

# Машинное обучение: предпроектное исследование

MADE academy  
Эмели Драль

# Прикладное машинное обучение

1. Предпроектное исследование: от постановки задачи до оценки потенциального эффекта
2. Проектная работа: оптимизация и валидация модели, демо-стенд, разработка сервиса
3. Поддержка и сопровождение сервиса. Чек-лист data-саентиста.

# Проектная работа

В индустрии стандартный формат работы – проектная работа.

Project - an individual or collaborative enterprise that is carefully planned to achieve a particular aim.

# Проектная работа

Основные отличия от других форм работы:

- Бизнес цель
- Формализованная задача
- Метрика качества
- Критерий успеха
- Проектная команда
- Сроки (и другие ресурсы)

# Проектная работа

Методики работы и управления проектом могут отличаться:

- Waterfall
- Agile (+ Scrum, Kanban)

Также существует специализированный стандарт для работы с проектами по анализу данных:

- CRISP DM (Cross-industry standard process for data mining)

мы не будем изучать методики проектной работы, а сфокусируемся на содержании проектов

# Этапы работы над проектом

- Постановка задачи
- Определение метрик и критериев успеха
- Оценка доступных данных
- Обучение моделей
- Тестирование моделей (эксперимент)
- Разработка сервиса
- Тестирование качества работы сервиса
- Мониторинг и поддержка качества сервиса, регулярное дообучение модели

# Этапы работы над проектом

- Постановка задачи
- Определение метрик и критериев успеха
- Оценка доступных данных
- Обучение моделей
- Тестирование моделей (эксперимент)
- Разработка сервиса
- Тестирование качества работы сервиса
- Мониторинг и поддержка качества сервиса, регулярное дообучение модели

# Проектная работа

Весь объем работы можно разделить на **три** стадии:

- **Предпроектное исследование**
- Работа над проектом
- Работа после окончания проекта

# Предпроектное исследование

1. Постановка задачи
2. Изучение предметной области
3. Обмен экспертизой
4. Определение метрик и критериев успеха
5. Оценка экономического потенциала

# Постановка задачи

# Активация пользователей

Проблема: после регистрации пользователи не активируются (не делают целевых действий), а значит не приносят пользу сервису

## Постановка задачи

Наблюдение: после того, как пользователь первый раз совершает целевое действие, он становится активным

# Активация пользователей

Проблема: после регистрации пользователи не активируются (не делают целевых действий), а значит не приносят пользу сервису

## Постановка задачи

Наблюдение: после того, как пользователь первый раз совершает целевое действие, он становится активным

Текущее решение: маркетинговая рассылка на всех пользователей

Идея оптимизации: с помощью модели предсказать, кто именно из пользователей активируется, и делать рассылку для них

## Постановка задачи

# Активация пользователей

Текущее решение: маркетинговая рассылка на всех пользователей

Идея оптимизации: с помощью модели предсказать, кто именно из пользователей активируется, и делать рассылку для них

Критика: email рассылка бесплатная, выгоднее всего попытаться активировать всех пользователей

Вывод: **для постановки задачи нужна как бизнес (или продуктовая), так и техническая экспертиза**

# Два взгляда на задачу

Постановка  
задачи

Business goal



Math problem statement



# Постановка задачи

## Два взгляда на задачу

### Бизнес-задача

- Сформулированная бизнес-цель
- Требует экспертных знаний в предметной области
- Обычно хорошо измеряется в деньгах

### Математическая задача

- Формальная постановка задачи в терминах анализа данных
- Требует экспертных знаний в математике
- Обычно хорошо измеряется в числах (точность, полнота, аккуратность)

# Конфликт постановок

Мы измеряем успех с точки зрения бизнес-цели  
А задачу оптимизации решаем математически

Постановка  
задачи

Поэтому, оптимизируя некоторые показатели качества,  
мы надеемся оптимизировать KPI бизнеса.

# Контрольный вопрос



Соответствует ли математическая постановка  
бизнес-задаче?

## Постановка задачи

# 3 шага для постановки задачи

В зависимости от специфики предметной области и имеющейся экспертизы, процесс постановки задачи может быть проще или сложнее.

