Brandon O'Briant
PREDICT 422-DL Practical Machine Learning
Assignment 3

Predicting Home Market Value: Regression Model Comparison

Introduction

Developing a home market valuation is an important step that, if done correctly, can lure in buyers and increase profits. To appropriately conduct home valuations traditional techniques work well, but can be time consuming and labor intensive. Combining traditional valuation with machine learning regression for residential real estate market valuation will assist in better valuations while promoting reduced turnaround times. Regression machine learning models considered in this study are: Elastic Net, Lasso, Linear, and Ridge.

Exploratory Data Analysis Methods

This evaluation of machine learning regression models to estimate real estate market value utilized census tract data from the Boston Housing Study. Data was imported as *boston_input* and contains 506 observations (rows) and 14 attributes (columns). Non-values were not discovered in the data. Descriptive statistics were calculated for *boston_input* data, along with histograms, boxplots, a scatter matrix, and a correlations matrix. *Neighborhood*, a categorical attribute, was removed and all numeric attributes were kept, ultimately resulting in a model data set, *model_data*, with 12 explanatory variables and one response variable *mv* (log median value of homes in thousands of 1970 dollars). Histogram and density plots were constructed for *model_data.mv*. Removal of outliers was not performed since the raw observations related to a meaningful scale. The root mean squared model evaluation metric was to gauge the performance of each model and allow for comparisons between each model's performance.

Brandon O'Briant
PREDICT 422-DL Practical Machine Learning
Assignment 3

Overview of Programming Work

Python was used exclusively for the analysis of the Boston housing market valuation data, including the use of the following packages: Pandas and Numpy for data handling and data visualization. Boston housing market data was presented to us as a CSV file and loaded into the program using Pandas. Subsets of data were obtained by dropping one categorical variable, *Neighborhood*. Descriptive statistics were calculated, via built in functions in the Numpy environment. The Pandas DataFrame, *boston*, was put into numpy arrays, *prelim_model_data*, so that it could be used within the Scikit Learn environment. Model data was obtain standardizing *prelim_model_data* using SciKit Learn StandardScaler(). All four regression models, Elastic Net, Lasso, Linear, and Ridge, along with a ten-fold K-fold cross validation design using root mean-squared error metric for performance evaluation were implemented within SciKit Learn environment. All the visual graphs were saved to separate pdf files and all printed text results were saved as separate .txt. files.

Results and Recommendations

The results from the 10-fold cross-validation in standardized units informs us that both the Linear and Ridge regression models performed the best with 0.50 and 0.50, respectively, under the root mean-squared error metric. Elastic Net regression model coming in next with a reporting 0.52 root mean-squared error metric and Lasso regression model taking last place with a root mean-squared error of 0.54. Therefore, I recommend from the regression models presented the Linear regression model using the log mean home market valuation for further valuation endeavors until other methods can be explored and compared.