

# LINEAR STATISTICAL MODELS

SYS 4021

---

## Project 2a Static Spam Filter

---

Shareen Arshad  
sa2bw@virginia.edu

Chandler Dalton  
pcd4hb@virginia.edu

Fan Feng  
ff9sd@virginia.edu

### Summary

Spam emails have increased significantly over the past few years, resulting in high volumes of spam wasting email users time and money. The purpose of this study is to build a static spam filter which is able to accurately classify emails as spam and non-spam using a spam data set. Within the spam data set, there were 57 predictor variables of spam (V1 to V57) that could be used to measure the relationship on the response variable, V58, which classifies the email as spam or not spam. After in-depth graphical analysis to determine the best predictor variables, 5 different candidate models were built for the static spam filter. The 5 candidate models were then evaluated through performance metrics, including the amount of false positive and false negatives of spam emails. Reducing the number of false positives was concluded as the most important metric because it would mean a crucial email would have been removed as spam. After the evaluation, the main effects with interaction model proved to be have the best performance, with the lowest amount of false positives at 40, false negatives at 58, and total error at 98, resulting in a prediction accuracy of 93.5%. This model employed 12 variables as the predictors of our generalized linear model: V5, V7, V16, V21, V23, V24, V25, V52, V53, V55, V56 and V57, as well as interaction terms between predictors V55, V56, and V57. This model proved to also be statistically significant with a residual deviance of 1806.4 on 4540 degrees of freedom and a p-value of less than  $2 \times 10^{-16}$ . As a result, the interaction model proves to be valid with the low p-value, indicating its statistical significance. Thus, the model can be used to build a static spam filter using its indicated predictor variables to classify emails as spam or not spam.

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

Shareen Arshad, Chandler Dalton, Fan Feng

# 1 Problem Description

## 1.1 Situation

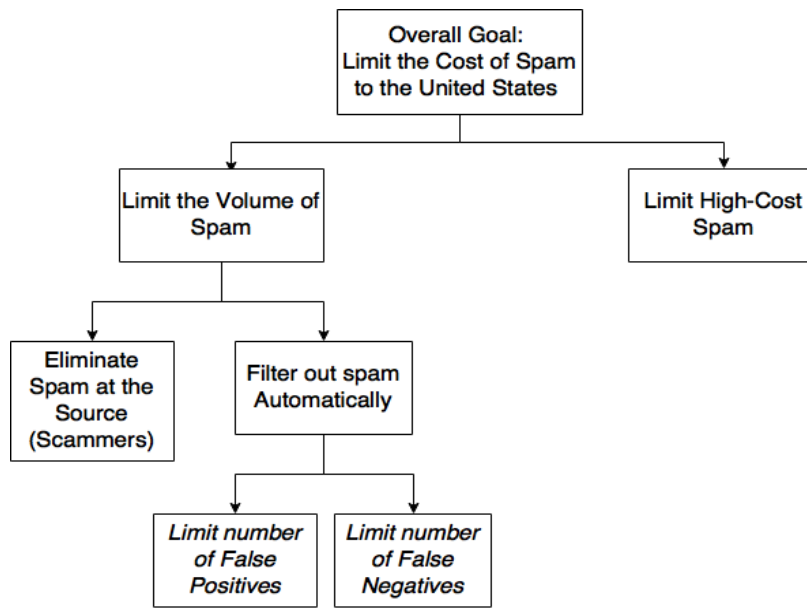
The term “spam” refers to “unsolicited commercial email (UCE) or unsolicited bulk email (UBE)” which is often affiliated with phishing scams, email fraud, foreign bank scams, pyramid schemes and other “Get Rich Quick” schemes, unapproved health products or remedies, ads, chain letters, and more [1]. Spam has increased significantly over recent years alongside the transition from paper letters to email as it costs spammers very little to send out billions of spam emails. In fact, spam is estimated to account for 60% of all emails sent, and even at a response rate of 0.00001%, spammers are still making money [2]. Spammers often use similar tricks to persuade people to open and read these emails, which generates two main costs: time wasted while reading spam emails and money lost through scams.

A 2004 National Technology Readiness Survey found that Internet users in the United States spend an average of three minutes deleting spam each day they use email, which, multiplied across 169.4 million adults and the average US salary, comes to \$21.58B lost annually in productivity. The same survey indicated that 4% of online adults have purchased a product or service advertised by spam, which represents almost 7 million people [3]. A further source estimated the cost of spammers at \$20B per year, which would lead our estimate for the total opportunity cost of spam to be about \$41.6B per year [4].

Therefore, there is a clear motivation to design a mechanism to limit the number of spam emails, both unwanted by the receiver and often containing costly scams [5]. Therefore, the purpose of this Static Spam Filter study is to use the dataset of variables and indicator of spam to deliver recommendations on how to detect spam emails and filter them from users’ email inboxes. In order to accomplish this, various generalized linear models will be applied to the predictor variables and the model will be evaluated through numerous performance metrics [6]. As a result, a final recommendation will be produced in order to identify components of a spam filter.

## 1.2 Goals

One powerful tool for eliminating the high cost and volume of spam is through filters, which detect spam by examining where the email came from, which software sent the message, and what is within the message. In this study, we aim to design a filter to detect the existence of spam based on the contents of the message, as indicated by variables in the dataset, which if successful, would limit the total cost of spam significantly (see Figure 1.2.1 below). Put explicitly, our goal is to build a model in R with predictor variables in the spam dataset to accurately predict the presence of spam within an email.



**Figure 1.2.1:** Objectives Tree for the Static Spam Filter

### 1.3 Metrics

Since the goal of this project is to build a model for prediction, we mainly used a confusion matrix in order to evaluate the performance of each model. For the two error types, false positives and false negatives, we believed that false positives are more costly because incorrectly detecting emails as spam runs the risk of deleting crucial emails for users. Conversely, if the filter fails to detect a spam, as is the case of false negatives, users can always delete it themselves. When we employed the test set method as the method for verifying our model, the prediction accuracy reflects the percentage of correct predictions among all the testing data. Additionally, AIC and BIC will also be used for our model selection because they both give insights about the performance of likelihood function whilst penalizing overly complicated models. However, AIC and BIC can only be used on models of the same scale. Thus, the confusion matrix indicating the number of false positives and false negatives for each model will be our primary metric for model comparison.

Figure 1.2.1 outlines the process of selecting a spam filter derived from the overall goal of limiting the cost of spam to the US in an overall objectives tree. This is then broken down into two subsections, which focuses on limiting the volume of spam and limiting high-cost spam. Since the goal of the project is to produce a static spam filter, this would directly apply to limiting the volume of spam because spam would be filtered out automatically. Given the fields of the dataset, designing a filter to limit both the number of false positives and false negatives is the most feasible objective for overall success in reducing cost.

