

# Инжиниринг признаков и выбор моделей

Наталья Баданина

Методист академических программ ML направления, Яндекс

Data Scientist



# Проверка связи





## Если у вас нет звука:

- убедитесь, что на вашем устройстве и колонках включён звук
- обновите страницу вебинара или закройте её и заново присоединитесь к вебинару
- откройте вебинар в другом браузере
- перезагрузите устройство и попробуйте зайти заново



## Поставьте в чат:

-  если меня видно и слышно
-  если нет

# Рекомендации

## → При просмотре с компьютера

- Используйте браузеры **Google Chrome** или **Microsoft Edge**
- Если есть проблемы с изображением или звуком, обновите страницу — **F5**

## → При просмотре с мобильного телефона или планшета

- Перейдите с мобильного интернет-соединения на **Wi-Fi**
- Если есть проблемы с изображением или звуком, перезапустите приложение на телефоне

# Правила участия

- 1 Приготовьте блокнот и ручку, чтобы записывать важные мысли и идеи
- 2 Продолжительность вебинара — 90 минут
- 3 Вы можете писать свои вопросы в чате
- 4 Запись вебинара будет доступна в личном кабинете



# Наталья Баданина

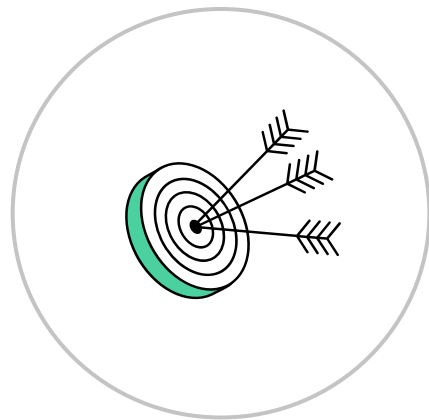
О спикере:

- Методист академических программ ML направления, Яндекс
- Data Scientist



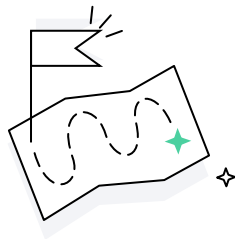
# Цели занятия

- 1 Подвести итоги Спринта 1
- 2 Рассмотреть подходы к трансформации данных
- 3 Изучить базовые методы инжиниринга признаков
- 4 Освоить методы отбора признаков, основанные на дисперсии и на результатах обучения модели классификации, и построить модель классификации на данных о недвижимости



# План занятия

- 1 Спринт 1, Итоги
- 2 Трансформация данных (Data Transformation)
- 3 Инжиниринг признаков
- 4 Практика по работе с признаками и алгоритмами по отбору признаков



# Итоги Спринта 1





# Спринт 1. Введение в Проектный практикум

Задачи Части 1:

- Сформировать команду
- Распределить роли в команде
- Изучить описание учебного кейса
- Сформировать таймлайн работы над проектом



# Спринт 2. «Предварительная обработка данных»

## Задачи Части 2:

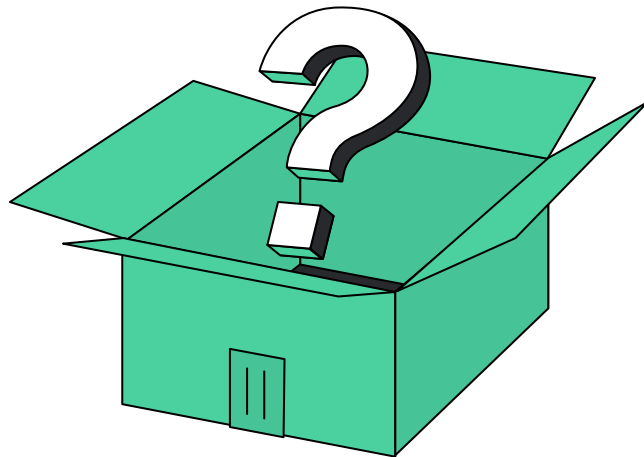
- Загрузить данные проекта с сайта кегл в среду разработки
- Провести предварительный анализ данных (без визуализации)
- Выявить пропуски в данных
- Принять решение по обработке найденных пропусков
- Выявить категориальные признаки
- Преобразовать категориальные данные
- Нормировать данные выбранным методом



# Подведение итогов

Ответьте на вопросы:

- Чего ваша команда достигла за текущий спринт?
- С какими трудностями вы столкнулись и как их преодолели?
- Что вы можете улучшить в следующем спринте?





**Ваши вопросы?**

# Трансформация данных (Data Transformation)



1

# Трансформация данных

1

Агрегирование  
(Aggregation)

2

Обобщение  
(Generalization)

3

Интеграция  
данных  
(Data Integration)

4

Нормировки  
(Data  
Normalization)

# Агрегирование (Aggregation)

Представляет собой процесс преобразования данных с высокой степенью детализации к более обобщенному представлению:

- составляющие суммы;
- замеры разными датчиками и т.п.

