Неудачные проекты в ML

Елена Кантонистова

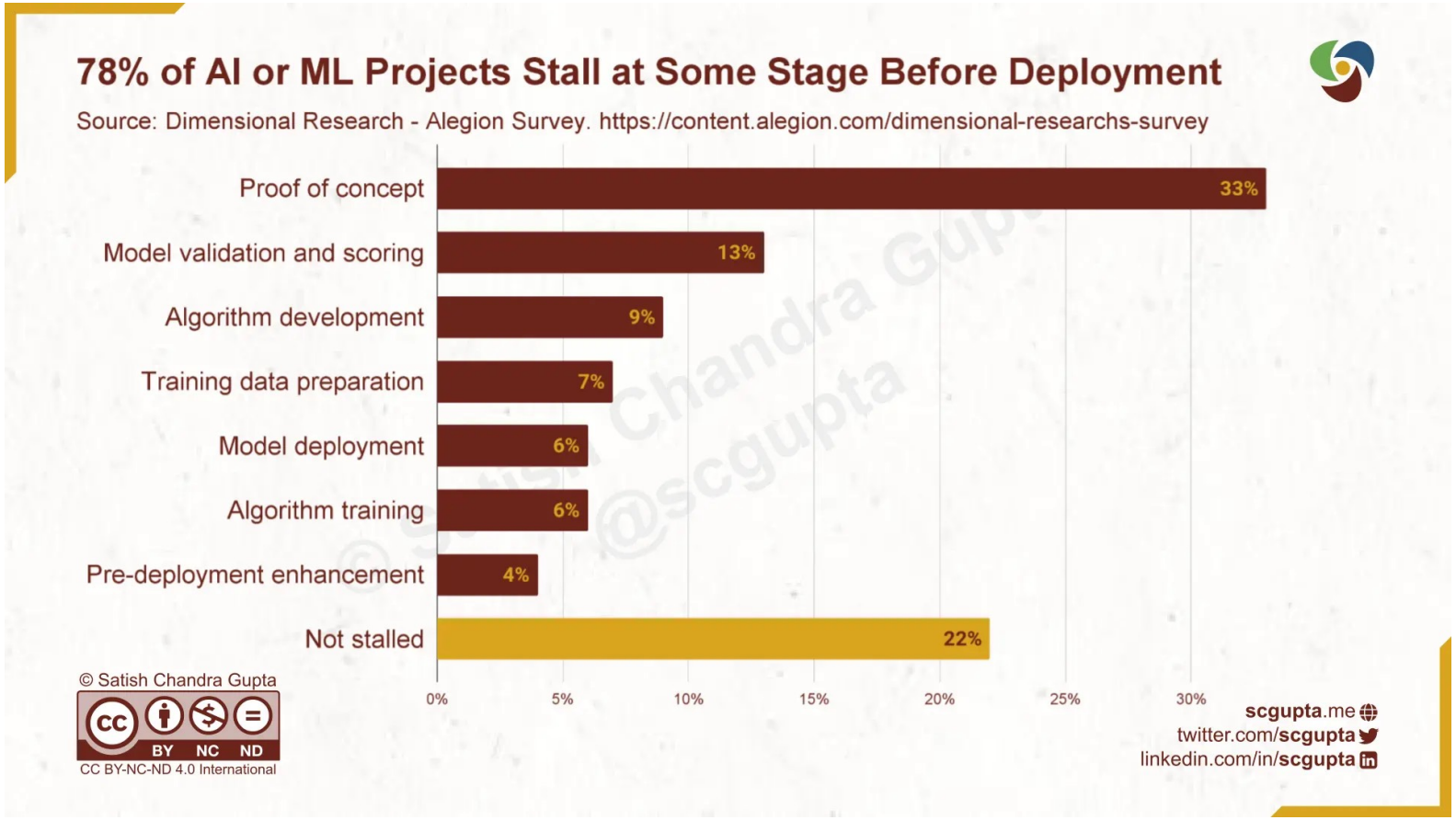
ekantonistova@hse.ru

2023

Как много неудачных проектов в ML?

Как много неудачных проектов в ML?

- Согласно исследованиям середины 2019 года, 78% всех ML-проектов терпят неудачу, не дойдя до внедрения



AI-система автоматического отбора резюме в Amazon (2018)

- Amazon потратил годы на создание автоматической системы отбора резюме
- Идея: автоматически отбирать среди опубликованных резюме наилучших кандидатов на открытые вакансии

AI-система автоматического отбора резюме в Amazon

- Систему постигла неудача! Так как в tech-области в основном работают мужчины, то обучаясь, алгоритм стал дискриминировать женщин:
  - Слово *woman* в резюме автоматически снижало его значимость
  - Слова *promoted* и *captured*, чаще используемые в резюме мужчин, повышали значимость резюме

AI-система от Google по обнаружению диабетической ретинопатии (2020)

Диабетическая ретинопатия — это повреждение сетчатки глаза, которое возникает при сахарном диабете.