## Постановка задачи

Рекомендую пройти следующие шаги:

1. Разобраться в предметной области
2. Обменяться экспертизой по анализу данных
3. Собрать потенциальные постановки задач для последующей валидации

Изучение предметной  
области

# Изучение предметной области

# Мотивация

Риски, связанные с отсутствием экспертизы в области решаемой задачи:

- коммуникация с продуктовой командой и пользователям
  - постановка математической задачи
  - ограничения и требования к решению
  - формулировка запроса на данные
  - валидация данных
  - feature engineering
  - дизайн сервиса и интеграция
  - дизайн эксперимента
- можно продолжать бесконечно =)

# Что делать?

Какие шаги можно предпринять, для того чтобы погрузиться в предметную область задачи?

Изучение  
предметной  
области

# Изучение предметной области

## Что делать?

Какие шаги можно предпринять, для того чтобы погрузиться в предметную область задачи?

Особенно, если задача из очень специфичной предметной области:

- Автоматизация подбора скважин для проведения гидроразрыва пласта
- Прогнозирование снижения производительности оборудования при производстве металла
- Оптимизация расхода ферросплавов и легирующих элементов при производстве стали

# Кандидаты на проведение гидроразрыва



# Кандидаты на проведение гидроразрыва



## Бизнес-цель:

Выбрать топ скважин для проведения гидроразрыва, для которого суммарная добыча нефти будет максимальной

## Математическая постановка:

- Спрогнозировать дебит нефти после гидроразрыва пласта для каждой скважины
- Отобрать в топ максимальные



# Кандидаты на проведение гидроразрыва

Дебиты нефти не являются независимыми!

- Если рвать соседние скважины по отдельности, дебит для каждой из них может быть большим
- Если рвать соседние скважины одновременно, дебит в некотором соотношении поделится между ними

# Кандидаты на проведение гидроразрыва

- Отбор кандидатов на основании одиночных прогнозов дебита не оптимален
- Значит, постановка задачи неправильная

Надо выбирать постановку, которая позволит получить оптимальный топ.

# Предсказание снижения производительности

## Задача

- Предсказать, для каких электролизеров будет снижена выливка алюминия

## Вопросы

- Что такое «сниженная выливка»?
- Какие действия можно предпринять на основе прогноза?

# Снижение расхода ферросплавов

## Задача

- Сократить расход ферросплавов в кислородно-конвертерном цехе

## Вопросы

- Не переносим ли мы потери на следующий этап?
- Действительно ли мы оптимизируем стоимость?

# Изучение предметной области

## Консультация с экспертом

- Экспертиза внутри команды
- Внешние эксперты
- Экспертиза заказчика (актуально, даже если заказчик внутренний)

# Изучение предметной области

## Экспертиза внутри команды

- Если у вас есть экспертиза в предметной области (образование, предыдущий опыт работы), скажите об этом и проведите встречу с проектной командой
- Если в команде есть эксперт, попросите провести воркшоп
- При наличии документации или отчетов о предыдущих проектах в данной области, их стоит прочесть =)