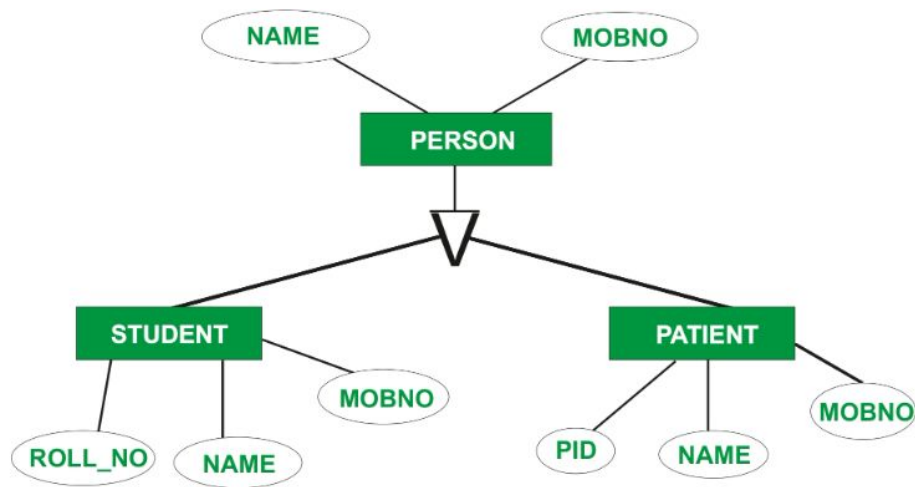
Совет: лучше использовать различные статистики

Лайфхак: отсортировать построчно показания

```
df['pr_mean'] = df[cols].mean(axis=1)
df['pr_std'] = df[cols].std(axis=1) #.round(2)
df['pr_max'] = df[cols].max(axis=1)
df['pr_min'] = df[cols].min(axis=1)
```

# Обобщение (Generalization)

Представляет собой создание описательных признаков





# Интеграция данных (Data Integration)

Представляет собой объединение данных, находящихся в различных источниках

```
df.merge(df2, how='left').merge(df3, how='left')
```

# Нормировки (Data Normalization)

Для большинства алгоритмов машинного обучения необходимо, чтобы все признаки были вещественными и «в одной шкале»:

- Стандартизация (Z-score Normalization / Variance Scaling)
- Нормировка на отрезок (Min-Max Normalization)
- Нормировка по максимуму
- Decimal Scaling Normalization
- Ранговая нормировка (tiedrank, rankdata)



**Ваши вопросы?**

# Инжиниринг признаков



2



Что такое инжиниринг  
признаков?

# Инжиниринг признаков

Инжиниринг (feature engineering / construction или генерация признаков) – процесс придумывания способов описания данных с помощью простых значений, отражающих характеристики объектов исследований, через которые могут выражаться целевые значения.

# Методы инжиниринга

1

Создание  
полиномиальных  
признаков

2

Отбор признаков  
на основе низкой  
вариации

3

Anova\*

\*ANOVA (дисперсионный анализ) — это статистический метод, который используется для сравнения средних значений двух или более выборок

# Важно понимать

Процесс создания признакового пространства зависит от модели, которую будем использовать:

- ОНЕ-кодирование\* предпочтительнее для линейных моделей;
- умное кодирование категорий – для деревьев;
- для робастной модели выбросы можно не удалять (и этапы предобработки данных тоже!)

Следует использовать:

- контекст (знание предметной области);
- EDA\*\*

\*ОНЕ-кодирование (One-Hot Encoding, быстрое кодирование) – процесс, с помощью которого категориальные переменные преобразуются в подходящую алгоритмам Машинного обучения (ML) форму

\*\*EDA (Exploratory Data Analysis, разведочный анализ данных) — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации



# Признаки (Features)

Признак – это функция на множестве объектов  $f : X \rightarrow S$

Признак: пол

Клиенты: {М, Ж}

Признак: доход

Клиенты: {..., 10 000, 20 000, ..., NA}

*Значения признака могут быть не определены – это тоже важная информация*

*Некоторые значения можно восстановить по другим признакам (например, пол)*

# Виды признаков

- Исходные (raw)
- Сгенерированные / производные (derived)

*Например: возраст = текущая дата – дата рождения*

Совет: даже если есть какой-то признак, сгенерируйте его по другим (пример с возрастом: есть дата рождения, текущая, возраст)

# Контекстные признаки

Это признаки, смысл которых явно прописан в постановке задачи или понятен из контекста.

Смысл определяет:

- Область значений
- Примерное распределение в этой области

# Практика



# Цель и задачи

## Цель:

1. Изучить практическую реализацию по работе с признаками
2. Реализовать алгоритмы по отбору признаков

## Задачи:

1. Понять, какие признаки можно сгенерировать
2. Реализовать методы по отбору признаков
3. Приступить к решению кейса по предсказанию популярности объявления о продаже дома



**Ваши вопросы?**

# Выводы

1. Трансформация данных предполагает агрегирование, обобщение, интеграцию и нормировку
2. К основным методам инжиниринга данных относятся создание полиномиальных признаков, отбор признаков на основе низкой вариации, Anova
3. Полезно генерировать новые признаки на основе уже существующих признаков и житейской смекалки
4. Среди всех признаков нужно проводить отбор и включать в модель наиболее полезные по какому-либо критерию

# Итоги занятия

- Рассмотрели подходы к трансформации данных
- Изучили базовые методы инжиниринга признаков
- Освоили методы отбора признаков, основанные на дисперсии и на результатах обучения модели классификации
- Приступили к построению модели классификации на данных о недвижимости





# Рефлексия

- Что изменилось? Раньше я думал(а), что..., а теперь...
- Какие вопросы у меня остались?



# Домашнее задание. ДЗ 2. Часть 1

1. Проведите встречу команды и распределите задачи по предварительной обработке данных для дальнейшего анализа, используя материалы Вебинара «Инжиниринг признаков и выбор моделей»
2. Реализуйте задачи предварительной обработки данных:
  - Сгенерировать новые признаки на основе существующих
  - Провести отбор сгенерированных признаков, а также тех, что уже были в датасете

Срок выполнения Части 1 и Части 2 задания — 7 дней с момента открытия задания

# Инжиниринг признаков и выбор моделей

Наталья Баданина

Методист академических программ ML направления, Яндекс

Data Scientist

