Project 3: Assess Learners

Jinelle Gilfillan
jgilfillan6@gatech.edu

Abstract—This project explores metrics regarding overfitting of trading data machine learning techniques. The three techniques used in the experiment are Decision Tree Learning, Random Tree Learning, and Bagging. The first experiment will explore overfitting for one Decision Tree, the second experiment will explore overfitting for the bagging of 20 Decision Trees, and the last experiment will explore the time to train and the time to query between Decision Tree Learning and Random Tree Learning.

1 INTRODUCTION

There will be three experiments conducting in this project, exploring overfitting and metrics for Decision Tree Learning, Random Tree Learning, and Bagging. Bagging was completed by sampling with replacement from the training dataset for each tree included.

The first experiment will explore overfitting with one Decision Tree. The hypothesis for this experiment is that the smaller the leaf size, the more chance for overfitting to occur. The reasoning for this is because with smaller leaf sizes, the data is more fit for just the training data and may not fit for the entire dataset.

The second experiment will explore overfitting with multiple Decision Trees using Bagging. The hypothesis for this experiment is that the error will be improved over the error seen in experiment 1. The overfitting of leaf size will be a smaller leaf size as well. This is because with random sampling with replacement of nodes within the training dataset, the results can be less biased and have less error overall.

The third experiment will cover two additional metrics comparing Decision Tree Learning and Random Tree Learning. The two additional metrics will be the time to train the tree and the time to query the tree. The hypothesis is the Random Tree training will be faster than the Decision Tree training. However, the Decision Tree could have less nodes as it is splitting on the best factor instead of random factor, so the time to query for Decision Tree Learning may be faster than Random Tree Learning.

2 METHODS

For the first experiment, the error calculation that will be used to determine overfitting is RMSE error. The degrees of freedom that will be used is leaf size of the decision tree. This experiment will be a comparison of the results of in-sample testing and out-of-sample testing. Overfitting will be determined at the point where the in-sample error increases over the leaf size and the out-of-sample error decreases.

For the second experiment, the set up will be the same as experiment 1, but with using 20 bags. Again, Bagging occurs with replacement from the training dataset for each tree. The error calculation and degrees of freedom will remain the same as experiment 1. Overfitting will also be determined using the same criteria as experiment 1.

The third experiment will time each of the Decision Tree method and Random Tree method on training and querying. The y axis will be the time and the x axis will remain the leaf size. The two lines on the graph will be the Decision Tree and the Random Tree. Results will be determined by which line sits higher on the graph and therefore takes more time to run. The time measured will be the time it takes to run a single tree.

3 DISCUSSION

3.1 Experiment 1

Using the above methods for experiment 1, the following result was generated:

