Assignment 3 - 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 like size of home, school district, location etc, 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 rows of data that reflect the response variable, median house values. The data file includes 13 different predictor variables that influence home values, where we were asked to keep in everything except neighborhood in the machine learning model we build out. 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.). The feature average room size seems to have the strongest correlation to median home price. EDA also assisted in determining need for scaling the dataset and if transforming dataset would improve the accuracy of models (log home value). A standard scaler was given in the base code, but we also experimented with others to improve our models.

Four different regression models were run (linear, ridge, elastic net, and lasso models) within a cross-validation design and evaluated using root mean-squared error (RMSE) as an index of prediction error. We first ran all models with every available feature to assess general performance. We then pruned the features and inputted selected features into the best fitting models. Next, we included gradient descent run which goes through the parameters iteratively looking to minimize the RMSE but the small size of the data limited its effectiveness. Finally, we performed an exhaustive search to optimize our models' parameters. The finalized models are then tested against a split off test data set.

Findings:

EDA: The standard scalar was given in the base code, but we also looked at normalizer and power transformer to see which scalar would provide the best model and help normalize our dataset. The Power transform scalar looked like it did the best job from histograms at normalizing the data and the features that stood out within the scaled dataset impacting median home value were air pollution concentration, average # rooms per home, weighted distance to employment centers, pupil / teacher ratio in public schools and % of population of lower socio-economic status.

After initial review of the four models with all the variables included and the standard scalar given, we notice that the linear and ridge models outperform the other models by reviewing the RSME scores. The best model was the linear regression by a slim margin against an optimized ridge regression in both the average mean of 5 cross validation runs and standard deviation between the runs. The log median house value transformation slightly improved the accuracy of the full model. After feature selection and transformation using power transformer, we saw similar results with the linear and ridge regression producing almost identical scores but lasso and elastic were significantly worse. Using the log median value on the home decreased performance of the model as the variables were already slimmed down to key features and normalized to fit the data through power transform scalar.

Conclusion and Recommendations:

Our final recommendation to management would be to utilize the linear regression model, track predictions, and continue to adapt the models as more data becomes available.

To improve the final model, we also recommend running multiple iterations on the key predictor variables one by one and evaluate if any result in further improvements in RMSE scores. For next week's assignment will be utilizing the same dataset and will explore a few other regression modeling methods to see if we can further improve our prediction error scores. Will also look at techniques from next week's learning with random forests and gradient boosting to see if we can enhance our model.