# Изучение предметной области

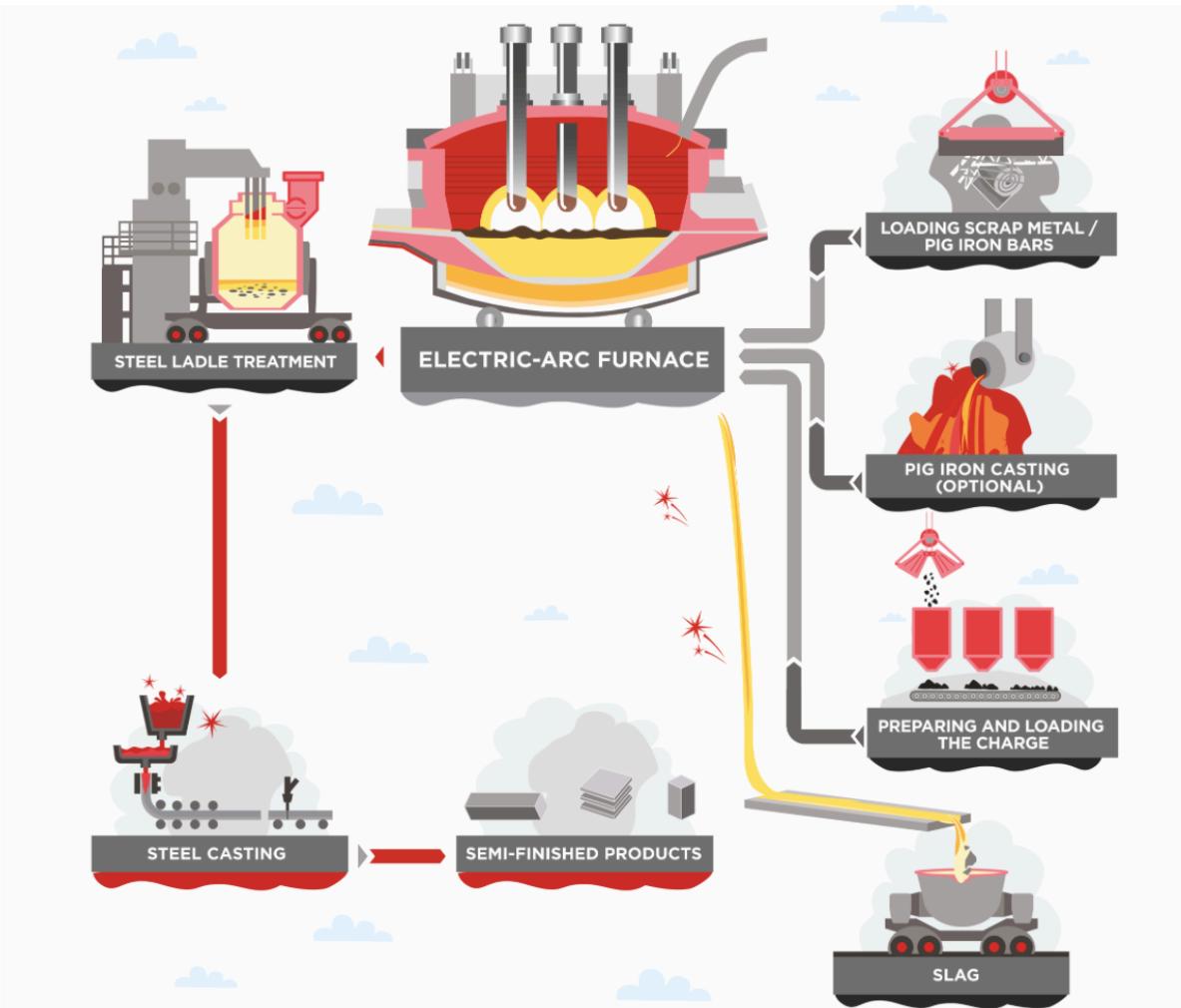
## Внешняя экспертиза

- Поиските информацию в поисковой системе, посмотрите обучающие видео
- Попросите организовать воркшоп с внешним консультантом

