

Прогнозирование стоимости квартир г. Магнитогорска

Цель: построение математической модели прогнозирования стоимости квартир г. Магнитогорска.

Задачи:

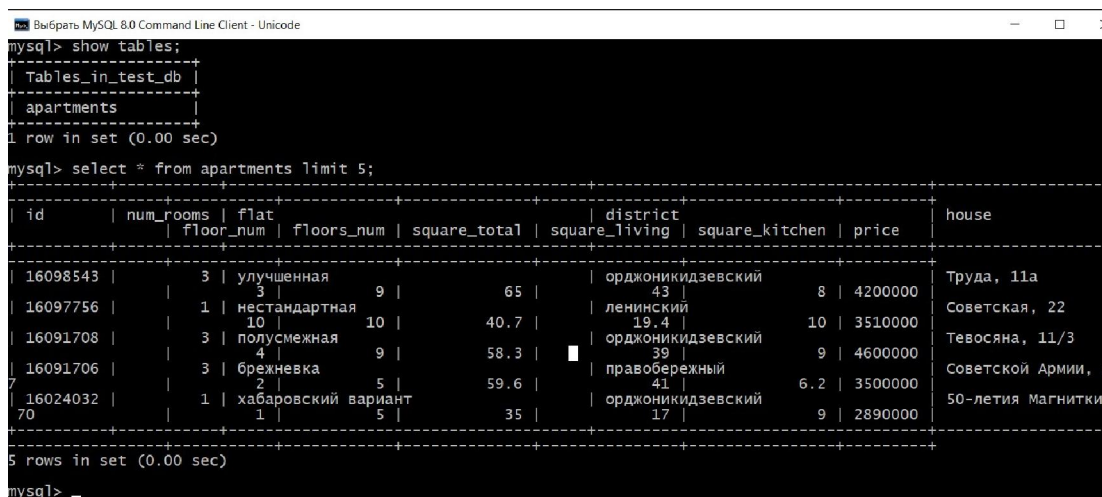
- 1) сбор исходных данных и их выгрузка в БД MySQL;
- 2) предобработка исходных данных и их анализ;
- 3) получение моделей прогнозирования стоимости квартир;
- 4) сравнение полученных моделей и выбор наилучшей;
- 5) создание HTTP API-сервера на основе REST API и FastAPI.

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.

В рамках проекта создана БД MySQL с исходными данными (рис. 1).



```
mysql> show tables;
+-----+
| Tables_in_test_db |
+-----+
| apartments         |
+-----+
1 row in set (0.00 sec)

mysql> select * from apartments limit 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| id      | num_rooms | flat      | floor_num | floors_num | square_total | square_living | square_kitchen | price | house |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 16098543 | 3 | улучшенная | 3 | 9 | 65 | 43 | 8 | 4200000 | Труда, 11а |
| 16097756 | 1 | нестандартная | 10 | 10 | 40.7 | 19.4 | 10 | 3510000 | Советская, 22 |
| 16091708 | 3 | полусмежная | 4 | 9 | 58.3 | 39 | 9 | 4600000 | Тевосяна, 11/3 |
| 16091706 | 3 | брежневка | 2 | 5 | 59.6 | 41 | 6.2 | 3500000 | Советской Армии, 3 |
| 16024032 | 1 | хабаровский вариант | 1 | 5 | 35 | 17 | 9 | 2890000 | 50-летия Магнитки, |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)

mysql>
```

Рис. 1. Screenshot экрана с исходными данными из БД

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' ($\approx 74\%$ объёма выборки) и 'district' ($\approx 45\%$ 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 соответствующий признак.

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.

Построение «ящиков с усами» (рис. 2) 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.

