International Journal of Research in Engineering, Science and Management
Volume 4, Issue 7, July 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Phishing Detection: Machine Learning Implementation

Bhandary Prajwal Gopal Krishna[1], D. S. Rajesh[2], K. Prashanth Kumar[3*]
[1,3]*B.E. Student, Department of Computer Science and Engineering, Srinivas Institute of Technology, Mangalore, India*
[2]*Associate Professor, Department of Computer Science & Engineering, Srinivas Inst. of Technology, Mangalore, India*

***Abstract*: Phishing attack is easiest way to obtain delicate information from innocent users. Aim of the phishers is to acquire crucial information like username, password and bank account details. Cyber security persons are now looking for reliable and steady detection techniques for phishing websites detection. The project uses machine-learning technology for detection of phishing URLs by extracting and analyzing various features of valid and phishing URLs. Decision Tree, Random forest algorithm is used to detect phishing websites. Aim of the project is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.**

***Keywords*: Phishing detection, Decision Tree, Machine learning.**

1. Introduction

In recent times, Phishing becomes an important area of concern for security researchers because it is not difficult to develop the phishing website, which looks so close to legitimate website. Experts can identify phishing websites but not all the users can identify the phishing website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account details and personal information. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. As per Index Report released in 2020, it was estimated that the annual worldwide impact of phishing could be as high as $1.6 million. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to reduce them but it is very important to enhance phishing detection techniques.

Social engineering and ransomware based are the phishing attacks, which contain malicious websites that are attached to E-mail, SMS or other interaction method to dupe people. Cybercrime or fraud uses spam email as a tool. Email spoofing or instant messaging carried out phishing. These emails and messages contain a URL link redirects users to another harmful website. It often directs users to enter personal information or sensitive information i.e., password, credits card details at a fake website, which look like legitimate site.

2. Literature Survey

Phishing is a social engineering attack that targets and exploiting the weakness found in the system at the user's end. This paper proposes the Agile Unified Process (AUP) to detect duplicate websites that can potentially collect sensitive information about the user. The system checks the blacklisted sites in dataset and learns the patterns followed by the phishing websites and applies it to further given inputs. The system sends a pop-up and an e-mail notification to the user, if the user clicks on a phishing link and redirects to the site if it is a safe website. This system does not support real time detection of phishing sites; user has to supply the website link to the system developed with Microsoft Visual Studio 2010 Ultimate and MySQL stocks up data and to implement database in this system.

Phishing costs Internet user's lots of money. It refers to misusing weakness on the user side, which is vulnerable to such attacks. The basic ideology of the proposed solution is use to all the three-hybrid solution blacklist and whitelist, heuristics and visual similarity. The proposed system carries out a set of procedures before giving out the results. First, it tracks all "http" traffic of client system by creating a browser extension. Then compare domain of each URL with the white list of trusted domains and the blacklist of illegitimate domains. Further various characters in the URL is considered like number of '@', number of '-' and many more. Next approach is to extract and compare CSS of doubtful URL and compare it with the CSS of each of the legitimate domains in queue. This method will look into visual based features of the phished websites and machine-learning classifiers such as decision tree, logistic regression, and random forest are applied to the collected data, and a score is generated. The match score and similarity score is evaluated. If the score is greater than threshold, then the URL marked as phishing and blocked. This approach provides a three level security block.

Phishing is a dangerous effort to steal private data from users like address, Aadhar number, PAN card details, credit or debit card details, bank account details, personal details etc. The various types of phishing attacks like spoofing, instant spam spoofing, Hosts file poisoning, malware-based phishing, Man-in-the middle, session hijacking, DNS based phishing, deceptive phishing, key loggers/loggers, Web Trojans, Data theft, Content-injection phishing, Search engine phishing,

*Corresponding author: prashanthkumar51782@gmail.com

Email /Spam, Web based delivery, Link Manipulation, System reconfiguration, Phone phishing, etc. are discussed in the paper. The recent approaches to prevent the attacks like heuristics approach, blacklist approach, fuzzy rule-based approach, machine learning approach etc. are also discussed and finally filtering all detection techniques based on accuracy and performance proposed a framework to detect and prevent phishing attacks. A combination of supervised and unsupervised machine learning techniques is used to detect malicious attacks.

3. Implementation

A. Feature Extraction

The feature extraction process is done from the URLs and corresponding binary values are given indicating whether the website is a phishing website or not. Below are the features that We can extract for detection of fraud URLs.

