DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

DATLA BHAVANI SRAVYA 19H51A1258

[1]Mrs.T. Lavanya ,[2]B SAI RAM,[3]P SAI&[4]D. BHAVANI SRAVYA

[1]Assistant Professor, Department of Information Technology, CMR College of Engineering & Technology

[2, 3, 4] B-Tech, Department of Information Technology, CMR College of Engineering &

Abstract

Phishing website is one of the internet security problems that target the human vulnerabilities rather than software vulnerabilities. It can be described as the process of attracting online users to obtain their sensitive information such as usernames and passwords. In this paper, we offer an intelligent system for detecting phishing websites. The system acts as an additional functionality to an internet browser as an extension that automatically notifies the user when it detects a phishing website. The system is based on a machine learning method, particularly supervised learning. We have selected the Support Vector Machine and LightGBM Algorithm due to its good performance in classification. Our focus is to pursue a higher performance classifier by studying the features of phishing website and choose the better combination of them to train the classifier. As a result, our prototype will have accuracy of 96.8% with combination of 87 features.

INTRODUCTION:

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details. Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents. Detecting Phishing Domains is a classification problem, so it means we need labeled data which has samples as phish domains and legitimate domains in the training phase. 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.

Fig 1 Phishing Demonstrating image

OBJECTIVE

Statistics have shown that the number of phishing attacks keeps increasing, which presents a security risk to the user information according to the Anti-Phishing Working Group(APWG) and recorded phishing attacks by Kaspersky Lab, which stated that it has increased by 47.48% from all of the phishing attacks that have been detected during 2016. Machine learning is a field of computer science, which is also a branch of artificial intelligence (AI) that performs tasks and is capable of learning or acting in an intelligent way. It has two different types of learning: Supervised learning and unsupervised learning. Supervised learning is based on training a model by giving it a set of measured features of data associated with a target label related to these data, And once the model is trained it can generate a new target label with unknown data. On the other hand, unsupervised learning is based on generating new data without giving any target label in the training process. In this project, the focus will be on the features combination that we get from Random

Forest (RF) technique, as it has high accuracy, is relatively robust, and has a good performance

IMPLEMENTATION

A Content-Based Approach to Detecting Phishing Web Sites using PHP & MYSQL. It is an implementation of a project our system will crawl the original site of bank and it will retrieve all URL's, location of bank's server and whois information. If user get any email with phishing attack link. Then our system will take that url as input and crawl the link, retrieve all url's and system will compare these url's with original banks url database, try to find url's are similar or not. Then system will find location of Phishing link URL and compare location with original banks location. After that system will find out Whois information of URL. System will analyze the information and show the results to the user

Flow Diagram

Fig 2 Flow Diagram of content based approach

PROPOSED SYSTEM

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colorful information at any time, from anywhere around the world. Phishing is the act of transferring an indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, and date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe. Individualizes and associations have lost a huge sum of plutocrat and private information through Phishing attacks. Detecting the phishing attack proves to be a challenging task. Tis attack may take a sophisticated form and fool even the savviest users: such as substituting a few characters of the URL with alike Unicode characters. By cons, it can come in sloppy forms, as the use of an IP address instead of the domain name. Nonetheless, in the literature, several works tackled the phishing attack detection

challenge while using artificial intelligence and data mining techniques [5–9] achieving some satisfying recognition rate peaking at 99.62%. However those systems are not optimal to smartphones and other embed devices because of their complex computing and their high battery usage, since they require as entry complete HTML pages or at least HTML links, tags and webpage JavaScript elements some of those systems uses image processing to achieve the recognition. Opposite to our recognition system since it is a less greedy in terms of CPU and memory unlike other proposed systems as it needs only six features completely extracted from the URL as input. In this paper, after a summary of this Feld key researches, we will detail the characteristics of the URL that our system uses to do the recognition. Otherwise we will describe our recognition system, next in the practical part we will test the proposed system while presenting the results obtained. Last DETECTION OF PHISING WEBSITES USING MACHINE LEARNING CMRCET B. Tech (IT) Page No 28 But not least we will enumerate the implications and advantages that our system brings as a solution to the phishing attack.

Support Vector Machine (Svm):

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyper plane. SVM chooses the extreme points/vectors that help in creating the hyper plane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane:

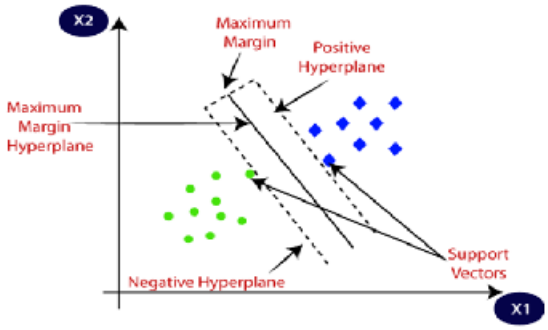

Fig 3. SVM Demonstration

RESULTS:

In below screen DJANGO webserver started and now open browser and enter URL http://127.0.0.1:8000/index.html and press enter key to get below output.

