Project 3 Access Learner Report – CS7637

Zixin Feng
zfeng305@gatech.edu

***Abstract*—This report intends to compare the pros and cons of multiple machine learning models.**

1 INTRODUCTION

Machine learning is magic to most people. It seems like a black box and as long as we feed something to it, it can predict and produce nice results. However, what can influence the accuracy of machine-learning models' results? What are the pros and cons of each machine-learning model?

This report intends to compare the pros, cons, and accuracy of multiple learners. We will first study if leaf size will influence overfitting in the Decision Tree Learner. We naturally expect the smaller the leaf size, the stronger the overfitting effect we will see.

The next goal is to check if bagging can reduce or eliminate overfitting. Since Bagging trains a series of models and produces a "balanced" result, we would expect the overfitting effect to be reduced with bagging.

In the end, we will compare the pros and cons of Decision Tree Learner and Random Tree Learner by 2 metrics: Mean Absolute error and time to train. We will expect Random Tree Learner to take less time in training but the Decision Tree Learner to have less MAE.

2 METHODS

This chapter will introduce how we design 3 experiments to implement the 3 goals we discussed above.

2.1 Leaf Size vs Overfitting in Decision Tree Learner

We will calculate both in-sample and out-of-sample RMSE of the Decision Tree learner with different leaf sizes. Out-of-sample RMSE is usually much higher than in-sample RMSE when there is strong overfitting so we can utilize this conclusion to summarize if overfitting occurs with respect to the leaf size.

2.2 Bagging vs Overfitting in Decision Tree Learner

Bagging trains a series of models. Some models produce biased results at one extreme angle while other models produce results at another extreme angle. As a result, the mean of all models' results is supposed to be very unbiased and balanced.

In this experiment, we will set up the bagging size to be 20 and run Decision Tree Learner, both with and without bagging. We will also record their RMSE with different leaf sizes.

2.3 Comparison of Decision Tree Learner and Random Tree Learner

In this experiment, we will mainly use 2 metrics: Mean Absolute error and time to train to compare Decision Tree Learner and Random Tree Learner.

Training time could be a possible pro for the Random Tree Learner because the Decision Tree Learner always needs to spend time finding the best feature to split on but the Random Tree Learner doesn't take that time. However, because the Random Tree Learner doesn't find the best-split feature, it might have a bigger mean absolute error.

3 DISCUSSION

3.1 Leaf Size vs Overfitting in Decision Tree Learner

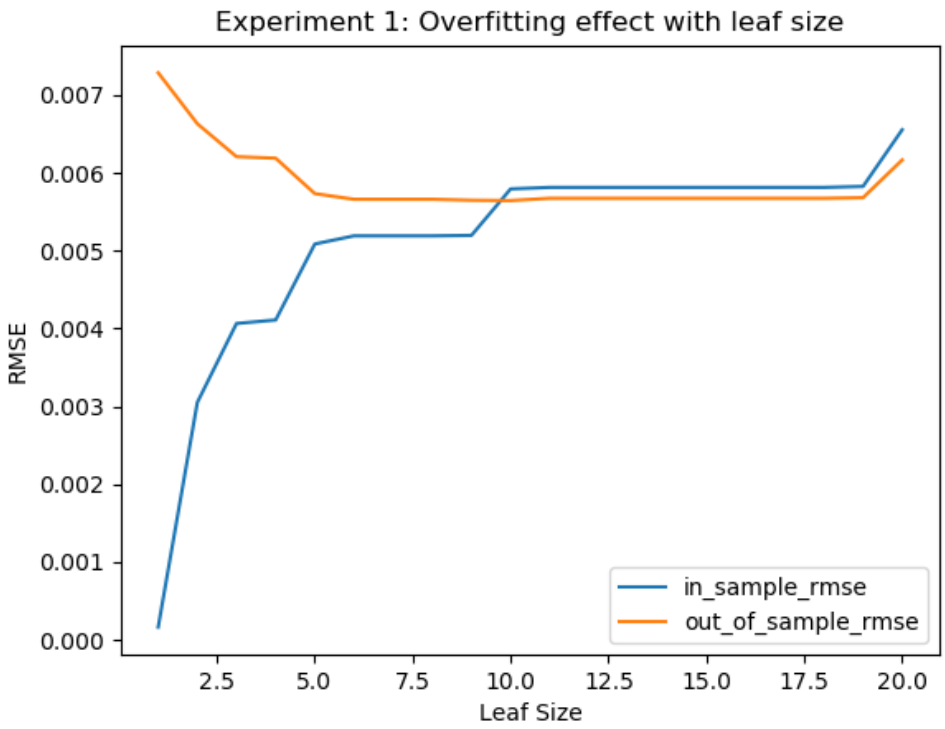

Figure 1—Overfitting effect with leaf size in Decision Tree Learner.

