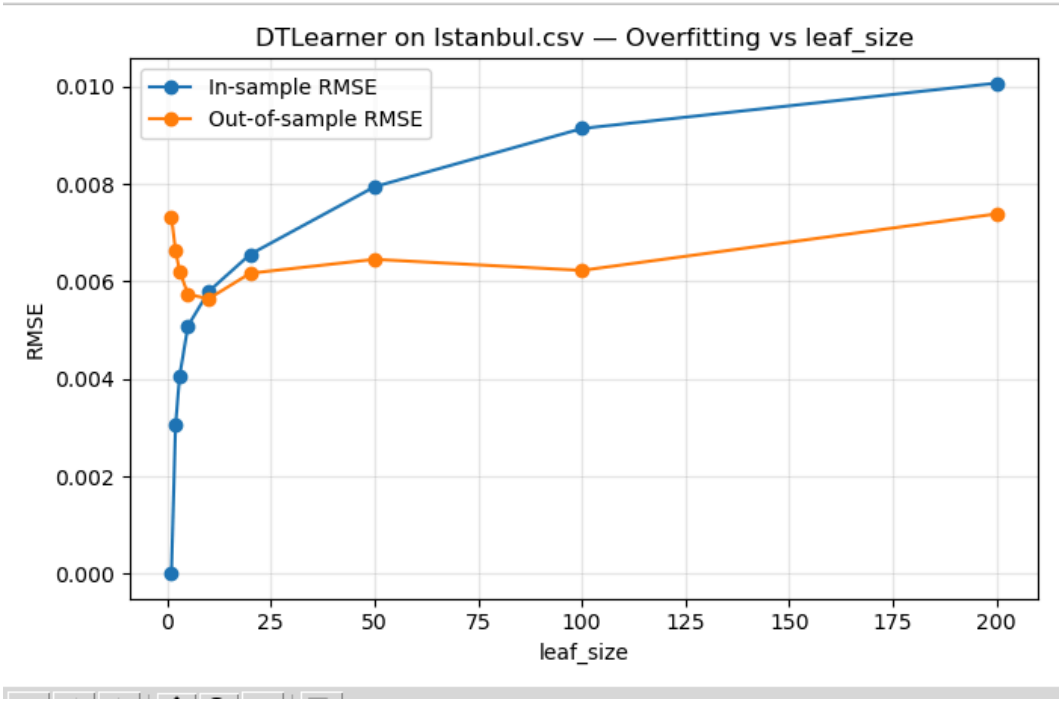
Project 3: Asses Learners

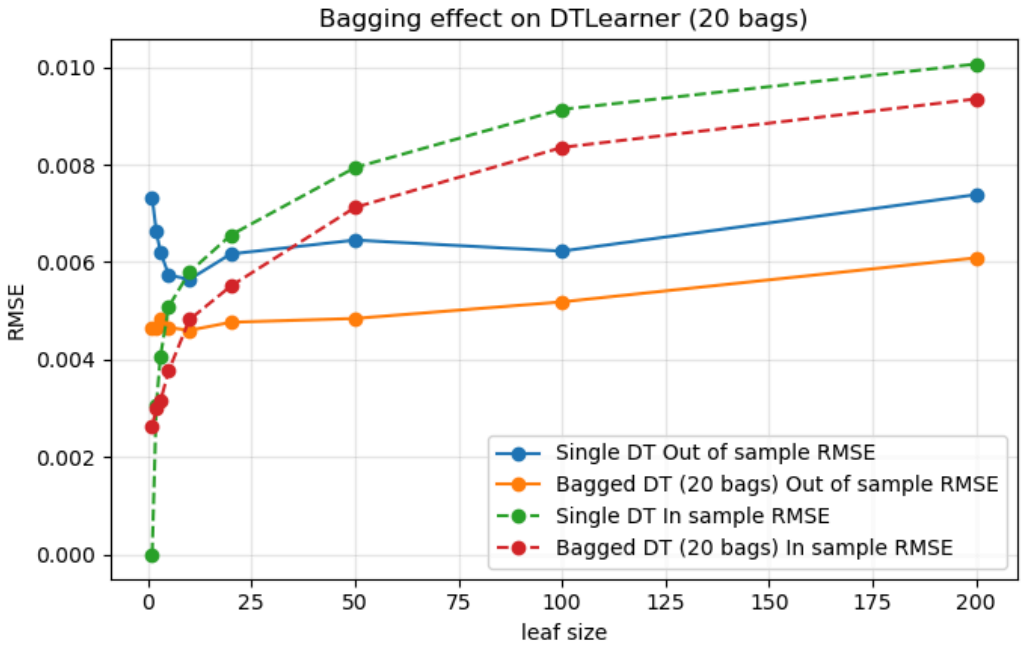
Cameron Railton
crailton3@gatech.edu

This In this project we implement four regression learners Decision Tree (DT), Random Tree (RT), Bagging, and an InsaneLearner (bags of RTs) and evaluate them on Istanbul index returns to predict EM. With randomized 60/40 train test splits, we report RMSE and correlation. Random Trees and bagging shrink the generalization gap versus Decision Trees, and the InsaneLearner achieves the best out-of-sample results (RMSE ≈ 0.004; corr ≈ 0.89)

