Detecting Email Spam

MGS 8040, Data Mining

Audrey Gies
Matt Labbe
Tatiana Restrepo

5 December 2011

INTRODUCTION

This report describes a model that may be used to improve likelihood of recognizing undesirable email commonly known as spam. We discuss the processes and results of building a linear discriminant regression model based on a set of given data. This model can be used to reduce the number of undesired emails that are allowed into a specific individual's inbox and thus improve their productivity. Actual mail filtered as spam (false positives) is very undesirable and this will be taken into account when making final recommendations. A successful model would classify the majority of spam emails correctly while having a minimal amount of non-spam emails misclassified as spam.

DATA

The dataset was obtained from the University of California's Machine Learning Repository, Center for Machine Learning and Intelligent Systems (http://archive.ics.uci.edu/ml/index.html). They currently maintain 211 data sets as a service to the machine learning community, and thus are cited in over 1,000 papers.

The particular dataset chosen contained data regarding spam emails (http://archive.ics.uci.edu/ml/datasets/Spambase). The data was collected and classified in 1999 and is specific to an individual, George a mail server administrator at Hewlett-Packard Labs, and contains both personal and work emails. Since this is specific to one individual, the word "George" and the year "1999" are indicators of actual emails.

The dependent variable is "spam," and denotes whether the e-mail was considered spam (1) or not (0). The outcome period for determining if an observation is spam is indefinite. There are 57 independent variables with 48 variables being the percent of words in the e-mail that match a given word. For example "word_freq_free" is the percentage of times the word "free" appears in the email. There are 6 variables displaying the percent of characters in the e-mail that match a given character, such as: exclamation points, semi-colons, and dollar signs. Lastly, there are 3 variables with statistics regarding the use of capital letters in the email. Overall there are 4,601 observations with 39.4% currently marked as spam. Please see Appendix A for a full data dictionary. No data scrubbing or variable creation was necessary. No variables were missing and outlier data, which is common when applied to word frequency in short text communication like emails, was retained.

METHODOLOGY

The process undertaken followed the traditional steps for linear discriminant regression analysis.

1. Import and Examine the Data

The raw data was imported into Excel in the CSV format and each column was labeled with the appropriate variable names. A univariate analysis was performed to find any missing, negative and unusual values.

2. Define dummies

After familiarizing ourselves with the data we did a 50/50 split of the data into a training and validation dataset. From there we defined our categories, created our crosstabs (frequency of each variable against the "spam" variable), calculated the good to bad ratio, and created the dummy breakpoints for each variable.

Table 1. Dummy Variable creation example

Table of word_freq_over by spam					
word_freq_over	spam			Good/Bad Ratio	Dummy Group
	0	1	Total		
0	1266 89.47	580 63.25	1846	0.707	N
0.01 TO 0.25	80 5.65	117 12.76	197	2.258	1
0.26 TO 0.50	36 2.54	94 10.25	130	4.035	2
0.51 OR MORE	33 2.33	126 13.74	159	5.897	3
Total	1415	917	2332		

3. Build regression model

Once the dummy variables were created the regression model was ready to be built. We started with all of the created dummy variables, 137 total, and then eliminated variables with high p-values. When conducting the final iterations of the model creation, we ensured all parameter estimates were significant at the 95% confidence level. Once the final model was selected the coefficients were actually significant at the 97% confidence level. We continued to evaluate the parameter estimates to ensure that they matched the behavior seen in the initial frequency analysis. All coefficients of the final model were felt to be meaningful. A collection of variables including "project" and "650" were removed and re-added to the model to ensure that they were contributing to the effectiveness of the model. Dummy variables with parameter estimates that were similar to neighboring ranges parameter estimates were combined to simplify the model and increase the model's applicability to the validation data set. The first and final model's complete regression output is displayed in Appendix B.

4. Score the model

To score all observations we ran a scoring program against the training dataset first. The scores were formatted so they would range from 0 to 1,000. These steps were repeated with the validation data once the KS test for the training set appeared to be within reasonable bounds.

5. Complete KS Test

The first KS test table was completed using the results from step 4 for the training data. We found the optimal point and felt it was within reasonable bounds of the desired 10% and continued to create the KS test table for the validation data. The final KS test results are shown in the "Results" section.

6. Create the scorecard

Once the model was finalized the scorecard was created and all variables were checked to ensure logical points and trending. The final scorecard is shown in the "Results" section.

PROJECT FLOW DIAGRAM

The diagram detailing the steps taken is shown below. These steps are described in more detail in the prior section. Additional detail regarding input and output files is shown in Table 2.