Нормальное зрение

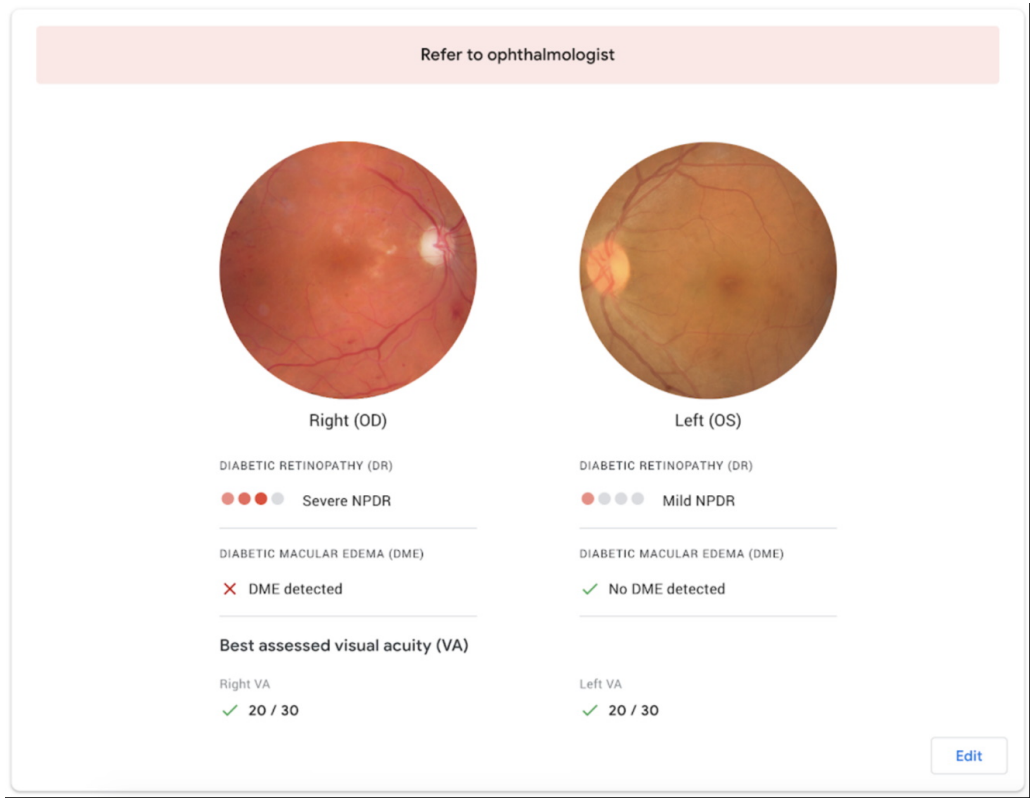
Зрение пациента
с диабетической ретинопатией

AI-система от Google по обнаружению диабетической ретинопатии

- Google Health разработали deep learning-систему, улучшающую диагностику ретинопатии у диабетиков
- Модель анализирует фотографии глаза и по ним выдает вероятность ретинопатии
- Алгоритмы обучались и были протестированы на данных из Тайских клиник - была доступна информация о более чем 4.5 миллионах пациентов!
- Точность обученных моделей на валидации превосходила **90%**!

Как работает диагностика в Таиланде

- Медсестра делает снимок сетчатки глаза пациента
- Когда набирается достаточное число снимков - они отправляются на экспертизу к врачу
- Через 4-5 недель (в виду загруженности) результаты возвращаются пациентам



- AI может сделать диагностику за секунды, рекомендации пациентам без задержки!

