May-June 2020
ISSN: 0193-4120 Page No. 4876-4881

Machine Learning based Phishing Website Detection

[1]Chandana T U, [2]Chandni B, [3]Eshwari C, [4]Cheathana M, [5]Shantala Devi Patil

[1,2,3,4,5]School of Computing and Information Technology, REVA University, Bangalore, India
[1]chandana021998@gmail.com, [2]chandsitara56@gmail.com, [3]eshwari1419@gmail.com,
[4]chethana41998@gmail.com, [5]shantaladevipatil@reva.edu.in

Article Info
Volume 83
Page Number: 4876-4881
Publication Issue:
May-June 2020

Abstract
Phishing is a technique to obtain or exploit the personal information of an individual by imitating an existing website or by offering interesting schemes through email, text messages. Phishes steal important and secured information like passwords, credit card details, phone numbers. Nowadays phishing attacks are increasing which is extremely problematic for social and economic websites. The prime focus of the paper is to build a powerful application that applies Machine Learning techniques and tools to identify phishing websites. Training with one classification model is not the best way in the case of predicting websites because accuracy plays an important role. Therefore, we consider various Machine Learning algorithms such as Random Forest(RF),Logistic Regression model(LR), Support Vector Machine(SVM) or maximum-margin classifier, Decision Tree(DT), Sequential Multilayer Perceptron(MLP), Naïve Bayes(NB). After reviewing each algorithm we select a classification model with the highest accuracy to detect new fake websites given by the user.

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Keywords: *Phishing attacks, features, Machine Learning techniques, Random Forest, Logistic Regression, Support Vector Machine, Decision Tree, Neural Network, Naïve Bayes*

1. Introduction

Phishing is descended from the word "fishing". Similar to fishing the scammer searches for the potential user who gets trapped for his trick. This trick involves sending a message or email with a link. This link generally imitates a trusted website like amazon or SBI bank or it lures the user to enter sensitive details. Sensitive details can be password, credit or debit card details, phone numbers or even security numbers. Phishing started long ago with phone calls, now it is done through emails, SMS and social media. [12] malware bytes tells how it is considered as the simplest cyberattack and, at the same time, the most dangerous and successful. That is because it's easy for attackers with zero skills to launch sophisticated phishing attacks. Phishing harms the organization's revenues, client relations, marketing efforts, and overall corporate image. One of the most well-known types of phishing attacks is deceptive phishing or clone phishing. Scammer mimic a bank or e-Commerce websites, those emails scare users into doing what the attackers want. Usually, these fraudsters use warnings or a sense of emergency or make their emails look like they are from any well-known domain. for example,[14]FBI conducted largest international phishing case called as Operation phish phry in 2009 where,The defendants targeted U.S. banks and robbed thousands of account holders by stealing their account details and used it to transfer about $1.5 million to fake accounts they managed. The next class of phishing attacks is called spear phishing, in contrast to deceptive phishing emails that use spam-like strategies to attack a large number of people in huge email campaigns, spear phishing emails aim specific people within a corporation. They

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 4876-4881

use special tactics to customize their attack emails with the victim's name, position, phone number, and related information. They may use subject lines that would interest the email recipients to trick them into clicking on links or attachments.[13] Director of research at the SANS Institute Allen Paller"more than 90% of all attacks on business networks are the result of successful spear phishing". Additionally, spear phishing where the target is powerful executives of an organization like CEOs is called as whaling. Whalers are interested in Business email compromise(BEC), where the attackers ask the employees and clients to transfer money or use it for the data breach. Whaling is dangerous considering victims can lose thousands of dollars or valuable information

As phishing attackers have advanced in years we need the best data security technologies to avoid cybercrime attacks. According to [4] Phishing detection approach is divided into two layers; the human layer and the software layer. Phishing can be avoided if the user is beware of malicious websites but this fails almost all the time. Therefore, we require technical solutions to detect these websites. There are many anti-phishing techniques over the years, such as blacklisting, content or visual-based detection, and Machine learning techniques. In blacklisting, the browser or the application checks whether the URL is listed as phished, if the URL is found in the list it alerts the user. Although it protects the system by blocking malicious websites it rarely detects zero-hour phishing websites. In the second technique, the visual and content of the websites are analysed to find the legitimacy of the website. The more effective and efficient way of finding phished websites is by machine learning techniques.

