Report of Assignment:2

A) observation for task 1: -

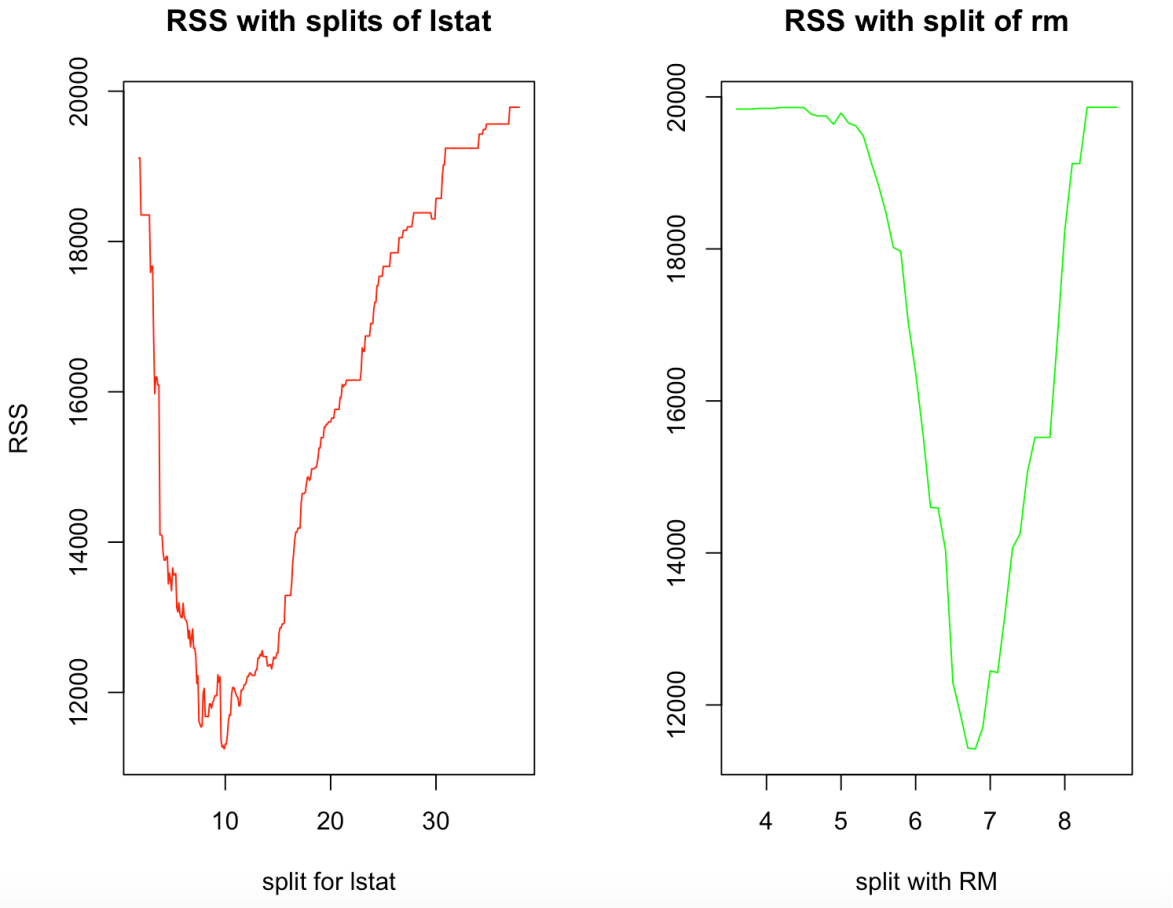
when random state=1811, for diff random state values may differ

