Задание на экзамен

Бершицкой Анастасии Сергеевны
team 5

Цель проекта

Разработать сервис, который сможет предсказывать стоимость недвижимости на основании имеющейся истории предложений, применив изученные инструменты языка Python, в срок до 01.06.202 включительно.

Задачи проекта

1. Изучить имеющийся датасет: содержание и форматы данных.
2. Произвести очистку данных: преобразовать данные к единому типу и формату по столбцам, удалить пропуски, разнести сгруппированные данные по столбцам.
3. Определить, какие данные необходимы для последующего анализа, максимально оцифровать их.
4. Воспользоваться графическими инструментами Python для последующего отбора данных в аналитику.
5. Обучить различные изученные на курсе модели прогнозирования данных, выбрать наилучшую.

Метрика проекта

В качестве метрики проекта выбран коэффициент детерминации модели (R-квадрат) величиной не менее 0,8.

Справочная информация

В качестве справочной информации использовался словарь риэлторских терминов, а также различная информация, взятая из сети интернет:

- словарь терминов (https://sun.iwu.edu/~finance/pages/docs/Glossary.pdf);
- определение MLS-ID (https://support.realtor.com/s/article/What-is-my-MLS-ID);
- особенности американских ванных комнат (http://begin-english.ru/article/chto-nuzhno-znat-ob-amerikanskom-dome/);
- данные об этажности (https://skyeng.ru/articles/poleznye-anglijskie-frazy-dlya-pokupki-ili-arendy-nedvizhimosti-za-granitsej).

Что получилось сделать

1. После прочтения исходного датасета, была собрана дополнительная информация (ссылки приведены в разделе выше) для понимания сведений, а также проанализированы форматы данных и их содержание.
2. После детального анализа была проведена очистка данных:
    - по полям target и sqft удалены точки, запятые, знак $ и прочие символы для приведения данных к единому формату (в т.ч. относительно денежного выражения);
    - в полях beds и baths удален текст, оставлена только информация о количестве ванных/спален;
    - проанализировано содержание дублирующихся столбцов о наличии бассейна и номере MLS-ID: в связи с тем, что данные не пересекаются, было принято решение попарно объединить данные столбцы;
    - столбцы status, propertyType и city приведены к единому виду на основании ТОП-30 значений в виду высокой уникальности данных;
    - данные столбцов schools и home_facts в виде словарей были очищены от лишних символов, а далее разнесены по отдельным столбцам (кроме стоимости за квадратный фут ввиду её прямой корреляции с площадью и полной стоимостью недвижимости).
3. Параллельно с п.2 данные были приведены в цифровой вид. Так, например, поле status было проранжировано в диапазоне от 0 до 1, в зависимости от вероятности скорой продажи объекта недвижимости. Аналогичный принцип был применен к полям private pool, fireplace, mls-id. Информация о городе и типе собственности не была оцифрована, т.к. была разнесена в столбцы категорий. Столбцы, не участвующие в анализе, как например state или zipcode, были удалены из датасета.
4. После очистки данных с помощью инструментов визуализации были проанализированы оставшиеся столбцы, «хвосты» распределяющих функций были также отсечены из датасета для корректного анализа. На финальном этапе построен графи корреляции данных.
5. На основании итогового датафрейма были обучены модели двух типов: линейная регрессия и градиентный бустинг. Валидация моделей показала, что модель градиентного бустинга показывает более качественный результат, однако, всё равно достаточно низкий для точного прогнозирования, т.к. коэффициент детерминации составил всего **0,46**.