**DETECTION OF MALICIOUS URL USING MACHINE LEARNING TECHNIQUES**

A PROJECT REPORT

*Submitted by*

**K.RohithKumar (2015103527)**

**S.Harish (2015103553)**

*for*

**SECURITY LABORATORY (CS7712)**

**COURSE INSTRUCTOR: Dr. Arockia Xavier Annie R, Asst. Prof, DCSE, CEG.**

*during*

**July –Oct 2018**

**COLLEGE OF ENGINEERING GUINDY**

**ANNA UNIVERSITY, CHENNAI – 600 025**

**ABSTRACT**

The World Wide Web (WWW) supports a wide range of criminal activities such as spam-advertised e-commerce, financial fraud and malware dissemination. Although, the precise motivations behind these schemes may differ, the common denominator lies in the fact that unsuspecting users visit their sites. These visits can be driven by email, web search results or links from other web pages. In all cases, the user is supposed to take action, by clicking on a desired Uniform Resource Locator (URL). In order to identify these malicious sites, the web security community has developed blacklisting services. These blacklists are in turn constructed by an array of techniques including manual reporting, honeypots, and web crawlers combined with site analysis heuristics. Inevitably, many malicious sites are not blacklisted either because they are too recent or were never or incorrectly evaluated Machine Learning helps solving this problem, through its specialized techniques .We take this problem as Binary Classification Problem, where a label of 0 indicates that the URL is benign ( non malicious ) and a label of 1 to indicate URL is malicious. We use Logisitic Regresssion model to classify the URLs into above classes.

**ARCHITECTURE DIAGRAM**

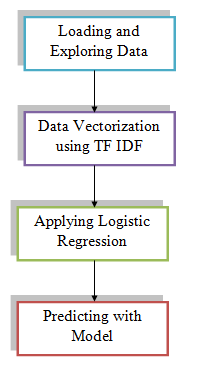


Figure 1 Architecture Diagram for Detection of Malicious URL using Machine Learning

**SYSTEM MODULES**

* **LOADING AND EXPLORING DATA**

The first step in this project is to load and explore the url data that we wish to process using Machine Learning. The dataset consists of 420654 rows and 2 columns. The two columns are namely the URL and the label of the URL, that is either good or bad. Since the data is already in preprocessed states, we can just peek around the top few observations in the dataset. The *input* to this module is the URL data and the *output* is the list of observations of input data.

DATASET <https://www.kaggle.com/antonyj453/urldataset>

* **DATA VECTORIZATION**

The second module is Data Vectorization. Since URL consists of character data, it is difficult to apply Machine Learning directly to the URL. So, we need to convert the character data to numeric suitable to apply logistic regression. For this, we use Term Frequency – Inverse Document Frequency methodology. The *input* to this module is the URL dataset and the *output* of this module is Matrix of numeric values.

* **APPLYING LOGISTIC REGRESSION**

The core module of this entire project is the applying of Logistic Regression to our URL dataset. We use Scikit Learn package, called LogisiticRegression() to perform this. The *input* to this module is the Numeric Matrix of values representing the features and the *output* is the Accuracy of the model.

* **PREDICTING WITH MODEL**

The final module in detection of malicious URL using machine learning is the testing our model, on unseen inputs.We need to verify whether it classifies correctly on new data, thereby checking our model is well trained or not. The input to this module is new URL data and the output of this module is Predicted Label for the given new input URL.

**SCREENSHOTS**

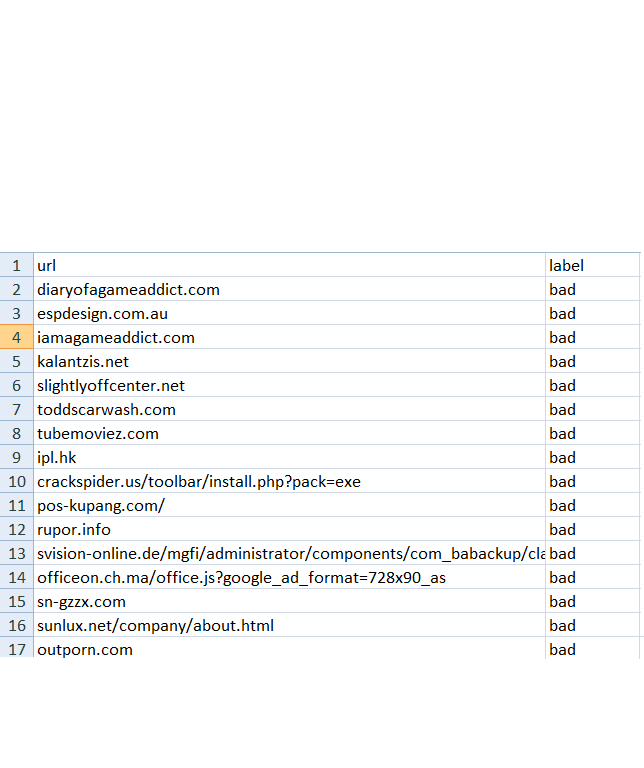


Figure 2 URL DATASET

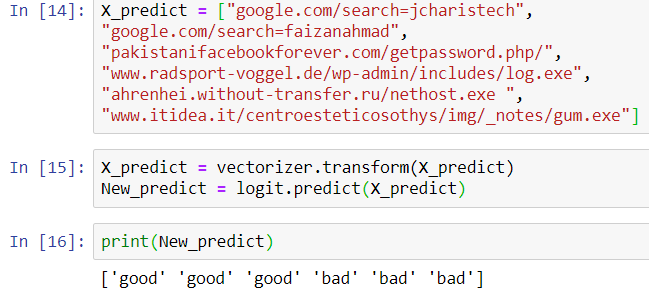


Figure 3 PREDICTION OF URL

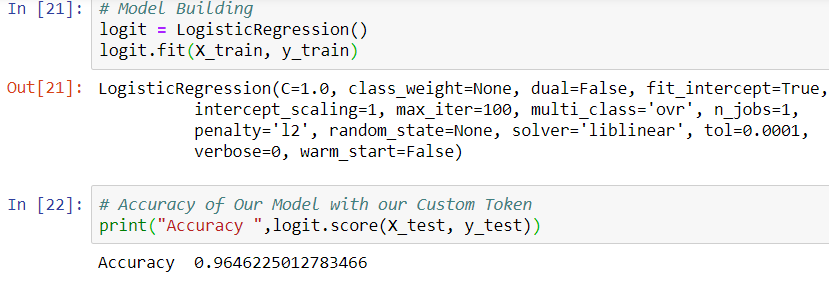


Figure 4 ACCURACY OF MODEL

**CONCLUSION**

Thus, We have been able to apply Machine Learning to Security of Users on the aspects of clicking of URL, using Logistic Regression. This project can be further extended by applying some advanced algorithms like Random Forest, Multi Layered Perceptron to achieve High Accuracy.