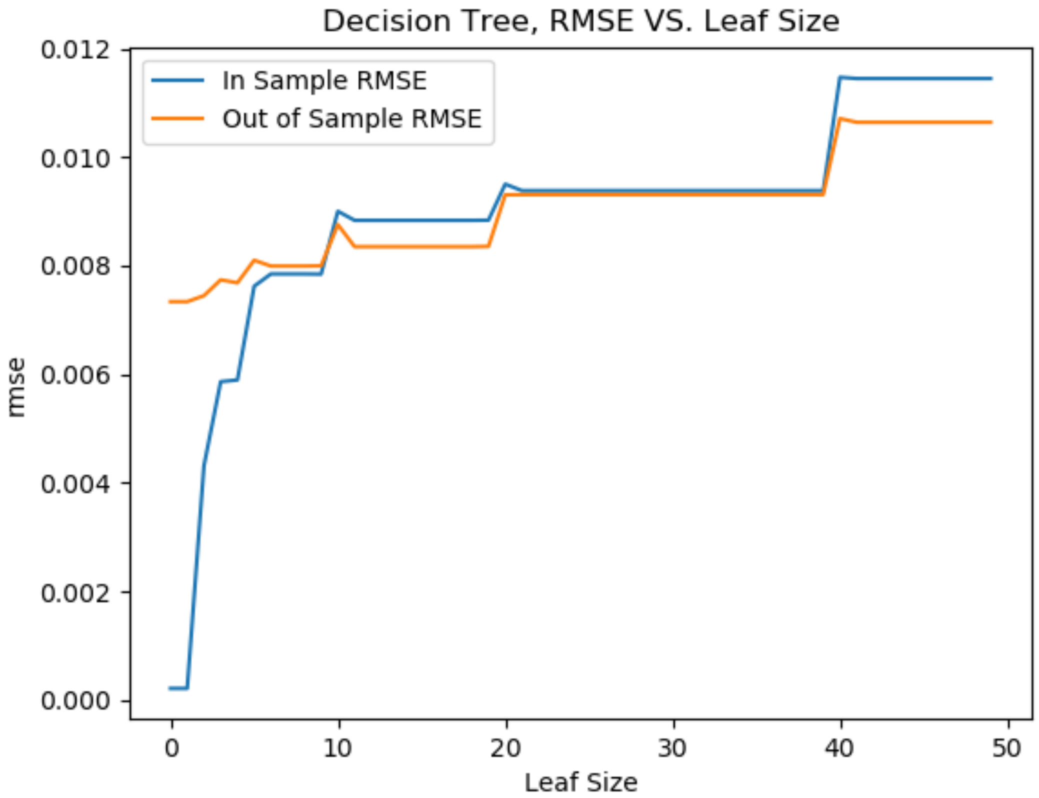
CS 7646 ML4T Project 3 Assess Learners

Chengqi Huang

chengqihuang@gatech.edu

Abstract – This report is to use the Decision Tree, Random Tree, and Bag learner that has been built in the assignment and discuss overfitting Decision Tree and Random Tree models. The report also conducts experiments of comparing Decision Tree model VS. Random Tree model

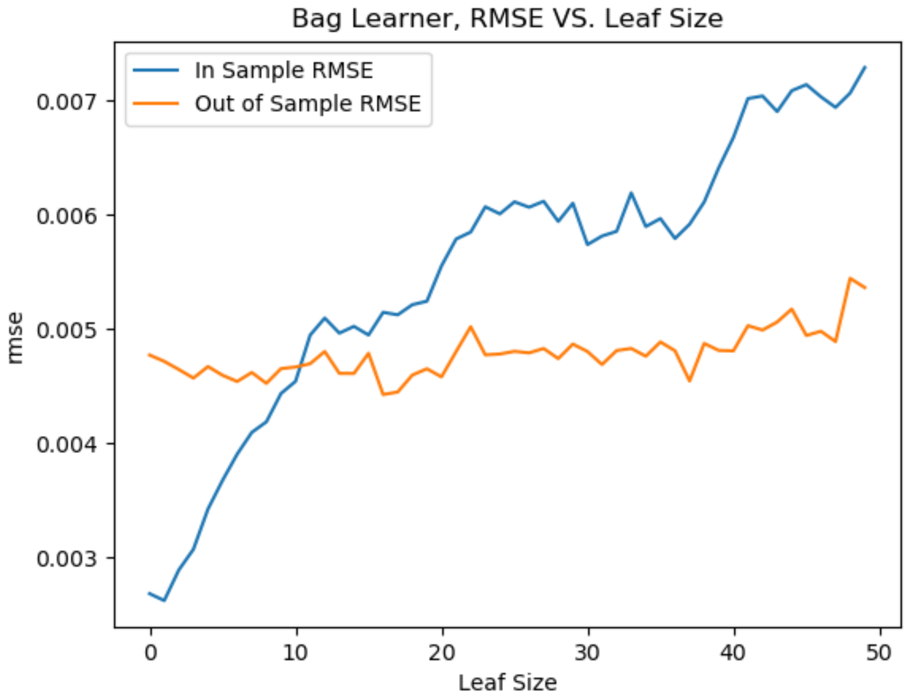
EXPERIMENT 1:



The chart above shows the RMSE of Decision Tree VS. Leaf Size. When leaf size is small (less than 8), we notice that in sample RMSE decrease drastically. Meanwhile the out of sample RMSE does not change a lot. We can say that there is over fitting when leaf size is small (less than 8). That's because the model fits too much to the training data, although in sample RMSE has a significant decrease, the out of sample RMSE stays the same.