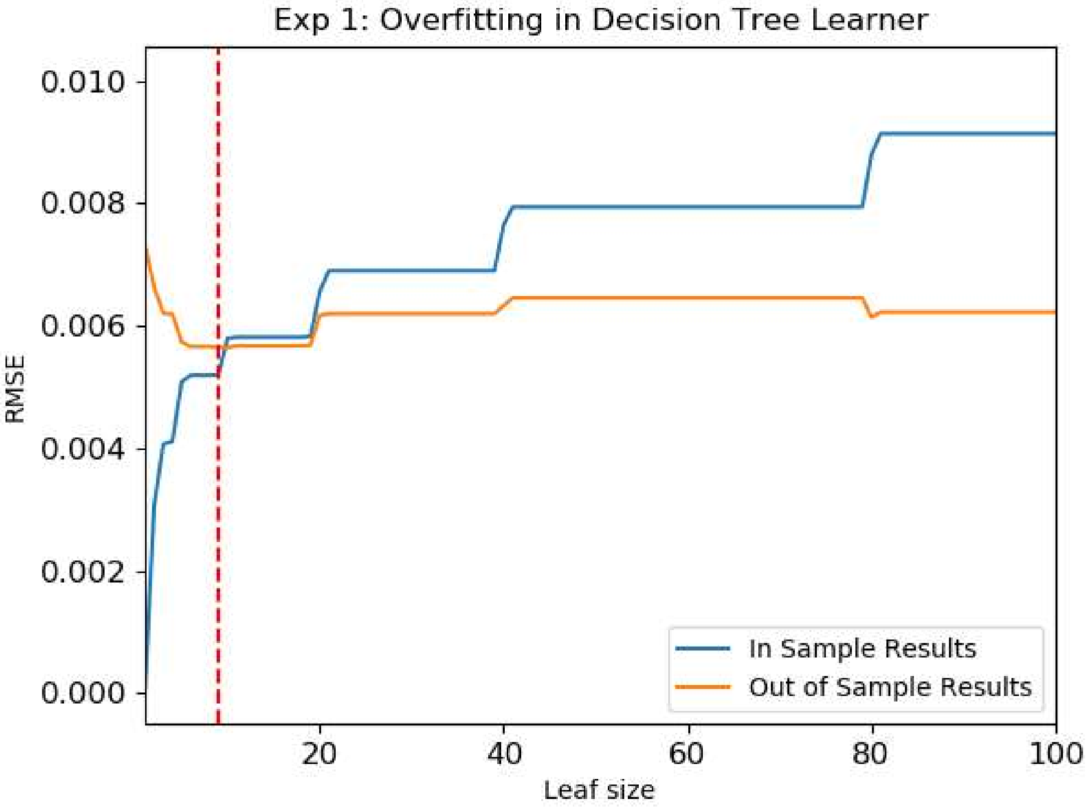

Figure 1—Experiment 1 results: overfitting between in sample and out of sample RMSE error

The axis of the leaf size was narrowed down to 0 – 100 in order to see the results of overfitting clearer. The red dotted line indicates the point of overfitting.

As assessed from the graph, there is a clear point at which overfitting occurs with respect to leaf size. It occurs at leaf size 9, indicated by the red dotted line. The direction of overfitting is going left, or from leaf size 1 to leaf size 9. You can see this in the graph because the out of sample RMSE error is decreasing while the in sample RMSE error is increasing in this leaf size range.

The results match the hypothesis because the smaller leaf sizes (leaf size 0-9) are the leaf sizes where overfitting occurs.

3.2 Experiment 2

Using the method above for experiment 2, here are the results:

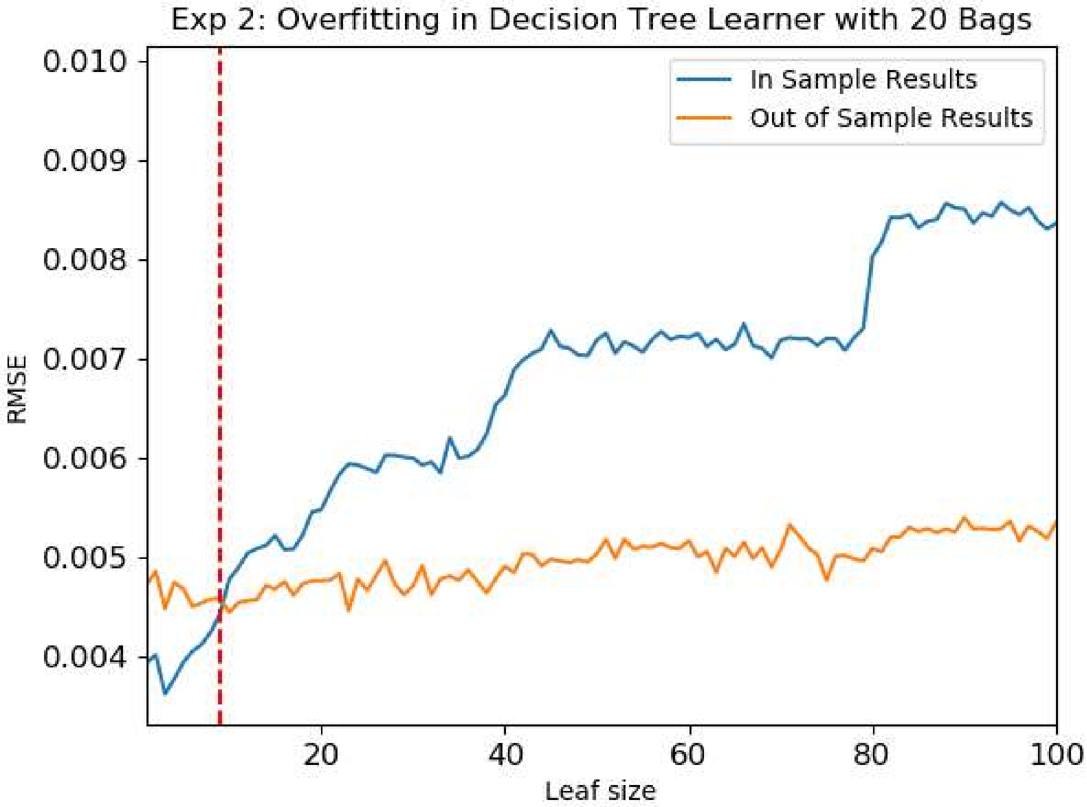

Figure 2—Experiment 2 results: overfitting analysis using decision tree learner and 20 bags

The axis of the leaf size was narrowed down to 0-100 in order to see the results more clearly. The point of overfitting is indicated with a red dotted line.

As assessed from the graph, there is a clear point at which overfitting occurs with respect to leaf size for this experiment as well. It occurs at leaf size 9 as well, indicated by the red dotted line. The direction of overfitting is going left, or from leaf size 1 to leaf size 9. You can see this in the graph because the out of sample RMSE error is decreasing (in general) while the in sample RMSE error is increasing (in general) in this leaf size range. It is also worth noting that the

RMSE errors for both the in sample and out of sample results are sitting lower than with one decision tree. Bagging with 20 bags is sitting between 0.004 and 0.005 RMSE while using one decision tree is sitting between 0.005 and 0.006 at leaf size of 9.

Looking at the results of this experiment, it looks like bagging does not reduce overfitting in regard to leaf size and does not eliminate overfitting. That part of the original hypothesis was wrong. It does, however, reduce the overall RMSE error that is shown for both in sample and out of sample data.

3.3 Experiment 3

The results for the training time between Decision Tree and Random Tree Learners is show below. The axis is 0-100 leaf size to see results better. The time factor is the time it takes a single tree to run.

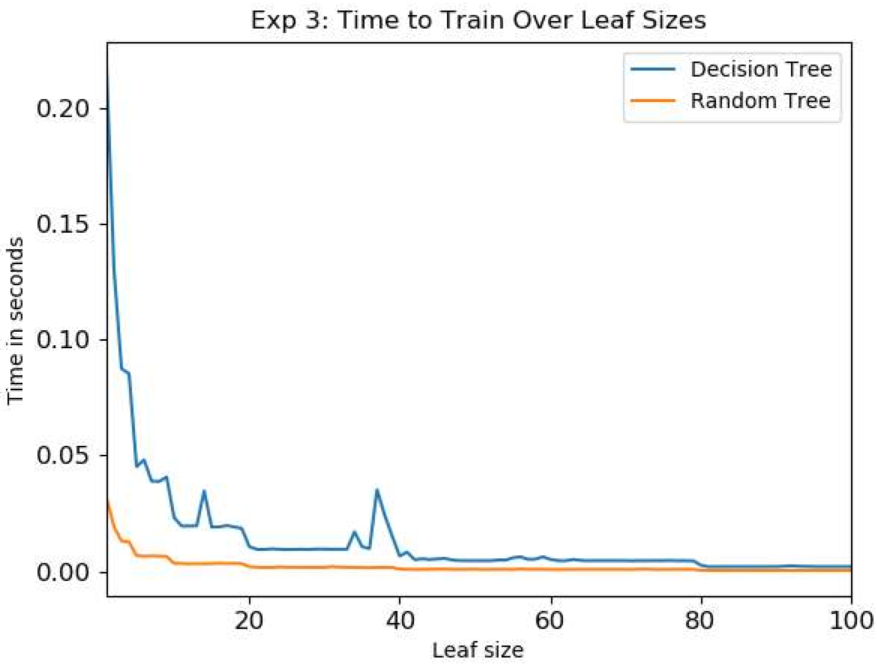

Figure 3—Time to train comparison between Decision Tree Learning and Random Tree Learning

