Steve Tang

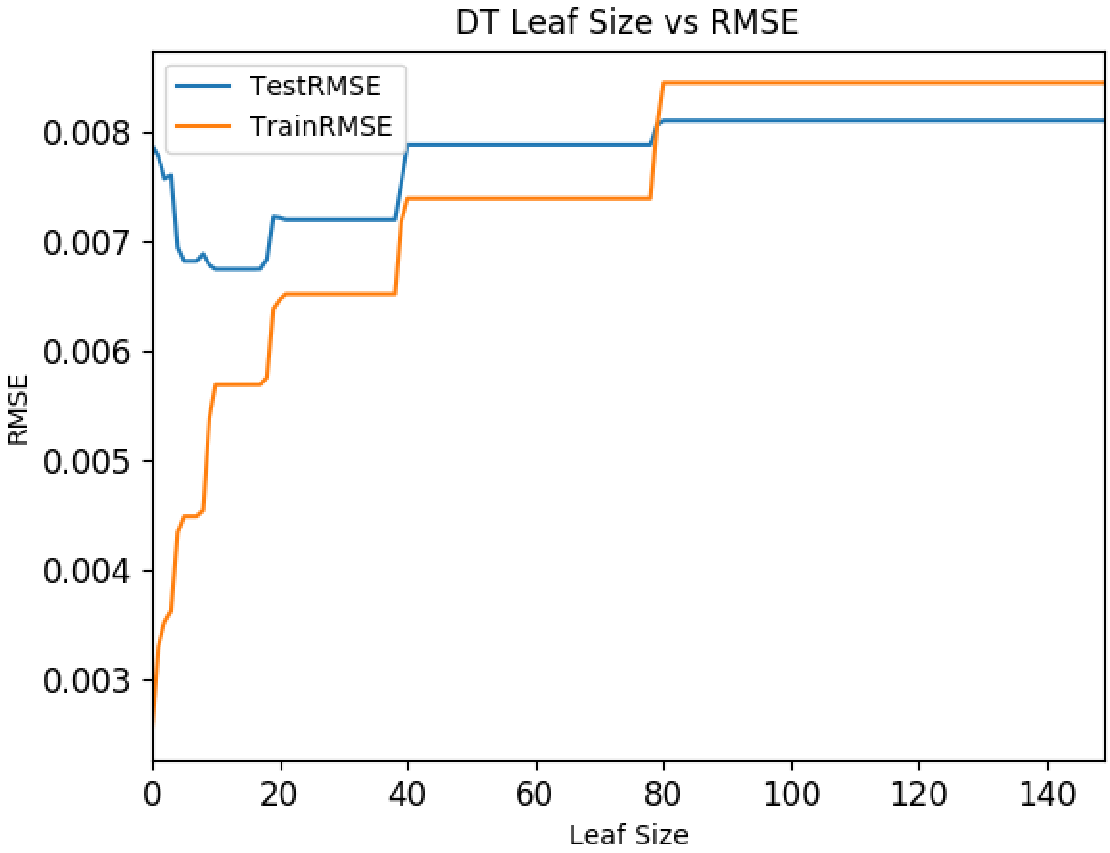
CS 7646

Assess Learners

Problem 1:

In the plot below, we have a graph showing the RMSE from the Istanbul dataset based on the leaf size. The question we would like to answer is if overfitting occurs with respect to leaf size. It can be seen from figure 1 that leaf size does influence the model overfitting. As we move from a larger to lower leaf size we see that out training error continues to get smaller while out test error grows. The values where this over fitting occurs would with leaf values of 1 to around 12 around when the training RMSE starts to drop from 0.00575.

