Project 3

Yingjie Qiu
yqiu322@gatech.edu

***Abstract*—**This report examined the performance and characteristics of three well-known supervised machine learning techniques. The results were based on three mini-experiment in Istanbul data. By comparing performances between decision trees and random trees, we get a comprehensive picture of underlying pathways of two trees.

1 INTRODUCTION

Decision trees are an important type of techniques for prediction in machine learning. With their transparent and efficient nature, they are widely used by scientists in many fields. To examine their characteristics and performance, I performed three mini-experiment to test the research hypothesis: 1: overfitting is associated with leaf size; 2. Bagging would reduce overfitting with respect to leaf size; 3. Decision trees outperforms random trees.

2 METHODS

Three mini-experiments were conducted to examine the proposed hypotheses.

The first mini-experiment was aimed to test the association between leaf size and overfitting for decision tree learners. Training data was set to 60% of Istanbul data, and rest of the data was used as testing set. The leaf size was set to be varied from 1 to 50, and the associated predictions for in-sample and out-of-sample were collected to calculate Root Mean Square Error (RMSE). The RMSE is a measurement of differences between actual value and predicted value. Generally, the model performs better if RMSE is lower.

$$RMSE = \sqrt{\sum_{i-1}^{N}(\hat{y}_i - y_i)^2 / N}$$

In the second mini-experiment, a similar examination as the first experiment with the use of bagging was conduct. The Istanbul was used with 60% of values being training sets and 40% of values being testing sets. The total number of bags was varied between 15 bags and 30 bags, with range of leaf sizes from 1 to 50. The RMSE was still used as metric to measure the fitting performances.

We compared the performance of classic Decision trees and Random tress in the third mini-experiment. Mean absolute error (MAE) and training time for out of samples were selected as the metrics for measurements. The same dataset as first two experiments with similar setting was used to test our hypothesis. Specifically, by changing the size of training sets that are passed to decision tree learners and random tree learners, we can collect the training time for those two learners from different size of training data. The MAE in different leaf size of two learners is calculated by the formula below.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - x|$$

3 DISCUSSION

3.1 Experiment 1

The RMSE were plotted against different leaf size in Figure 1. From the graph, I conclude that overfitting is associated with leaf size. To be specific, out of sample RMSE decreases as leaf size increases from 1 to 10 while in sample RMSE increases by the increase of leaf size, and they cross at leaf size around 10. The results suggest the overfitting starts when leaf size is less than 10. Additionally, out of sample RMSE remains constant when leaf size is 10 to 20. Since the in sample RMSE is larger than out of sample RMSE during this leaf range, we conclude underfitting occurs.

3.2 Experiment 2

Figure 2 and Figure 3 is generated with bag size 20 and 40. We could see that in sample RMSE keep increasing as the leaf size increase, and exceed the RMSE of out of sample when the leaf size is around 10. In the other hand, the out of sample RMSE shows a similar stable and constant trend in two different bag sizes. Since the RMSE of out of samples are stable when the RMSE of in sample increases, we conclude that bagging can eliminate overfitting with respect to leaf size.

3.3 Experiment 3

MAE for out of samples with different leaf size is plotted in Figure 4. Since MAE of DT learner is smaller than that of RT learner for majority of leaf sizes. We conclude that DT is better than RT in terms of prediction accuracy. Additionally, various sizes of training data sets are passed to DT and RT learners to test the

efficiency of algorithm. In Figure 5, we could see that the RT learners works faster in all possible leaf sizes. This may due to the fact that DT learners always pick the best feature but RT learners choose a random feature each time. By considering all the metrics, I prefer DT learners for prediction accuracy as generally the computing speed can be improved by using supercomputer.

4 SUMMARY

To summarize, the three mini-experiments provide us with comprehensive picture of these learners. Overfitting occurs with respect to leaf size and bagging is an effective wat to eliminate it. DT learners outperforms RT leaners in prediction accuracy although DT learners would cost more time on computing. Generally, depending on our use, we could choose which learners to use by doing simulation.

