

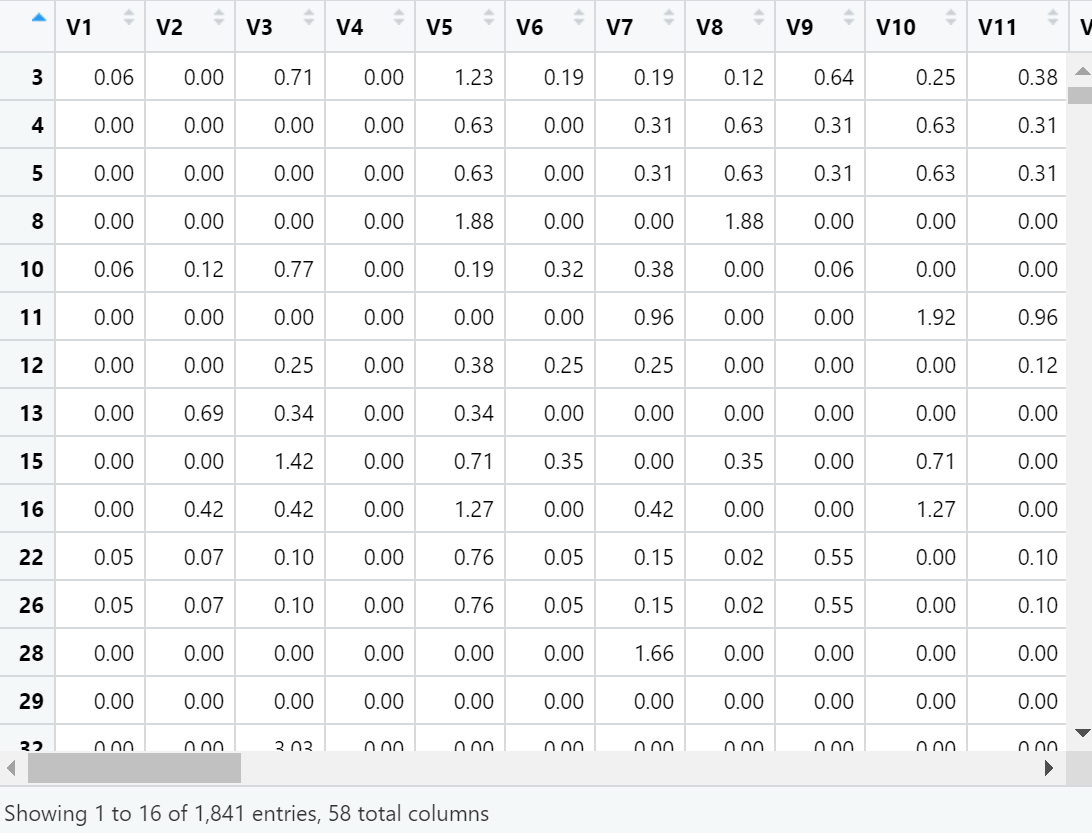
a)

Process:

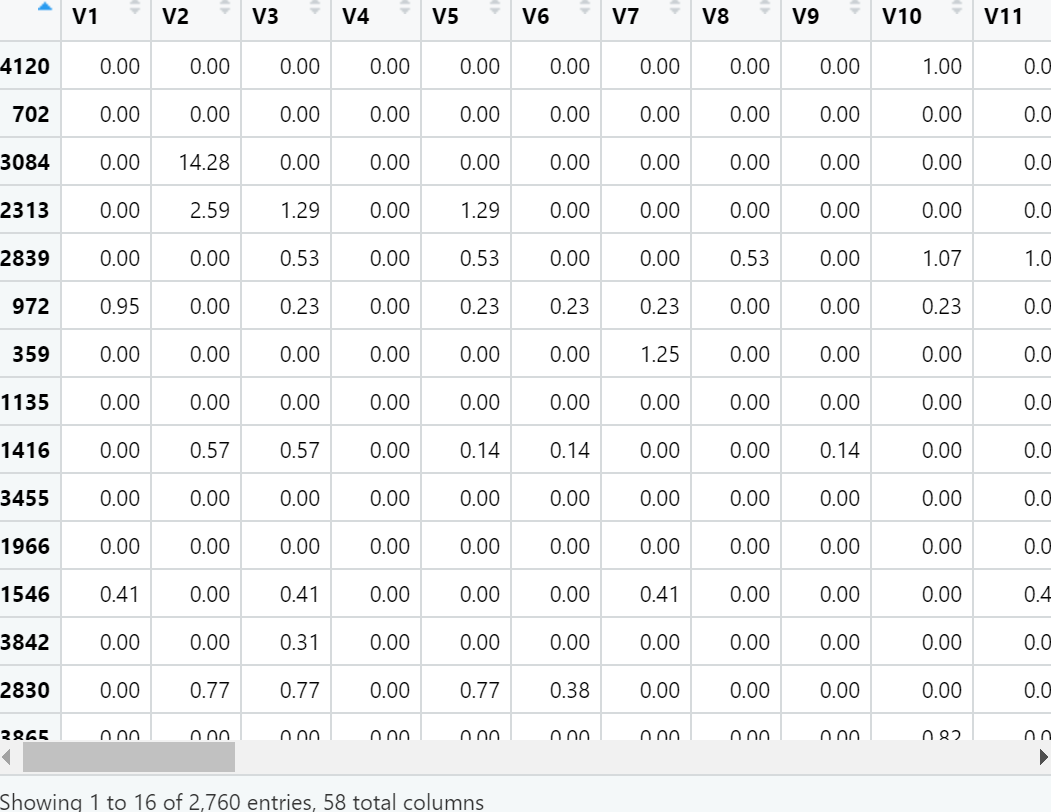
1. **Load data:** I read ‘spam’ dataset
2. **Split the Train and Test data**: I divided 70% of data into train and 30 % of data into the test.
3. **OOB error by the mtry values:** By using tuneRF function I OOB error by the number of mtry values. And Findout the range of the mtry that indicates minimum OOB error.
4. **Test Accuracy:** By the mtry range that I got from step3, I fit the Random forest models and calculated accuracy to the test data. Also, I compared test accuracy for the mode by the number of trees.

