

Предварительная обработка данных

Наталья Баданина
Data Scientist, эксперт Нетологии



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 попытайтесь 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

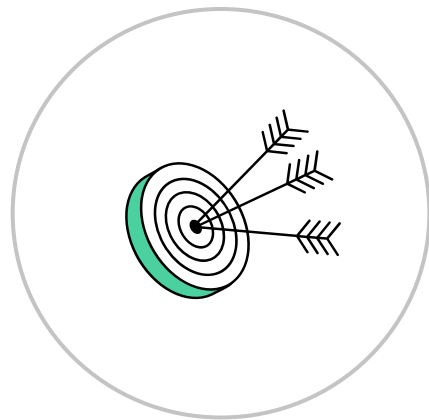
О спикере:

- Data Scientist в Чистой линии
- эксперт по ML и DS в Нетологии
- соавтор курсов по Data Science, R



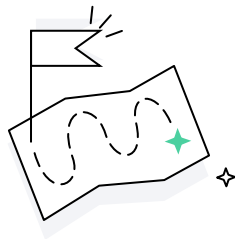
Цели занятия

- 1 Подвести итоги Спринта 1
- 2 Познакомиться с фундаментальными свойствами и видами данных
- 3 Рассмотреть основные этапы предварительной обработки данных
- 4 Понять подходы к улучшению свойств данных с помощью методов очистки и трансформации данных



План занятия

- 1 Итоги Спринта 1
- 2 Фундаментальные свойства и виды данных
- 3 Предобработка данных
- 4 Очистка данных (Data Cleaning)
- 5 Трансформация данных (Data Transformation)



Итоги Спринта 1



Спринт 1. Введение в Проектный практикум

Задачи:

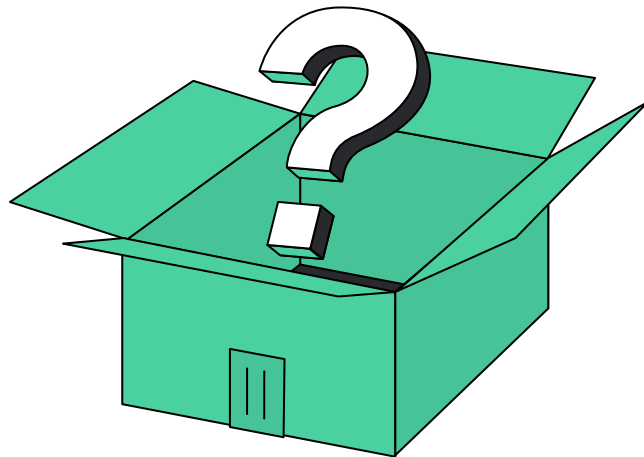
- Сформировать команду
- Распределить роли в команде
- Изучить описание учебного кейса
- Сформировать таймлайн работы над проектом



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?

Фундаментальные свойства и виды данных



1



Как вы думаете, что такое
«данные»?



В информатике и информационных технологиях данные — это:

- поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи или обработки (ISO/IEC 2382:2015);
- формы представления информации, с которыми имеют дело информационные системы и их пользователи (ISO/IEC 10746-2:1996).

Фундаментальные свойства данных

1. Доступность (Accessibility)
2. Актуальность (Timeliness)
3. Ценность (Value added)
4. Истинность (Believability)

Сбор данных (Data Collection)

На что смотреть:

- | | | |
|------------------|---------------------|--------------------|
| ✓ Размеры | ✓ Семантика данных, | ✓ Структура данных |
| ✓ Число | ✓ Идентификация | ✓ Режим доступа к |
| элементарных | отдельных | данным (online / |
| объектов | элементов и | offline) |
| ✓ Разреженность, | порций данных (id | ✓ Способ доступа |
| ✓ Ролнота | объектов, связи | ✓ Источник данных |
| | между таблицами и | (source) |
| | т.п.) | |

Виды данных

- Признаковые описания (матрица «объект-признак»)
- Измерения:
 - одномерные сигналы (ряды, звук и т.п.), последовательности, тексты;
 - изображения;
 - видео
- Метрические данные
- Данные в специальных форматах:
 - графы;
 - XML-файлы;
 - пространственно-временные данные;
 - «сырые» логи

