




Анализ данных

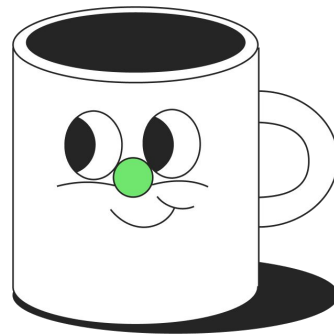
Урок о том, как работать с данными





Что будет на уроке сегодня

-  Обсудим этапы анализа данных
-  Рассмотрим часть этапов на примере
-  Ответы на вопросы





Анализ данных





Анализ данных

Это процесс обработки данных с целью получения полезной информации и её интерпретации.

- Перед тем как начать, нужно определить, какой результат вы хотите получить.
- Не всегда, чтобы получить удовлетворяющий вас результат, нужна ML модель.
- Данные могут содержать ложную информацию.
- Желательно разбираться в предметной области, в которой вы проводите анализ данных.





Этапы анализа данных





Данные



Какие они могут быть

Признак (feature) — это измеримая характеристика объекта.

- Номинальные — это признаки, которые можно сравнить на равенство. Например, статус здоровья сотрудника (больной или здоровый).
- Порядковые — признаки, в которых важен порядок. Мы можем сравнить их на «больше»/«меньше». Например, оценка в школе.
- Количественные — признаки, между которыми можно измерить разницу и выполнять все мат. операции. Например: дата, рост, напряжение, кол-во эритроцитов.

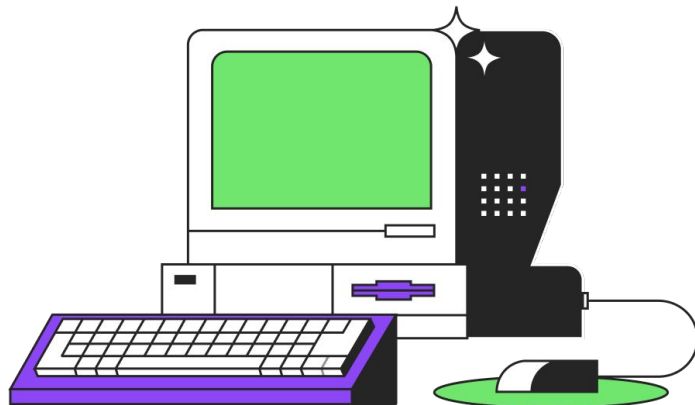


Со стороны компьютера

Все данные для вычислительной техники — числа, байты, биты. Будь это текст, персональные данные человека или картинка.

- Даты — число кол-ва секунд или микросекунд от опорной даты (unix time — от начала 1970 года).
- Текст — набор байт.
- Картинка — набор чисел, описывающих каждую точку, как комбинацию цветов.
- Звук — набор чисел, описывающих амплитуду.

Для корректного анализа номинальных и порядковых признаков рекомендуется переводить их в числа.





Сбор данных

Чтобы что-то анализировать, это что-то надо где-то достать.

- Использование открытых датасетов.
- Покупка готовых датасетов.
- Автоматический сбор из доступных источников.
- Ручной сбор данных.

И разметка данных, если этого требует задача. Например, проставление отметки того, какое животное находится на изображении.





Подготовка данных

Чтобы ваш компьютер мог их корректно посчитать.

- Очистка данных — удаление данных, которые могут ввести в заблуждение (ложные, неполные данные).
- Перевод в интерпретируемый формат (csv, xls, pascal format, coco format).
- Нормализация данных (например, все данные одной колонки таблицы должны иметь один и тот же тип данных).





Анализ

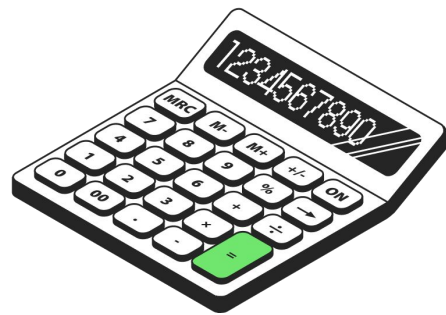




Как же это делать

Самый первый шаг — это «пощупать» данные самому. Иногда стандартные операции уже могут принести желаемый результат.

- Просмотр, какие есть данные и что они означают.
- Группировка данных.
- Получение среднего, мат. ожидания, медианы.
- Выравнивание данных.
- Визуализация (графики, тепловая карта).
- Поиск корреляций.