## 1.4 Hypotheses

For this project, our hypothesis is as follows: By using some chosen predictors (within V1 to V57), we can build a generalized linear model to predict if an email is spam or not (V58) on a significant level, resulting in a recommendation for a static spam filter.

## 2 Approach

### 2.1 Data

The spam data set was contributed by the machine learning repository from the University of California, Irvine (UCI). The creators of the set were Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt [7]. The donor included George Forman [7]. The data was donated January 7th, 1999. The spam data set was a collection of emails which came from the creator's postmaster and individuals who had filed spam. The collection of non-spam emails were retrieved from field work and personal emails.

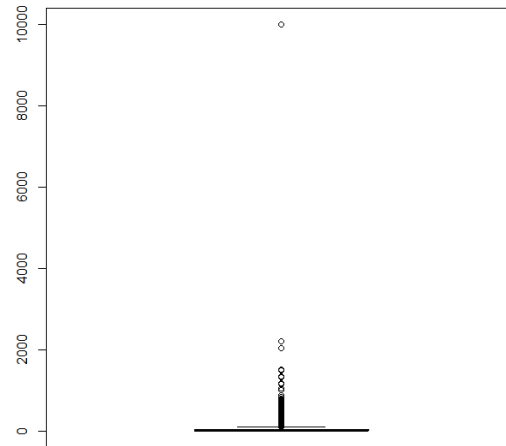
There are a few general data set characteristics that should be noted about the spam data set. The data set characteristics are multivariate and the attribute characteristics are either integer or real. Moreover, there are associated tasks that can be linked to the data set, such as classification of whether an email is spam or not. For the purpose of this report, the spam data set will be used in generating a static spam filter to accurately classify if an email is spam or not, utilizing predictors in the data set itself.

There are many specific features about the spam data set that were noted during analysis. There are 4601 total number of email instances. Of these instances, 1813 or 39.4% are classified as spam. For each one of these instances, there are 57 attribute variables, with the 58th variable being a binary variable that classifies the email as either spam, which is denoted with a 1, or not spam, which is denoted with a 0. A description of these variables can be found within the Appendix in Table 2.1, as well as summary statistics for each variable within the Appendix in Table 2.2.

For the purpose of creating a static spam filter, exploratory analysis was performed on the spam data set in order to provide motivation for the creation of candidate models. After becoming familiar with the data and the predictors, the relationship between various predictors, V1 to V57, and the response variable, V58, was performed. Most of this exploratory analysis was included within Section 2.2.1, as this also provided the basis and justification for various model creations. Within the exploratory data analysis, this included the creation of scatter plot matrices to identify correlations (see Figure 2.2.1) between predictor and the response variable. After identifying which predictors were most correlated, boxplots were also created to identify if the predictors were discriminatory of spam (see Figure 2.1.2), as well as boxplots on log transformed predictors as well (see Figure 2.1.3). Biplots were also analyzed of all of the predictor variables (see Figure 2.1.5), as well as on separate graphs for sets of variables. Therefore, this data analysis provided the motivation for the creation of the various candidate models and is explained more in detail within Section 2.2.1.

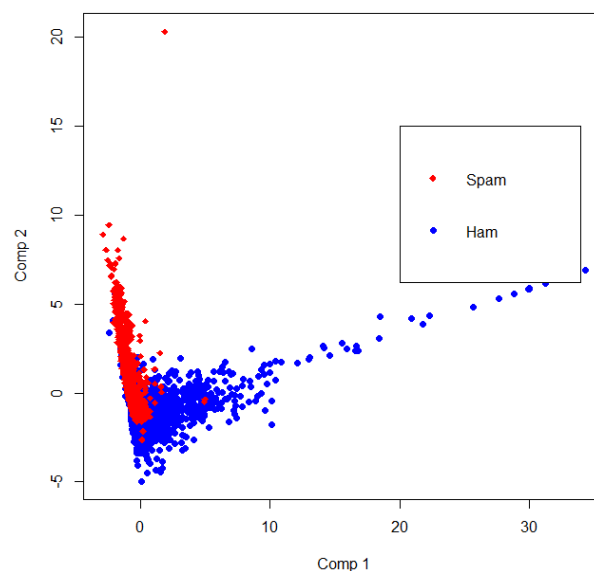
Finally, the dataset used does not include missing data. However, in terms of possible biases, the existence of multicollinearity between predictors and outliers was considered. To account for multicollinearity, interaction terms were included in one of the

candidate models, as shown in Figure 2.1.4 and Table 2.1.7. To account for outliers, the biplot from Figure 2.1.5 was further investigated by creating a box plot of the 2nd principal component loadings, as shown in Figure 2.1.



**Figure 2.1** Boxplot of PCA Loadings with Identified Outlier

This identified outlier was further explored in a biplot of spam vs “ham” emails, where ham represents non-spam emails.



**Figure 2.2** Spam vs Ham Emails within Data Set

In Figure 2.2, the identified outlier is still present and is classified as spam. In creating a static spam filter, it was concluded that this large outlier, identified as observation point 1754, could certainly skew results and represent bias. In order to