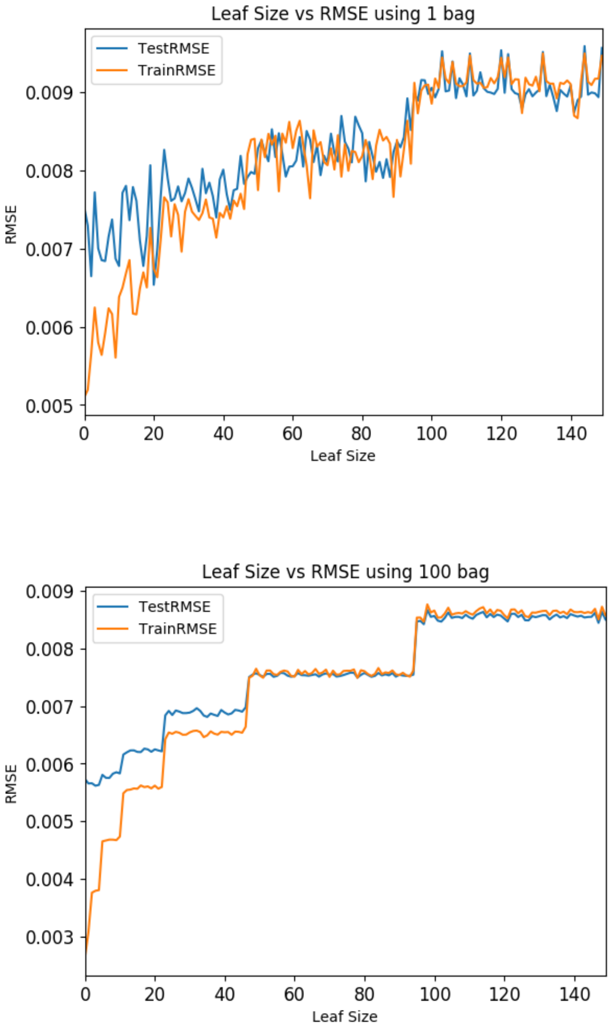
Figure 1: Leaf Size vs RMSE



Problem 2:

In the plots below, we have the Istanbul dataset running again with Decision Trees but this time we added the used of bagging. Figure two has example results with the number of bags set at 1, 10 and 100. Each bag randomly selected 60 percent of the training data at random with replacement. In the figures below, they follow the same trend as in figure1 where train RMSE decreases with lower leaf size but one different is that Test RMSE is also decreasing with leaf sizing meaning that the model has not over fit yet. Having many bags will allow us to sample the data and build multiple models and average them together to smooth out any harsh bias that may exist in a single model. The benefit of sampling from the data and combining multiple models means that we will have a model that will generalize the problem better compared to not using bagging.