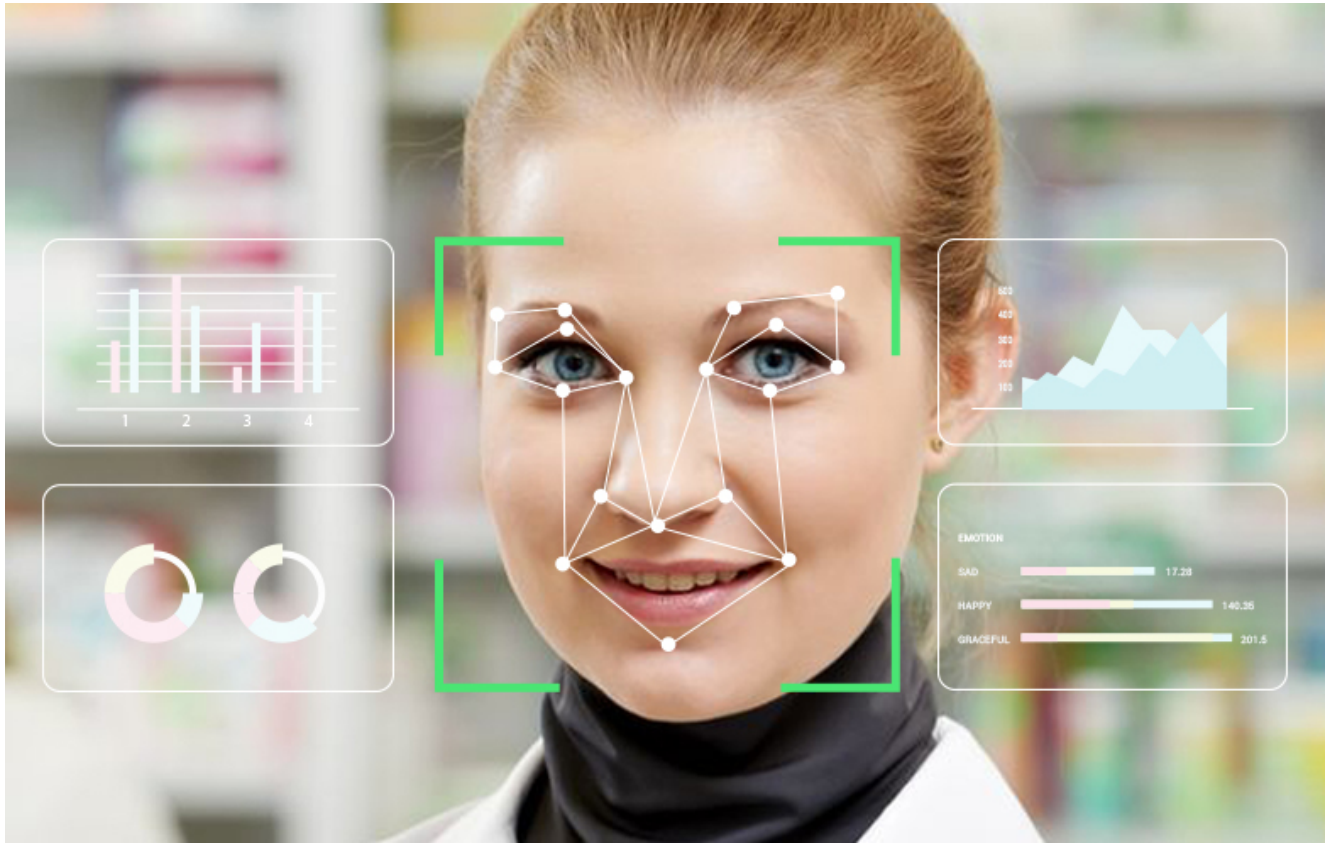
Почему алгоритм потерпел неудачу?

- Только в 2 из 11 клиник были специальные темные комнаты для получения качественных снимков
- Все снимки делались в разных условиях
- Некоторые были частично размыты

- И даже в тех случаях, когда специалист по снимку однозначно видит ретинопатию, AI из-за плохого качества снимка выдавал некорректный результат!
- Кроме того, всплыла проблема медленного интернета в тайских клиниках - из-за которого отправка снимков и получение ответа занимали критически долгое время

Amazon’s Rekognition (2018)

- Amazon разработал AI-систему автоматического распознавания лиц



Amazon’s Rekognition

Система проявила себя крайне неудачно:

- Она определила 28 членов конгресса как преступников
- Кроме того, система дискриминировала чернокожих: из 28 членов конгресса определенных как преступники - 11 были чернокожими (а это почти 40%)!

Ответ Amazon

- Компания заявила, что плохие результаты - следствие плохо откалиброванной системы
- Кроме того, порог уверенности системы 80%, что больше подходит *для распознавания животных, но не людей*. Для легализованных приложений компания рекомендует увеличить порог уверенности до 95%

IBM Watson - диагностика рака (2022)

- IBM разработали AI-систему для диагностики рака
- Однако, система часто давала неверные и даже опасные рекомендации пациентам!

Source: GAO. | GAO-22-104629

IBM Watson - причины неудач

- Из-за специфической области система обучалась на искусственных данных
- Данные были сгенерированы группой врачей, работающих в одной клиники

- Поэтому система ориентировалась на данные с субъективными оценками врачей
- Также врачи не смогли осветить все случаи болезни, и потому система их также не научилась выявлять

IBM Watson - диагностика рака

- IBM разработали AI-систему для диагностики рака
- Однако, система часто давала неверные и даже опасные рекомендации пациентам!

В чем причины неудачных проектов?

- Поставлена неверная задача
  - Задача может не учитывать целей бизнеса
  - Задача может не учитывать потребности пользователей

В чем причины неудачных проектов?

- Проблемы с данными
  - Нет доступа к нужным данным
  - Собираются неверные (ненужные) данные
  - Низкое качество данных

В чем причины неудачных проектов?

- Неверно или не распределена ответственность за результат проекта в команде
  - Data Scientist обучил модель и на этом его дело сделано
  - Инженеры на этапе внедрения могут неверно что-то сделать (может быть, внести незначительные изменения в код модели для увеличения эффективности/скорости), и качество модели упадет

В чем причины неудачных проектов?

