Machine Learning Project

Decision Tree Regressor Using R

Min Seong Kim

January 2017

1. Introduction

This report analyzes the Boston Housing Prices dataset. The data was collected from the "MASS" library in R, and a decision tree "regressor" algorithm was applied to forecast a housing price; this method was chosen over a "classification" model, as the data contained 13 mixed features (11 continuous variables and 2 categorical variables) and the predicted outcome is not binary. The total number of houses was 506.

```
      crim                zn             indus        chas          nox             rm       
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850   Min.   :3.561  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490   1st Qu.:5.886  
 Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380   Median :6.208  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547   Mean   :6.285  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240   3rd Qu.:6.623  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710   Max.   :8.780  

      age              dis              rad             tax             ptratio      
 Min.   :  2.90   Min.   : 1.130   24     :132   Min.   :187.0   Min.   :12.60  
 1st Qu.: 45.02   1st Qu.: 2.100   5      :115   1st Qu.:279.0   1st Qu.:17.40  
 Median : 77.50   Median : 3.207   4      :110   Median :330.0   Median :19.05  
 Mean   : 68.57   Mean   : 3.795   3      : 38   Mean   :408.2   Mean   :18.46  
 3rd Qu.: 94.08   3rd Qu.: 5.188   6      : 26   3rd Qu.:666.0   3rd Qu.:20.20  
 Max.   :100.00   Max.   :12.127   2      : 24   Max.   :711.0   Max.   :22.00  
                                   (Other): 61                                  
     black            lstat            medv      
 Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median :391.44   Median :11.36   Median :21.20  
 Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

Figure (1) A Basic Statistical Analysis of the Boston Housing Prices Dataset

In the dataset, the variables "chas" and "rad" are categorical, while the remainder are continuous. The "medv" variable is the target variable.

2. Evaluating Model Performance

We split the dataset into a training set and a test set and conducted a cross-validation (10-fold) in order to tune the model. If we had failed to divide the dataset and had instead built a model based on all the data, we would have run the risk of creating an overfitting issue; it should be noted that such models can have solid predictive power in terms of in-sample data, but can demonstrate poor performance when applied to out-of-sample data. Additionally, we used the **Predicted Mean Squared Error (PMSE)** factor as a performance measure and compared these errors across training and test datasets by gradually increasing the size of the training datasets.

3. Cross-Validation

This study utilized a 10-fold cross-validation method; accordingly, the dataset was evenly split into 10 sub-datasets. One was considered a testing dataset and the remainder of the sub-datasets were deemed training samples. Then, we conducted 10 iterations of each step in order to tune a predictive model. The cross-validation function allowed us to conduct a grid search of an entire dataset and prevented us from generating any biased results that may have been caused by the random selection of data.

Additionally, we coded a function of cross-validation without using any built-in packages in R. This practice was adopted mainly because it allowed us to develop a better understanding of how the cross-validation process works and gave us more flexibility to test datasets.

4. Analyzing Model Performance

In this portion of the study, we created charts using the "ggplot2" library in R and compared the predictive powers of the trained models based on the various levels of max depth in the decision tree.