Figure 1. Process Flow Diagram

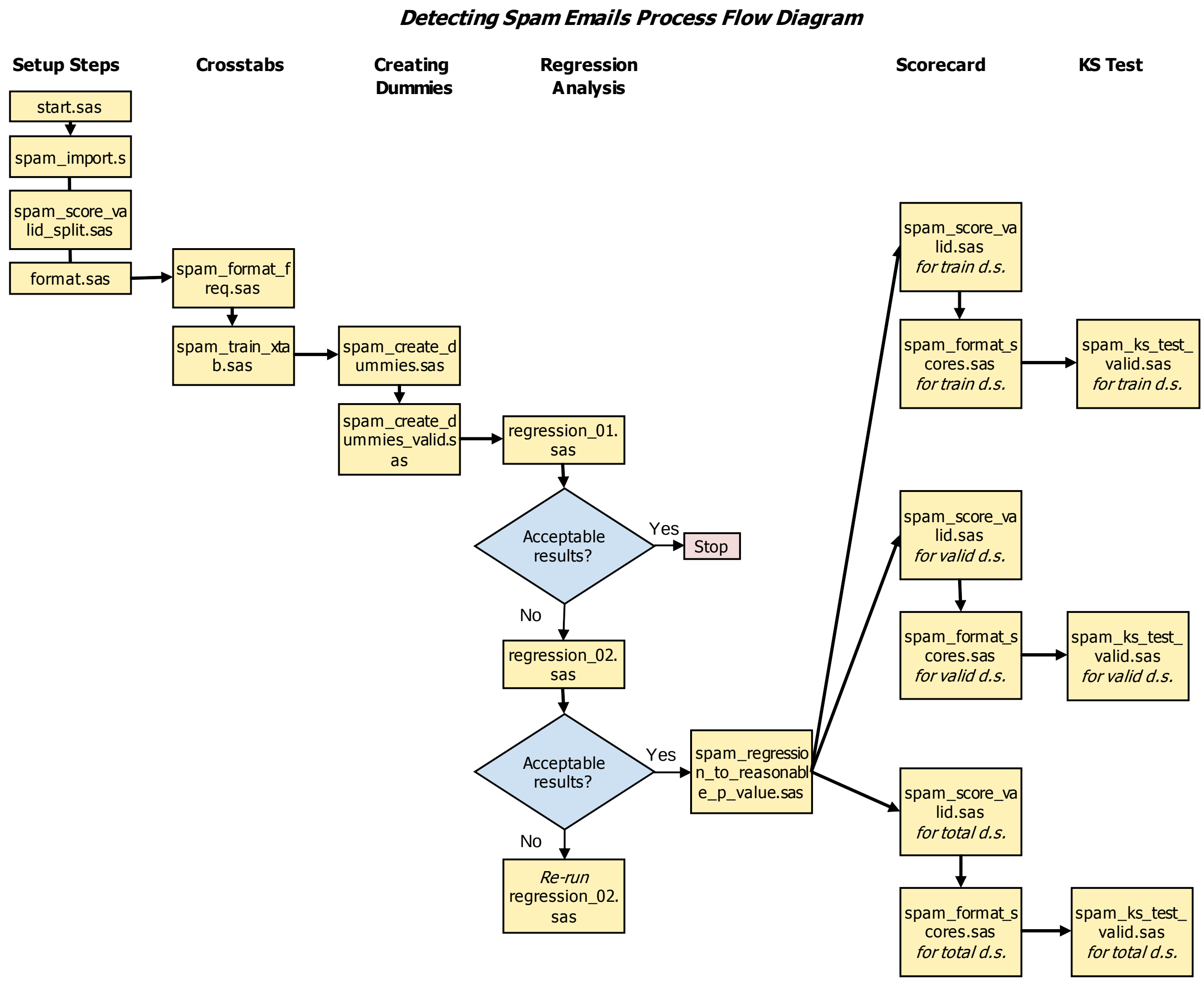


Table 2. Process Flow Chart

	Step	Description	Input Files	Output Files
1	start.sas	Assign SAS libraries	N/A	N/A
2	spam_import.sas	Import Spam dataset	spam_data.csv	spam.sas7bdat
3	spam_score_valid_split.sas	Splits dataset into training and validation	spam.sas7bdat	spam_train.sas7bdat spam_valid.sas7bdat
4	format.sas	Creates format library	N/A	N/A
	spam_freq.sas	Testing of one variable crosstab	spam_train.sas7bdat	
5	spam_format_freq.sas	Creates frequency table for each individual variable and applies format	spam_train.sas7bdat	spam_train_freq.html
6	spam_train_xtab.sas	Creates crosstabs and applies format	spam_train.sas7bdat	cross_tab_spam_train.html
7	spam_create_dummies.sas	Defines and creates the dummy category variables on training set	spam_train.sas7bdat	spam_train2.sas7bdat
8	spam_create_dummies_valid.sas	Defines and creates the dummy category variables on validation set	spam_valid.sas7bdat	spam_valid2.sas7bdat
9	regression_01.sas	Runs the regression for all dummy variables	spam_train2.sas7bdat	estfile.sas7bdat (temp d.s.) reg_01_all_dummies.html
10	regression_02.sas	Runs the regression for specific dummy variables	spam_train2.sas7bdat	estfile.sas7bdat (temp d.s.) reg_02_all_dummies.html reg_02.html
11	spam_regression_to_resonable _p_value.sas	Runs the regression for the final set of dummy variables	spam_train2.sas7bdat	estfile.sas7bdat (temp d.s.) reg_15_resonable_p_vals.html
12	spam_score_valid.sas	Scores the regression model , ran on training then validation	spam_train2.sas7bdat spam_valid2.sas7bdat	spam_train_scr.sas7bdat spam_valid_scr.sas7bdat
13	spam_ks_test_valid.sas	Creates a crosstab of the scores to the spam variable, ran on training then validation and then on the total dataset	spam_train_scr.sas7bdat spam_valid_scr.sas7bdat spam_total_scr.sas7bdat	spam_KS_train.html spam_KS_valid.html spam_KS_total.html
14	spam_format_scores.sas	Applies the format to the regression scorecard	spam_total_scr.sas7bdat	spam_total_scr.sas7bdat

RESULTS

Final Scorecard

Table 2 contains the final scorecard which shows results that one would expect. Words that positively impact (i.e. has an increased likelihood of being spam) in an increasing positive manner are: remove, internet, order, report, address, free, you, font, 000, and money, and 650. Words that negatively impact the model (i.e. predict actual emails) are: HP, HPL, George, Data, 85, 1999, meeting, project, RE, EDU, and conference. Most of the negative words are highly specific to George and his interests. The exclamation point was increasingly positive on its impact to the model. As the number of capital letters increased the trend went from negative to positive. In other words and email with a high number of capital letters is more likely to be spam whereas an expected number of capital letters is likely to be an actual email.