# Изучение предметной области

# Внешняя экспертиза

Пример: производство стали в электродуговой печи



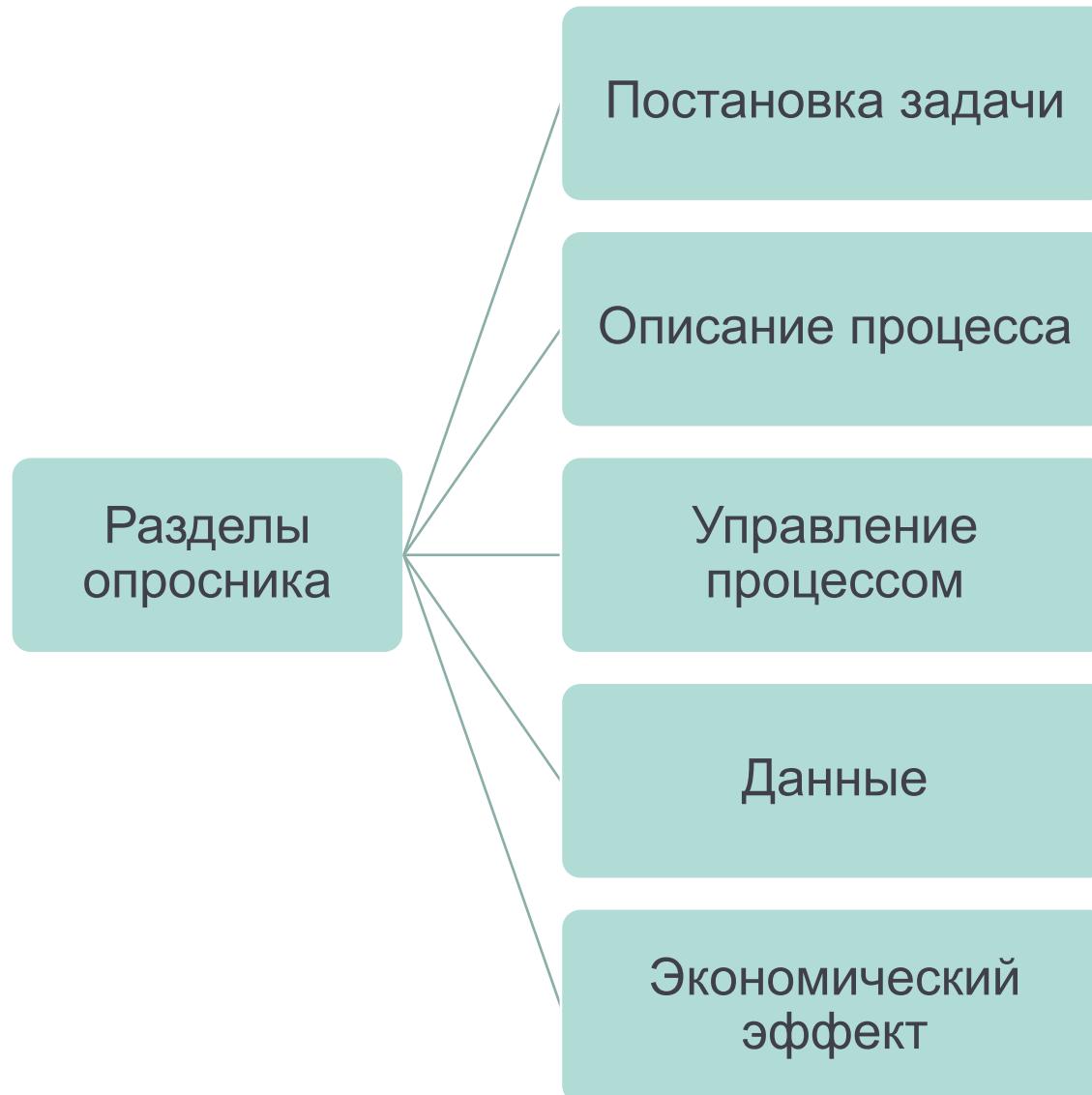
# Изучение предметной области

## Экспертиза заказчика

- Попросите поделиться с вами публично доступной информацией
- Предложите воркшоп с внутренними специалистами заказчика
- Создайте список вопросов и попросите подготовить ответы в письменном виде

