Финальный проект

Прогнозирование стоимости дома по характеристикам

План проекта

1. Постановка задачи

Необходимо создать модели прогнозирования стоимости объекта недвижимости по заданным характеристикам и определить наиболее точную модель по метрике качества Для обучения моделей будет использован датасет https://drive.google.com/file/d/1An8LtTSdv6pB5nSpNBwH1Njcwh2SpYW6/view

Сложность заключается в том, что представленный датасет не имеет описания. Поэтому значительная часть времени затрачена на анализ данных датасета.

2. Выбор метрики

Для определения качества модели будем использовать метрику MAPE. Эта метрика показывает, на сколько процентов в среднем наше предсказание отклоняется от реального значения.

3. EDA

На начало анализа датасета у нас не было описания полей и данных в них. После анализа датасета мы определили и создали следующие признаки:

1. Target – целевая переменная. Преобразовали в числовое значение. Удалили записи со значениями более 3000000
2. Status - статус жилья. Всего было вариантов - 159. Мы преобразовали все возможные варианты к основным вариантам:

   1) 'for sale' - на продажу;

   2) 'active' - активно используется;

   3) 'unknown' - значение пропущено или непонятно;

   4) 'new' - новое жилье;

   5) 'for rent' - сдается в аренду

3. private pool и PrivatePool – наличие частного бассейна. Объединили оба признака в один PrivatePool, привели его к числовому значению
4. propertyType - Признак 'тип недвижимости'. Всего было вариантов – 1253. Мы преобразовали значения в следующие категории типа недвижимости:
   - houses - различные типы домов;

- condos/coops - кономиниумы и кооперативы;
- unknown - неизвестный или непонятный тип недвижимости;
- lots/land - участки земли;
- townhomes - таунхаус;
- multi-family - дом на несколько семей;
- other style; 8) manufactured - новый дом

5. homeFacts - Признак "Факты о доме". Из этого признака выделили новые признаки:
   - "Year built" - год постройки
   - "Remodeled year" - год реконструкции
   - 'Heating' - отопление
   - 'Cooling' - кондиционер
   - 'Parking' - парковка
   - 'lotsize' - площадь участка
   - 'Price/sqft' - цена за 1 ед площади участка
6. street - Признак- адрес. Из этого признака выделили новые признаки:
   - house_num – номер дома
   - street_name – название улицы
7. baths - Признак "ванные комнаты". Преобразовали в числовой признак. Значения более 30 заменили на медианное значение для каждого значения propertyType
8. fireplace - Признак "камин". Преобразовали в числовой признак
9. city - Признак "город". Отсутствующие значения заменили на «nocity»
10. schools - Признак - 'школа'. Данные о ближайших школах. Выделили новые признаки:
    - elem_scl_mean_count - Количество начальных школ
    - elem_scl_mean_rat - Средний рейтинг начальных школ
    - elem_scl_mean_dist - Среднее расстояние до начальной школы
    - mid_scl_mean_count - Количество средних школ
    - mid_scl_mean_rat - Средний рейтинг средних школ
    - mid_scl_mean_dist - Среднее расстояние до средней школы
    - high_scl_mean_count - Количество высших школ
    - high_scl_mean_rat - Средний рейтинг высших школ
    - high_scl_mean_dist - Среднее расстояние до высшей школы
    - priv_scl_mean_dist - Количество частных школ
    - priv_scl_mean_count - Среднее расстояние до частной школы

    Заменили расстояние до школ свыше 35 миль и пропущенные значения на среднее значение Заменили пропущенные значения рейтингов на 0
11. sqft - Признак - "площадь дома в квадратных футах". Преобразовали в числовой признак. Значения представленные в акрах преобразовали в квадратные футы Заменили значения более 7000 и стоимостью менее 1500000 на медианные значения

12. zipcode - Признак "почтовый индекс". Преобразовали в 5ти значное значения типа строки. С помощью библиотеки uszipcode создали новые признаки связанные с почтовым индексом:
    - population - население
    - population_density - плотность населения
    - land_area_in_sqmi - площадь земли
    - water_area_in_sqmi - площадь воды
    - housing_units - количество жилых единиц
    - occupied_housing_units - количество заселенных жилых единиц
    - median_home_value - медианная стоимость жилья
    - median_household_income - медианная доход домохозяйства