Table 3. Final Scorecard

Variable	Range	Points
Intercept		+327
% Remove	0%	0
	0.01-0.25%	+98
	0.26-0.50%	+176
	0.51-1.00%	+263
	1.01%+	+354
% Internet	0-.025%	0
	0.26-1.00%	+49
	1.01%+	+199
% Order	0%	0
	0.01-0.50%	+57
	0.51%+	+99
% Report	0%	0
	0.01%+	+52
% Addresses	0%	0
	0.01-0.50%	+72
	0.51%+	0
% Free	0%	0
	0.01-0.25%	+105
	0.26%+	+153
% You	0-2.00%	0
	2.01-4.50%	+37
	4.51%+	+68
% Font	0%	0
	0.01%+	+111
% "000"	0%	0
	0.01-0.50%	+82
	0.51%+	+172
% Money	0%	0
	0.01%+	+103
% HP	0%	0
	0.01-0.50%	-180
	0.51%+	-193
% HPL	0%	0
	0.01%+	-82
% George	0%	0
	0.01%+	-143
% 650	0%	0
	0.01-1.00%	+133
	1.00%+	+115
% Data	0%	0
	0.51%+	-94
% 85	0%	0
	0.01%+	-131
% 1999	0%	0
	0.01-0.50%	-130
	0.51%+	-67
% Meeting	0%	0
	0.01-1.50%	-88
	1.51%+	-135
% Project	0%	0
	0.01-0.50%	-78
	0.51%+	0
% RE	0%	0
	0.01-0.50%	-101
	0.51%+	-83
% Edu	0-0.25%	0
	0.26%+	-213
% Conference	0%	0
	0.01%+	-88
% Character Exclaimation Point	0-0.075%	0
	0.076-0.400%	+111
	0.401-0.600%	+175
	0.601%+	+269
Capital Letter Run Length	0-20	-236
	21-50	-111
	51-100	0
	101+	+103

In some instances we found combining the groups led to diminished predictive power of the model and the variables were separated again. For variables such as "RE," commonly used to denote replies, and 1999, the current year the data was collected, a value of 0% was neutral and a very low percentage meant that the email was not likely to be spam. However, if the percentage increased beyond a certain level the email was less likely to be non-spam. We believe this may occur because longer email messages containing these terms could indicate the occasional presence of spam, especially emails that overuse the current year.

Alternatively, short emails containing the words "addresses" could indicate spam, possible selling email address lists, while longer emails could be legitimate. Emails that sparingly refer to the word "project" are classified as legitimate non-spam emails as opposed to en email that lack the word or overuses the word. George's projects at his employer, HP, likely contribute to this occurrence. The variable "650" was quite surprising positive as that is the email recipient George's area code. We surmise that exceptionally short emails from local businesses are very likely to be spam, and longer emails from those same sources are less likely. It would be helpful in this instance to have access to the email corpus to validate these findings.

KS Test

The following is the KS test for the training data. The optimal point determined by the KS test was a score range of 400 to 449. This score would have 7.42% false positives and correctly filter 94.33% of the incoming spam emails.

Table 4. KS test for Training Data

The Kolmogorov-Smirnov (K-S) Test					TRAINING Data		
			Cumulative		Cumulative Percent		
Score Range	Not Spam	Spam	Not Spam	Spam	Not Spam	Spam	Difference
1000 OR MORE	2	240	2	240	0.14%	26.17%	0.260
950 TO 999	0	70	2	310	0.14%	33.81%	0.337
900 TO 949	0	64	2	374	0.14%	40.79%	0.406
850 TO 899	1	82	3	456	0.21%	49.73%	0.495
800 TO 849	0	54	3	510	0.21%	55.62%	0.554
750 TO 799	3	87	6	597	0.42%	65.10%	0.647
700 TO 749	2	50	8	647	0.57%	70.56%	0.700
650 TO 699	7	49	15	696	1.06%	75.90%	0.748
600 TO 649	10	52	25	748	1.77%	81.57%	0.798
550 TO 599	8	42	33	790	2.33%	86.15%	0.838
500 TO 549	23	26	56	816	3.96%	88.99%	0.850
450 TO 499	15	24	71	840	5.02%	91.60%	0.866
400 TO 449	**34**	**25**	**105**	**865**	**7.42%**	**94.33%**	**0.869**
350 TO 399	48	12	153	877	10.81%	95.64%	0.848
300 TO 349	62	8	215	885	15.19%	96.51%	0.813
250 TO 299	83	12	298	897	21.06%	97.82%	0.768
200 TO 249	110	5	408	902	28.83%	98.36%	0.695
150 TO 199	141	7	549	909	38.80%	99.13%	0.603
100 TO 149	112	1	661	910	46.71%	99.24%	0.525
50 TO 99	216	5	877	915	61.98%	99.78%	0.378
0 TO 49	122	1	999	916	70.60%	99.89%	0.293
NEGATIVE	416	1	1415	917	100.00%	100.00%	0.000

The KS test shown below is for the validation dataset. The optimal point determined by the KS test is a score between 400 to 449, which yields 8.16% false positives. This optimal point was within 4% of the training dataset which indicates that our model could be highly applicable to additional emails to George.

Table 5. KS test for Validation Data

The Kolmogorov-Smirnov (K-S) Test					VALIDATION Data		
			Cumulative		Cumulative Percent		
Score Range	Not Spam	Spam	Not Spam	Spam	Not Spam	Spam	Difference
1000 OR MORE	1	231	1	231	0.07%	25.78%	0.257
950 TO 999	1	63	2	294	0.15%	32.81%	0.327
900 TO 949	1	48	3	342	0.22%	38.17%	0.380
850 TO 899	3	81	6	423	0.44%	47.21%	0.468
800 TO 849	0	50	6	473	0.44%	52.79%	0.524
750 TO 799	2	67	8	540	0.58%	60.27%	0.597
700 TO 749	5	44	13	584	0.95%	65.18%	0.642
650 TO 699	7	47	20	631	1.46%	70.42%	0.690
600 TO 649	12	49	32	680	2.33%	75.89%	0.736
550 TO 599	6	32	38	712	2.77%	79.46%	0.767
500 TO 549	26	46	64	758	4.66%	84.60%	0.799
450 TO 499	21	17	85	775	6.19%	86.50%	0.803
400 TO 449	**27**	**40**	**112**	**815**	**8.16%**	**90.96%**	**0.828**
350 TO 399	44	25	156	840	11.36%	93.75%	0.824
300 TO 349	40	14	196	854	14.28%	95.31%	0.810
250 TO 299	81	14	277	868	20.17%	96.88%	0.767
200 TO 249	112	15	389	883	28.33%	98.55%	0.702
150 TO 199	110	5	499	888	36.34%	99.11%	0.628
100 TO 149	113	0	612	888	44.57%	99.11%	0.545
50 TO 99	217	6	829	894	60.38%	99.78%	0.394
0 TO 49	108	0	937	894	68.24%	99.78%	0.315
NEGATIVE	436	2	1373	896	100.00%	100.00%	0.000

