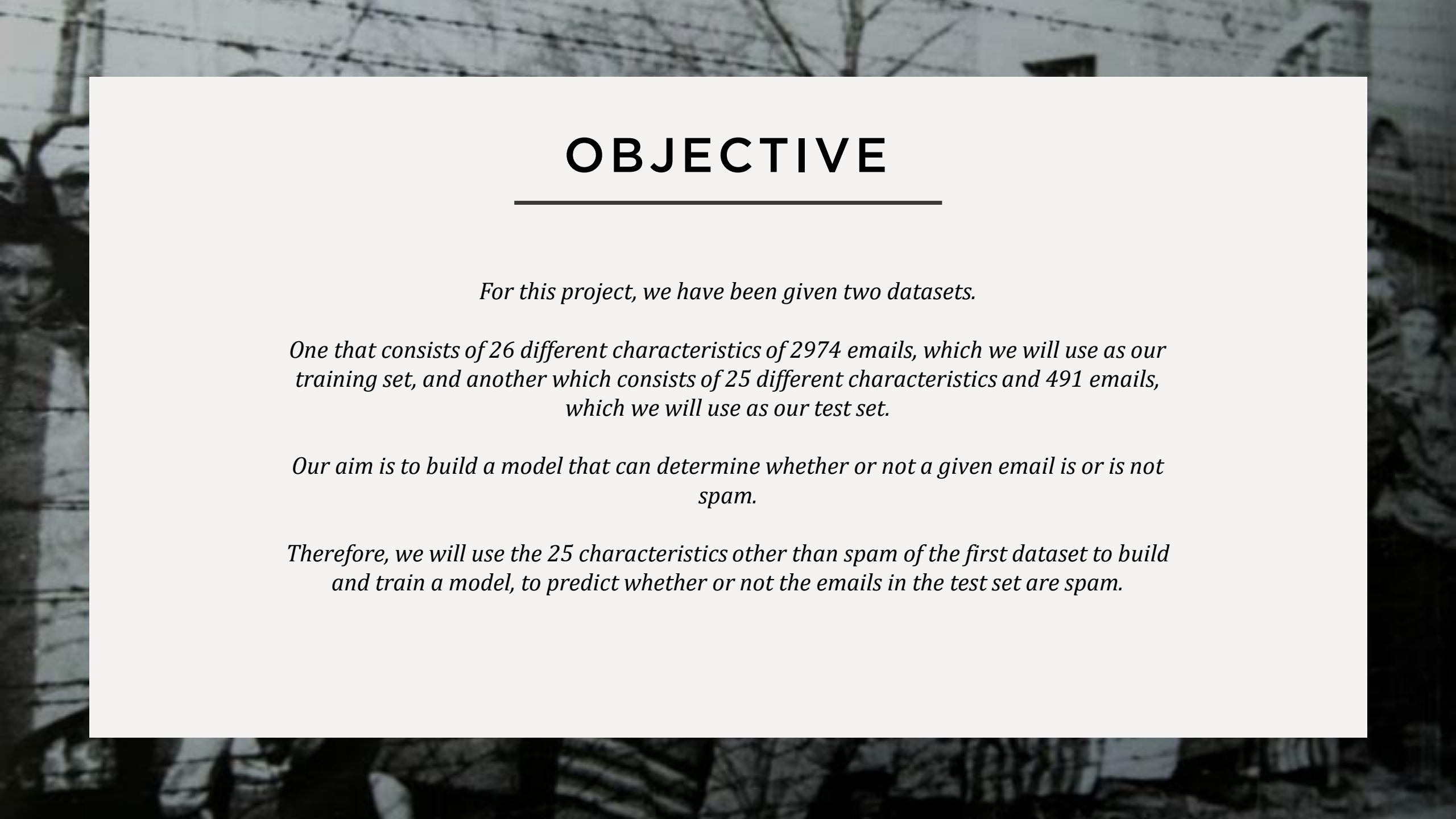
# SPAM

## OR NOT SPAM?

ETC3250 PRESENTATION

ERNEST JEREMY KUSAL MAX



# OBJECTIVE

*For this project, we have been given two datasets.*

*One that consists of 26 different characteristics of 2974 emails, which we will use as our training set, and another which consists of 25 different characteristics and 491 emails, which we will use as our test set.*

*Our aim is to build a model that can determine whether or not a given email is or is not spam.*

*Therefore, we will use the 25 characteristics other than spam of the first dataset to build and train a model, to predict whether or not the emails in the test set are spam.*

# DATA EXPLORATION

## "WEEKDAY" VARIABLE

What we see from the "Weekday" variable is that the percentage of spam on weekdays is consistently around the mid-30s.