Источники данных

1. Proprietary data sources — часто нельзя получить доступ
2. Government data sets — Data.gov
3. Academic data sets — используются при написании публикаций
4. Web search «Scraping» — есть лимиты и условия использования
5. Sensor data — относительно дешёвы, но специфичны

Ваши данные – самые ценные

Свойства данных

Свойство	Корректность
Что мешает этому свойству	Аномалии (выбросы + шум), «некорректности»
Причины нарушения свойства	Погрешность приборов, ошибки при заполнении
Средство борьбы	Очистка данных (Data Cleaning)

Свойства данных

Свойство	Полнота (Completeness)
Что мешает этому свойству	Пропуски, разреженность
Причины нарушения свойства	Недоступность данных, ошибки при заполнении, сбои при записи
Средство борьбы	Очистка данных (Data Cleaning)

Свойства данных

Свойство	Непротиворечивость (согласованность, Consistency)
Что мешает этому свойству	«Противоречия»
Причины нарушения свойства	Различные источники данных
Средство борьбы	Интеграция (Data Integration)

Свойства данных

Свойство	Безызбыточность
Что мешает этому свойству	<ul style="list-style-type: none">• Дубликаты• Шум• Излишняя дискретизация
Причины нарушения свойства	Особенности интеграции, ошибки при заполнении
Средство борьбы	<ul style="list-style-type: none">• Сокращение данных (Data Reduction)• Трансформация (Data Transformation)

Свойства данных

Свойство	Ясность (Interpretability)
Что мешает этому свойству	«Неясности»
Причины нарушения свойства	Плохие хранение и подготовка данных
Средство борьбы	Трансформация (Data Transformation)

Свойства данных

Свойство	Структурированность и однородность
Что мешает этому свойству	«Сырые» данные
Причины нарушения свойства	<ul style="list-style-type: none">• Нет признаков описаний• Признаки в разных шкалах
Средство борьбы	<ul style="list-style-type: none">• Генерация признаков (Feature engineering)• Трансформация (Data Transformation)



Ваши вопросы?

Предобработка данных



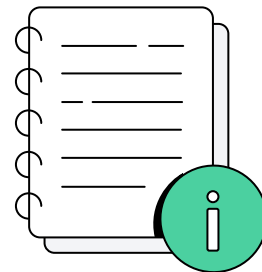
2



Предобработка данных (Data Preprocessing / Preparation) — это процесс преобразования данных в форму, удобную для анализа; замена, модификация или удаление частей набора данных с целью повышения непротиворечивости, полноты, корректности и ясности набора данных, а также уменьшения избыточности.

Предобработка данных

- Выполняется на полном наборе данных (в том числе и на контрольных объектах)
- Важно не допустить утечки информации, не доступной при функционировании модели



Предобработка данных

1

Очистка данных
(Data Cleaning)

2

Сокращение
данных
(Data Reduction)

3

Трансформация
данных (Data
Transformation)

4

Интеграция
данных
(Data Integration)

Очистка данных (Data Cleaning)

Обнаружение (и удаление / замена):

- аномалий / выбросов (Anomaly Detection);
- пропусков (Missing Data Imputation);
- шумов (Noise Identification);
- некорректных значений (Correct Bad Data / Filter Incorrect Data)

Сокращение данных (Data Reduction)

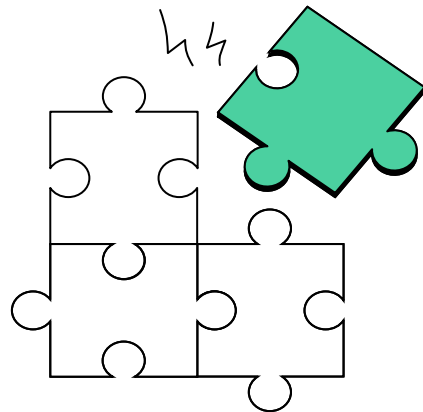
- Сэмплирование (Sampling)
- Сокращение размерности (Dimensionality reduction)
- Отбор признаков (Feature subset selection)
- Отбор объектов (Instance Selection)
- Удаление дубликатов

Трансформация данных (Data Transformation)

- Переименование признаков, объектов, значений признаков, преобразование типов
- Кодирование значений категориальных переменных
- Дискретизация (Discretization / Binning)
- Нормализация (Normalization)
- Сглаживание (Smoothing)
- Создание признаков (Feature creation)
- Агрегирование (Aggregation)
- Обобщение (Generalization)
- Деформация значений

Интеграция данных (Data Integration)

- Объединение данных из разных источников





Ваши вопросы?

