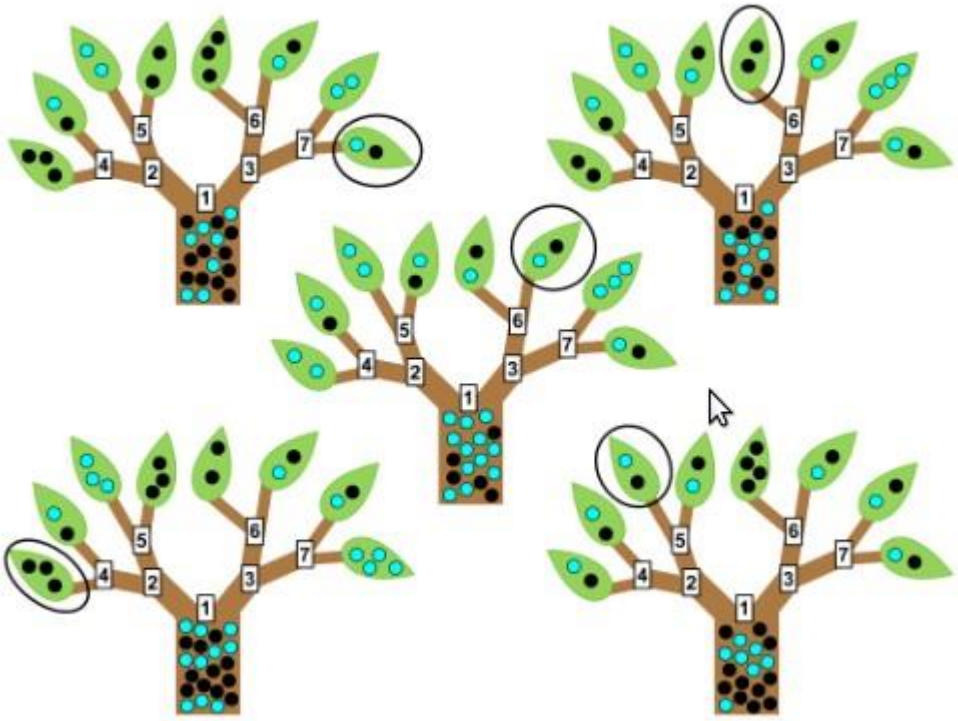

EVALUATING RANDOM FORESTS

Boston Housing Study

ASSIGNMENT #4
Kimberly Kaminsky
Predict 422 Section 59

Introduction

The purpose of this study was to advise a real estate brokerage firm in its attempt to utilize machine learning for assessing the market value of residential real estate. This study extended a previous study that compared four modelling techniques, (Linear Regression (LR), Ridge Regression (RR), Lasso Regression (LaR), and Elastic Net (EN)). This paper added a fifth technique, Random Forest (RF). The best hyperparameters from the first study were applied in this round, along with an investigation of hyperparameters from the Random Forest. The best models for each technique were compared utilizing the root mean squared error (RMSE). A cross validation design was used in order to select a method providing good generalization.

Data Exploration and Visualization

An extensive data exploration exercise was performed for the first study so is not being reported for this study. However, it is worth repeating that for this study the explanatory variables were all converted to a standard scaling and the response variable was log transformed to correct a skewness issue. Also, some of the variables had outliers, such as the rooms variable, shown in Figure 1. Since the object of this study was not to optimize the solution but to compare machine learning methods to traditional methods no effort was made to clean up the data, especially since RF is known for its ability to work with data "out of the box".

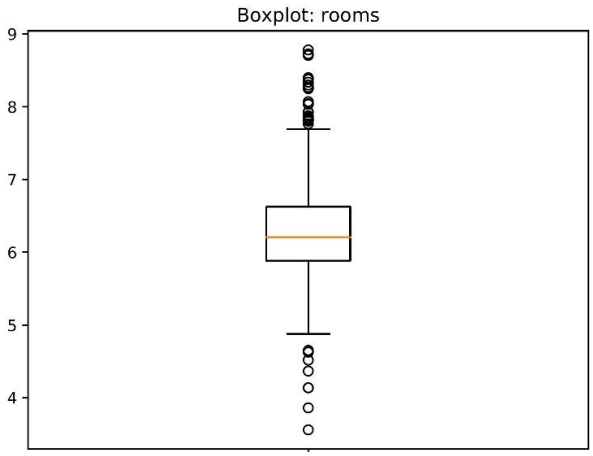

Figure 1: Boxplot of Rooms Variable

Research Design and Modeling Methods

In the last study, the RR, LaR, and EN models had performed differently depending on their alpha value. The best models for RR, LaR and EN were selected in the previous study. In this