On the other hand, the percentage of spam on weekends is in the high 50s.

Thus, the percentage of spam on weekends is higher than that of weekdays.

# FIGURE 1.

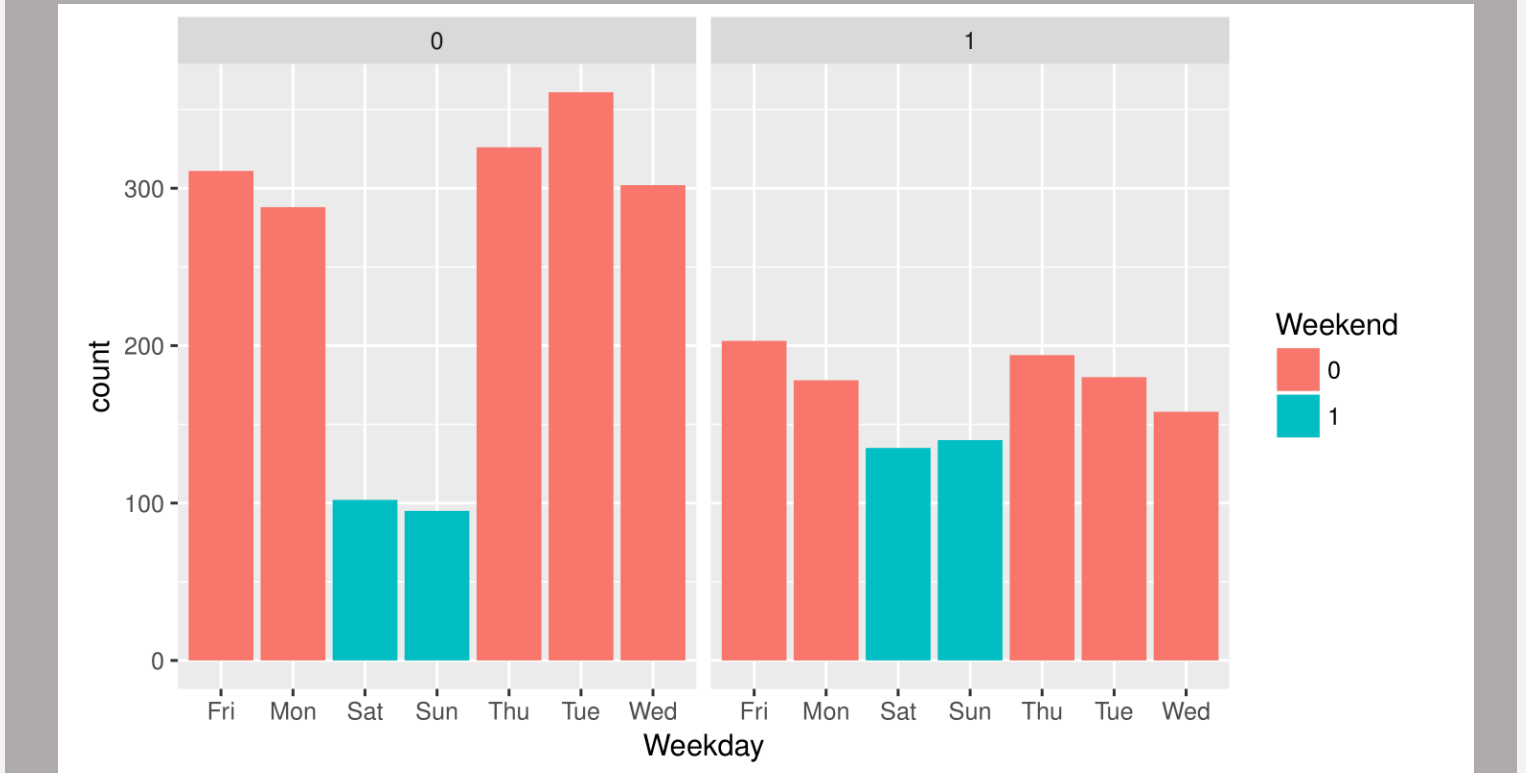



Figure 1. Count of spam emails on each of the days of the week, red representing weekdays and blue representing the weekend.

