могут помочь с поиском данных. Однако у инженеров данных много своих задачи, и помощи в поиске данных часто приходится ждать.

Главные инструменты инженера данных:

- SQL

Он нужен для проектирования баз данных.

- Системы обработки больших данных: Spark, Hadoop и другие

Они нужны для преобразования больших объемов данных вычислительно эффективно.

- Python или более производительный язык программирования (C++, Java, Scala)

Он используется для операций по преобразованию данных, когда можно обойтись без Spark/Hadoop или в сочетании с этими инструментами.

От инженера данных ожидается хорошее понимание области, в которой работает бизнес, и знание ИТ. Однако инженеры данных меньше работают с математикой, поэтому будет достаточно математических основ.

- **Инженер машинного обучения (ML engineer, МО-инженер).**

Это человек, который отвечает за моделирование. Главный продукт МО-инженера — модель, автоматизирующая какой-либо процесс, например решающая, выдать кредит человеку или отказать. Инженеры машинного обучения занимаются в основном подготовкой данных для обучения, моделированием и оценкой качества модели.

Главные инструменты инженера данных:

- Python

Используется для экспериментов с моделями, т. к. в нем есть большое количество библиотек, и можно быстро собрать и проверить модель.

- Математика

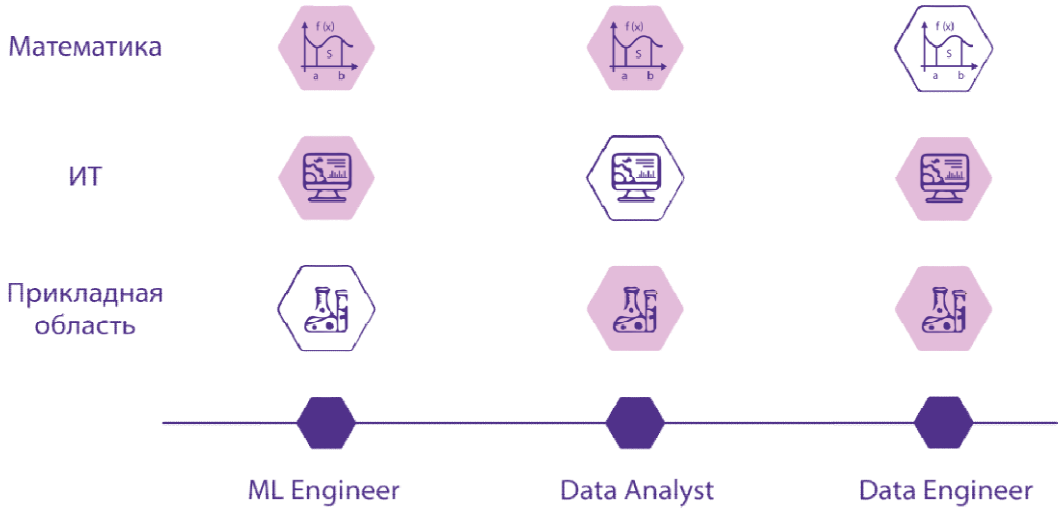
Инженеру машинного обучения важно понимать, как работают модели изнутри, какие есть преимущества и недостатки каждой модели. Также математика нужна для анализа качества моделей и преобразования данных.

- Более производительный язык программирования: C++, Java, Scala

Иногда МО-инженеры могут помогать программистам внедрять модели или даже внедряют их самостоятельно. Для этого надо знать вычислительно-эффективный язык программирования, чтобы модель работала быстро. Однако это требование не так важно по сравнению с первыми двумя.

Соответственно, для инженера машинного обучения важно в первую очередь знать математику и ИТ. Понимание предметной области тоже нужно, но не так, как для инженера данных или аналитика.

Необходимые компетенции схематично отражены на картинке ниже



Ключевые навыки в специальностях по анализу данных

Другие профессии в анализе данных

Однако остается открытым вопрос: кто занимается разметкой данных, созданием новых моделей и внедрением моделей? В командах средних размеров эти обязанности разделяются между профессиями выше или кем-то не из команды по анализу данных: за разметку отвечают аналитики либо МО-инженеры, за внедрение — бэкенд-разработчики либо инженеры машинного обучения. Созданием новых методов обычно в средних компаниях не занимаются, но в больших компаниях профессий по анализу данных еще больше.

