Project 3: Assess Learners

Allen Worthley
mworthley3@gatech.edu

Abstract—This project assesses the performance and effect of overfitting on various learning algorithms including decision trees, random decision trees, and ensemble learners.

1 INTRODUCTION

This project assesses the performance and efficacy of various learning algorithms including decision trees, random decision trees, and ensemble learners. The project performs several experiments to answer the following questions how does leaf size affect overfitting in a decision tree model, how do ensemble learners impact overfitting, and how does the decision tree algorithm perform compared to random tree leaner algorithm. In theory, overfitting occurs when leaf size is small in a decision tree, ensemble learners mitigate the effects of overfitting, and the random tree algorithm should outperform the decision tree algorithm in time but not accuracy.

2 METHODS

Three experiments are setup to test various learning algorithms and their effectiveness. Experiment 1 seeks to test the effect of overfitting with respect to leaf size on the decision tree algorithm. A decision algorithm models a target data set by splitting the training set into subdivisions based on the best feature within the training set. In this model, the best feature is the training factor the highest correlation to the target data set. The experiment trains 50 different decision tree models corresponding to varying leaf sizes on the same training data. The experiment then compares the in-sample and out-of-sample root mean squared error ("RMSE") across the different models.

Experiment 2 is similar to experiment 1, however, the setup focuses on the effect of ensemble learners on overfitting. Specially, the ensemble learner used is a Bag Learner leveraging 20 decision tree models with 20 corresponding "Bags." Each bag samples the training set with replacement and trains a decision tree model with varying levels of leaf size. The experiment compares RMSE for both in- and out-of-sample across the "Bags" of models with a set leaf size.

Experiment 3 seeks to assess the performance of the decision tree model compared to the random tree model. The random tree model is similar to the decision tree model except that the best feature is chosen at random. The experiment creates a series of 50 independent training sets called trials to train both the random and decision tree models. Two metrics are used to quantify performance: mean absolute error ("MAE") of out-of-sample data and time to build. Note that time to build is the elapsed time between the system clock at which the algorithm started to the time the algorithm built the model. Time is highly dependent on the local environment.

3 DISCUSSION

Most of the below experiments seek to explore the effects of overfitting. Overfitting occurs when in-sample error decreases but out-of-sample error increases. This occurs because as more parameters are introduced into a model to explain observed sample data, the mathematical relationships observed in sample do not generalize well to the population of data.

3.1 Experiment 1

The results from experiment 1 show that overfitting does occur as leaf size decreases in the decision tree models. As shown in Figure 1, in-sample RMSE continues to decrease with out-sample-error until leaf size is 10. After 10, overfitting is apparent for this training set as in-sample error quickly declines towards zero and out-of-sample error steeply inclines to its max for the models at around 0.0075.

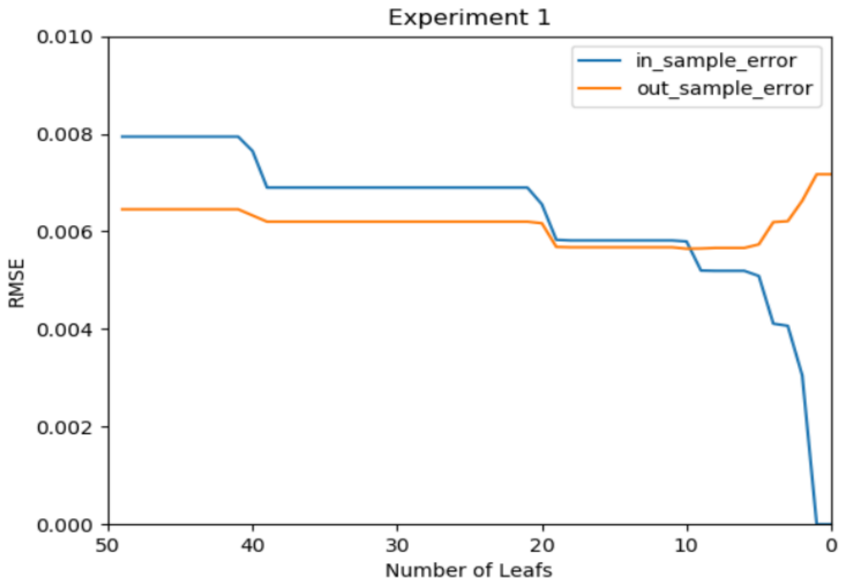

Figure 1—RMSE results of experiment 1 as shown from largest leaf size (50) to smallest (1).