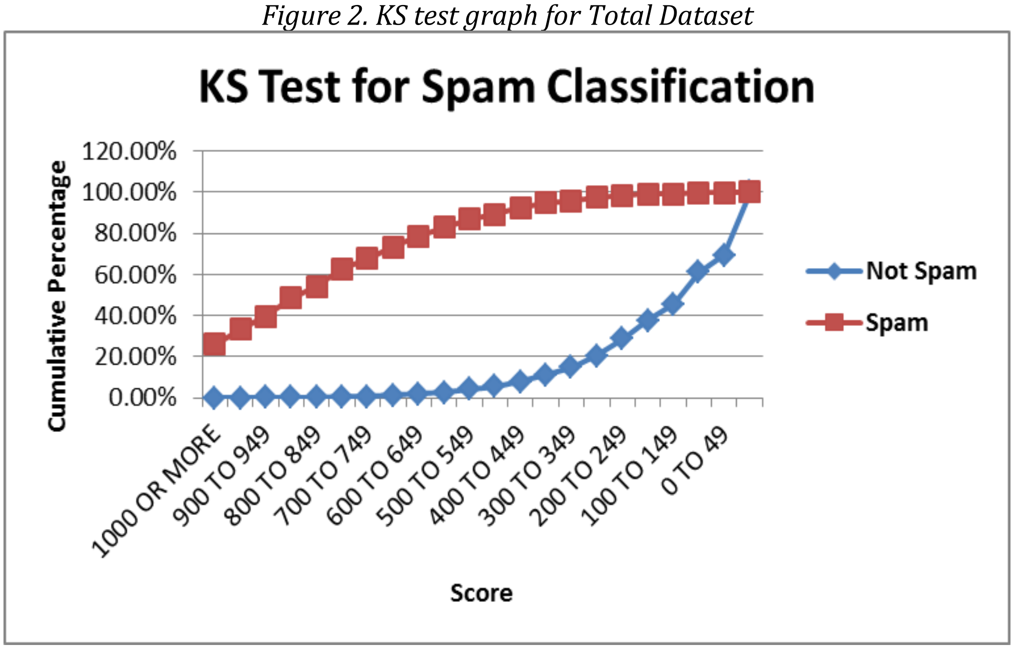
The optimal point calculated for the validation dataset maybe a point which is acceptable for other data but this would cause 8.16% of actual emails to be filtered as spam. Actual spam is correctly filtered out 90.96% of the time. Our customer may be dissatisfied with such a high percent of false positives. Our team recommends raising the cutoff to resolve the issue of a high false positive percentage. A score cutoff from 600 to 649 would greatly improve the false positive score while decreasing the correct filtering of spam.

The KS test was also calculated for the total dataset and results are shown below. Score cutoffs are similar to the validation set with the optimal cutoff from 400 to 449 and our recommended cutoff is slightly higher. Using a score of 600 as the cutoff would allow slightly over 2.5% of all legitimate emails to be classified as spam, which our team has selected as a reasonable false positive rate.

Table 6. KS test for Total Dataset

The Kolmogorov-Smirnov (K-S) Test			COMPLETE Data				
			Cumulative		Cumulative Percent		
Score Range	Not Spam	Spam	Not Spam	Spam	Not Spam	Spam	Difference
1000 OR MORE	3	471	3	471	0.11%	25.98%	25.87%
950 TO 999	1	133	4	604	0.14%	33.31%	33.17%
900 TO 949	1	112	5	716	0.18%	39.49%	39.31%
850 TO 899	4	163	9	879	0.32%	48.48%	48.16%
800 TO 849	0	104	9	983	0.32%	54.22%	53.90%
750 TO 799	5	154	14	1137	0.50%	62.71%	62.21%
700 TO 749	7	94	21	1231	0.75%	67.90%	67.15%
650 TO 699	14	96	35	1327	1.26%	73.19%	71.94%
600 TO 649	22	101	57	1428	2.04%	78.76%	76.72%
550 TO 599	14	74	71	1502	2.55%	82.85%	80.30%
500 TO 549	49	72	120	1574	4.30%	86.82%	82.51%
450 TO 499	36	41	156	1615	5.60%	89.08%	83.48%
400 TO 449	61	65	217	1680	7.78%	92.66%	84.88%
350 TO 399	92	37	309	1717	11.08%	94.70%	83.62%
300 TO 349	102	22	411	1739	14.74%	95.92%	81.18%
250 TO 299	164	26	575	1765	20.62%	97.35%	76.73%
200 TO 249	222	20	797	1785	28.59%	98.46%	69.87%
150 TO 199	251	12	1048	1797	37.59%	99.12%	61.53%
100 TO 149	225	1	1273	1798	45.66%	99.17%	53.51%
50 TO 99	433	11	1706	1809	61.19%	99.78%	38.59%
0 TO 49	230	1	1936	1810	69.44%	99.83%	30.39%
NEGATIVE	852	3	2788	1813	100%	100%	0.00%

Figure 2. KS test graph for Total Dataset



IMPLEMENTATION

In conclusion, we recommend scoring each incoming email using our model's score card. Any email with a score above 600 should be immediately routed to a spam folder. Only emails with a score under 600 should be routed to George's inbox. We believe is a score of 600 is a reasonable cutoff level that eliminates most spam, but keeps a low false positive rate. A high false positive rate would result in an unsatisfied customer due to the possible loss of actual emails. With a cutoff score of 600 our customer can expect to see about 83% of the spam correctly filtered to another spam mail folder and only 17% of the spam emails entering their actual inbox. Without using our model nearly 40% of the emails in the customer's inbox would be spam and with our model that percentage would drop to around 6.7%. Using a score of 600 will also only filter actual emails to the spam folder 2.5% of the time. Encouraging our customer to occasionally review his spam folder could prevent the loss of real emails.

