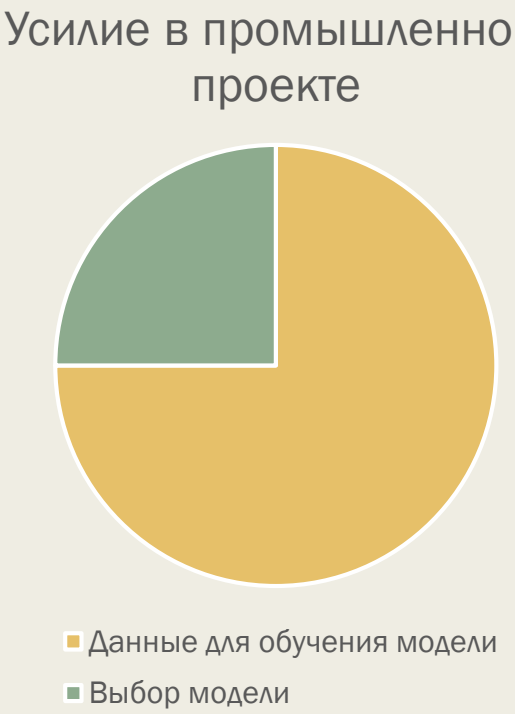
TGPO Consult

БОЛЬШИЕ ДАННЫЕ ИЛИ БОЛЬШОЙ ШУМ?

Дженнифер Трелевич
Исполнительный директор, TGPO Consult
d.trelevich@tgpo.ru

Почему даже «опытные» DS ошибаются в планировании проектов?[1]



[1] Andrey Karpathy на https://www.figure-eight.com/train-ai/

Ошибка №1 – сэкономить на категоризацию данных

Jetpac[2]: как Yelp, с автоматической сборкой данных

- Наняли недорогую рабочую силу из южной Азии на категоризацию фоток отпусков
- Они выбрали фотки конференц-залов, а не пляжей, поскольку им пляж подсказывает о труде...

TensorFlow[2]: данные для распознавания речи

- Поправка категоризации записей улучшила точность алгоритма с 85% до 90% *без изменения данных или моделей.*

Категоризация данных для обучения моделей является критической задачей.

- Скучный и утомительный труд, но...
- Владелец продукта и инженер должны участвовать в процессе
- Стоп-сценарии, проверка предположений и use-cases

[2] *https://petewarden.com/2018/05/28/why-you-need-to-improve-your-training-data-and-how-to-do-it/*

Ошибка №2 – 1-10-100

Проблема в клиентских данных в одной крупной российской компании из-за колл-центра:

- Чтобы удалить персональные данные, операторы изменят ФИО.
- Чтобы быстро закончить звонок, введут «нулевые данные» в запись.
- Нет мотивации исправить ошибки в данных в течение разговора.

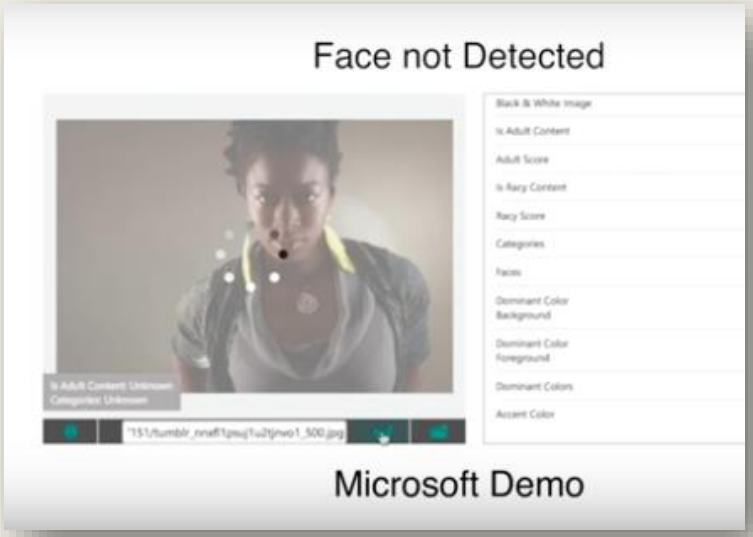
Эмпирическое правило[3]:

- Если стоимость поправки ошибок оператором – 1 ₽,
- То стоимость поправки в базе данных – 10 ₽,
- И стоимость поправки после последующей генерации отчетов – > 100 ₽.