# DATA EXPLORATION

## "HOUR" VARIABLE

Looking at the different hours during which an email was sent, we can see an interesting pattern arise.

As expected, there is a significant increase in the number of emails sent during office hours (9am -5pm), but this increase consists of, for the most part, non-spam emails.

Furthermore, we can also see that the number of legitimate emails sent between 12am and 6pm is lower, while the number of spam emails remains at the same level.

Essentially, this graph demonstrates that legitimate emails follow a cycle dictated by our days (fewer emails sent during sleeping hours, more during working hours), while the spam emails are sent at an almost uniform cycle.

# FIGURE 2.

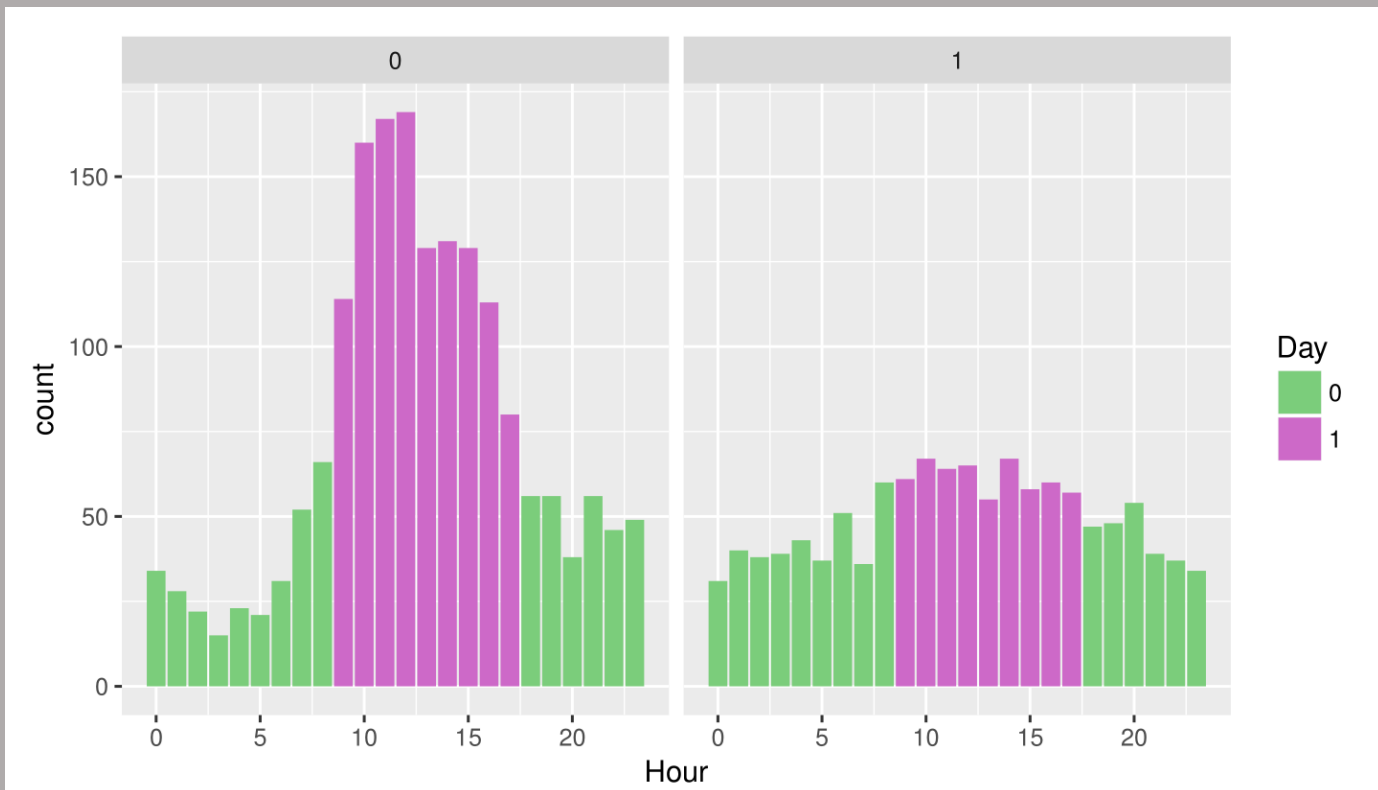



Figure 2. Count of spam emails at each hour of the day, office hours in purple and out-of-office hours in green.

