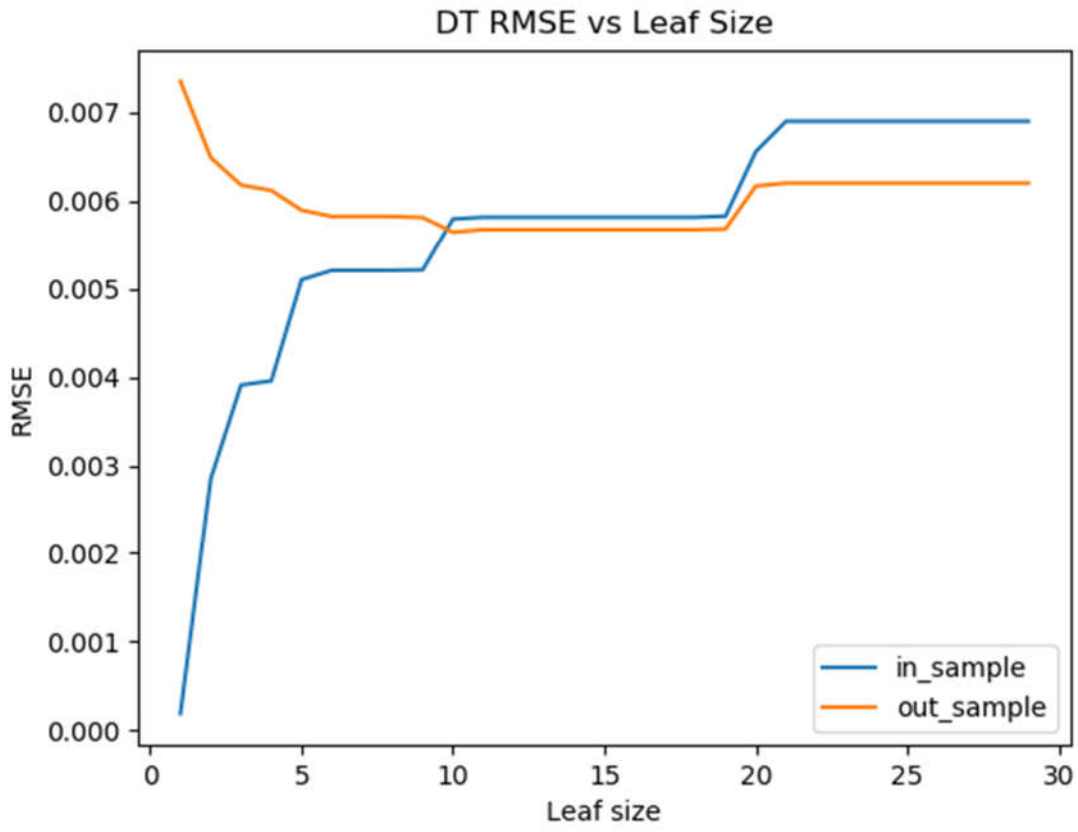
Name: Kok Jian Yu GTID: 903550380

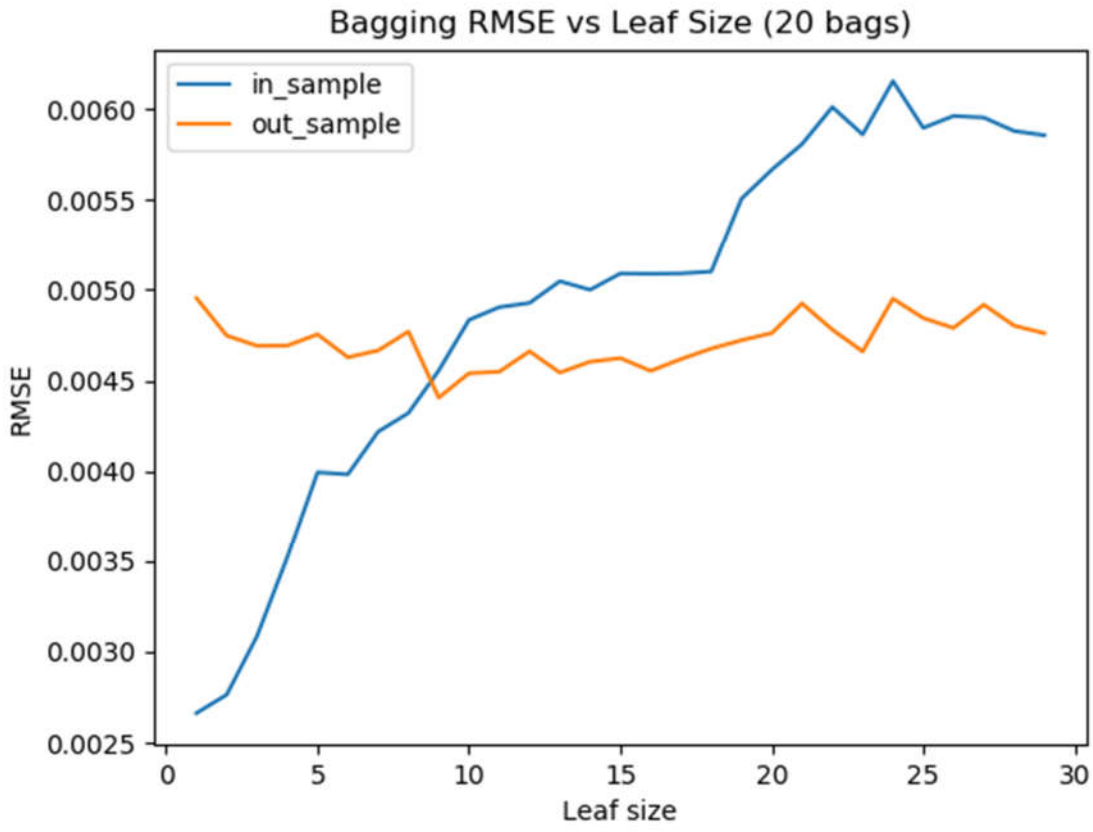
- Does overfitting occur with respect to leaf_size? Use the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting. Support your assertion with graphs/charts. (Don't use bagging).

