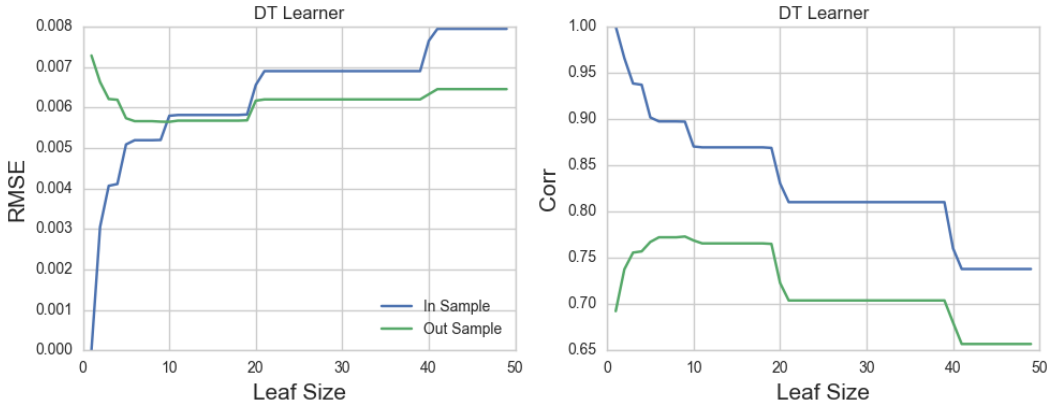
**Assess Learners Project Report**

**CS7646 ML4T Georgia Tech**

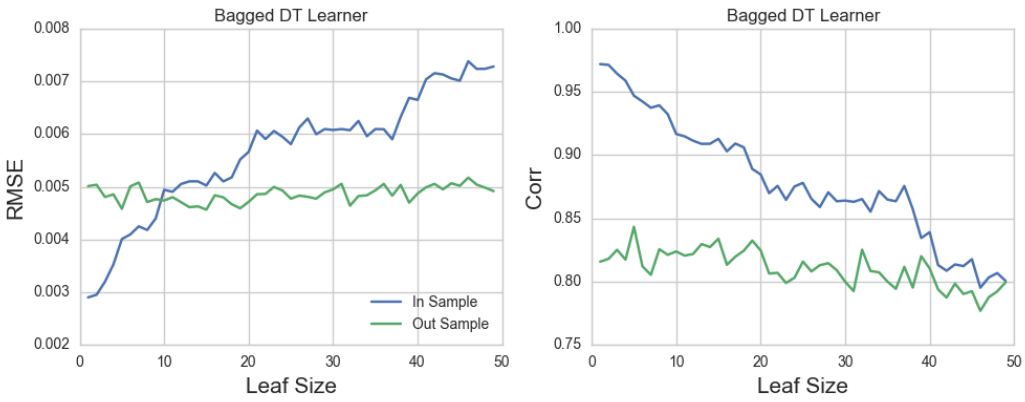
T. Ruzmetov

1. Does overfitting occur with respect to leaf_size? Consider the dataset istanbul.csv with DTLearner. For which values of leaf_size does overfitting occur? Use RMSE as your metric for assessing overfitting.



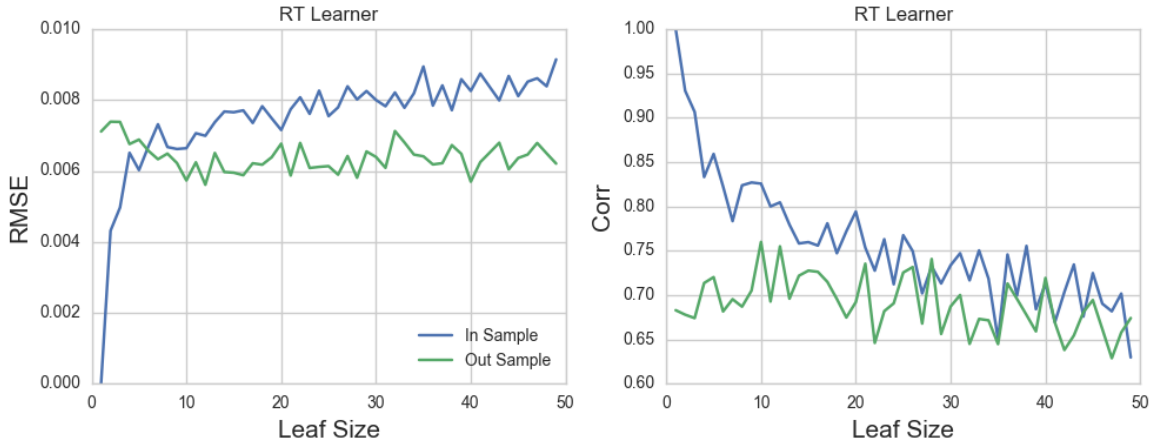

Overfitting does occur for leaf size values less than 10 based on RMSE vs leaf size plot. In sample error increaseas we increase leaf size but out of sample error decrease. In addition, correlations between predicted and actual outcome are very high for training set at small leaf size, then it steadily decrees.

2. Can bagging reduce or eliminate overfitting with respect to leaf_size? Again consider the dataset istanbul.csv with DTLearner. To investigate this choose a fixed number of bags to use and vary leaf_size to evaluate.




Based on RMSE vs leaf size plot, for leaf size values less than 10, difference in in sample and out of sample errors seem to decrease that suggests slight reduction in overfitting. On the other hand, out of sample error is nearly constant with leaf size. So, bagging does help avoid overfittin.

3. Quantitatively compare "classic" decision trees (DTLearner) versus random trees (RTLearner). In which ways is one method better than the other?




Random Tree method results show high fluctuations for both RMSE and correlation values between predicted and actual outcome values. High fluctuations mean high uncertainty, even if error metric on average is same compared to classic DT method. Classic DT method chooses best feature based on highest correlation with the outcome that's why it should perform better. But, it depends on type of problem under consideration. Correlation is not always causation meaning some features might be highly correlated with outcome by chance. If this is the case, then RT method could be more reliable.