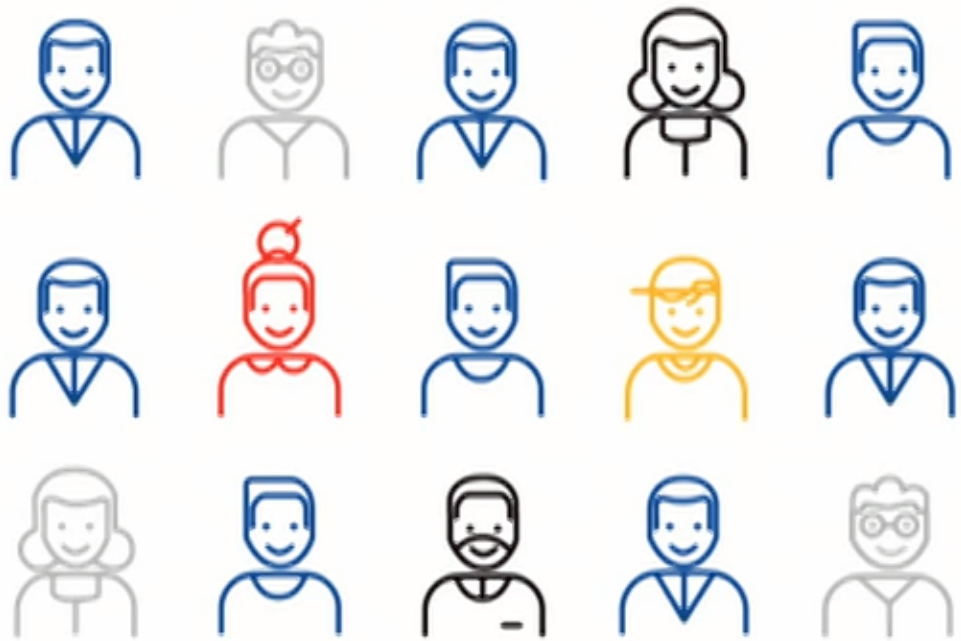
- Трудности с внедрением
  - Слишком высокая стоимость внедрения
  - Медленная работа (inference) модели
  - Неинтерпретируемость модели вопреки требованиям

В чем причины неудачных проектов?

- Изменение в поведении данных (Data Drift)
  - Когда модель внедрена, необходим мониторинг ее эффективности по различным метрикам
  - Рано или поздно поведение данных меняется (например, клиентам перестают нравиться рекомендации), и модель надо пере- или дообучать

Опыт Yandex Data Factory

Активация пользователей онлайн-сервиса



Активация пользователей онлайн-сервиса

- Активация планируется при помощи emailов

  В чем может быть проблема?