# DATA EXPLORATION

## "DOMAIN" VARIABLE

53 tw 0.9354839 62

There are a variety of different domains from which emails are sent (55 different domains in the training data), but we can see that a large proportion of them (42, in fact) are used 10 times or less in the data set of almost 3000.

Further, it seems that the most common domain .com is largely impartial at 54% spam.

Some of the more useful domains appear to be .edu (typically not spam), and .net and .tw (typically spam).

# TABLE 1.

| | Domain | Percent.Spam | Total.Count |
|---|---|---|---|
| 53 | tw | 0.9354839 | 62 |
| 11 | de | 0.3125000 | 80 |
| 39 | org | 0.4040404 | 99 |
| 34 | net | 0.8956522 | 115 |
| 13 | edu | 0.0822050 | 1034 |
| 9 | com | 0.5413589 | 1354 |

Table 1. The six most common domains and their relevant counts and percentage spam.

# DATA EXPLORATION

## KEYWORDS

A couple of keywords seem to be fairly good indicators of a spam email. In the table below we can see that pharm , discreet and prescription all have above 90% spam while still having reasonably high counts.

Surprisingly, porn is the keyword with the lowest percentage instance of spam.

| | Keyword | Percent.Spam | Total.Count |
|---|---|---|---|
| 13 | asseenon | 1.0000000 | 3 |
| 3 | sucker | 0.2500000 | 4 |
| 2 | porn | 0.2285714 | 35 |
| 6 | drugs | 0.7500000 | 36 |
| 5 | prescription | 0.9770115 | 87 |
| 9 | discreet | 0.9646018 | 113 |
| 14 | discount | 0.7345133 | 113 |
| 4 | pharm | 0.9337748 | 151 |
| 1 | credit | 0.5094340 | 159 |
| 11 | sell | 0.4545455 | 220 |
| 7 | save | 0.4812030 | 266 |
| 12 | sale | 0.5071942 | 278 |
| 8 | sex | 0.4414716 | 299 |
| 10 | free | 0.4619377 | 578 |

TABLE 2. THE SELECTED KEYWORDS AND THEIR RELEVANT COUNTS AND PERCENTAGE SPAM.

# DATA MANIPULATION

## DEALING WITH DOMAIN

Due to the number of levels not working with some models, as well as the presence of certain domain names being present in the training set and not in the test set, we had to find a way to remove certain levels without impacting the data too much.

We can see that the Domain variable in the train set has 55 levels, as opposed to the 25 in the test set.
As a result, we removed the irrelevant levels from the train set and substituted them with com.
As com had such a high count and was largely impartial (roughly 50% spam, 50% not spam), we didn't think this would impact the data too much.

# DATA MANIPULATION

## "WEEKEND" AND "DAY"

We also included a binary variable distinguishing between the weekend and weekdays through the **Weekend** variable, as well as a binary variable distinguishing between office hours and out-of-office hours through the **Day** variable.

# MODEL SELECTION

## LOGISTIC REGRESSION

*Reasonably good predictor of whether or not the emails were spam.*

*Blind use results in about 60 to 70 percent accuracy.*

*With refinement, we obtained a decent model that gave an accuracy of around 0.80.*

## KNN

*When we looked at using KNN, we also saw a mediocre outcome in our test set predictions.*

*While we tried a number of values for K, the optimal value of K = 3 resulted in an accuracy of around 0.70.*

