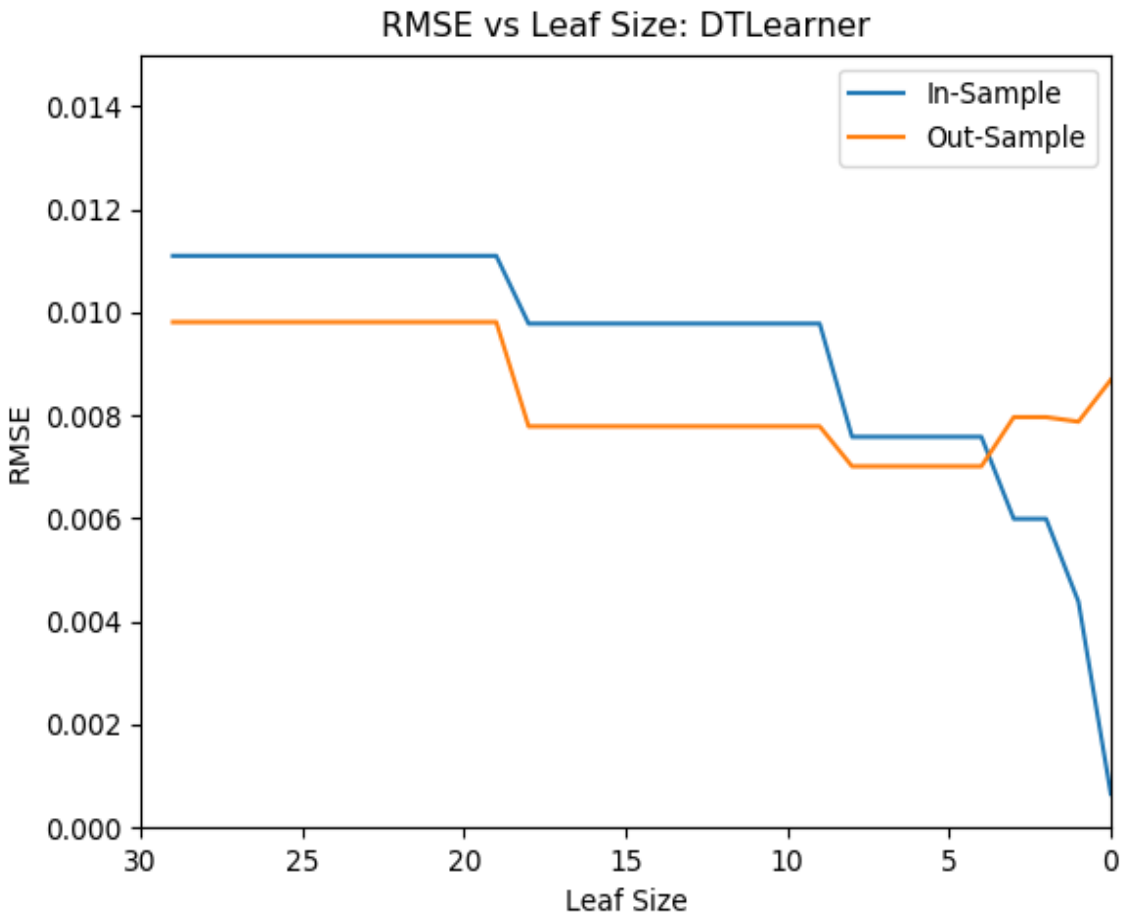
Tarun Pasumarthi
903085537
ML4T

Assignment 3: Report

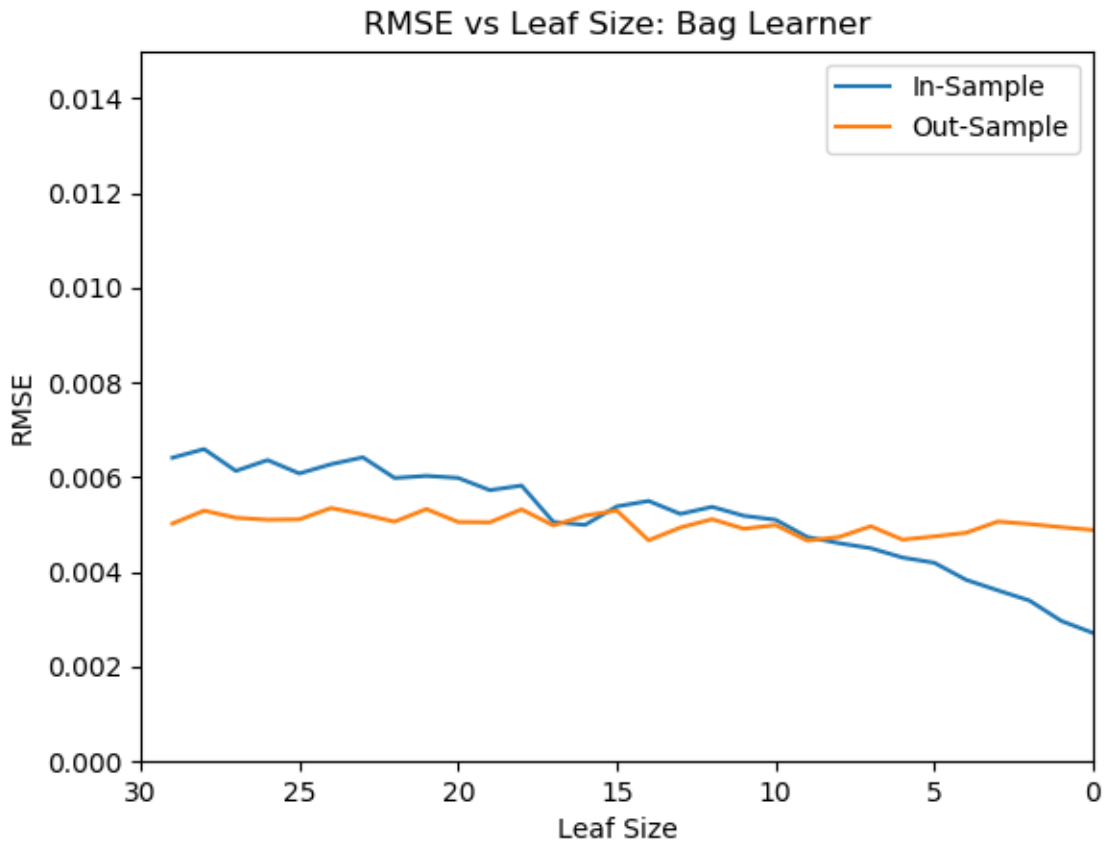
Does overfitting occur with respect to leaf_size? Use the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).



Yes, overfitting occurs for the DTLearner with respective to leaf size. In fact, according to the graph above, we see that as leaf size decreases from 30 to 0, the in-sample error continually decreases, but the out-sample