2018FA_MSDS_422-DL_SEC58

Practical Machine Learning Assignment 4

Prabhat Thakur

1. Problem Definition

The Boston Housing Study is a market response study of sorts, with the market being 506 census tracts in the Boston metropolitan area. The objective of the study was to examine the effect of air pollution on housing prices, controlling for the effects of other explanatory variables. The response variable is the median price of homes in the census track.

Evaluate different regression machine learning models and recommend best model to estate brokerage firm for predicting median home price for new housing data.

2. Research Design and methods

The dataset has 506 records of Boston hosing data with 14 variables. Neighborhood variable has been excluded from the study. I evaluated performance and predictive power of a different regression models trained and tested on collected data. A model trained on this data that is seen as a good fit could then be used to make certain predictions about a home’s median value.

As part of this assignment, I added Random Forest Regression model along with Linear, Ridge, Lasso, and ElasticNet regression machine learning models for this study. There are 12 explanatory variables and 1 target variable ‘mv’. 10 fold k-fold cross-validation design is used for evaluate accuracy of these methods.

Python Scikit Learn package was used for conducting this research.

3. Implementation and Programming:

Python programming language and different python packages like numpy, pandas and matplotlib for data preparation and visualization are used for this analysis. Various methods from sklearn package is also used for building machine learning regression models and evaluation.

The first step is to read the hosing data from comma delimited txt file into python data-frame object. Drop neighborhood variable data and convert remaining variables to two dimensional array object so that it can be used in sklearn package. Scale features using sklearn StandardScaler function.

Four linear regression models: Linear, Ridge, Lasso, and ElasticNet and Ensemble model RandomForestRegressor are initialized and passed through 10 fold k-fold cross-validation to evaluate their performance using Root mean-squared error values. Additional code is added to perform exploratory data analysis and visualization. Used GridSearch view to find best parameters for all 5 models and changes the parameters in the model to increase the model performance.

4. Findings and recommendations

From the data we can see that median value has been capped at $50,000. From correlation matrix and scatter plots we can see that Average number of rooms per home (rooms) and Percentage of population of lower socio-economic status (lstat) variables are the most influential followed by Pupil/teacher ratio in public schools (ptratio) and Percent of business that is industrial or nonretail (indus) variables. From RandomForestRegressor modal grid search view analysis following 6 explanatory variables are most important in predicting home prices listed in ascending order of their importance.

rooms - Average number of rooms per home

lstat - Percentage of population of lower socio-economic status

dis - Weighted distance to employment centers

nox - Air pollution (nitrogen oxide concentration)

crim - Crime rate

indus - Percent of business that is industrial or nonretail

Recommendation: Based on the five regression models evaluated, real estate agency should use **RandomForestRegressor machine learning model and above 6 variables** to predict market value of residential real estate with typical prediction error of $3904. The house median price from 1970 is definitely not a good basis to predict current housing prices, inflation should be factored in.