Final Project

Joshua Firestone

May 05, 2020

Abstract

For this project I am doubling up on a project for my *Supply Chain Analytics* course. It uses Boston housing data which is a fairly common dataset that is often used as a sample case (Harrison Jr and Rubinfeld 1978). For the class project our task was to utilize multiple models and methods that we have learned this semester to analyze the data. My portion was regression trees and random forests, so that will be all I include here.

Contents

1 Overview

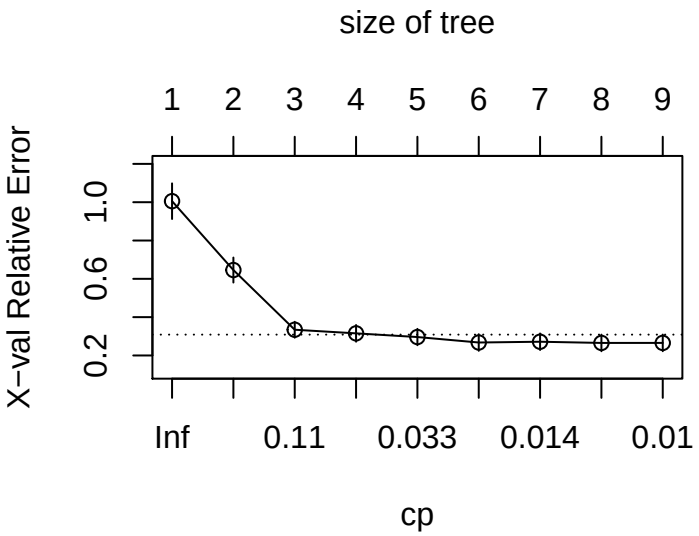
For this project I will be using regression trees and random forests to analyze a dataset conaining housing data from the Boston area. Overall, the goal is to compare model performance and arrive at one with the greatest predictive power.

2 Regression Trees

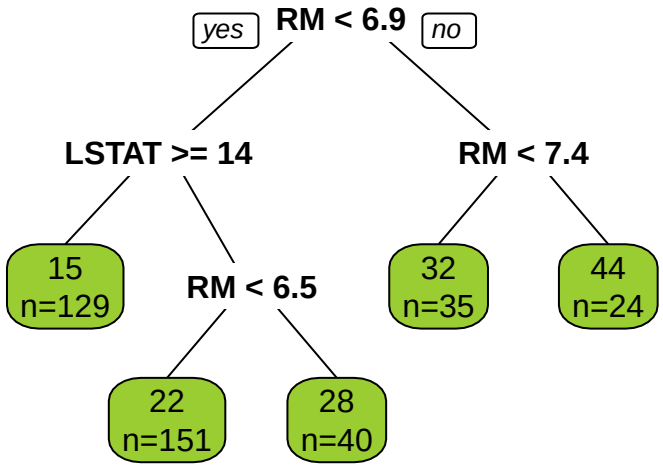
Regression Trees use recursive splits to create a series of internal nodes that result in decision pathways predicting some outcome variable. They split along variables in a dataset using a series of true/false criteria, either putting an observation to the left or right. Ultimately, the trees end in "leaves" that represent this prediction.

2.1 Model One

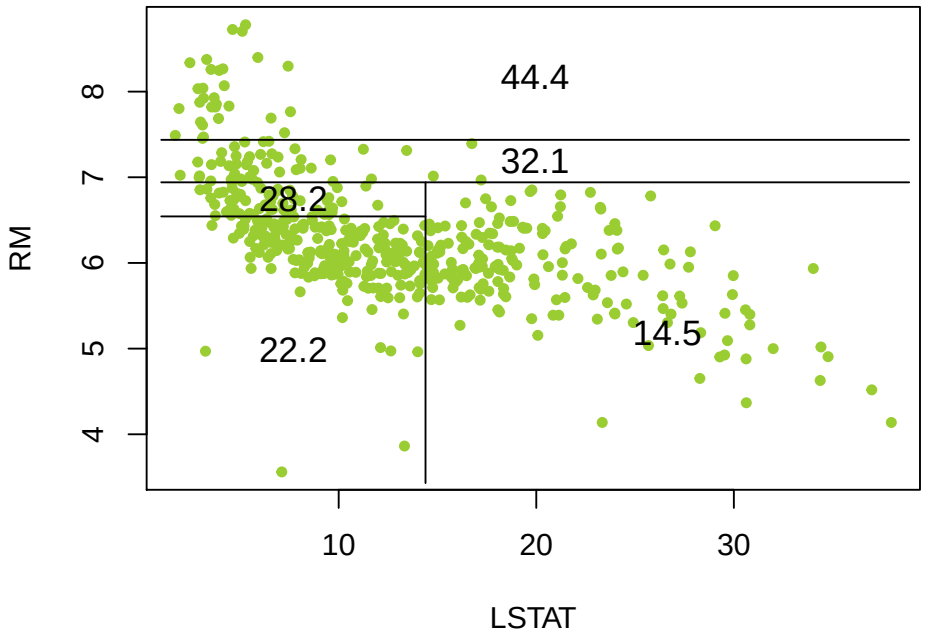
Using `rpart` we created a tree with all of the dataset's variables. But this results in a tree with many internal nodes. In order to scale this down we can "prune" our tree. One method for pruning a tree is to use a complexity parameter plot and choose the prune at the point which the graph falls below the x-value relative error line.