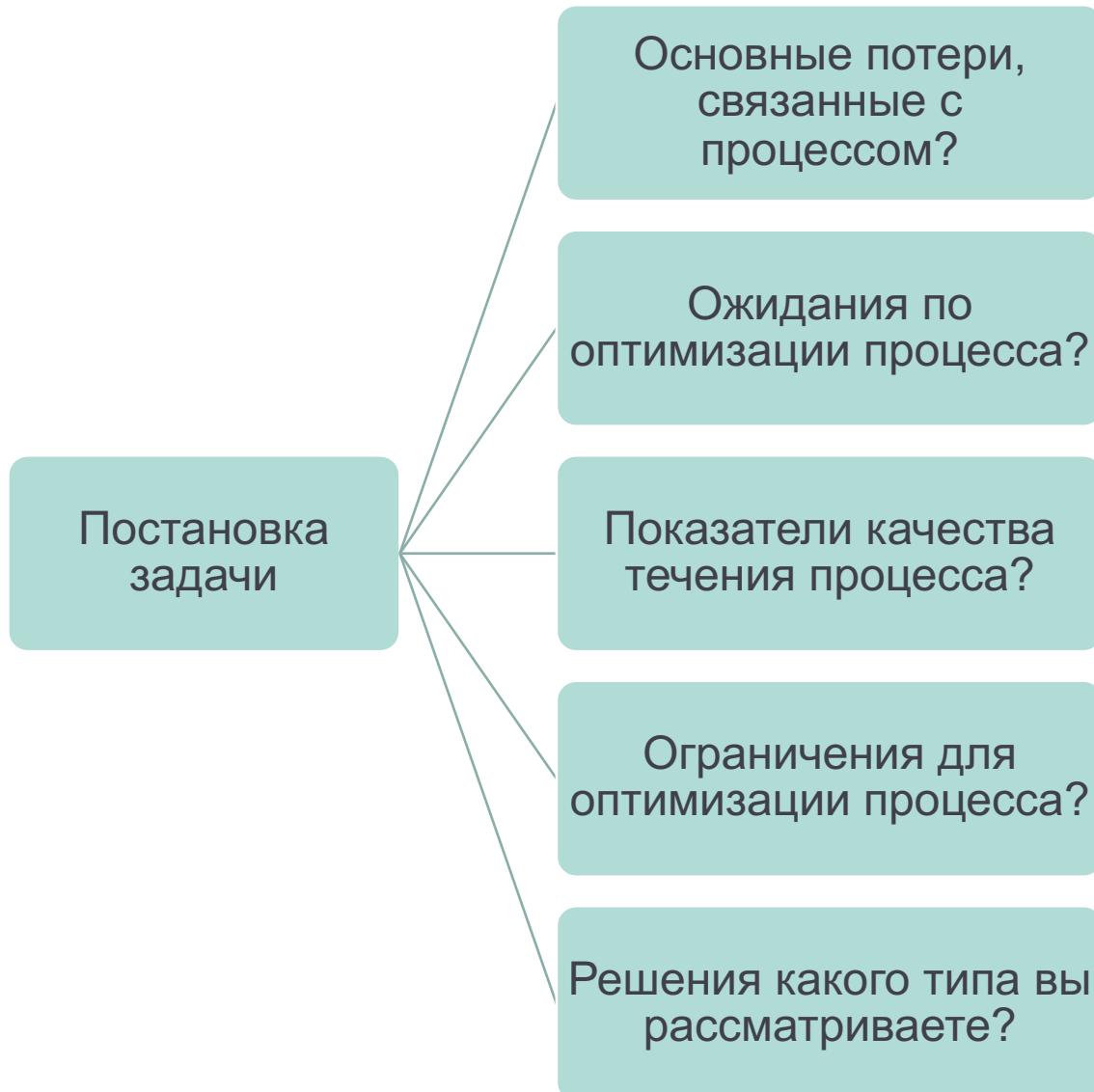
# Изучение предметной области

## Список вопросов



# Изучение предметной области

## Список вопросов



# Изучение предметной области

## Список вопросов



Вопросов может быть гораздо больше;  
Важно **кастомизировать** вопросы для каждого проекта.

