# Heuristic Based Phishing Detection Using URL Feature Extraction

[*]Aurelia Fernandes, [#]Alisha Gonsalves, [$]Janisa Colaco

[*,#]UG Student, [$]Professor, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India.

[*]aurelia.fernandes97@gmail.com, [#]gonsalves.alisha26@gmail.com, [$]janisa.colaco@frcrce.ac.in

**Abstract - Phishing is a type of social engineering attack often used to steal essential user information. In this attack the attacker masquerades as a trusted entity and dupes the system into thinking it is the actual user . Sometimes the user is sent a spam mail by the attacker . The user is may be tricked into clicking a malicious link which can cause the system to freeze till the user reveals sensitive information. For users, this includes the stealing of bank account credentials or unauthorised purchase , etc. Very often phishing is carried out through fake websites or through fake websites disguised as a legitimate site. We will be using a heuristic based phishing detection technique that uses uniform re-source Locater (URL) features. The method tends to improve the performance of the systems by choosing fewer URL features and makes it more efficient by using Logistic regression for classification and identification.**

***Keywords — data mining , feature extraction , heuristics , logistic regression , Phishing , URLs***

## I. INTRODUCTION

Digital fraud is all around us and if you are not careful with online transactions, you are leaving yourself vulnerable to be cheated off. Phishing is at its core just another con game, and attackers are nothing more than tech-savvy con artists/identity thieves. They use spam emails and malicious fake websites to trick people into giving out sensitive information. Phishing is defined as mimicking a creditable company's website aiming to take private information of a user. In order to eliminate phishing, different solutions were proposed. However, only one single magic bullet cannot eliminate this threat completely [1].

Data mining is a promising technique used to detect phishing attacks. In this project, an intelligent system to detect phishing attacks will be used. Data mining techniques are generally used to extract the features from the websites to find patterns as well as relationships between them. Data mining algorithms are highly imperative for decision-making, since decisions can be made based on the rules accomplished from a data-mining algorithm. Different classifiers in order to construct an accurate intelligent system for phishing, website detection will be used.

## II. LITERATURE SURVEY

### *A. Phishing URL Detection using URL Ranking:*

In this paper URLs are classified based on their lexical and host-based features. Classification requires proper encoding for different types of values associated with predictor variables to obtain a feature vector that accurately describes the URL. For instance, feature hashing is used in order to encode raw feature data into feature vectors [3]. The classifier achieves 93-98 percent accuracy by detecting a large number of phishing hosts, while maintaining a modest false positive rate[2]. URL clustering, URL classification, and URL categorization mechanisms work in conjunction to give URLs a rank.

### *B. Intelligent Phishing Website Detection using Random Forest Classifier:*

An intelligent system to detect phishing attacks was presented. We used different data mining techniques to decide categories of websites: legitimate or phishing. Different classifiers were used in order to construct accurate intelligent system for phishing website detection. Classification accuracy, area under receiver operating characteristic (ROC) curves (AUC) and F-measure is used to evaluate the performance of the data mining techniques. Results showed that Random Forest has outperformed best among the classification methods by achieving the highest accuracy 97.36 percent [4]. But in this paper only known and commonly used data mining algorithms were researched and used.

### *C. Fraud Website Detection using Data Mining:*

This project provides intelligent solution to phishing attack. This application extracts some characteristics from URL and source code of a website. These features are used for classification. RIPPER algorithm is used to classify the websites. Netbeans is used for this purpose. Weka.jargles are used for implementing the algorithm. 1250 URLs are given as input for training set. The algorithm is tested on 250 URLs containing 125 legal and 125 fraud URLs. The

 DOI : 10.18231/2454-9150.2019.0321 

accuracy obtained is 86.4 percent[5]. Nowadays, it is important to detect fraud website on zero day as the fraud websites are short lived because the government may identify these websites and shut them down. They are designed to create maximum damage before getting tagged and listed as black listed website. List based detection is unable to detect fraud websites on zero day or before the fraud website is blacklisted. Use of heuristic-based detection approach in Fraud website detection application, enables it to detect fraud websites before they are blacklisted. But the drawback in this paper is that too many features are used .

## III. Overview of Features

Some features that are used to distinguish fraud websites from legitimate ones are as follows:

1) Address Bar based Features: These include various features such as IP address, '@' symbol, length of the URL, redirecting using '//', prefixes & suffixes, Sub-domains, domain registration length.

