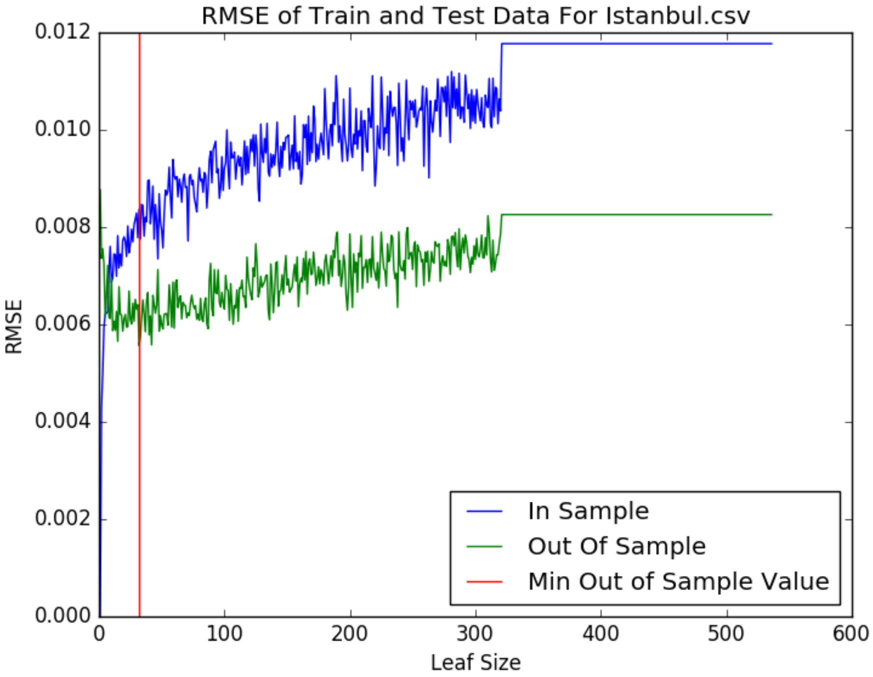
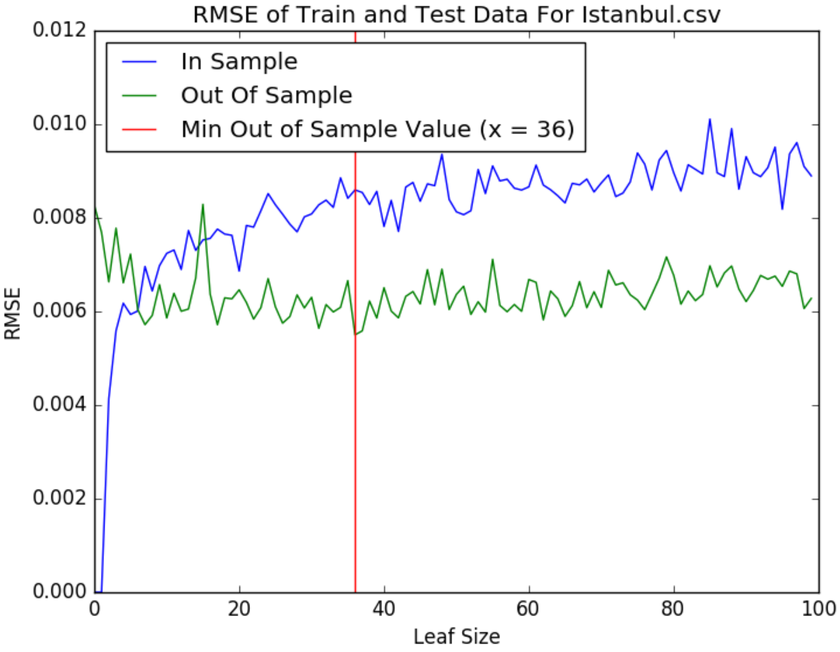
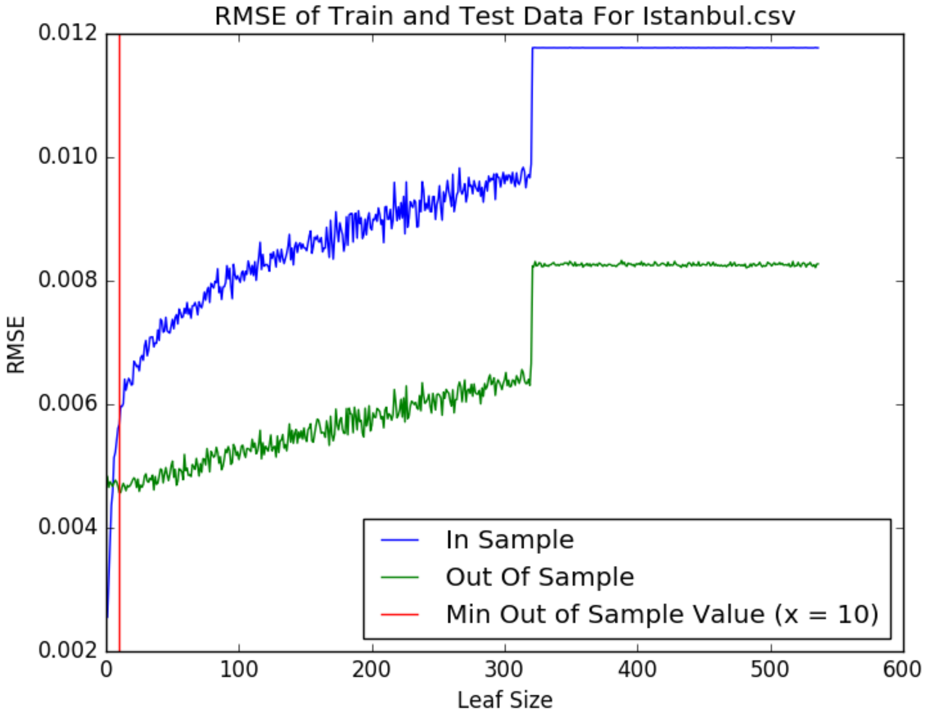
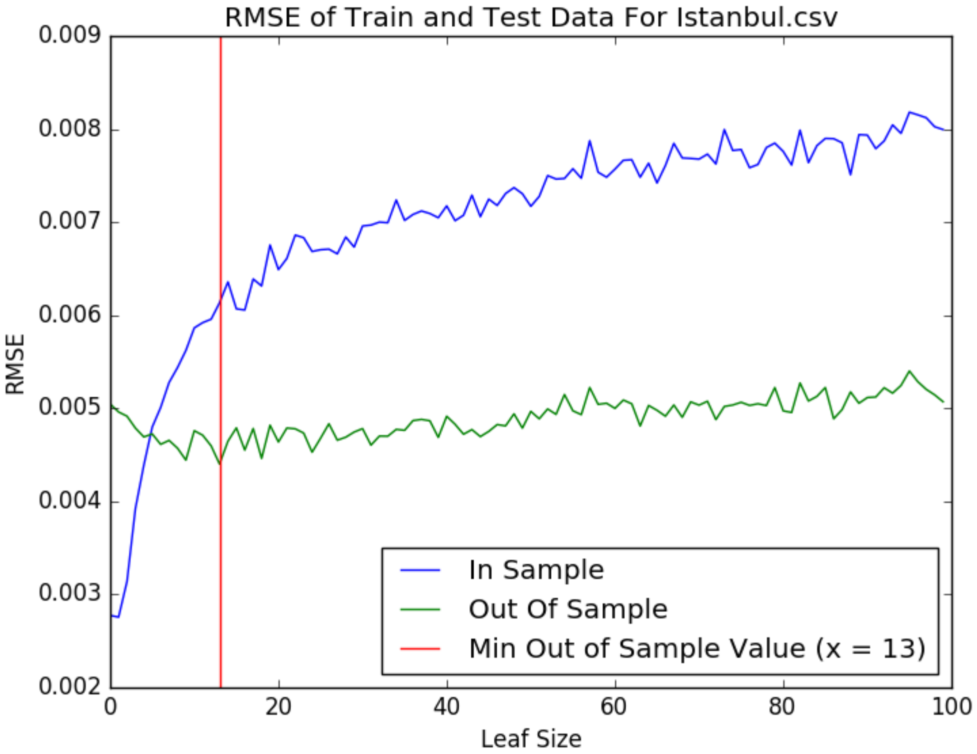




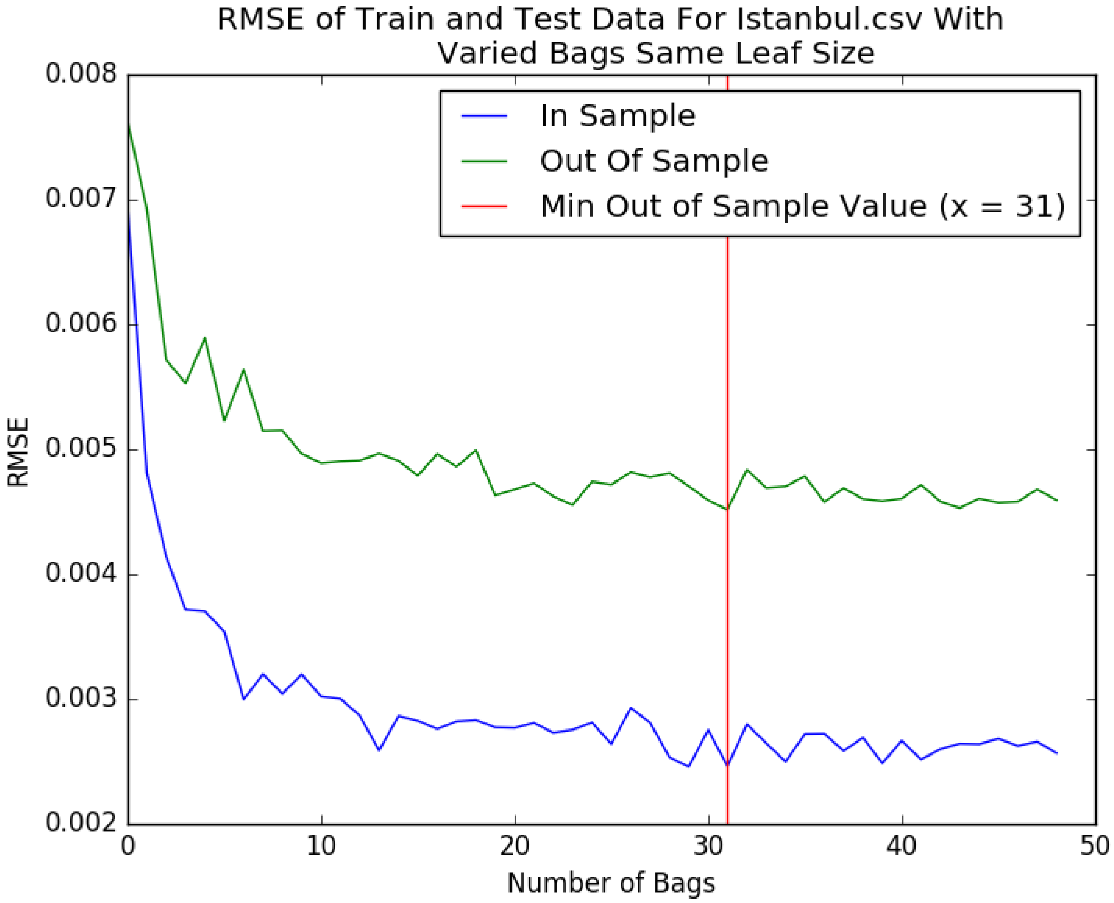
The first graph above shows the RMSE values of in sample as well as out of sample data for Istanbul.csv plotted versus the leaf size of the random tree learner. It can be seen from it, that the minimum RMSE value of the out of sample data occurs between leaf size of 0 and 100. The second plot shows a similar version of the first plot over a shorter range of leaf size.

Using analysis of the above two plots, it can be shown that overfitting occurs with respect to leaf size. For in sample data, as the leaf size increases the predicted y values match less closely to the actual y values. Hence for in sample data, the RMSE increases as the leaf size increases as shown in the plots. The general trend for out of sample data as shown by the plots is that initially as the leaf size increases the RMSE value decreases, but after reaching a leaf size of 36, where the minimum out of sample RMSE value occurs, the RMSE value starts to generally increase until it reaches