study, several RF hyperparameters were selected for analysis to find the best RF: n_estimators (number of trees), max_features (maximum number of features per tree), max_depth (maximum depth of the tree), bootstrap (Boolean indicating the use of boostrap sampling), and warm_start (Boolean indicating if previous best solution should be iteratively incorporated). The values selected from each parameter are as follows: n_estimators (12, 100, & 506), max_features (1, log2, & auto), max_depth (none, 5 & 8), bootstrap (true & false), warm_start (true & false). This created 13 different parameter settings, with a total of 108 scenarios, which were run using a 20-fold cross validation. The best model out of these 108 runs utilized the following parameter settings: max_depth = 8, max_features = log2, warm_start = true, bootstrap = false, n_estimators = 12. The RMSE for this model was .1885 and the worst model was .265.

After selecting the RF model, it was rerun, along with the best four models from the original study, and verified utilizing a k-fold, (20 fold), cross validation. The RMSE was used as the measure to compare the four methods. Table 1 displays the method and the RMSE value.

Table 1: Comparison of Regressors

Method	Root mean-squared error
Linear_Regression	0.471614
Ridge_Regression	0.46553
Lasso_Regression	0.470976
ElasticNet	0.468129
Random_Forest	0.143454

The RMSE values range from .1435 to .4716. The RR, LaR, and EN models beat the LR by small amounts. The RF model beat out all the other models by a large degree.

Another advantage of RF is that it easily shows feature importance. Figure 2 provides a bar chart showing the most important features. Lstat, the percentage of population of lower socio-economic status contributed the most importance to the model closely followed by the number of rooms. The Boolean variable indicating if the house is on the Charles river contributed the least.

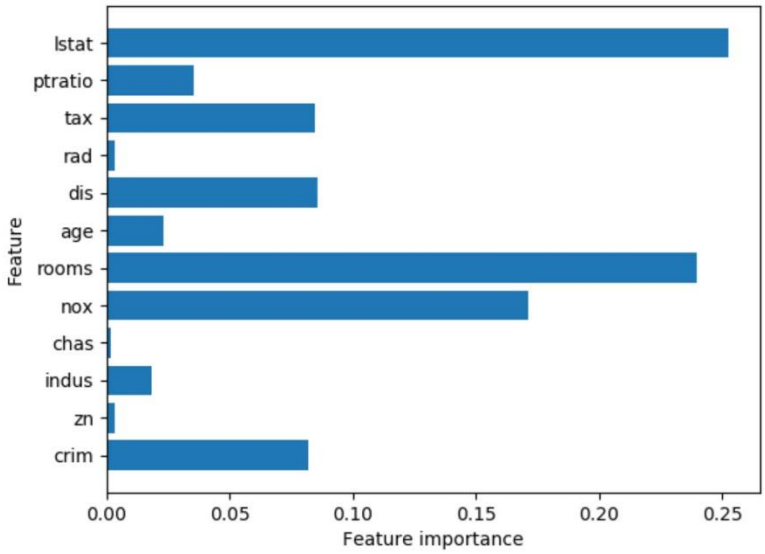

Figure 2: Boston Housing Data Feature Importance

Recommendations

The results above show the power of RF. Could the other models have been optimized through feature selection and variable transformation? Maybe, but it would take a lot of work. Also, it is likely that the RF could also be further optimized through further tweaking of the parameters. The real beauty of this approach is the accuracy of the results without any time spent on data cleanup. Since the method was run through a 20-fold cross validation the model should be generalizable as long as the new data is in sample. When the model was run against the entire dataset the RMSE was .1389, which was an improvement over the test results indicating that the method is indeed robust.