2) Abnormal based Features: URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Abnormal URL, Server Form Handler (SFH)

3) HTML and Javascript based Features: Website Forwarding, Status Bar Customization, IFrame Redirection, Using Pop-up Window

4) Domain based Features: Age of domain, DNS record, Website traffic, Page rank, Google index, Number of Links Pointing to Page

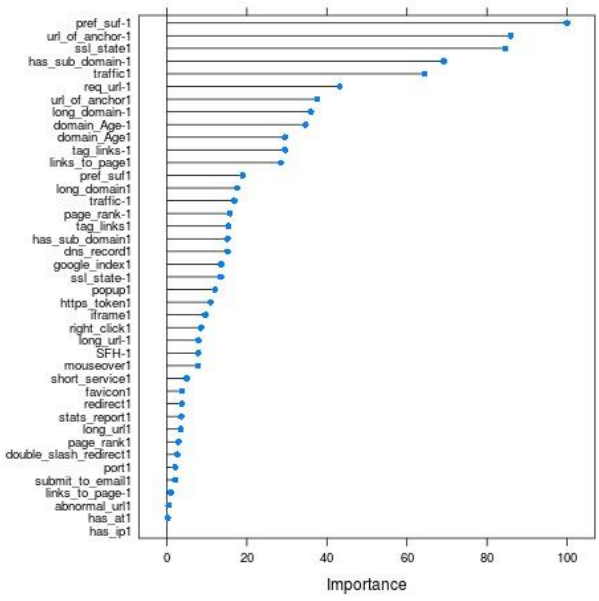


Figure 1: Importance of Features [9]

## IV. Proposed System

The objective is to train and develop a system that is able to distinguish between a phishing and a legitimate website based on the training done using an acquired dataset.

Build classifier: After acquiring a dataset, Split the dataset into (train-test split) 70 percent (training data)-30 (testing data). Use Logistic Regression algorithm to train the model and build the classifier. After Training the classifier , test the classifier to check the accuracy of the classifier. Repeat the training and testing till we get minimum error

Predicion: After getting the Url submitted by the user, Extract Features from the URL using following steps : Check its ssl state (whether its has ssl or not). Get DNS info of the URL. Check if the site content has iframes and alert used. Check if the Url is running on port default 80 port or not. Similar to above mentioned all other features will be extracted such as Google page rank, DNS record, google index , etc. Feature extracted will be stored in a csv _le and later used for prediction. Pass the extracted features and the url to the classifier, which is prior trained and tested. Send the output from the classifier to the Frontend

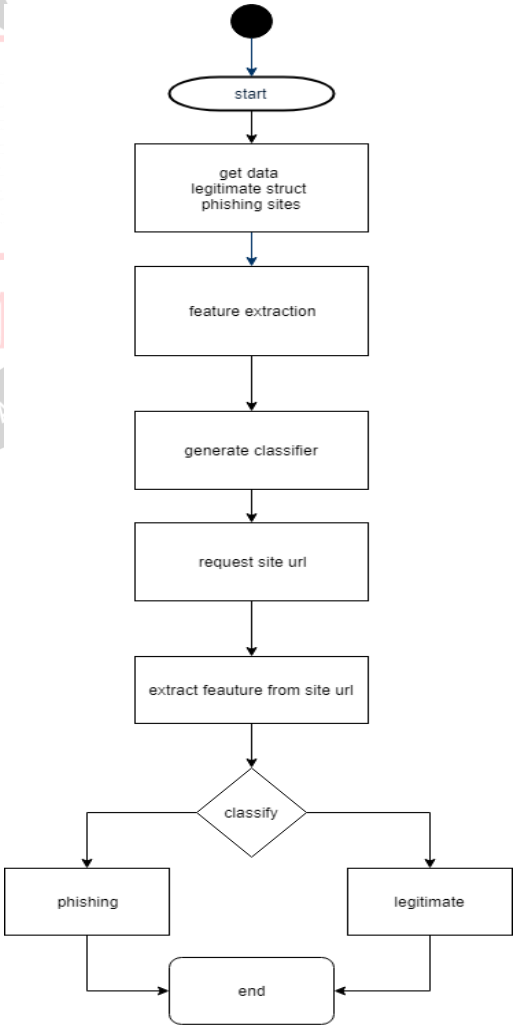


Figure 2. Flow chart of proposed system

Apart from Training and testing the system , we also need to inculcate additional features like image classification,

 DOI : 10.18231/2454-9150.2019.0321 

google safe browsing and local data base search. If possible we also need to find a way to identify phishing urls that are embedded within legitimate sites.