As accuracy is the most sought out feature in the case of detecting zero-hour phishing websites. It is a known fact that machine learning classifications predict and improve its performance with experience. The accuracy of the classification model to a great extent depends on the features of the URL that are considered for the phishing campaign. In that context, it is imperative to provide a trustful and precise tool using machine learning algorithms to identify this threat before falling into the trap. This paper focuses on various machine learning techniques and also provides a user-friendly solution used for phishing detection.

2. Related Works

In [1] Vaibhav Patil, Tushar Bhat, Pritesh Thakkar, Chirag Shah, Prof S. P. Godse proposed a browser extension system that goes through a three-layer of protection to prevent phishing attacks. This model monitors all the URLs and compare domain with white-list of known domains and also the black-list of malicious domains. We can cheat on many people if the fake websites look exactly like the targeted website. Therefore CSS of the target website is compared with the CSS of all of the non-phished domains in the queue. Furthermore, all the features of the websites are extracted and various machine learning classifier like Decision Tree, Logistic Regression, and the Random Forest is applied to mark as phishing and block it. According to the paper, Using different approaches collectively will improve the efficiency of the system..

In [2], According to APWG's 3rd quarter report number of phishing websites is highest which was last seen in 2016. SaaS and webmail sites continued to exist as the biggest targets of phishing whereas cloud storage and eCommerce remained less popular.40% of Business Email Compromise(BEC) phishing make use of domains registered by offenders. In a BEC, intruder aims employees with access to administration and management department of the organization. Over 66.7% of phishing websites used SSL protection, which was recorded as highest ever. Average of 313 brands were attacked per month in Q2 whereas it increased to 400 brands in Q3. More and more phished websites are created daily using new and improved methods, therefore new models must be implemented to detect new and zero-hour phishing websites.

In [3] Experian estimated the possible data breach trends, based on their expertise with more people joining the like-minded social media group there are new potentially unsuspected targets that can be easily manipulated by scammers. These scammers use user convincing smishing attacks in the form of donations or offers. Increasing mobile payments spikes the possibility of identity theft significantly.

In [4] Ebubekir Buber, OnderDemir, Ozgur KoraySahingoz presents phishing as a major threat and compares two major ways to prevent it that is black/whitelist approach and machine learning techniques. It is seen that the black/whitelist approach does not identify zero-hour phishing websites, which are new types of phishing attacks, whereas machine learning techniques are an effective and efficient way. Features used for phishing URL detection are classified asURL, Domain, Page, and Contentbased features. This paper aims to list important features used for the machine learning approach to detect whether a website is legitimate or not.

In [5]Jatin Shad, Siddharth Gaur, Gagandeep Kaur, IshantTyagi, Shubham Sharma proposed a newmethod that detects phishing URLs by machine learning solutions. This paper states that Machine learning is a persuasive tool that can be used to identify phishing attacks. This system uses algorithms such as Decision Tree(DT), Random Forest(RF), Gradient Boosting(GBM), Generalized Linear Model(GLM). These algorithms are compared on the bases of accuracy, recall, and performance by R programming language.30 Features are extracted using python language. From the comparison table consisting of recall, precision, and accuracy, the random forest is a more effective model.

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 4876-4881

In [6] Dr.SumilaGodara, Meenu proposed a framework using various machine learning techniques on the Microsoft Azure stage. A filter-based feature selection method along with feature hashing is used to enhance the efficiency of the classification model by selecting the best features and finds the best results, besides the comparison of various machine learning techniques also enhance the accuracy of the detection. The system comparison is done by considering f1 score, accuracy, precession, and recall. According to the paper, Improved Logistic Regression has the highest accuracy.

In [7] Ba Lam To, Minh Hoang Nguyen,HuuKhuong Nguyen,Luong Anh Tuan Nguyen proposed a new Heuristic URL-based approach. This system derives four features from URL and six metrics (heuristics)such as primary domain, subdomain, pathdomain, pagerank, alexarank and alexareputation are calculated for each component. Websites with negative heuristic value is considered as phished and websites with positive value is said to be real.

