Математическая модель стоимости жилья в зависимости от параметров этого жилья.

Постановка задачи

Необходимо построить модель предсказания цены квартиры по ее параметрам. Источник данных - данные сайта магнитогорской недвижимости www.citystar.ru.

Поставленная задача – задача регрессии, относится к задачам обучения с учителем.

Параметры квартиры, на основании которых формируется предсказание стоимости:

Название параметра	**Описание**	**Принимаемые значения**
district	район, в котором расположена квартира	Орджоникидзевский, Орджоникидзевский (левый берег), Ленинский, Правобережный, район может быть не указан
adress	улица, на которой расположен дом	можно передавать только название улицы, можно добавить номер дома
floor	сведения об этаже, на котором расположена квартира	формат данных: этаж квартиры / общее число этажей в доме
total_square	общая площадь квартиры	десятичная дробь
living_square	жилая площадь квартиры	десятичная дробь
kitchen_square	площадь кухни	десятичная дробь
num_of_rooms	количество комнат	целое число
flat_type	тип планировки	брежневка, старой планировки и д.р. Поле может быть пустым

Выбор и получение исходных данных

В качестве данных для обучения модели использованы данные о продаже квартир на сайте www.citystar.ru. Использована старая версия сайта, так как она предоставляет больше возможностей. В новой версии сайта можно сохранить все объявления в файл .excel, но при этом теряется информация о районе нахождения квартиры.

Данные сайта www.citystar.ru получены с помощью парсинга. Код парсера находится в файле parsing.py.

Данные собираются в автоматическом режиме. Для каждого объявления отбираются только те поля, которые содержат информацию, необходимую для обучения и тестирования модели.

Всего на сайте размещено 456 объявлений о продаже квартир.

После скачивания данные сохраняются в базу данных SQLight. Выбор базы данных обусловлен тем, что количество данных небольшое и тем, что она будет храниться локально на компьютере на время разработки.

Выбор метода решения

Для контроля качества модели и выбора решения использовалась метрика MAE (Mean Absolute Error) – средняя ошибка предсказания модели, чем она меньше, тем лучше. Цена квартиры на сайте измеряется в тысячах рублей.

Данные разбиты на подвыборки: обучающую и валидационную в отношении 4 : 1. Качество моделей тестировалось на валидационной выборке.

В ходе разработки модели были опробованы три подхода, их описание представлено в таблице:

Подход	**Достоинства**	**Средняя величина ошибки (MAE), тыс. рублей**
Написана собственная модель линейной регрессии, которая находит решение задачи регрессии в явном виде.	• высокая скорость предсказания • возможность внесения изменений в код модели	670
Использована модель дерева принятия решений из библиотеки Scikit Learn.	• высокая точность	584
Использована модель решающего леса из библиотеки Scikit Learn.	• более высокая точность за счет использования ансамбля из решающих деревьев.	484

В результате сравнения качества предсказаний моделей решено использовать модель решающего леса из библиотеки Scikit Learn.

Описание алгоритма решения

Алгоритм предсказания цены квартиры.

1. **Предварительная обработка данных**

   1.1. **Очистка в трансформере DataCleaner** (написан на основе анализа дефектов в данных):
   - 1.1.1. название района приводится к единому виду: убираются опечатки, разный регистр букв заменяется на единый
   - 1.1.2. пропуск в названии района заменяется значением «неизвестно»
   - 1.1.3. из адреса квартиры убирается номер дома и обозначение «ул.»

   1.2. **Уточнение признаков в трансформере FeaturesTransformer** (написан на основе формата описания квартиры, принятого на сайте www.citystar.ru):
   - 1.2.1. поле с данными об этаже квартиры заменяется на два новых поля – с номером этажа, на котором расположена квартира, и с общим количеством этажей в доме
   - 1.2.2. поле с данными о типе квартиры заменяется на два новых поля – с количеством комнат и с типом планировки

2. **Преобразование категориальных признаков** с помощью OrdinalEncoder из библиотеки Scikit Learn.

3. **Масштабирование признаков** с помощью StandardScaler из библиотеки Scikit Learn

4. **Формирование предсказания** моделью RandomForestRegressor из библиотеки Scikit Learn, обученной на тренировочной выборке.

Шаги алгоритма с 1 по 3 собраны в PipeLine и сохранены в файл pickle. Обученная модель RandomForestRegressor также сохранена в файл pickle, что позволяет использовать ее для развертывания в продакшн.