    Удалили пропущенные значения
13. beds - Признак - "количество спален". В признаке имелись значения 'Baths' и другие примеси. Преобразовали признак в числовой и убрали примеси. Значения с примесями заменили на медианное
14. state - Признак - "штат"
15. stories – Признак "количество этажей". Преобразовали в числовой признак. Пропущенные значения заменили на медианные
16. mls-id и MlsId - это признаки идинтификатора включения в единую базу данных участников на рынке недвижимости США. На основании этих признаков создали новый признак MLS со значениями 1 или 0
17. Year built - Признак - 'год постройки'. Отсутствующие значения заменили на 2100. Создали новый признак "возраст дома" (age_house)
18. Remodeled year – признак «год реконструкции». Отсутствующие значения заменили на 2100. Создали новый признак "возраст реконструкции" (age_remodeled)
19. Heating - Признак "отопление". Было много вариантов – 1887. Преобразовали все варианты в следующие категории отопления:
    - central
    - forced air
    - electric
    - gas
    - other
    - unknown
20. Cooling - Признак - "охлаждение". Было много вариантов - 1324. Преобразовали все варианты в следующие категории охлаждения дома:
    - central
    - has cooling
    - ceiling fan
    - other
    - no cooling
    - unknown

21. Parking - Признак - "парковка". Создали новые признаки из этого поля:
    - attached_garage - пристроенный гараж
    - detached_garage - отделенный гараж
    - carport - Навес
    - parking_space - парковочное место

    Признаки числовые. Если признак равен 0, то означает что он отсутствует, если значение более 0, то это количество машиномест в этом признаке. Удалили записи со значением parking_space более 100
22. lotsize - Признак 'размер участка'. Преобразовали в числовое значение. Значения выраженные в акрах привели к квадратным футам
23. Price/sqft - признак 'стоимость квадратного фута площади дома'. Привели в числовой формат

4. Дубликаты

В датасете удалили дупликаты по следующим столбцам: 'city', 'street', 'target'

5. Выбросы

- Удалено выбросов в столбце target - 24350
- Удалено выбросов в столбце lotsize - 36140
- Удалено выбросов в столбце beds - 27938
- Удалено выбросов в столбце baths - 15492
- Удалено выбросов в столбце sqft - 15440
- Размер датафрейма до удаления выбросов - 303100, после удаления - 220286

6. Выбор значимых признаков.

6.1. Корреляция числовых признаков

Числовые признаки, которые наиболее коррелируют с целевой переменной:

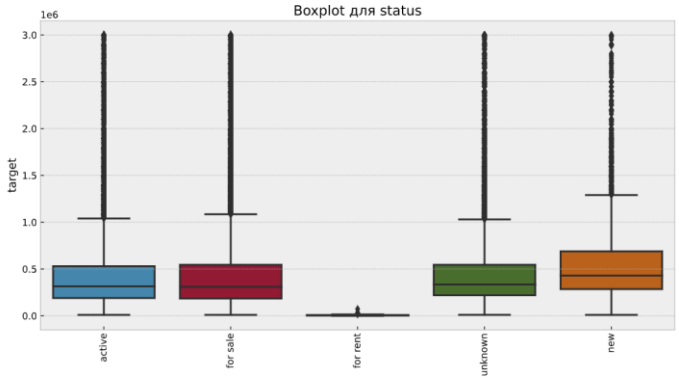
- zip_median_home_value - 0.585197
- baths - 0.585197
- zip_median_household_income - 0.341109
- zip_population_density - 0.264796
- beds - 0.204898
- high_scl_mean_rat - 0.115149

- zip_land_area_in_sqmi - 0.112175
- elem_scl_mean_rat - 0.105801
- parking_space - 0.093926
- sqft - 0.092984

Вполне ожидаемо, что на первом месте показатель медианного стоимости жилья для почтогого индекса с корреляцией 59%, но достаточно странно, что у площади жилья корелляция с целевой переменной всего 9%. Возможно в признаке sqft есть ошибки.

