**ТРЕБОВАНИЯ К ОФОРМЛЕНИЮ**

1. Загрузите проект на *github* и поделитесь ссылкой с ментором.
2. Проект должен содержать: краткое описание и код, который может быть представлен в виде файла с расширением *.py* или в виде *Jupyter*-ноутбука.

В **описании**кратко укажите:

* + основные цели и задачи проекта;
  + краткую информацию о данных;
  + этапы работы над проектом.

**Ноутбук с решением** должен содержать:

* + цели и задачи проекта;
  + поэтапное описание работы;
  + промежуточные выводы;
  + основные итоги работы.

**Краткое описание проекта**

*Основные цели и задачи проекта* заключаются в построении модели прогноза стоимости домов по данным об истории предложений.

В проекте проведен комплекс работ по изучению данных и по исследовательскому анализу Dataset, т.е. выполнены работы:

* Exploratory Data Analysis, (EDA);
* Data preprocessing, (DP);
* Feature Engineering, (FE) и Feature Selection.

На этих этапах выполнены работы:

* по исправлению ошибок и неточностей;
* по изменениям, по структурированию;
* по пополнению недостающих данных;
* по очистке от ненужных и излишних данных.

Исследовано влияние признаков на разные модели прогноза, при необходимости созданы новые, и/или преобразованы исходные признаки. Сделаны варианты выделения важных признаков с помощью модулей:

* sklearn.feature\_selection.SelectKBest;
* sklearn.feature\_selection. f\_regression;
* sklearn.feature\_selection.SelectFromModel;
* sklearn.ensemble.ExtraTreesClassifier.

В результате проведенных экспериментов выбраны и построены модели для прогноза стоимости домов.

В дипломном проекте используются базы данных Kaggle:

[house\_prices | Kaggle](https://www.kaggle.com/datasets/lkatran/house-prices/code)

Основные этапы работ соответствуют шагам вышеописанных процессов и задач с рекурсивным возвратом назад в ходе проведения исследовательских работ.

**Тема дипломного проекта**

«Разработка сервиса, который будет предсказывать стоимость домов, основываясь на истории предложений» (База данных: [house\_prices | Kaggle](https://www.kaggle.com/datasets/lkatran/house-prices/code))

**Цель и задачи работы**

***Цель проекта:*** Построить модель прогноза стоимости домов с использованием данных Dataset Kaggle ([house\_prices | Kaggle](https://www.kaggle.com/datasets/lkatran/house-prices/code))

***Задачи проекта:***

* + 1. Проведение полного комплекса работ по разведочному анализу и чистке исходных данных Dataset (Exploratory Data Analysis, EDA), и по предварительной обработке данных (Data preprocessing, DP).
    2. Изучение влияния признаков на разные модели прогноза, при необходимости, создание новых, и/или преобразование признаков, используя методы Feature Engineering, (FE) и Feature Selection.
    3. Проведение экспериментальных расчетов по выбору моделей и подбор подходящих признаков для прогноза.
    4. Оформление проекта.

**Поэтапное описание работы**

***Этап EDA***

В начальном этапе в ходе первого знакомства с данными Dataset пришлось уточнить и осмыслить специфические термины, использованные в базе и в сфере продажи-покупки недвижимости в западных англоязычных странах. В результате знакомства с данными Dataset выявлено, что данные настолько «грязные», что даже некоторые признаки содержат данные, несвойственные для рассматриваемого признака, и/или содержат данные других признаков. Большинство признаков задано на свободном английском неструктурированном языке, порою по смыслу не относящихся к данному признаку. К тому же, очень много нестандартных сокращений, и неспецифические жаргонные слова. Числовые признаки заданы в свободном неформатном виде в разных единицах измерения с различной сигнатурой. Поэтому была предпринята попытка использовать NLP, но тщетно, и пришлось использовать «грубый» ручной парсинг. В отдельных случаях пришлось перекидывать данные одних признаков в другой признак. Результат этапа сохранен в файле, чтобы не повторять все расчеты с начала, если возникнет необходимость возврата к подготовленной базе на данном этапе.

***Этап DP***

На этапе Data preprocessing проведен более глубокий анализ структуры и содержания данных по всем признакам Dataset с использованием методов Pandas Dataframe и средств Matplotlib.pyplot , Seaborn. Выявили и исправили выбросы данных, при необходимости с возвратом на этап EDA.В результате все данные приведены к форме, приемлемой в прогнозных моделях.

Результат этапа также сохранен в файле, чтобы не повторять все расчеты сначала, если возникнет необходимость возврата к подготовленной базе данных на данном этапе. Такой рекурсивный подход потребовался постоянно при проведении экспериментальных расчетов.

***Этап FE***

На этапе Feature Engineering, (FE) и Feature Selection проведена работа по определению состава числовых и категориальных признаков, переведены категориальные признаки в Dummy-переменные, и числовые признаки в стандартную и масштабированную формы. Сделаны варианты выделения важных признаков с помощью модулей:

* sklearn.feature\_selection.SelectKBest,
* sklearn.feature\_selection. f\_regression,
* sklearn.feature\_selection.SelectFromModel
* sklearn.ensemble.ExtraTreesClassifier

Также предусмотрена возможность прологарифмировать данные числовых признаков, т.к. многие данные асимметрично распределены, прослеживается необходимость приведения их к нормальной форме распределения.

***Этап экспериментальных расчетов по выбору моделей и подбор подходящих признаков для прогноза.***

Данный этап в процессе эксперимента. На днях завершаются работы по выбору моделей и их параметров, а также подбором необходимых признаков в соответствующей структурной форме. И будут описаны итоги работ и выводы.

Замечания постараюсь полностью учесть.

***Оформление проекта***

Работа по оформлению проекта выполниться в соответствии с предложенными требованиями