## V. EXPERIMENTAL WORK

### A. Feature Extraction:

We considered 30 different features that would help as deciding variables such as Using Non-Standard Port, The Existence of "HTTPS" Token in the Domain Part of the URL, Request URL, URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Server Form Handler (SFH), Submitting Information to Email, Abnormal URL, Website Forwarding, Status Bar Customization, Disabling Right Click, Using Pop-up Window, IFrame Redirection, Age of Domain, DNS Record, Website Traffic, PageRank, Google Index , Number of Links Pointing to Page , Statistical-Reports Based Feature

| Sr No. | Features | Rules |
|---|---|---|
| 1. | Using the IP Address | Rule: IF{If The Domain Part has an IP Address → Phishing Otherwise→ Legitimate |
| 2. | Long URL to Hide the Suspicious Part | Rule: IF{$URLlength$<54 → $feature$=Legitimate $elseifURLlength$≥54 $and$ ≤75 → $feature$=$Suspiciousotherw$ $se$→ $feature$=Phishing |
| 3. | Using URL Shortening Services "TinyURL" | Rule: IF{Tiny URL → Phishing Otherwise→ Legitimate |
| 4. | URL's having "@" Symbol | Rule: IF {Url Having @ Symbol→ Phishing Otherwise→ Legitimate |
| 5. | Redirecting using "//" | Rule: IF {The Position of the Last Occurrence of "//" in the URL > 7→ Phishing Otherwise→ Legitimate |
| 6. | Adding Prefix or Suffix Separated by (-) to the Domain | Rule: IF {Domain Name Part Includes (−) Symbol → Phishing Otherwise → Legitimate |
| 7. | Sub Domain and Multi Sub Domains | Rule: IF {Dots In Domain Part=1 → Legitimate Dots In Domain Part=2 → Suspicious Otherwise→ Phishing |
| 8. | HTTPS | Rule: IF{Use https and Issuer Is Trusted and Age of Certificate≥ 1 Years → Legitimate Using https and Issuer Is Not Trusted → Suspicious Otherwise→ Phishing |
| 9. | Domain Registration Length | Rule: IF{Domains Expires on≤ 1 years → Phishing Otherwise→ Legitimate |
| 10. | Favicon | Rule: IF{Favicon Loaded From External Domain→ Phishing Otherwise→ Legitimate |

### B. Algorithm

Logistic regression is another machine learning technique from the field of statistics. Logistic regression is a method that can determine the output based on one or more independent variables. It is generally used in binary classification problems . The outcome is measured with a dichotomous variable i.e there are only two possible outcomes .

In logistic regression, the dependent variable is binary or divided, i.e. it only contains data coded as true or false. The goal of this algorithm is to find a model that best fits . If you ever draw a logistic regression graph you will see that it follows a straight line of equation y=mx+c. That means Logistic regression algorithm helps us find values that best fit the straight line equation.

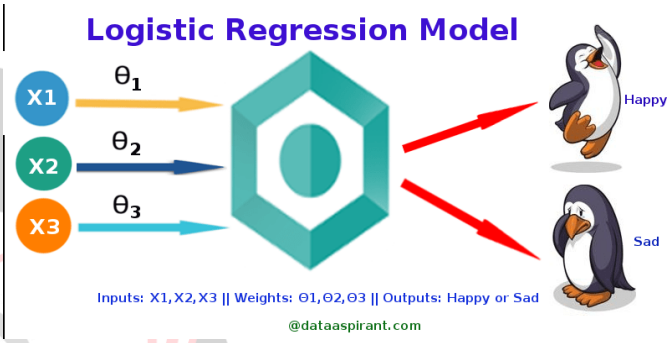


Figure 3: Logistic Regression Problem [8]

### C. Training and Testing

1) Extract features from the URL and Load Dataset
2) Remove unnecessary ID column from CSV dataset
3) Get X and Y from the dataset to fit Logistic Regression equation.
4) Using cross validation, split data for training and testing then find the values that fit the equation: y=mx+c
5) Do prediction on the testing data to test accuracy of the system that we have designed
6) Create confusion matrix by calculating error i.e. to find how well the system is performing
7) Model is trained to classify between clean and phishing websites. Populate the result and display on frontend

## VI. IMPLEMENTATION DETAILS

We conducted classifier training and testing through a dataset of more than 5000 URL's found on the UCI website. The dataset was divided by cross-validation technique into 75% for training and 25% for testing. During training we found the accuracy of the system to be 0.9023 which seemed pretty good.

 DOI : 10.18231/2454-9150.2019.0321 


Figure 4. Front-end display

The system is first designed to check if the URL entered is a valid URL .After the URL entered satisfies all the conditions , its features are extracted and whether the URL is phishing or not is verified using Logistic Regression as shown in the Figure below. Tensor flow was used to train and test the model based on the 30 features mentioned in table1 .