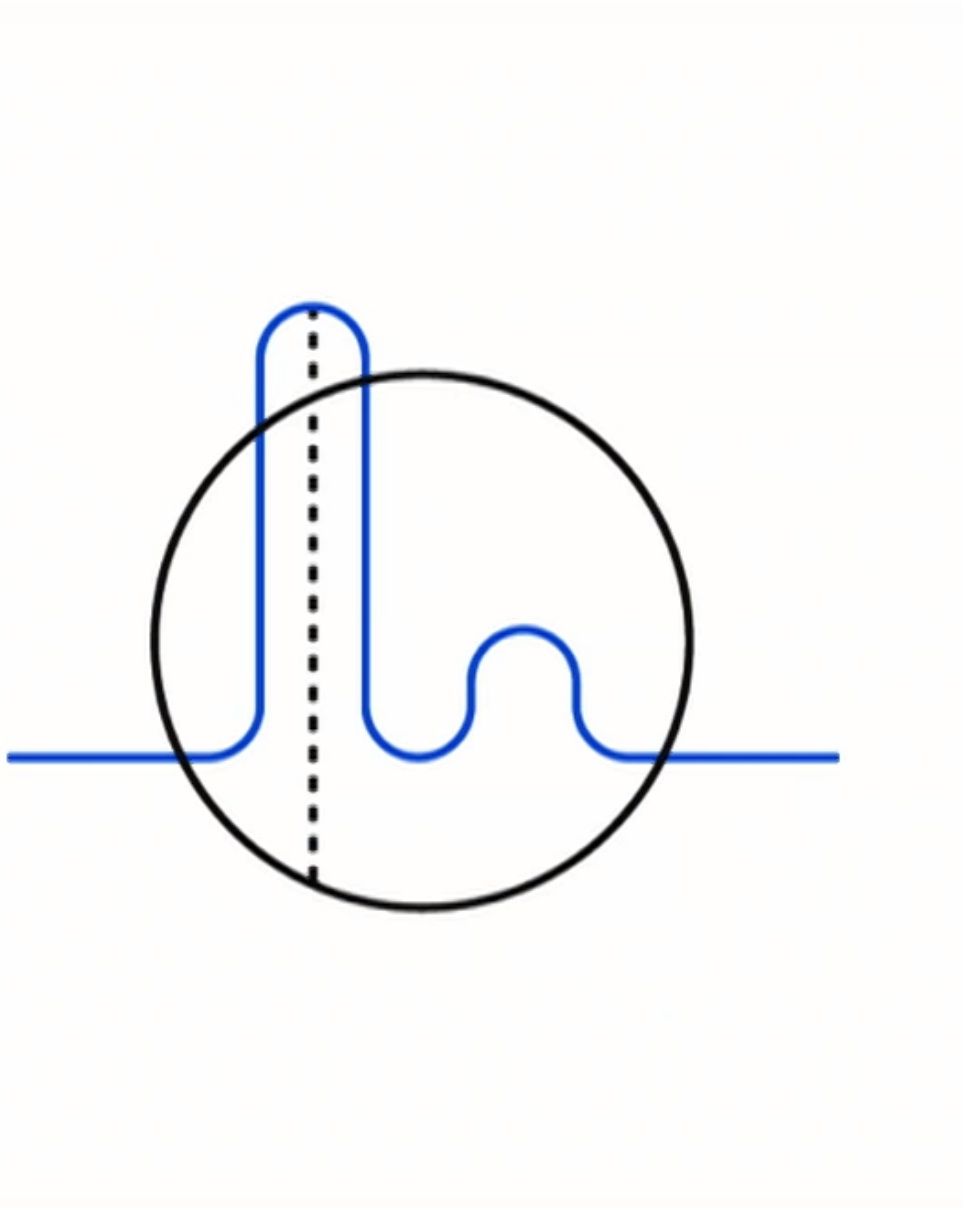
Активация пользователей онлайн-сервиса

- Активация планируется при помощи emailов

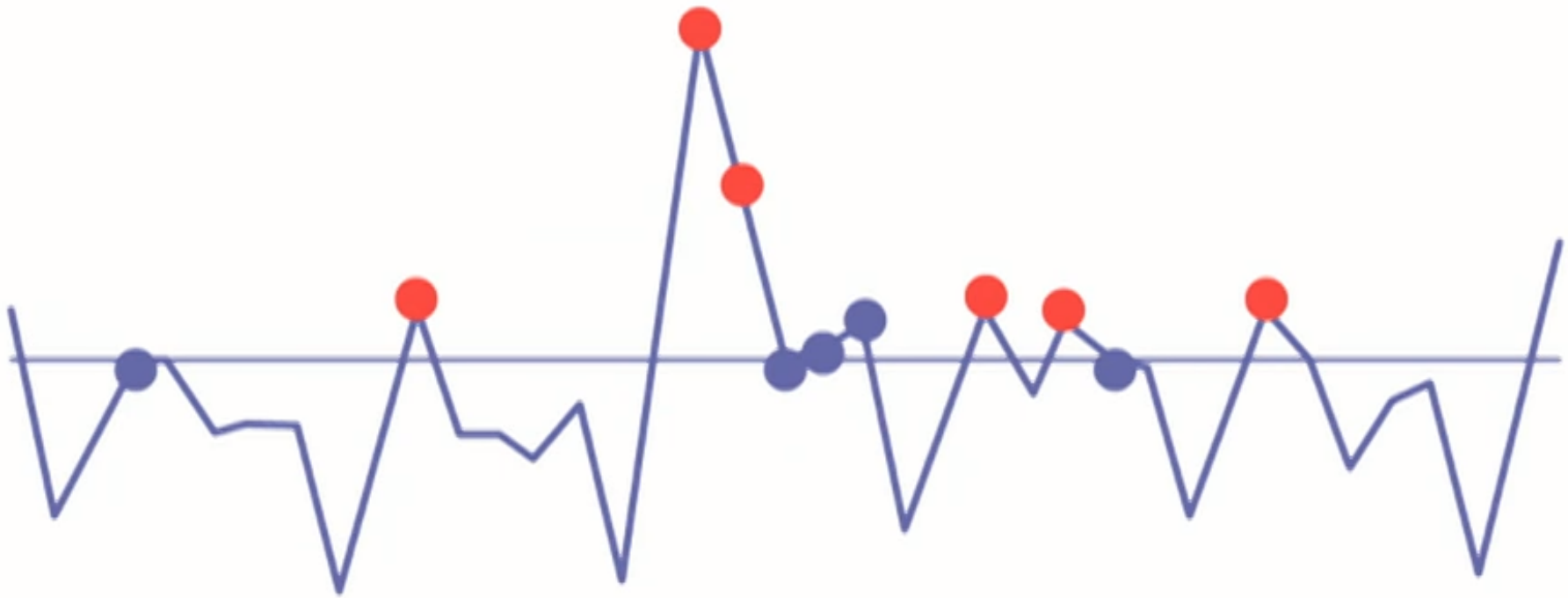
  В чем может быть проблема?

  Emailы бесплатные, и выгоднее пытаться активировать всех

Предсказание пиков посещаемости для компании из сферы услуг



Предсказание пиков посещаемости



Пики посещаемости

- Предсказываем 5 пиковых дней в месяце
- В эти дни предлагается выводить дополнительных сотрудников

Какие могут быть проблемы?