- **MLOps (MachineLearningOperations)**

люди, отвечающие за внедрение моделей и поддержку работы моделей. МО-инженеры не всегда пишут читаемый и эффективный код, а программисты не всегда понимают нюансы моделей. Поэтому появилась отдельная профессия. MLOps — это инженер машинного обучения с уклоном в инженерную составляющую: он не подбирает математически оптимальную модель, но делает так, чтобы выбранная модель работала оптимально с точки

зрения ИТ и кода. MLOps работают с такими инструментами, как AirFlow или MLFlow, Docker, Kubernetes.

- **Researchscientist (исследователь)**

Роль этого специалиста — изучение и изобретение новых моделей машинного обучения. Иными словами, Researchscientists — это ученые в области анализа данных. Основной продукт их деятельности — статьи на конференциях и в международно признанных журналах. Крупные компании тратят большие деньги на исследования, чтобы оставаться конкурентоспособными. Исследователи нужны бизнесу, чтобы быстрее адаптировать самые новые и точные модели под свои задачи. Часто компании выделяют исследователей в отдельные команды, лаборатории или корпоративные университеты. Эти люди ближе всего по задачам к МО-инженерам, но они меньше работают с решением бизнес-проблем и больше с математическим пониманием моделей машинного обучения.

- **Специалисты по разметке данных (Crowdsolutionsarchitects)**

Это специалисты, задача которых — правильно организовать процесс разметки данных. Они выбирают способ разметки, ставят задачу для разметчиков (иногда их называют асессорами), контролируют качество разметки. Специалисты по разметке нужны для минимизации ошибки разметки и экономии ресурсов компании, т. к. разметка — очень дорогой процесс.

Для лучшего понимания можно еще раз ознакомиться со схемой (рис. 2):

Размер команды и роли в DataScience

С кем взаимодействуют аналитики данных

Помимо вышеперечисленных профессий, аналитики данных взаимодействуют с менеджерами проектов, бизнес-заказчиками, владельцами данных и архитекторами ИТ-инфраструктуры (таблица 1).

Таблица 1: этапы проектов по анализу данных и роли, с которыми взаимодействует аналитик данных.

Этап проекта по CRISP-DM	С кем взаимодействует аналитик данных
Бизнес-понимание задачи	Бизнес-заказчик, менеджер проекта
Понимание данных	Владелец данных или его подчиненные: аналитики
Подготовка данных	
Моделирование	
Валидация	Бизнес-заказчик, менеджер проекта
Внедрение	ИТ-архитектор и его подчиненные: инженеры ИТ-инфраструктуры

- **Менеджер проекта**

У каждого проекта должен быть управляющий. В ИТ-проектах эта роль часто отведена менеджеру проекта. Если компания работает по методологии управления проектами Scrum, этот человек может называться Scrum-мастер. Менеджер проекта не обязательно является ИТ-специалистом. Это человек, который помогает определять цели проекта, контролирует ход проекта, сроки и результаты.