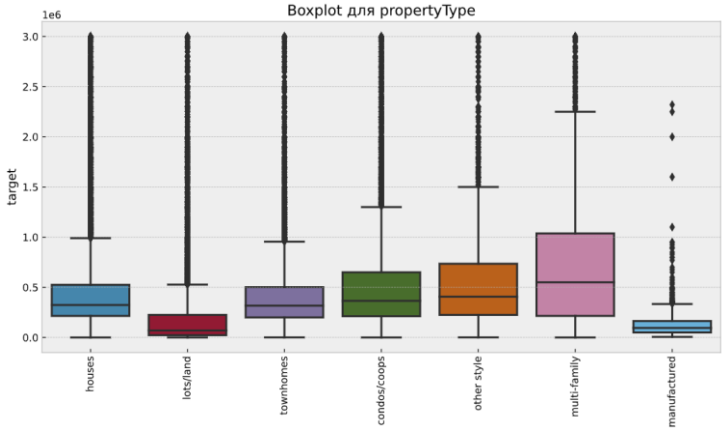
6.2. Категорийные признаки с небольшим количеством категорий

Признак status:



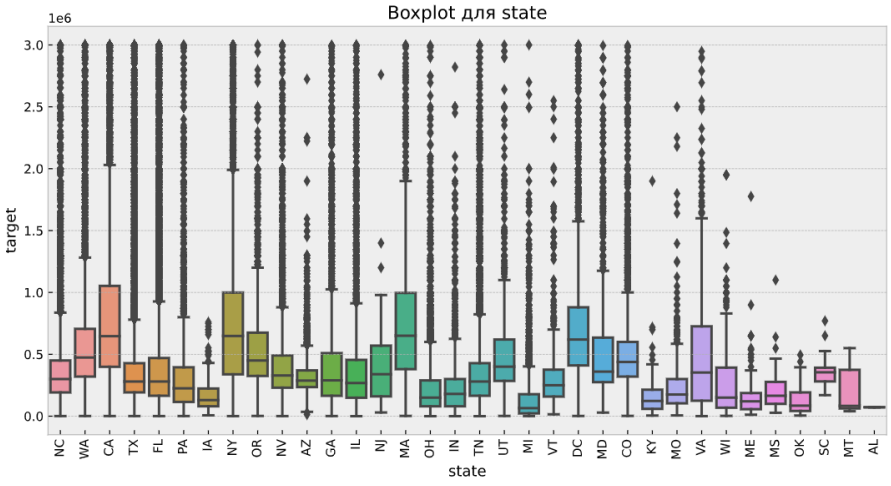
- минимальное значение target для категории аренда. Разница между ней и остальными категориями в 2 порядка
- максимальная верхняя граница значения target для категории new ~1.3 млн.
- категории active, for sale, unknown - практически одинаковые 1й, 2й и 3й квартили
- для всех категорий признака имеются "выбросы". Здесь и далее под выбросами мы будем понимать не ошибку в данных (разное жилье может стоить и 50 тыс. и 50 млн.), а статистические выбросы. Хотя, возможно, где-то в данных имеются ошибки.

Признак propertyType:



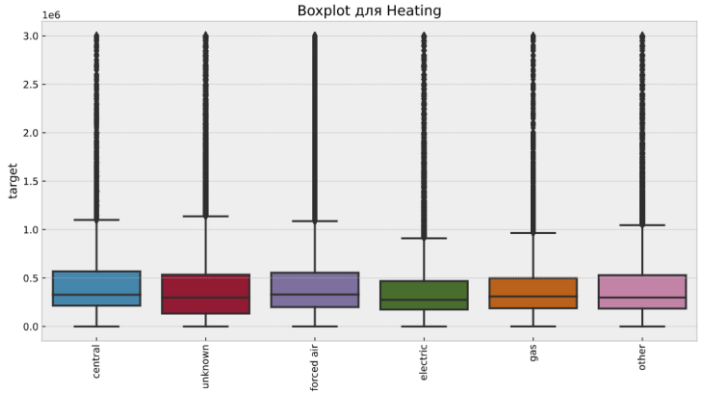
- максимальная верхняя граница значения target для категории multi-family ~2.25 млн.
- минимальная верхняя граница значения target для категории manufactured ~0.33 млн.
- для всех категорий признака имеются "выбросы"

Признак state:



