CS 7646 Project 3 – Assess Learners

Guangqing Tai
gtai3@gatech.edu

Abstract— ***This project evaluates multiple supervised learning algorithms, namely Decision Tree (DT), Random Tree (RT), Bag Learner (BL), and Insane Learner (IL), to assess their performance and potential overfitting behavior. The experiments were conducted using a time-series financial dataset to investigate how different parameters impact the learners' ability to generalize, specifically focusing on leaf size, bagging techniques, and comparing classic decision trees with random trees.***

Note: This work is a re-submission of Summer 2024 version, which may contain partial previous work.

1 INTRODUCTION

The primary goal of this project was to implement and evaluate different machine learning algorithms to understand their behavior concerning performance and overfitting. Specifically, we explored the DT, RT, BL, and IL algorithms. The project investigated how leaf size influences overfitting, how bagging affects the performance of decision trees, and how DT and RT learners compare using Mean Absolute Error (MAE) and other metrics. We hypothesized that ensemble methods like BL would reduce overfitting compared to simpler learners (DT and RT) and that RT would be less prone to overfitting than DT due to its random feature selection.

2 METHODS

The experiments were conducted using the Istanbul dataset, where the data was split into 60% for training and 40% for testing. Three main experiments were performed.

2.1 Experiment 1: DT Learner Overfitting Analysis

This experiment aimed to observe how leaf size affects overfitting in the DT algorithm. RMSE (Root Mean Squared Error) was used as the performance metric. We kept leaf size from 1 to 50 and analyzed how RMSE changed for both in-sample (training) and out-of-sample (testing) data. The experiment identified the leaf size at which overfitting started to occur.

2.2 Experiment 2: Bag Learner and Overfitting

For this experiment, we evaluated how bagging (using BagLearner with DTLearner as the base learner) influenced overfitting. Bagging combined 20 decision trees, each trained on random subsets of the training data. RMSE was calculated for varying leaf sizes to determine if bagging reduced overfitting and if it eliminated it entirely.

2.3 Experiment 3: DT vs. RT Comparison

The third experiment compared DTLearner and RTLearner using Mean Absolute Error (MAE) as a metric over multiple trials. We also measured consistency by evaluating how MAE varied across 20 trials. Additionally, other metrics such as Maximum Error (ME) were considered to analyze the differences between the two learners.

2.4 Dataset

The dataset used for this project is the "Istanbul.csv," which contains historical financial data with multiple worldwide index returns over several days. This dataset includes various columns representing different financial indices, such as the S&P 500 (SP), DAX, FTSE, and MSCI Emerging Markets (EM), among others. The goal is to predict the return of the MSCI Emerging Markets (EM) index (the target variable, Y) based on the other index returns (the feature variables, X). For the analysis, the date column was removed, and the remaining data was split into 60% training and 40% testing sets.

