International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume: 3 | Issue: 4 | May-Jun 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 - 6470

Detecting Phishing using Machine Learning

Nakkala Srinivas Mudiraj

M.Tech, Computer Science & Technology (Cyber Security), Central University of Punjab, Bathinda, Punjab, India

How to cite this paper: Nakkala Srinivas Mudiraj "Detecting Phishing using Machine Learning" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-4, June 2019, pp.488-490, URL: https://www.ijtsrd.com/papers/ijtsrd23755.pdf

IJTSRD23755

ABSTRACT

Phishing is a social engineering Technique which they main aim is to target the user Information like user id, password, credit card information and so on. Which result a financial loss to the user. Detecting Phishing is the one of the challenge problem that relay to human vulnerabilities. This paper proposed the Detecting Phishing Web Sites using different Machine Learning Approaches. In this to evaluate different classification models to predict malicious and benign websites by using Machine Learning Algorithms. Experiments are performed on data set consisting malicious and benign, In This paper the results shows the proposed Algorithms has high detection accuracy.

Keywords: *Benign; Malicious; machine Learning; Phishing; Social Engineering*

INTRODUCTION

Social Engineering is the one of the big threat that steals the sensitive and confidential information of the users without being detection social engineering techniques, for instance, phishing email, Attackers send emails containing a phishing link to a malicious website or an attachment that contains malicious programs to target users. Then, attackers deceive target users to install a malicious program and then control the target host to steal sensitive information or cause damage. Phishing and website spoofing is also one of the social engineering techniques. The user who are not have knowledge on website usage are the more vulnerabilities to this kind of attacks.

Phishing is performed by using the social engineering toolkit. Which spoof the link which will redirect the user to the fake website. The spoofed link is placed in the popular pages like Gmail. That the user trust to the link which make to open the link lead to the fake web page. Thus rather than redirecting to the real Web server it redirect to the attacker server.

Phishing can be detected by two software methods are blacklist and machine learning approach. In this paper we are apply the machine learning approaches for detection of phishing.

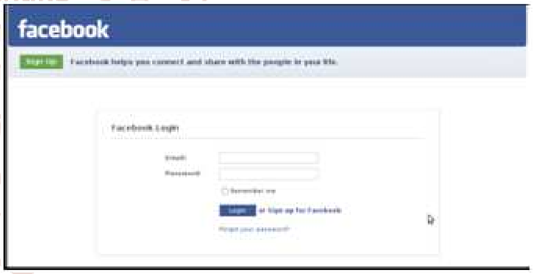

Figure2. Phishing web page [2]

PHISHING TECHNIQUES

The attacker who wanted to steal the sensitive information from the user, he will first create a replica of the website which is exactly look like the real website. The of the fake website is send to the user through the email and make him to Believe as the original website in order to login to the website leads to the loss of sensitive information.

The logo, Templates of the web page is make the user to believe. The Internet growing day by day which lead the many fake website that make the user mislead. The Figure1 and 2 show the difference of the original and fake Web page.

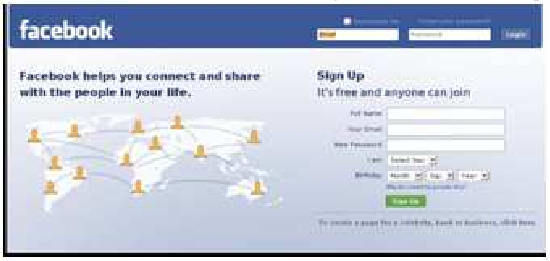

Figure 1.Original Facebook Web page

The above figures2 shows how the phishing is performed by the attackers.

Related Work

Many researchers made analysis on the suspicious URLs in many ways. We review the previous work of URL detection.

Ma et al. [1, 2] compared several batch-based learning algorithms for classifying phishing URLs and showed that the Combination of host-based and lexical features results in the highest classification accuracy. Also they compared the performance of batch-based algorithms to online algorithms when using full features and found that online algorithms, especially Confidence-Weighted (CW), outperform batch-based algorithms.

The work by Garera et al. [3] uses logistic regression over hand-selected features to classify phishing URLs. The features include the presence of red flag keywords in the URL, features based on Google's Page Rank, and Google's

Web page quality guidelines. It is difficult to make a direct comparison with our approach without access to the same URLs and features.

McGrath and Gupta [4] did not construct a classifier, but performs a comparative analysis of phishing and non-phishing URLs with respect to data sets. They compared non phishing URLs drawn from the DMOZ Open Directory Project [5] to Phishing URLs from Phish Tank [6]. The features they analyze include IP addresses, WHOIS thin records containing date and registrar-provided information, geographic information, and lexical features of the URL such as length, character distribution, and presence of predefined brand names [4].