We observed a huge difference between in-sample RMSE and out-of-sample RMSE when leaf size is lower than 9, which indicates strong overfitting when leaf size is small.

3.2 Bagging vs Overfitting in Decision Tree Learner

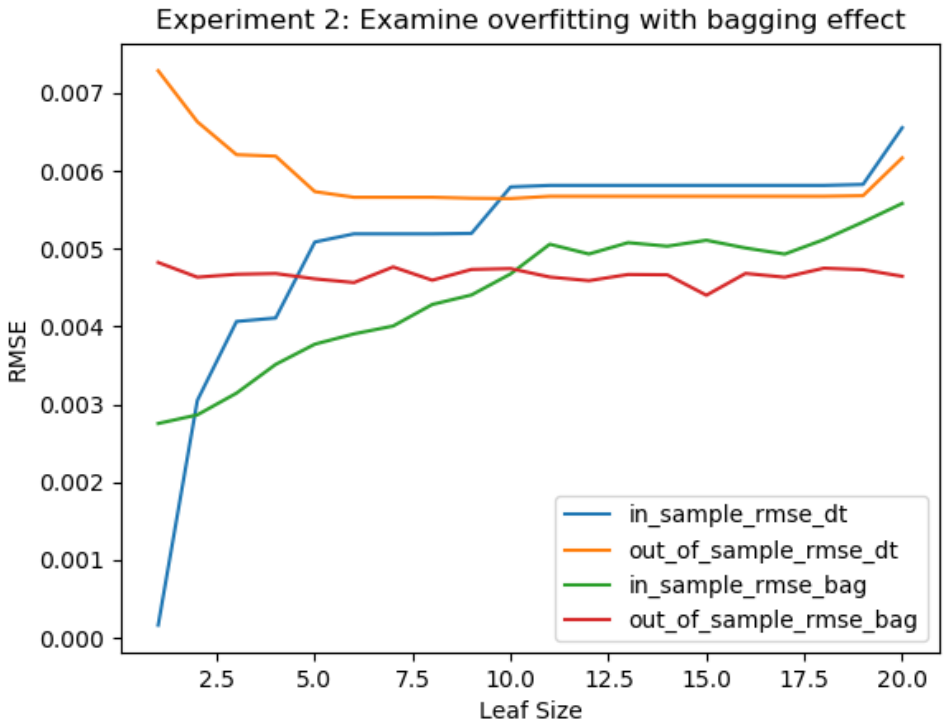

Figure 2—Overfitting effect bagging in Decision Tree Learner.

After introducing bagging, we can observe a much weaker overfitting effect even when the leaf size is smaller.

In the above graph Figure 2, green and red lines represent in-sample RMSE and out-of-sample RMSE of Decision Tree Learner with bagging effect. Blue and yellow lines represent in-sample RMSE and out-of-sample RMSE of Decision Tree Learner without bagging effect. We can obviously see a much smaller difference between lines with the bagging effect when the leaf size is smaller than 10. This is quite strong proof that bagging can help reduce overfitting when the leaf size is small.