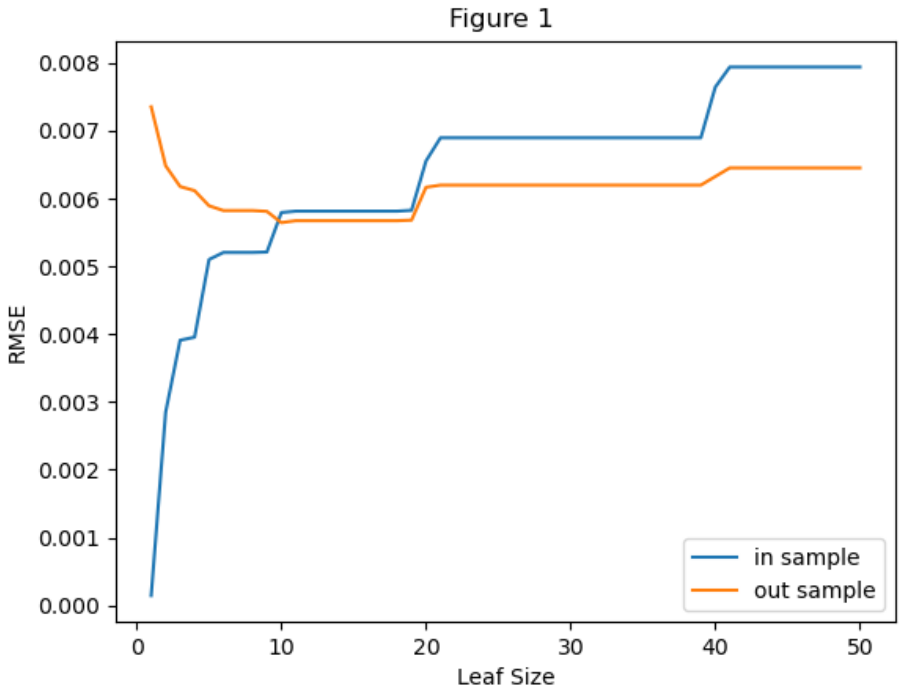

Figure 1. RMSE of DT learners with respect to leaf size

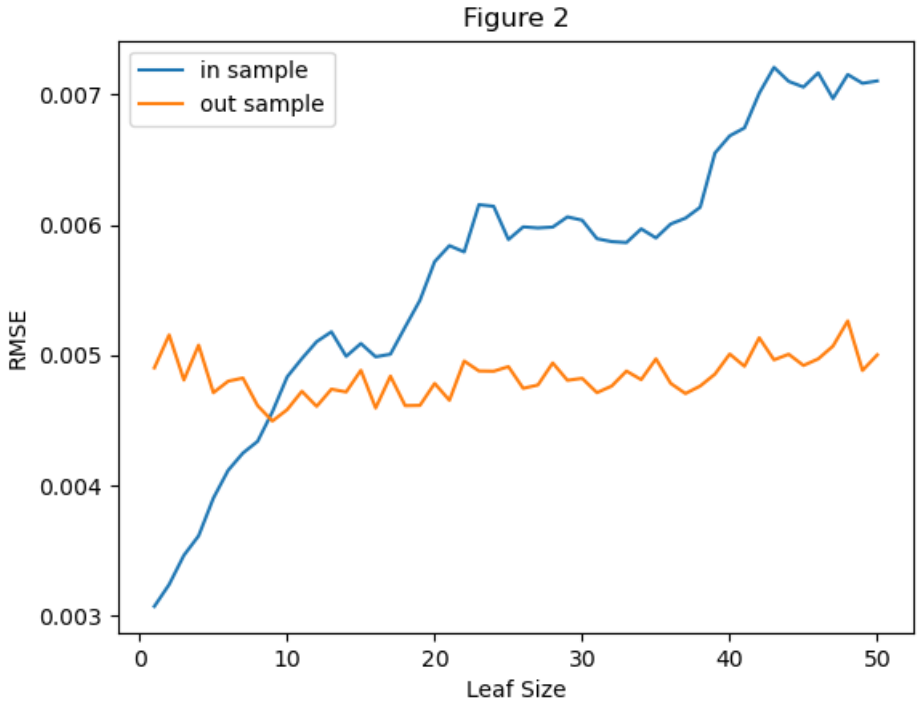

Figure 2. RMSE of DT learners with respect to leaf size (15 bags)

