**Classification Metrics**

For the binary classification problem at hand, with goal of predicting if Email is SPAM or NOT, we have following confusion matrix possibility:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | IsSPAM = TRUE | IsSPAM = FALSE |
| Predicted  (Classifier Outcome) | IsSPAM = TRUE | TP | FP |
| IsSPAM = FALSE | FN | TN |
|  |  |  |  |

Considerations on different classification outcomes

Accuracy (True Positive and True Negative) metric determines ability to classify email a spam or not and needs to be as high as possible.

False Positive (FP) rate is the number of times Email is classified as SPAM, while actually it was not. This needs to be minimized, as otherwise the SPAM filter could end of deleting lot of important emails that are not SPAM. This implies that specificity given by TN / (TN+FP) = ~1, or as high as possible.

False Negative (FN) rate is the number of times Email is classified as Ham, while actually it was SPAM. This need not be penalized as much as FP rate, since it implies we can occasionally allow SPAM messages to reach end user, as long as we get better model with very low FP rate. It would be expected that the end person can exercise judgement and categorize such emails as SPAM, and take corrective actions as per company’s IT policy.

The best model is one which provides best accuracy and highest specificity

**Basic Tree Model & Parameters**

Parameters tuned for our basic tree are mentioned below along with description:

minsplit - the minimum number of observations that must exist in a node in order for a split to be attempted

minbucket - the minimum number of observations in any terminal <leaf> node

cp - complexity parameter. Any split that does not decrease the overall lack of ﬁt by a factor of cp is not attempted. A high cp value implies a split is attempted only if information gain is larger by amount cp. Choosing a higher value of cp reduces number of tree splits, as fewer splits will meet the high bar on information gain as set by high cp parameter.

In our analysis above parameters are varied.