Machine Learning In-Depth Analysis

The supervised machine learning algorithms that were best suited for my dataset were regression models since the response being predicted, home price, is ordered and continuous. The first machine learning classifier utilized was linear regression and a total of five different feature selection combinations were chosen utilizing different feature selection methods in order to create the best model for home price prediction.

The first feature combination included the four highest correlated features with price as seen on the correlation heatmap. The four highest correlated features were: grade, sqft\_above, sqft\_living, and sqft\_living15. The linear regression model was evaluated based on the coefficients, the mean squared error, the root mean squared error and the R^2 value. The R^2 score was 0.57 and cross validation score was .57.

In order to improve the model further, each of the 19 individual features were evaluated utilizing OLS regression results. The features with AIC/BIC values that showed to be the weakest price predictors were:bedrooms, year, floors, sqft\_basement and yr\_renovated. The features with almost zero R^2 value were: long, zipcode, year and yr\_renovated.The highest R^2 value features were: **grade, sqft\_living, sqft\_above and sqft\_living15.** Finally, the features chosen were: grade, sqft\_above, sqft\_living, sqft\_living15, bathrooms, lat, and view. Utilizing these seven features, greatly reduced the MSE, RMSE and increased the R^2 value to 0.72.

Since adding in additional features utilizing OLS results improved the model, an additional five features were added for a total of 12 features. The 12 total features utilized were: grade, sqft\_living, sqft\_above, sqft\_living15, bathrooms, bedroom, sqft\_basement, view, lat, sqft\_lot, floors, and waterfront and were also based on OLS results. Interestingly, even though there was an improvement, it was very minor with slight reduction of MSE and RMSE and very slight increase in R^2 to 0.73 and 10-fold cross-validation score of 0.73.

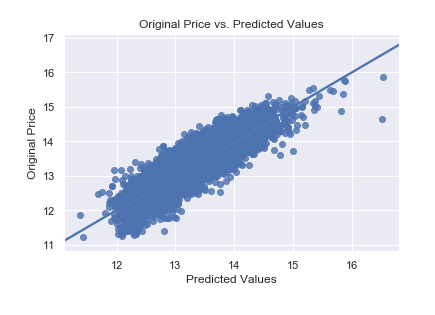
The next feature selection method utilized was sklearn feature selection. Out of the 19 total features, sklearn removed the following 5 features:  **bedrooms, year, floors, sqft\_basement, and yr\_renovated. Consequently, year, yr\_renovated, floors and bedrooms agreed with data exploratory analysis results as not being the best predictors of home price. But in data exploratory analysis, it was unclear whether sqft\_basement was correlated with price or not. The sklearn feature selection removal of sqft\_basement was an important finding since it indicated that sqft\_basement was not a significant home price predictor in King County.** Sklearn feature selection also showed that all **location-related features were important to the housing price prediction model. Utilizing these 14 features showed a further reduction in MSE and RMSE and an increase in R^2 to 0.76**

**Finally, a combination of features from both sklearn selection,  R^2, AIC/BIC, p-value, coefficients, RMSE and MSE, were chosen to see if the model would improve even further. The features chosen were: grade, bathrooms, bedrooms, sqft\_living, sqft\_lot, floors, waterfront, view, condition, yr\_built, zipcode, sqft\_basement, lat, long, and sqft\_living15. Sklearn feature selection method removed year and yr\_renovated. OLS analysis removed sqft\_lot15 and sqft\_above. Combining the feature selections together gave the best model! The RMSE and MSE were further reduced and the R^2 value was 0.77. The 10-fold average cross validation score was 0.77.**

**Next, the random forest regressor model was utilized and gave accuracy on the training subset of 0.98 and accuracy on the test subset of 0.88. The same 15 features also created the best home price prediction model.**

**Next, the gradient boosting regressor model was chosen and gave accuracy on the training subset of 0.94 and accuracy on the test subset of 0.90. This indicated that the gradient boosting regressor model is the best classifier of the three machine learning models for home price prediction. The same 15 features again, also, gave the best result.**

**Finally, the Ridge and Lasso regularization testings were utilized and both gave a model score of 0.77. When the original vs predicted price values were plotted on a scatter plot, the result was highly correlated and indicated that the machine learning model is doing a good job of predicting King County home prices.**



By Roshnee Raval, June 2019