A number of modifications during the implementation could increase the utility of the solution. In reality no one would want any of their actual email to be placed in the "Spam" folder. If 2.5% false positives are unacceptable the cutoff may be modified to decrease that number with the downside of letting more spam through. The system would also need to be modified to understand that the "1999" variable should reflect the current year and "650" should reflect the zip code. This would increase the durability of the model.

MONITORING REPORTS

Monitoring Report A

The performance of the model has to be monitored to ensure it remains effective. In the first report, the differences between the Expected Score Distribution, as predicted by our model, and the Actual Score Distribution, as observed in the future, are able to be monitored. Our model was built on past data, so there might be some changes to the characteristics of spam nowadays that could require adjustments in the model. Spammers continually look for ways of bypassing filters to reach their targets so word frequencies and use of capital letters could change. This model is also highly tuned to the personal characteristics of George and a change in his job, location, or other data could significantly change the results. Therefore we recommend at least a quarterly evaluation of the existing model. A significant number of misclassifications of spam or valid emails should also trigger the use of this report.

Table 7. Monitoring Report A - Actual vs. Expected Score Distribution

Score Range	Expected Score Distribution	Actual Score Distribution	Difference
>1000	10.32%		
>950	12.85%		
>900	15.80%		
>850	18.80%		
>800	21.52%		
>750	24.93%		
>700	27.19%		
>650	29.65%		
>600	32.34%		
>550	34.41%		
>500	36.82%		
>450	38.54%		
>400	41.19%		
>350	43.99%		
>300	46.84%		
>250	50.92%		
>200	55.94%		
>150	61.99%		
>100	66.83%		
>50	76.35%		
>0	81.31%		
> Low	100.00%		

Once the expected versus observed differences are calculated for each score range, then it should be determined if they are statistically significant. The minimum required difference at a 95% confidence level has to be determined. If all of the differences are below this number, then the fluctuations are among what is expected. If one or more of the differences are found to be significant, then a second monitoring report should be evaluated: Actual vs. Expected Characteristic Distribution.

Monitoring Report B

In this monitoring report, the differences between the observed and expected frequencies of each dummy variable are evaluated, to see what variable and which specific category are causing the large difference in the Score Distribution. With more detailed information regarding the variable that is affecting the performance of the model, the client and analyst can determine if there are changes in the characteristics defined for the variable that require modifications in the model. A new category, for example, might be required to better describe the current conditions of what is being modeled.

Table 8. Monitoring Report B - Actual vs. Expected Characteristic Distribution

Variable	Intervals	Points	Actual Frequency	Expected Frequency
% Remove	0%	0		81.8%
	0.01-0.25%	+98		5.7%
	0.26-0.50%	+176		4.2%
	0.51-1.00%	+263		4.6%
	1.01%+	+354		3.7%
% Internet	0-.025%	0		83.0%
	0.26-1.00%	+49		14.0%
	1.01%+	+199		3.0%
% Order	0%	0		83.2%
	0.01-0.50%	+57		9.3%
	0.51%+	+99		7.5%
% Report	0%	0		91.6%
	0.01%+	+52		8.4%
% Addresses	0%	0		92.8%
	0.01-0.50%	+72		3.9%
	0.51%+	0		3.3%
% Free	0%	0		72.9%
	0.01-0.25%	+105		5.9%
	0.26%+	+153		21.1%
% You	0-2.00%	0		63.2%
	2.01-4.50%	+37		30.0%
	4.51%+	+68		6.8%
% Font	0%	0		97.4%
	0.01%+	+111		2.6%
% "000"	0%	0		85.4%
	0.01-0.50%	+82		7.3%
	0.51%+	+172		7.2%
% Money	0%	0		84.5%
	0.01%+	+103		15.5%
% HP	0%	0		76.5%

	0.01-0.50%	-180		3.9%
	0.51%+	-193		19.5%
% HPL	0%	0		82.5%
	0.01%+	-82		17.5%
% George	0%	0		82.8%
	0.01%+	-143		17.2%
% 650	0%	0		90.0%
	0.01-1.00%	+133		5.9%
	1.00%+	+115		4.1%
% Data	0%	0		91.1%
	0.51%+	-94		8.9%
% 85	0%	0		89.6%
	0.01%+	-131		10.4%
% 1999	0%	0		83.2%
	0.01-0.50%	-130		8.1%
	0.51%+	-67		8.7%
% Meeting	0%	0		92.7%
	0.01-1.50%	-88		4.2%
	1.51%+	-135		3.1%
% Project	0%	0		92.2%
	0.01-0.50%	-78		4.3%
	0.51%+	0		3.5%
% RE	0%	0		72.9%
	0.01-0.50%	-101		11.5%
	0.51%+	-83		15.6%
% Edu	0-0.25%	0		91.1%
	0.26%+	-213		8.9%
% Conference	0%	0		95.9%
	0.01%+	-88		4.1%
% Character Exclamation Point	0-0.075%	0		57.5%
	0.076-0.400%	+111		21.0%
	0.401-0.600%	+175		7.8%
	0.601%+	+269		13.7%
Capital Letter Run Length	0-20	-236		59.0%
	21-50	-111		19.3%
	51-100	0		16.4%
	101+	+103		5.3%

Monitoring Report C

As spammers adapt and the subject matter of George's emails change, new words should be considered for inclusion in the variable list. Each existing word variable should be considered in the context of overall occurrence across all emails. The report below can be used to track the existing occurrences. New words that exceed a particular threshold, such as 0.04% should be considered in future redevelopments of the model. Words that no longer occur with regularity may be removed from consideration. This report should be considered during model redevelopment and does not need to be run at a regular frequency.