error, initially decreases, and then starts to increase. The overfitting occurs when the out sample RMSE start to increase, and this happens approximately around leaf_size=4.

Tarun Pasumarthi
903085537

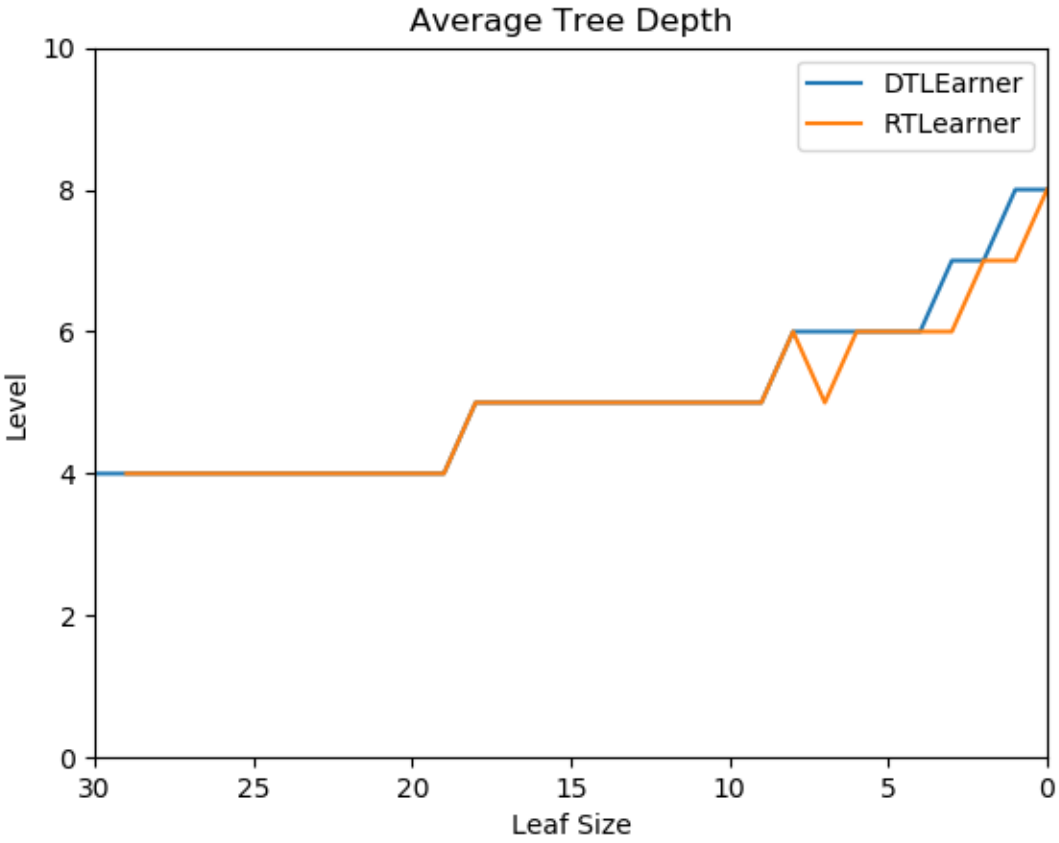
Can bagging reduce or eliminate overfitting with respect to leaf_size? Again use the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf_size to evaluate. Provide charts to validate your conclusions. Use RMSE as your metric.



