**Предсказание стоимости домов в США**

Цель и задачи проекта: подготовить данные, обучить модель для предсказания стоимости домов в США, получить высокое качество модели.

1. **Анализ данных.**

Датасет содержит информацию по продажам домов в городах США, имеет 377185 записей и 18 колонок.

Пример записи представлен на рисунке 1.

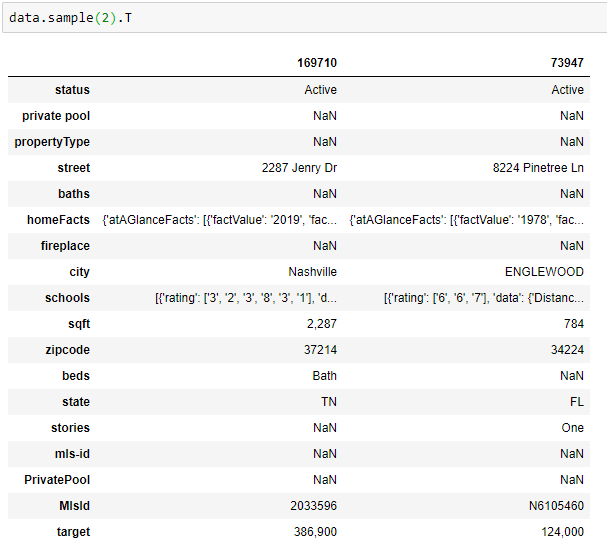


Рис. 1 Некоторые записи датасета

Статистика по колонкам представлена на рисунке 2, по данной информации можно судить, что колонки имеют строковый тип.

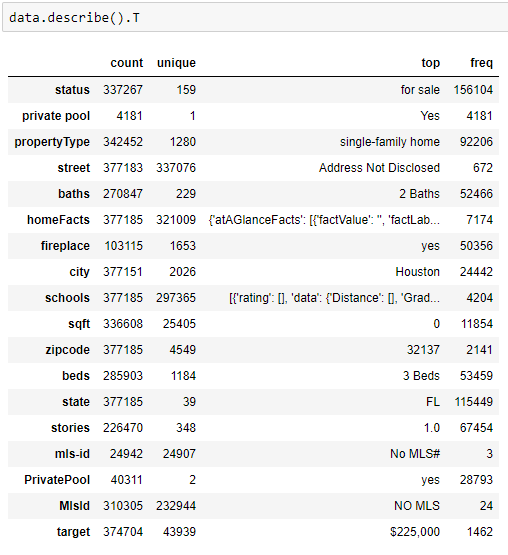


Рис.2 Статистическая информация по колонкам.

На рисунке 3 представлена информация о пропущенных данных.

|  |  |
| --- | --- |
| Желтый – пропущенные данные, синий – не пропущенные | **Процент пропущенных данных в колонках**  status - 10.583%  private pool - 98.892%  propertyType - 9.208%  street - 0.001%  baths - 28.193%  homeFacts - 0.0%  fireplace - 72.662%  city - 0.009%  schools - 0.0%  sqft - 10.758%  zipcode - 0.0%  beds - 24.201%  state - 0.0%  stories - 39.958%  mls-id - 93.387%  PrivatePool - 89.313%  MlsId - 17.731%  target - 0.658% |

Рис. 3 Информация о пропущенных данных

По данным изображениям видно, что только в четырех колонках нет пропусков: homeFacts, schools, zipcode, state. Есть повторяющиеся колонки: private pool и PrivatePool, mls-id и MlsId с разным уровнем пропусков. Также колонки homeFacts и schools содержат еще признаки в json формате (рисунок 1, 2).

На рисунках 3, 4 представлена информация о ценах домов в разрезе штатов и городов.

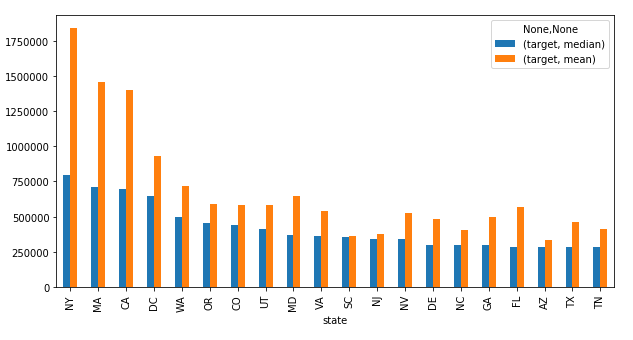


Рис. 4. Топ 20 штатов с высокой стоимостью дома.