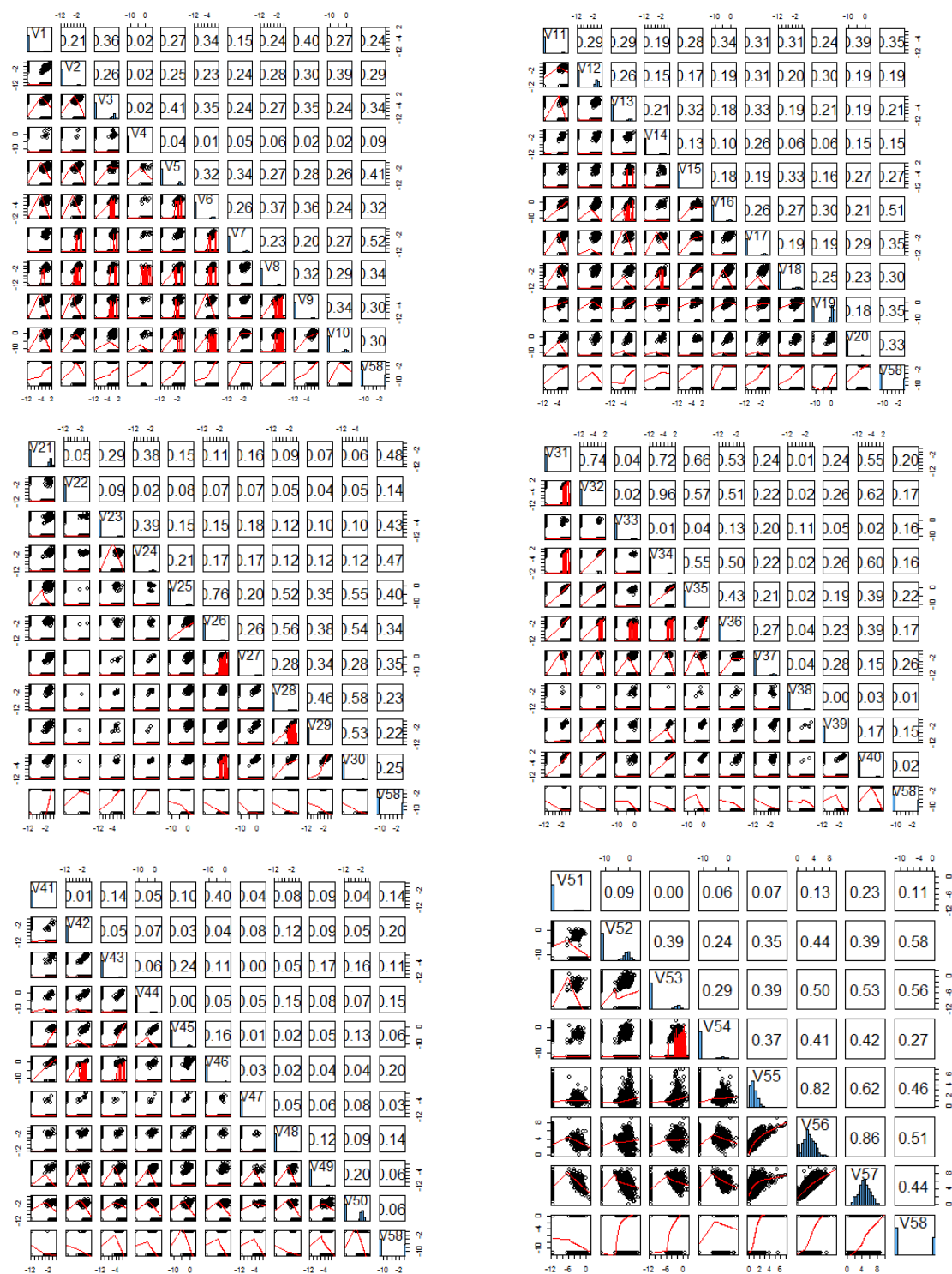
account for this, a Cook's distance plot was generated for each model in Section 2.2.2 and if the outlier was an identified influential point in the data by having a large Cook's Distance, it was extracted from the model before performance metrics were generated in model comparison. Also, the chosen model was compared before and after to see if performance metrics would change as the outlier was extracted.

## **2.2 Analysis & Evidence**

### **2.2.1 Static Analysis for Spam Filter Design**

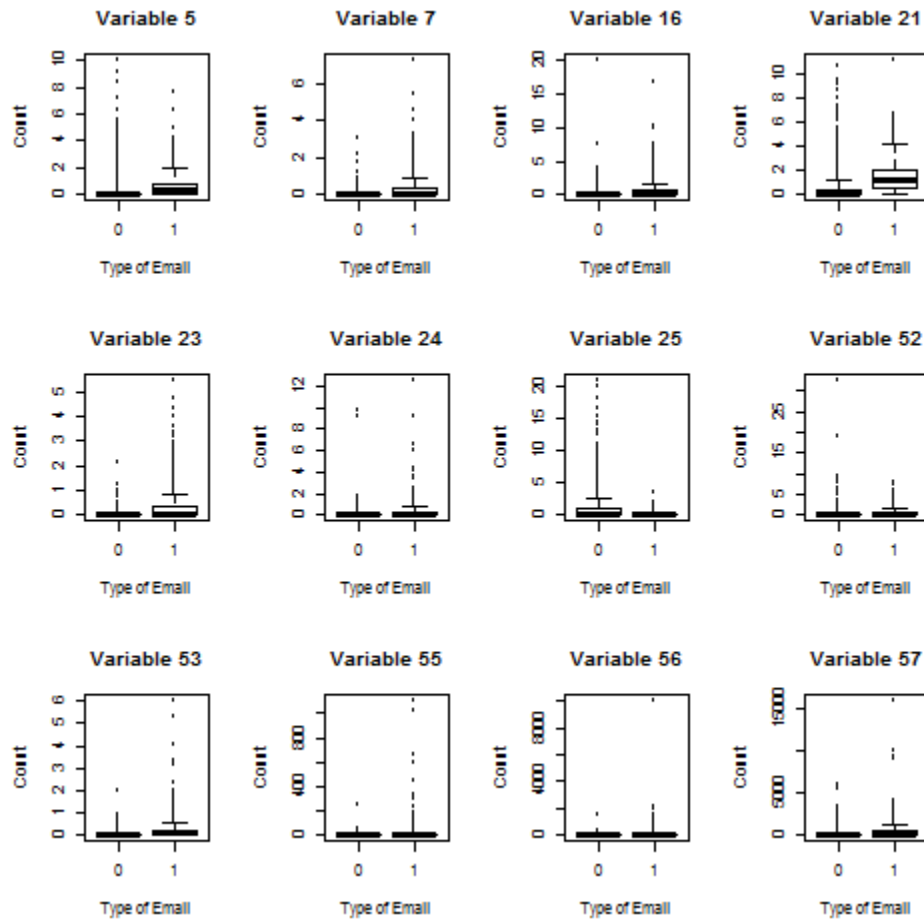
Since the goal of this project is to build a model for predicting if an email is spam or not, we first choose our predictors according to how correlated the predictors are with the response variable (V58). We do so by creating the scatter plot matrix of all the variables in the dataset in Figure 2.1.1.

By reading predictors' correlation factors from the scatter plot matrix, we picked predictors that at least have a correlation factor of 0.4 and above with the response variables: V5, V7, V16, V21, V23, V24, V25, V52, V53, V55, V56 and V57. As a result, a main effects model was built from these variables (see Table 2.1.7). We then created the factor plots of these 12 variables to see if these variables are discriminatory in terms of spam vs. ham, as seen in Figure 2.1.2.



