PROJECT 3: Assess Learners

By: Zi Liu
GTID: 903476881

ABSTRACT

In this project, multiple experiments are performed to explore the advantages and disadvantages for the classic decision trees, random trees and bagged trees. There are different comparisons made among these machine learning algorithms. Different characteries such as overfitting, accuracy and time efficiency will also be discussed throughout this report.

INTRODUCTION

Decision tree is one of the most powerful predictive modelling approaches in machine learning. The tree model uses the observation values and makes the predictions on the item's target values; these algorithms are appliable to solve both classification and regression problems. Note this project considers regression problem only, not classification. Therefore the predictions are a continuous numerical result instead of a discreet result.

METHOD

This project implements and evaluates the classic Decision Tree learner, a Random Tree learner, a Bootstrap Aggregating learner by using Python. The algorithm used in Decision Tree learner is based on the paper written by JP Quinlan. The experiments are made based on the dataset 'Istanbul.csv". The overall objective is to return the Emerging Market Index (known as "EM") using the several different X features provided in the dataset. The outcomes from the different decision tree models will be examined against the actual results in order to evaluate the accuracy for these learning models. Also note in this experiment, 60% of the dataset were randomly selected for training purpose and 40% were utilized for testing.

DISCUSSION

Experiment 1

Research and discuss overfitting as observed in the experiment. Support your assertion with graphs/charts. Does overfitting occur with respect to leaf_size? For which values of leaf_size does overfitting occur? Indicate the starting point and the direction of overfitting. Use RMSE as your metric for assessing overfitting.

The following graph 'Figure_1' showcases the effects of the size of the leaf in the classic decision tree model. There are 20 learners examined in this experiment, using 60% of the training data from Istanbul.csv, and the leaf size is ranging from 1 to 20. Both of the in-sample and out-of-sample errors were calculated using RMSE as a metric.

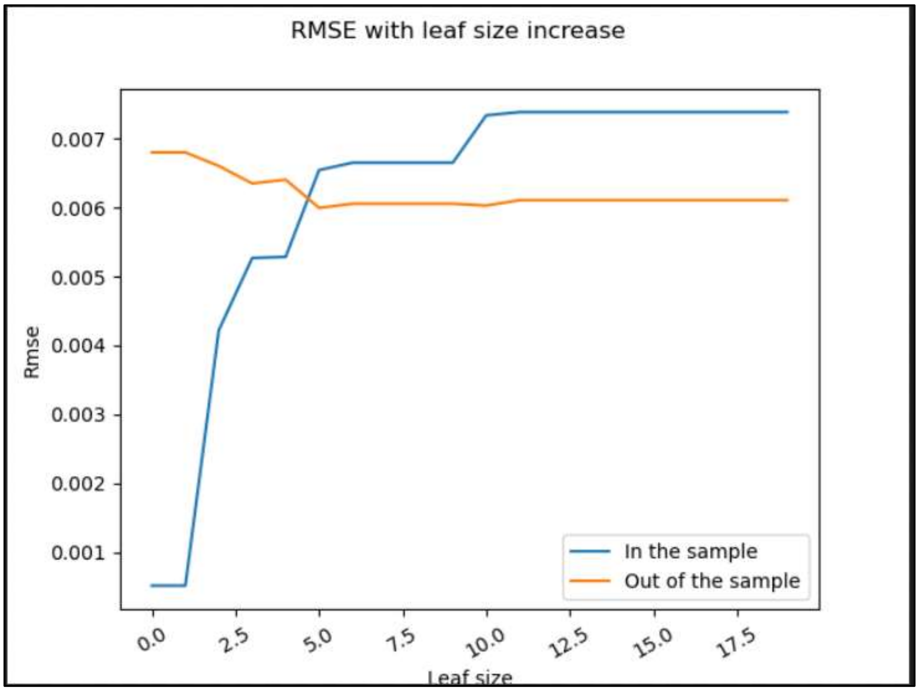

Figure_1

As the graph presented, there is a clear indication of overfitting in this model. The overfitting can be identified as the area where as the in-the-sample RMSE increases, and the out-of-sample RMSE slightly decreases. As the leave_size gets smaller, the overfitting impacts more greatly. Approximately the overfitting occurs on leaf_size equals to 5.

