Christopher Rico
MSDS 422
Assignment 4
10/20/19

Introduction

To assess the value of real estate, many different factors must be taken into consideration. Several factors, such as home size, neighborhood, and school district are typically strong predictors of home value. However, many of these valuation metrics are time-consuming to research manually.

By leveraging a machine-learning backed model to predict home value based on data collected from home listings, it may be possible to reduce the amount of research needed to estimate home value. This would reduce the operational overhead of the brokerage firm and allow for more accurate home value predictions when machine learning predictions can complement more traditional valuation techniques.

Research Design and Statistical Methods

The dataset provided is a collection of about 500 census tracts within the Boston metropolitan area in 1978. The original objective of this dataset was to examine the effect of air pollution on home value while controlling for 13 explanatory variables such as proximity to the Charles river, home age, highway access, etc.

The ‘neighborhood’ variable was removed from the dataset, which leaves 12 explanatory variables from which to build and test several linear regression machine learning models. The ultimate goal of this work is to compare different machine learning models and recommend the best one to a real estate brokerage firm who wishes to augment their conventional methods for assessing market value of residential real estate.

Programming Work

This analysis was performed in the cloud using a Google Colaboratory notebook. Scikit-Learn was the primary modeling environment. Several linear regression models from the scikit-learn library were trained on the sample data: linear regression, ridge regression, lasso, elastic net, random forest, and gradient boosted. Each of these regressors have various characteristics that make them suitable for different types of datasets, so to ‘level the playing field,’ the data was pre-processed by running it through a standard feature scaler that helped normalize the input variables.

Christopher Rico
MSDS 422
Assignment 4
10/20/19

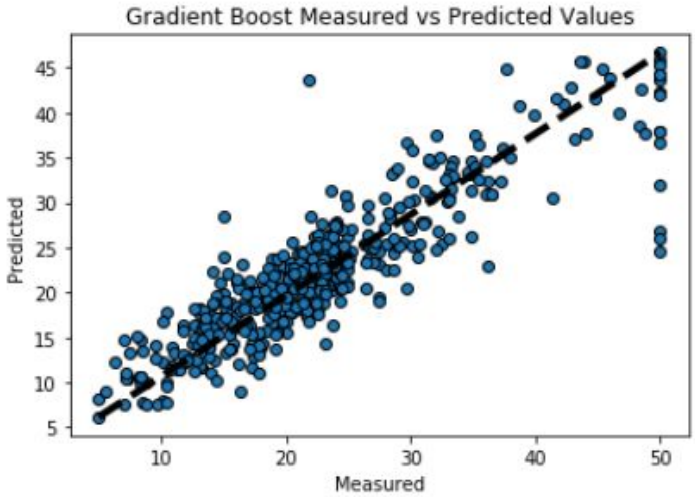
Each regressor was then tested for its ability to predict home values, and validated using a multiple k-folds method. Root mean square error was used as a metric to compare the accuracy of the regressors.

Results and Recommendation

My recommendation to the firm is to use a gradient-boosted regression model, as this model trained on the sample data gives passably accurate predictions regarding home price. If the firm wishes to target homes with high resale value it should target home with a large number of rooms and higher socioeconomic neighborhood makeup, as these are the strongest predictors of home prices.

Model	Train RMSE	Test RMSE	Train Score	Test Score
Linear	4.569606	5.990063	0.746679	0.383111
Ridge	4.569608	5.988431	0.746679	0.383584
Lasso	4.621793	5.945974	0.740780	0.410406
Elastic Net	4.653609	5.763485	0.737233	0.447256
Random Forest	2.982805	5.019145	0.891739	0.578182
Gradient Boost	1.207231	4.198912	0.982195	0.702870

I would also suggest that the brokerage firm expand its dataset to be able to train a more accurate model. 500 points is, frankly, a rather small set of data to make meaningful predictions from.



A shared version of the interactive Jupyter notebook used to run this analysis can be found at:

https://colab.research.google.com/drive/1mslPkyF7XzoGxnpbTRdBjRtLEoFDlTEY