### **Apply Machine Learning to the Capstone project**

## Objective of Project

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 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 (Spambase data set) is acquired from UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/spambase>).

## 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.

### Building the Predictive Model and Performance Evaluation

Since the outcome variable spam has binary levels (0 or 1), logistic regression will be used to build the predictive model **using all of the independent variables (attributes 1 to 57)**. The data will be divided into a training and testing set with 75/25 ratio. The set.seed variable will be used to make sure the dependent variable will be well balanced in both the training and testing sets. Once the model will be developed with the training data, it will be subsequently tested on the test data to determine its accuracy.

### Performance Evaluation Parameters

The performance evaluation will be done using classification matrix as shown in the Table below and by calculating the accuracy, sensitivity and specificity of the model.

**Table: Classification Matrix:**

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

The description of parameters are as follows:

* 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

### Selecting a Threshold for Filter

By using a threshold value, the outcome of logistic regression which are probabilities can be converted into predictions. If the probability of an email being spam is greater than the threshold, then the prediction is that the email is spam. If it’s below, then prediction is that the email is not a spam. Selecting a right threshold is often challenging. A Receiving Operator Characteristic (ROC) curve will be created to decide which value of the threshold could be best.

### Evolution of Model using Test Data Set

The testing data set contains 1380 observations and will be used to evaluate the model. The threshold value be determined to create the classification matrix as shown in the table above to calculate accuracy, sensitivity and specificity of the model.

The accuracy of the model will be compared with baseline method accuracy.