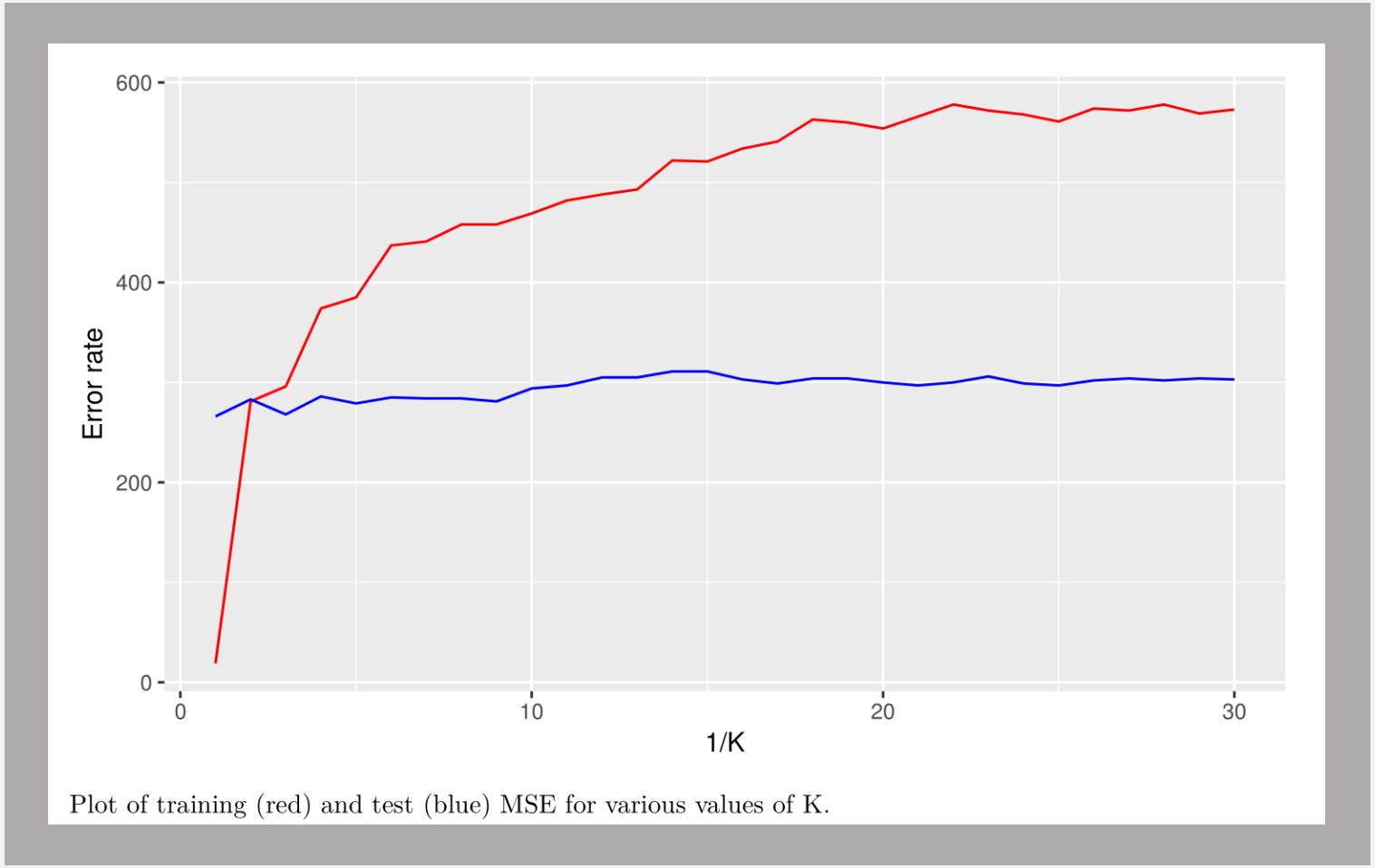
## RANDOM FOREST

*From the outset, random forest returned very good accuracy.*

*By default model instantly returned a high accuracy of 89%.*

*It is for that reason, that we decided to focus on refining and improving the accuracy of our random forest model.*

# FIGURE 3.




Plot of training (red) and test (blue) MSE for various values of K.

# MODEL REFINEMENT

Following the decision to refine the random forest model, we attempted to improve the model, by increasing the number of tree used. However, upon closer inspection, it became clear that at the default value of 500, there would be as good as no additional benefit in adding any additional trees.