As show by the graph above, the Bag Learner does, in fact, reduces overfitting. In this graph, the Bag Learner runs with a bag size of 15 with varying leaf sizes from 30 to 0. Comparing this graph to the DTLearner graph in part 1, we can see that overfitting still occurs around tree_size=8, where the in-sample error decreases but the out-sample error does not. However, unlike the DTLearner graph, the rmse does not shoot up, but rather levels off as the leaf size further decreases, so clearly the effects of overfitting are mitigated by the Bag Learner.

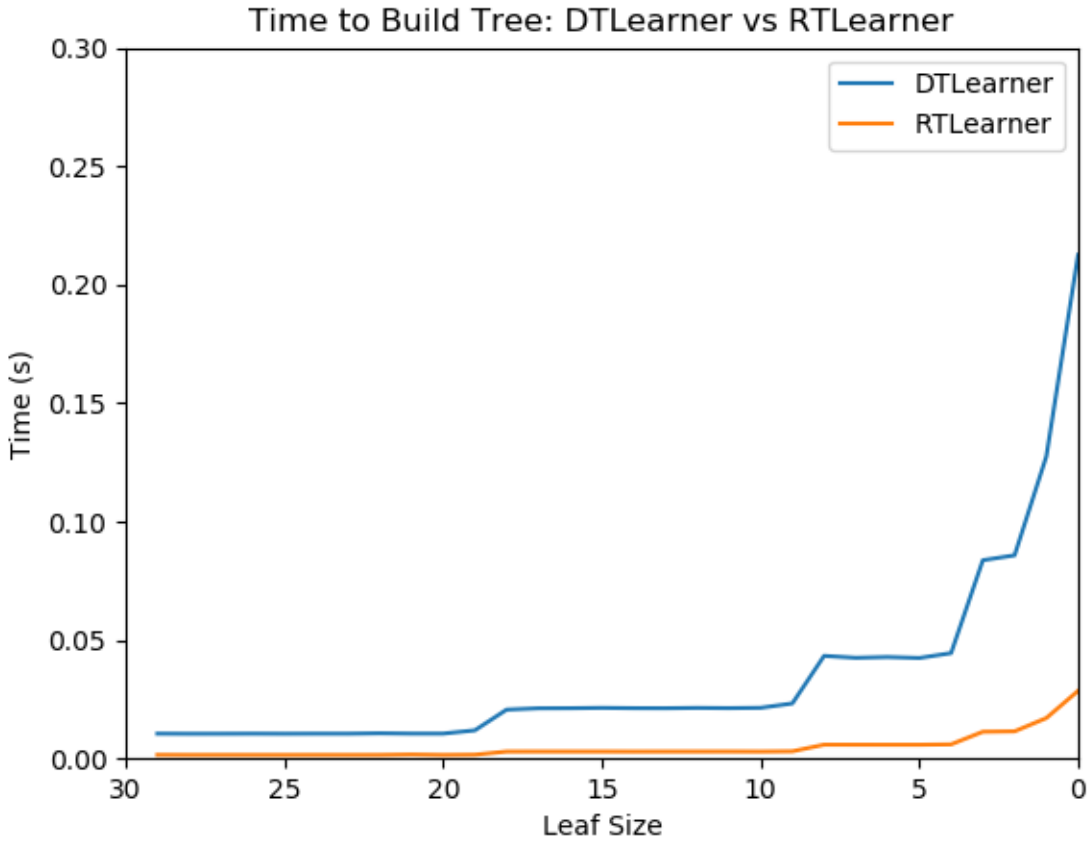
Tarun Pasumarthi
903085537

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).



One of the two quantitative metrics that I am measuring is the average depth of two trees as the leaf size increases. Since the decision tree and the random tree are both binary trees, I calculate the depth by taking the log (base 2) of the total number of leaves. As the graph above shows, both the Decision Tree and the Random Tree have about the same average depth at any given leaf size, and that both overall have more depth on average as the leaf size decreases. This intuitively makes sense as a smaller leaf size would require more nodes and decisions, which would increase the tree depth.

Tarun Pasumarthi
903085537
ML4T



The second quantitative metric that I measure is time to build either of the two trees. As seen from the graph above, the decision tree takes significantly more time than the random tree. This intuitively makes sense as the decision tree takes a longer time to pick which factor to make the decision on since it needs to compute the feature with the highest correlation. Additionally, as the leaf size decreases, the decision tree takes an even greater time than the random tree, as the decision tree needs to compute the correlation more frequently.