Hope Frost
Capstone project

House Price Project Report

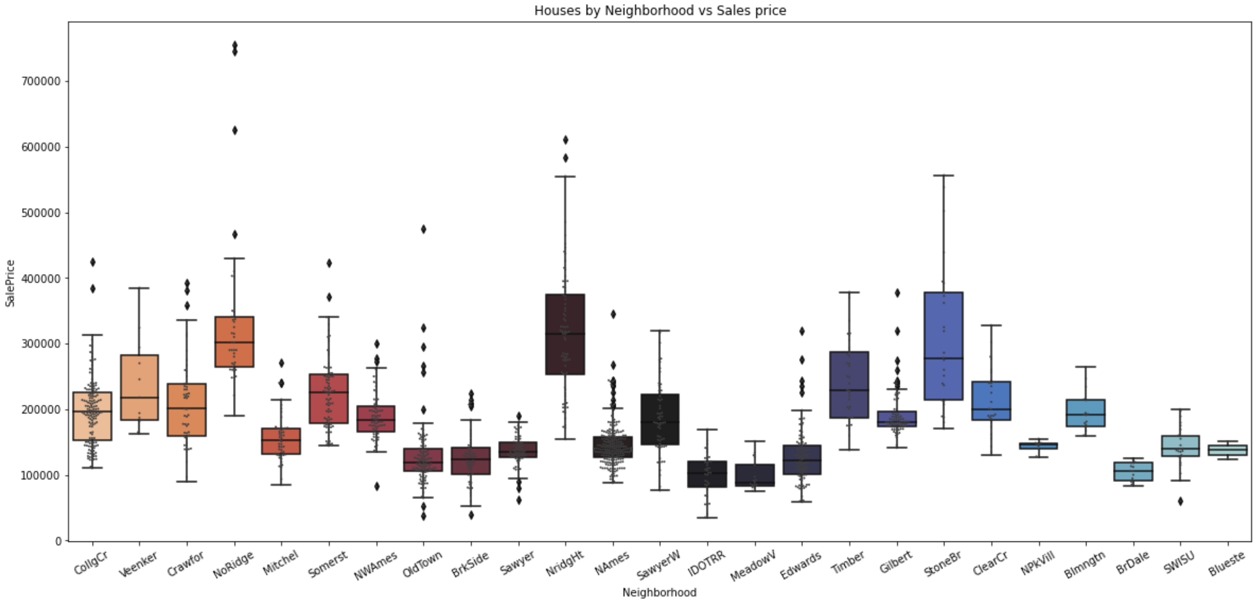
The Price of a House

When a house goes on the market, how much will it really sell for? Homeowners want to know what they're home is worth. Realtors want to price homes aggressively to sell. Home buyers want to know they're not overpaying. Everyone involved is interested in market prices and what they're neighborhood houses are selling for.

"The Ames Housing dataset was compiled by Dean De Cock for use in data science education" (https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview) The Kaggle competition dataset contains 80 qualitative and quantitative features on 1459 houses. A second dataset excluding the sales price provides another 1459 houses to be used as a final testset to submit to the competition.

