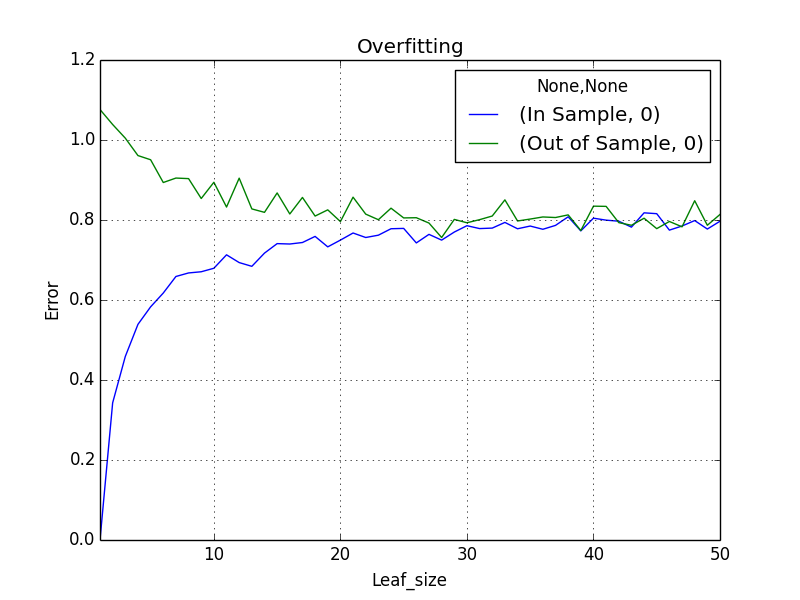
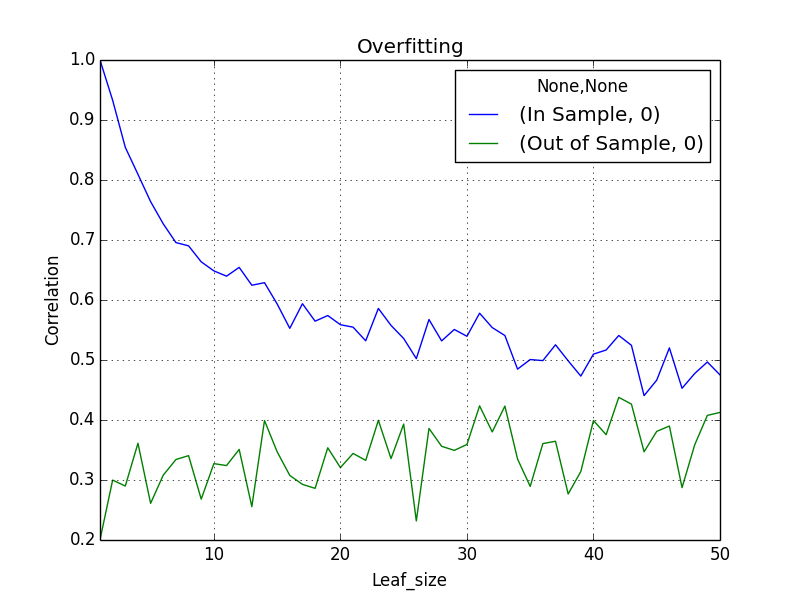
Overfitting occurs when the learner is too specific to the training dataset. Overfitting is defined as the area where the in sample error is decreasing and the out of sample error is increasing. I have decided to use root mean squared error and correlation to help me determine when over fitting is occurring. To show these graphs more python files were generated to determine graph and average data points. In all of the files each case is ran 15 times and the values are averaged to get an accurate value that is more consistent with the learner. This was done because random numbers are being used in the random tree learner to determine features and split values.

The first test case uses the winequality-white.csv dataset. The leaf size is being changed to determine when overfitting occurs. The leaf size helps determine how specific of a value is located at the leaf. A greater leaf size will give more of an average value for the training set and a lesser leaf size will give a more specialized value. The leaf size is varied from 1-50. The endpoint 50 was determined after running multiple tests that determined the data begins to run to an asymptote. **Error! Reference source not found.** below shows the graph of the Root Mean Square Error (RMSE) as the leaf size changes. Using the definition stated above for overfitting the RMSE cannot determine where overfitting occurs. After this chart was found to be inconclusive another metric, Correlation was used. The overfitting chart for correlation is show below in **Error! Reference source not found.**. This chart shows overfitting to occur when In sample correlation is high and out of sample is low. This area can be found between 0 and 20 leaves and is shown in the charts below. This is also the region where the slope tends to decrease at a slower rate.