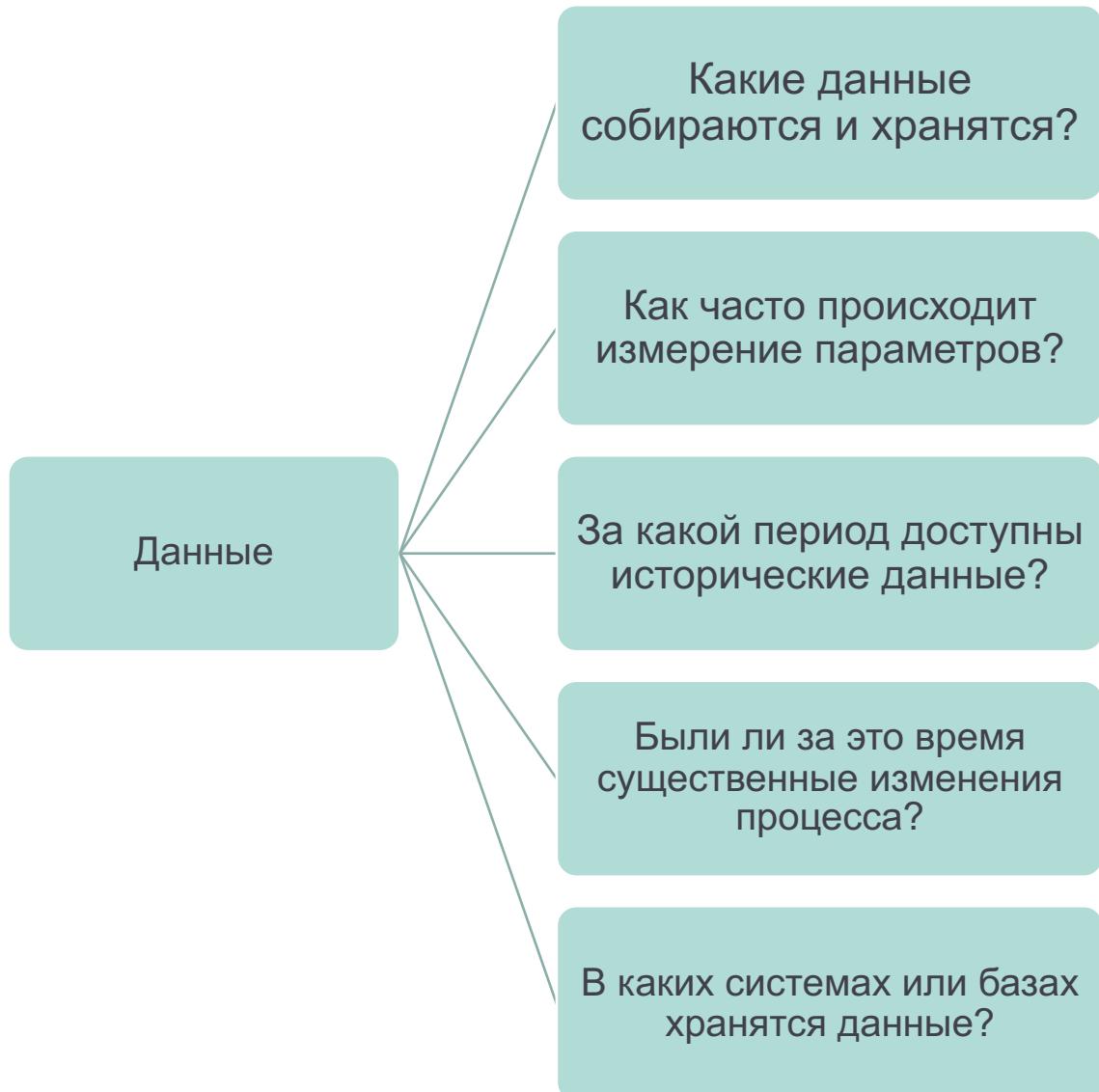
# Изучение предметной области

## Список вопросов



# Изучение предметной области

## Список вопросов



Здорово, если вы можете предложить список потенциально полезных источников данных:

- данные о сырье
- данные производственных датчиков
- лабораторные данные и пр.

# Изучение предметной области

## Список вопросов

Экономический эффект

Можете ли вы оценить возможный экономический эффект за месяц?

Можете ли вы оценить возможный эффект, например, при улучшении прогноза на 1%?

Потенциальный эффект крайне важен, но аккуратно его оценить сложно. По этой причине наряду с самой оценкой стоит обсудить методику её расчета.

# Изучение предметной области

## Подготовка опросника

Заполнение опросника – большая инвестиция времени со стороны заказчика, поэтому имеет смысл:

- Подготовить хорошо структурированный документ
- Кратко объяснить, как эта информация повлияет на ход проекта
- После получения ответов внимательно проанализировать их, не задавать те же вопросы повторно

# Обмен экспертизой

# Обмен Экспертизой

## Мотивация

Риски, связанные с отсутствием экспертизы в области анализа данных у заказчика:

- коммуникация с технической командой
- нереалистичные ожидания от решения
- ограничения и требования к решению
- предоставление релевантных данных

Тоже можно продолжать очень долго =)

# Обмен экспертизой

## Что делать?

1. Важно убедиться, что в вашей команде есть достаточная экспертиза. Если требуется – делитесь экспертизой со своей (технической) командой
2. Поделитесь экспертизой с заказчиком, если это требуется:
  - Поделитесь адаптированными материалами
  - Покажите релевантные примеры завершенных проектов
  - Можно провести встречу и рассказать про технологии (с адекватным уровнем детализации)
  - Часто, лучше всего работает демо сервиса

# Определение метрик и критериев успеха

# Метрики и критерии успеха



# Метрики и критерии успеха

## Метрики качества

Метрик качества очень много, одно и то же решение можно оценить сразу несколькими.

Например, метрики для качества прогноза:

- MAE
- MSE
- RMSE
- MAPE
- WAPE
- SMAPE

и пр.