The reasoning behind this behavior is that the algorithm forces to keep partitioning the samples into smaller group until the desired leaf size is. In the extreme example where the leaf_size equals to 1, in-the-sample RSME is at the lowest point because prediction points are grouped to their own nodes. As a result, the model fits the training data perfectly, however it generates a poor outcome for the out-of-sample data (The RSME in out-of-sample is at its peak when leaf size equals to 1).

Experiment 2

Research and discuss the use of bagging and its effect on overfitting. (Again, use the dataset Istanbul.csv with DTLearner.) Can bagging reduce overfitting with respect to leaf_size? Can bagging eliminate overfitting with respect to leaf_size? To investigate this, choose a fixed number of bags to use and vary leaf_size to evaluate. If there is overfitting, indicate the starting point and the direction of overfitting.

Experiment 2 utilizes the bagging technique in the model to test whether it can reduce, or even eliminates the overfitting issue. In the bagged tree, fixed number of 20 "bags" are used based on the randomly selected training data. The results of comparing all learners using in-the-sample and out-of-sample RMSE are shown in the following "Figure_2".

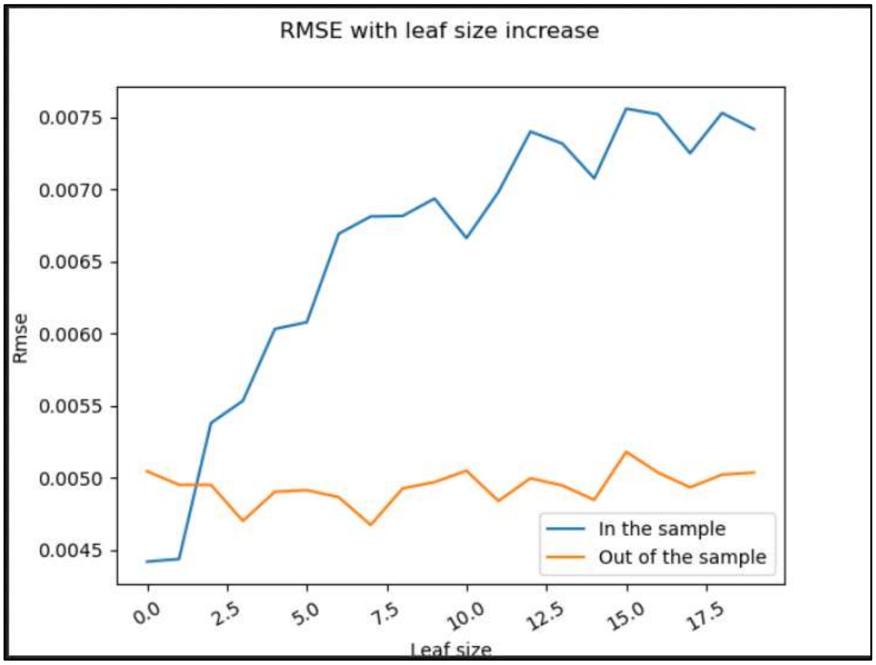

Figure_2

As the graph presented, there is no significant differences for the in-the-sample line comparing to the one in experiment 1. However, the out-of-sample line flattens from leaf size 1 to 20, meaning its RMSE did not change greatly as the in-the-sample changed. The bagged tree model definitely has a more promising outcome which the overfitting issue gets reduced significantly. Another observation from this experiment is that the RMSE of in-the-sample increases quickly as the leaf size increases. This is an indication that the decision tree model is underfitting the data and it does not produce predictions with high accuracy. On the other hand, the RMSE of the out-of-sample does not changes greatly as the leaf size increases, meaning the prediction outcome remains quite independently regardless of the leaf size. In this case the model does not underfitting the data. The contradicted observations from the both lines lead to a possibility that the Istanbul data is not distributed evenly among the bags, which can create a bias result. Therefore, more investigations are high recommended before drawing any conclusion.