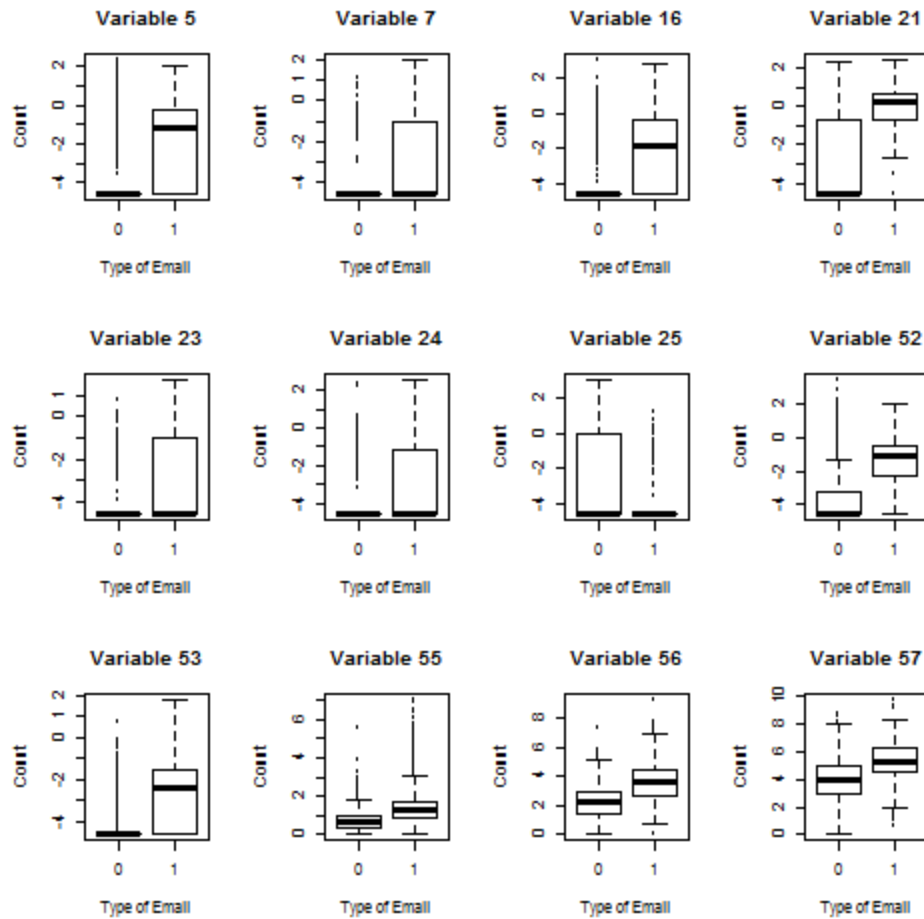
**Figure 2.1.1: Scatter Plot Matrix of All Predictor Variables vs. Response Variable**





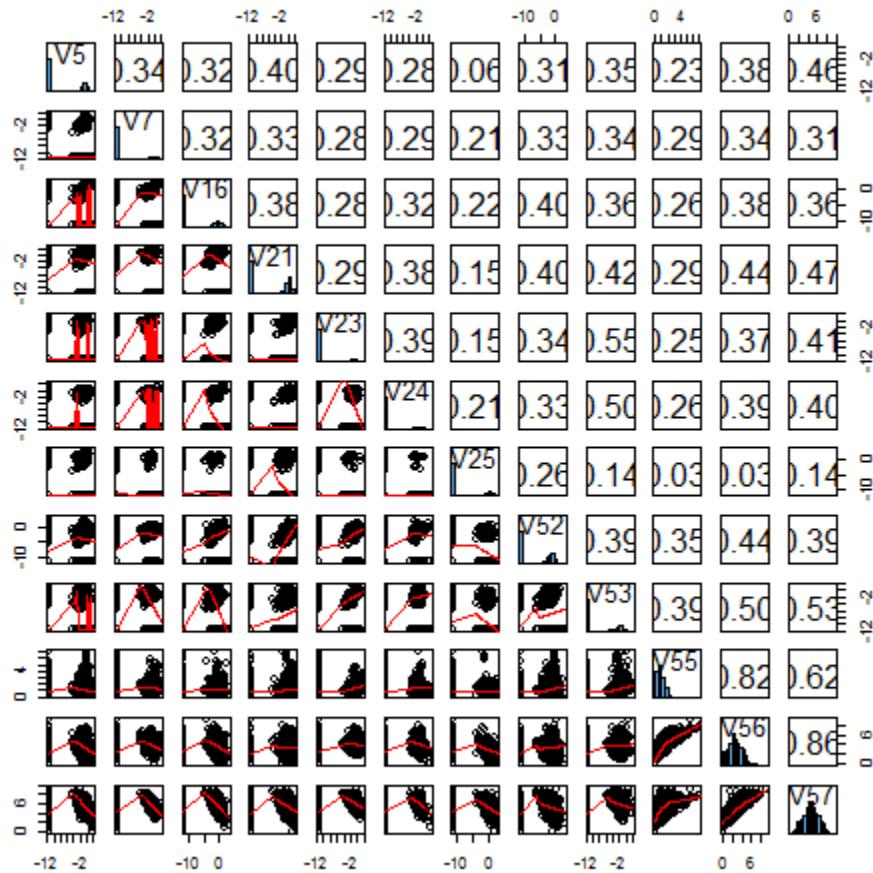
**Figure 2.1.2:** Factor Plot for 12 Selected Predictors

According to Figure 2.1.2, most of our selected predictors fail to show discriminatory results of the response variable. Therefore, we changed the scale of predictors to logarithm with 0.01 added as offset to reproduce the factor plots with log-transformed predictors, shown in Figure 2.1.3.



**Figure 2.1.3:** Factor Plot of 12 Selected Log-Transformed Predictors

As we can see from Figure 2.1.3, all the variables we select with a correlation factor greater than or equal to 0.4 show discriminatory results on the response variable. We then decided to build a main effects model with the log transformed predictors (see Table 2.1.7). Additionally, in order to account for possible multicollinearity between predictors, we generated the scatter plot matrix of the 12 selected predictors, as shown in Figure 2.1.4,

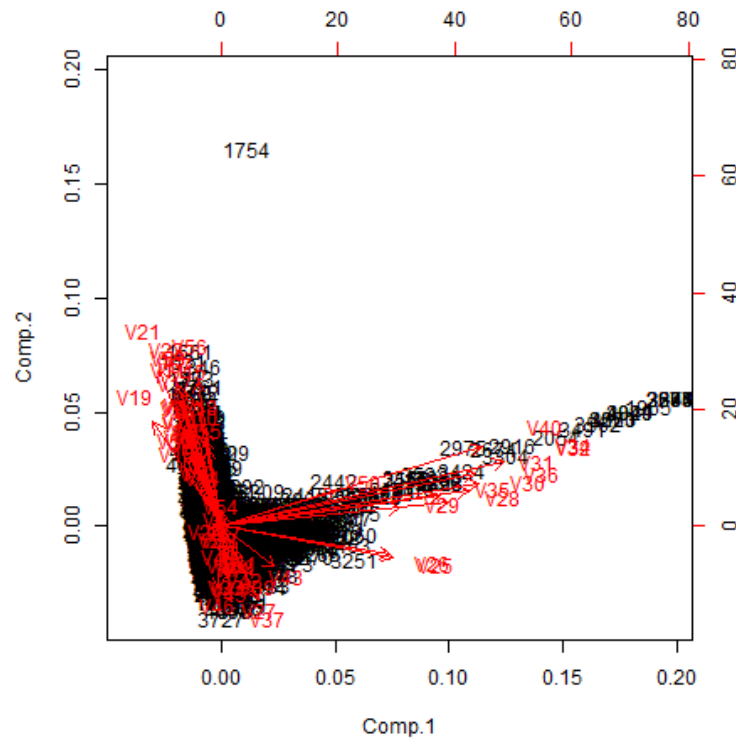


**Figure 2.1.4:** Scatter plot Matrix for 12 Selected Predictors

We used 0.5 as a threshold of whether or not to be included as part of an interaction term. Thus, we chose V55, V56 and V57 to be our interaction terms. According to the interpretation of the data set, this choice makes sense because V55, V56 and V57 correspond, respectively, to `capital_run_length_average`, `capital_run_length_longest` and `capital_run_length_total`, which can be highly correlated by intuition and because they represent different attributes within the data set. Thus, a main effects with interaction terms model was built with these 12 predictors and interactions between these three predictors (see Table 2.1.7).