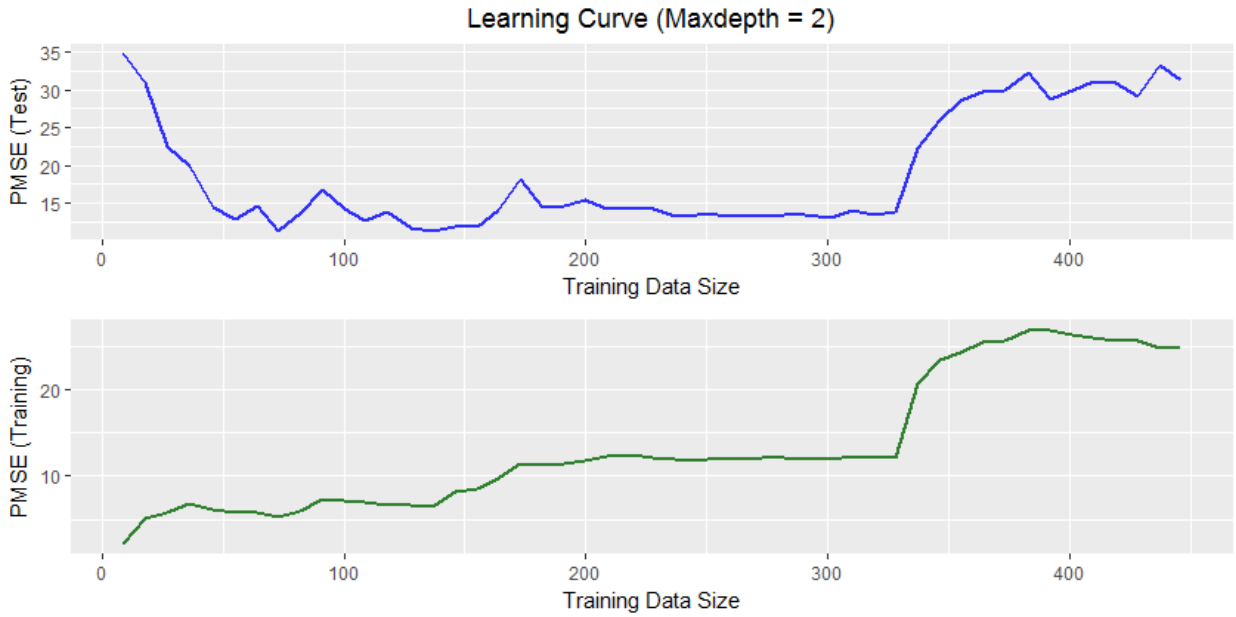

Figure (2) Decision Tree Regressor: Learning Curve (Max Depth = 2)

Figure (2) illustrates that both test and training errors are very large when the training data size is relatively small. Although test errors decrease with a growing training data size, the training errors remain high. This example demonstrates the problem of underfitting. The model is not well-tuned because its complexity is not sufficient to fit the training samples; thus, it fails to capture the pattern recognition of the underlying dataset. Therefore, the trained model has little predictive power.

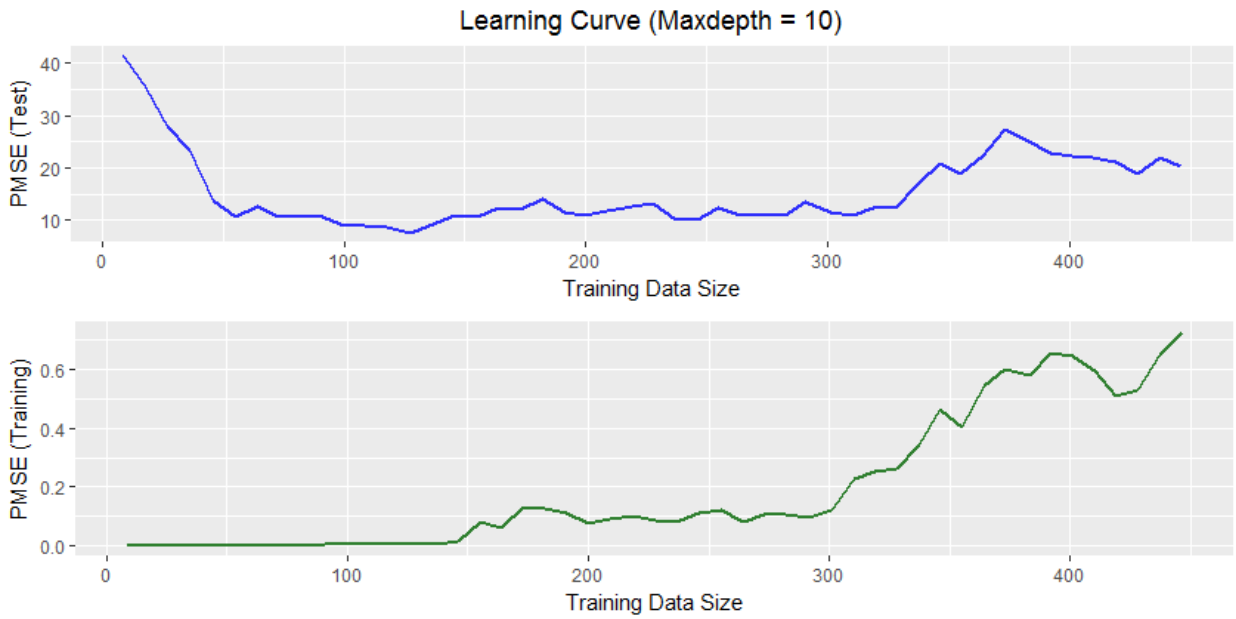

Figure (3) Decision Tree Regressor: Learning Curve (Max Depth = 10)

Figure (3) exemplifies the fact that the trained model demonstrates overfitting, given that training errors are very small and test errors