CS7646 Machine Learning For Trading

Project 3 Assess Learners

Yijie Zhao
yzhao633@gatech.edu

Abstract— This report has 4 sections: introduction, methods, discussion and summary. 4 tree learner models implemented and 4 quantitative metrics calculated to analyze and discuss.

1 INTRODUCTION

This project includes implementation of 4 main learners models (Decision tree learner model, Random tree learner model, Bag learner(bag size = 20) model as well as Insane Learner model). Another file testlearner.py has code for plotting graphs. There learners work for the data file Istanbul.csv.

The testlearner.py can be run by:

Python testlearner.py Data/Istanbul.csv

2 METHODS

X:

ISE-TL	ISE-USD	SP	DAX	FTSE	NIKKEI	BOVESPA	EU

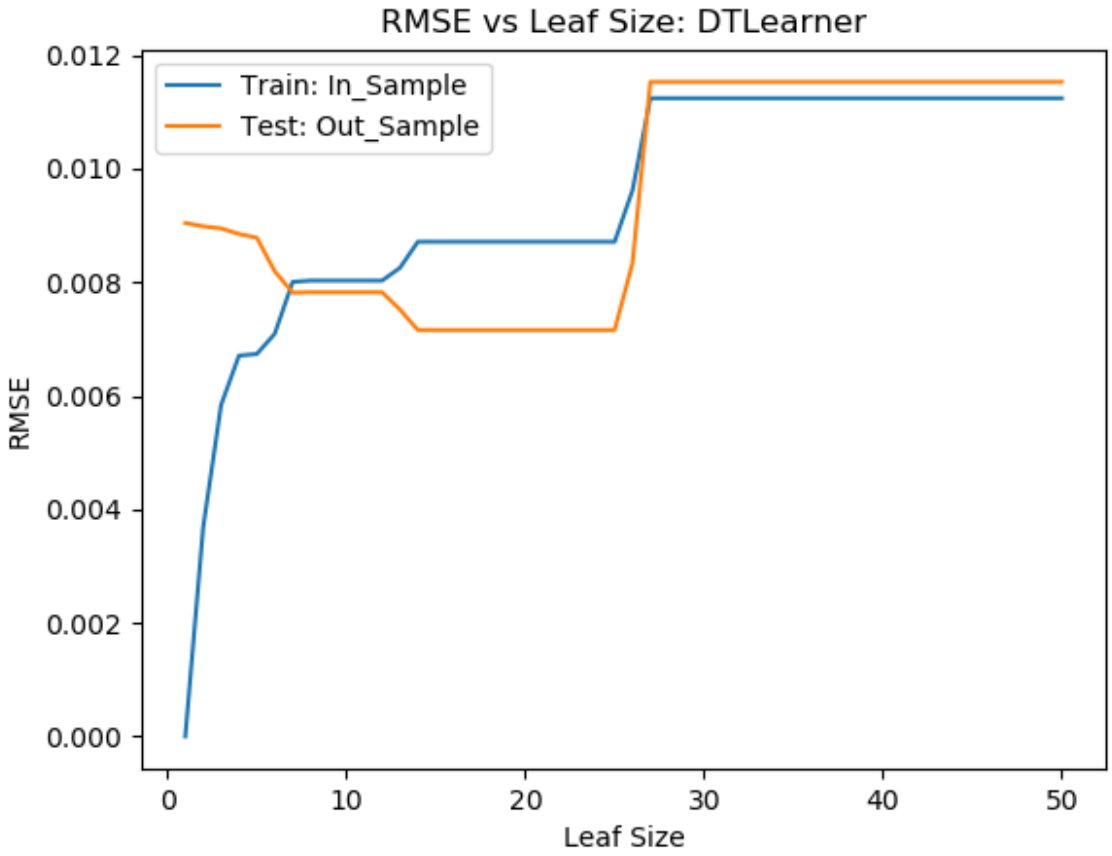
Y:

EM

The quantitative matrix RMSE, MEA, build time and average depth are calculated to analyze the overfitting in terms of leaf size and comparison of decision tree learner and random tree learner.