# Метрики и критерии успеха

## Метрики качества

Важно, чтобы метрики успеха:

- Были адекватны математической постановке задачи
- Отвечали потребностям бизнеса
- Были зафиксированы до начала разработки и тестирования
- Не пересматривались в процессе или после тестирования (особенно актуально для пилотов, АБ-тестов)

# Виртуальные анализаторы

Виртуальное измерение  
состава входного газа при  
фракционировании

- Состав входного газа меняется
- Значения важны для оптимального управления процессом
- Задача: в реальном времени оценивать состав



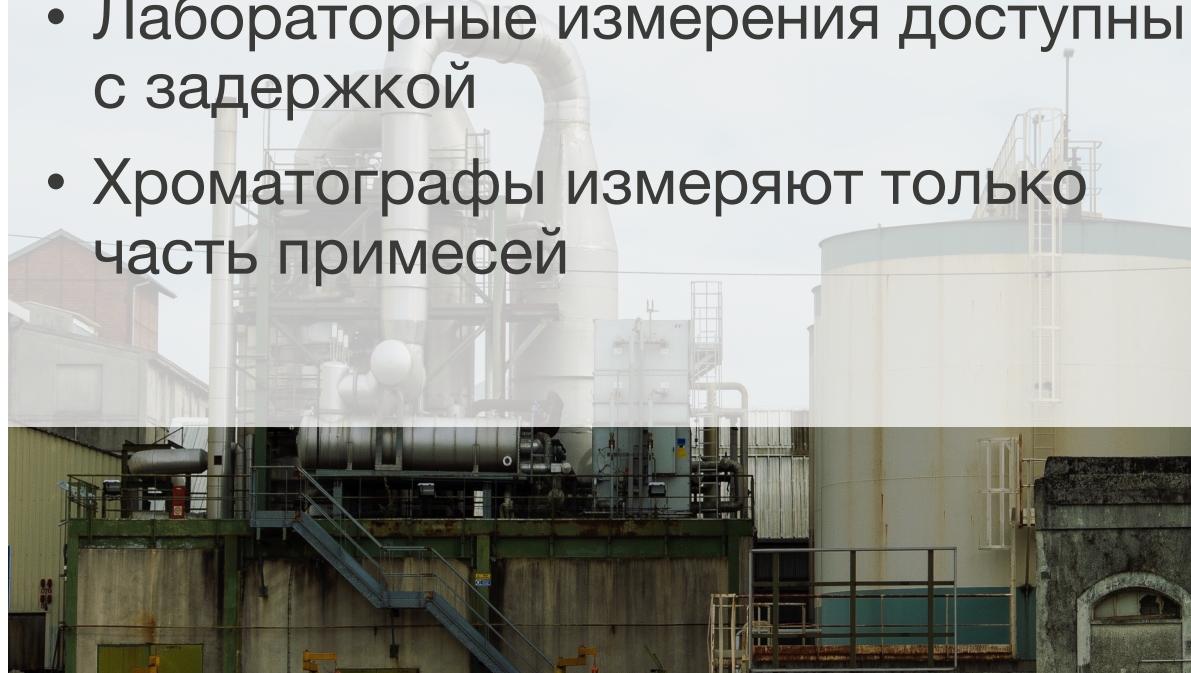
# Виртуальные анализаторы

Виртуальное измерение  
состава входного газа при  
фракционировании

- Состав входного газа меняется
- Значения важны для оптимального управления процессом
- Задача: в реальном времени оценивать состав

Специфика:

- Есть точные термодинамические модели, но для них нужны все данные
- Лабораторные измерения доступны с задержкой
- Хроматографы измеряют только часть примесей



# Виртуальные анализаторы

Как измерить качество  
виртуального измерения?



# Виртуальные анализаторы

Как измерить качество виртуального измерения?

**Оцениваем качество ВА как отклонение показаний от хроматографов**

- Показания хроматографов доступны с меньшей частотой
- Хроматографы измеряют не все интересующие показатели
- Хроматографы врут



# Виртуальные анализаторы

Как измерить качество виртуального измерения?

- Вместо непосредственно «точности» ВА надо оценивать то, насколько лучше мы управляем колонной
- Например, эффективность разделения газа

Надо оценивать конечный, а не промежуточный шаг.