Поправьте бизнес-процессы, генерирующие данные.

[3] https://economia.icaew.com/opinion/july-2013/the-unseen-cost-of-bad-data

Ошибка №3 – обучающие данные плохо изображают use-cases



4 коммерческих системы для классификации лиц (IBM, Microsoft, Face++, Kairos)[4]

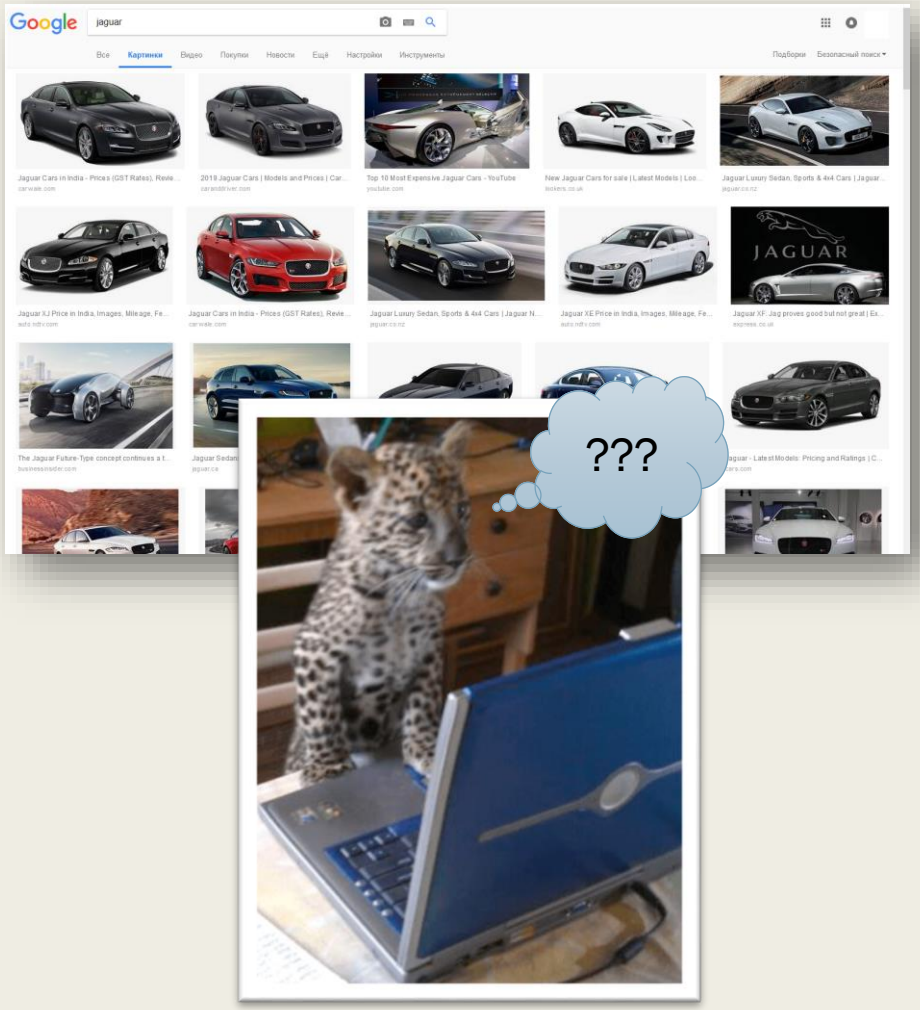
- Ошибались до 50% по лицам женщин и разных рас
  - *Либо не нашли лица*
  - *Либо не распознавали пол*
- Плохо, если такие демографические группы среди ваших клиентов…

Другой пример – Watson for Ontology

Правильный **подбор обучающих данных** определяет успех продукта.

[4] J. Buolamwini и T. Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, *Proceedings of Machine Learning Research* 81:1–15, 2018.