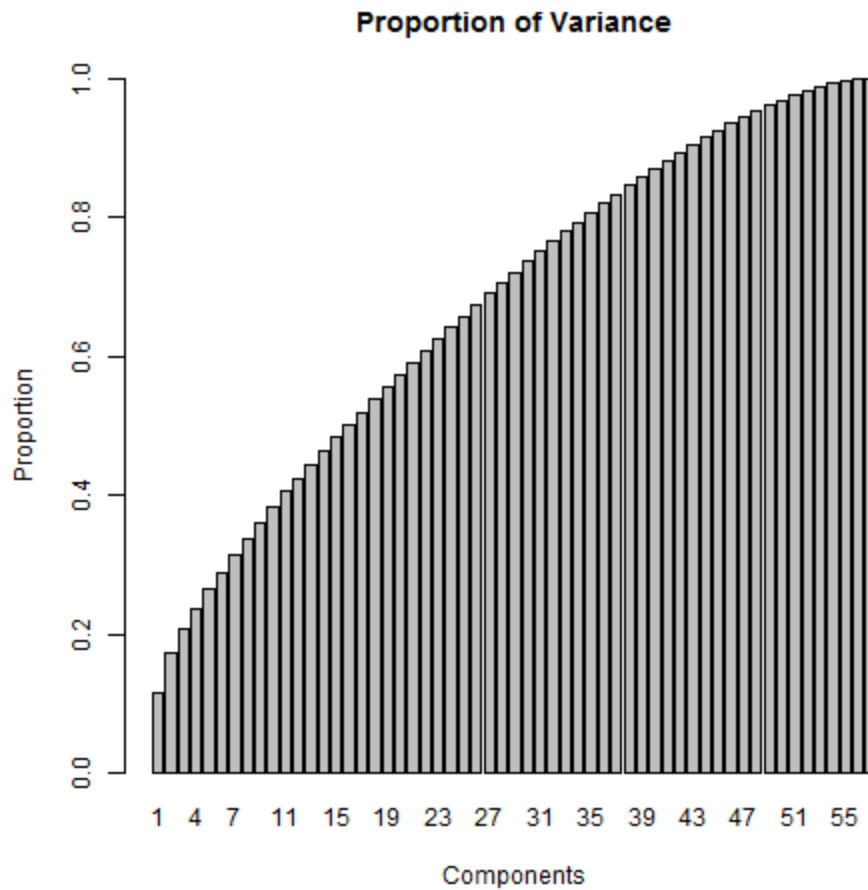
Apart from the above three models, we also included the stepwise model of the main effect model to reduce the number of predictors. After running the stepwise function, we found the only variable eliminated from the main effects model was V55, and this result accords with the Chi-square test we ran on the main effect model with the `drop1` function in R. All the predictors in the main effect model had 3 stars, meaning significance level of 0, in the Chi-square test except for V55. Therefore, this variable was excluded from the stepwise regression model (see Table 2.1.7).

Finally, we used PCA to account for possible latent characteristics in the data and reduce dimensionality. A PCA biplot is shown in Figure 2.1.5.



**Figure 2.1.5:** Biplot of PCA of All Predictors

In order to choose the appropriate number of components in building the model, we chose the components that account for 95% of variance in the dataset, which according to the cumulative plot, are the first 48 components in PCA as shown in Figure 2.1.5. In terms of the other PCA models that accounted for 90% and 98% of the variance, the 95% variance model performed the best and used less amount of predictors as the 98% variance PCA model. Thus, the 95% variance PCA model was chosen as one of the five candidate models (see Table 2.1.7). Compared to the other candidate models, the PCA regression model has many more predictors, and we took this factor into consideration while choosing between the different models presented.



**Figure 2.1.6:** Cumulative plot of PCA of all predictors

Therefore, through the data analysis of the spam data set and the exclusion of various predictors, the following five candidate models were built as possible static spam filters. All models proved to have statistical significance, proving to be good contenders for candidate models for a static spam filter.

Model Number	Response Variable	Predictors
1 (Main Effects Model)	V58	V5,V7, V16, V21, V23, V24, V25, V52, V53, V55, V56, V57
2 (Main Effects Model with Log-transformed Predictors)	V58	$\log(V5) + 0.1, \log(V7) + 0.1, \log(V16) + 0.1, \log(V21) + 0.1, \log(V23) + 0.1, \log(V24) + 0.1, \log(V25) + 0.1, \log(V52) + 0.1, \log(V53) + 0.1, \log(V55) + 0.1, \log(V56) + 0.1, \log(V57) + 0.1$
3 (Interaction Model)	V58	V5,V7, V16, V21, V23, V24, V25, V52, V53, V55, V56, V57, V55*V56, V55*57, V56*V57
4 (Stepwise Main Effects Model)	V58	V5,V7, V16, V21, V23, V24, V25, V52, V53, V56, V57
5 (PCA Model)	V58	Principal Components that account for 95% of variance (first 48 components)

