# Общее описание

# домашнего задания

**Часть 1. EDA**:

1.EDA.ipynb,

файлы в подкаталоге 1.EDA (описаны в 1.EDA.ipynb notebook),

после старта домашней работы были занятия по PCA/UMAP, поэтому добавлен анализ:

FeatureSelectionPCA.ipynb — построена диаграмма абсолютных значений весов компонент по features, это помогло лучше понять/подтвердить: какие features более информативны (в них больше дисперсии), а какие ближе к константным и не были использованы PCA для компонент;

UMAP.ipynb — построена визуализация данных, по которой можно сделать вывод: структура негативного класса более сложная, там явно есть внутренние подгруппы;

UMAPplusKmeans.ipynb — построена визуализация данных с детализацией с помощью модели Kmeans, по которой можно сделать вывод: и позитивный и негативный класс имеют выраженные внутренние подгруппы, более сложный негативный класс, чем позитивный и однозначно классифицировать набор данных трудно.

**Часть 2. Preprocessing & Feature Engineering**

Данная часть выполнялась чередуясь с EDA, поэтому информация о предобработке данных и Feature Engineering вошли также в  
1.EDA.ipynb и файлы в подкаталоге 1.EDA (описаны в 1.EDA.ipynb notebook).

**разбейте данные на train-test**

«2.Split dataset to train-test.ipynb»

и подкаталог «2.Split dataset to train-test» содержит файлы тренировочной и тестовой подвыборок со стратификацией по целевой колонке.

**Часть 3. Who's the mightiest of them all?**

«3.Who s the mightiest of them all.ipynb» + два эксперимента в конце notebook.

**Кто в итоге победил?**

CatBoostClassifier — как и говорили на занятии, данная модель сильнее других на наборах данных с множеством категориальных features. В данной задаче как раз такой набор данных.