Machine Learning Framework for Detecting Phishing Websites

Jaya Siva Shankaran G
UG Scholar, Department of IT, Velammal Engineering College,Chennai
Email: ajaysiva01@gmail.com
Gowtham Chand Narrayanan S A
UG Scholar, Department of IT, Velammal Engineering College,Chennai
Email: sagowtham.android@gmail.com
Bala Kumar S
UG Scholar, Department of IT, Velammal Engineering College,Chennai
Email: balakumar.1997@gmail.com
Vijayan A
Assistant Professor, Department of IT, Velammal Engineering College, Chennai.
Email: vijay.avm@gmail.com

----------ABSTRACT----------

Phishing costs Internet users billions of dollars per year. It refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting Internet users.Phishers use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords. Phishing is an example of social engineering techniques being used to deceive users. Users are often lured by communications purporting to be from trusted parties such as social web sites, auction sites, banks, online payment processorsor IT administrators. Typically carried out by email spoofing or instant messaging, it often directs users to enter personal information at a fake website, the look and feel of which are identical to the legitimate site.

Keywords - **Phishing, machine learning**

I. INTRODUCTION

Social engineering attack is a common security threat used to reveal private and confidential information by simply tricking the users without being detected. The main purpose of this attack is to gain sensitive information such as username, password and account numbers. According to, phishing or web spoofing technique is one example of social engineering attack. Phishing attack may appear in many types of communication forms such as messaging, SMS, VOIP and fraudster emails. Users commonly have many user accounts on various websites including social network, email and also accounts for banking. Therefore, the innocent web users are the most vulnerable targets towards this attack since the fact that most people are unaware of their valuable information, which helps to make this attack successful. Typically phishing attack exploits the social engineering to lure the victim through sending a spoofed link by redirecting the victim to a fake web page. The spoofed link is placed on the popular web pages or sent via email to the victim. The fake webpage is created similar to the legitimate webpage. Thus, rather than directing the victim request to the real web server, it will be directed to the attacker server. The current solutions of antivirus, firewall and designated software do not fully prevent the web spoofing attack. The implementation of Secure Socket Layer (SSL) and digital certificate (CA) also does not protect the web user against such attack. In web spoofing attack, the attacker diverts the request to fake web server. In fact, a certain type of SSL and CA can be forged while everything appears to belegitimate.

II. PROPOSED SYSTEM

This section describes the proposed model of phishing attack detection. The proposed model focuses on identifying the phishing attack based on checking phishing websites features, Blacklist and WHOIS database. According to few selected features can be used to differentiate between legitimate and spoofed web pages. These selected features are many such as URLs, domain identity, security & encryption, source code, page style and contents, web address bar and social human factor. This study focuses only on URLs and domain name features. Features of URLs and domain names are checked using several criteria such as IP Address, long URL address, adding a prefix or suffix, redirecting using the symbol "//", and URLs having the symbol "@".These features are inspected using a set of rules in order to distinguish URLs of phishing webpages from the URLs of legitimate websites.

III. URL BASED

3.1 Using the IPAddress

If an IP address is used as an alternative of the domain name in the URL, such as "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal sensitive

information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/payp al.ca/index.html".
Rule: IF The Domain Part has an IP Address → Phishing Otherwise→ Legitimate

3.2 Long URL to Hide the Suspicious Part Phishers can use long URL to hide the doubtful part in the address bar

For example
http://federmacedoadv.com.br/3f/aze/a
b51e2e319e51502f416dbe46b773a5e/?
cmd=_home&dispatch=11004d58f5b7
4f8dc1e7c2e8dd4105e811004d58f5b7
4f8dc1e7c2e8dd4105e8@phishing.we bsite.html

To ensure the accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL function that might be used for this purpose is the "mailto:" function.

Rule: IF Using ""mail()\" or \"mailto:\" Function to Submit User Information" → PhishingOtherwise
→ Legitimate

3.4 Black list based

A Blacklist is created in the proposed model in which the website detected as phishing is saved for the future use a to keep a track record and data of the phishing website this can be useful in analyzing the phishing website to increase the efficiency of thesystem.

classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total datasetsize.

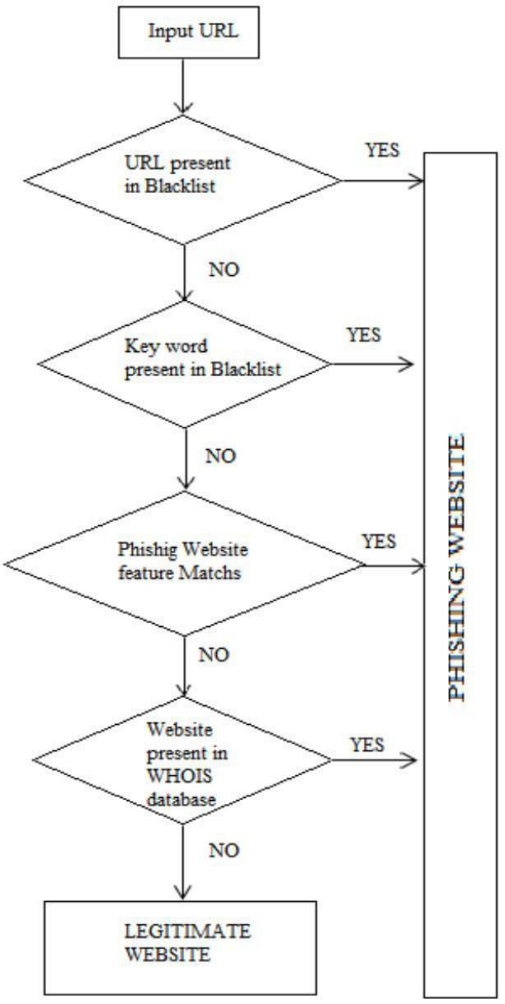
Rule: IF URLlength is ≤ 75→legitimate

Otherwise→Phishing

We have been able to update this feature rule by using a method based on frequency and thus improving upon itsaccuracy.

3.3 Submitting Information toEmail

Web form allows a user to submit his personal sensitive information that is directed to some server for processing. A phisher might redirect the user's information to his personal email. To that end, a server-side script language might be used such as "mail()" function in PHP. One moreclient-side



3.5 WHOISDatabase

The life of phishing site is very short, therefore; this DNS information may not be available after some time. If the DNS record is not available anywhere then the website is phishing. If the domain name of the suspicious webpage is not match with the WHOIS database record, then webpage considers asphishing.

The process of registration was established in RFC 920. WHOIS was standardized in the early 1980s to look up domains, people and other resources related to domain and number registrations. As all registration was done by one organization at that time, one centralized server was used for WHOIS queries. This made looking up such information very easy.

IV. CONCLUSION

The most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should also be trained not to blindly follow the links to websites where they have to enter their sensitive information. It is essential to check the URL before entering the website. In Future System can upgrade to automatic Detect the web page and the compatibility of the Application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. PhishChecker application also can be upgraded into the web phone application in detecting phishing on the mobileplatform.

REFERENCES

[1] https://en.wikipedia.org/wiki/Machine_learning
[2] https://www.sas.com/en_us/insights/analytics/machine-learning. html
[3] https://towardsdatascience.com/phishing-domain-detection-with -ml-5be9c99293e5
[4] https://www.researchgate.net/publication/269032183_Detection_of_phishing_URLs_using_machine_learning_techniques
[5] https://www.sciencedirect.com/book/9780128029275/a-machine-l earning-approach-to-phishing-d etection-and-defense
[6] https://www.hindawi.com/journals/wcmc/2018/4678746/