Очистка данных (Data Cleaning)



3

Переименования

Названия переменных (и их значения) должны быть интуитивны (они используются в том числе при передачи данных коллегам, презентации результатов и т.п.)

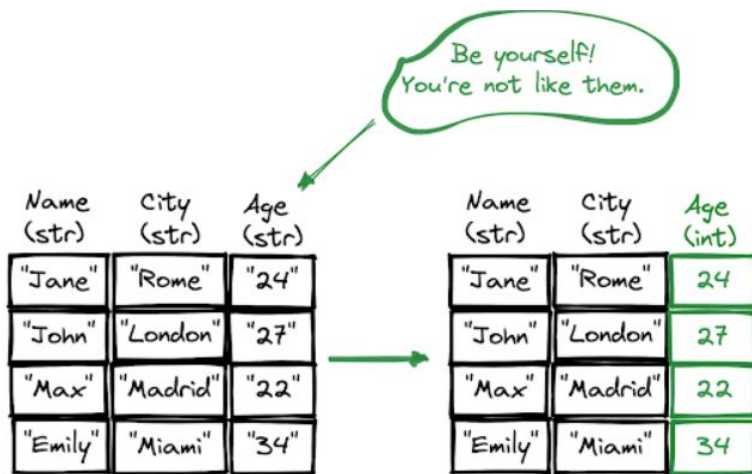
country	pop*
USA	322,179,605
China	1,403,500,365
Japan	127,748,513
India	1,324,171,354

`.rename()`

country	population*
USA	322,179,605
China	1,403,500,365
Japan	127,748,513
India	1,324,171,354

Преобразования типов данных

Нужно использовать типы, которые поддерживает ваша среда программирования



Кодировки

Как правило, компьютер работает с числами, поэтому категории представляем числами (векторами)

"red"	1
"yellow"	2
"red"	1
"blue"	0
"yellow"	2
"blue"	0
"red"	1

Пропуски

Пропуски – это отсутствующие значения в наборе данных, которые могут выглядеть как:

- специальные значения (NA, NaN, null, ...)
- специальный код (–999, mean, число за пределами значения признака)

direct bilirubin mg/dL	iron	oxygen saturation %	ferritin
0.8	50	16	20
22.1	?	?	?
1	?	?	?
?	?	?	?
0.3	?	?	?

[Источник](#)

Пропуски. Что делать?

- **Оставляем** *(но не все модели могут работать с пропусками)*
- **Удаляем** описания объектов с пропусками / признаки *(радикальная мера, которая редко используется)*

```
df.dropna(how='any', axis=1)
```

- **Заменяем** на фиксированное значение (например, если признак бинарный, то на 0.5). *Значение -999, как правило, «плохое», то есть является выбросом*

```
df.fillna(-1)
```

- **Заменяем** на легко вычисляемое значение (среднее, медиана, мода)

```
df.fillna(df.mean()) # , inplace=True
```


Пропуски. Что делать?

- **Восстанавливаем** значения *(требуется построение специальной модели для восстановления)*

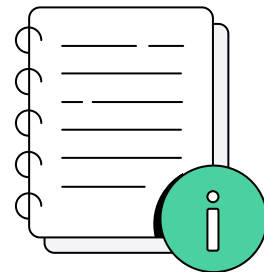
```
import numpy as np
from sklearn.impute import SimpleImputer

imp_mean = SimpleImputer(missing_values=np.nan, strategy='mean')
imp_mean.fit([[7, 2, 3], [4, np.nan, 6], [10, 5, 9]])
```

- **Производим** экспертную замену

Пропуски. Тонкости

- Если добавлять характеристический признак пропусков «is_nan», то тогда модель сама определит оптимальное значение для заполнения
- Заполнять пропуски лучше после генерации признаков.
иначе возникают дополнительные неопределённости



