Assignment 4 - Boston Housing Study

Executive Summary

A real estate firm in Boston is interested in deploying machine learning models to help assess market value of residential homes. Deploying the best model will enable the firm to advise clients in optimal listing prices for homes to reduce list time on market. Machine learning models will help reduce manual effort in assessing median value of home prices using limited predicted variables and help with operational efficiencies for the real estate firm.

Research and Design

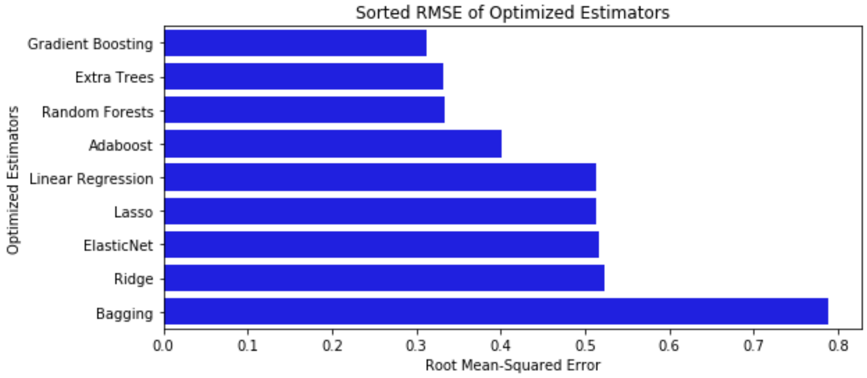
In order to develop machine learning models to predict market value of homes, we utilize data from Boston Housing Study from 1970s; this data set includes 506 observations that reflect the response variable, median house values. The data file includes 13 different predictor variables that influence home values. We were asked to keep all predictor variables except neighborhood in the machine learning model. Our analysis was to utilize the 12 remaining predictor variables and determine the most appropriate conventional regression model using machine learning for assessing market value of residential homes in Boston.

Technical Overview

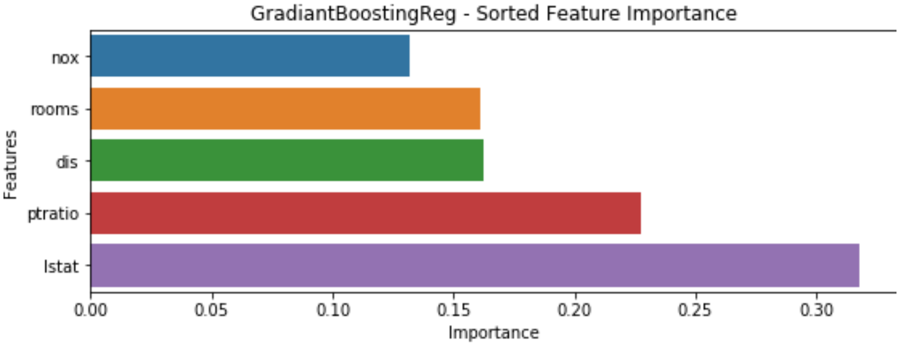
We used Python to perform EDA and run the regression models. EDA helped us identify influence of variables on median home values through correlation matrices and various graphing techniques (e.g.: heatmaps, pairplots, etc.). In addition to the models we explored in week 3, we explored tree models - random forest, bagging, extra trees, along with boosting techniques - gradient, adaboost in week 4 to see if we can improve the model with better predictive power. The approach taken for these models are identical to week 3 where we ran all the models in a cross validation design with all features to assess general performance. We then pruned the features and inputted selected features into all the models using "GridSearchCV" & "RandomizedSearchCV" functions to tune the models and optimize for the lowest root mean-squared error.. To conclude the assignment we looked at feature importance for the best model to see what had the biggest impact on RMSE score. We also ran a voting ensemble experiment to see if we could return an even better score. Voting ensembles work by combining the predictions of several base features built with a given learning algorithm in order to improve generalizability / robustness over a single estimator through averages.

Findings:

For this week, we included tree and boosting models in addition to the models tested last week. Finally, we ran voting regressors, however it did not improve our score despite a very lengthy processing time. As seen in the chart below gradient boosting produced the best result but all tree / boosting models tested, resulted in significant improvements over all the linear models tested last week when comparing RMSE scores.



To understand which explanatory variables are most important in predicting home price, the following chart shows the Feature Importance for the top performing model - Gradient Boosting. We notice that “lstat” and “ptratio” are the most important followed by “dis”, “rooms”, “nox”.



Conclusion and Recommendations:

After reviewing all the models, we can clearly notice that tree models tested this week are better suited for this data set and improved our scores. Our final recommendation to management would be to utilize gradient boosting regression models to predict housing values in the Boston area. The model could see some slight improvements if we wanted to run a more exhaustive model tuning process but due to time and CPU limitations we kept it to a randomized shuffled subset.