3.2 Experiment 2

Bagging does reduce overall RMSE. This is apparent when comparing RMSE values in Figure 1 to Figure 2. It reduces the overall effect of overfitting as shown in Figure 2. As leaf size decreases, out-of-sample error remains relatively stable hovering around a RMSE of 0.005. This is surprising as the decision tree model saw sharp inclines in out-of-sample RMSE at around a leaf size of 10. However, overfitting does not eliminate overfitting entirely, there are instances of overfitting like in ranges 9 to 0. While out-of-sample error remains flat, in-sample error decreases relatively quickly. Depending on the application, in-sample error may have a negative impact on the analysis. However, the Bag learner does a fantastic job of reducing the out-of-sample errors.

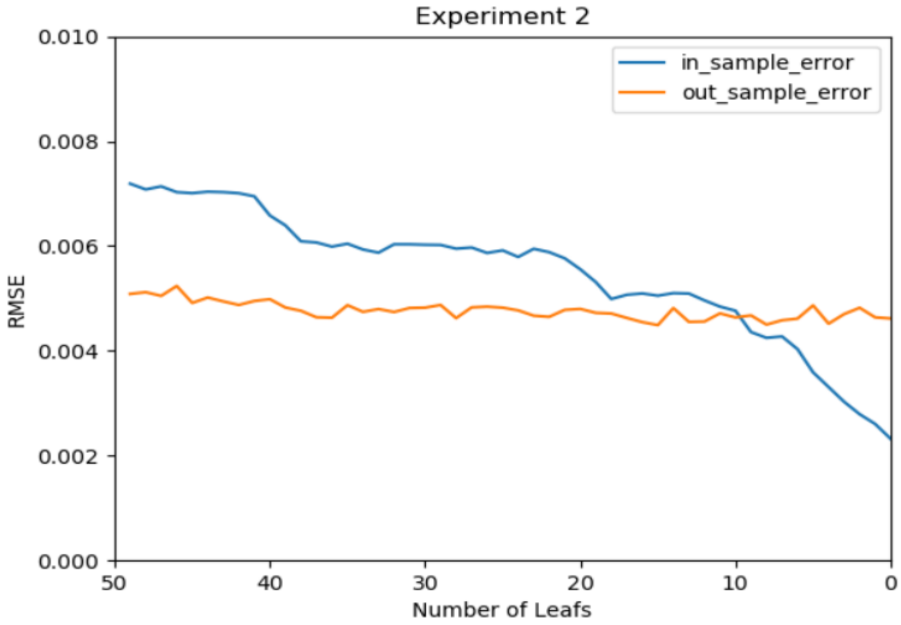

Figure 2 — RMSE results of experiment 2 as shown from largest leaf size (50) to smallest (1).

3.3 Experiment 3

On average, the decision tree model has less mean absolute error ("MAE") than the random tree model by 0.0002. While small, the decision tree model still has more explanatory power than the random tree model. As shown in figure 3, the decision tree algorithm is mostly below the random tree model.

In Figure 4, the time to build or time to construct the tree for the random tree algorithm is almost twice as fast as the decision tree algorithm.

The random tree model seems superior in that there is relatively low MAE difference to the decision model and the time to build is quite fast. This is likely due to the limited processing it takes to randomly choose a number representing a factor. The decision tree on the other hand has to process unique edge cases and perform correlation calculations on each step in the recursive algorithm. Though, the decision tree model is still marginally better than the random forecast because it has lower MAE for most of the trials.

Yet, no model is unanimously better in every case. Decision tree algorithms are easy to interpret and understand. Random tree algorithms are essentially black boxes. If the need for an algorithm was to be efficient with mild accuracy, the random tree model would be better. If the need was to get as close to 0 error as possible, then the decision tree model would be better.