Table 9. Monitoring Report C – New Word Frequency

Word	Existing Percentage	New Percentage
make	0.21%	
address	0.28%	
all	0.07%	
3d	0.31%	
our	0.10%	
over	0.11%	
remove	0.11%	
internet	0.09%	
order	0.24%	
mail	0.06%	
receive	0.54%	
will	0.09%	
people	0.06%	
report	0.05%	
addresses	0.25%	
free	0.14%	
business	0.19%	
email	1.66%	
you	0.09%	
credit	0.81%	
money	0.55%	
hp	0.27%	
hpl	0.77%	
george	0.13%	
conference	0.04%	

APPENDIX A

Data Dictionary

Variable	***Description***
word_freq_make word_freq_address word_freq_all word_freq_3d word_freq_our word_freq_over word_freq_remove word_freq_internet word_freq_order word_freq_mail word_freq_receive word_freq_will word_freq_people word_freq_report word_freq_addresses word_freq_free word_freq_business word_freq_email word_freq_you word_freq_credit word_freq_your word_freq_font word_freq_000 word_freq_money word_freq_hp word_freq_hpl word_freq_george word_freq_650 word_freq_lab word_freq_labs word_freq_telnet word_freq_857 word_freq_data word_freq_415 word_freq_85 word_freq_technology word_freq_1999 word_freq_parts word_freq_pm word_freq_direct word_freq_cs word_freq_meeting word_freq_original word_freq_project word_freq_re word_freq_edu word_freq_table word_freq_conference	48 continuous real attributes of type word_freq_WORD = percentage of words in the e-mail that match WORD, i.e. (100 * (number of times the WORD appears in the e-mail) / 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.
char_freq_semicolon char_freq_open_paren char_freq_open_bracket char_freq_excl_point char_freq_dollar_sign char_freq_hash	6 continuous real attributes of type char_freq_CHAR] = percentage of characters in the e-mail that match CHAR, i.e. (100 * (number of CHAR occurences) / total characters in e-mail)
capital_run_length_average	1 continuous real attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters
capital_run_length_longest	1 continuous integer attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters
capital_run_length_total	1 continuous integer attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail
spam	1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

APPENDIX B

REGRESSION RESULTS - 1ST ITERATION

The REG Procedure
Model: bgscore
Dependent Variable: spam

Number of Observations Read	2332
Number of Observations Used	2332

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	137	418.52839	3.05495	48.61	<.0001
Error	2194	137.88456	0.06285		
Corrected Total	2331	556.41295			

Root MSE	0.25069	**R-Square**	0.7522
Dependent Mean	0.39322	**Adj R-Sq**	0.7367
Coeff Var	63.75272		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.45993	0.06862	6.70	<.0001
word_freq_make1	**1**	0.06958	0.03695	1.88	0.0598
word_freq_make2	**1**	0.03072	0.05073	0.61	0.5448
word_freq_make3	**1**	0.03259	0.03998	0.82	0.4150

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_address1	**1**	-0.03004	0.02340	-1.28	0.1993
word_freq_address2	**1**	0.00097101	0.02384	0.04	0.9675
word_freq_all1	**1**	0.01347	0.01721	0.78	0.4339
word_freq_all2	**1**	-0.01098	0.02262	-0.49	0.6273
word_freq_all3	**1**	-0.01235	0.02327	-0.53	0.5958
word_freq_3d1	**1**	0.13179	0.05289	2.49	0.0128
word_freq_our1	**1**	-0.05556	0.02665	-2.08	0.0372
word_freq_our2	**1**	0.01367	0.02721	0.50	0.6154
word_freq_our3	**1**	0.04281	0.03528	1.21	0.2251
word_freq_our4	**1**	0.04977	0.03508	1.42	0.1562
word_freq_over1	**1**	-0.03416	0.02687	-1.27	0.2037
word_freq_over2	**1**	0.01118	0.02642	0.42	0.6723
word_freq_over3	**1**	0.03554	0.02452	1.45	0.1474
word_freq_remove1	**1**	0.09854	0.02940	3.35	0.0008
word_freq_remove2	**1**	0.14265	0.03123	4.57	<.0001
word_freq_remove3	**1**	0.18413	0.02855	6.45	<.0001
word_freq_remove4	**1**	0.30219	0.03112	9.71	<.0001
word_freq_internet1	**1**	0.03090	0.02942	1.05	0.2937
word_freq_internet2	**1**	0.05862	0.02312	2.54	0.0113
word_freq_internet3	**1**	0.15787	0.03351	4.71	<.0001
word_freq_order1	**1**	0.04626	0.02356	1.96	0.0497

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_order2	**1**	0.05980	0.02600	2.30	0.0215
word_freq_mail1	**1**	-0.02938	0.02879	-1.02	0.3077
word_freq_mail2	**1**	0.00501	0.01900	0.26	0.7922
word_freq_mail3	**1**	-0.03433	0.02284	-1.50	0.1330
word_freq_receive1	**1**	-0.00705	0.03042	-0.23	0.8167
word_freq_receive2	**1**	0.00499	0.02281	0.22	0.8268
word_freq_will1	**1**	-0.02463	0.01677	-1.47	0.1422
word_freq_will2	**1**	0.00601	0.02166	0.28	0.7815
word_freq_will3	**1**	0.02871	0.02255	1.27	0.2031
word_freq_people1	**1**	-0.05025	0.02833	-1.77	0.0763
word_freq_people2	**1**	-0.04703	0.02973	-1.58	0.1138
word_freq_people3	**1**	-0.06524	0.02349	-2.78	0.0055
word_freq_report1	**1**	0.03671	0.02357	1.56	0.1195
word_freq_addresses1	**1**	0.06855	0.03305	2.07	0.0382
word_freq_addresses2	**1**	-0.02795	0.03874	-0.72	0.4707
word_freq_free1	**1**	0.11912	0.02891	4.12	<.0001
word_freq_free2	**1**	0.13327	0.01638	8.14	<.0001
word_freq_business1	**1**	0.03253	0.02686	1.21	0.2260
word_freq_business2	**1**	-0.00080086	0.02631	-0.03	0.9757
word_freq_business3	**1**	0.02886	0.03242	0.89	0.3735
word_freq_email1	**1**	-0.02796	0.02152	-1.30	0.1940