3 DISCUSSION

3.1 Experiment 1 : DT Learner and Overfitting Analysis

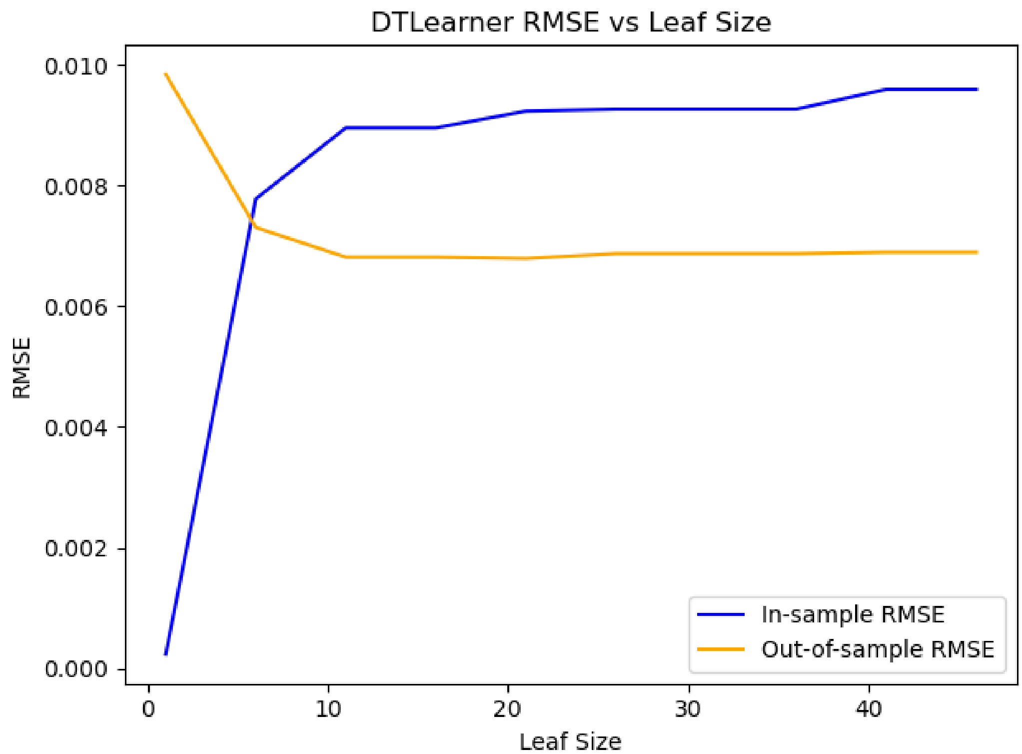

Figure 1: Experiment 1 - Decision tree Learner

Figure 1 indicates that as leaf size increased, the in-sample RMSE increased as well, suggesting lower training accuracy. However, beyond a certain point (around leaf size 6), the out-of-sample RMSE started to increase, indicating overfitting. This result confirms that smaller leaf sizes enable the model to capture finer patterns but also noise, making it less generalizable to unseen data. Hence, setting the leaf size too low results in overfitting, leading to high error rates on out-of-sample data.

3.2 Experiment 2

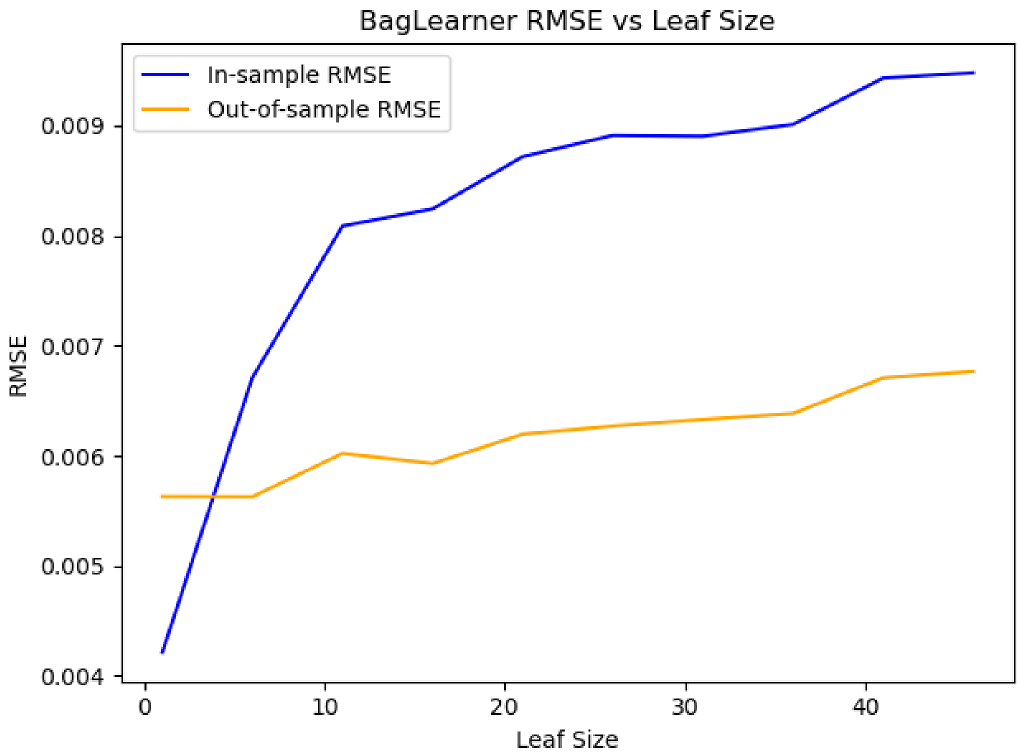

Figure 2 Experiment2 Bag Learner

Figure 2 illustrates that while the in-sample RMSE increases as the leaf size grows, the out-of-sample RMSE remains relatively stable across all leaf sizes. This stability indicates that bagging effectively reduces overfitting, unlike the DTLearner from Experiment 1, where overfitting was evident at smaller leaf sizes. The slight increase in out-of-sample RMSE at larger leaf sizes suggests that bagging doesn't completely eliminate overfitting but significantly mitigates it. Overall, bagging enhances model generalization, maintaining more consistent performance on unseen data compared to using DTLearner alone.