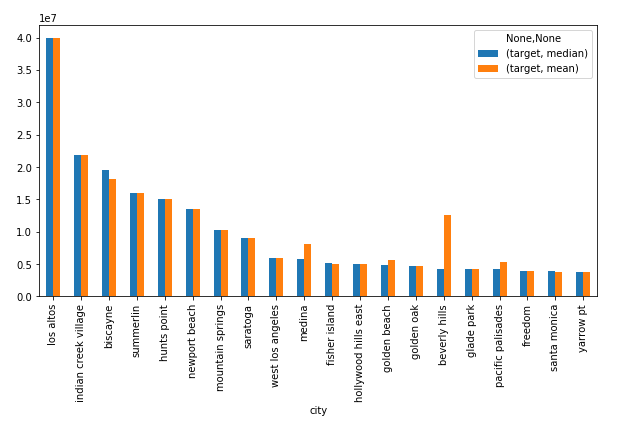


Рис.5. Топ 20 городов с высокой стоимостью дома.

В тройке дорогих штатов большая разница между средним и медианным значением. Два города из топа имеются только в одной записи (рисунок 6), соответственно судить о стоимости всего жилья не стоит.

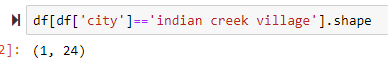
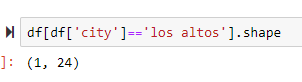


Рис. 6

Строить модель по всей стране думаю не стоит в виду большого разброса цен по штатам и городам, поэтому остановимся на одном штате или городе, например Florida и Houston как самые распространенные в датасете (рисунок 7).

|  |  |
| --- | --- |
| Количество записей в датасете по штатам | Количество записей в датасете по городам |

Рис. 7. Количество записей

Будем предсказывать стоимость дома в городе Houston. Для этого оставим в датасете только записи с городом Houston предварительно переведя в нижний регистр значения колонки city.

1. **Предобработка и очистка данных.**

Предобработка данных представлена в notebook «Sokolov\_final\_preprocessing.ipynb», где были выполнены следующие действия:

* Объединил колонки: private pool и PrivatePool, mls-id и MlsId.
* Получил дополнительные данные из колонок homeFacts и schools:

homeFacts – разобрал на колонки,

schools – посчитал общее количество школ на районе, среднее расстояние, средний рейтинг, количество школ по классам (preschool, elementary, midle и тп.).

* Дополнил координатами почтового отделения. Файл US.txt взят здесь <http://download.geonames.org/export/zip/>.
* Дополнил координатами центров городов. Файл worldcitiespop.csv взят здесь <https://www.kaggle.com/max-mind/world-cities-database>.
* Дополнил координатами адреса, использовал библиотеку geopy.
* Дополнил данными по населению района. Датасет Azure взял здесь <https://azure.microsoft.com/en-us/services/open-datasets/catalog/us-decennial-census-zip/#AzureNotebooks>.

После удаления дубликатов получили датасет с 14595 записями (почти в 2 раза меньше, чем исходный датасет по городу).

1. **Модель.**

Baseline и модель представлены в notebook «Sokolov\_final\_baseline.ipynb».

В качестве целевой переменной выберем значение цены на квадратный фут площади дома (Price/sqft).

Проанализировал получившиеся данные на выбросы (таблица 1) и избавился от выбросов (IQR).

Таблица 1.

| **Колонка** | **Выбросы** | **IQR** |
| --- | --- | --- |
| sqft |  |  |
| stories |  |  |
| beds |  |  |
| lotsize |  |  |
| Price/sqft |  |  |
| year\_built |  |  |
| distanceToPost |  |  |
| distanceToCentre |  |  |

Матрица корреляций числовых признаков датасета houston представлена на рисунке 8.