3 DISCUSSION

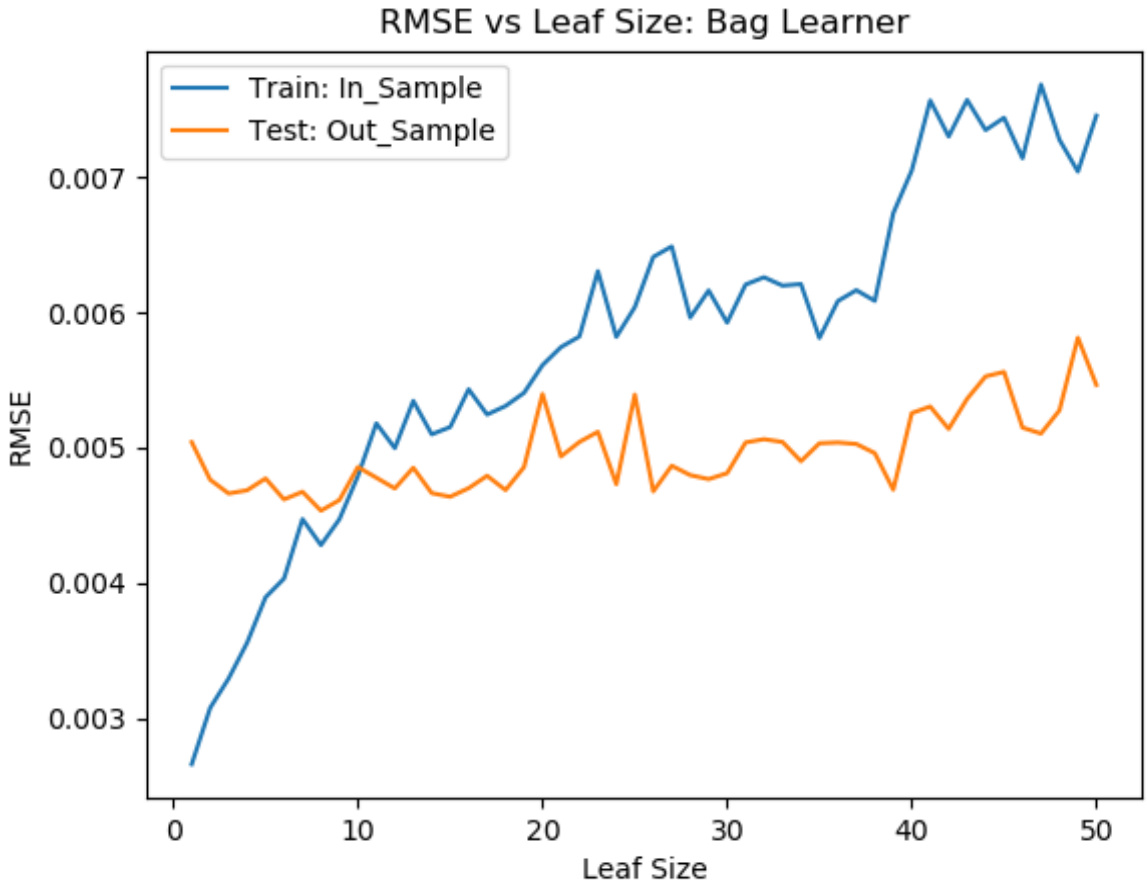
3.1 Experiment 1



Istanbul.csv dataset was trained and tested in the Decision Tree model as leaf size ranging from 1 to 50. The Root Mean Square Error(RMSE) is plotted for the Training set and Testing set for comparison.

As per the graph, overfitting does exist in terms of leaf size. The RMSE of in the sample data increases as we increase the number of leaf size, and The RMSE of out of the sample data increases as we increase the number of leaf size at the point (left size >= 25), which indicates the overfitting.