a constant value of around 0.008. Hence, the leaf size with the minimum RMSE value for out of sample data is the optimal leaf size for predicting y values of the data set using random tree learners. Hence, overfitting occurs for every leaf size less than 36 for the Istanbul.csv dataset, as the lesser the leaf size, the prediction is drawn based on less number of observations. For leaf size greater than 36 overfitting does not occur, as the RMSE for in sample data is high for all those values, indicating that the learner is not a good fit (is under fit) for leaf size greater than 36.





As the above two plots show bagging seems to reduce overfitting. Overfitting occurs for the above two plots for leaf size lower than 13 (or below 10 according to the first plot. The two were plotted during different run times) with bagging. The two plots were plotted at different times to demonstrate that the optimal leaf size is consistently low, indicating that the range of values over which overfitting occurs while using bagging, is reduced. Overfitting does occur, however, its' reduced as the range of leaf size values over which overfitting occurred was 0 to 36 without bagging and with bagging the range was reduced to 0 to 13. This makes sense since bagging takes input from multiple random tree learners and also uses sampling with replacement while building the random tree learners. Here the bag size was 20. Hence, the higher the number of random trees used to build the bag learner, the higher number of observations that the learner can use to extrapolate the data and predict a result. This means that not many observations are needed as part of a leaf size to achieve the optimum leaf size where the RMSE value is the minimum. Optimal number of observations to predict a y-value for a given data point is achieved early as more observations to draw a prediction from means less error.



Overfitting does not occur in the above plot as the when the leaf size is fixed and the number of bags are varied the in sample data initially produces output y-values that are not close to the y-values of the actual data. Thus, the RMSE for in sample data is high initially and the RMSE for out of sample data is also high initially. Both RMSE decrease as the number of bags increases. There is no overfitting for fixed leaf size with varied number of bags.