**Table 2.1.7:** Variables Utilized in 5 Candidate Models

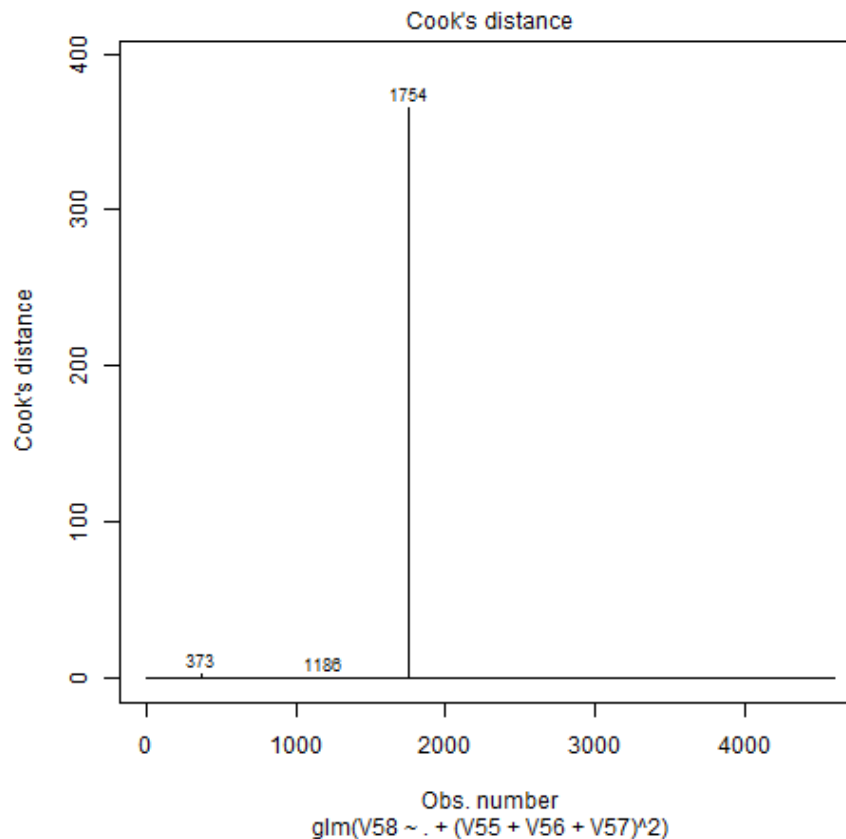
### 2.2.2 Static Filter Design

In order to compare the different models generated in the section above, we first came up with the performance metrics for each of the models. Some of these metrics include generating the AIC, BIC, and Residual Deviance for the models; however, untransformed models cannot be compared with log transformed models through AIC and BIC.

Instead, a confusion matrix was generated in order to identify the number of false positive and false negatives within each model. Our primary focus is the number of false positives generated because this would be a worst-case scenario, where a non-spam email would be classified as spam, as identified in Section 1.3. Thus, in creating the confusion matrix for each model, a test set was created that used 33% of the spam data. Then, each model was then applied to this same test set in order to create a trained spam data set which classified the emails as spam or non-spam. After these trained sets were created, a

confusion matrix was developed to record the number of false positive, false negatives, total error, and prediction accuracy. The results for each model were recorded in Table 2.2.9.

However, before assessing performance metrics of each model, the influential points within each model were assessed and removed if there was a Cook's Distance greater than 1.0. After analyzing all the models, an influential point was only identified for the main effects and interaction model, as demonstrated in Figure 2.2.8.



**Figure 2.2.8:** Cook's Distance Plot for the Interaction Model

In Figure 2.2.8, an abnormally large Cook's Distance can be found at observation 1754. This was the same outlier identified in section 2.1 as an outlier in the data and it is clear that this identified outlier is an influential point in this model's performance. To rectify this problem within the model, the model was adjusted by extracting this observation point from this model's data before pursuing model evaluation to avoid skewing the results.

Next, the performance metrics comparison of the 5 Candidate models was performed as outlined at the beginning of section 2.2.2. The results are shown in Table 2.2.9 below.

**Table 2.2.9: Performance Metrics Comparison of 5 Candidate Models**

Model Number	AIC	BIC	Residual Deviance	False Positive	False Negative	Total Error	Prediction Accuracy
1	2745.323	2828.965	2719.3	56	108	164	89.2%
2	1957.132*	2040.774 *	1931.1	48	62	110	92.8%
3	1928.392	2320.868	1806.4	40	58	98	93.5%
4	2743.42	2820.628	2719.4	56	108	164	89.2%
5	2052.623	2367.891	1954.6	45	62	107	92.9%

\* Not appropriate for comparison due to different scale (log-transform)

Key

- 1: Main Effects Model
- 2: Main Effects Model with Log-transformed Predictors
- 3: Main Effects Model with Interaction Terms
- 4: Stepwise Main Effects Model
- 5: PCA Model

From the chart above, it can be seen that the Main Effects Model with Interaction Terms, model 3, performs the best in terms of all of the performance metrics in comparison to the other models. It gives the lowest number of false positives, which is the performance metric we are most concerned about for our spam filter. Moreover, the interaction model gives the lowest number of false negatives and total error as well, resulting in the model's prediction accuracy to be the highest, with 93.5%. Also, in terms of AIC, BIC, and Residual Deviance in comparison to the untransformed models, the interaction model once again has the lowest value, with an AIC of 1928.4 and BIC of 2320.9, ranking it better in terms of these performance metrics. Therefore, the interaction model is the best model choice for the spam filter.

The threshold for the confusion matrix was also analyzed for the interaction model. Thus, a range of values were chosen to assess how the threshold would determine the number of false positives, false negatives, and total error produced, which is shown below.

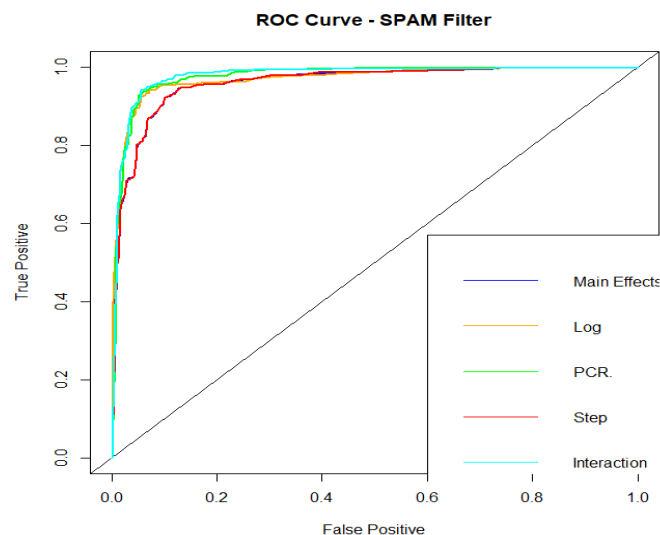


Confusion Matrix Threshold	False Positive	False Negative	Total Error
0.35	72	35	107
0.40	60	40	100
0.45	53	46	99
0.50	40	58	98
0.55	35	65	100
0.60	34	71	105
0.65	30	83	113

**Table 2.2.10:** Altered Confusion Matrix Threshold for Interaction Model

Within Table 2.2.10, we can see that the number of false positives decreases with an increase in the confusion matrix threshold in the interaction model. Therefore, 0.65 would be the best confusion matrix threshold to reduce the number of false positives, but there is also a tradeoff in the fact that the number of false negatives would increase, resulting in a larger total error. Thus, the 0.5 confusion matrix threshold proves to be the optimal threshold because it results in the least total error with 98, but also the least amount of false positives with 40.

Furthermore, in order to focus more on the false positives generated by each model, an ROC curve was created on each train set for each model. This visualization for comparison can be seen below.



**Figure 2.2.11:** ROC curve of 5 Candidate Models

In the above ROC curve generated for all of the models, the main effects model is denoted by the blue line, the in log transformation main effects by the orange line, the interaction model by the cyan line, the stepwise model by the red line, and the PC regression model of 95% variance by the green line. From the ROC curve above, it is clear that the cyan line generated by the interaction model gives the best ROC curve, as it reaches the top left corner of the graph the most. As a result, the false positives are reduced the most by this model and the true positives are maximized, meaning that less non-spam emails will be classified as spam under this model compared to other models. Thus, the interaction model serves as the best model to generate the spam filter.

