MSDS 422: Practical Machine Learning

Professor Anil D. Chaturvedi

Assignment 4: Random Forests and Gradient Boosting

Tiffany Duong

Summary

Supervised machine learning methods such as linear and tree-based models are a great tool when it comes to understanding the relationships among the target and feature variables. In this assignment, a real estate brokerage firm is looking for advice in its attempt to employ machine learning methods that would complement conventional methods for assessing the market median value for housing in the Boston metropolitan area.

Research Design

To give advice on what machine learning methods to employ, we must explore the data. The predictive variable is *mv* (median value) and there are 12 features that can be used to help predict the predictive/target variable. From initial data explorations, we found out that the number of rooms, the percentage of land zoned for lots, the weighted distance to employment centers, and whether or not the house is located on the Charles River have positive correlations to the median housing value, while the percentage of the population being of a lower socioeconomic status had a high value of negative correlation.

Technical Overview

Analysis on the survey data was performed using traditional statistical methods and Python, a high-level general program language. Several Python packages were used, including NumPy, pandas, matplotlib, and seaborn, and scikit-learn. Scikit-learn was heavily used when setting up the linear and tree-based models as well as performing the cross validation to see model performance. The predictive variable we will be looking at is the median housing variable. There were 12 features from the dataset that will be used as explanatory variables for the linear and tree-based models. The most basic linear model OLS will be used as well as regularized

models (ridge regression, lasso regression, and ElasticNet. Regularized models are a good way to prevent and reduce overfitting of the model. We will also use Random Forests and Gradient Boosting as tree-based models are power machine learning models that have several advantages such as being interpretable and they perform very well on large datasets. We will be comparing to see which machine learning model performs better and to conduct the comparison, we will evaluate the models using k-fold (5 folds in this case) cross validation, using the root mean square error (RMSE) as an index of prediction performance. In addition, feature selection will be used along with Gini impurity to see the important features that predict *mv*.

Results

When looking at the performance scores, Random Forest (0.55) and Gradient Boosting (0.64) outperformed the other models, while Decision Trees performed poorly. In terms of RMSE, the models shared similar results, with Decision Tree having the highest RMSE at (0.75) and Random Forest (0.51) and Gradient Boosting (0.48) having lower RMSE than the rest. Gradient Boosting had the lowest RMSE and highest performance score, so it was employed to the full data set. The model score was 0.99 and the RMSE was 0.002. In Gradient Boosting, each single tree uses the tree before to recursively correct itself, therefore it is extremely powerful, and it is less prone to overfitting – this is my recommendation of a modeling method. Furthermore, feature selection was done using Gradient Boosting and the features that were considered most important were the *rooms*, *dis*, *lstat*, *age*, and *crim*.