As hypothesized, Decision Trees take longer to train. This makes sense because Random Trees are selecting the factor to train on randomly while Decision Trees are taking the extra time to determine the best factor to train on.

The results for the querying time between Decision Tree and Random Tree Learners is show below. The axis is 0-100 leaf size to see results better. The time factor is the time it takes to query against a single tree.

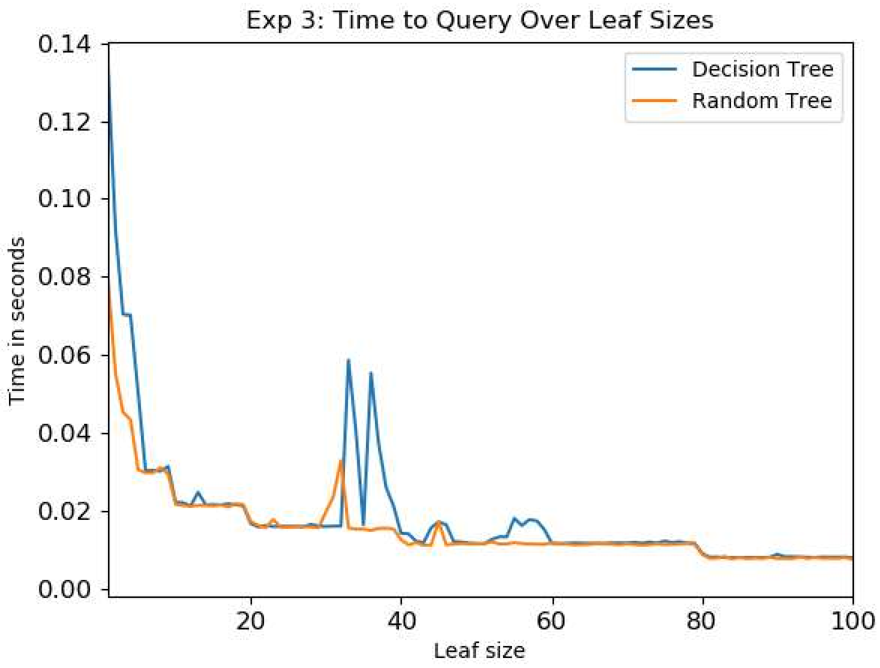

Figure 4—Time to query comparison between Decision Tree Learner and Random Tree Learner

These results do not match the hypothesis for this metric. Assuming that the Decision Tree Learner is selecting the best factor to train on, it should theoretically have less leaf nodes and therefore require less time to query. However, what are actual results show is that the two, for the most part, take the same time to query. There are spikes for when Decision Tree Learning takes a little longer.

Strictly looking at timing results without comparing training results, it appears that Random Tree Learning is faster, and therefore a better approach to use.

4 SUMMARY:

Experiment 1 demonstrated that overfitting can occur when training datasets and that the leaf size selected directly impacts the risk to overfitting. Therefore, leaf size selection should be carefully analyzed when creating and running learner models.

Experiment 2 demonstrated that overfitting is still an issue when using bagging. Even though multiple trees are being created and used in the process, selecting an appropriate leaf size is still very important.

Experiment 3 demonstrated the timing difference between using Random Tree and Decision Tree methods. However, in order to truly determine the better method, further analysis should occur comparing the error of each of the methods as well. In addition, the results are returned for a single tree, using the bagging method and averaging the time across multiple trees could provide better insight into the timing differences between the different methods.