## Springboard Introduction to Data Science Capstone Project Data Story

For Project

Machine Learning Classifier for Detecting

Email Spams

## **Introduction**

Email is one of the most efficient and effective mode of communication with one another. Today a serious problem for web users and web services is caused by inflow of large number of spam emails. Spam emails are called the unwanted emails or unsolicited emails or bad emails which user receives without any prior information of the sender. Spam emails are usually trying to get the recipient to buy some product or services, spreading viruses, advertisements, for fraud in banking and for phishing. An estimation shows that close to 80% of all the emails are spam.

A spam filter is a software that keeps spam emails from entering the in-box. Hence, it predicts if an email is considered spam or no-spam, and decides if the email should be displayed in the in-box or be junked.



Existing spam filtering techniques use classification. Classification is a type of data analysis that extracts models describing important data classes or concepts. Classification mainly consists of two steps. First is the learning step: where a classification model is constructed and second is the classification step: in this step the extracted model is used to predict the class labels for new data or unknown data depending on the learning step. Machine learning algorithms are used for classification of objects of different classes. Such algorithms have proved to be efficient in classifying emails as spam or not spam.

The objective of this project will be to 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 answers to the following questions :

* 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 the model making false predictions?

### The Data Set

The Spambase data set is acquired from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/spambase>) which will be used for this project.

### Data Exploration

The Spambase dataset contains:

* 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).
* Most of the attributes indicate whether a particular word or character was frequently occuring in the email. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. The definitions of the attributes are described in table below:

**Table 1 : Attributes in Spambase Data**

|  |  |  |  |
| --- | --- | --- | --- |
| **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:

**Table 2: Summary of Outcome Variable ‘spam’**

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

### Data Cleaning

The data cleaning was steps includes as described below:

* Find and remove any missing values
* Change the name of the below attributes which have special characters in their name as below:
  1. char\_freq\_; to char\_freq\_semic
  2. char\_freq\_( to char\_freq\_openp
  3. char\_freq\_[ to char\_freq\_openb
  4. char\_freq\_! to char\_freq\_excl
  5. char\_freq\_$ to char\_freq\_dollar
  6. char\_freq\_# to char\_freq\_pound
* Since the column names were not present in the data, the column names were added to the data after above step.

## Data Analysis

Since outcome variable in this data set is ‘spam’ which is a binary or dichotomous, i.e. it only contains data coded as 1 (TRUE) or 0 (FALSE), the Logistic Regression algorithm will be used to classify if an email is spam or not a spam.

### What is 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).

*Logit Transformation*

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 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:

Logistic regression equation

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=p/(1-p)

And *logit transformation* of probability p is:

Logit(p)=ln(p/(1-p))

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

*Odds-ratio*

Equation 2 shows that if the probability of the outcome variable spam is between [0,1], the odds will be non-negative. If *odds > 1* the probability of an email being spam is greater than the probability of an email being no-spam.

*Definitions:*

### Dependent variable

The variable whose values you want to predict. The dependent variable must be binary or dichotomous, and should only contain data coded as 0 or 1.

Independent variables

The independent variables are the variables which are expected to influence the dependent variable.