Therefore, we went and looked back at the data exploration to see what changes to the data we could make in order to improve accuracy on the test set.

From the findings in our data exploration, we could see that a few variables (asseenon, sucker, drugs and porn) had very low frequency in the data. These variables have the lowest mean decrease in accuracy if they were removed (generally less than 1 or 2 and at times negative).

In order to reduce the number of dimensions and to improve the model, these variables were not selected

```
SPAM.FINAL <- RANDOMFOREST(SPAM ~ . -ASSEENON-SUCKER-DRUGS-PORN, DATA =
TRAINSET, MTRY=5, IMPORTANCE=TRUE)
```

# FIGURE 3.

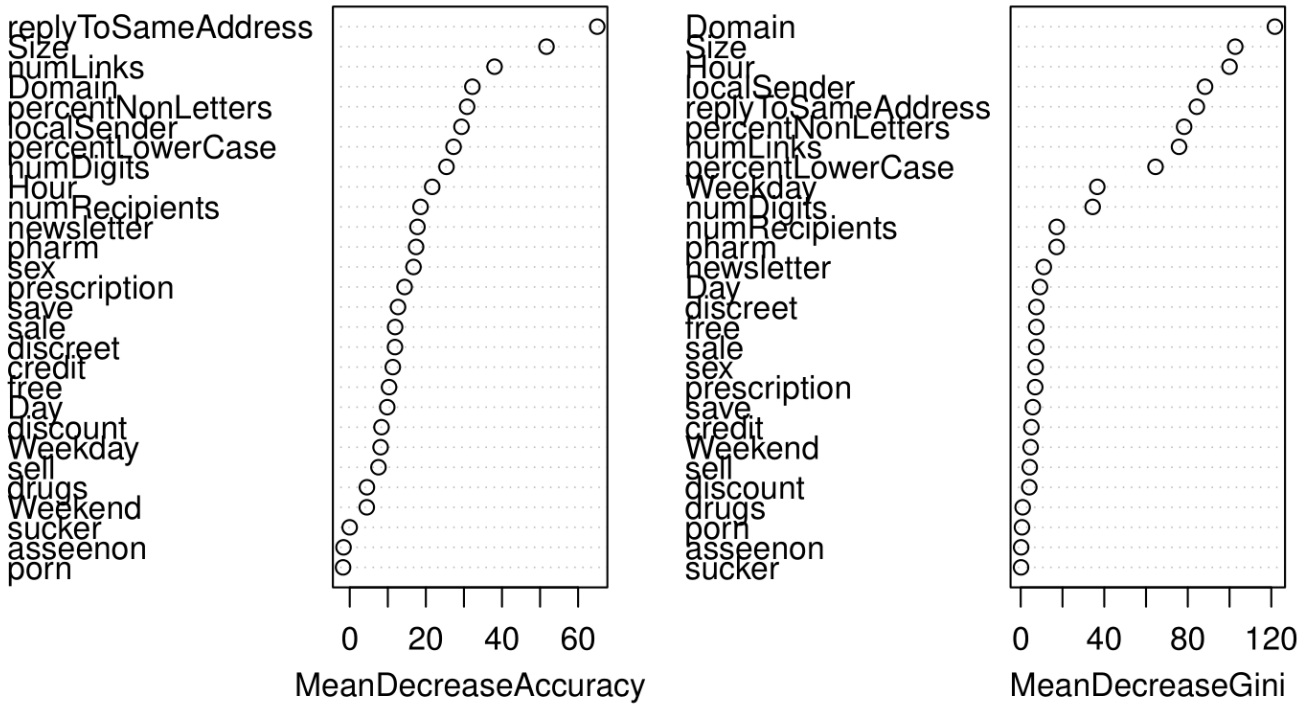



Figure 3. Importance plot from the random forest model.

# INTERESTING FINDINGS

*Strangely low number of instances of spam with the word porn. In our minds, the two seem highly correlated. However, the actual percentage of emails with the word porn that were spam was only 23%.*

*We thought perhaps that these legitimate emails could be pornographic newsletters people had voluntarily signed up to, but alas only 7 of all of the porn =1 observations were also newsletters.*

*This begs the question of how many people in the world are sending legitimate, porn-related emails.*