Experiment 3

Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). Provide at least two new quantitative measures in the comparison. Using two similar measures that illustrate the same broader metric does not count as two separate measures. (For example, do not use two measures for accuracy.) Provide charts to support your conclusions. In which ways is one method better than the other?

For comparing the DTLearner model vs the RTLearner model quantitatively, two different metrics are utilized in this experiment: 1. the RMSE comparison for in both testing data and the training data, and the time difference spent in training and query for the two models.

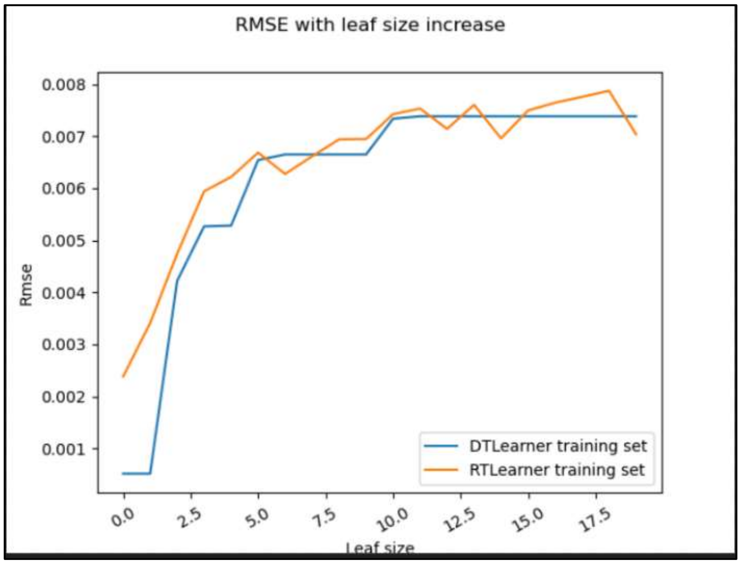

Figure_3

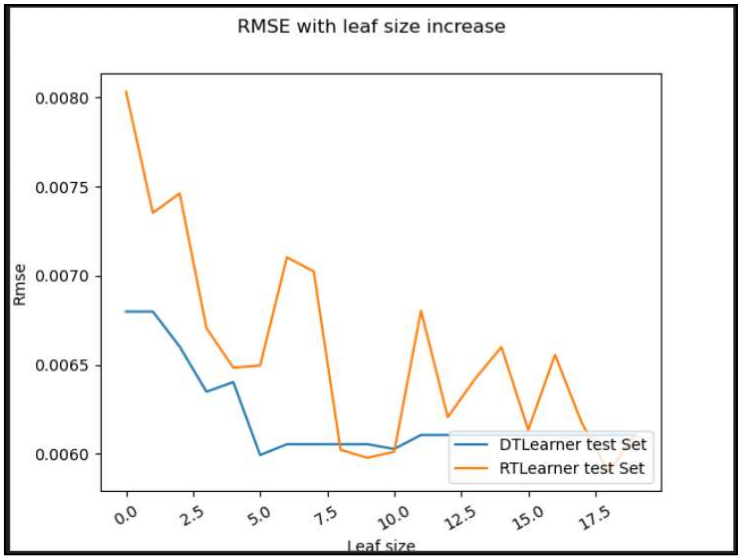

Figure_4

Figure_3 and Figure_4 above indicate the RMSE for the in-the-sample and the out-of-sample results in corresponding to leaf_size from 1 to 20 using the two models. For the in-the-sample graph (Figure_3), both learner lines increase as the leaf_size increases. It is expected and the reasoning has been explained in experiment 1. From Figure_4 using the out-of-sample testing, it is obvious that the RMSE for DTLearner is generally less than the one for RTLearner. It is an indication that the DTLearner might generate a more accurate prediction and its outcomes are relatively more stable.

The second metric used in experiment 3 is the time performance difference for both models. Figure_5 indicates the time differences between DTLearner and RTLearner when training the Istanbul data. As shown in the graph, the time difference line is above 0, it means DTLearner in general consumes more time comparing to RTLearner in the training process. This observation makes sense because in DTLearner, it requires to calculate the correlations between X factors and Ys to determine the best X feature on each level of the tree splitting. This step is voided in the RTLearner since the split takes place randomly. As a result, the accuracy of the prediction in RTLearner is sacrificed in returns for the greater efficiency during data training.