This model relates back to our goals mentioned in section 1.2. The predictors within the model as mentioned in Table 2.1.7 are important predictors to account for in a static spam filter, as well as the interaction terms between the 3 predictors, V55, V56, and V57. Therefore, in order to limit the amount of spam volumes within email mailboxes, a static spam filter must be able to search through these predictors, as well as the interactions between these predictors in order to classify the email as spam or not spam accurately. The model proved this through a high accuracy rate of 93.5% and a low total error amount of 98 in comparison to the other candidate models.

In terms of the statistical confidence of the model, the interaction model was found to have statistical significance when compared to the main effects model through the partial likelihood test. Although all the models were statistically significant, this ultimately became the chosen model due to high performance in the performance metrics and it is important to provide the statistical significance of this model. The results are shown in Table 2.2.12.

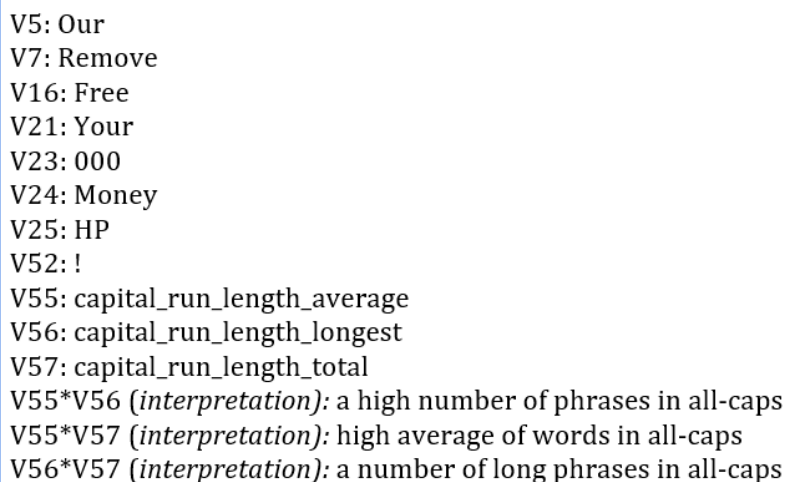
Partial Likelihood Test for Interaction Model				
Res.Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1 4588	2719.3			
2 4540	1806.4	48	912.93	< 2.2e-16 ***

**Table 2.2.12: Interaction Model Statistical Significance**

Due to the low p-value of less than 2.2e-16, it is clear that the main effects model with interaction terms is holds statistical significance, meaning that the larger model must be chosen which includes the interaction terms. Thus, the hypothesis made in section 1.4 can be confirmed and narrowed down to the predictors present in the interaction model along with the interaction terms in order to create the optimal static spam filter.

### 3 Recommendation

Based on our analysis above, our recommendation for the spam filter would include variables from model 3, the Interaction Model (see Table 2.1.7). These variables are listed below in Figure 3.1. It can be inferred that these variables would intuitively be in most spam emails, as they address the words money, free, and words in all capital letters, which are very common in many spam emails. A Spam Static Filter designed with these components would, if the dataset was representative of the typical American email box, be expected to have a spam detection accuracy of about 93.5%. The model also proved to have a high statistical significance with a p-value of less than  $2.2e-16$  (see Table 2.2.12). Outliers were also accounted for within the model from the spam data set used (Figure 2.2.8). This filter would tend to have more false negative errors than false positives, where the spam emails are better off being displayed in the email inbox versus a false positive error where a critical email is removed as spam. Further, it proved to be the best filter design across a variety of model performance metrics.



V5: Our  
V7: Remove  
V16: Free  
V21: Your  
V23: 000  
V24: Money  
V25: HP  
V52: !  
V55: capital\_run\_length\_average  
V56: capital\_run\_length\_longest  
V57: capital\_run\_length\_total  
V55\*V56 (interpretation): a high number of phrases in all-caps  
V55\*V57 (interpretation): high average of words in all-caps  
V56\*V57 (interpretation): a number of long phrases in all-caps

**Figure 3.1:** Filter Design Components

The interaction model performed the best in terms of the performance metrics. These results are recorded in Table 3.2. The optimal confusion matrix threshold proved to be 0.5 to reduce the amount of false positives at 48, false negatives at 50, total error overall at 98 (Table 2.2.9).

Model Number	AIC	BIC	Residual Deviance	False Positive	False Negative	Total Error	Prediction Accuracy
3	1928.392	2320.868	1806.4	40	58	98	93.5%

***Table 3.2: Performance Metrics of Interaction Model***

Given the predicted accuracy of the model (93.5%) and the estimated opportunity cost of spam emails (\$41.6B per year in the US--\$21.6B from loss in productivity and \$20B from explicit costs), we would estimate the total savings to the US to be about \$38.9B<sup>1</sup>. This would be less than 2.65% of false positive emails, which are those that are not spam but labeled as “spam” by the filter, which are difficult to quantify in opportunity cost. This is a significant yield in savings, and would indicate tremendous success if the study is able to be translated to general United States email servers; however, clear limitations come in the form of implementation costs, as well as the highly variable cost in false positives.

---

<sup>1</sup> (Gross Savings - False Negatives) = \$41.6B\*(0.935)

## 4 References

- [1] "What is Spam?" *Indiana University Knowledge Base*. (Modified Oct 2017). Retrieved from: <https://kb.iu.edu/d/afne>
- [2] Conner, Katie (Jul 2017). "Spam Costs Businesses, Consumers Billions of Dollars Per Year; How to Protect Yourself From Spam." Retrieved from: <http://www.abc15.com/news/local-news/water-cooler/spam-costs-businesses-consumers-billions-of-dollars-per-year-how-to-protect-yourself-from-spam>
- [3] Claburn, Thomas (Feb 2005). "Spam Costs Billions." *InformationWeek*. Retrieved from: <https://www.informationweek.com/spam-costs-billions/d/d-id/1030111>
- [4] "The Economics of Spam?" *Stop Junk Mail*. (Aug 2012). Retrieved from: <https://stopjunkmail.org.uk/blogs/diary/2012/08/economics-spam>
- [5] L. E. Barnes, "Project 2: Static spam filter," Class project in SYS 4021, 2017.
- [6] L.E. Barnes, "Project 2 template," Class template in SYS 4021, 2017.
- [7] Data source: SPAM E-mail Database.  
Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt.  
Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304  
Donor: George Forman (gforman at nospam hpl.hp.com) 650-857-7835  
Generated: June-July 1999