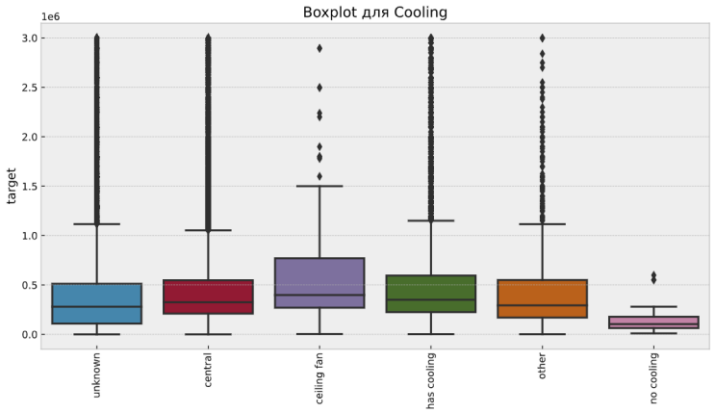
- максимальная верхняя граница значения target для штатов CA ~2.05 млн и NY ~2 млн.
- минимальная верхняя граница значения target для штата AL ~2.05 млн NY ~0.1 млн.
- для всех штатов кроме AL имеются "выбросы"

Признак Heating:



- максимальная верхняя граница значения target для категории unknown ~1.15 млн.
- минимальная верхняя граница значения target для категории electric ~0.9 млн.
- для всех категорий признака имеются "выбросы"

Признак Cooling:



- максимальная верхняя граница значения target для категории ceiling fan ~1.5 млн.
- максимальная верхняя граница значения target для категории no cooling ~0.3 млн.
- для всех категорий признака имеются "выбросы"

Для категорийных признаков с небольшим количеством категорий было использовано кодирование One-hot encoding

6.3. Категорийные признаки с большим количеством категорий

Признаки city, zipcode, street_name имеют большое количество категорий. Для их кодирования было использовано кодирование **FeatureHasher**, которое в проходимом курсе не изучали

6.4. Тест Стьюдента

Статистически значимые различия на тесте Стьюдента были обнаружены для всех категорийных признаков

7. Model 0: Наивная модель

Точность наивной модели - MAPE: 58.34%

8. Model 1: CatBoost

По сравнению с наивной моделью (MAPE: 58.34%) с помощью CatBoost мы улучшили метрику до 20.96%

9. Model 2: RandomForestRegressor

По сравнению с CatBoost (MAPE: 20.96%) с помощью RandomForestRegressor мы улучшили метрику до 20.29%

10. Model 3: GradientBoostingRegressor

Метрика MAPE: 24.97% ухудшилась по сравнению с RandomForestRegressor (MAPE: 20.29%)

11. Model 4: Простая полносвязная нейросеть

С помощью простой полносвязной нейросети на 2х этапах (1й с шагом обучения 0.01, 2й с шагом обучения 0.001) мы достигли метрики MAPE 20.61% и не смогли улучшить результат RandomForestRegressor MAPE: 20.29%

12. Blending

- Были смешаны результаты моделей: CatBoost, RandomForestRegressor, Простая полносвязная нейросеть

Метрика MAPE: 18.73%

Результат метрики существенно улучшился по сравнению с лучшей метрикой

RandomForestRegressor(MAPE: 20.29%)

13. Общие выводы

Что мы делали в проекте:

1) анализ и обработка датасета - 80% времени

- определение признаков, создание новых категорий
- создание новых признаков на основе существующих
- анализ числовых признаков
- анализ категорийных признаков
- определение и удаление дубликатов записей
- определение и замена (удаление) пропусков
- определение и удаление выбросов

2) определение метрики качества (MAPE)

3) Создали несколько моделей для прогнозирования стоимости недвижимости:

- CatBoost
- RandomForestRegressor
- GradientBoostingRegressor
- Простая полносвязная нейросеть

Лучший одиночный результат MAPE: 20.29% был достигнут на модели RandomForestRegressor

Далее по ухудшению метрики:

- 20.61% - простая полносвязная нейросеть
- 20.96% - CatBoost
- 24.97% - GradientBoostingRegressor

4) С помощью Blending существенно была увеличена метрика качества MAPE до 18.73%. Смешивались следующие модели:

- CatBoost
- RandomForestRegressor
- Простая полносвязная нейросеть

Что не удалось сделать: улучшить метрику качества моделей до 15%. Вероятно проблема во входном датасете и создании на его основе новых категорий.

Для улучшения качества модели предлагается улучшить качество входных данных и дополнить их фотографиями недвижимости, сделать перебор гиперпараметров и признаков.