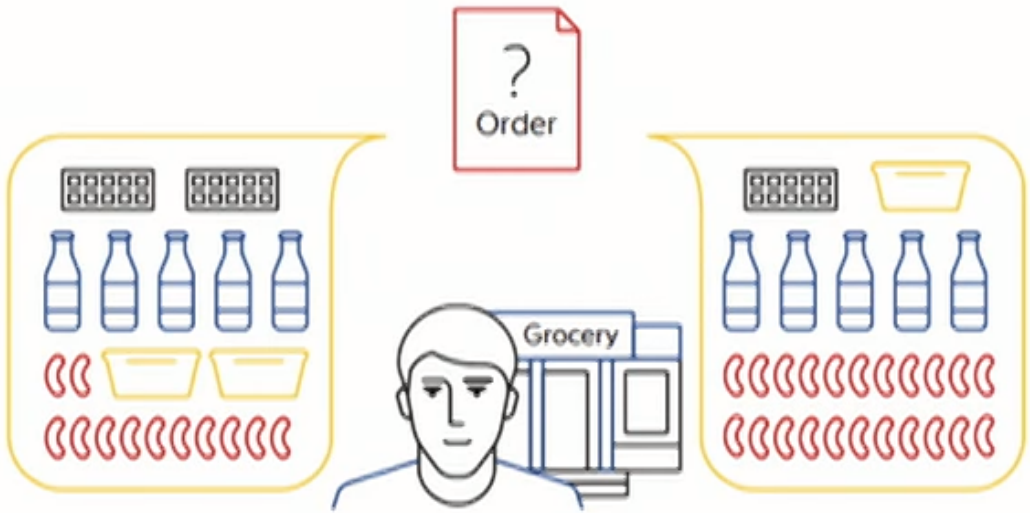
Пики посещаемости

- Предсказываем 5 пиковых дней в месяце
- В эти дни предлагается выводить дополнительных сотрудников

Какие могут быть проблемы?

- **Количество пиковых дней варьируется**
- **Часто даже 7-10 пиковых дней по загрузке мало отличаются друг от друга**
- **Выбранная метрика не позволяет понять, в какие из дней персоналу требуется подкрепление**

Прогнозирование спроса для ритейла



Прогнозирование спроса в ритейле

- Задача: спрогнозировать продажи товаров, чтобы удовлетворить спрос и не потерять в объемах продаж, при этом минимизировав страховой запас

$$WAPE = \frac{\sum_{i,k} |Predict_{i,k} - Fact_{i,k}|}{\sum_{i,k} Fact_{i,k}} \cdot 100\,\%$$

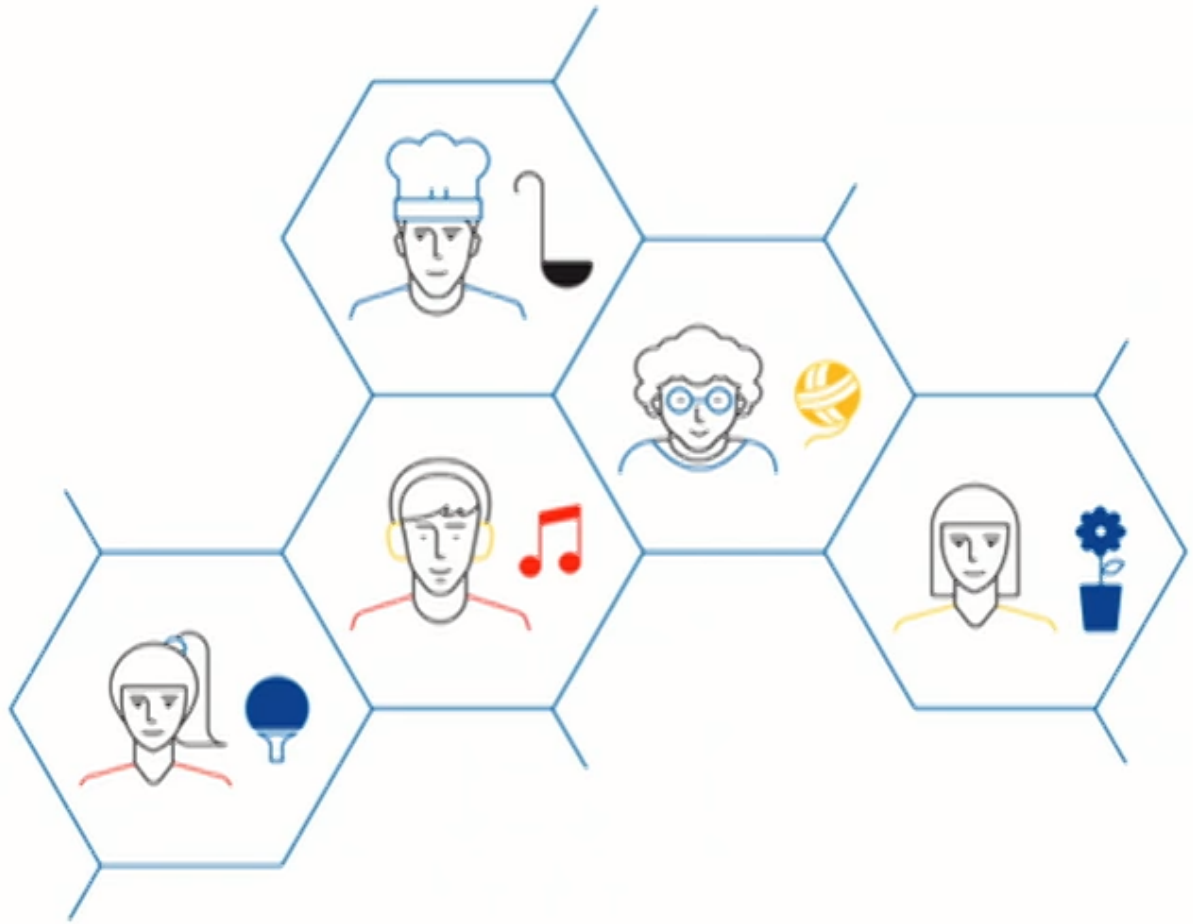
Проблемы?

