CS7646 - Project 3: Assess Learners

Apurva Gandhi
agandhi301@gatech.edu

Abstract—The study examines diverse patterns of overfitting in machine learning models with regard to various leaf sizes. It demonstrates that overfitting occurs for small leaf sizes but decreases as leaf number rises. Bagging reduces overfitting, especially when there are more bags. In addition, runtime analysis reveals that the Random Tree (RT) Learner is more time-effective during training while the Decision Tree (DT) Learner exhibits reduced Mean Absolute Error (MAE), indicating higher prediction accuracy.

INTRODUCTION

This research project explores the behavior of decision tree learners, random tree learners, and bagged ensembles in the context of assessing overfitting. The study involves constructing trees using these learners and querying them for specific values. Evaluation metrics include the correlation and root mean square error (RMSE) calculated for both in-sample and out-of-sample data sets. Subsequently, we create a bagged learner using a decision tree with a bag size of 20. An "insane" learner is then formed by aggregating 20 bagged learner instances, each composed of 20 linear regression learners. The hypothesis for the three conducted experiments is that overfitting will occur in the DTLearner model when the leaf size is too small, resulting in a decrease in performance as measured by RMSE. Additionally, Bagging will reduce overfitting in the DTLearner model when compared to only DTLearner, resulting in improved RMSE values. Finally, Random Trees (RTLearner) will perform better than Classic Decision Trees (DTLearner) in terms of two new quantitative measures, run time execution and Mean Absolute Error.

EXPERIMENT 1

Method

The objective is to investigate the occurrence of overfitting in a Decision Tree Learner (DTLearner) model when varying the leaf size parameter. The dataset employed for this analysis is "Istanbul.csv." The independent variable, leaf size, will be systematically modified, while the dependent variable, Root Mean Squared Error (RMSE), will serve as the metric to assess overfitting. The process involves training the DTLearner with different leaf size values, and the resulting RMSE values will be plotted against leaf size. The primary hypothesis suggests that overfitting will manifest when leaf size is too small, leading to a deterioration in model performance as indicated by increased RMSE values. The analysis will discern the precise values of leaf size at which overfitting commences and the direction of its impact on RMSE.

Discussion

There is overfitting with regard to leaf size. Overfitting begins when the leaf size is less than 5. Figure 1 shows that the Out Sample RMSE declines from 1 to 5 as leaf size increases, indicating overfitting before 5. The RMSE grows after 10, which denotes undercutting. The decision tree model is susceptible to overfitting due to its complex and lengthy decision chain based on a subset of features. It becomes highly specific to the training data, making it less capable of generalizing to new, unseen data points. Overfitting is exacerbated when the tree is too deep or requires excessive splits, indicating the need for a larger leaf size to prevent overfitting.

EXPERIMENT 2

Method

Building on the results of Experiment 1, Experiment 2 investigates the efficacy of bagging in reducing overfitting in the context of a Decision Tree Learner (DTLearner) model. We'll use the same "Istanbul.csv" dataset. The dependent variable for determining overfitting in this experiment is RMSE, but the independent variable is still leaf size, which will be adjusted consistently. The DTLearner model's implementation of bagging is where there is the biggest difference. Using different leaf size values, the DTLearner is trained with bagging as part of the procedure, and the resulting RMSE values are plotted versus leaf size. The basic assumption is that bagging, when compared to Experiment 1, will lessen overfitting, leading to improved RMSE values.

