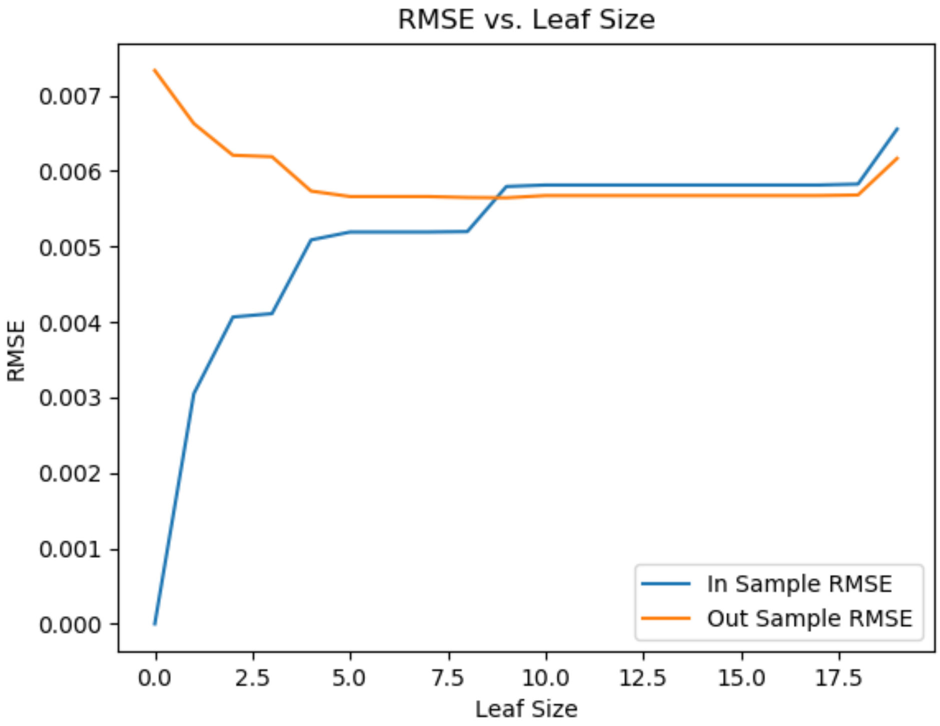
CS 7646 Machine Learning for Trading

Assignment 3: Assess Learners report
Name: Sagarika Srishti
GTID: ssrishti3

1. Does overfitting occur with respect to leaf_size? Use the dataset instanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use the RMSE as your metric for assessing overfitting. Support your assertion with graph/charts. Don't use bagging.

