IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Phishing Website Detection Using Data Mining Classification And Prediction In Machine Learning Algorithms

Rajitha M

Research Scholar, Department of Computer Science,Sree Narayana Guru College, K.G.Chavadi, Coimbatore - 641 105, Tamil Nadu, India.

Dr. R. Priya

Associate Professor & Head,Department of Computer Science

Sree Narayana Guru College, K.G.Chavadi, Coimbatore - 641 105, Tamil Nadu, India.

Abstract—The article discusses the reviews and issues of effective measures taken to enforce cyber security. The phishing website has evolved as one of the main cyber security threat in recent times. The phishing websites,malware, ransomware, host spam drive-by exploits. A phishing website almost look-alike a very popular website and lure an unsuspecting user to fall victim to the trap. The victim of the scams incurs a monetary loss, loss of confidential information and loss of identity. Hence, it is imperative to find a solution that could eliminate such security threats in a timely manner. Typically, the detection of phishing websites is done using blacklists. There are many popular websites which host a list of blacklisted websites, e. g. PhisTank. The blacklisting technique lack in two aspects, blacklists might not be exhaustive and can't detect a newly generated phishing website. In recent times machine learning techniques have been used in the classification,prediction and detection of phishing websites. In this paper comparison of different machine learning techniques for the phishing URL classification task has performed and achieved. In recent years, with the increasing usage of mobile services, there is a growing trend to move almost real-world tasks to the cyber world. Although this makes easy our daily lives, it also brings many security threats due to the anonymous structure of the Internet.This system compares and predict which algorithm have higher accuracy to predict a given URL is cyber threat or not.

Keywords—cyber threats,Phishing,Data mining,machine learning

I. INTRODUCTION

Phishing is a form type of a cyber security attack where an attacker gains control on sensitive website user accounts by learning sensitive information such as login credentials, credit card information by sending a malicious URL in email or masquerading as a reputable person in email or through other communication channels. The person receives a message from well known contacts, persons or organizations and looks like very much genuine in its appeal. The message received might contain malicious links, software that might target the user computer or the malicious link may direct the user to some fake website which is similar in look and feel of a popular website, and the user may reveal his personal information e.g. credit card details, login and password credentials and other sensitive informations like account credentials etc. Phishing is the most popular type of cyber security attack and very common among the attackers. Phishing attacks are comparatively easy as most of the victims are not well aware of the issues and implications in the web applications and computer networks and its technologies.This lead to getting tricked or spoofed easily.

II.CONCEPTUAL DESIGN

Data Mining: Data mining allows users to sift through the enormous amount of information available in data warehouses and extracting the necessary knowledge which can be used for the decision making purpose. Data Mining allows backend processors to analyze data from many different dimensions, categories it and summarize the relationships identified. Most of the companies in the retail business are using Data Mining in one-way or the other. Jyothi Pillai in "User centric approach to item set utility mining in Market Basket Analysis" describes Business intelligence is information about a company's past performance that is used to help predict the company's future performance. Association rule mining is one of the technique used an in data mining research where the aim is to find interesting correlations among sets of items in databases.

Cyber Threat: Cyber Threat or cyber security threat is a malicious act that seeks to damage data, steal data, or disrupt digital life in general.Cyber threats include computer worms,viruses, Denial of Service (DoS) attacks etc. A successful cyber attack that primarily aims to gain unauthorized access, damage, disrupt, or steal an information technology asset, computer network, intellectual property or any other form of confidential or sensitive data. Cyber threats can occur within an organization by trusted users or from remote locations by unknown sources.

Phishing: Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers. It occurs when an attacker, behaves like a trusted entity, dupes a victim into opening an email, instant message, or text message. The recipient is then tricked into clicking a malicious link, which can lead to the installation of malware, the freezing of the system as part of a ransomware attack or the revealing of sensitive information.An attack can have non recoverable results. For individuals, this includes unauthorized purchases, the stealing of funds, or identify theft.

Moreover, phishing is sometimes used to gain a foothold in governmental or corporate networks as a part of a larger attack, such as an advanced persistent threat event. In this latter scenario, employees are often compromised in order to bypass security perimeters and distribute malware within a closed environment, or gain privileged access to secured data.

An organization succumbing to such an attack typically sustains severe financial losses in addition to declining market share, reputation, and consumer trust. Depending on the scope, a phishing attempt might lead into a security incident from which a business will have a difficult time to recover.

III. APPROACH

Below mentioned are the steps involved in the completion of this project: Collect dataset containing phishing and legitimate websites from the open source platforms.The next step is to write code to extract the proposed or required features from the URL database. Study and analyze and then preprocess the dataset by using varoius EDA techniques. Divide the whole dataset into two categories such as training and testing sets. Run selected machine learning and deep neural network algorithms like SVM, Random Forest, Knn,XGBoost on the dataset. The next step is to write a code for displaying the evaluated result considering accuracy metrics. Compare the obtained results for trained models and specify which is the better.

Presence of IP address in URL: If IP address is 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 sometimes indicates that the attacker is trying to steal confidential or sensitive information.

Presence of symbol @ in URL: If the symbol @ present in URL then the feature is set to 1 else set to 0. Phishers mostly add special symbol @ in the URL which leads the browser to ignore everything preceding the symbol "@" and the real address often follows the "@" symbol .