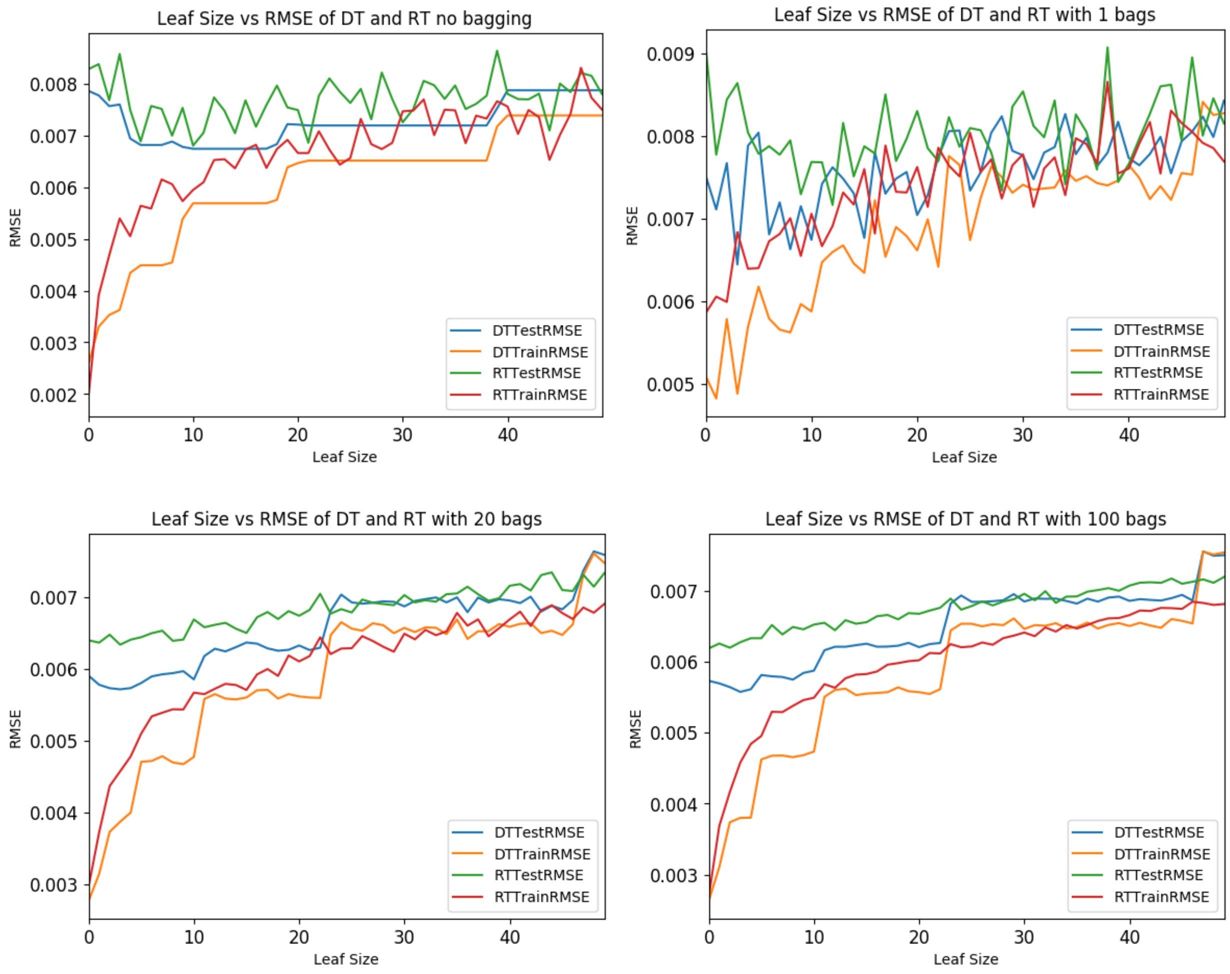
Figure 2: Leaf Size vs RMSE Given n Number of Bags



Problem 3:

If we were to be given no labels and were asked to pick out 2 lines that represent the Random tree and the other two lines that are from the decision tree this would seem like a simple and do able task. As we get to higher leaf size we can see the DT has longer flat lines while the random tree follows the trend but has a bit of noise. On one hand the great benefit of decision trees is that they are easy to follow and are simple to explain. A decision tree will require the least number of node to get to a leaf, as we are splitting at the best possible factor and splitting the data in half each time while with the random tree we are selectin two random values from the data and then averaging them mean there would be a lot of variance between our left and right tree splits. Since decision trees are split in half at each node they are shorted and easier to follower so when we need to query the speed will be very fast while with random forest our trees will be larger and the paths down the tree will be much longer making look up times slower. The advantages of random forest come in the training of the data. Rather than burn all the time what the best feature is, we can just choose it randomly to make it faster. Split value, rather than compute the median buy looking over all the values we can just select two random numbers. Generally, without any bagging we can see in figure 3 that design trees perform generally better then random trees but it may not be by much. When we add in bagging we can see that random trees give s more of a consistent result and the data from 100 makes it seem like random forest will perform better with more bags and larger leaf size. Since decision trees are smoother trends of over fitting is easy to spot where we may be a bit unsure of if random forest are also over fitting at lower leaves buts seem likely.

Figure 3 Comparing DT and RT with some bagging


Leaf Size vs RMSE of DT and RT no bagging
RMSE
Leaf Size
DTTestRMSE
DTTrainRMSE
RTTestRMSE
RTTrainRMSE
Leaf Size vs RMSE of DT and RT with 1 bags
RMSE
Leaf Size
DTTestRMSE
DTTrainRMSE
RTTestRMSE
RTTrainRMSE
Leaf Size vs RMSE of DT and RT with 20 bags
RMSE
Leaf Size
DTTestRMSE
DTTrainRMSE
RTTestRMSE
RTTrainRMSE
Leaf Size vs RMSE of DT and RT with 100 bags
RMSE
Leaf Size
DTTestRMSE
DTTrainRMSE
RTTestRMSE
RTTrainRMSE