Прогнозирование спроса в ритейле

- Неходовые товары: по метрике WAPE их продажи выгодно предсказывать как 0
- Из-за этого метрика оказывается не самой удачной

Нужно аккуратно выбирать метрику, исходя из особенностей задачи

Рекомендательная система для увеличения кросс-продаж



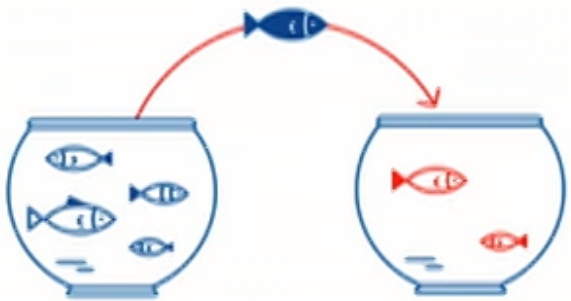
Рекомендательная система для увеличения кросс-продаж

- Провели A/B-тестирование для сравнения эффекта

Рекомендательная система для увеличения кросс-продаж

- Провели A/B-тестирование для сравнения эффекта
- **Одной группе SMS отправили в пятницу, другой - в субботу (результаты невозможно корректно сравнить)**

Предсказание оттока пользователей для телекома



Предсказание оттока

- Была внутренняя модель телекома
- И новая модель Яндекса

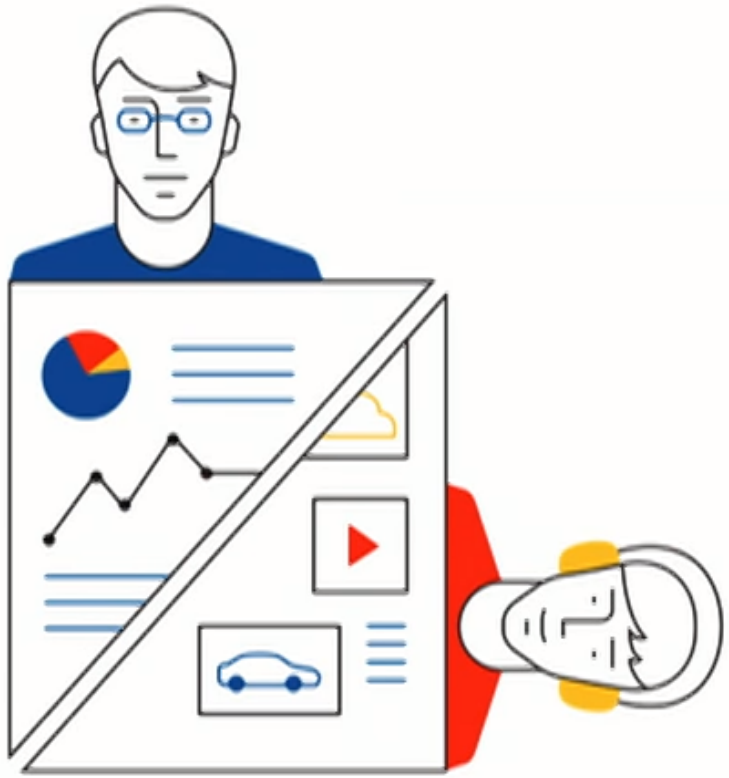
Провели A/B-тестирование, и ни одна модель не показала значимого удержания!

Предсказание оттока - проблемы

Провели A/B-тестирование, и ни одна модель не показала значимого удержания!

- Ранее такой метод удержания не использовали - **скорее всего не работает**
- Обычно метод зависит от сегмента пользователя
- Данные для обучения запаздывали на неделю - **клиенты уже успевают уйти**

Рекомендательная система для крупного интернет-магазина



Рекомендательная система

- По внутренним метрикам модель Яндекса отличная, но на практике не сработала…

Почему?

Рекомендательная система

- По внутренним метрикам модель Яндекса отличная, но на практике не сработала…

  Почему?