3.2 Experiment 2

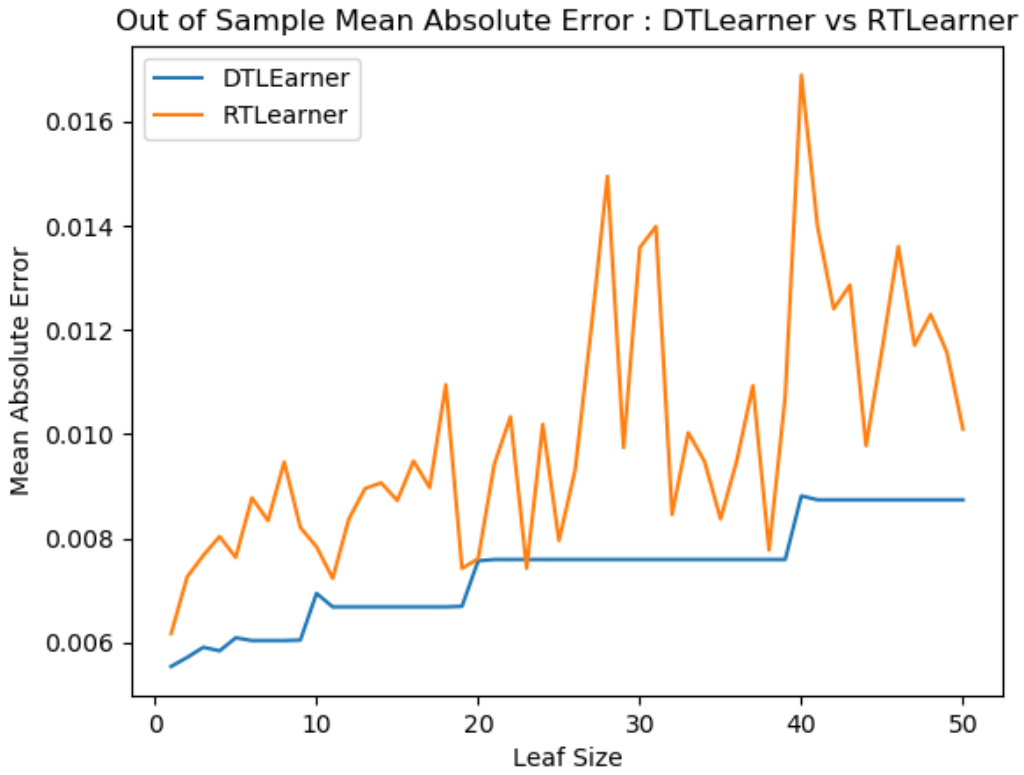


Istanbul.csv dataset was trained and tested in the Decision Tree model as leaf size ranging from 1 to 50 with a bag size of 20.The Root Mean Square Error(RMSE) is plotted for the Training set and Testing set for comparison.

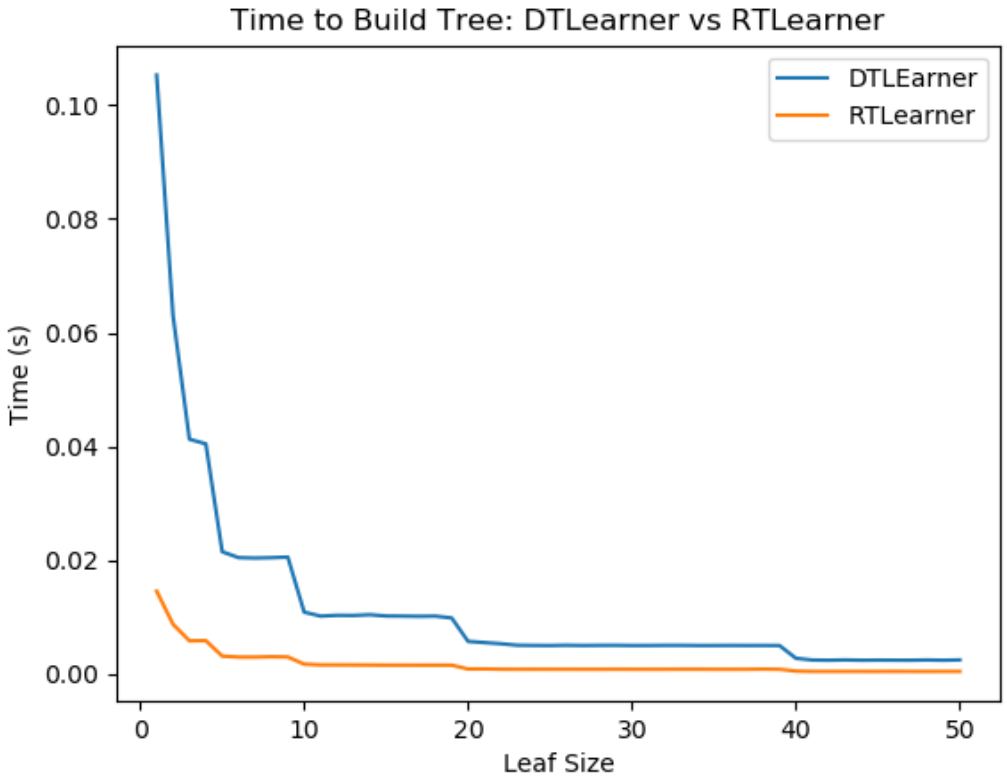
As per the graph, overfitting does exist in terms of leaf size. The RMSE of in the sample data increases as we increase the number of leaf size, and The RMSE of out of the sample data increases as we increase the number of leaf size at the point (left size >= 9), which indicates the overfitting.

