Machine Learning Classifier for Detecting Email Spams

Springboard Introduction to Data Science Capstone Project

By

Savita Shrivastava

26 Dec 2017

Introduction

- ‘Spamming’
  - The action of sending unwanted messages in bulk quantity without obtaining explicit permission of the recipient.

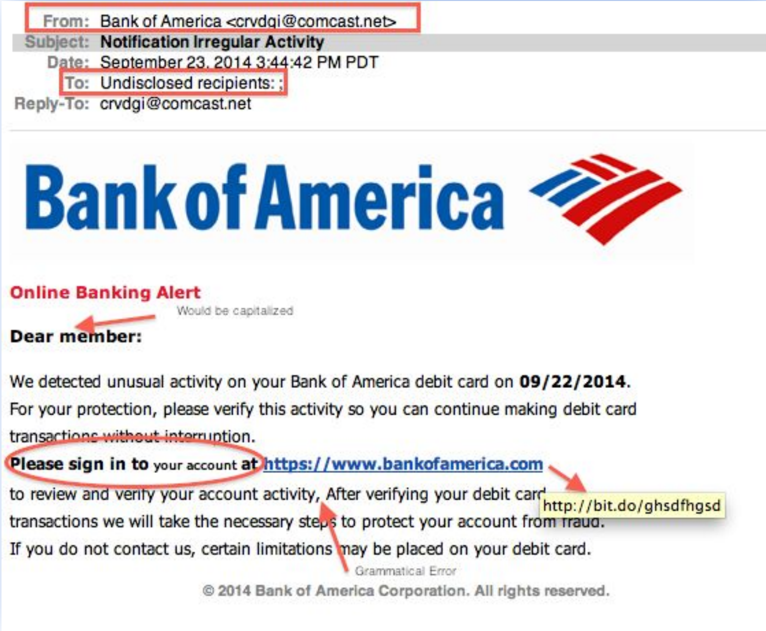
- Examples of Spam:
  - Email spam
  - Instant messaging spam
  - Usenet newsgroup spamt etc.

Email Spam

- Email spam refers to sending irrelevant, inappropriate and unrequested email message to several people.
- Mostly email spam are commercial in nature
- The purpose of email spam is advertising, promotion, spreading viruses, phishing or baking fraud.

The Problem

- Around 80% emails are spam.



From: Bank of America <crvdgi@comcast.net>
Subject: **Notification Irregular Activity**
Date: September 23, 2014 3:44:42 PM PDT
To: Undisclosed recipients: ;
Reply-To: crvdgi@comcast.net

Bank of America

Online Banking Alert

Would be capitalized

Dear member:

We detected unusual activity on your Bank of America debit card on **09/22/2014**. For your protection, please verify this activity so you can continue making debit card transactions without interruption.

Please sign in to your account **at** https://www.bankofamerica.com to review and verify your account activity, After verifying your debit card transactions we will take the necessary steps to protect your account from fraud.

http://bit.do/ghsdfhgsd

If you do not contact us, certain limitations may be placed on your debit card.

Grammatical Error

© 2014 Bank of America Corporation. All rights reserved.

Types of Spam Content

- Product advertisement
- Financial
- Adult
- Internet
- Pharmacy
- Health
- Scams
- Leisure
- Fraud
- Political etc.

Spam Filter

- A spam filter is a software that keep spam emails from entering the inbox
- It predicts if an email is spam or no-spam an
- Statistical and machine learning based classification system
    - Uses relevant features for classification

My Objective

- Build a model using a machine learning method which can predict the outcome if an email is spam or no-spam and based on that the spam emails can be filtered out.
- The project will try to give answet to following question
  - How can we construct a spam filter, given the data set
  - What factors alter the probability of an email being a spam-email
  - How to create an accurate model that can predict if an email is spam
  - What is the risk of model making false prediction

Data Acquisition

- The Spambase data set was acquired from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/spambase)

Data Exploration

- Number of Instances: 4601
- Number of attributes: 58
- Number of missing data points: None
- The last column of 'spambase.data' named ‘spam’ denotes whether the email was considered spam (1) or not spam (0).