1. *IP address in URL:* If IP address present in URL then the feature is set to 1 else set to 0. Most of the legitimate sites do not use IP address in an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to collect sensitive information.
2. *'@' symbol in URL:* If @ symbol present in URL then the feature is set to 1 else set to 0. Hackers add special symbol @ in the URL leads the browser to ignore everything before the at the rate (@) symbol and the real address often follows the "@" symbol.
3. *Prefix or Suffix separated by (-) to domain:* If domain name separated by dash (-) symbol then the feature is set to 1 else to 0. This '-' symbol is rarely used in legitimate URLs. Phishers add hyphen symbol (-) to the domain name, so that users feel that they are dealing with a legitimate Webpage. For example, site is http://www.onlineamazon.com but phisher can develop another fake website like http://www.onlineamazon.com to trick the innocent users.
4. *Length of Host name:* Average length of the benign URLs is found to be a 25, If URL's length is more than 25 then the feature is set to 1 else to 0.
5. *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 fool users.
6. *URL redirection:* If "//" present in URL path then the feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be direct to another website.

B. Random Forest Algorithm

Random Forest is a machine-learning algorithm that belongs to the supervised learning technique; it can be applied for Classification and Regression problems. It is based on Phishing Website Detection using Machine Learning System Implementation the concept of Associative learning, which is a process of combining many different classifiers to solve a complex problem and to improve the efficiency of the model. As the name suggests, Random Forest is a classifier that contains a many number of decision trees on various subsets of the given dataset and takes the average to increase the predictive accuracy of that dataset. Instead of depending on one decision tree, the random forest takes the prediction from each tree and based on the maximum votes of predictions, it decides the final output.

C. Decision Tree

Decision Tree is a supervised learning technique that can be used for classification and Regression problems, but mainly it is preferred for solving Classification problems. It is a tree structured classifier, which internal nodes represent the features of a dataset, branches indicate the decision rules and each leaf node indicates the outcome. In the Decision tree, there are two nodes, one is the Decision Node and other is Leaf Node. Decision nodes are used to make many decision and have different branches, whereas Leaf nodes are the output of those decisions and decisions and do not contain any further branches.

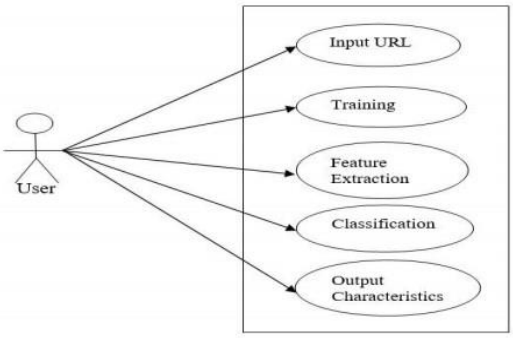

Fig. 1. Use Case diagram

The figure shows use case diagram for the proposed system.

4. Conclusion

The proposed system enables the internet users to have a safe browsing and transactions. Its helps users to save their important private details that should not be leaked. Phishing website detection which is used to detect illegal website using machine learning project. We have extracted the different feature of the URL and decided the given URL is legitimate or not. We have taken into the consideration of the dataset which consists of 10000 URL's consisting 5000 legitimate and 5000 phishing websites. We have used Machine Learning domain and implemented the Random forest and decision tree algorithms which is used because it is one of the most efficient methodology in machine learning which has delivered about 81% accuracy. We used HTML and CSS for front end of the project.

Future work should focus on direct implementation of project to the chrome extension so that as the user clicks on the particular URL and if that URL is phishing site then the user gets a pop up warning message.

References

[1] Mahajan, Rishikesh & Siddavatam, Irfan. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications. 181. 45-47.

[2] Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset.

[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey," IEEE, 2013.

[4] Rishikesh Mahajan (2018) "Phishing Website Detection using Machine Learning Algorithms."

[5] Purvi Pujara, M. B. Chaudhari (2018) "Phishing Website Detection using Machine Learning: A Review."

[6] S. Abu-Nimeh and T. M. Chen. Proliferation and detection of blog spam. Security & Privacy, IEEE, 8(5):42–47, 2010.

[7] Jalil Nourmohammadi Khiarak (2017) "What is Machine Learning."

[8] Tenzin Dakpa, Peter Augustine (2017) "Study of Phishing Attacks and Preventions."

[9] Sadia Afroz, Rachel Greenstadt (2017) "PhishZoo: Detecting Phishing Websites by Looking at them."

[10] Srushti Patil, and Sudhir Dhage, "A Methodical Overview On Phishing Detection Along with an Organized Way to Construct an Anti-Phishing Framework", 2019 5th International Conference on Advanced Computing & Communication System (ICACCS), pp. 1-6.