Report – CS 7646

Name: Arpit Patel

GT ID: 902891506

Report: CS 7646

Name: Arpit Patel

0.0 Methodology

In this project, we develop a random tree learner based on Cutler method that develop a random decision tree and helps us train a learning model. Throughout this project, we would be using the Istanbul.csv data file and will use that for both training and testing the model. We would be using a 60-40 split and will use correlation co-efficient and RMSE error to validate the learner. Further, we will implement bag learner that implements bagging on an ensemble of learners with an aim to cancel random bias and reduce overfitting.

1 Random Tree Learner

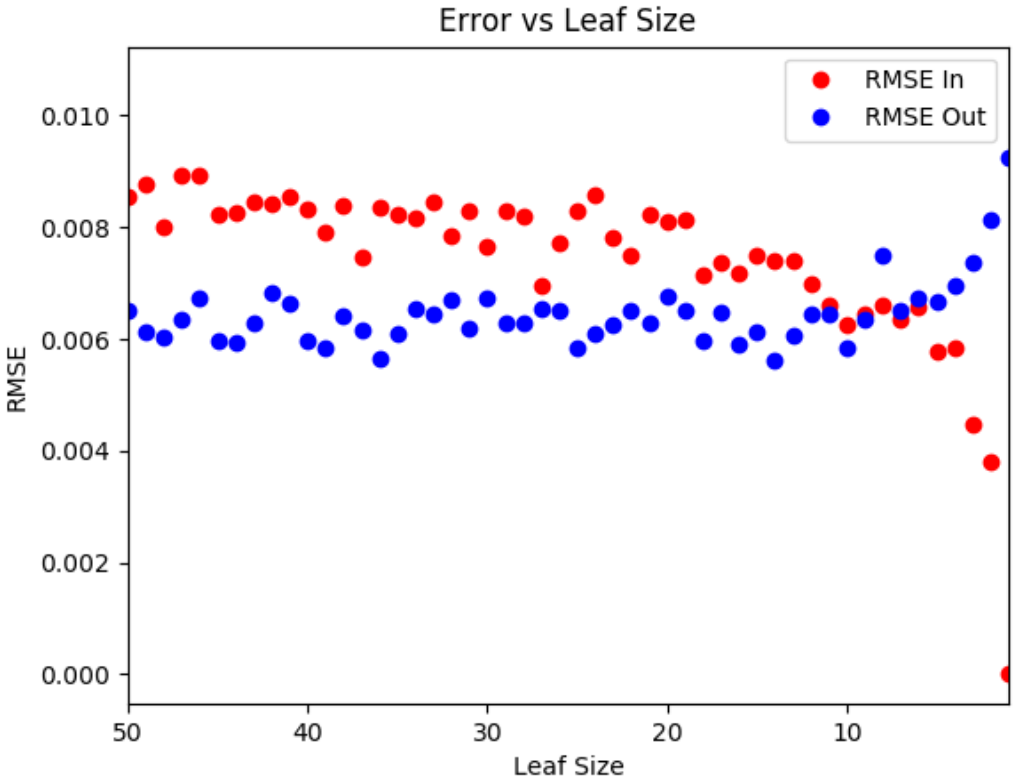
1.1 Overfitting

Leaf Size vs RMSE for Istanbul.csv: As the number of leaves increases, the model no longer fits perfectly to the data. This can be verified by an increase in training error (RMSE) with increasing leaf –size. However, this reduces the amount of over-fitting since the model is more resilient to noise in the training data (verified by decreasing RMSE In). Thus, the RMSE error decreases as the leaf size of the learner is increased. As we lower the leaf size below 10 , the testing error(RMSE Error) increases at the expense of decreasing training error(RMSE in), resulting in overfitting of data.

Report – CS 7646

Name: Arpit Patel

GT ID: 902891506



Report – CS 7646

Name: Arpit Patel

GT ID: 903901504

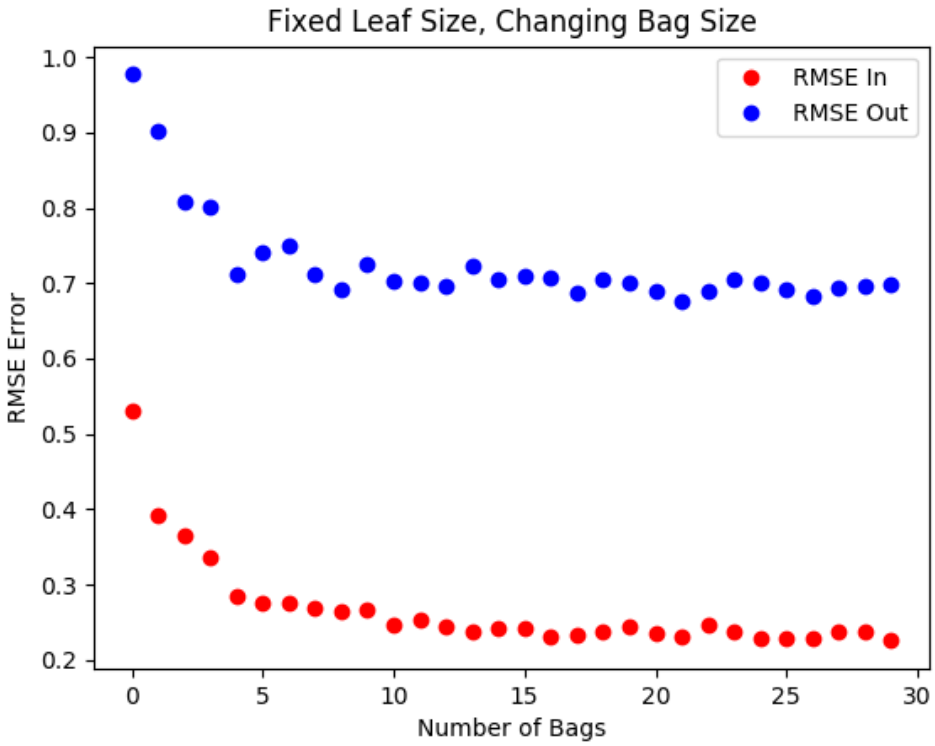
2 Bag Learner

2.1 Overfitting

Bag Size vs RMSE for Istanbul.csv: To determine the relationship between overfitting and the number of bags, RSME was measured as leaf size was kept at a constant value and as the number of bags was varied. Ideally speaking, as the bag size increases, the over-fitting should decrease since by using an ensemble of learners we should be able to cancel the independent biases by different learners. As seen from the figure, the error of training and testing data is very high with low bag sizes (both RMSE-In is high and the RMSE-out is quite high as well). As the bag size increases to beyond 15, both the RMSEs start to converge to a minima and the error decreases low. This is due to increase in model resilience to noise in data, due to mixing from multiple learners and mixing of data within the learners itself. However, despite our prediction, we do not see any significant worsening on Our-Sample(testing) error at the cost of In-Sample(testing) error and we will thus conclude that overfitting does not occur.

Report – CS 7646

Name: Arpit Patel



2.2 Bag Learner with varied Leaf Size

Effect of Leaf Size on Bagging: To determine whether bagging can reduce or eliminate overfitting with respect to leaf size, the number of bags was fixed as leaf size was varied.As the number of leafs increases for a fixed bag sized bag learner, RMSE Error on the training data increases while the error on the testing data remains low. It seems that after a point(leaf size =60), both the error on the training and testing data remains low despite changing the hyper parameter(leaf-size) and the model doesn't result in any significant overfitting

Report – CS 7646

Name: Arpit Patel

GT ID: 902891506