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_email2	**1**	0.02553	0.01916	1.33	0.1829
word_freq_you1	**1**	0.01126	0.01779	0.63	0.5268
word_freq_you2	**1**	-0.00313	0.02247	-0.14	0.8892
word_freq_you3	**1**	0.04165	0.01750	2.38	0.0174
word_freq_you4	**1**	0.08976	0.02563	3.50	0.0005
word_freq_credit1	**1**	0.02262	0.03883	0.58	0.5601
word_freq_credit2	**1**	-0.02379	0.02835	-0.84	0.4015
word_freq_your1	**1**	-0.01509	0.02770	-0.54	0.5860
word_freq_your2	**1**	0.02388	0.02994	0.80	0.4252
word_freq_your3	**1**	0.03187	0.02804	1.14	0.2559
word_freq_font1	**1**	0.11532	0.04078	2.83	0.0047
word_freq_0001	**1**	0.04883	0.02665	1.83	0.0671
word_freq_0002	**1**	0.11351	0.02734	4.15	<.0001
word_freq_money1	**1**	0.06076	0.02133	2.85	0.0044
word_freq_hp1	**1**	-0.18162	0.03299	-5.51	<.0001
word_freq_hp2	**1**	-0.20560	0.02313	-8.89	<.0001
word_freq_hpl1	**1**	-0.06974	0.02378	-2.93	0.0034
word_freq_george1	**1**	-0.14263	0.01825	-7.81	<.0001
word_freq_6501	**1**	0.13677	0.03478	3.93	<.0001
word_freq_6502	**1**	0.08101	0.03877	2.09	0.0368
word_freq_lab1	**1**	-0.01318	0.02786	-0.47	0.6362

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_labs1	**1**	0.02915	0.02724	1.07	0.2846
word_freq_telnet1	**1**	-0.03428	0.03800	-0.90	0.3671
word_freq_8571	**1**	-0.01064	0.09907	-0.11	0.9145
word_freq_data1	**1**	0.00126	0.03140	0.04	0.9681
word_freq_data2	**1**	-0.07545	0.02686	-2.81	0.0050
word_freq_4151	**1**	0.12022	0.08940	1.34	0.1789
word_freq_851	**1**	-0.09235	0.03700	-2.50	0.0126
word_freq_852	**1**	-0.15104	0.03503	-4.31	<.0001
word_freq_technology1	**1**	0.00498	0.02820	0.18	0.8597
word_freq_technology2	**1**	0.06575	0.02886	2.28	0.0228
word_freq_19991	**1**	-0.12167	0.02454	-4.96	<.0001
word_freq_19992	**1**	-0.05884	0.02359	-2.49	0.0127
word_freq_parts1	**1**	0.08044	0.03958	2.03	0.0423
word_freq_pm1	**1**	-0.04570	0.03306	-1.38	0.1669
word_freq_pm2	**1**	-0.07902	0.02704	-2.92	0.0035
word_freq_direct1	**1**	-0.00830	0.03280	-0.25	0.8003
word_freq_direct2	**1**	-0.00148	0.03262	-0.05	0.9638
word_freq_cs1	**1**	-0.00781	0.03525	-0.22	0.8246
word_freq_meeting1	**1**	-0.07868	0.02867	-2.74	0.0061
word_freq_meeting2	**1**	-0.13201	0.03200	-4.13	<.0001
word_freq_original1	**1**	0.05375	0.03835	1.40	0.1612

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_project1	**1**	-0.07343	0.03342	-2.20	0.0281
word_freq_re1	**1**	-0.06364	0.03121	-2.04	0.0415
word_freq_re2	**1**	-0.07935	0.02447	-3.24	0.0012
word_freq_re3	**1**	-0.08711	0.01651	-5.28	<.0001
word_freq_edu1	**1**	-0.18938	0.02162	-8.76	<.0001
word_freq_table1	**1**	-0.01625	0.04991	-0.33	0.7448
word_freq_conference1	**1**	-0.06852	0.02995	-2.29	0.0223
char_freq_semicolon1	**1**	0.01467	0.02866	0.51	0.6087
char_freq_semicolon2	**1**	0.00215	0.01877	0.11	0.9088
char_freq_open_paren1	**1**	0.00731	0.01988	0.37	0.7132
char_freq_open_paren2	**1**	-0.04563	0.02698	-1.69	0.0910
char_freq_open_paren3	**1**	-0.01001	0.01645	-0.61	0.5430
char_freq_open_paren4	**1**	-0.00657	0.01906	-0.34	0.7303
char_freq_open_bracket1	**1**	-0.00405	0.03357	-0.12	0.9041
char_freq_open_bracket2	**1**	-0.05012	0.03837	-1.31	0.1916
char_freq_open_bracket3	**1**	-0.05342	0.03445	-1.55	0.1212
char_freq_excl_point1	**1**	0.02144	0.02612	0.82	0.4118
char_freq_excl_point2	**1**	0.12276	0.02928	4.19	<.0001
char_freq_excl_point3	**1**	0.12562	0.02999	4.19	<.0001
char_freq_excl_point4	**1**	0.17578	0.03203	5.49	<.0001
char_freq_excl_point5	**1**	0.25308	0.03019	8.38	<.0001