# Оценка экономического потенциала

# Оценка эффекта

## Экономический эффект

Решение математической задачи оценивается с помощью метрик качества модели:

- precision/recall
- ROC AUC
- MAE, MSE, logloss

Решение бизнес задачи оценивается по размеру экономического эффекта:

- ???

## Оценка эффекта

# Экономический эффект

Решение математической задачи оценивается с помощью метрик качества модели:

- precision/recall
- ROC AUC
- MAE, MSE, logloss

Решение бизнес задачи оценивается по размеру экономического эффекта:

- К сожалению стандартных формул для такой оценки нет
- Но есть подход, который можно пробовать обобщать для конкретных задач

# Рассмотрим пример

Решается задача снижения оттока для онлайн сервиса

Отток: отказ пользователя от продукта или услуги

Оценка  
эффекта



## Оценка эффекта

# Отток и удержание

Решается задача снижения оттока для онлайн сервиса

- Больше пользователей -> больше прибыли
- Удерживать всех пользователей дорого -> **адресное** удержание
- Удержание пользователей происходит не мгновенно  
-> прогноз с **солидным горизонтом**

# Давайте оценим эффект



- Как бы вы подошли к оценке потенциального эффекта?
- Какая информация еще потребуется?

Оценка  
эффекта

# Оценка эффекта

## Давайте оценим эффект



- Как бы вы подошли к оценке потенциального эффекта?

Допустим, что

- Мощность кампании - **N пользователей**, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого **С денег**
- **p – доля оттока** среди удерживаемых
- **ARPU** – сколько в среднем нам приносит клиент

# Оценка эффекта

## Давайте оценим эффект

### Задача

- Пусть мы удерживаем **N пользователей**, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого **C денег**
- **p – доля оттока** среди удерживаемых
- **ARPU** – сколько в среднем нам приносит клиент

### Вопросы:

- Как оценить потенциальный экономический эффект?
- Как этот эффект оптимизировать?

# Оценка эффекта

## Давайте оценим эффект

### Задача

- Пусть мы удерживаем **N пользователей**, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого **C денег**
- **p – доля оттока** среди удерживаемых
- **Nr – количество удержанных пользователей**, если удерживаем со 100% успехом

# Оценка эффекта

## Давайте оценим эффект

### Задача

- Пусть мы удерживаем **N пользователей**, наиболее вероятно уходящих в отток по прогнозу нашей модели
- Тратим на удержание каждого **С денег**
- **r – доля оттока** среди удерживаемых
- **Nр** – количество удержанных пользователей, если удерживаем со 100% успехом
- **Npr** – количество удержанных пользователей, если **удержание успешно с вероятностью r**

# Давайте оценим эффект

Экономический эффект: ARPU \* N\*p\*r – C\*N

Оптимизация эффекта: максимизация р

Оценка  
эффекта

# Давайте оценим эффект

Экономический эффект: ARPU \* N\*p\*r – C\*N

Оптимизация эффекта: максимизация р

Оценка  
эффекта

Корректна ли данная оценка эффекта?

## Оценка эффекта

# Давайте оценим эффект

Экономический эффект: ARPU \* N\*p\*r – C\*N

- На какое время мы удерживаем клиента?
- Снизили ли мы ARPU клиентов, которых ошибочно приняли за отток?

# Давайте оценим эффект

## Задача

Как измениться оценка, если мы удерживаем пользователей на M месяцев снижая в процессе ARPU на X процентов?

Оценка  
эффекта

# Давайте оценим эффект

## Задача

Как измениться оценка, если мы удерживаем пользователей на M месяцев снижая в процессе ARPU на X процентов?

## Оценка эффекта

$$\text{ARPU} * N * p * r - C * N$$

vs

$$N * p * r * \text{ARPU}(m) * M * (1 - X) - C * N$$

# Оценка эффекта

## Оценка эффекта

- Способ оценки потенциального экономического эффекта строится отдельно для каждой задачи
- Не смотря на отсутствие общего подхода, можно выделить общие принципы построения такой оценки
- Существует trade-off между точностью оценки и сложностью её получения
- Часто эффективнее взять менее точную оценку и пессимизировать её
- На базе оценки, например,  $ARPU * N*p*r - C*N$ , можно получить ограничения на качество модели, в данном случае,  $p$  - доля оттока

# Машинное обучение: предпроектное исследование

Спасибо!  
Эмели Драль