PROBLEM OVERVIEW

URLs are the web link which is used to location the information on the website. Our aim is to build the classification model that detect the phishing websites with analysis of lexical and host-based features of URLs. In this we analyze the different machine learning algorithms by using the python.

DESIGN FLOW

In this we analyze the host based and lexical features of the URLs. First we collect the phishing URLs and the benign URLs then apply the host-based and lexical features to form a data set values. To this data set apply the different classification machine learning algorithms in python. After evaluating the classifier, a classical is selected and implemented in python.

A. COLLECTING URLs

We collected URLs of benign websites from www.alexa.com [9] www.dmoz.org [7] and personal web browser history. The phishing URLs were collected from www.phishtak.com [8]. The data set consists of 17000 phishing URLs and 20000 benign URLs. We obtained Page Rank [10] of 240 benign websites and 240 phishing websites by checking Page Rank individually at PR Checker [11]. We collected WHO IS [12] information of 240 benign websites and 240 phishing websites.

B. Host based analysis

Host based is nothing but where the phishing websites are hosed, who are they managed and how they can Implemented. The features of phishing websites can be used because they are less hosting centers. The properties of the host that are identified are include as WHOIS, Geographic, Blacklist membership.

C. Lexical analysis

Textual properties of the Lexical features are URLs itself, which are not the content of the web page. Generally URLs are the human readable string which is used that they will locate the server's location of the website. That translate the machine for the process.

URLs standard syntax as following:

<Protocol>://<hostname><path>

EX: http://accounts.google.com

Here the <protocol> which refers the network protocol that used to fetch the request. They are different types of protocols are available some of them are http, https, ftp, smtp and so on.

<hostname> which refers the web server in the Internet and sometime as ip address which machine readable but mostly it Will be in the human readable domain name.

<Path> which the path of the file in the local computer. The path token consists of the different types of Delimited, slashes, comma, dots which shows the how the site is organized.

The Methodology used is to extract the lexical features of the URLs. Here we collecting the different URLs of the www.alexa.com, www.dmoz.org and put them in the data set in the format of the CSV or the excel file and load the data set into the R studio. Here we take the benign URLs into the R. Now we set a Decision vector 0 and 1. It will compare the features and analysis the features of host name, host length, path and make the decision as 0 or 1. By this way it will classify the malicious websites or not. By using the different classification methods of Machine Learning Algorithms.

MACHINE LEARNING ALGORITHMS

They are many classification Machine Learning Algorithm are available in order to classify the malicious websites by using the R, Python or MATLAB.

The Input data is given in the format of the CSV, txt, Excel, XLS and so on. It will compute the Input data and analyze the features of the data and make the decision vector 0 or 1 based on the given features of the data. In this the data set will be split as training data and the testing data. The split of the data is in the 60% train data and 40% of the test data.

The Following are Machine Learning Algorithms

Support Vector machine:

SVM is a Classification Algorithm by the finds the hyper plan to increase the margin between two classes. The vectors in the hyper plan are the support vectors.

Logistic Regression:

It is a classification Algorithm by which that the y variable is binary categorical. It has the two values 0 or 1.

Decision Tree:

It is also the classification Algorithm in which the target value will be depend on the various available data.

These are the some of the classification algorithms in which the practical Implementation can be performed in order to check the accuracy of the detection phishing.

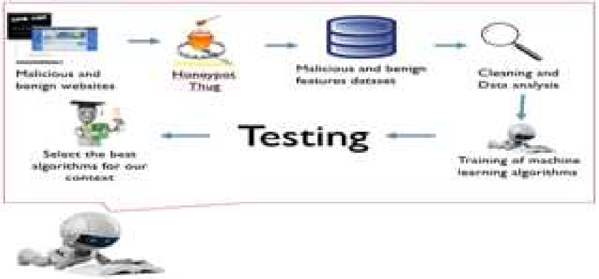
Malicious and benign websites
Honeypot Thug
Malicious and benign features dataset
Cleaning and Data analysis
Training of machine learning algorithms
Testing
Select the best algorithms for our context

Figure 3: The Flow Diagram of the Phishing Detection

Experimental results

In order to evaluate the Approach using different Machine Learning Algorithm. First we are collected the benign and the malicious website URLs. We loaded the data set to the R environment. We can do this by using R, python, WEKA. In this performing by using R. After Load the data the next process is the preprocessing like the cleaning the data, removing the noise in the data. Which will improve the performance of the model. After the preprocessing built the mode for the prediction of the model. In the give data set we are split the data set for training and the testing data. Then we train the machine by using the data set features are trained to the machine. By using the training data set we test the test data and calculate the accuracy of the analysis using the confusion matrix. We calculate the different accuracy of the algorithms with the train and the test data using the confusion matrix.