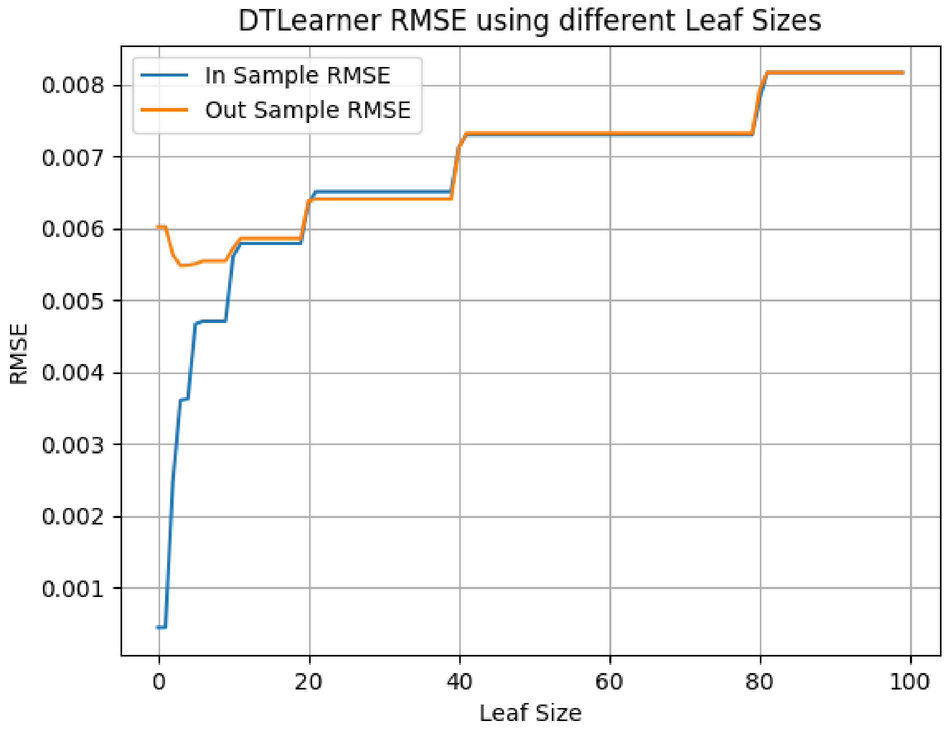

Figure 1—Leaf Size and Overfitting with DTLearner

Discussion

As the number of leaves rises, bagging does lessen overfitting within a dataset, but it does not appear to be able to completely remove overfitting. However, it appears that overfitting is less common with more bags. Additionally, when there are more bags, the root square mean error decreases. The following Figure 2 illustrates this using 1, 10, 20, and 50 bags respectively. More bags also result in a smoother graph since the RMSE variance for close leaf sizes is considerably more comparable. In comparison to 50 bags, the RMSE vs. leaf size data for 1 bag vary greatly. However, it is evidently obvious that there is far more overfitting without bags than with them when comparing the out-of-sample RMSE curve between a bag size of 1 and a bag size of 50 in Figure 2. In fact, there is essentially little to no overfitting with 50 bags since, as illustrated, the orange out-of-sample RMSE line doesn't have a steeply declining slope right away.