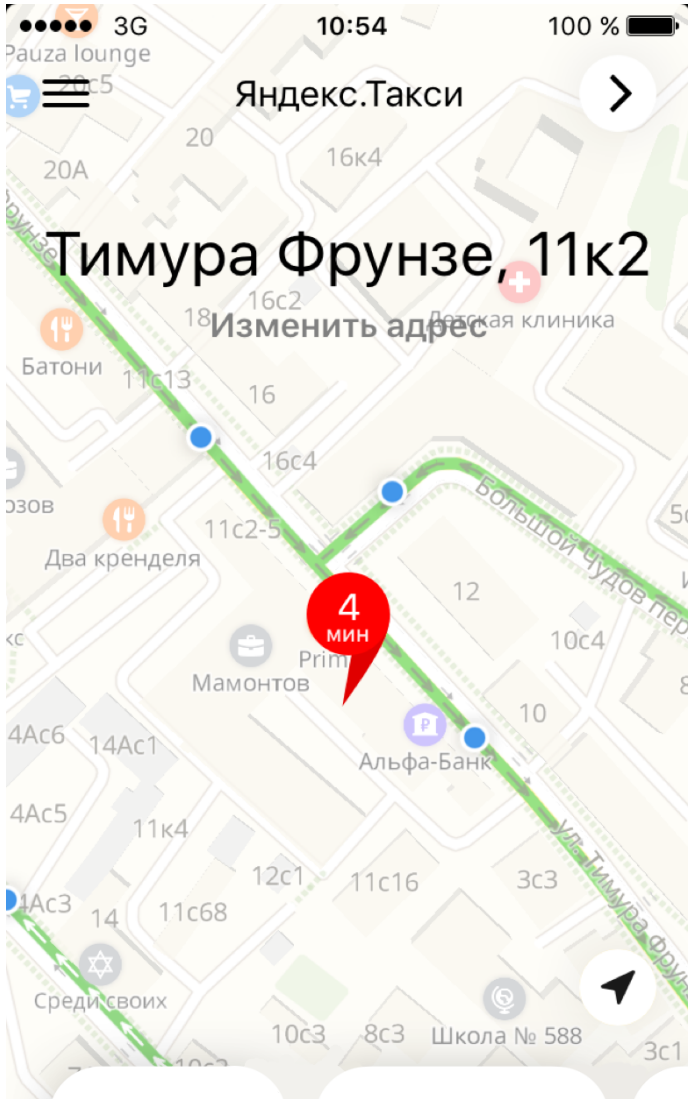
  - **Сверхдорогие товары и оптовики приносят значительную часть прибыли**
  - **Их нужно фильтровать при построении рекомендаций**

Какие ошибки влияют на результат?

- Ошибки в постановке задачи
  - Нужна ли задача в принципе?
  - Можно ли задачу в таком виде решить хорошо?Ошибки в постановке задачи
- Метрика качества: что просишь, то и получишь
  - Учтены ли в метрике все пожелания к результату?
- Проведение A/B-экспериментов и другие сложности
  - В A/B-тесте отличается только модель?
  - Есть ли крупные внешние события, которые могли повлиять на тест?

Успешный кейс!

ML для прогноза времени подачи такси в Яндекс Такси



ML для прогноза времени подачи такси в Яндекс Такси

- Объекты — это пользовательские сессии (признаками объекта это числовые параметры, известные до заказа: количество водителей и пользователей приложения рядом с пином, расстояние до ближайших автомобилей сервиса и так далее)
- Ответы — время, через которое фактически приехала машина

Хотим, чтобы в среднем модель ошибалась мало

ML для прогноза времени подачи такси в Яндекс Такси

В принципе работает неплохо, но довольно часто имеем большие ошибки в прогнозах:

	Mean Absolute Error	Ошибка более 1 минуты	Ошибка более 2 минут	Ошибка более 5 минут
Исходное ETA	82,082	29,95	18,12	3,7

ML для прогноза времени подачи такси в Яндекс Такси

В чем проблема?

- До назначения машины вокруг вас много других пользователей заказывают такси - это влияет на поведение водителей
- Мы точно не знаем какая машина будет назначена - от этого зависит время в пути до клиента
- Водитель может задержаться в пути по разным причинам

Очень шумные данные, поэтому прогноз получается плохой!

ML для прогноза времени подачи такси в Яндекс Такси

Выход - замена целевой переменной:

предсказываем время в пути после того, как водитель назначен!

	Mean Absolute Error	Ошибка более 1 минуты	Ошибка более 2 минут	Ошибка более 5 минут
Исходное ETA	82,082	29,95	18,12	3,7
Текущая модель	79,276 (–3,4)	29,33	16,98	3

ML для прогноза времени подачи такси в Яндекс Такси

Прирост на 1% для бизнеса - это много денег!

	Mean Absolute Error	Ошибка более 1 минуты	Ошибка более 2 минут	Ошибка более 5 минут
Исходное ETA	82,082	29,95	18,12	3,7
Текущая модель	79,276 (–3,4)	29,33	16,98	3