3.3 Experiment 3

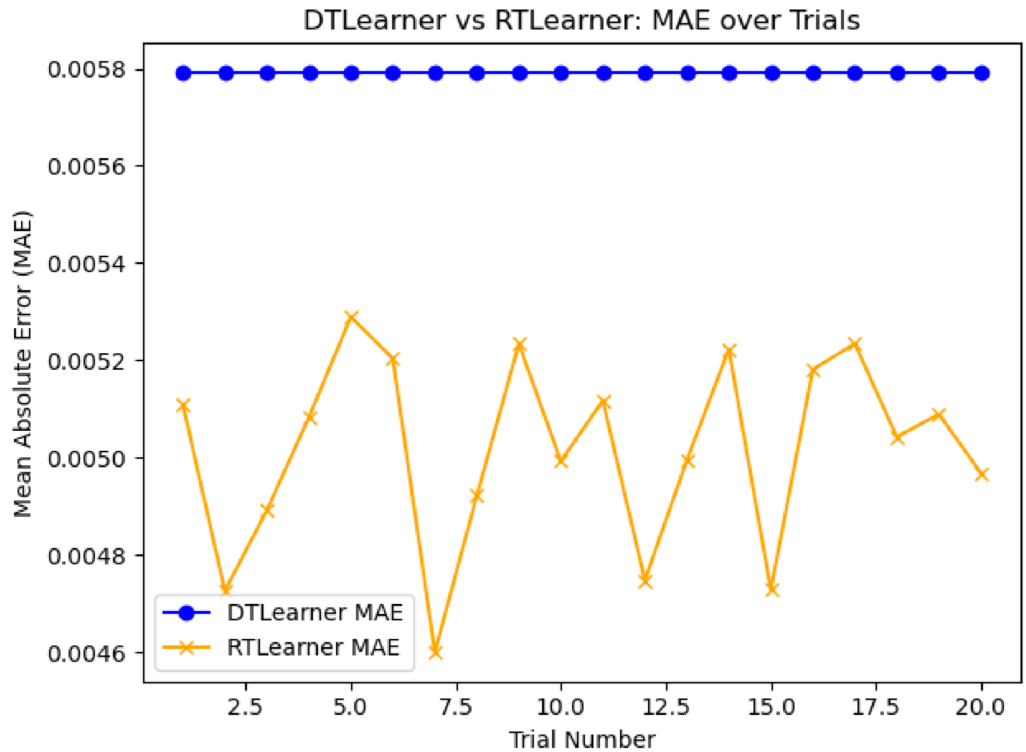

Figure 3 Experiment 3 DT vs RT MAE

Figure 3 reveals that RTLearner exhibits more variability in Mean Absolute Error (MAE) across trials compared to DTLearner, which maintains a consistent MAE throughout. Despite this variability, RTLearner generally achieves a lower MAE than DTLearner in most trials, indicating better predictive performance. The consistency of DTLearner suggests that it might be less influenced by random fluctuations, but its higher MAE means it is less accurate on average. In contrast, the random feature selection in RTLearner allows it to capture more diverse patterns, leading to overall improved accuracy, albeit with some fluctuations. This suggests that RTLearner may offer better performance in scenarios where adaptability and accuracy are prioritized over consistency.

4 SUMMARY

In this project, we analyzed the performance and overfitting behavior of various learners, including DTLearner, RTLearner, BagLearner, and InsaneLearner. Experiment 1 showed that DTLearner tends to overfit with smaller leaf sizes, resulting in decreased out-of-sample performance as the model complexity increased. In Experiment 2, we demonstrated that bagging significantly reduced overfitting, with BagLearner maintaining a more stable and lower out-of-sample RMSE across different leaf sizes compared to DTLearner. Experiment 3 revealed that RTLearner generally performed better in terms of MAE, despite exhibiting more variability, while DTLearner was more consistent but less accurate. Overall, the results highlight that ensemble methods like bagging are effective in mitigating overfitting, and random trees can provide more adaptable and accurate predictions. For real-world applications, these findings suggest that using ensemble learners or random tree methods may offer more robust and reliable performance, especially in complex datasets.