are relatively high in all training data sizes. This overfitting issue arose when we excessively fit the model; thus, the model was able to adapt to the noise of the training data which is unnecessary. An examination of Figures (2) and (3) can lead to the legitimate assumption that the optimal max depth lies between 2 and 10.

Model Complexity

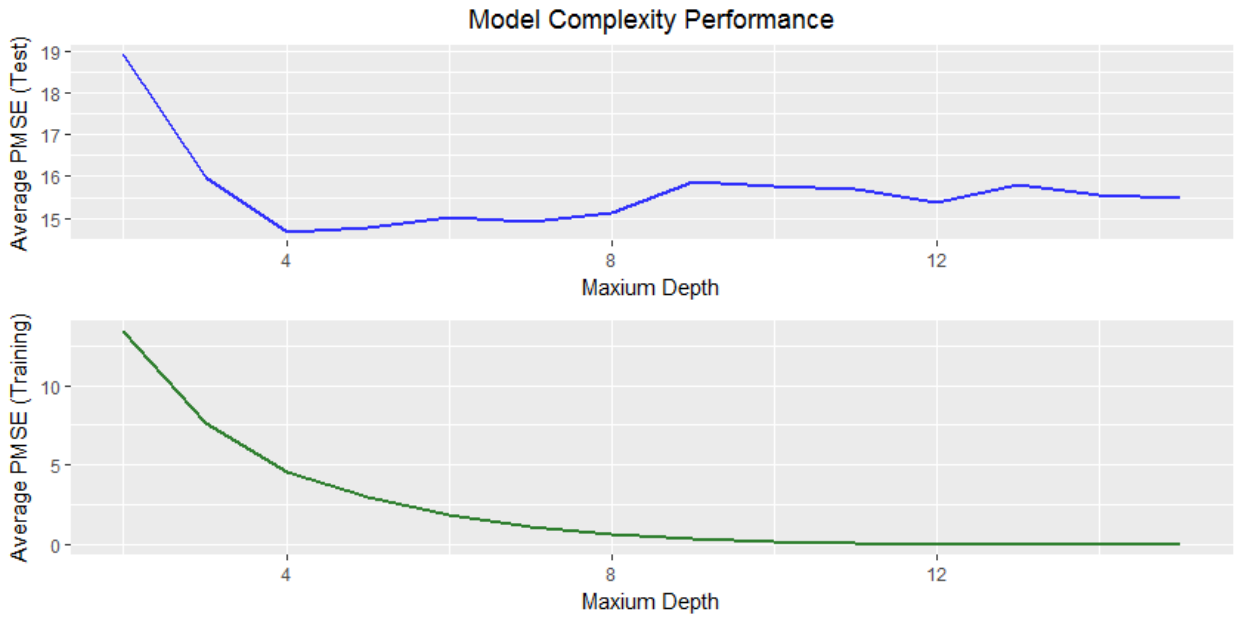

Figure (4) Decision Tree Regressor: Model Complexity

Based on Figure (4), we can conclude that the model with a maximum depth of 7 is best suited to predict a housing price. First, in this model, the training error continues to decline when the maximum depth is equal to 7; additionally, the test error is also at its lowest level. Moreover, the test error begins to rebound after a maximum depth of 7, indicating that this is the optimal tree depth for the predictive model.