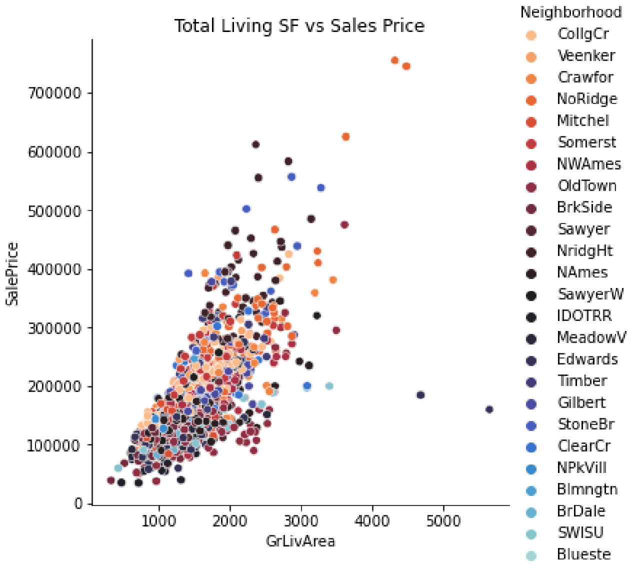
Data Wrangling

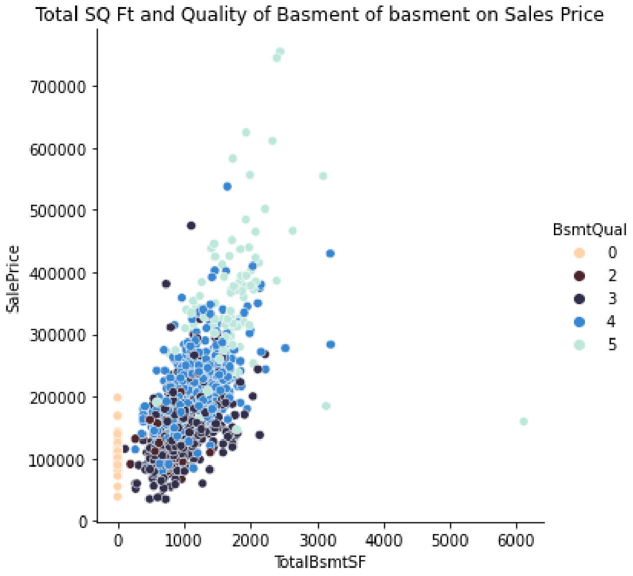
The Dataset had lots of categorical data scaling quality and condition of attributes. These features were simplified into 0-5 integer scales. Many Features had missing data to impute, relying on the data description to either fill mode from the train set or 0 due to features not present in house. The dataset has houses spread over 25 neighborhoods. Here you see the neighborhoods by price:



Exploratory Data Analysis

To get a better idea of the features provided each feature was plotted with Sales price to visualize their effect. Almost every plot showed some extreme outliers. Most features are correlative to sales price but few have a very high correlation. Since our aim is to predict the sales price rather than show a trend it's best to give the model as much information as we can. Take a look at these plots showing the correlation of square footage of living space on the sales price by neighborhood, and the square footage of basement and quality of basement on sales price.





