**Malicious Web Detection using Machine Learning Techniques**

Prashant Kumar 2018187

Shashank Kumar 2018228

Sumit Kumar 2018250

**ABSTRACT**

*The primitive usage of URL (Uniform Resource Locator) is to use as a Web Address. However, some URLs can also be used to host unsolicited content that can potentially result in cyberattacks. The inability of the end user system to detect and remove the malicious URLs can put the legitimate user in vulnerable condition. Furthermore, usage of malicious URLs may lead to illegitimate access to the user data by adversary. The main motive for malicious URL detection is that they provide an attack surface to the adversary. It is vital to counter these activities via some new methodology.*

*In literature, there have been many filtering mechanisms to detect the malicious URLs. Some of them are Black-Listing, Heuristic Classification etc.*

*These traditional mechanisms rely on keyword matching and URL syntax matching. Therefore, these conventional mechanisms cannot effectively deal with the everevolving technologies and web access techniques. Furthermore, these approaches also fall short in detecting the modern URLs such as short URLs, dark web URLs. In this paper, we propose a novel classification method to address the challenges faced by the traditional mechanisms in malicious URL detection.*

*The proposed classification model is built on sophisticated machine learning methods that not only takes care about the syntactical nature of the URL, but also the semantic and lexical meaning of these dynamically changing URLs. The proposed approach is expected to outperform the existing techniques*

**Keywords:** ***Malicious Phishing attack, Machine learning, Malicious attack***

**1.INTRODUCTION**

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

The general method to detect phishing websites by updating blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as “blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack.

Heuristic based detection which includes characteristics that are found to exist in phishing attacks in reality and can detect zero-hour phishing attack, but the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high.

To overcome the drawbacks of blacklist and heuristics based method, many security researchers now focused on machine learning techniques. Machine learning technology consists of a many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

**2.FEATURE EXTRACTION**

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

**Address Bar Based Features:**

* **Domain of URL:** We are just extracting the domain present in the URL to properly verify that the url is legitimate or not.
* **Presence of IP address in URL:** If IP address present in URL, then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.
* **Presence of @ symbol in URL**: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.
* **Length of URL**: Average length of the benign URLs is found to be a 25, If URL’s length is greater than 25 then the feature is set to 1 else to 0.
* **URL redirection:** If “//” present in URL path then feature is set to 1 else to 0. The existence of “//” within the URL path means that the user will be redirected to another website.
* **HTTPS token in URL:** If HTTPS token present in URL, then the feature is set to 1 else to 0. Phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, [http://https-www-](http://https-www-/) paypal-it-mpp-home.soft-hair.com .
* **URL Shortening Services “TinyURL”:** TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0
* **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is [http://www.onlineamazon.com](http://www.onlineamazon.com/) but phisher can create another fake website like [http://www.online-amazon.com](http://www.online-amazon.com/) to confuse the innocent users.

**HTML and JavaScript based Features**

* **IFrame Redirection:** IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the “frameBorder” attribute which causes the browser to render a visual delineation.

If the iframe is empty or repsonse is not found then, the value assigned to this feature is 1 (phishing) or else 0 (legitimate).

* **Status Bar Customization:** Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the “onMouseOver” event, and check if it makes any changes on the status bar. If the response is empty or onmouseover is found then, the value assigned to this feature is 1 (phishing) or else 0 (legitimate).

**3. MACHINE LEARNING ALGORITHM**

Three machine learning classification model Linear Regression has been selected to detect phishing websites.

**3.1 Linear Regression Algorithm**

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

We are Using Linear regression for classification.

Applying linear regression for classification is not an absurd idea but logistic regression or other classification methods are preferred over linear regression.

You can apply linear regression for classification by assigning a threshold, given below is an example from an online course by Andrew NG where he fitted a line to the data set and used .5 as threshold for classification.

**4. RESULT**

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set 80 :20 ratios respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers.

After creating a Linear Regression model it should be trained using the dataset. After that, by evaluating all the modules the accuracy can be found then Concluding by comparing the accuracy of the model with high performance.

**REFERENCES**

1. Gunter Ollmann, “The Phishing Guide Understanding & Preventing Phishing Attacks”, IBMInternet Security Systems, 2007.
2. https://resources.infosecinstitute.com/category/enterprise

/phishing/the-phishing-landscape/phishing-data-attack- statistics/#gref

1. Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
2. Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016
3. <http://dataaspirant.com/2017/01/30/how-decision-tree-> algorithm-works/
4. <http://dataaspirant.com/2017/05/22/random-forest-> algorithm-machine-learning/
5. https:/[/www](http://www.kdnuggets.com/2016/07/support-vector-).[kdnuggets.com/2016/07/support-vector-](http://www.kdnuggets.com/2016/07/support-vector-) machines-simple-explanation.html
6. [www.alexa.com](http://www.alexa.com/)
7. [www.phishtank.com](http://www.phishtank.com/)