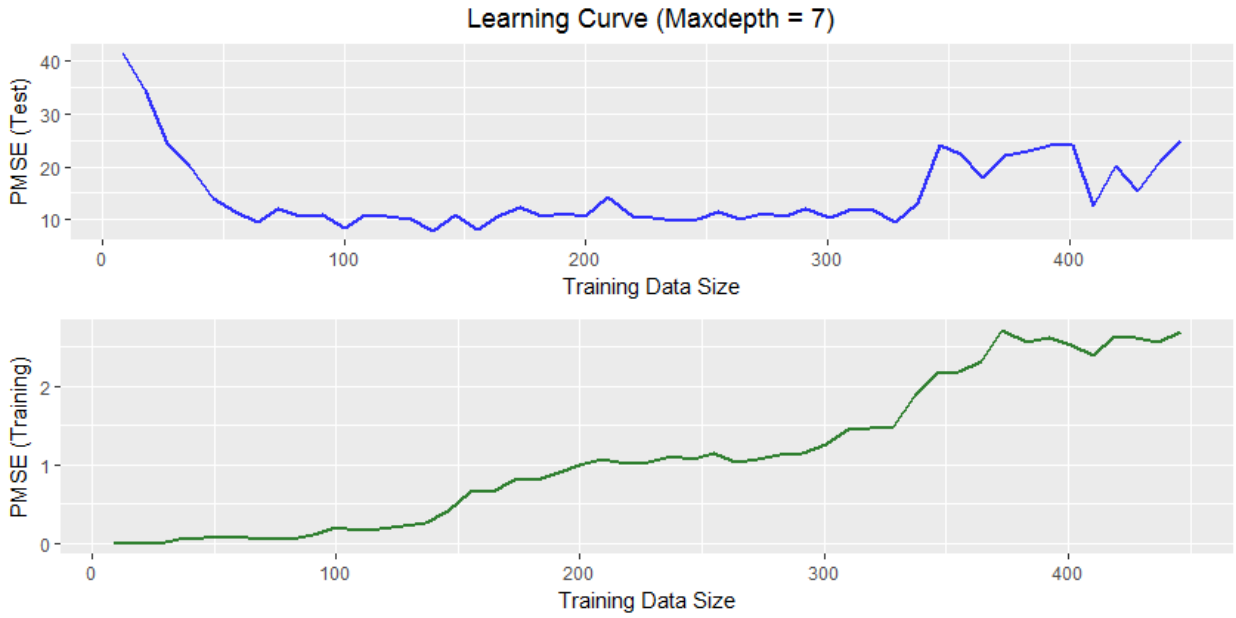

Figure (5) Decision Tree Regressor: Learning Curve (Max Depth = 7)

5. Conclusion

In this report, we constructed a predictive model to forecast a housing price in Boston by implementing a 10-fold cross-validation process and investigating the performance of the trained models according to varying levels of maximum depth in a decision tree. We also coded the function in R, which turns out to effectively capture the cross-validation process. Future researchers might consider applying the "Decision Tree Regressor" algorithm to other datasets, such as those containing financial, health, and demographics statistics with diverse and mixed features.

6. Appendix

Attribute Information:

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's

Resource: https://archive.ics.uci.edu/ml/datasets/Housing