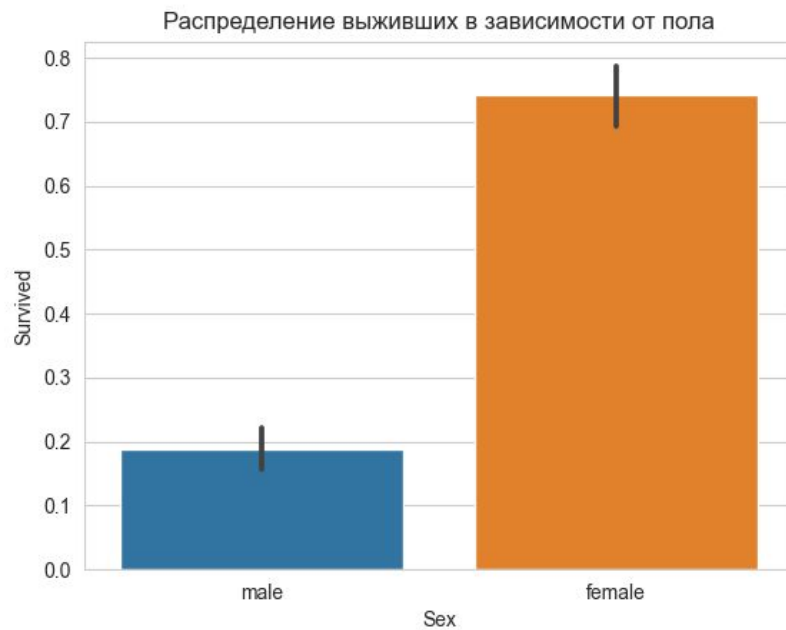
Пример кода

Например, можно посчитать статистику выживших на Титанике и понять, какой бы был у вас шанс, если бы вы туда поехали.

```
1 def survived_by_gender(df):
2     sns.barplot(x="Sex", y="Survived", data=df)
3     plt.title("Распределение выживших в зависимости от пола")
4     plt.show()
5
6     total_survived_females = df[df.Sex == "female"]["Survived"].sum()
7     total_survived_males = df[df.Sex == "male"]["Survived"].sum()
8
9     print("Всего выживших: " + str((total_survived_females + total_survived_males)))
10    print("Процент выживших женщин:")
11    print(total_survived_females / (total_survived_females + total_survived_males))
12    print("Процент выживших мужчин:")
13    print(total_survived_males / (total_survived_females + total_survived_males))
```



Результат





Предсказания





Предсказательные модели

- Самый простой пример — интерполяция.
- После анализа данных требуется разделить данные, на которых мы будем обучать модель и проверять, насколько корректно она работает.
- Далее требуется выбрать, с помощью какого метода или методов мы будем делать модель, и определить метрики, с помощью которых хотим проверить корректность (регрессия, деревья решений, KNN, K-mean, случайный лес, глубокие нейросети, PCA, Lasso).





Предсказательные модели

- После выбора и обучения модели требуется проверить её с помощью тестового датасета.
- Если результат не устраивает, то требуется вернуть предыдущие шаги и, например, поменять метод или его параметры, метрики или дополнительно проанализировать данные.
- При дополнительном анализе данных можно строить новые признаки. Например, вместо даты рождения, указывать возраст или вместо времени начала и конца, указывать продолжительность.





Метрики

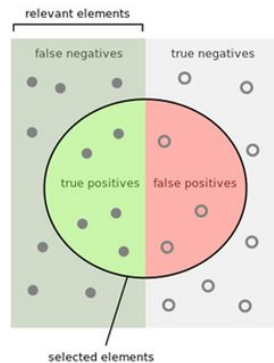
TP — верный положительный результат

FP — верный отрицательный результат

TP — ложно положительный результат

FP — ложно отрицательный результат

- Accuracy = $TP+TN/TP+FP+FN+TN$
- Precision = $TP/TP+FP$
- Recall = $TP/TP+FN$
- F1 Score = $2*(Recall * Precision) / (Recall + Precision)$



How many selected items are relevant?



Precision =

How many relevant items are selected?



Recall =



Разбивка датасета и регрессия

```
1 from sklearn.naive_bayes import GaussianNB
2 from sklearn.metrics import accuracy_score
3 from sklearn.model_selection import train_test_split
4
5 def split(train_df):
6     predictors = train_df.drop(['Survived', 'PassengerId'], axis=1)
7     target = train_df["Survived"]
8     x_train, x_val, y_train, y_val = train_test_split(predictors, target, test_size = 0.22,
9     random_state = 0)
9     return x_train, x_val, y_train, y_val
10
11
12
13 def regression(x_train, x_val, y_train, y_val):
14     logreg = LogisticRegression()
15     logreg.fit(x_train, y_train)
16     y_pred = logreg.predict(x_val)
17     acc_logreg = round(accuracy_score(y_pred, y_val) * 100, 2)
18     print(acc_logreg)
```



Развёртывание





Что под этим может подразумеваться

- **Подготовка модели для среды, где она будет выполняться**
- **Подготовка инфраструктуры для модели**
- **Подготовка обёртки для модели**
Чтобы система могла взаимодействовать с моделью
- **Мониторинг модели**
Отслеживания того, какие результаты даёт модель во время работы





Практическое задание

- Получите из seaborn датасет ирисов.
- Проанализируйте данные с помощью группировки и получите статистику с помощью pandas.
- Постройте графики по этим данным с помощью seaborn.

```
1 import seaborn as sns
2
3 iris = sns.load_dataset('iris')
4 iris.head()
```



Вопросы?

Вопросы?



Вопросы?