Attributes

Attribute Column Number	Attribute name	Attribute Type	Attribute Description
1 to 48	word_freq_WORD	continuous real [0,100]	percentage of words in the email that match WORD
49 to 54	char_freq_CHAR	continuous real [0,100]	percentage of characters in the email that match CHAR
55	capital_run_length_average	continuous real [0,100]	average length of uninterrupted sequences of capital letters
56	capital_run_length_longest	continuous integer [1,...]	length of longest uninterrupted sequence of capital letters
57	capital_run_length_total	continuous integer [1,...]	total number of capital letters in the email
58	spam	nominal {0,1}	denotes whether the email was considered spam (1) or not (0)

Outcome or dependent variable will be ‘spam’ and all other attributes from column 1 to 57 will be independent variables. Below is the summary of spam variable

Outcome variable ‘spam’

- Summary of outcome variable ‘spam’

Spam	Frequency	Percent
0 (not spam)	2788	60%
1 (spam)	1813	39%

Data Analysis Method - Logistic Regression

- Logistic regression is a simple classification algorithm to analyze a dataset in which there are one or more independent variables that determine an outcome. In logistic regression the outcome or dependent variable is coded a 1 (TRUE) or 0 (FALSE).
- The goal of logistic regression is to find the best fitting model to describe the relationship between the dependent variable (response or outcome variable) and a set of independent (predictor or explanatory) variables.

Logistic Regression - *Logit Transformation*

- Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of an email being spam:

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

- where p is the probability of presence of characteristic of interest (an email being spam). The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

- *Logit transformation* of probability p is:

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Data Cleaning

- Change the name of the below attributes which have special characters in their name as below:
    - char_freq_; to char_freq_semic
    - char_freq_(to char_freq_openp
    - char_freq_[to char_freq_openb
    - char_freq_! to char_freq_excl
    - char_freq_$ to char_freq_dollar
    - char_freq_# to char_freq_pound

Building the Predictive Model

- Logistic regression was used to build the predictive model using all of the independent variables (attributes 1 to 57)
- Data was divided into a training and testing set with 75/25 ratio

Performance Evaluation Parameters

- True Positive (TP): Spam emails are correctly predicted as spams
- True Negatives (TN) : No-spam emails are correctly predicted as no-spam emails
- False Positive (FP) : No-spam emails are incorrectly predicted as spam emails
- False Negative (FN) : Spam emails are incorrectly predicted as no-spam emails
- Accuracy : (True Negatives (TN) + True Positive (TP)) / Total number of observations
- Sensitivity (True Positive Rate) = True Positive (TP) / (True Positive (TP) + False Negative (FN))
- Specificity (False Positive Rate) = True Negatives (TN) / True Negatives (TN) + False Positive (FP))
- Error rate = (False Positive (FP) + False Negative (FN)) / Total number of observations

	Predicted = 0	**Predicted = 1**
Actual = 0	True Negatives (TN)	False Positive (FP)
Actual = 1	False Negative (FN)	True Positive (TP)

Coefficient Output of Training Model

- Showed that the independent variables which are significantly affecting the model and outcome variable
- The negative estimate value of some of the significant variables such as ‘word_freq_george’, ‘word_freq_hp’, ‘word_freq_hpl’ and ‘word_freq_edu’ clearly showing that these are no-spam email related words
- ‘word_freq_free’, ‘word_freq_000’, ‘word_freq_our’, ‘word_freq_remove’, ‘word_freq_re’ and ‘char_freq_dollar’ have high positive value indicating that these are spam emails.