Fig- 4 Django webserver

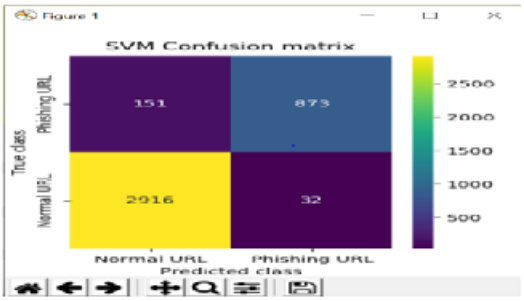

Fig- 5 Confusion Matrix of SVM Algorithm

In above screen we can see SVM confusion matrix where x-axis represents predicted class and y-axis represents TRUE class and we can see SVM predict 2977 records DETECTION OF PHISING WEBSITES USING MACHINE LEARNING CMRCET B. Tech (IT) Page No 51 correctly as NORMAL and only 145 are incorrect prediction and it predict 824 records as PHISHING URL and only 26 are incorrect prediction and now close above graph to get below output.

CONCLUSION

This proposed model deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. The SVM (support vector machine), Lightgbm (Light Gradient Boosting Method) machine learning algorithms are used to detect phishing websites. The main objective of this project is to detect any URL into phishing URL or the legitimate URL. We are mainly using the supervised machine learning techniques out of which classification algorithms are used to classify the URLs to be genuine or fake, finally we are going to narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

FUTURE EXTENSIONS:

In future we would like to extend this proposed model in the following ways:-

- This model can be further developed in such a way that it could detect the URL, whether it is phishing or legitimate also it defines which kind of phishing and the amount of phishing.
- This model can be embedded with other social media like Facebook, Twitter etc
- For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

REFERENCES

[1] Matthew Dunlop, Stephen Groat, David Shelly (2010) " GoldPhish: Using Images for Content-Based Phishing Analysis”

[2] Rishikesh Mahajan (2018) “Phishing Website Detection using Machine Learning Algorithms”

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

[4] David G. Dobolyi, Ahmed Abbasi (2016) “PhishMonger: A Free and Open Source Public Archive of Real-World Phishing Websites”

[5] Satish.S, Suresh Babu.K (2013) “Phishing Websites Detection Based On Web Source Code and URL in the Webpage”

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

[7] Satish.S, Suresh Babu.K (2013) “Phishing Websites Detection Based On Web Source Code and URL in the Webpage”

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

[9] Ping Yi (2018) “Web Phishing Detection Using a Deep Learning Framework”

[10] Revathy, G., Gurumoorthi, E., Sasikala, C., & Latha, T. M. (2023, June). Training superbot with learning automata and multi kernel SVM. In AIP Conference Proceedings (Vol. 2782, No. 1). AIP Publishing.

[11] Poongodai, P. Singh, K. Soujanya and R. Muthukumar, "A Novel Decision Support System for the Prognosis of Parkinson Disease," *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Dharan, Nepal, 2022, pp. 1083-1089, doi: 10.1109/I-SMAC55078.2022.9986506.

[12] Madhavi Latha, C., Soujanya, K.L.S. (2021). Secure IoT Framework Through FSIE Approach. In: Singh, P.K., Veselov, G., Vyatkin, V., Pljonkin, A., Dodero, J.M., Kumar, Y. (eds) Futuristic Trends in Network and Communication Technologies. FTNCT 2020. Communications in Computer and Information Science, vol 1395. Springer, Singapore. https://doi.org/10.1007/978-981-16-1480-4_2

[13] Patel, P., Sivaiah, B., Patel, R., 2022, Relevance of Frequent Pattern (FP)-Growth-Based Association Rules on Liver Diseases, Lecture Notes in Networks and Systems, 10.1007/978-981-19-0901-6_58

[14] Behera, A.K., Panda, M., Nayak, S.C., Dash, C.S.K., 2022, An Artificial Electric Field Algorithm and Artificial Neural Network-Based Hybrid Model for Software Reliability Prediction, Smart

Innovation, Systems and Technologies, 10.1007/978-981-16-9447-9_21

[15] Lu, Y., Khan, M., Ansari, M.D., 2022, Face recognition algorithm based on stack denoising and self-encoding LBP, Journal of Intelligent Systems, 10.1515/jisys-2022-0011