## Springboard Introduction to Data Science Capstone Project Proposal

Machine Learning Classifier for Detecting

Email Spams

By

# Savita Shrivastava

# 

## **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.

## **Objective**

The objective of this project is to build a model using a machine learning method which can predict the outcome if an email is spam or no-spam. 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 a model that can predict if an email is spam?
* What is the risk of the model making false predictions?

## **Data Acquisition**

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

## **Method**

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).

The goal of logistic regression is to find the best fitting (yet biologically reasonable) 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 presence of the characteristic of interest:

Logistic regression equation

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

Odds=p/(1-p)

and

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.

**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.

**Project Deliverables:**

Project deliverables will be:

* Code
* Report in PDF format
* Slides