Coefficients

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.006e+00  2.112e-01  -9.499  < 2e-16 ***
word_freq_make           -4.501e-01  2.887e-01  -1.559 0.119051
word_freq_address        -1.516e-01  8.264e-02  -1.834 0.066600 .
word_freq_all             6.749e-02  1.459e-01   0.462 0.643753
word_freq_3d              3.616e+00  2.249e+00   1.607 0.107957
word_freq_our             8.113e-01  1.360e-01   5.968 2.41e-09 ***
word_freq_over            1.329e+00  3.193e-01   4.164 3.13e-05 ***
word_freq_remove          2.998e+00  5.080e-01   5.902 3.59e-09 ***
word_freq_internet        4.624e-01  1.574e-01   2.939 0.003295 **
word_freq_order           3.857e-01  3.787e-01   1.018 0.308453
word_freq_mail           -2.065e-02  1.241e-01  -0.166 0.867867
 word_freq_receive        -2.970e-01  3.590e-01  -0.827 0.408124
word_freq_will           -1.279e-01  9.024e-02  -1.418 0.156268
```

Coefficients contd...

word_freq_people	-5.160e-02	3.063e-01	-0.168	0.866241	
word_freq_report	5.376e-02	1.470e-01	0.366	0.714624	
word_freq_addresses	8.097e-01	7.600e-01	1.065	0.286692	
word_freq_free	8.309e-01	1.753e-01	4.741	2.13e-06	***
word_freq_business	1.083e+00	2.839e-01	3.815	0.000136	***
word_freq_email	1.062e-01	1.606e-01	0.662	0.508269	
word_freq_you	6.335e-02	4.493e-02	1.410	0.158555	
word_freq_credit	1.335e+00	9.340e-01	1.429	0.153040	
word_freq_your	1.575e-01	6.644e-02	2.370	0.017791	*
word_freq_font	-5.869e-02	2.086e-01	-0.281	0.778427	
word_freq_000	3.237e+00	7.330e-01	4.417	1.00e-05	***
word_freq_money	3.038e-01	1.343e-01	2.263	0.023664	*
word_freq_hp	-2.046e+00	3.884e-01	-5.269	1.37e-07	***
word_freq_hpl	-1.004e+00	5.307e-01	-1.893	0.058418	.
word_freq_george	-1.850e+01	3.980e+00	-4.647	3.37e-06	***
word_freq_650	6.413e-01	2.932e-01	2.187	0.028737	*

Coefficients contd...

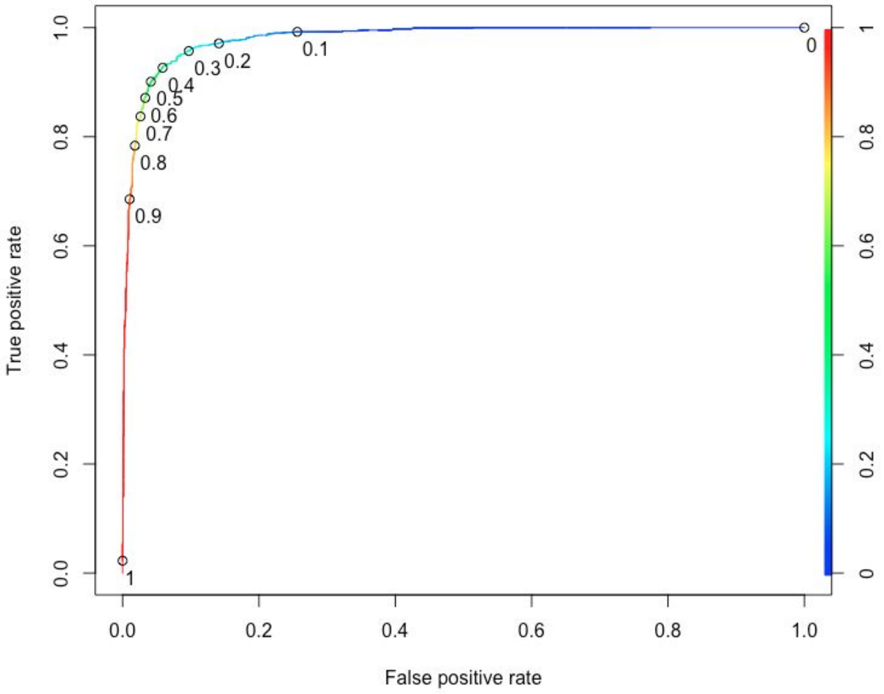
```
word_freq_lab               -2.350e+00  1.742e+00  -1.349 0.177233
word_freq_labs              -4.709e-01  5.470e-01  -0.861 0.389270
word_freq_telnet            -3.886e+00  3.131e+00  -1.241 0.214524
word_freq_857                1.940e+00  3.710e+00   0.523 0.601048
word_freq_data              -6.652e-01  3.640e-01  -1.827 0.067645 .
word_freq_415                1.641e-01  1.708e+00   0.096 0.923455
word_freq_85                -2.705e+00  1.022e+00  -2.647 0.008110 **
word_freq_technology         7.257e-01  3.831e-01   1.894 0.058172 .
word_freq_1999              -3.035e-01  2.522e-01  -1.203 0.228963
word_freq_parts             -5.981e-01  5.995e-01  -0.998 0.318437
word_freq_pm                -7.256e-01  5.117e-01  -1.418 0.156183
word_freq_direct            -3.735e-01  4.439e-01  -0.842 0.400063
word_freq_cs                -5.676e+02  2.071e+04  -0.027 0.978141
word_freq_meeting           -2.270e+00  8.923e-01  -2.544 0.010973 *
word_freq_original          -1.474e+00  1.173e+00  -1.257 0.208771
word_freq_project           -1.641e+00  7.555e-01  -2.172 0.029857 *
word_freq_re                -6.978e-01  1.627e-01  -4.290 1.79e-05 ***
```

Coefficients contd...

