**Прогнозирование стоимости квартир г. Магнитогорска**

Цель: построение математической модели прогнозирования стоимости квартир г. Магнитогорска

Задачи:

1. сбор исходных данных и их выгрузка в БД MySQL;
2. предобработка исходных данных и их анализ;
3. получение моделей прогнозирования стоимости квартир;
4. сравнение полученных моделей и выбор наилучшей;
5. создание HTTP API-сервера на основе REST API и FastAPI.

**Выбор и получение исходных данных**

В качестве источника исходных данных использованы данные сайта магнитогорской недвижимости <http://magnitogorsk-citystar.ru/>. К сайту невозможно подключиться через API, поэтому для сбора данных использовали библиотеку bs4. Объём исходной выборки составил 446 продаваемых квартир. Каждый элемент выборки описывается следующими признаками:

1. id ‒ идентификатор объявления;
2. num\_rooms ‒ число комнат;
3. flat ‒ тип планировки;
4. district ‒ район расположения;
5. house ‒ адрес;
6. floor\_num ‒ номер этажа;
7. floors\_num ‒ число этажей в доме;
8. square\_total ‒ общая площадь, м2;
9. square\_living ‒ жилая площадь, м2;
10. square\_kitchen ‒ площадь кухни, м2;
11. price ‒ цена, руб.

**Предобработка исходных данных и их анализ**

Предобработка включала в себя:

1. работу с пропущенными значениями;
2. работу с редкими и аномальными значениями;
3. создание новых признаков на основе имеющихся;
4. трансформация признаков: кодирование категориальных и масштабирование числовых.

Большинство пропусков содержалось в столбцах ‘flat’ (≈74% объёма выборки) и ‘district’ (≈45% объёма выборки). Заполнить пропуски в столбце ‘flat’ не представляется возможным, поэтому было принято решение его удалить. Пропуски в столбце ‘district’ возможно заполнить, сопоставляя геоданные районов Магнитогорска и адреса домов, но в данной работе этот вариант не прорабатывался из-за ограниченного срока выполнения тестового задания. Было принято решение удалить строки с пропусками в столбце ‘district’ и сохранить его в качестве обучающего признака.

Построение «ящиков с усами» для числовых признаков выявило небольшое количество редких значений, выходящих за пределы «усов». Эти редкие значения заменили граничными, поскольку модель линейной регрессии чувствительна к подобным выбросам.

|  |  |
| --- | --- |
|  |  |
|  |  |

Вместо признаков ‘floor\_num’ и ‘floors\_num’ для обучения модели создали новый: ‘floor\_cat’, который включает 3 значения: «верхний», «нижний» и «промежуточный». Отказ от признака ‘floor\_num’ строится на предположении о том, что на цену сильно влияет не номер этажа, а факт того, является ли он крайним.