Overfitting

**Figure 1.** RMSE for the random tree learner as leaf size is changed for the winequality-white.csv dataset.

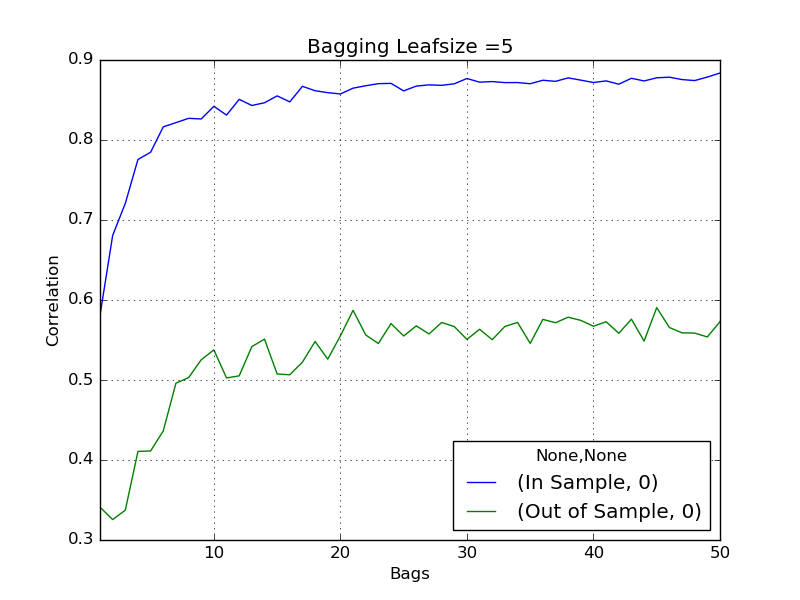
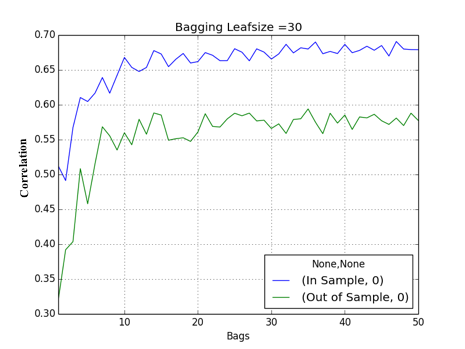
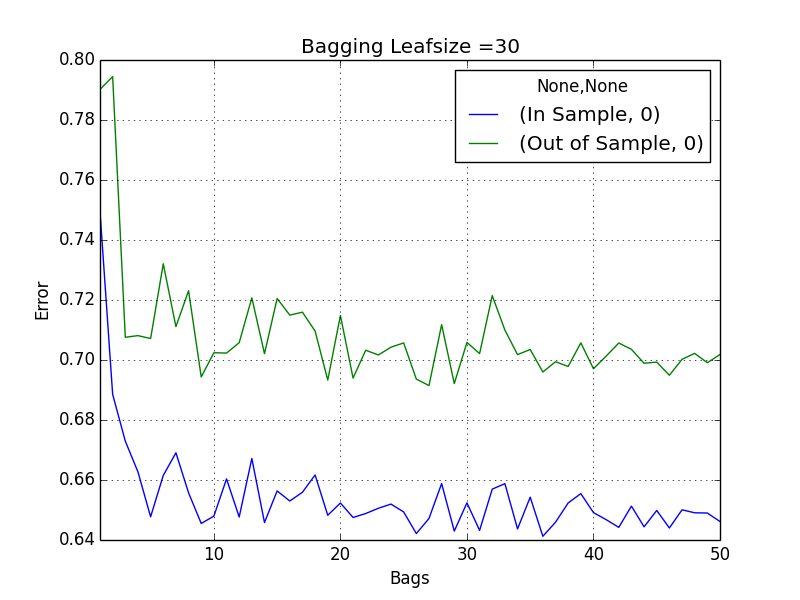
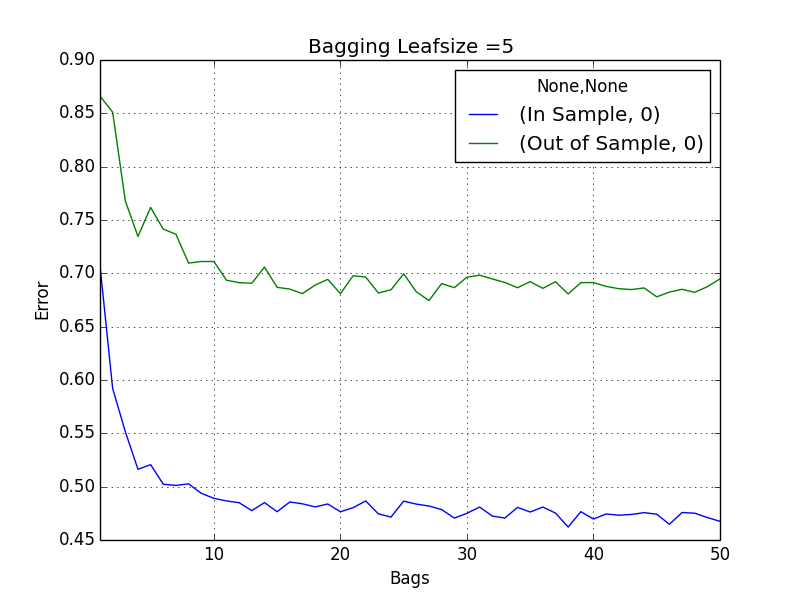


Overfitting

**Figure 2.** Correlation for the random tree learner as leaf size is changed for the winequality-white.csv dataset.

The second test case uses the winequality-red.csv dataset. The number of bags is being changed to determine when overfitting occurs. The leaf size is kept at a constant 5. This number was chosen arbitrarily. A greater bag size will give more of an average value for the learner for the training set. The bag size is varied from 1-50. The endpoint 50 was determined after running multiple tests that determined the data begins to run to an asymptote and this amount of data correctly shows this occurring. Performance quality is increased as the bag number is increased which is shown in the charts below. Overfitting does not occur as the number of bags increase. This is shown in plots below.

**Figure 3.** Charts showing the bags increasing with a steady leafsize.



After analyzing the charts below I have determined that bagging can reduce the overfitting in the dataset. This is shown below as the correlation does not look similar to the charts above.