# P4-SpamData

*Jeetendra Gan*

*11/30/2019*

The data has slightly imbalanced classes with around 39% marked as spam.

```
## [1] "The percentage of spam rows in train data is: 39.391304"
```

```
## [1] "The percentage of spam rows in test data is: 39.443962"
```

So, our test data and train are only slightly imbalanced.

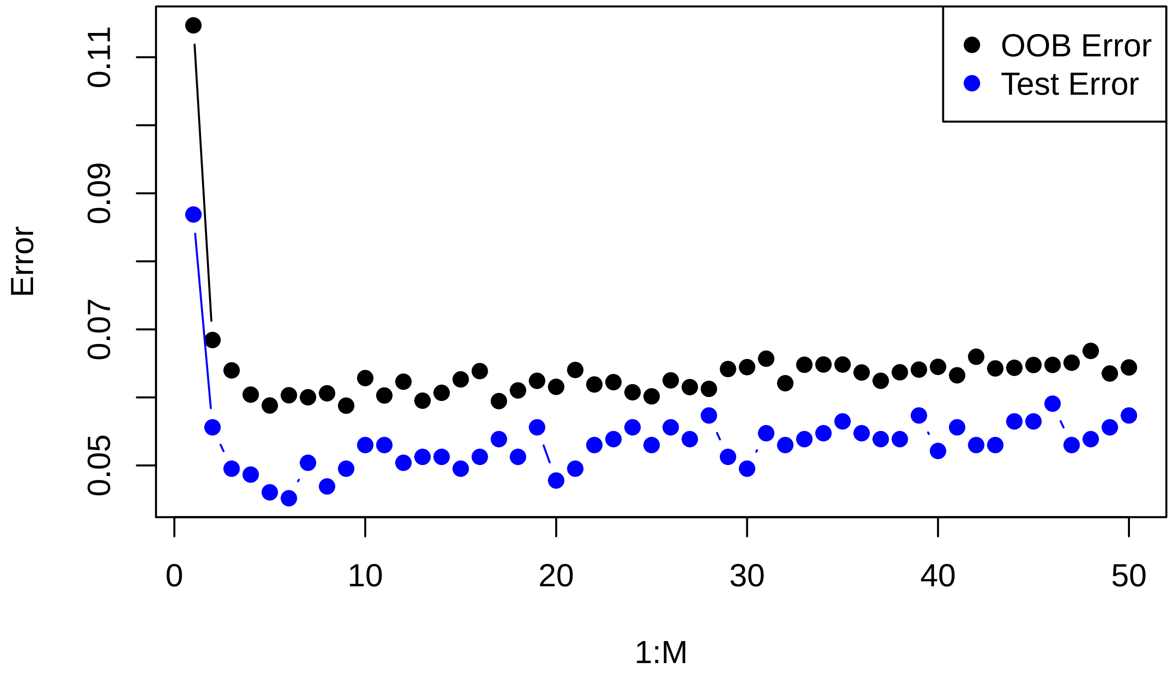
### Applying random forest

### OOB Vs Test Error

I have used the following algorithm for computing the OOB estimate for each value of m

- for every value of m
  - get the mean of errors for all trees in a forest produced by mtry = m

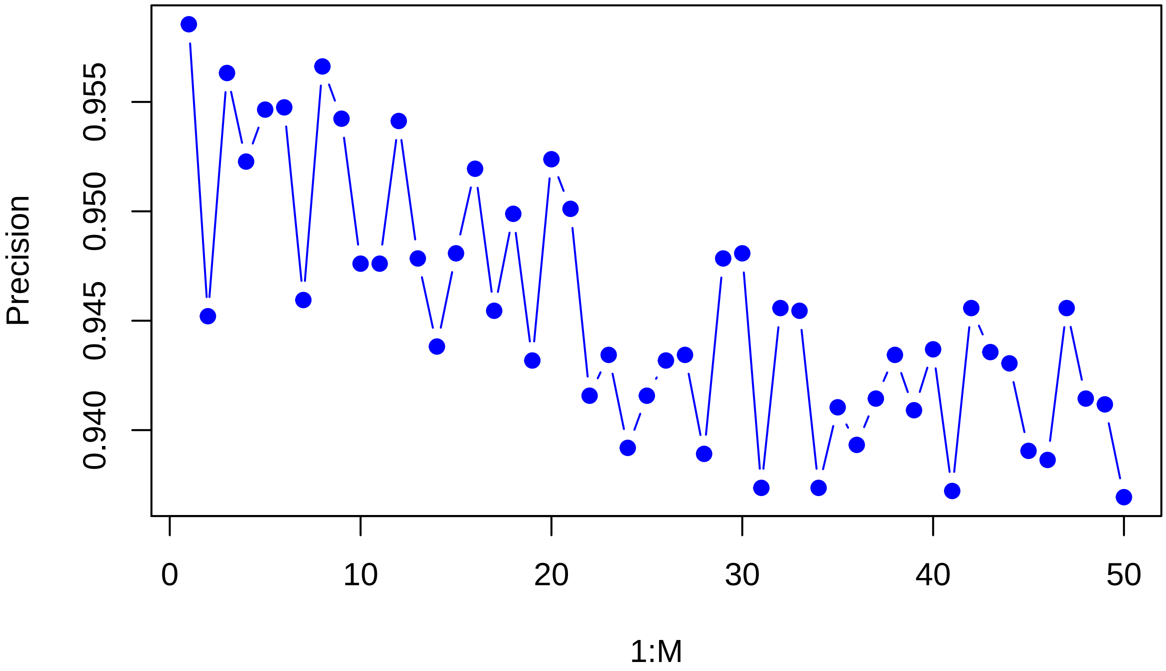
plot the collected mean for each forest against the interaction depth values