Outputs:

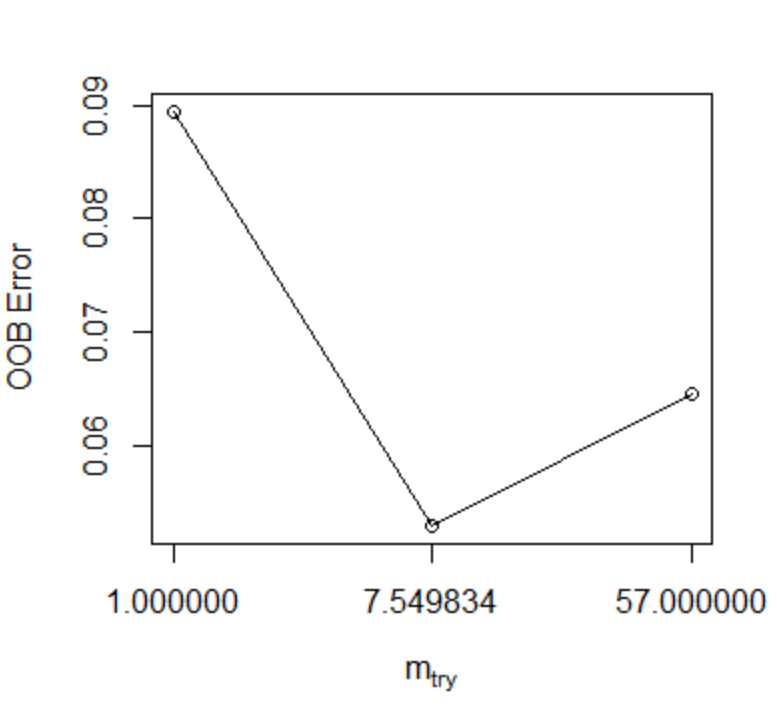
**Test data**



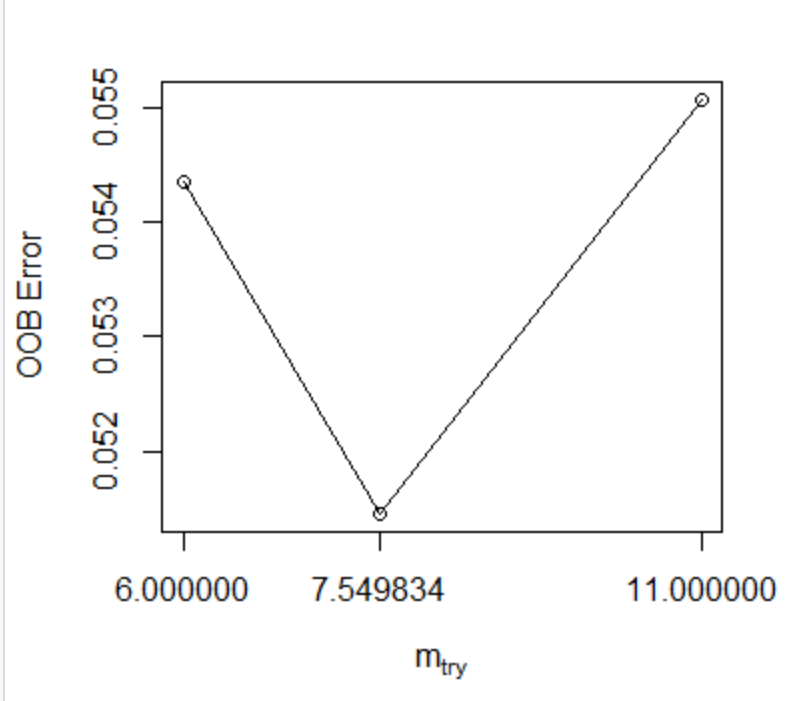
**Train data**



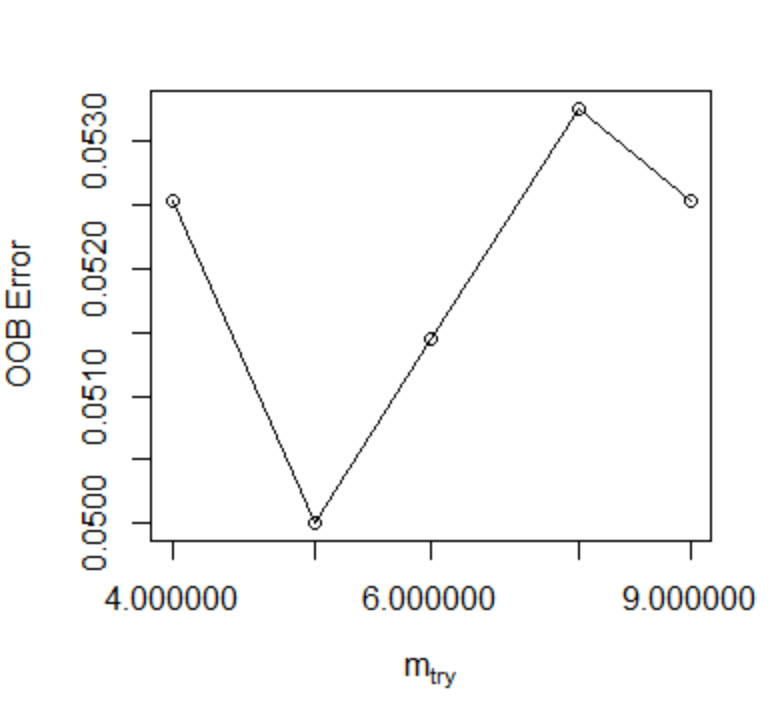
**OBB error (range 1:57)**



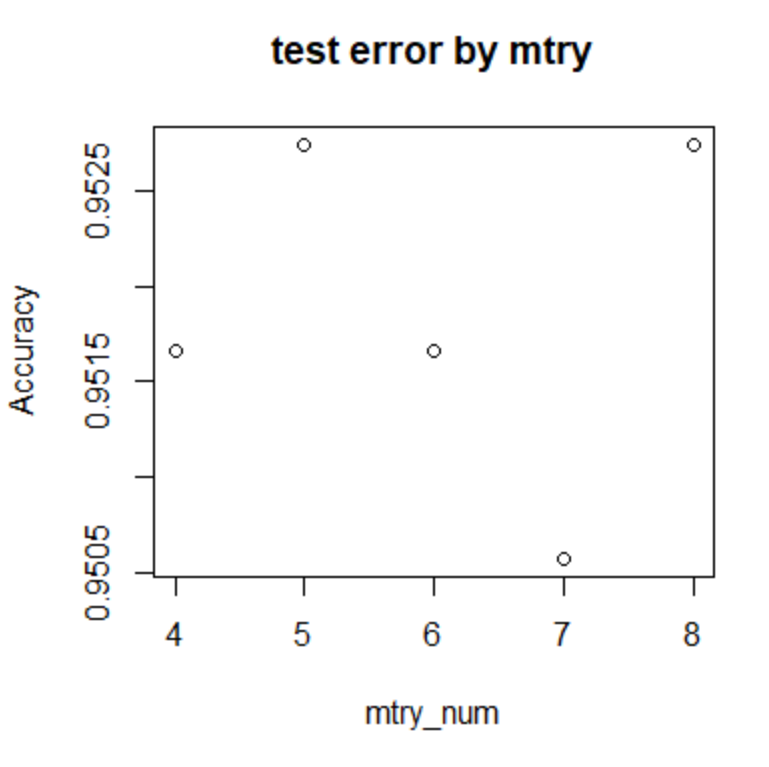
**OBB error (range 6:11)**



**OBB error (range 4:9)**



**The Accuracy by the Random Forest models(mtry\_num 4:8)**



the Random Forest models’ test accuracy and error for the by the number of ntree (mtry = 5 fixed) (ntree=100,200,300,400,500)





Discussion:

***Mtry***

Mtry is the number of randomly selected input for each tree. Impiridicaly, there is an approximate value for the Mtry that minimize the OOB errors

Classification: (total number of variables)^0.5

Regression: Total number of variables/3

In our data set, there are 57 variables. 57^0.5 is about 7.549(classification). According to the result, mtry num of 5 indicated the minimum OOB error, and mtry num of 5, 8 show the highest classification accuracy for the test data.

This indicated that the theoretically optimize ‘mtry’ value for the model is similar to the ‘mtry’ value that gives minimum result for test error and OOB error.