In this case our graph dips below the relative error line at about 0.035 so this is the complexity parameter we chose to create our first pruned tree, which results in the tree below. This tree produces an RMSE of 5.12.

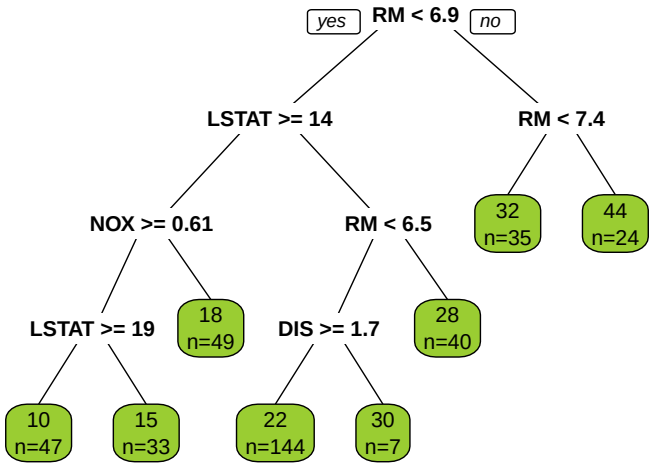


Regression trees are pretty straighforward and intuitive in their interpretation already. But it can also be helpful to think about how the decisions and partitions are being crafted within the space of the data observations. Using a partition plot like the one below we get a glimpse of how the splits from the tree above are being made over the data points. Each of the sections in this graph correspond with the terminal nodes from our previous regression tree.



2.2 Model Two

The first model seemed pretty basic and only included two of the thriteen total variables. So, we thought we might be able to improve upon this tree if we used a different pruning criteria. And indeed this was the case! For this second tree we used a prune that minimized the relative x-error value from the complexity parameter table. Using this as pruning criteria we achieved an RMSE of 4.74.

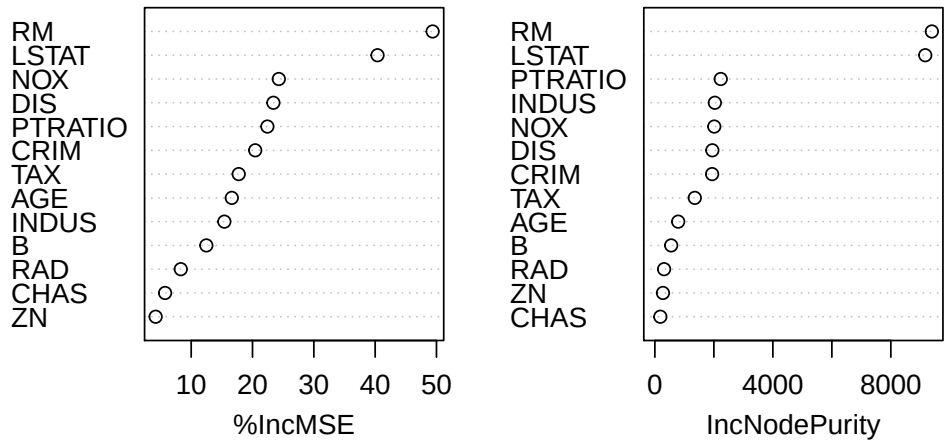


Of note in this tree is that we went from four to seven internal nodes and we added in two additional variables, *NOX* and *DIS*.

3 Random Forests

As we know, we can typically do even better than a regression tree with random forests. A random forest builds hundreds or even thousands of regression trees, randomly making the splits with different variables and at different points, then taking the best average of all of these trees. For our model we used the `randomForest` package and specified the number of trees to be 1000. This model performed the best of all of our models, achieving an RMSE of 3.16.

Although a random forest does not produce a readily digestible output similar to regression trees, what we can get is a description of variable importance. Using `varImpPlot` we can get a sense for how the forest is creating its decision pathway and discover which variables have the most weight in the model.



Here we can see two importance plots, one ordering by `%IncMSE` and the other by `IncNodePurity`. They are similar in the first two variables they selected, but then begin to differ as they move through the decision pathway. Notably, the `%IncMSE` plot has *NOX* and *DIS* as its next two variables, which aligns with our second regression tree from above.

4 Conclusion

For the entire project we took a preliminary look at our data using *Tableau* to inspect any early relationships. From there we utilized **Linear Regressions**, **Regression Trees**, **Random Forests**, and **Neural Networks** to further explore and analyze our data and ultimately come up with a best model. In this case, we found that **Random Forests** outperformed all the other models and would therefore be the method of choice in making predictions about this dataset.

References

Harrison Jr, David and Daniel L Rubinfeld (1978). "Hedonic housing prices and the demand for clean air". In: *Environmental Economics and Management* 05.01, pp. 81–102. URL: https://www.sciencedirect.com/science/article/pii/0095069678900062.