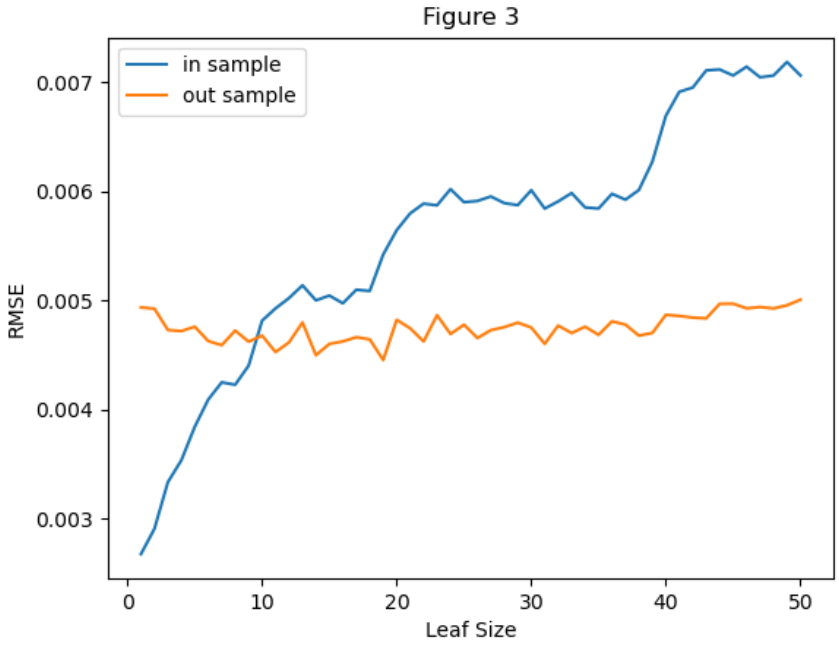

Figure 3. RMSE of DT learners with respect to leaf size (15 bags)

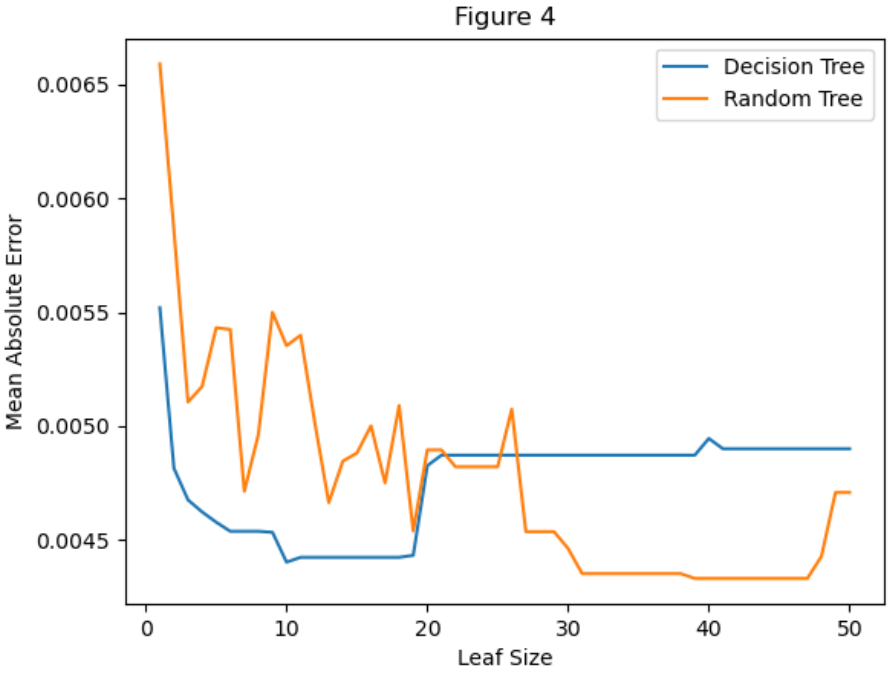

Figure 4. MAE with respect to leaf size (DT vs RT learner)

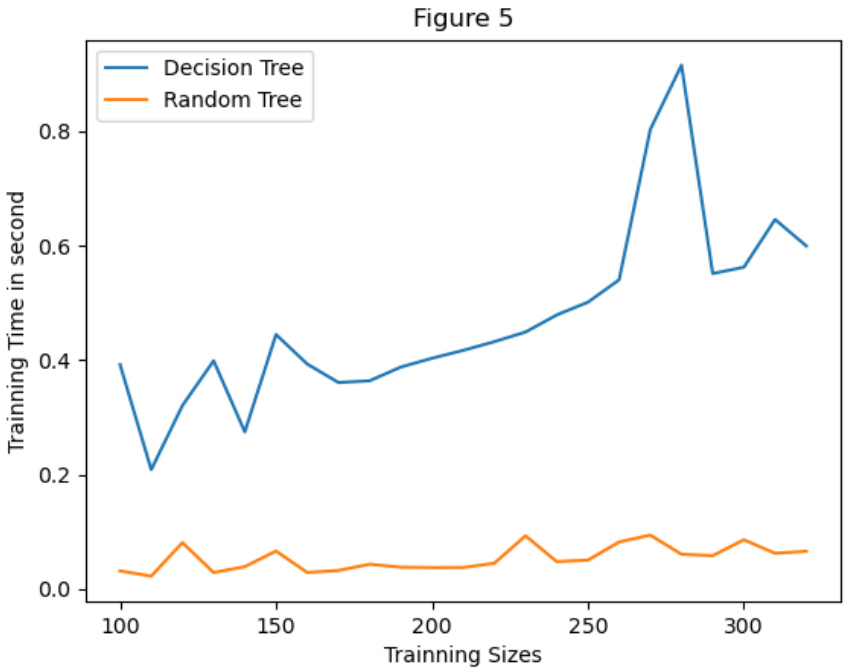

Figure 5. Training time with respect to training sizes (DT vs RT learner)