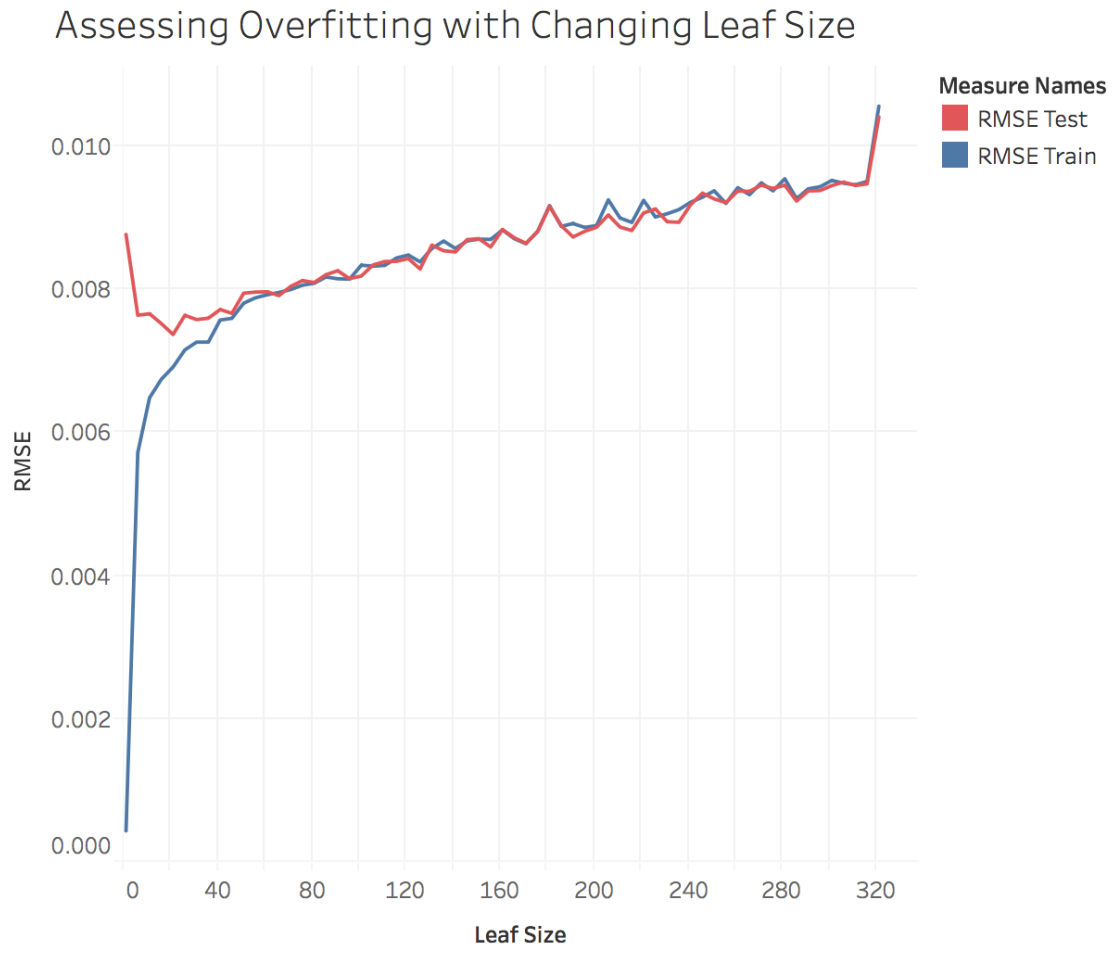
CS7646 MC3-Project-1

ywang3234

Overfitting with Respect to Leaf Size

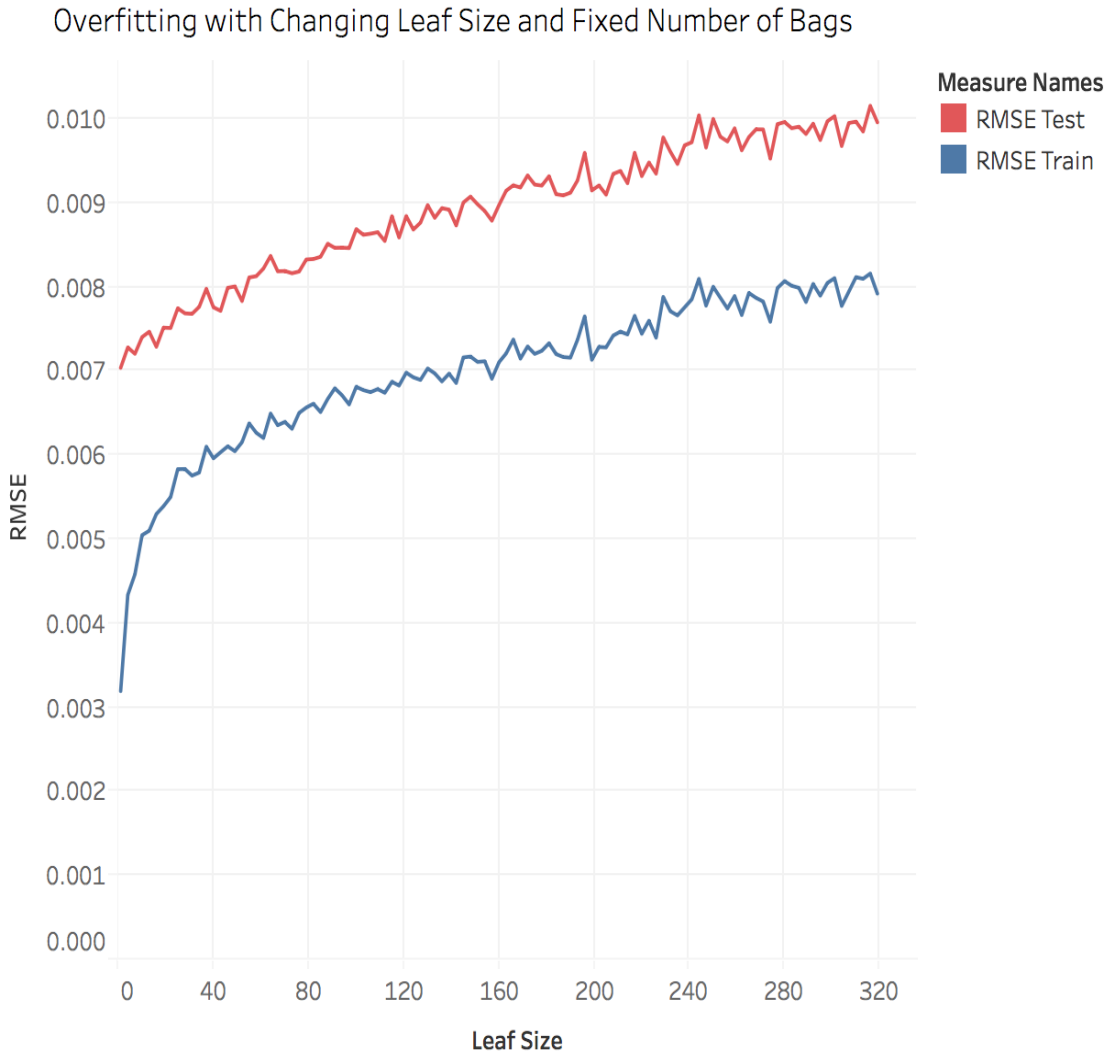
Dataset: istanbul.csv
Training Set: 60 % of the data
Testing Set: 40 % of the data
Model: RTLearner
Experimental Methodology:
The leaf size of RTLearner Model changes from 0 to 320. RMSE of training set and testing set are calculated to assess the overfitting.



Conclusion:
The graph demonstrates that there exits overfitting in this situation. The red line represents the RMSE of testing set and the blue line describes training set.
The overfitting occurs when the leaf size is between 1 and 20 and the distance between two lines is the result of overfitting.
When the leaf size is 1, the RMSE of training set is 0. However, the RMSE of testing set is very high, which is near 0.009. When the leaf size increases, the RMSE of testing set tends to decrease and the RMSE of training set begins to grow. When the leaf size grows to 20, the RMSE of testing set reaches its lowest value. Then, the RMSE of testing set also starts to increase. When the leaf size is above 60, the RMSE of both sets are almost the same and continues to grow.
The lowest of RMSE of testing set is about 0.0075.

Overfitting with Changing Leaf Size and Fixed Number of Bags

Dataset: istanbul.csv
Training Set: 60 % of the data
Testing Set: 40 % of the data
Model: RTLearner, BagLearner
Experimental Methodology:
The leaf size of RTLearner Model changes from 0 to 320. The number of bags of BagLearner Model is fixed at 20. RMSE of training set and testing set are calculated to assess the overfitting.