Описание модели

В ходе исследования обучена модель RandomForestRegressor (библиотека Scikit Learn) с гиперпараметрами

- max_depth=10,
- n_estimators=100,
- random_state=12345

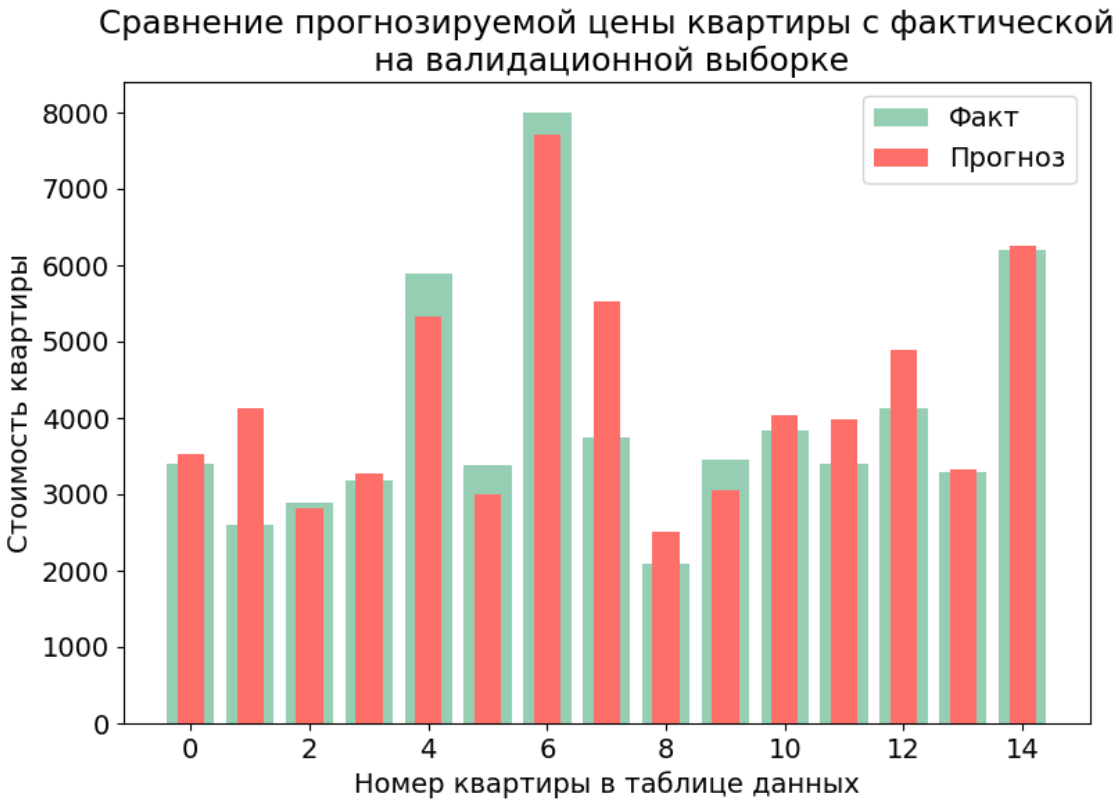
Гиперпараметры подбирались с помощью GridSearchCV.

Описание качества модели и результатов тестирования модели

Средняя ошибка на прогнозе 484 тыс рублей, что составляет примерно 14,4 % от цены квартиры (величина MAPE - mean absolute percentage error – на валидационной выборке).

После обучения модели проведен анализ важности параметров квартиры, на основании которых модель предсказывает ее цену. Самыми важными оказались:

- общая площадь
- площадь кухни
- общее количество этажей в доме



Общий вывод

В ходе проведенного исследования построена модель предсказания цены квартиры по ее параметрам.

Источник данных - сайт магнитогорской недвижимости www.citystar.ru. Объявления с сайта получены с помощью парсинга. Код парсера находится в файле parsing.py. Данные сохраняются в базу данных SQLight. Написаны собственные трансформеры для предварительной обработки данных – DataCleaner и FeaturesTransformer. Все шаги предварительной обработки данных собраны в PipeLine.

На тренировочной выборке обучена модель RandomForestRegressor с гиперпараметрами

- max_depth=10,
- n_estimators=100,
- random_state=12345

На вход модели подаются параметры квартиры в формате JSON на выходе получается цена квартиры в формате JSON. Средняя ошибка на прогнозе 484 тыс рублей, что составляет примерно 14,4 % от цены квартиры.

Программный код хранится в репозитории на GitHub: https://github.com/VeraMeln/magnito_price.git