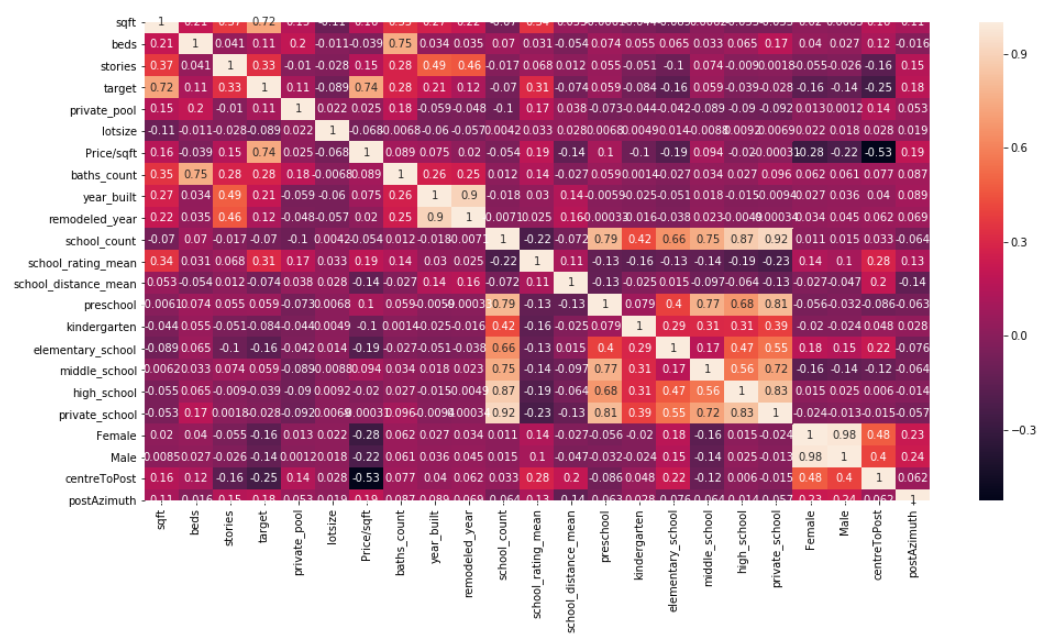
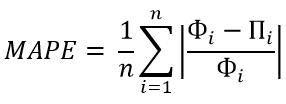


Рис.8. Полная матрица корреляций.

Для оценки качества модели использовал две метрики.

Средняя абсолютная процентная ошибка (MAPE)— показывает среднее арифметическое значение всех абсолютных (взятых по модулю) процентных ошибок прогноза (Ф — фактическое значение стоимости, П — прогнозное значение стоимости). Рассчитывается по формуле:



Медианная абсолютная процентная ошибка (MedAPE) - та же идея вычисления абсолютной процентной ошибки, но вместо среднего арифметического берем более устойчивую к выбросам медиану.

В качестве модели использовал RandomForestRegressor, ExtreTreeRegressor, XGBRegressor, CatBoostRegressor затем усреднение предсказаний.

После очистки данных исходного датасета получил MAPE 21%, что явно недостаточно. Добавил в датасет координаты почтовых отделений, городов, адреса, население по почтовым отделениям, посчитал расстояние до центра, до почты и азимут. В итоге было получено 29 признаков, после экспериментов с моделью оставил только числовые признаки в количестве 16 штук.

Величина метрик составила:

MAPE: 14.52

MedAPE: 7.83

В ходе экспериментов выяснилось, что расстояние от адреса до центра и азимут адреса ухудшают результат, замена пропусков расстояния средним или медианным значением, а также рандомным значением на основе расстояния до почты, результатов не улучшило. Теоретически эти признаки должны быть наиболее важными в этом датасете. Также видно из рисунков 9, 10, что наиболее важный признак – расстояние от центра до почты. Возможные причины: данные о координатах (почта, центр города, адрес) собирались из разных источников, что внесло большую погрешность в вычислении расстояния. В итоге удалил эти признаки для обучения модели.