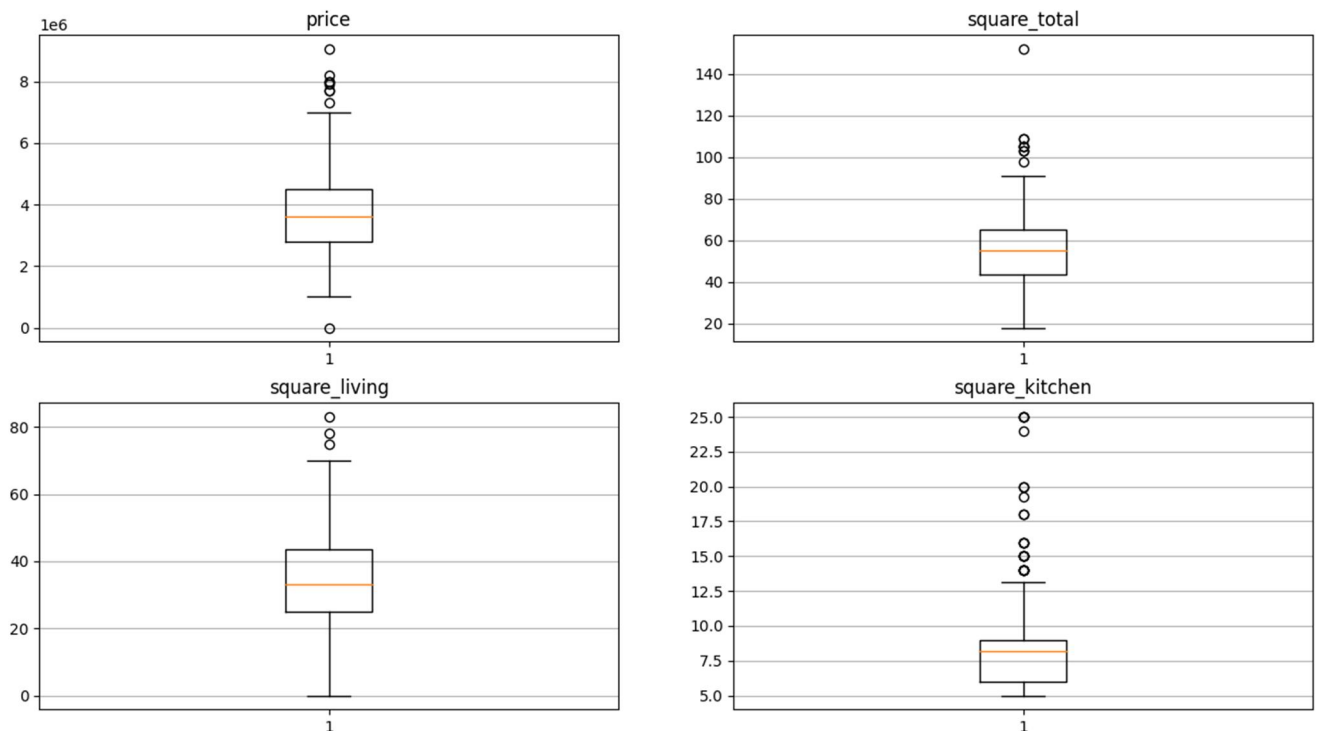


Рис. 2. Распределение значений числовых признаков

Дальнейшим этапом подготовки данных была их трансформация. Числовые признаки масштабировали с помощью StandardScaler, категориальные кодировали методом OneHotEncoder.

Получение моделей прогнозирования стоимости квартир

В рамках проекта было обучено несколько моделей линейной регрессии и CatBoost, использующая метод градиентного бустинга. Подбор гиперпараметров CatBoost осуществляли с помощью метода GridSearchCV().

В тестовом задании не указаны метрики сравнения моделей, поэтому использовали часто используемые в подобных задачах среднее квадратичное отклонение (MSE) и коэффициент детерминации (R2). Результаты сравнения моделей приведены на рис. 3.

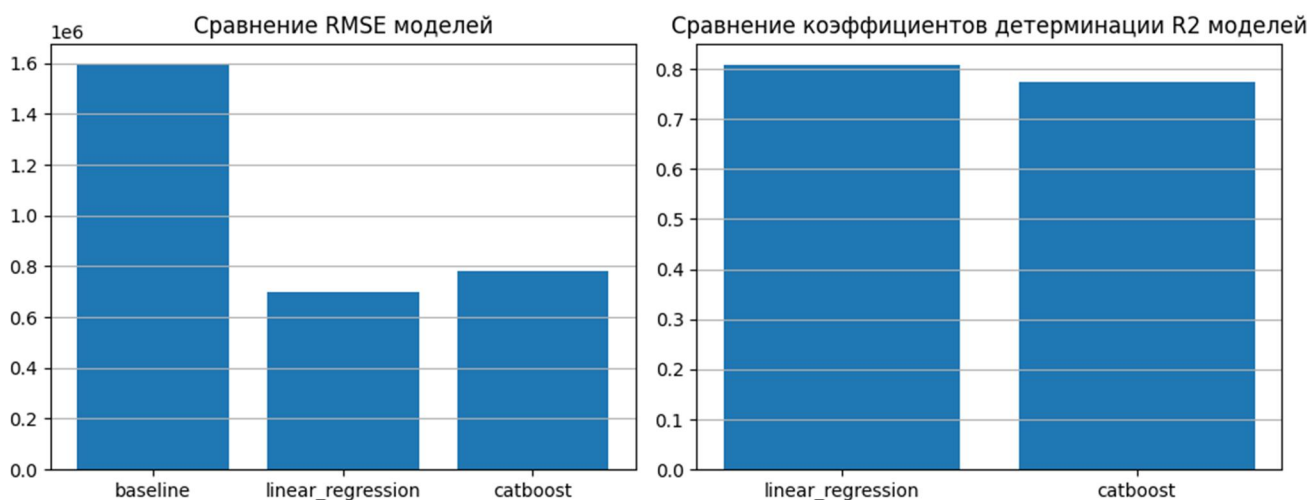


Рис. 3. Оценки качества обученных моделей

Заключение

В ходе проекта было проведено обучение нескольких моделей и их последующее сравнение. Основываясь на метрике R2, наилучшие результаты показала модель линейной регрессии с показателем R2 равным 80%. Это означает, что модель может объяснить 80% вариации целевой переменной на основе предоставленных входных данных. Такой высокий показатель R2 свидетельствует о том, что модель хорошо демонстрирует свою эффективность на конкретных данных, используемых в этом проекте, и может быть использована для достоверного прогнозирования стоимости квартир в г. Магнитогорске.

В рамках проекта созданий HTTP API сервера на основе REST API и FastAPI. Скриншот запуска сервера представлены на рис. 4.

```
sanya@DESKTOP-B2GPR2L MINGW64 ~/Downloads/Telegram Desktop/test_project (conda
$ python main.py
INFO:      Started server process [14840]
INFO:      Waiting for application startup.
INFO:      Application startup complete.
INFO:      Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit)

$ python request.py
{"prediction":3689856.54973729}
{"prediction":3474552.1253916696}
{"prediction":3558387.0390023026}
```

Рис. 4. Результаты отправки запросов на сервер информации о прогнозной стоимости 3-х квартир

На вход модели подаются параметры квартиры в формате JSON. На выходе получается цена квартиры в формате JSON.