  Yes. Overfitting occurs at around leaf sizes that is smaller than 10. This can be seen by looking at the graph below that visualizes the RMSE error of both the in_sample data and out_sample data. When leaf size is lower than 10, it can be seen that the RMSE for in_sample data is lower compared to RMSE for out_sample data. This is a sign of overfitting.



- 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.

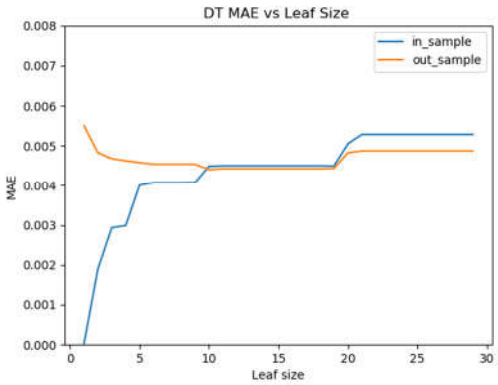
  No. When using bagging, the results still show signs of overfitting when leaf_size is lesser than 10. Therefore, this shows that while bagging does successfully reduce the overall RMSE, it is still not able to reduce or eliminate overfitting.

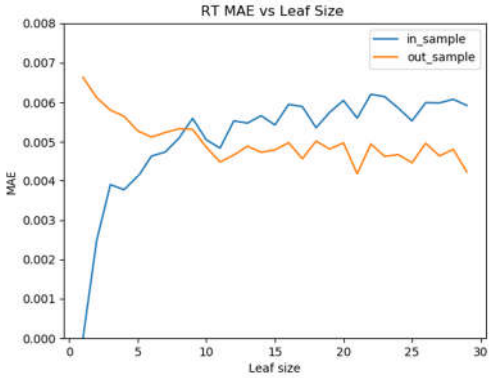


- Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other? Provide at least two quantitative measures. Important, using two similar measures that illustrate the same broader metric does not count as two. (For example, 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).

  Using Mean Absolute Error as the metrics, I got the following 2 graphs for MAE vs Leaf size for DT and RT.

  This is done by training the DT and RT with Istanbul dataset, and calculating the MAE using the prediction result.

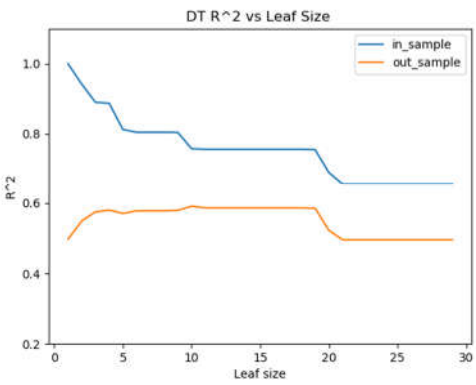


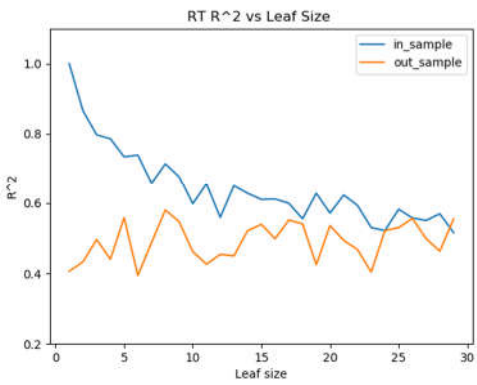


From the chart, It can be seen that DT error is smoother compared to RT, which means that it performs more consistently compared to RT with regards to leaf size. When the leaf size is small, DT performs better as it has a lower MAE for the out_sample. However, when the leaf size increases, the MAE becomes roughly the same. Therefore, taking into account both the error rate given small leaf size, and also the inconsistency in result, DT is better when comparing with MAE.

Another metrics I used is the R-squared error

This is done by training the DT and RT with Istanbul dataset, and calculating the R-Squared by calculating the correlation coefficient and squaring it.





R squared error shows how well the model prediction fits to the actual dataset. The higher, the better it fits. Therefore, when both the trees has a leaf size of 1, it completely fits the in_sample data set. From the two graphs, we can see that the R-squared error against the leaf size is more consistent for the DT than the RT. Furthermore, the R-squared error for the out_sample data is higher for the DT for most part, until when the leaf size reaches around 18, where the 2 model now have similar R-squared error.

From these observations, we can see the DT is better compared to the RT as the R-squared error is generally higher for the out_sample in DT compared to RT.

This shows that the DT's prediction is a closer match with the actual dataset compared to RT.