Number of dots in the Hostname: Phishing URLs may contain many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is three. If the number of dots in URLs is greater than three then the feature is set to 1 else to 0.

Prefix or Suffix separated by symbol (-) to domain: If domain name separated by symbol dash (-) then feature is set to 1 else to 0. The symbol dash is rarely used in genuine URLs. Phishers might add dash symbol (-) to the domain name so that users feel that they are dealing with a valid webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

URL redirection**:** If "//" present in the specified URL path then feature is set to 1 else to 0. The presence of "//" symbol within the URL path means that the user will be redirected to another website

Information submission to Email: Phisher might use "mail()" or "mailto:" procedures or functions to redirect the user's information to his personal email. If such procedures are present in the URL then feature is set to 1 else to 0.

URL Shortening Services such as "TinyURL": TinyURL services are used by phisher to hide long phishing URL by making it short. The major goal is to redirect user to

phishing websites. If the URL is formed using shortening services (like bit.ly) then the feature is set to 1 else 0
Length of Host name: Recommended 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
Presence of sensitive words in URL: majority of Phishing sites uses sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Following are the frequently used words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc;

IV.DATA COLLECTION

Legitimate URLs are collected from the dataset provided by University of New Brunswick. From the collection of URL, 5000 URLs are randomly picked. Phishing URLs are collected from opensource data set service called PhishTank . This data set service provide a set of phishing URLs in multiple formats like csv, json etc. which gets updated hourly. From the obtained collection, 5000 URLs are randomly picked.

V.MACHINE LEARNING MODELS

This is a supervised machine learning task. There are two major classification of supervised machine learning problems, such as classification and regression. This selected data set comes under the classification problem, as the input URL is classified as phishing (1) or legitimate (0) URLs. The machine learning models (classification) considered to train the dataset in this notebook are. Random Forest. XGBoost. K NN • Support Vector Machines.

VI .FINDINGS

- Phishing sites can be identified and precautions can be taken
- It is very useful to identify which algorithm gives accurate results in prediction
- The performance level of each model is measured and compared to gain efficient results
- We can use the analyzed data to keep track of vulnerabilities in the system
- Data sets can be elaborated using training in the cases of new threats happening
- We can make use of the efficient algorithm to predict whether a url is malicious or not

VII. CONCLUSIONS AND FUTURE SCOPE

To conclude that it is possible to use datamining techniques like classification and prediction as well as machine learning algorithms to predict a given url is a cyberthreat or not. In the recent years, with the increasing use of mobile and network technologies, there comes a growing trend to move almost real-world tasks or operations to the cyber world. Albeit this makes easy our daily tasks, it also brings many security implications due to the anonymous structure of the web. antivirus programs and firewall installed in systems can prevent most of the attacks. Even though experienced attackers target on the weakness of the computer users by trying to phish them using fake web pages. These pages looks like some popular banking, social media, e-commerce sites etc. sites to steal some sensitive information such as, user-ids, passwords, bank account, credit card numbers, etc. Phishing detection is a challenging problem.Further development of such a system can be continued by adding statistical tools and machine learning tools to obtain additional information about threats on their own, which will reduce the requirements for data sources and allow you to expand the set of analyzed indicators.

REFERENCES

1. S.Vijaylakshmi, V. Mohan, S. Suresh Raja, "*Mining of users access behavior for frequent sequential pattern from web logs*" International Journal of Database Management System (IJDM) Vol 2, August 2010.
2. Edi Winarko and John F. Roddick, "*Discovering Richer Temporal Association Rules from Interval-Based Data, Data Warehousing and Knowledge Discovery*" 2005., LNCS 3589
3. Shrivastava A.,Sahu R "*Efficient Association Rule Mining for Market Basket Analysis*" Global Journal of e-Business & Knowledge ManagementYear:2007,Volume:3,Issue:1
4. Aggarwal C.C, "*Mining association with the collective strength approach*", Yu,P.S. 2001
5. Jianying Hu, Aleksandra Mojsilovic,"*High-utility pattern mining: A method for discovery of high-utility item sets, Pattern Recognition*" Volume 40, Issue 11, November 2007
6. N.R. Srinivasa Raghavan. "*Data Mining in E-commerce: A Servey*".Sadhana, vol,30, no.2, 2005
7. Mobasher. B,"*Web Usage Mining and Personalization*" Practical Handbook of Internet Computing (ed.) M P Singh (CRC Press), 2004.
8. Jiauei Han , Michele Kamber,Simon Fraser,"*Data mining Concepts and Techniques*" University ISBN 1-55860-489-8-2001

9. Randall S. Sexton, Richard A. and Michael A. "*Predicting Internet/e-commerce use*", Internet Research, vol.5,2002
10. J Hu, A Mojsilovic, "*High-utility pattern mining: A method for discovery of high-utility item sets*", Pattern Recognition 2007
11. https://www.stealthlabs.com/blog/cyber-security-threats-all-you-need-to-know/
12. https://www.microstrategy.com/us/resources/introductory-guides/data-mining-explained
13. https://www.upguard.com/blog/cyber-threathttps://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications
14. https://towardsdatascience.com/association-rules-2-aa9a77241654