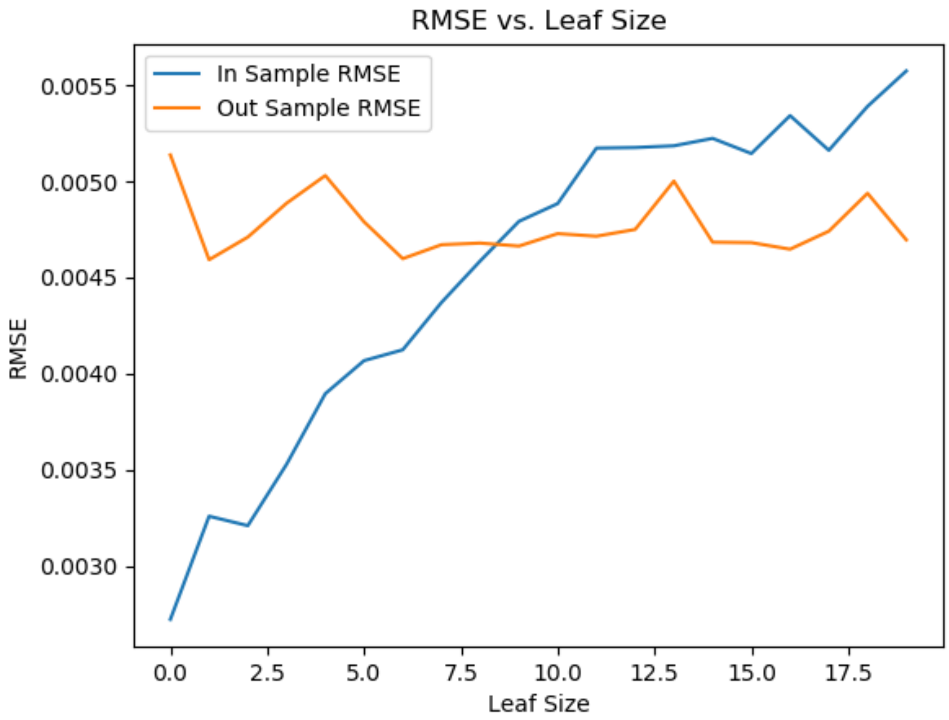


Ans:

Fig. 1: RMSE Error vs. Leaf Size for DT Learner

As can be seen from the graph, overfitting occurs with respect to leaf size. For leaf sizes of around 8 or less, RMSE of in sample is very low, and RMSE error of out sample is high. Very low RMSE error for training data and high RMSE error for test data is a characteristic of overfitting. This means that the model is fitting too well for the training data, at the cost of accuracy for test data. Hence, overfitting occurs for DT Learner for leaf sizes values less than 8.

2. Can bagging reduce or eliminate overfitting with respect to leaf_size? Again use the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf_size to evaluate. Provide charts to validate your conclusions. Use RMSE as your metric.



Ans:

Fig. 2: RMSE Error vs. Leaf Size for DT Learner with 10 bags

As can be seen from the above graph, Bagging does not eliminate overfitting but it helps in reducing it. After running BagLearner on DTLearner with 10 bags, we can see that while overfitting still occurs for DT Learner for leaf sizes less than 9, the difference between the RMSE errors for training and test data has decreased, which means that overfitting has decreased. While the initial difference between the in-sample and out-sample RMSE errors was almost 0.008 (as seen in Fig. 1 in the answer for Q1 above), the initial difference between the two RMSE errors in the case of bagging is around 0.005 (as seen in Fig. 2 above).

3. Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other? Provide at least two quantitative measures. Important, using two similar measures that illustrate the same broader metric does not count as two. (For example, do not use two measures for accuracy.) Note for this part of the report you must conduct new experiments, don't use the results of the experiments above for this(RMSE is not allowed as a new experiment).

Ans: Below is a plot between the training time and the training data size for DTLearner and RTLearner. As can be seen from the plot, the training time of RTLearner is considerably lesser than that of DTLearner. This is because in DTLearner, we need to calculate the correlation values to determine the best feature of each row, the time needed for which keeps increasing with the training data size. On the other hand, for RTLearner, the best feature is just selected at random. Hence, the difference between the time complexities of both the learners.