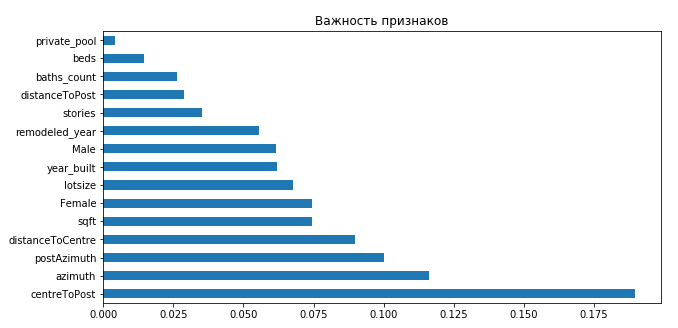


Рисунок 9. RandomForestRegressor важность признаков.

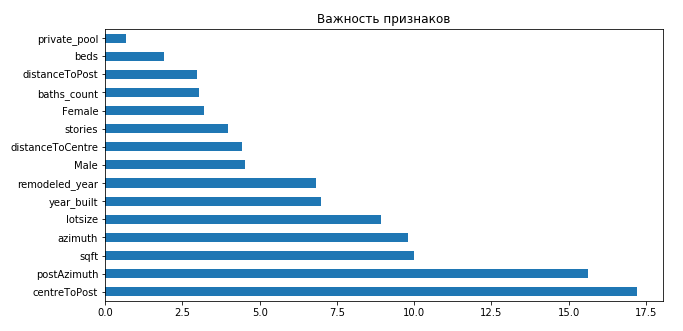


Рисунок 10. CatBoostRegressor важность признаков.

1. **Прототип.**

Прототип представляет собой сервер на Flask.

Cостоит из следующих файлов:

- server.py – инициализация сервера, получение данных с формы, вызов метода для предсказания, обновление формы.

- house\_predict.py – загрузка моделей, предобработка данных с формы, предсказание.

- form.html – шаблон формы.

- файлы обученных моделей (catbr.pkl, etr.pkl, rfr.pkl, xgb.pkl) – на гитхаб не влезли.

- файлы с данными по координатам и населению (US.txt, US\_cities.csv, zip\_sex\_pop.csv) - на гитхаб не влезли.

При инициализации сервера происходит загрузка ранее обученных моделей, и датафреймов с координатами и данных по населению.

После загрузки сервера доступна форма (рисунок 11)

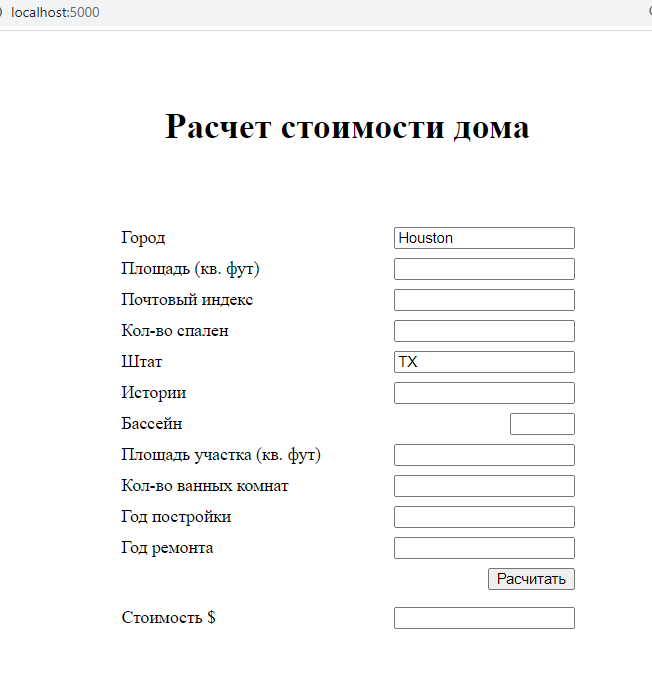


Рисунок 11. Форма

1. **Выводы.**

Датасет требует большей подготовки. Получение координат по адресу с бесплатных сервисов требует много времени. Необходимо дополнить новыми данными например: получить параметры дома по Mls\_ID, получить дополнительные данные по адресу (перепись, криминальная обстановка, наличие социально значимых мест и тп.).