In [8] Scammer aims the attack on a massive scale or targeted users, the latter being difficult to detect. Therefore Jake Drew and Tyler Moore suggested a combined clustering method to find criminal websites. The dataset of criminal websites considered here ishigh-yield investment programs (HYIPs) and fake-escrow services. Information on the websites is taken into consideration and features such as webpage text and HTML tags are selected. The distance matrix is calculated by pairwise similarities between websites for each attribute using hierarchical and agglomerative algorithms. Besides, a generalised distance matrix is calculated by individual distance matrices, which is again applied to Hierarchical andagglomerative methods to yield the final result.

In [9] Peng Yang, Guangzhen Zhao, Peng Zeng proposed a multidimensional feature phishing detection(MFPD) method using deep learning and based on fast detection method. In this approach, an URL goes through two steps of feature extraction. Initially character sequence and local correlation attributes are extracted using CNN(convolutional neural network) and LSTM(long short-term memory) network is applied to obtain context semantic and dependency attributes of URL. The detection time of this step is less therefore obtained result is checked by using Softmax classifier in deep learning. If the result is greater than the specified threshold the process is stopped else it goes to the second step where they combine webpage text, code features, URL statistical features and the result into multidimensional fields which are classified using a gradient boosted decision tree called as XGBoost. MFPD ensures high detection speed and high accuracy.

In [10]Stephen Groat,Matthew Dunlop, David Shelly proposed a scheme called GoldPhish which uses screenshot, optical character recognition and search engine for the detection of fake websites. First, the user's web browser captures the image of the current website and the screenshot is converted from Bitmap to TIFF image. The second step is to process the image to text by Optical Character Recognition software. Later Google search retrieves the website PageRank, if the PageRank is low that is the site has only been up for a short time is considered as the phished websites. It is stated that although GoldPhish has limitations like time delay, textual contents, and font of the page, images to verify the domain and logos it has provided accuracy of 98% with 0% false positive and 2% false negative report better than heuristic-based tools.

In [11] R. Kiruthiga, D. Akila provides a survey on the features considered and detection algorithms using machine learning for the detection of phishing websites. According to the papermany works are done is by using machine learning algorithms such as Naïve Bayes, Support Vector Machine, Random Forest. Few authors proposed a novel approach like PhishScore and PhishChecker which uses fewer features of the dataset from PhishTank and Yahoo directory set for detection.

We have considered many ways by which phishing attacks are avoided. Some of the paper gives novel ideas on detection. Whereas, Many have used any one of the machine learning concepts in predicting the legitimate level of websites which makes accuracy limited. Even though some have considered two or more techniques it is neither user friendly nor the features considered helps in finding newer malicious websites.

3. Proposed Work

The proposed method uses Machine learning techniques along with the use of a dataset with binary class and features to predict malicious websites. The project is classified into two parts. We briefly describe each step below:

Selection of Machine Learning Algorithm

From the related works, it is noted that machine learning with feature selection is more effective than other existing techniques. Because machine learning is known to learn and enhance its performance, hence to detect a new URL it must be thoroughly trained with various algorithms. Here we have considered six well-known algorithms:

A. Logistic Regression: The logistic regression model is a binary classification model i.e., it classifies into two categories like whether a website is phished or not. Unlike linear regression, logistic regression fits an S-shaped curve called *Sigmoid.* Logistic regression is applied after removing *id* from the dataset. Also to increase the accuracy Grid Search is used to select optimal parameters for the model.

B. Support Vector Machine: Like logistic regression, it is a supervised and binary classification. The output of this model is a hyperplane which is a line in two

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 4876-4881

dimensions. Among many hyperplanes, one which has a maximum distance from the nearest point is selected as a solution. Also to improve the efficiency data cleaning and grid search is used to select optimal parameters for the model.

C. Decision Tree: Decision tree is a tool that uses trees for classification and prediction. Each attribute of the dataset is drawn by internal nodes and classes are drawn by leaves. It splits on the feature with the highest value.