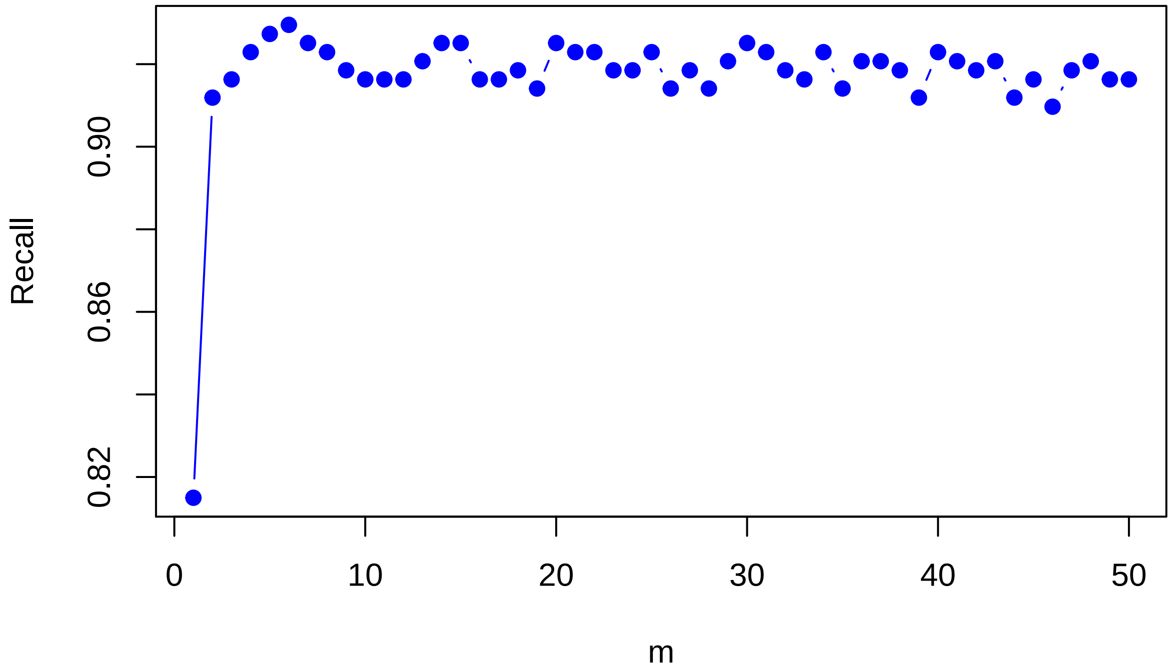
It can be seen that the OOB error and test error is very close. Test error is slightly lesser than OOB error. OOB error is somewhat equivalent to the cross validation error estimate. Hence generalizes better than the test error.

### Model sensitivity to m

The precision on the test data follows the trend as shown below.



The recall on the test data is shown below.



As the value of m increases, the error decreases. An increase in m helps us capture greater variance in the data. But after a point, that increase is of no use as the information has already been captured by other smaller models. Random forest does not overfit, hence the error does not flucturte too much after some point. Looking at the graph of m vs Error, it can be seen that a value of 5 is good.

Sensitivity to precision and recall

Precision is decaying very very slowly. Even though fluctuations can be seen in the graph, they are happening over a very small range, so the difference is negligible.

A clear pattern can be seen for recall, It increases significantly till a value of 5 and then remains constant more or less.