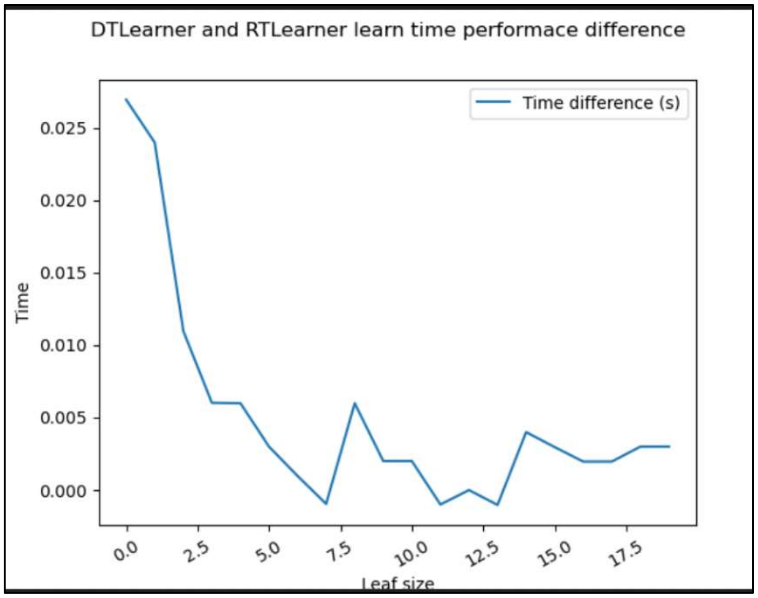

Figure_5

Figure_6 indicates the time differences between DTLearner and RTLearner when making the query for prediction in the Istanbul data. In this case, majority of the data points are below 0 meaning in general the DTLearner spent less time in the query performance in comparison to RTLearner. This observation corresponds to the mathematical theory behind these two models as well because once the model is trained, DTLearner only consists one tree and therefore the query can be made fairly quickly. For RTLearner, it is required to first calculate all trees, compute the mean value and eventually return the final result. This is definitely more time consuming.

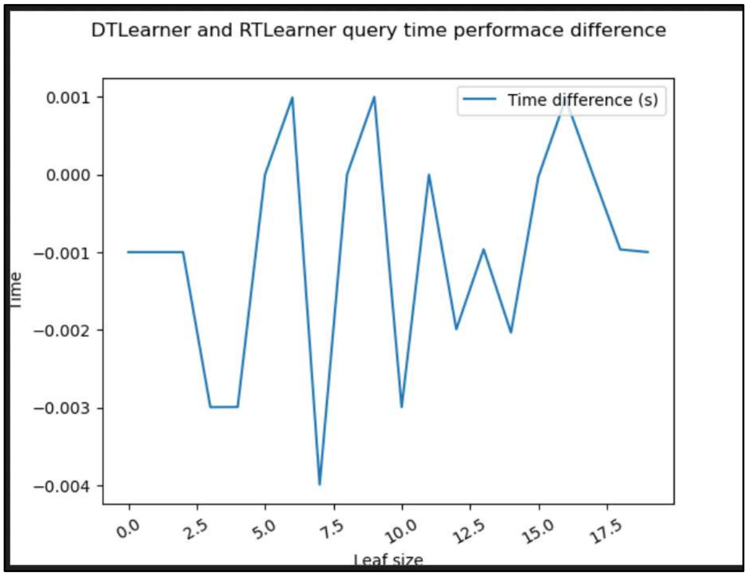

Figure_6

SUMMARY

In summary, different types of decision trees are examined in this project. The experiments include how the leaf size impacts the overfitting of the models, and whether the bootstrap method in the bagged tree helps reduce this impact and produce a more accurate outcome. On the other hand, there are also comparisons made between the classic decision tree vs. the random forest. The experiment results indicate that in the model training process, classic decision tree model is time consuming comparing to the random forest, but it becomes more time efficient in the later query process.