**Objective:**

The candidates have to read the data and create a model based on the data analysis to identify if the website is legitimate or a phishing website. The Result will be determined by the two values [1, -1] where 1 represent the legitimate and -1 represents phishing

**Data Description:**

The data set consists of 30 features of a phishing website. The value of attributes can be [1, 0, -1] except the key value which is incremental.

* [1] represents the legitimate
* [0] represents suspicious
* [-1] represents phishing

**Features:**

There are a total 32 columns present in the dataset with 8955 number of rows. Data columns are based on address bar, domain and HTML and Javascript features.

**Modeling and training:**

In the starting phase we split the data into 60 - 40 i.e. 5373 training sample data and remaining i.e. 3582 testing sample data. From the dataset, From the objective and given dataset format we can identify it is a supervised task hence, there are two steps of supervised machine learning problems we have to follow which are called classification and regression.

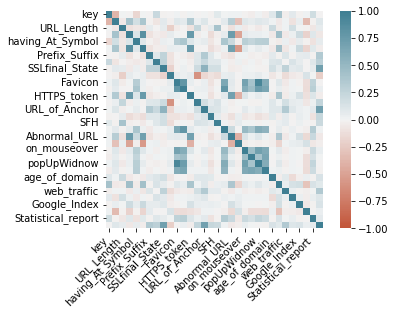
The given objective comes under classification problems and according to the input URL is classified as phishing (-1) or legitimate (1). The supervised machine learning models (classification) considered to train the dataset in this project are:

1. XGBoost
2. Random Forest
3. Decision Tree
4. Multilayer Perceptrons

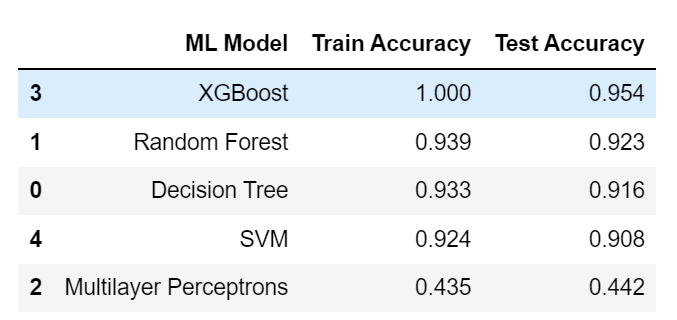
**Approach:**

* Analyze and preprocess the dataset and do the basic EDA on the dataset for correlation plot.
* Divide the data into train and test split dataset.
* Run the selected models.
* Do the performance evaluation
* Compare the performance of the models
* Save the top performing model and run it on sample test data
* Save the results

**How columns are affecting results?**

****

**Model evaluation table:**

****

From the above table we clearly can see that XGBoost is performing best on both the training and test data.