IS 4300 Final Project

Camille Sloan

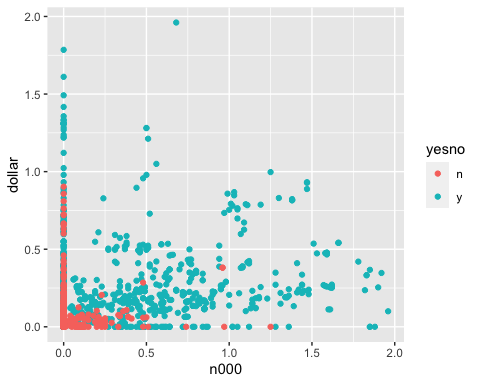
## Dataset Introduction

In this report, we will create a model using the *spam7* dataset to determine when an email is likely to be identified as spam or not. There are 4,601 total records of data in this dataset with 8 variables to choose from for input to the model. The target variable the model will predict is *yesno*, and it specifies the spam status of the record as *n*, not spam, or *y*, spam.

The two variables within the dataset that we will use as input to the model are *n000*, which holds the number of times the content of the email had a string that included ‘000’, and *dollar*, which holds the number of ‘$’ symbols were in the content of the email.

## Visualizations

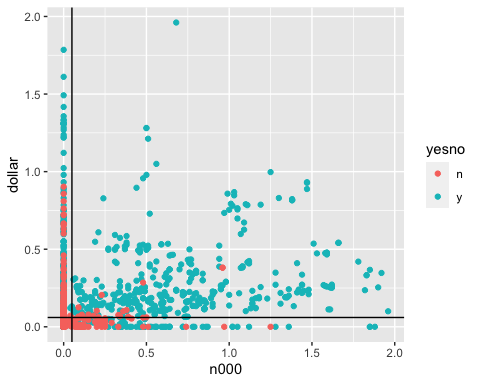
Below is a plot that will help to give a better idea of how these variables appear on a graph. The values are colored red for *n*, or **not** spam, and blue for *y*, or **yes** spam. Outliers have been filtered out to better show the spread of the values.



## Creating the Model

To create a decision model for this dataset, lines will be placed on both the horizontal and vertical axes at the best position possible to indicate when a value would likely be determined *n* or *y*.

After some time exploring the options, I concluded that the lines should be placed at 0.05 for the x axis and 0.06 for the y axis. Here is what that looks like:



These lines of determination mean that the model will predict any value of *n000* less than 0.05 to be considered *n*, not spam, and anything above it *y*, spam, and any value of *dollar* less than 0.06 to be considered *n*, not spam, and anything above it *y*, spam.

## Model Analysis

Below is a table that displays the predicted values compared to the actual:

## Actual  
## Predicted n y  
## n 2609 734  
## y 179 1079

This shows that the model predicted *n* correctly 2,609 times out of the total 3,343 occurrences of *n*, and it predicted *y* correctly 1,079 times out of the total 1,258 occurrences of *y*.

To further analyze the results, let’s look at the model’s accuracy, sensitivity, and specificity.

## Metric Value  
## 1 Accuracy 0.802  
## 2 Sensitivity 0.595  
## 3 Specificity 0.936

## Conclusion

To create this model I explored numerous combinations of the 7 potential input variables. I tested various lines of determination across the possible variable combinations, and, while many came close, found that these two variables- *n000* and *dollar*- with these lines of determination- x axis at 0.06 and y at 0.05- to be the best decision model possible. The accuracy is 80.2%, which indicates a strong model. The specificity is an impressive 93.6%. The sensitivity is a bit low, at 59.5%, but, based on my analysis, this decision model is the strongest possible for predicting when an email is determined as spam or not spam in this dataset.