1) found that best split was when rm is having 10 value

2) training MSE=45.10

2) testing MSE=49.67

3) graph below represent RSS for different LSTAT vs Rss different split of RM



observation for task 2: -

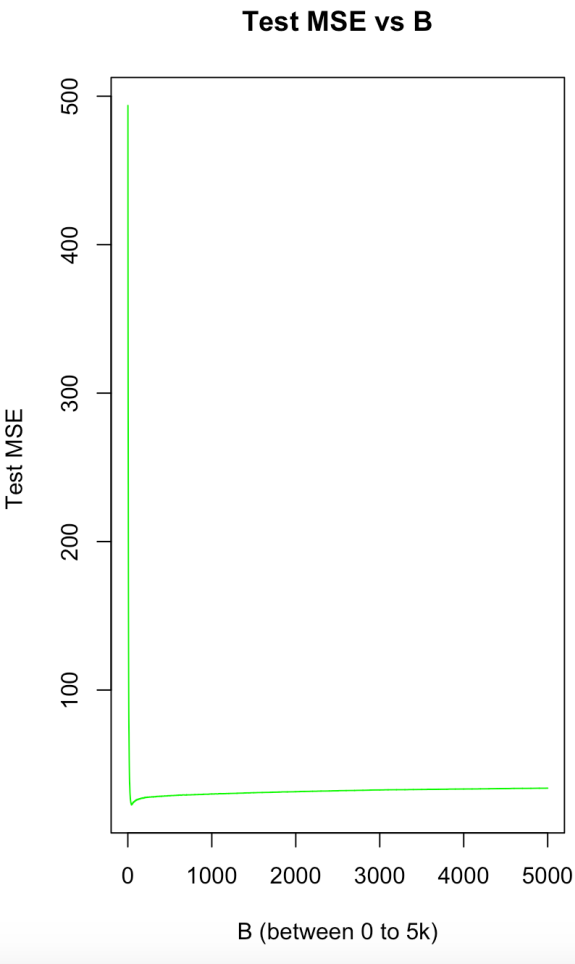
1) MSE of RM is lesser than MSE of LSTAT

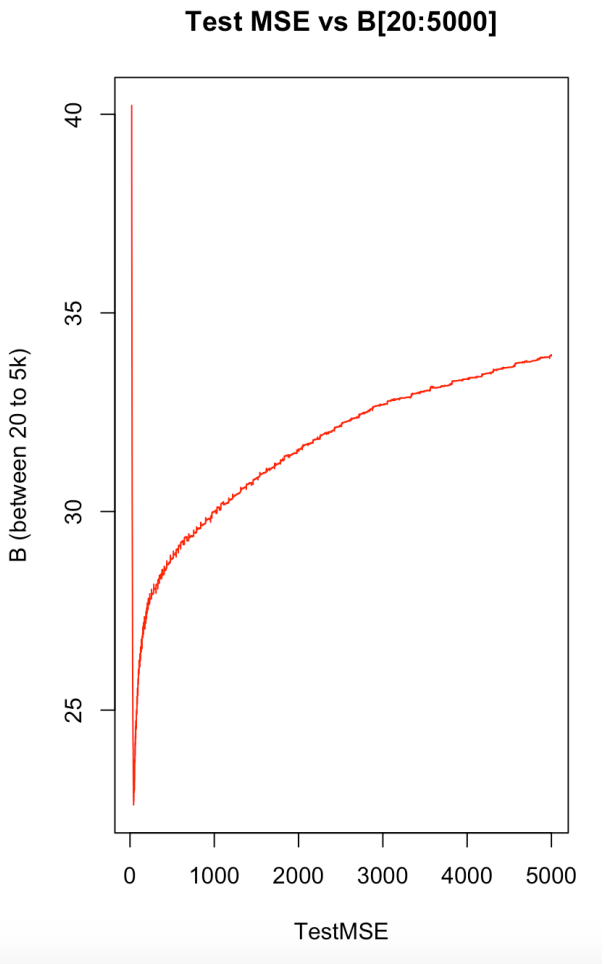
2)When training rate=0.01 given and B=1000 test MSE=25.81 (when random state= 1811)

observation for task 3: -

1) when training rate =0.1 and B=5000 to check either its having overfitting.

2) below graph represent number of trees used to calculate the test mse vs test mse against b [20:5000]





3) Both graphs are similar only 2nd graph is not plotting values till 19 test MSE's

4) This is because test MSE started at high number, so we missed out initial values.

3) second graph is having overfitting issue

observation as per the Parameters we passed: -

1. The number of trees B. Unlike bagging and random forests, boosting can overfit if B is too large (although overfitting tends to occur slowly if at all). We can use a validation set or cross-validation to select B

2. The learning rate η, a small positive number. Typical values are 0.01 or 0.001, and the right choice can depend on the problem. Very small η can require a very large B.

3. The number d of splits in each tree, which controls the complexity of the boosted ensemble. Often d = 1 works well, in which case such a tree is called a (decision) stump.