Пропуски. Тонкости

Важно понимать природу пропуска:

- значение может не быть доступно *(например, клиент банка не указал в анкете свой возраст);*
- значение может не существовать *(например, «Доход» для детей моложе 18 лет (=0));*
- значение не является числом *(например, $0/0 = NaN$);*
- значение вызвана предобработкой данных *(например, при конкатенации* таблиц – несуществующие колонки, при обработке дат – исключение);*
- можно посмотреть, зависит ли факт пропуска от других данных

***Конкатенация** (лат. *concatenatio* «присоединение цепями; сцепление») — операция склеивания объектов линейной структуры, обычно строк. Например, конкатенация слов «микро» и «мир» даст слово «микромир».



Ваши вопросы?

Трансформация данных (Data Transformation)



4

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.

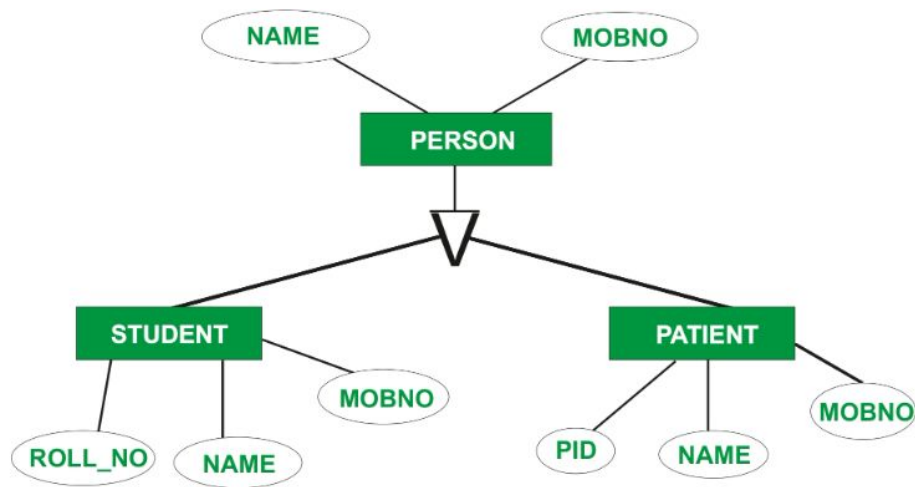
Совет: лучше использовать различные статистики

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)

Практика



Цель и задачи задания

Цель: изучить практическую реализацию по работе с датасетом

Задачи:

1. Понять как работать с пропусками в данных
2. Понять как работать с категориальными признаками
3. Понять как осуществлять полезные преобразования данных



Ваши вопросы?

Выводы

- Данные — это формы представления информации в различных видах: признаковые описания, измерения, метрические данные, данные в специальных форматах.
- К фундаментальным свойствам данных относятся доступность (Accessibility), актуальность (Timeliness), ценность (Value added), истинность (Believability).
- Предобработка данных (Data Preprocessing / Preparation) представляет собой процесс преобразования данных в форму, удобную для анализа, путём их очистки, сокращения, трансформации, интеграции.
- Очистка данных предполагает их переименование, преобразование, кодировку и работу с пропусками. Трансформация данных предполагает агрегирование, обобщение, интеграцию и нормировку.

Итоги занятия

- Подвели итоги Спринта 1
- Познакомились с фундаментальными свойствами и видами данных
- Рассмотрели основные этапы предварительной обработки данных
- Поняли подходы к улучшению свойств данных с помощью методов очистки и трансформации данных



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 «Предварительная обработка данных»
2. Реализуйте задачи предварительной обработки данных:
 - Загрузите данные проекта своей команды в среду разработки
 - Проведите предварительный анализ данных (без визуализации)
 - Выявите пропуски в данных
 - Примите решение по обработке найденных пропусков
 - Выявите категориальные признаки
 - Преобразуйте категориальные данные
 - Нормируйте данные выбранным методом

Срок выполнения Части 1 и Части 2 задания — 7 календарных дней

Предварительная обработка данных

Наталья Баданина
Data Scientist, эксперт Нетологии