Our main finding are:

- Phishing URLs and the domain name having the different URLs Lengths when compared the URLs and Domain names in the Internet.
- Many URLs contains the brands they tagged
- Find the website malicious or not
- Different accuracy rate detection
- Lexical Analysis
- Host based Analysis

Figure 4 show the histogram. Which the target value is type and the count are plotted. By the histogram shows the target values are how computed by the count values. The values are computed based on the data that we are loaded.

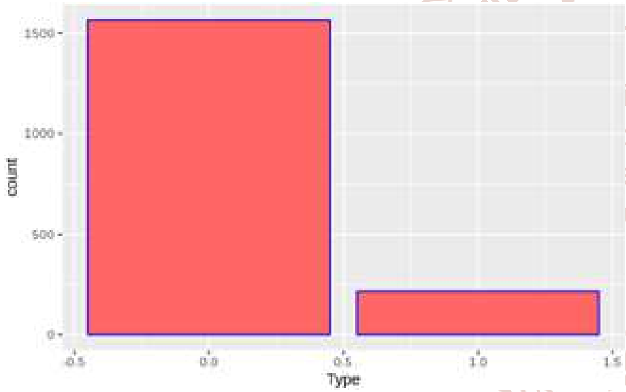

Figure 4 Type Vs. Count

Figure 5 show that the histogram. Which the URLs Length and the count are shown on the graph. By this we can understand the way of the count and URLS Length computing.

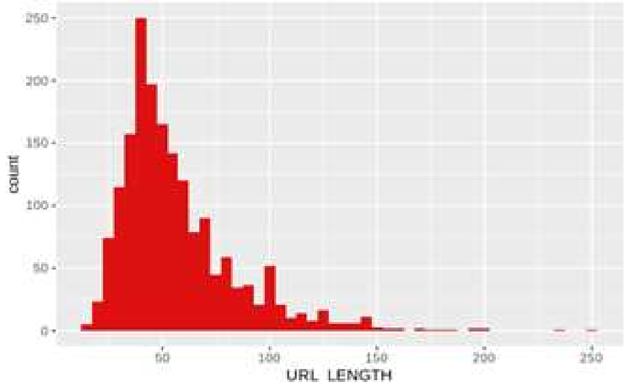

Figure 5 URLs Length Vs. Count

The Following Table 1 show the Different accuracy rates of detection of phishing using machine learning algorithms.

Classifiers	Accuracy(%)	Error Rate(%)
SVM	90	10
Logistic Regression	80	20
Decision Tree	94	6

Table 1 Different Classification Algorithm Accuracy

The experimental result of analysis of the Input data which the malicious and benign are show by the table with different accuracy rates.

The accuracy rate of the SVM are 90% which gives the error rate of 10%. Where Logistic regression are which the error rate is. And the Decision Tree is which the error rate is .

Among the algorithms the highest accuracy rate is in which the error rate is less. That means the mode will predict the detection of phishing is good. It predict the accuracy by the train data with test data using the confusion matrix.

Conclusion AND FUTURE work

In This Paper presented that the Detection of phishing websites using machine learning approaches. By taking the different features of the URLS, lexical analysis, host based analysis, model is built to detecting phishing or not in R and calculated accuracy rate of different algorithm. This approaches will detect the phishing websites and reduces the Social Engineering attacks. In order to improve the performance of the detection with the future work is carry out by using the advanced algorithms and apply the Deep Learning; Neural Networks are applied to improve the detection accuracy.

References

[1] J. Ma, L. K. Saul, S. Savage and G. M. Voelker, "Learning to Detect Phishing URLs", AC CM Transactions on Intelligent Systems and Technology, Vol. V 2, No. 3, Article 30, Publication date: April 2011

[2] J. Ma, L. K. Saul, S. Savage and G. M. Voelker," Beyond Blacklists: Learning to Detect Phishing Web Sites from Suspicious URLs", Proc. of SIGKDD '09.

[3] Garera S., Provos N., Chew M., Rubin A. D., "A Framework for Detection and measurement of phishing attacks", In Proceedings of the ACM Workshop on Rapid Malcode (WORM), Alexandria, VA.

[4] D. K. McGrath, M. Gupta, " Behind Phishing: An Examination of Phisher Modi Operandi", In Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)

[5] DMOZ Open Directory Project. http://www.dmoz.org.

[6] Phish Tank. http://www.phish htank.com.

[7] The Web Information Company, www.alexa.com.

[8] WHOIS look up, www.whoi is.net, www.whois.com

[9] Urcuqui, C., Navarro, A., Osorio, J., & Garcıa, M. (2017). Machine Learning Classifiers to Detect Malicious Websites. CEUR Workshop Proceedings. Vol 1950, 14-17.

[10] I. Rogers, "Google Page Rank – White Paper".

[11] http://www.sirgroane.net/google-page-rank/PR Checker, http://www.prchecker.info/check_page_rank.php

[12] WHOIS look up, www.whois.net, www.whois