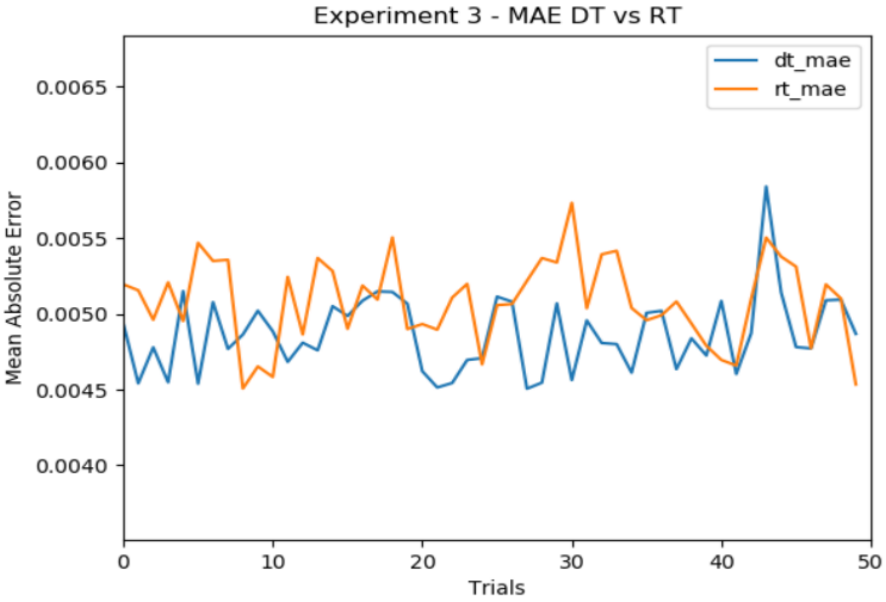

Figure 3 — MAE results for experiment 3 for 50 trials comparing the decision tree algorithm ("dt") to the random tree algorithm ("rt").

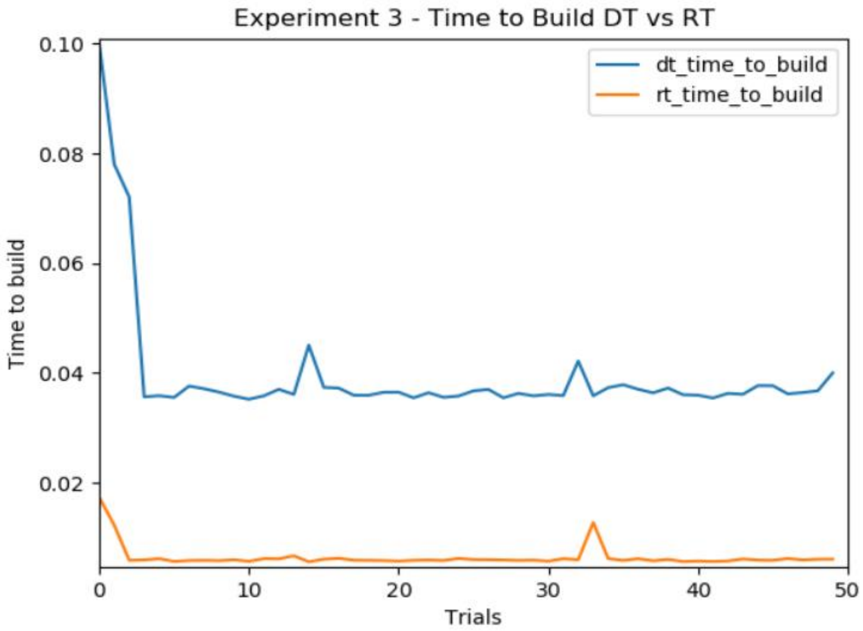

Figure 4— Time to train the model results for experiment 3 for 50 trials comparing the decision tree algorithm ("dt") to the random tree algorithm ("rt").

4 SUMMARY

In conclusion, overfitting does occur as leaf size decreases in the decision tree algorithm. Ensemble learners reduce and mitigate some of the overfitting effects compared to single models. Random tree models are fast but have limited accuracy compared to decision tree models.