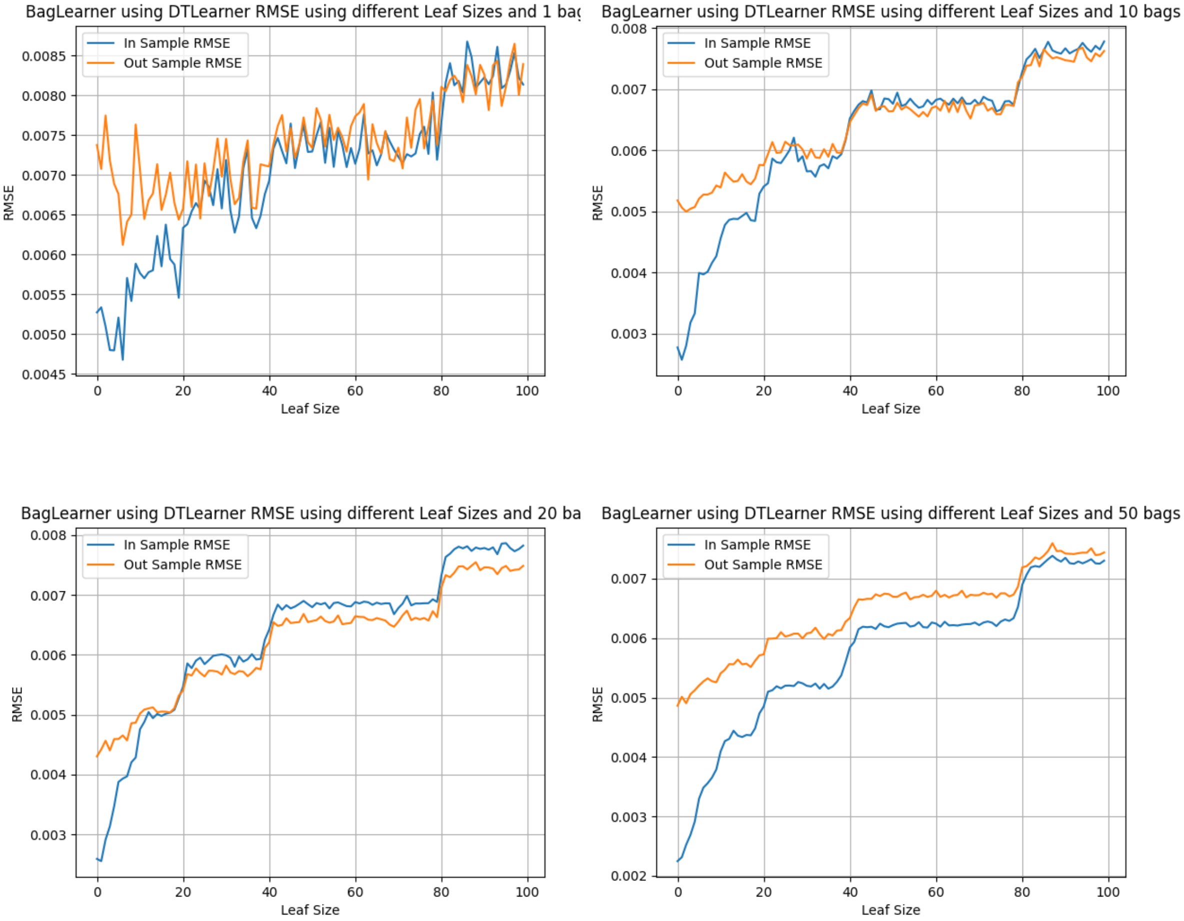

***Figure* 2**—Leaf Size and Overfitting with Bag Learner using Different bag sizes fo 1, 10, 20 and 50

EXPERIMENT 3

Methods

Experiment 3 introduces a comparative study between Classic Decision Trees (DTLearner) and Random Trees (RTLearner) using a different dataset distinct from "Istanbul.csv." Independent variables are not introduced, as the primary focus is on comparing the two models. Instead, the experiment seeks to establish the superiority of one model over the other based on two novel quantitative measures: Run Time and Mean Absolute Error. The method comprises training both DTLearner and RTLearner on the chosen dataset and subsequently calculating the Run time and Mean Absolute Error for each model. The results will be visualized through charts to facilitate comparison. The hypothesis asserts that RTLearner will outperform DTLearner.

Discussion

The first metric used to compare two learners is run time execution. In accordance with figure 3, the analysis of the query time shows that there are no significant differences between the two models, with the exception of a few peaks, both graph lines mostly coinciding. As a result, it is impossible to choose a better model based purely on query time. The Decision Tree (DT) Learner consistently needs more time to train the model than the Random Tree (RT) Learner, according to measures for training time. As a result, in terms of time efficiency, the RT model performs better than the DT model. Another finding from the original graph is that the DT learner's time consumption increases significantly when the leaf size is further decreased to 10.

The second metric used to compare two learners is Mean Absolute Error. As per figure 4, the graph shows that the mean absolute error (MAE) of the decision tree learner (DTL) is almost consistently lower than that of the random tree learner (RTL), regardless of the leaf size. This means that the DTL is making more accurate predictions than the RTL. This is likely because the DTL is less likely to underfit the training data than the RTL. It could also be due to the nature of decision trees, which tend to create more complex models that fit the training data more closely, potentially leading to more accurate predictions. Similarly, the Decision Tree Learner (orange line) also appears to be better than the Random Tree Learner (blue line) in terms of Mean Absolute Error (MAE) for out-sample as seen in Figure 5. The MAE of the Decision Tree Learner is consistently lower, indicating that its predictions are, on average, closer to the actual values. The Random Tree Learner, on the other hand, has a higher MAE for most leaf sizes and a sharp spike in error at a leaf size of around 80.

It's unlikely that one learner will always be superior to another. The performance of machine learning algorithms can vary greatly depending on the specific characteristics of the dataset, including its size, dimensional, noise level, and underlying patterns. While the Decision Tree Learner performed better in this case, there may be other scenarios or data sets where the Random Tree Learner could outperform. The DTL is likely to perform better than the RTL on small to medium-sized data-sets. However, the RTL may perform better than the DTL on large data-sets.