Ошибка №4 – несовместимость в категоризации



Крупная российская промышленная корпорация

- Каждый филиал имел свою базу данных
  - *Свои инвентарные номера*
  - *Свою структуру характеристик*
- Миграция «зоопарка» не предусмотрена в ТЗ
  - *Проект BI бесконечно задерживается*

До проекта: чистка зоопарка

В течение: ручная проверка кластеризации

Выгодное применение на российском рынке

ООО «Политех-Плюс»[5]

- 40% поломок на производстве – задержки и потерь заказов
- Предиктивное обслуживание
  - Собрали исторические данные
  - Инженеры вводят параметры о работе станок
- Общую эффективность оборудования с 60 до 75–79%

[5] https://www.gd.ru/articles/9073-mashinnoe-obuchenie
[6] https://spark.ru/startup/digital-contact/blog/38210/primenenie-mashinnogo-intellekta-v-rossijskom-biznese

Выгодное применение на российском рынке

ООО «Политех-Плюс»[5]

- 40% поломок на производстве – задержки и потерь заказов
- Предиктивное обслуживание
  - Собрали исторические данные
  - Инженеры вводят параметры о работе станок
- Общую эффективность оборудования с 60 до 75–79%

банк «Уралсиб» [6]

- Минимизировать риски, предотвращать мошенничество, проверять заемщиков, оценивать их платежеспособность и делать прогнозирование более точным
- Выдали в 2017 г.у в 2,9 раза больше кредитов физлицам, чем в 2016 г.
- Качество выдач при этом возросло.

[5] https://www.gd.ru/articles/9073-mashinnoe-obuchenie
[6] https://spark.ru/startup/digital-contact/blog/38210/primenenie-mashinnogo-intellekta-v-rossijskom-biznese

Выгодное применение на российском рынке

ООО «Политех-Плюс»[5]

- 40% поломок на производстве – задержки и потерь заказов
- Предиктивное обслуживание
  - Собрали исторические данные
  - Инженеры вводят параметры о работе станок
- Общую эффективность оборудования с 60 до 75–79%

банк «Уралсиб» [6]

- Минимизировать риски, предотвращать мошенничество, проверять заемщиков, оценивать их платежеспособность и делать прогнозирование более точным
- Выдали в 2017 г.у в 2,9 раза больше кредитов физлицам, чем в 2016 г.
- Качество выдач при этом возросло.

ООО «ИТ Сервис» (Digital Contact)[6]

- Персонализированные email-рассылки
  - Кому отправить письмо
  - Подбирает тему и предложение для каждого получателя отдельно
- Открытие писем повышается на 20-25% по сравнению с обычными рассылками
- Один из российских банков
  - Увеличил кликов с рассылок в 5 раз
  - Конверсию в заявки – в 3 раза

[5] https://www.gd.ru/articles/9073-mashinnoe-obuchenie
[6] https://spark.ru/startup/digital-contact/blog/38210/primenenie-mashinnogo-intellekta-v-rossijskom-biznese

Выгодное применение на российском рынке

ООО «Политех-Плюс»[5]

- 40% поломок на производстве – задержки и потерь заказов
- Предиктивное обслуживание
  - Собрали исторические данные
  - Инженеры вводят параметры о работе станок
- Общую эффективность оборудования с 60 до 75–79%

банк «Уралсиб» [6]

- Минимизировать риски, предотвращать мошенничество, проверять заемщиков, оценивать их платежеспособность и делать прогнозирование более точным
- Выдали в 2017 г.у в 2,9 раза больше кредитов физлицам, чем в 2016 г.
- Качество выдач при этом возросло.

ООО «ИТ Сервис» (Digital Contact)[6]