- **Бизнес-заказчик**

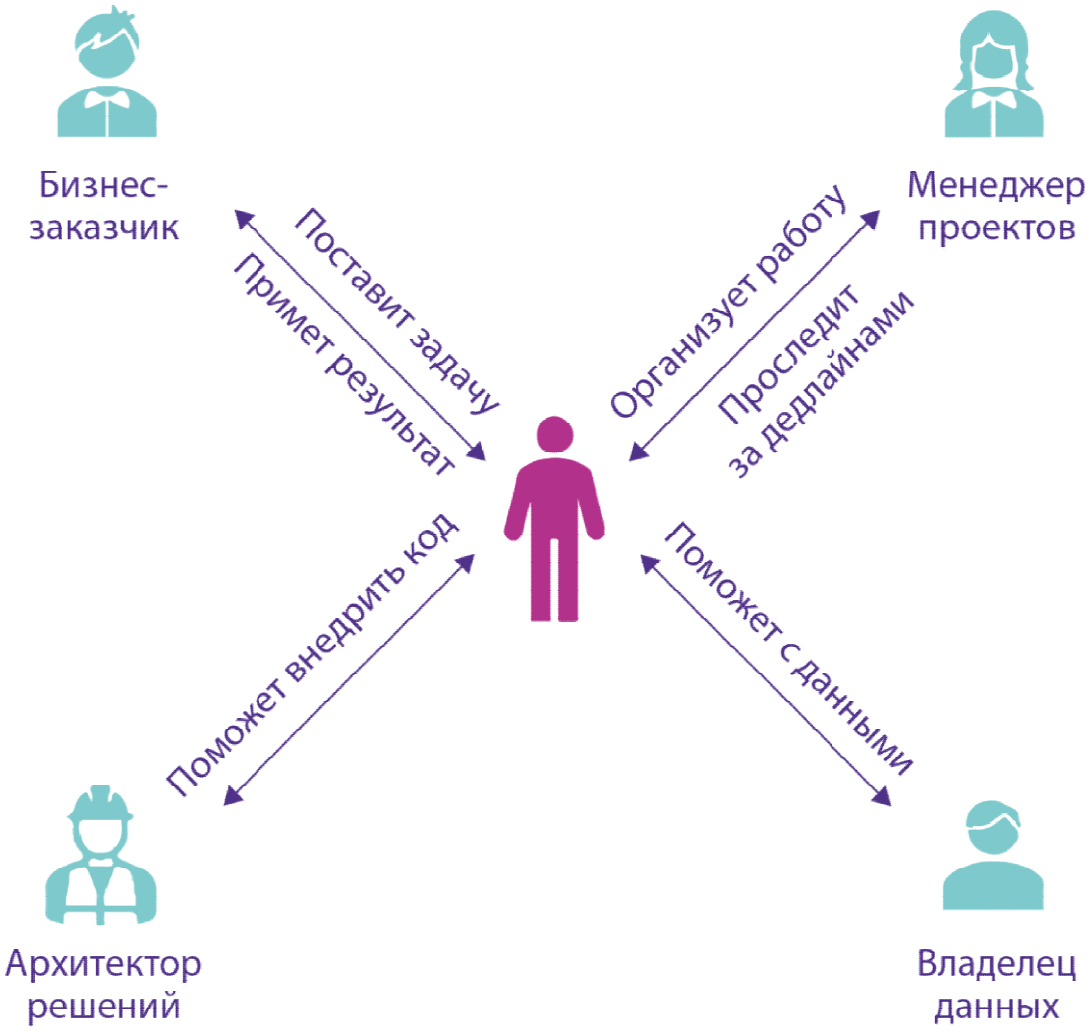
Это руководитель бизнес-подразделения компании, который нуждается в помощи аналитиков данных для решения своей проблемы. Бизнес-заказчики создают запрос на начало проекта, помогают достичь бизнес-понимания задачи, а также принимают модель на этапе валидации. В проекте по внедрению новой модели кредитного скоринга бизнес-заказчиком может быть руководитель кредитного направления банка или его подчиненные.

- **Владелец данных (dataowner)**

В больших компаниях существует большое количество подразделений, в которых создаются и хранятся разные данные. Человек, который отвечает за данные в подразделении, называется **владельцем данных**. Именно к нему мы приходим, когда ищем данные для обучения модели.

- **Архитекторы решений**

ИТ-решения компании состоят из большого количества блоков. За то, какие это блоки и как они выстроены, отвечает архитектор ИТ-решений. Наша модель машинного обучения — это тоже один из множества ИТ-блоков. Поэтому для внедрения модели нам необходимо обсудить с архитектором ИТ-решений или его подчиненным, **инженером ИТ-инфраструктуры** компании, нюансы нашего ИТ-блока с моделью.



С кем взаимодействует аналитик данных