Business Understanding:

Email is a service that has all but replaced traditional mail in the 21st century. Unfortunately, junk mail, called spam mail in the context of email, has filled inboxes as email becomes more ubiquitous. Spam mail is significantly cheaper to send than junk mail, so there is very little stopping advertisers from sending spam emails to as many people as possible. Spam mail can range from mostly harmless advertising to phishing or malware links. The range of outcomes from clicking a spam link requires a system to properly identify spam. A human educated about spam mail can often spot spam emails at a glance, but what about a less tech savvy operator? An automated system to detect spam mail can be used to unclutter inboxes and help protect those who have trouble identifying spam mail.

In this analysis, we will create an automated system to identify spam emails using machine learning. More specifically, we will use information contained in the email such as the number of characters in the body, whether the email is a reply email, if the email body has images, and other characteristics. Using the extracted information, we will create a tree based model in order to predict if an email is spam or not. Since we were given the data and did not perform the data extraction ourselves, we are assuming that the data extraction was done in a robust manner. The following analysis is only as robust as the data that we were provided.

Modeling Preparations:

As stated in the Business Understanding, we will be using a tree based algorithm in order to determine whether an email is spam or not. Decision trees are fairly useful for binary classification problems. The direction an observation flows through the tree structure is determined by decision nodes. If the decision node is based on a binary variable, the observation flows right or left based on the value of the variable for the observation. If the decision node is based on a continuous variable, the observation flows right or left based on if the observation’s value is above or below a certain threshold. The tree model has a binary structure and is a non-parametric model, so the variables used in the model don’t need to follow any distribution. This makes a tree model extremely useful for binary classification problems such as this one.

In order to find the “best” decision tree, you need to test every possible tree for the model. In order to avoid testing every possible tree, we decided to leverage the random forest model. The random forest takes a number of decision trees, which each make a decision if the email is “spam” or “ham”. If enough of trees classify the email as “spam”, then the email is classified as spam. If the number of trees classifying the email as spam is below a certain threshold, then we classify the email as ham. Utilizing random forest allows us to test fewer trees in the model creation step and allows us to have a more robust model.

In order to build our tree, we have decided to leverage CART and XGBoost. CART is an algorithm that allows for building of classification and regression trees. CART trees are created by splitting nodes based on what creates the biggest improvement to the classification or regression. The algorithm determines when to stop making splits either while building the tree by determining if the split gains enough information to be useful, or after building the tree by pruning. XGBoost is a gradient boosting algorithm that simultaneously speeds up the model building process and improves model performance. We believe that comparing these two methods will allow us to build the most accurate model to predict spam emails.

To determine the best model, we are using a few evaluation metrics. Accuracy, which is the number of correct predictions divided by the number of total observations, is important, but it doesn’t necessarily tell the entire story. It is important to minimize the number of false negatives, but a large number of false positives is more problematic. It is a bigger problem to classify a real email as spam, rather than classifying a spam email as not spam. If a spam email gets through our filter, then a human will waste a couple seconds deleting the email from their inbox. If an email gets incorrectly classified as spam, a potentially important email may not be seen. Therefore, we are also interested in looking at sensitivity, the true positive rate, and specificity, the true negative rate. More weight will be given to specificity, but sensitivity will also be considered.

We will consider ROC-Curves and AUC to determine our best models as well. The ROC Curve plots false positive rate on the x-axis and true positive rate on the y-axis. AUC is the “area under the curve” of this ROC plot. An AUC close to 1 indicates a model with a high true positive rate with a low false positive rate, which is a desired model. F1 score will be considered as well, which is the reciprocal of the mean of precision and sensitivity, where precision is number of true positives divided by the number of true positives plus the number of false positives. It is important to note that we consider “spam” to be the positive class.

If you are interested in reading more about the algorithms and accuracy measures used in this analysis, here are a few useful links:

XGBoost: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Decision Tree Learning: https://en.wikipedia.org/wiki/Decision\_tree\_learning

Accuracy Measures: <https://en.wikipedia.org/wiki/Sensitivity_and_specificity>

CART: <https://en.wikipedia.org/wiki/Predictive_analytics#Classification_and_regression_trees_.28CART.29>