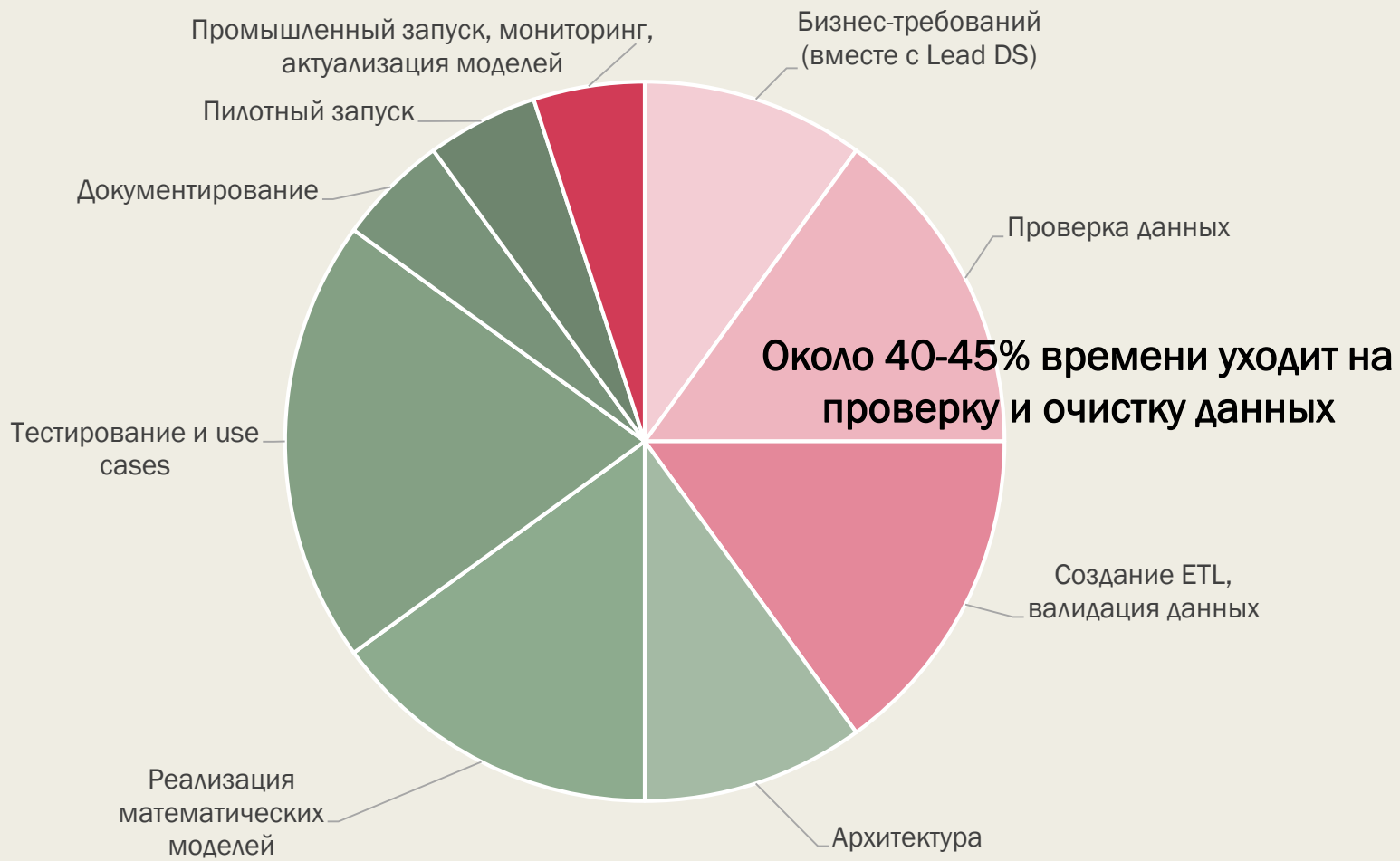
- Персонализированные email-рассылки
  - Кому отправить письмо
  - Подбирает тему и предложение для каждого получателя отдельно
- Открытие писем повышается на 20-25% по сравнению с обычными рассылками
- Один из российских банков
  - Увеличил кликов с рассылок в 5 раз
  - Конверсию в заявки – в 3 раза

Superjob.ru[5]

- Автоматическое предсказание з\п по резюме
- Объединение похожих вакансий
- Снизили нагрузку на сотрудников без расширения штата

[5] https://www.gd.ru/articles/9073-mashinnoe-obuchenie
[6] https://spark.ru/startup/digital-contact/blog/38210/primenenie-mashinnogo-intellekta-v-rossijskom-biznese

Где мы ожидаем тратить времени проекта vs. в самом деле...



Дженнифер Трелевич

https://TGPO.ru

d.trelevich@tgpo.ru

Спасибо за внимание!