3.3 Comparison of Decision Tree Learner and Random Tree Learner

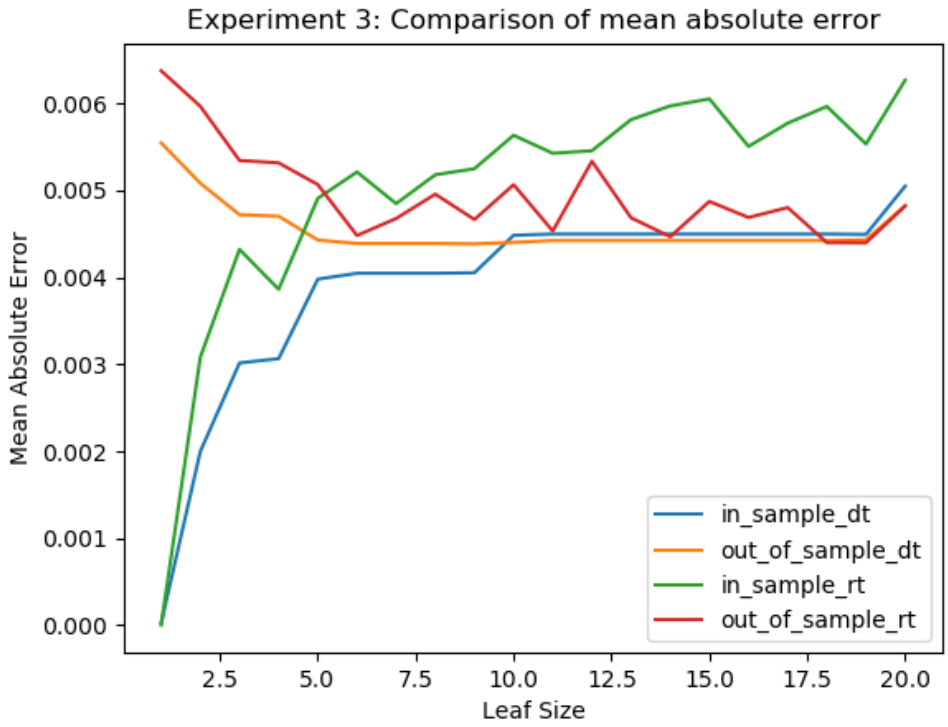

Figure 3—Comparison of MAE between Decision Tree Learner and Random Tree Learner.

In the above graph Figure 2, green and red lines represent in-sample MAE and out-of-sample MAE of Random Tree Learner. Blue and yellow lines represent in-sample MAE and out-of-sample MAE of Decision Tree Learner.

The Decision Tree Learner in general has a much smaller MAE value, no matter whether we talk about in-sample or out-of-sample mean absolute error.

The overfitting effect of Random Tree Learner happens when the leaf size is lower than 5 but the overfitting effect of Random Tree Learner happens when the leaf size is lower than 10.

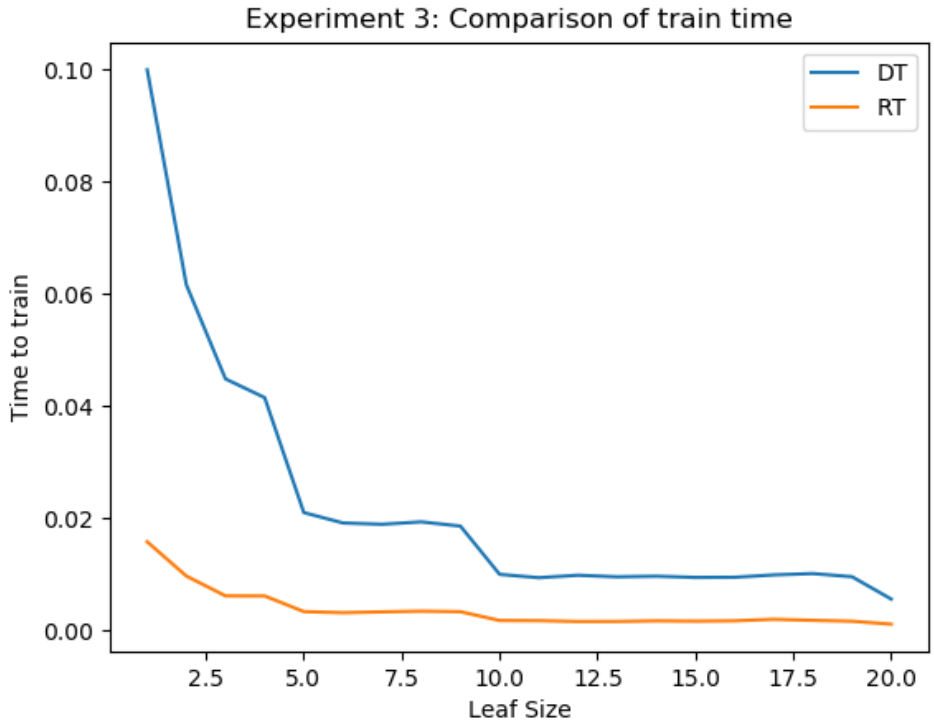

Figure 3—Comparison of train time between Decision Tree Learner and Random Tree Learner.

The train time of a Random Tree Learner is much shorter than that of a Decision Tree Learner in all leaf sizes. The smaller the leaf sizes, the more differences in train time.

4 CONCLUSION

Through this 3-part study, we can make following conclusions:

1. Overfitting occurs with respect to the leaf size. The smaller the leaf size, the larger the overfitting effect.
2. Bagging can significantly help reduce overfitting with respect to the leaf size.
3. Random Tree Learner and Decision Tree Learner have their pros and cons. Random Tree Learner takes a shorter time to train and Decision Tree Learner produces more precise results.