D. Random Forest: Random Forest in simple words contains many decision trees for the classification. Considering one decision tree for classification is complex. Random forest lessens the risk of overfitting and the expected training time. Additionally, it offers a high level of accuracy. it is applied along with grid search and after the data cleaning process. Besides, the random forest algorithm is used to know the value of each feature using the "feather_importance_" attribute for a better understanding of the dataset.

E. Gaussian Naïve Bayes: Gaussian Naïve Bayes is a type of Naïve Bayes classification model. It is a probabilistic model based on Bayes theorem and Normal Distribution.

F. Neural Network: Neural Network is a software implementation of the human brain for various problem-solving.It is a foundation for deep learning. The Neural network used here is sequential multilayer perceptron network, this modelhas 2 hidden layers with 40 and30 neuronsand an output layer with 1 output. Training continues until there is no change in weights and accuracy with less loss.

Feature Extraction and Dataset

Dataset plays an important role in the classification and prediction model. the dataset is taken from the database UCI Machine learning Repository which has many datasets and other information regarding machine learning. It has cited many papers and helped researchers in better understanding of machine learning. It also provides an easily understood dataset description. Dataset used here has 30 URL features which are classified as binary classes i.e., -1 for phishing and 1 for legitimate websites. The URL features are distributed as 1,-1 and 0 representing non-phished, phished and suspected respectively. The dataset has 4898 phished websites and a 6157 non-phished class. For better understanding on how attackers think when they develop a phishing domain, we must know the URL structure(Fig.1).

Figure 1: URL Structure

The Phisher tries to manipulate the protocol, domain and, path. Sometimes they cheat the victim by imitating the visual features of the real websites along with similar domain names. Based on the dataset and URL submitted different features are extracted. Each type of feature is described below.

A. Addressbar based features: This type of attributes deals with URL length, use of shortening services, favicon, domain registration length, use of non-standard ports, a certificate from a nontrusted issuer, etc.

B. Abnormal-based features: Phishers try to load images, sounds and various anchor tags from other domains and most of the time fake websites submit the user information to the phisher's mail. This type of feature extraction goes through all the images, script, anchor tags to classify websites.

C. HTML and JavaScript based features: We extract HTML and java script features like, how many times the websites have been redirected to another page, whether the host has used iframe and hid the additional webpages, use of onMouseOver to change status bar, etc.

D. Domain-based features: Domain information like Google Page Rank,Index, DNS record, no. of links pointing to the page are extracted in this type of feature extraction. Fig.2 shows the info of the datset features

```
RangeIndex: 11055 entries, 0 to 11054
Data columns (total 32 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   id                           11055 non-null  int64
 1   having_IP_Address            11055 non-null  int64
 2   URL_Length                   11055 non-null  int64
 3   Shortining_Service           11055 non-null  int64
 4   having_At_Symbol             11055 non-null  int64
 5   double_slash_redirecting     11055 non-null  int64
 6   Prefix_Suffix                11055 non-null  int64
 7   having_Sub_Domain            11055 non-null  int64
 8   SSLfinal_State               11055 non-null  int64
 9   Domain_registeration_length  11055 non-null  int64
 10  Favicon                      11055 non-null  int64
 11  port                         11055 non-null  int64
 12  HTTPS_token                  11055 non-null  int64
 13  Request_URL                  11055 non-null  int64
 14  URL_of_Anchor                11055 non-null  int64
 15  Links_in_tags                11055 non-null  int64
 16  SFH                          11055 non-null  int64
 17  Submitting_to_email          11055 non-null  int64
 18  Abnormal_URL                 11055 non-null  int64
 19  Redirect                     11055 non-null  int64
 20  on_mouseover                 11055 non-null  int64
 21  RightClick                   11055 non-null  int64
 22  popUpWidnow                  11055 non-null  int64
 23  Iframe                       11055 non-null  int64
 24  age_of_domain                11055 non-null  int64
 25  DNSRecord                    11055 non-null  int64
 26  web_traffic                  11055 non-null  int64
 27  Page_Rank                    11055 non-null  int64
 28  Google_Index                 11055 non-null  int64
 29  Links_pointing_to_page       11055 non-null  int64
 30  Statistical_report           11055 non-null  int64
 31  Result                       11055 non-null  int64
```

Figure 2: URL Features

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 4876-4881

After successful training of the model, it is ready to detect new URLs for this purpose we created a User Interface where can run the application and enter new URLs. The application is hosted using the *Flask* framework andfor prediction, we make use of the machine learning model selected after the study and obtained results.