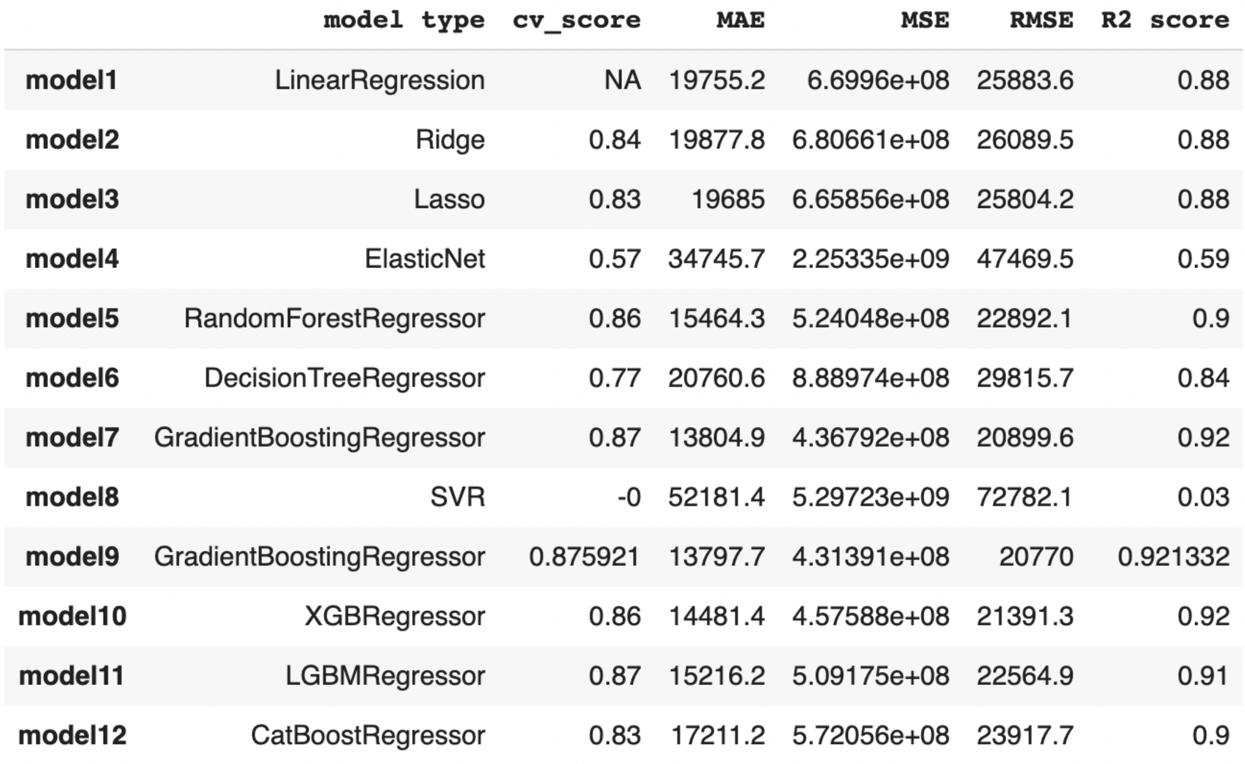
Feature engineering and Pre-processing

A total of 12 outliers were dropped bringing our dataset to 1459 rows. The features age of house, total sq ft of house, total bathrooms, and total porch sq ft were imputed from the given data bringing our training features to 83. After one hot encoding this came to 237 columns to pass to the ML models. The train dataset was split with ¼ allotted to the test set to give the model as many samples as possible. Then the RobustScaler is fit to the train set to scale and standardize the data.

Modeling

This project tested linear models first and advanced to non-linear models to further our study. In total 11 models were used to engineer tests using randomized search and 5 folds to select the best hyperparameters.

Linear models: Linear Regression, Ridge, Lasso, and ElasticNet
Non-linear models: RandomForestRegressor, DecisionTreeRegressor, GradientBoostingRegressor, SVR, XGBRegressor, LGBMRegressor, and CatBoostRegressor



	model type	cv_score	MAE	MSE	RMSE	R2 score
model1	LinearRegression	NA	19755.2	6.6996e+08	25883.6	0.88
model2	Ridge	0.84	19877.8	6.80661e+08	26089.5	0.88
model3	Lasso	0.83	19685	6.65856e+08	25804.2	0.88
model4	ElasticNet	0.57	34745.7	2.25335e+09	47469.5	0.59
model5	RandomForestRegressor	0.86	15464.3	5.24048e+08	22892.1	0.9
model6	DecisionTreeRegressor	0.77	20760.6	8.88974e+08	29815.7	0.84
model7	GradientBoostingRegressor	0.87	13804.9	4.36792e+08	20899.6	0.92
model8	SVR	-0	52181.4	5.29723e+09	72782.1	0.03
model9	GradientBoostingRegressor	0.875921	13797.7	4.31391e+08	20770	0.921332
model10	XGBRegressor	0.86	14481.4	4.57588e+08	21391.3	0.92
model11	LGBMRegressor	0.87	15216.2	5.09175e+08	22564.9	0.91
model12	CatBoostRegressor	0.83	17211.2	5.72056e+08	23917.7	0.9

The GradientBoostRegressor had the best results so it was repeated with grid search to further tune its hyperparameters. This refined GBR has the best overall scores. When comparing how the model performed on predicting the test set vs the train set we find it is slightly overfit. Far more it shows a bias that the model is not predicting the extremes on either the test or train set. The final model is predicting over 50% of the house prices within $9,000. For house price estimates this is not far off.

The competition evaluates submissions on the RMSE between the logarithm of the predicted value and the logarithm of the actual sales price. Our GradientBoostRegressor model scored a 0.14051 on the competition, with a Rank of 6352 at the time of submission. This is within the 50th percentile of the current entries. (The rolling leaderboard only keeps entries for 2 months.)

Further steps

This project could be furthered in many ways. One possibility would be to engineer more features such as polynomials, or combining quality and condition features of many attributes of the house. An automatic Engineering API such as Feature tools could also be considered. Some of the numeric features show some skew and could be corrected with logtransform to increase accuracy. This project explored a wide range of models but only scratched the surface on the more complex models. XGBoost in particular has many more options such as Dart boost and a RandomForest that could be explored.