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
char_freq_dollar_sign1	**1**	-0.11423	0.03100	-3.68	0.0002
char_freq_dollar_sign2	**1**	0.00740	0.03644	0.20	0.8391
char_freq_dollar_sign3	**1**	0.00923	0.03920	0.24	0.8139
char_freq_dollar_sign4	**1**	0.07164	0.03746	1.91	0.0560
char_freq_dollar_sign5	**1**	0.02444	0.04192	0.58	0.5600
char_freq_dollar_sign6	**1**	0.05899	0.04088	1.44	0.1491
char_freq_hash1	**1**	-0.00824	0.02959	-0.28	0.7806
char_freq_hash2	**1**	-0.06104	0.03142	-1.94	0.0522
char_freq_hash3	**1**	-0.05145	0.02488	-2.07	0.0388
capital_run_length_average1	**1**	-0.02246	0.03607	-0.62	0.5334
capital_run_length_average2	**1**	-0.02415	0.03389	-0.71	0.4761
capital_run_length_average3	**1**	-0.05757	0.02287	-2.52	0.0119
capital_run_length_average4	**1**	-0.06629	0.02083	-3.18	0.0015
capital_run_length_average5	**1**	0.01127	0.02219	0.51	0.6116
capital_run_length_average6	**1**	0.03544	0.03129	1.13	0.2575
capital_run_length_longest1	**1**	-0.16816	0.02850	-5.90	<.0001
capital_run_length_longest2	**1**	-0.12740	0.02584	-4.93	<.0001
capital_run_length_longest3	**1**	-0.02949	0.02269	-1.30	0.1938
capital_run_length_longest4	**1**	-0.05572	0.02376	-2.35	0.0191
capital_run_length_longest5	**1**	-0.09493	0.02868	-3.31	0.0009
capital_run_length_longest6	**1**	-0.13490	0.03973	-3.40	0.0007

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
capital_run_length_total1	**1**	-0.11081	0.03901	-2.84	0.0045
capital_run_length_total2	**1**	-0.10470	0.03399	-3.08	0.0021
capital_run_length_total3	**1**	-0.03447	0.02473	-1.39	0.1635
capital_run_length_total4	**1**	0.00753	0.02557	0.29	0.7684
capital_run_length_total5	**1**	0.06834	0.02440	2.80	0.0051
capital_run_length_total6	**1**	0.05512	0.02527	2.18	0.0292
capital_run_length_total7	**1**	0.10727	0.04347	2.47	0.0137
capital_run_length_total8	**1**	0.10404	0.04587	2.27	0.0234

REGRESSION RESULTS - FINAL ITERATION

The REG Procedure
Model: bgscore
Dependent Variable: spam

Number of Observations Read	2332
Number of Observations Used	2332

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	42	397.96765	9.47542	136.89	<.0001
Error	2289	158.44530	0.06922		
Corrected Total	2331	556.41295			

Root MSE	0.26310	**R-Square**	0.7152
Dependent Mean	0.39322	**Adj R-Sq**	0.7100
Coeff Var	66.90768		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.32506	0.01758	18.49	<.0001
word_freq_remove1	**1**	0.09765	0.02812	3.47	0.0005
word_freq_remove2	**1**	0.17614	0.02981	5.91	<.0001
word_freq_remove3	**1**	0.26323	0.02777	9.48	<.0001
word_freq_remove4	**1**	0.35425	0.03030	11.69	<.0001

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_internet2	**1**	0.04926	0.02146	2.30	0.0218
word_freq_internet3	**1**	0.19916	0.03309	6.02	<.0001
word_freq_order1	**1**	0.05694	0.02258	2.52	0.0118
word_freq_order2	**1**	0.09859	0.02330	4.23	<.0001
word_freq_report1	**1**	0.05204	0.02221	2.34	0.0192
word_freq_addresses1	**1**	0.07159	0.03084	2.32	0.0204
word_freq_free1	**1**	0.10534	0.02659	3.96	<.0001
word_freq_free2	**1**	0.15358	0.01610	9.54	<.0001
word_freq_you3	**1**	0.03730	0.01352	2.76	0.0059
word_freq_you4	**1**	0.06817	0.02314	2.95	0.0032
word_freq_font1	**1**	0.11083	0.03571	3.10	0.0019
word_freq_0001	**1**	0.08206	0.02476	3.31	0.0009
word_freq_0002	**1**	0.17231	0.02347	7.34	<.0001
word_freq_money1	**1**	0.10299	0.01954	5.27	<.0001
word_freq_hp1	**1**	-0.18020	0.03267	-5.52	<.0001
word_freq_hp2	**1**	-0.19346	0.02243	-8.62	<.0001
word_freq_hpl1	**1**	-0.08160	0.02330	-3.50	0.0005
word_freq_george1	**1**	-0.14330	0.01785	-8.03	<.0001
word_freq_6501	**1**	0.13324	0.03380	3.94	<.0001
word_freq_6502	**1**	0.11454	0.03570	3.21	0.0014
word_freq_data2	**1**	-0.09373	0.02641	-3.55	0.0004

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
word_freq_4151	**1**	0.12537	0.03500	3.58	0.0003
word_freq_8512	**1**	-0.13146	0.02786	-4.72	<.0001
word_freq_19991	**1**	-0.13035	0.02326	-5.60	<.0001
word_freq_19992	**1**	-0.06709	0.02164	-3.10	0.0020
word_freq_meeting1	**1**	-0.08845	0.02862	-3.09	0.0020
word_freq_meeting2	**1**	-0.13472	0.03229	-4.17	<.0001
word_freq_project1	**1**	-0.07838	0.03280	-2.39	0.0169
word_freq_re12	**1**	-0.10096	0.01929	-5.24	<.0001
word_freq_re3	**1**	-0.08272	0.01633	-5.07	<.0001
word_freq_edu1	**1**	-0.21338	0.02083	-10.24	<.0001
word_freq_conference1	**1**	-0.08769	0.02975	-2.95	0.0032
char_freq_excl_point23	**1**	0.11125	0.01581	7.04	<.0001
char_freq_excl_point4	**1**	0.17497	0.02303	7.60	<.0001
char_freq_excl_point5	**1**	0.26893	0.01899	14.16	<.0001
capital_run_length_total12	**1**	-0.23643	0.02070	-11.42	<.0001
capital_run_length_total3	**1**	-0.11064	0.01931	-5.73	<.0001
capital_run_length_total5678	**1**	0.10276	0.01707	6.02	<.0001