EXPERIMENT 2

The bag learner combines 50 Decision Tree models. The chart below shows the RMSE vs. Leaf Size for in sample and out of sample RMSE:

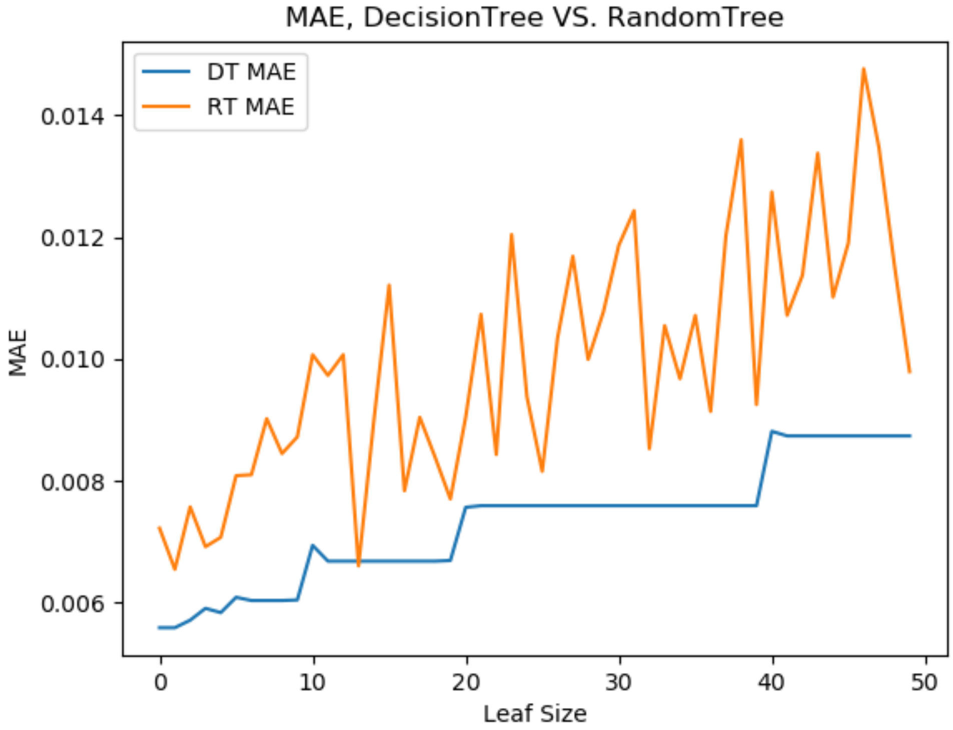


Still, there is over fitting when Leaf Size < 8, as in sample RMSE still decreases and out of sample RMSE does not change a lot when leaf size decreased. On the other hand, we can say that Bag learner does reduce overfitting a little bit. The trend that RMSE decreases is not as fast as decision tree learner. The bag learner does not eliminate overfitting.

EXPERIMENT 3

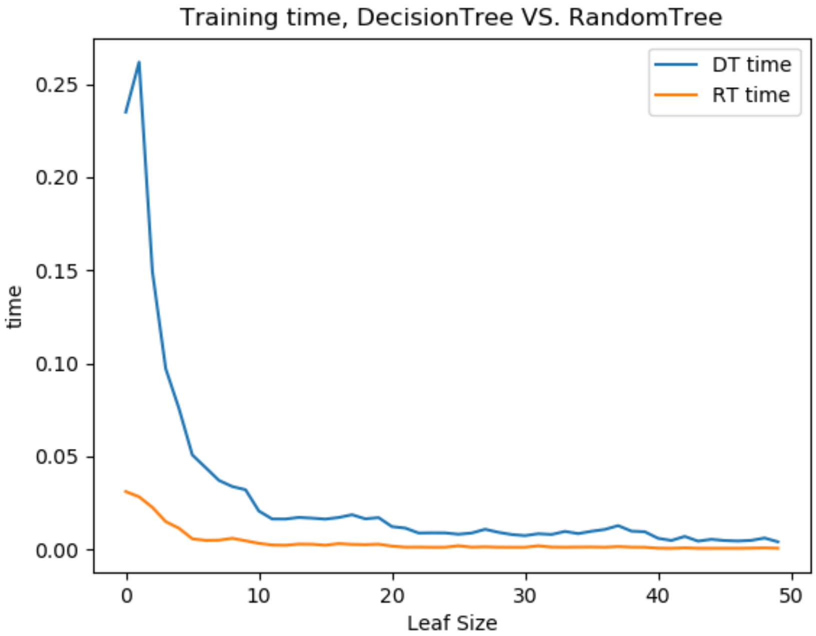
I choose2 metrics: Mean Absolute Error (MAE) and training time to measure the accuracy and cost of time to train the 2 models: Decision Tree and Random Tree

MAE:



The MAE chart shows the out of sample MAE for both Decision Tree and Random Tree model. It's clearly shown in the chart that the MAE of Decision Tree is less than Random Tree. We can say from the chart that if we use MAE to measure accuracy, decision tree is better than random tree. However, the difference is not significant. They are in the same level of accuracy in terms of MAE.

Training Time:



The chart above shows training time of decision tree and random tree. Random tree always takes less time to train, especially when leaf size is small (leaf size < 10). That's because decision tree needs to find the highest correlated factor in each branch, and it takes longer than randomly choose a factor. When the model gets more complexed(when leaf size goes down, or the amount of training data set is huge), random tree has a big advantage of taking far less time to train.