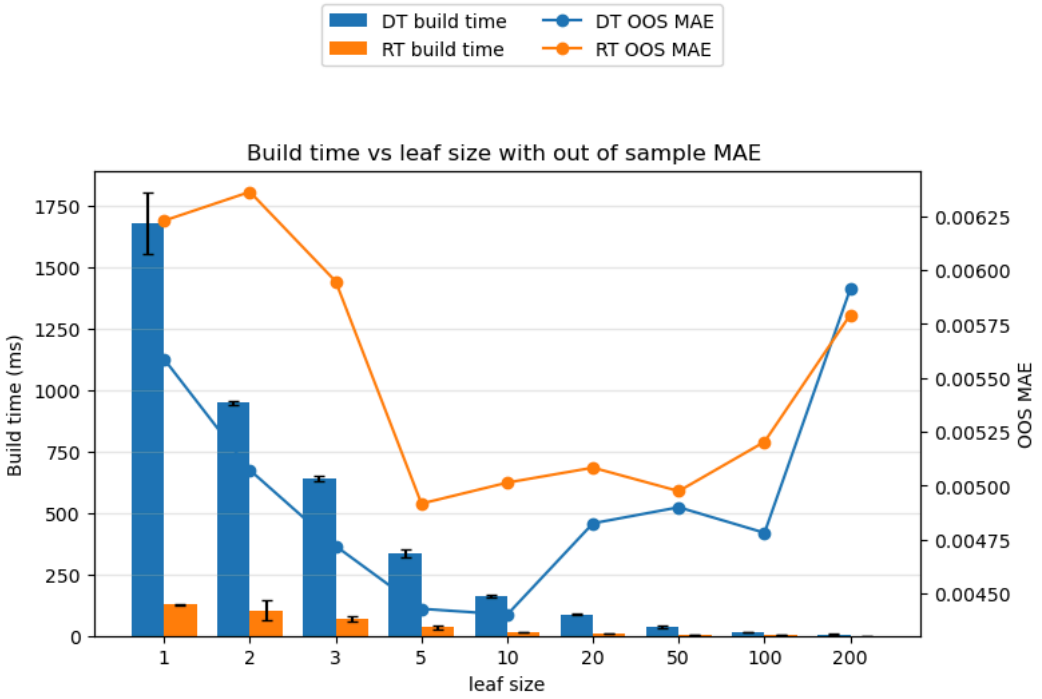


Yes, overfitting does occur with respect to leaf size. As we can see in the chart above the in sample rmse continues to trend upwards as leaf size increases. While on the other hand the out of sample rmse relatively stays flat an is minimized around a size of 10 leaves and goes. Starting off the model is an overfitting as we can see the RMSE is starting at a low rmse with 1 leaf but as

more leaves are added and passing 10 leafs the model starts to underfit. Overfitting occurs here since with a low leaf size since the model having more splits and less averaging of outputs picks up more noise from the training data. Overfitting because the model picking up too much noise from the training set makes it worse at generalizing new data as we can see with the out of sample rmse above. Allowing for more leafs is one way we can mitigate overfitting as it reduces the noise from the training data.



Bagging can reduce overfitting with respect to leaf size. In the graph above we compare RMSE for In sample and out of sample for a single bagged Decision Tree Learner and a Bagged learner with 20. As we can see above the RMSE of the 20 bagged learner consistently shows lower RMSE compared to the single bag learner. However it does not fully eliminate overfitting as having bags does not fully get rid of the noise from the training dataset but it does reduce the error since it makes the model generalize better by averaging results.



The measures I selected were build time and mean average error. In the chart above we see no clear winner for mae. They both stop overfitting at similar leaf sizes. With the Random Decision tree in this run reaching the lowest mae at 5 and the Decision tree reaching it at 10. Although small it does seem like the DT performed slightly in terms of MAE. As far as build time Rt has a distinct advantage as it does not have to waste time finding the best column to split on and chooses randomly. In terms of buid time RT will always be superior as it is always skipping this step.