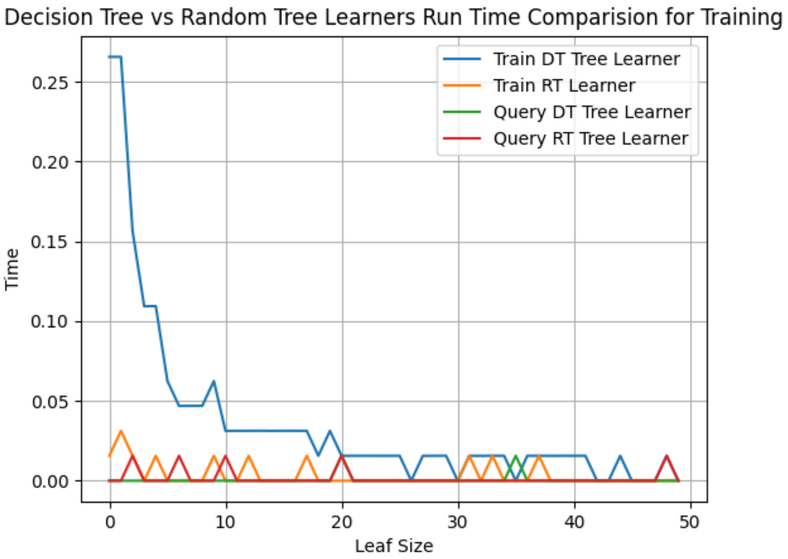

Figure 3—Run Time Comparison of training and querying between DT and RT Learner

SUMMARY

The data acquired from the studies show that DT Learner has a propensity towards overfitting. Additionally, it was discovered that the training time for the Random Tree (RT) learner is more efficient than the training time for the Decision Tree (DT) learner. In contrast, the DT learner outperforms the RT learner in terms of accuracy for the dataset and leaf sizes presented. However, it is anticipated that when utilized in a bagging framework to train the model, the RT learner will outperform the DT learner in terms of prediction accuracy.

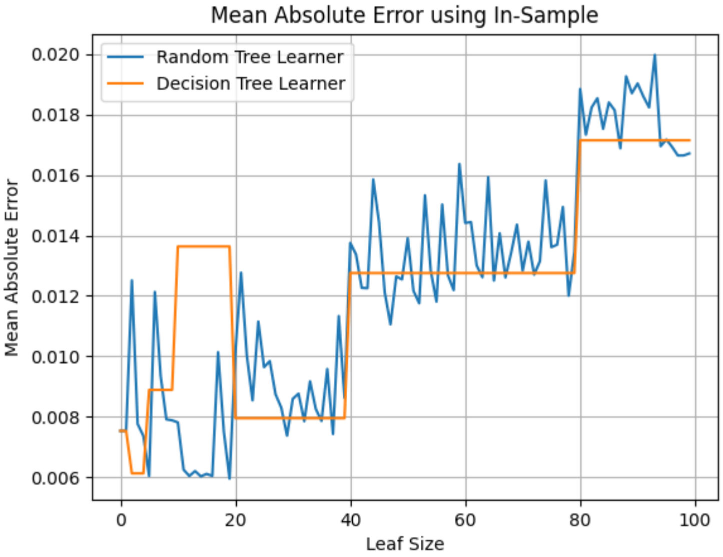

Figure 4—Mean Absolute Error between DT and RT Learner using In Sample data

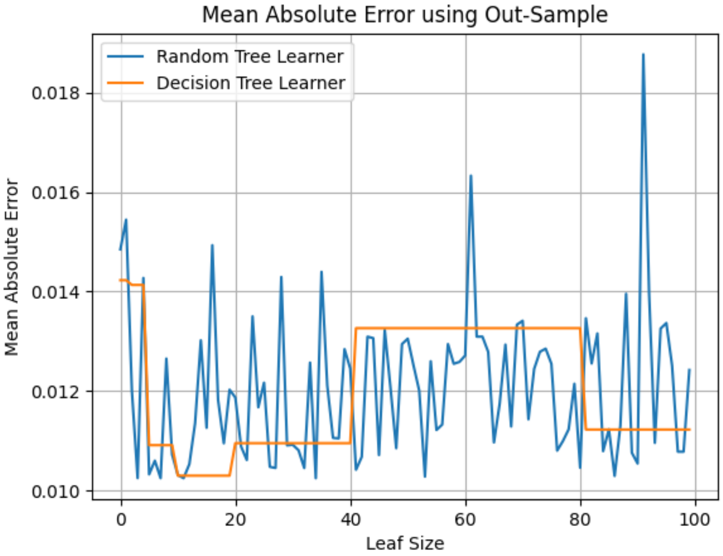

Figure 5—Mean Absolute Error between DT and RT Learner using Out Sample data