Compared with Experiment 1, I concluded that bagging does reduce overfitting with respect to leaf_size and bagging cannot eliminate overfitting with respect to leaf_size.

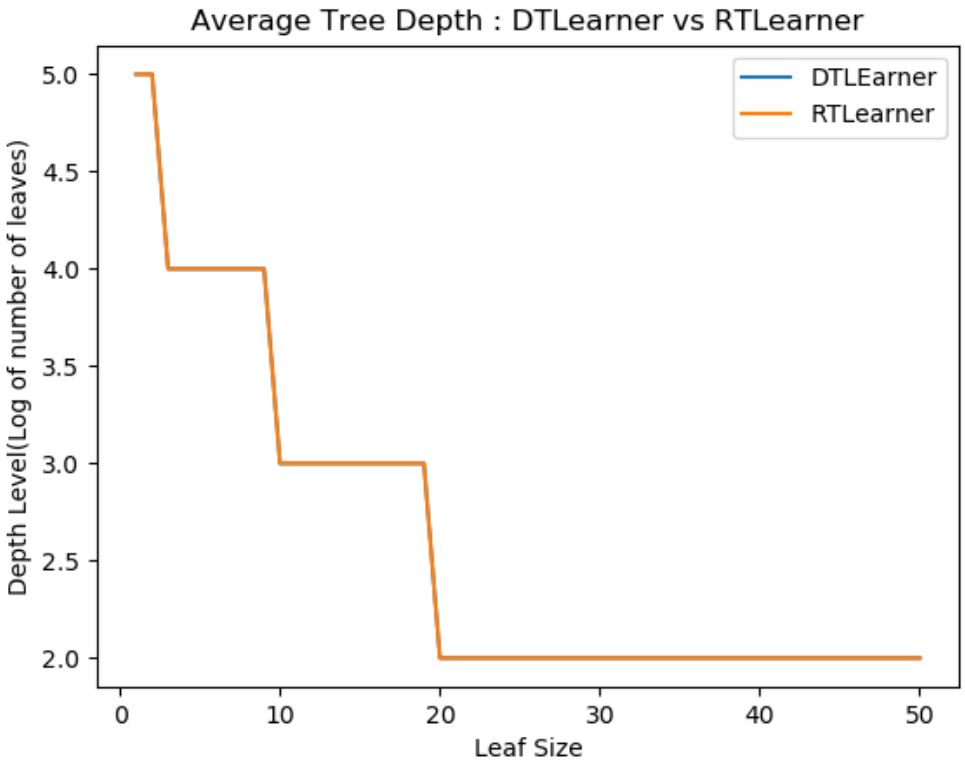
3.3 Experiment 3



I used out of sample Mean Absolute Error as my first quantitative metric to compare Decision Tree Learner and Random Tree Learner. The graph illustrates that the Decision tree model has better accuracy than the Random tree model since the Mean Absolute Error of Random tree is significantly and consistently above that of Decision tree.



I used Time to Build Tree as my second quantitative metric to compare Decision Tree Learner and Random Tree Learner. As per the graph above, the decision tree takes significantly more time than the random tree as the leaf size decreases. This can be explained by the fact that in the decision tree, we calculate correlation value every time for all the points in the dataset, but on the other hand, we decided to just do a random split for the random tree.



I used Average Tree Depth as my third quantitative metric to compare Decision Tree Learner and Random Tree Learner. They are both binary trees so I take the log(exp) of the total number of leaves. As per the graph above, the decision tree and the random tree have almost the same downward average depth as the leaf size increases. This can be explained by the fact that a smaller leaf size would require more nodes and decisions, which would increase the tree depth.

4 SUMMARY

In terms of overfitting, bagging does reduce overfitting with respect to leaf sizes but bagging cannot eliminate overfitting with respect to leaf sizes.

In terms of Decision Tree Learner and Random Tree Learner comparison, Decision tree has better accuracy but more building time, while Random Tree has worse accuracy but less building time. As for tree depth, both models have the same downwards average depth as the leaf size increases.