MSDS 422 Practical Machine Learning
John Moderwell
Assignment 3

Summary

The Boston Housing Study was a study of 506 census tracts in the Boston metropolitan area. The purpose of the study was to analyze the effect of air pollution on housing prices. However, the data can also be used to calculate valuations for the residential real estate market. MSDS 422 students evaluated several machine learning regression models as they were applied to the dataset. The models included in this study were Elastic Net, Lasso, Linear and Ridge.

Research Design

The dataset consisted of 506 observations and 14 variables. These variables represented factors that would influence the valuation of a residential house as well as descriptive characteristics. They include: name of the Boston neighborhood (neighborhood), median value of home in 1970 dollars (mv), air pollution (nox), crime rate (crim), percent of land zone for lots (zn), percent of business that is industrial (indus), proximity to Charles River (chas), average number of rooms per home (rooms), percentage of homes built before 1940 (age), distance to employment centers (dis), accessibility to highways (rad), tax rate (tax), student/teacher ratio (ptratio) and percentage of population of lower socio-economic status (lstat). The jumpstart code provided to students dropped the neighborhood variable, standardized scores across all variables and arranged the data into a numpy array. As part of the exploratory data analysis, histogram and density plots were created. Descriptive statistics were calculated as well. To apply machine learning methods to the data, median value of the house (mv) was used as the

response variable and the other 12 variables were used as the explanatory variables. The root mean squared error (RMSE) evaluation metric was used to compare the regression models.

Programming Overview

For this assignment, the programming language Python was used and executed in the Jupyter Notebook environment. Several packages were utilized throughout the project including: Pandas and Numpy for data handling/manipulation. This included the loading of the 'Boston.csv' file and subsequent conversion to a Numpy array. Descriptive statistics were calculated using Numpy built-in functions. The package Seaborn was used to visualize data including histogram/density plots as well as a correlation matrix. The machine learning portion of the assignment was done entirely using the Sci-Kit Learn package. This included the regression methods (LinearRegresion, Ridge, Lasso, ElasticNet), metrics to evaluate the models (mean squared error, R2 score) and KFold for cross validation. The sqrt function was imported from the math package to calculate root mean-squared error.

Recommendations

After training the data to the regression models and then testing it using 10-fold cross validation, it was observed that on average Lasso Regression resulted in a lower root mean squared error (.56051). Linear Regression was a close second with an RMSE score of .56194. Ridge Regression and ElasticNet Regression followed up with RMSE scores of .56808 and .58738. Therefore, it is recommended that Lasso Regression be used to determine residential house value.