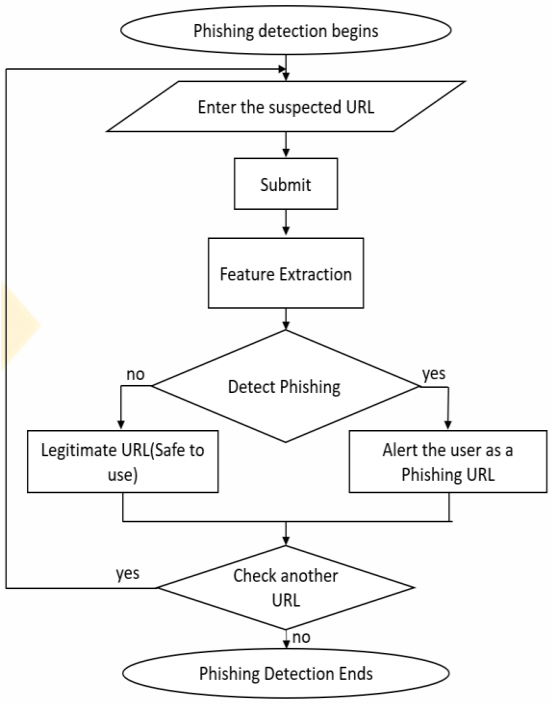

Figure 3: Detection of Phishing website Flowchart

The flow control of the application is demonstrated through flow chart (Fig) and the following steps take place in the project to detect websites legitimate level:-

1. Phishing campaign starts when run the web server and enter the local host *ip* in any browser.
2. After the front-end is loaded user can enter any websites URL and click on submit.
3. URL is fed to the feature extraction method and all the features are extracted as values -1, 0, 1.
4. The selected classification model predicts the website's legitimate level by considering the attributes extracted.
5. The result is displayed in the browser, if we wish to detect another website we can enter another URL or stop the process.

4. Results

As mentioned in Section III we have applied six machine learning algorithms on the dataset. By analysing the accuracy of each algorithm the result obtained are given below:

1. Logistic regression has an accuracy of 92.32% after removing unwanted attributes from the dataset.
2. Support Vector Machine showed an accuracy of 96.74% after Grid Search selects C parameter as 1000 and gamma as 0.2 .
3. Random Forest showed an accuracy of 96.85% after Grid Search selects criterion as Gini impurity, max features as log2, and n estimators as 100.Random forest is also applied to find the feature importance and as seen in Fig.3, SSL certificate issuer and validity have relatively high importance among various URL features.

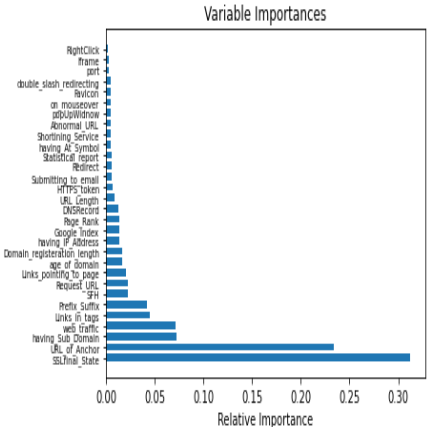

Figure 4: Relative importance of URL features from the dataset

4. When Gaussian Naive Bayes is applied accuracy is found to be 61.17%
5. The Decision tree has an accuracy of 96.09% on the dataset.
6. The Neural network selected here is the Sequential Multilayer perceptron. 128 epochs are applied where it stops at 22 with a loss of -0.43 with an average accuracy of 79.33%.