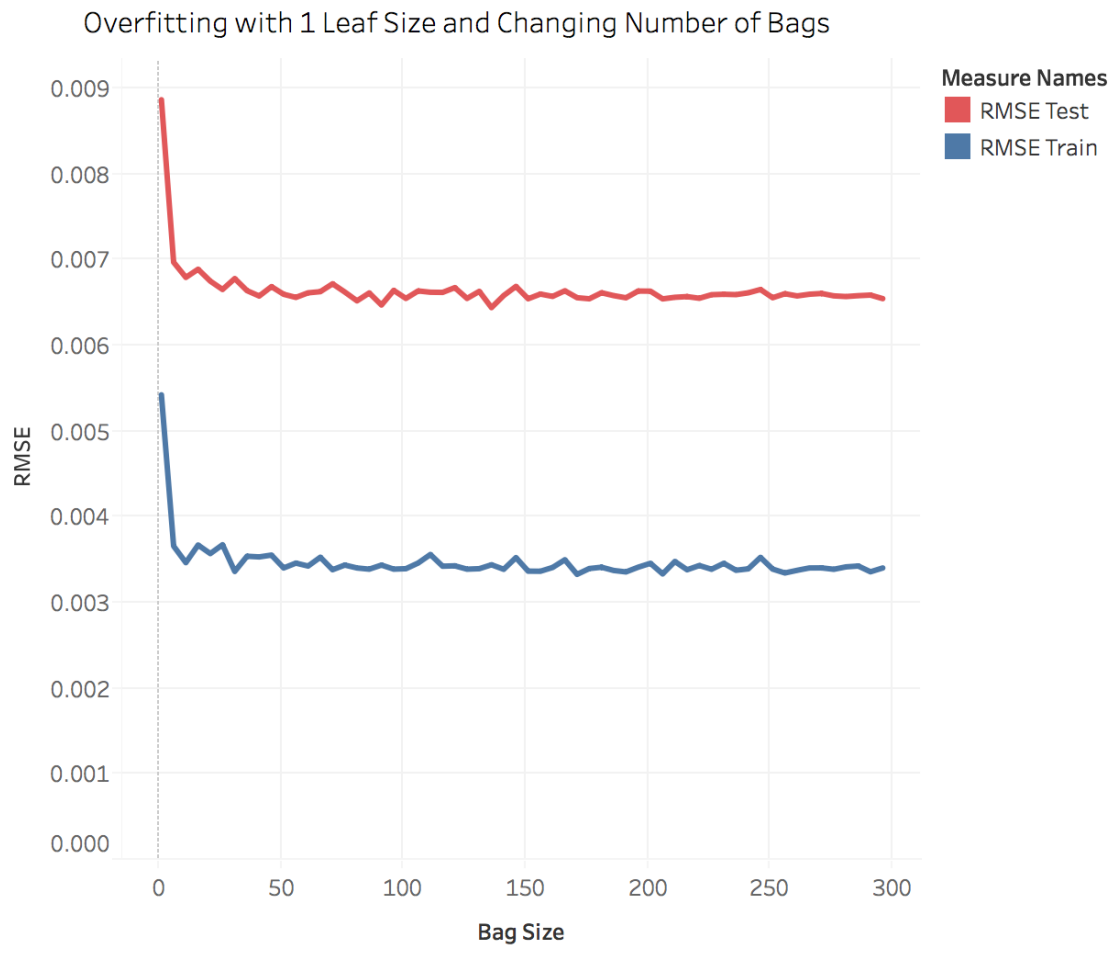
Conclusion:
This graph shows that the overfitting is eliminated when the number of bags is fixed. The red line represents the RMSE of testing set and the blue line describes training set.
When the leaf size starts to increase, the RMSE of both sets tend of grow. Thus, it is impossible that the overfitting occurs. Compared with no bagging, the two lines are somewhat parallel, so the distance between the two lines are almost the same and there is no interaction between the two lines. The RMSE of testing set is always higher than training set's. However, under the no bagging situation, the RMSE of both sets are similar when the leaf size is large. The minimum and maximum values also change significantly in this situation. Here are their comparisons.