```
word_freq_edu              -1.409e+00  2.963e-01  -4.756 1.98e-06 ***
word_freq_table            -4.017e+00  2.575e+00  -1.560 0.118749
word_freq_conference       -6.014e+00  3.249e+00  -1.851 0.064132 .
char_freq_semic            -1.085e+00  5.476e-01  -1.981 0.047570 *
char_freq_openp            -3.150e-01  3.654e-01  -0.862 0.388588
char_freq_openb            -2.397e+00  1.725e+00  -1.390 0.164645
char_freq_excl              7.186e-01  1.657e-01   4.337 1.44e-05 ***
char_freq_dollar            5.446e+00  9.474e-01   5.748 9.01e-09 ***
char_freq_pound             2.187e+00  1.645e+00   1.330 0.183522
capital_run_length_average  2.839e-01  7.102e-02   3.997 6.42e-05 ***
capital_run_length_longest -1.043e-03  3.413e-03  -0.306 0.759940
capital_run_length_total    1.515e-03  3.196e-04   4.739 2.15e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 4329.5  on 3220  degrees of freedom
Residual deviance: 1159.7  on 3163  degrees of freedom
AIC: 1275.7
Number of Fisher Scoring iterations: 23
```

Selecting a Threshold for Filter Using ROC Curve

- Threshold cutoff 0.9999 was selected.
- With threshold cutoff 0.9999, specificity was higher and sensitivity is lower



Classification Matrix of Training Model using Threshold 0.99

	Predicted = 0	**Predicted = 1**
Actual = 0	1933	7
Actual = 1	697	584

Accuracy, Sensitivity and Specificity of Training Model using Threshold 0.99

- Accuracy : 78.14%
- Sensitivity : 45.6%
- Specificity : 99.6%

Area Under the ROC (AUROC)

- The metric for AUROC ranges from 0.50 to 1.00.
- Value above 0.80 indicate that the model does a good job in discriminating spam and no-spam emails.
- The AUROC for this training model is 0.98 which is above then 0.80.

Evaluating the Model on Testing Data Set

- Classification Matrix using Testing Data Set to Evaluate the Model

	Predicted = 0	**Predicted = 1**
Actual = 0	803	48
Actual = 1	45	484

Accuracy, Sensitivity and Specificity using Testing Data Set to Evaluate the Model

- Accuracy : 93.26%
- Sensitivity : 91.5%
- Specificity : 94.35%
- Overall error rate : 6.7%

The model exhibit high sensitivity as well as high specificity for predicting spam and no-spam emails and exhibit that this model is a better predictor than the baseline method where accuracy was 60.8%

Conclusion

- Email spam classification has received a tremendous attention by majority of the people as it helps to identify the unwanted information and threats
- The results demonstrate that logistic regression model has 93.26 % accuracy in spam detection than baseline methods 60.8% accuracy.
- The careful selection of attributes in building model can increase the accuracy of the model.