**Modeling Preparations**

*Approaches:* Two approaches were under consideration for email SPAM classification problem, (1) Naïve Bayes, and (2) decision tree (DT) based approach. These are discussed below and a case is made for use of decision tree for modeling problem. Naïve Bayes (NB) requires or assumes an underlying probabilistic model. This is not a requirement for decision tree. In contrast to NB approach, decision tree algorithms are relatively easier to understand in terms of understanding classification algorithm, and also provides easy identification of important features, and handling of missing values. In DT based approach it is easier to accommodate additional derived features from an email (e.g. number of forwards, number of capital letters in subject line, etc.), and not just the word content. To explore and capture above, DT based approach is pursued in this work.

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

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

While accuracy is a default metric for better classification, for current problem we need to discuss importance of FP (False Positive) and FN (False Negative) scenarios to identify metrics:

* False Positive (FP) rate needs to be minimized.
  + It is the number of times Email is classified as SPAM, while actually it was not.
  + Goal is that the SPAM filter should not automatically block emails by classifying them as SPAM, when they actually are not. Otherwise, users could end up loosing a lot of important emails.
  + This implies that specificity given by TN / (TN+FP) = ~1, or as high as possible.
* False Negative (FN) rate should be minimized, but is less critical than FP rate.
  + This is the number of times Email is classified as Ham, while actually it was SPAM.
  + FN rate is less critical than FP rate, as we can occasionally allow SPAM messages to reach end user, as long as we get better model, i.e. a very low FP rate.
  + It is assumed that that the end person can exercise judgement and categorize such emails as SPAM, and take corrective actions as per company’s IT policy.

Above arguments lead us to select following metrics to be considered while choosing the best model:

* Specificity, or True Negative Rate (TNR)
  + Given by .
  + Minimizing FP, implies we need to select model with highest TNR values
* Precision, or Positive Predictive Value (PPV)
  + Given by .
  + Minimizing FP, implies we need to select model with highest PPV values

There is additional flexibility in choosing model with a slightly poorer recall, or sensitivity (= , as cost of FP is larger than FN, and focus is minimizing FP over FN.

Lastly, we have an imbalanced classification problem at hand. Refer Figure <XYZ> above, that shows that ~75% outcomes are HAM, while ~25% outcomes are SPAM in modeling dataset. Due to above, and to provide a better balance between precision and recall, F1 metric is additionally chosen as a metric for choosing the best model. F1 is harmonic mean of ‘Precision’ and ‘Recall’ and given by

F1 score penalizes if any one of precision or recall falls low, which may also happen due to imbalanced classification problem at hand when attempting k-fold cross validation.

**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.