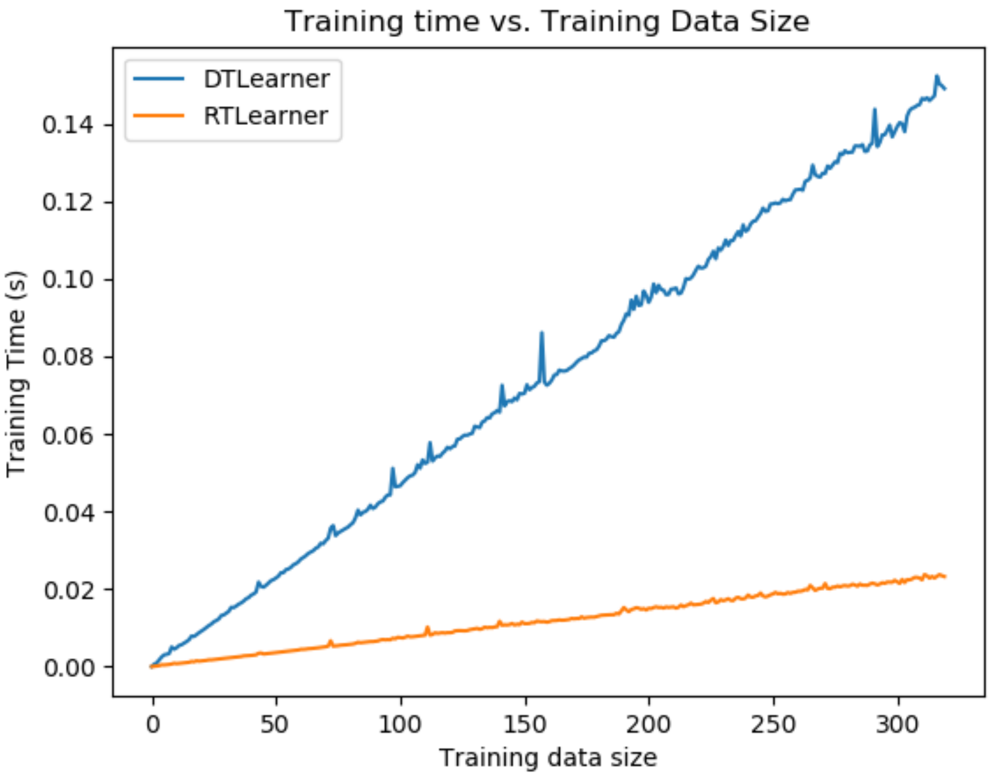


Fig. 3: Training Time vs. Training data size for DTLearner and RTLearner

Below is a plot between the tree size and training data size for DTLearner and RTLearner. As seen from the graph, the tree sizes for both the learners is almost the same for various training data sizes, except for a few variations in between. Hence, we can say that the space complexities for both DTLearner and RTLearner is almost the same.

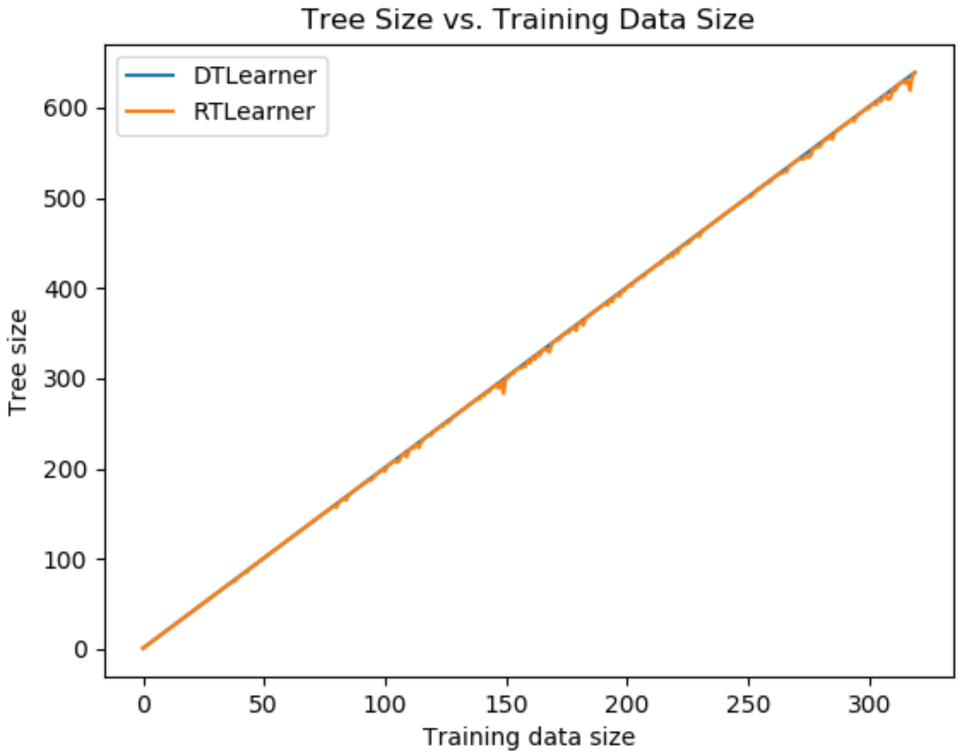


Fig. 4: Tree size vs. Training data size for DTLearner and RTLearner