## 5 Optional Appendix

**Table A.1: Variable Descriptions**

<b>Variable Number</b>	<b>Type</b>	<b>Continuous</b>	<b>Real/Integer</b>
1	word_freq_make	Continuous	Real
2	word_freq_address	Continuous	Real
3	word_freq_all	Continuous	Real
4	word_freq_3d	Continuous	Real
5	word_freq_our	Continuous	Real
6	word_freq_over	Continuous	Real
7	word_freq_remove	Continuous	Real
8	word_freq_internet	Continuous	Real
9	word_freq_order	Continuous	Real
10	word_freq_mail	Continuous	Real
11	word_freq_receive	Continuous	Real
12	word_freq_will	Continuous	Real
13	word_freq_people	Continuous	Real
14	word_freq_report	Continuous	Real
15	word_freq_addresses	Continuous	Real
16	word_freq_free	Continuous	Real
17	word_freq_business	Continuous	Real
18	word_freq_email:	Continuous	Real
19	word_freq_you:	Continuous	Real
20	word_freq_credit:	Continuous	Real
21	word_freq_your:	Continuous	Real

22	word_freq_font:	Continuous	Real
23	word_freq_000:	Continuous	Real
24	word_freq_money:	Continuous	Real
25	word_freq_hp:	Continuous	Real
26	word_freq_hpl:	Continuous	Real
27	word_freq_george:	Continuous	Real
28	word_freq_650:	Continuous	Real
29	word_freq_lab:	Continuous	Real
30	word_freq_labs:	Continuous	Real
31	word_freq_telnet:	Continuous	Real
32	word_freq_857:	Continuous	Real
33	word_freq_data:	Continuous	Real
34	word_freq_415:	Continuous	Real
35	word_freq_85:	Continuous	Real
36	word_freq_technology:	Continuous	Real
37	word_freq_1999:	Continuous	Real
38	word_freq_parts:	Continuous	Real
39	word_freq_pm:	Continuous	Real
40	word_freq_direct:	Continuous	Real
41	word_freq_cs:	Continuous	Real
42	word_freq_meeting:	Continuous	Real
43	word_freq_original:	Continuous	Real
44	word_freq_project:	Continuous	Real
45	word_freq_re:	Continuous	Real
46	word_freq_edu	Continuous	Real
47	word_freq_table:	Continuous	Real
48	word_freq_conference:	Continuous	Real

49	char_freq_:	Continuous	Real
50	char_freq_(	Continuous	Real
51	char_freq_[	Continuous	Real
52	char_freq_!	Continuous	Real
53	char_freq_\$	Continuous	Real
54	char_freq_#	Continuous	Real
55	capital_run_length_average:	Continuous	Real
56	capital_run_length_longest:	Continuous	Integer
57	capital_run_length_total:	Continuous	Integer
58	Spam {1} vs no spam {0}	N/A	Nominal

\* Variables 1-48: percentage of words in the e-mail that match WORD, i.e.  $100 * (\text{number of times the WORD appears in the e-mail}) / \text{total number of words in e-mail}$ . A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

\* Variables 49-54: percentage of characters in the e-mail that match CHAR, i.e.  $100 * (\text{number of CHAR occurrences}) / \text{total characters in e-mail}$

\* Variable 55: average length of uninterrupted sequences of capital letters

\* Variable 56: length of longest uninterrupted sequence of capital letters

\* Variable 57: sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

\* Variable 58: denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.



**Table A.2: Variable Summary Statistics**

	sd	min	median	mean	max
V1	0.30535756	0	0.000	1.045534e-01	4.540
V2	1.29057519	0	0.000	2.130146e-01	14.280
V3	0.50414288	0	0.000	2.806564e-01	5.100
V4	1.39515137	0	0.000	6.542491e-02	42.810
V5	0.67251277	0	0.000	3.122234e-01	10.000
V6	0.27382408	0	0.000	9.590089e-02	5.880
V7	0.39144135	0	0.000	1.142078e-01	7.270
V8	0.40107145	0	0.000	1.052945e-01	11.110
V9	0.27861586	0	0.000	9.006738e-02	5.260
V10	0.64475540	0	0.000	2.394132e-01	18.180
V11	0.20154466	0	0.000	5.982395e-02	2.610
V12	0.86169847	0	0.100	5.417018e-01	9.670
V13	0.30103580	0	0.000	9.392958e-02	5.550
V14	0.33518383	0	0.000	5.862639e-02	10.000
V15	0.25884345	0	0.000	4.920452e-02	4.410
V16	0.82579170	0	0.000	2.488481e-01	20.000
V17	0.44405533	0	0.000	1.425864e-01	7.140
V18	0.53112242	0	0.000	1.847446e-01	9.090
V19	1.77548066	0	1.310	1.662100e+00	18.750
V20	0.50976689	0	0.000	8.557705e-02	18.180
V21	1.20080981	0	0.220	8.097609e-01	11.110
V22	1.02575559	0	0.000	1.212019e-01	17.100
V23	0.35028642	0	0.000	1.016453e-01	5.450
V24	0.44263553	0	0.000	9.426864e-02	12.500
V25	1.67134934	0	0.000	5.495045e-01	20.830
V26	0.88695534	0	0.000	2.653836e-01	16.660

V27	3.36729180	0	0.000	7.673049e-01	33.330
V28	0.53857604	0	0.000	1.248446e-01	9.090
V29	0.59332660	0	0.000	9.891545e-02	14.280
V30	0.45668155	0	0.000	1.028516e-01	5.880
V31	0.40339250	0	0.000	6.475331e-02	12.500
V32	0.32855888	0	0.000	4.704847e-02	4.760
V33	0.55590720	0	0.000	9.722886e-02	18.180
V34	0.32944533	0	0.000	4.783525e-02	4.760
V35	0.53225988	0	0.000	1.054119e-01	20.000
V36	0.40262314	0	0.000	9.747664e-02	7.690
V37	0.42345137	0	0.000	1.369528e-01	6.890
V38	0.22065079	0	0.000	1.320148e-02	8.330
V39	0.43467205	0	0.000	7.862856e-02	11.110
V40	0.34991598	0	0.000	6.483373e-02	4.760
V41	0.36120470	0	0.000	4.366659e-02	7.140
V42	0.76681944	0	0.000	1.323386e-01	14.280
V43	0.22381178	0	0.000	4.609867e-02	3.570
V44	0.62197557	0	0.000	7.919583e-02	20.000
V45	1.01168723	0	0.000	3.012236e-01	21.420
V46	0.91111906	0	0.000	1.798240e-01	22.050
V47	0.07627427	0	0.000	5.444469e-03	2.170
V48	0.28573465	0	0.000	3.186916e-02	10.000
V49	0.24347133	0	0.000	3.857466e-02	4.385
V50	0.27035537	0	0.065	1.390304e-01	9.752
V51	0.10939416	0	0.000	1.697587e-02	4.081
V52	0.81567163	0	0.000	2.690709e-01	32.478
V53	0.24588201	0	0.000	7.581069e-02	6.003
V54	0.42934209	0	0.000	4.423821e-02	19.829
V55	31.72944874	1	2.276	5.191515e+00	1102.500
V56	194.89130953	1	15.000	5.217279e+01	9989.000
V57	606.34785072	1	95.000	2.832893e+02	15841.000
V58	0.48869765	0	0.000	3.940448e-01	1.000