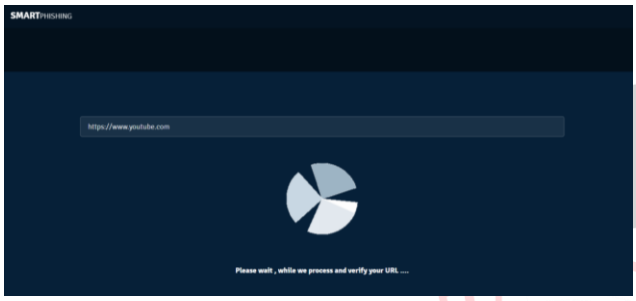


Figure 5. URL processing and verification

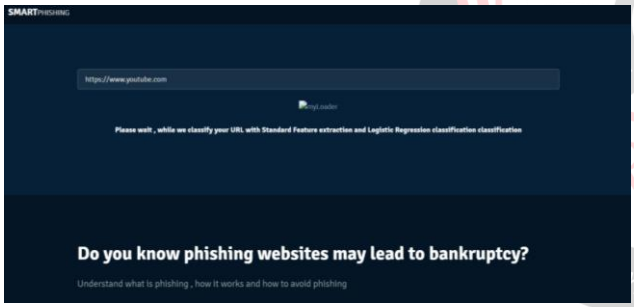


Figure 6. Logistic Regression Classification

In addition ,to increase system accuracy (i.e to make it more precise) we also included local database search, image classification and Google safe browsing. As a result it should be noted that classification just based on features is not enough. Our system is able to identify youtube, google, gmail and other common known websites too.

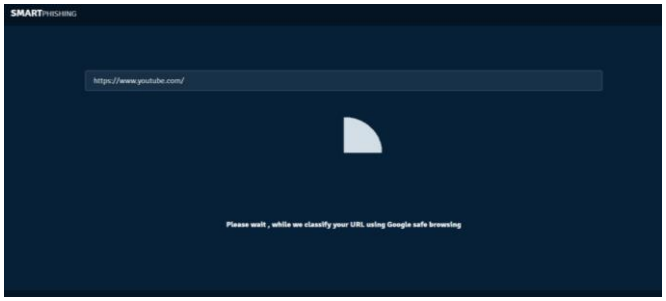


Figure 7. Google safe browsing

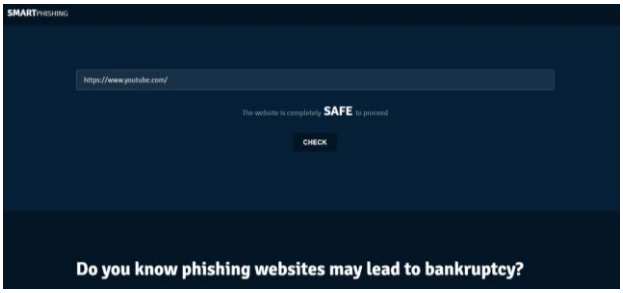


Figure 8. Result

## VII. CONCLUSION

Logistic Regression is an intelligent ML algorithm that is used in binary classification . But it may not be the best algorithm to classify phishing and legitimate websites.

In this study our system is able to detect the phishing websites by using our feature extraction technique. But our proposed system should be able to achieve more accurate predictions.

We also determine that to be able to predict more accurately or with more precision we need to consider other options such as google safe browsing, Local database search and Image classification. Using this process our accuracy has really improved compared to normal approach of feature extraction. Although logistic regression's runtime is not as fast as we would expect but it is still able detect phishing websites.

## REFERENCES

[1] https://us.norton.com/internetsecurity-online-scams-how-to-protect-against-phishingscams.html

[2] Phishing URL detection using URL Ranking by Mohammed Nazim Feroz, Susan Mengel

[3] Brad Green and Shyam Seshadri, "AngularJS", O'Reilly Media Inc., September 201

[4] Intelligent Phishing Website Detection using Random Forest Classifier by Abdul hamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery

[5] Fraud Website Detection using Data Mining by Urvashi Prajapati, Neha Sangal, Deepti Patole

[6] Phishing Website Features by Rami Mohammad, Fadi Thabtah, Lee McCluskey

[7] Website Phishing Detection using Heuristic Based Approach by Jaydeep Solanki , Rupesh C. Vaishnav

[8] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[9] http://rishy.github.io/projects/2015/05/08/phishing-websites-detection/

[10] Machine Learning Based Phishing Detection from Urls by O. Sahingoz, E. Buber, O. Demir, B. Diri

[11] Design and implementation of heuristic based phishing detection technique by Patel, Jaynish

[12] Phish safe url feature based phishing detection system using machine learning by AK Jain , BB Gupta

 DOI : 10.18231/2454-9150.2019.0321 