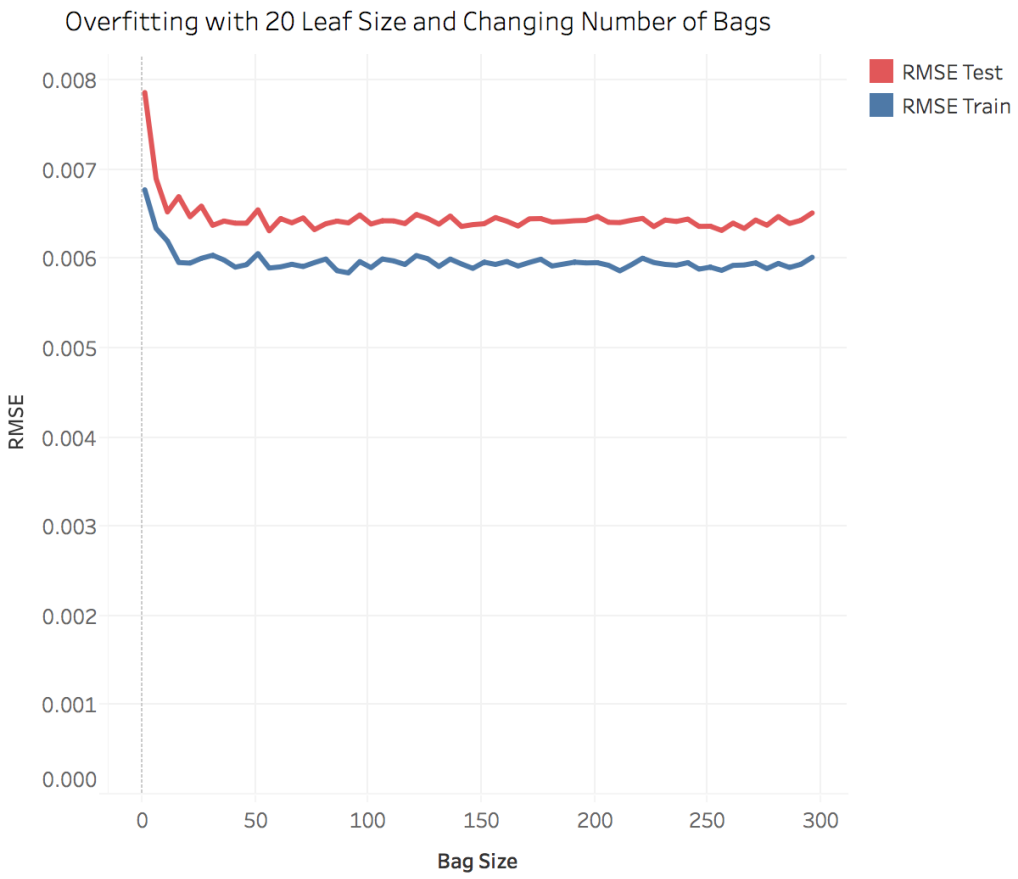
	Changing Leaf Size with no bagging	Changing Leaf Size with Fixed Bags
RMSE Train Minimum	0	0.003566061
RMSE Train Maximum	0.009363872	0.008557899
RMSE Test Minimum	0.006390422	0.006232732
RMSE Test Maximum	0.009445242	0.008720577

The RMSE of training set can reach 0 when the leaf size is 0 with no bagging. However, the lowest value of training set can only reach about 0.0036 with bagging. No matter if bagging is utilized, the minimum values of testing set's RMSE are similar. The table also shows that the minimum RMSE of both sets are lower with bagging when the leaf size is large.

Overfitting with Fixed Leaf Size and Changing Number of Bags

Dataset: istanbul.csv
Training Set: 60 % of the data
Testing Set: 40 % of the data
Model: RTLearner, BagLearner
Experimental Methodology:
The bag size of BagLearner Model changes from 0 to 300. The leaf size of RTLearner Model is fixed at 1 or 20. RMSE of training set and testing set are calculated to assess the overfitting.





Conclusion:

This two graph shows that the overfitting does not occur when the bag size varies.

In both graphs, the red line represents the RMSE of testing set and the blue line describes training set.

When the leaf size is fixed and the bag size tends to grow, both RMSE will decrease firstly. However, when the bag size becomes larger, both RMSE remain unchanged. Thus, it is impossible that the overfitting occurs. We can conclude that when the bag size is large, other factor should be discovered to decrease the RMSE.

The graph also demonstrates that there always exits gap between training set's RMSE and testing set's RMSE. When the leaf size is 1, the gap is biggest. The gap will shrink as the leaf size becomes larger. When the leaf size is 1 and 20, the RMSE of testing set are almost the same. However, the RMSE of training set is affected significantly.