After successful training of the all the above models Fig.4 shows the accuracy graph. From that graph it is clear that Naïve Bayes has least accuracy compared to other algorithms. Whereas, random forest, decision tree and SVM has higher accuracy. Therefore for higher of the project we make use of random forest model or support vector machine.

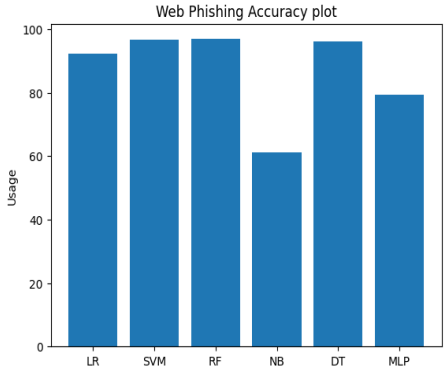

Figure 5: Machine Learning accuracy graph

Published by: The Mattingley Publishing Co., Inc.

May-June 2020
ISSN: 0193-4120 Page No. 4876-4881

5. Conclusion

Phishing has become more threatening to society as we advance to complete digital transactions. Therefore the aim is to create an effective tool to avoid this threat. In this paper, We have analysed and tested various features of an URL and also the importance of each feature. From the results obtained various machine learning algorithms we either apply Random Forest or Support Vector Machine to find malicious websites.

References

[1] Vaibhav Patil, Tushar Bhat, Pritesh Thakkar, Chirag Shah, Prof S. P. Godse "Detection and Prevention of Phishing Websites using Machine Learning Approach" in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, 2018.

[2] The Anti-Phishing Working Group "Phishing Activity Trends Report 3rd Quarter 2019" on November 4, 2019.

[3] Experian Data Breach Resolution "Data Breach Industry Forecast2020"https://www.experian.com/content/dam/marketing/na/assets/data-breach/white-papers/Experian-Data-Breach-Industry-Forecast-2020.pdf

[4] Ebubekir Buber, OnderDemir, OzgurKoray Sahingoz "Feature Selection for the Machine Learning based Detection of Phishing websites" 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, 2017.

[5] IshantTyagi, Jatin Shad, Shubham Sharma, Siddharth Gaur, Gagandeep Kaur "A Novel Machine Learning Approach to Detect Phishing Websites" 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, 2018.

[6] Dr.SumilaGodara, Meenu "Phishing Detection using Machine Learning Techniques" International Journal of Engineering and Advanced Technology (IJEAT), Volume-9 Issue-2, December 2019.

[7] Luong Anh Tuan Nguyen, Ba Lam To, HuuKhuong Nguyen, Minh Hoang Nguyen "A novel approach for phishing detection using URL-based heuristic" 2014 International Conference on Computing, Management and Telecommunications (ComManTel), Da Nang, 2014.

[8] Jake Drew, Tyler Moore "Automatic Identification of Replicated Criminal Websites Using Combined Clustering" 2014 IEEE Security and Privacy Workshops, San Jose, 2014.

[9] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, 2019.

[10] M. Dunlop, S. Groat and D. Shelly, "GoldPhish: Using Images for Content-Based Phishing Analysis," 2010 Fifth International Conference on Internet Monitoring and Protection, Barcelona, 2010.

[11] R. Kiruthiga, D. Akila "Phishing Websites Detection Using Machine Learning" International Journal of Recent Technology and Engineering, Volume-8, Issue-2S11, September 2019.

[12] Malwarebytes.com, 'What is Phishing? Types of Phishing and Examples'.[Online]. Available: https://www.malwarebytes.com/phishing/ [Accessed: 14-04-2020].

[13] Neal Weinberg, "How to blunt spear phishing attacks", 2013. [Online]. Available: https://www.networkworld.com/article/2164139/how-to-blunt-spear-phishing-attacks.html. [Accessed: 14-04-2020].

[14] The Federal Bureau of Investigation, "Operation 'Phish Phry' Major Cyber FraudTakedown",2009.[Online].Available:https://archives.fbi.gov/archives/news/stories/2009/october/phishphry_100709. [Accessed: 16-04-2020]

Published by: The Mattingley Publishing Co., Inc.