Проект по умолчанию «**Разработка сервиса, предсказывающего стоимость домов, основываясь на истории предложений**».

Для анализа выбрал метрики MAPE (mean absolute percentage error) и MAE (mean absolute error).

Предоставленные данные были очень «грязными» с отсутствием информации или же перемешанной по всем 17 признакам. Всего для работы было предоставлено 377 185 строк.

После долгой обработки данных- генерации информации из имеющихся признаков и добавления новых признаков получили 15 признаков с 249 832 строками. Анализ строил на трех моделях:

1. **RandomForestRegressor**MAPE 54.44145587111288 %

MAE 298 954.95250632265 $

1. **XGBOOST**

MAPE 56.87272670232019 %

MAE 2989 54.95250632265 $

1. **GradientBoostingRegressor**

MAPE 53.37723667717164 %

MAE 27 9315.4420384513 $

Лучший результат показала модель **GradientBoostingRegressor**

GradientBoostingRegressor(max\_depth=15, min\_samples\_leaf= 15, min\_samples\_split= 10, max\_features= 'auto', random\_state=RANDOM\_SEED)

Результаты были не очень хорошими, от чего возникла идея совсем по-другому подойти к выполнению задания. То есть не тратить большое количество времени на заполнение отсутствующих значений, а в большинстве своем просто удалять их. Так же дополнил данные новыми признаками – самые дорогие города, популяция городов, ранга штата.

После обработки данных по второму варианту выполнения задания на выходе получили 26 признаков с 73 233 строками.

Анализ второго варианта строил на обучении моделей:

1. **HistGradientBoostingRegressor**

MAPE 36.97016089568272 %

MAE 263 473.643443036 $

1. **GradientBoostingRegressor**

MAPE 24.26046465253648 %

MAE 206 917.21741424745 $

1. **RandomForestRegressor**

Mape 25.08840806328279%

MAE 219 704.37061252055 $

Лучший результат снова показал **GradientBoostingRegressor**

GradientBoostingRegressor(max\_depth=15, min\_samples\_leaf= 15, min\_samples\_split= 10, max\_features= 'auto', random\_state=RANDOM\_SEED)

По влиятельности признаков на модель видим, что основные 5 признаков это: размер дома, количество ванных комнат, населенность штата, расстояние до школы и год постройки.

Из двух вариантов наилучший результат показал второй, где данные обрабатывались грубо (удалил NAN), модель GradientBoostingRegressor обучалась на дополненных данных о характеристиках мест, где расположены дома (города, штаты).

Можно сделать вывод, что возможно некорректно были обработаны данные в 1 варианте , где можно было бы добиться более положительных результатов.

Характеристики дома, такие как камин, бассейн или парковочное место конечно влияют на стоимость жилья, но в большей степени на стоимость влияет географическое расположение жилья.   
Проще говоря стоимость большого жилья в дорогих и густонаселенных городах выше!