IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 1, January 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001016 |

Detection of Phishing Sites using ML

Om More, Shruti Sawant, Payoshni Holey, Atish Sawant, Rohini Agawane

UG Student, Dept. of Computer Engineering, KJCOEMR, Pune, Maharashtra, India

Assistant Professor, Dept. of Computer Engineering, KJCOEMR, Pune, Maharashtra, India

ABSTRACT: 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. [3].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.[3]

KEYWORDS: Detection, Phishing, Legitimate, Algorithms, Blacklists, Phishing Attacks, Machine Learning.

I. INTRODUCTION

Phishing is that the fraudulent plan to obtain sensitive information like username, password and credit card details, often malicious purposes, by disguising as a trustworthy entity in an electronic communication [1]. 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. [3] In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that 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. [3] In this attack, Phisher makes a fake web page by copying contents of the legitimate page, so that a user cannot differentiate between phishing and legitimate sites. Social engineering schemes prey on unwary victims by fooling them into believing they are dealing with a trusted, legitimate party, such as by using deceptive email addresses and email messages. [1] 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.[3]

II. PROPOSED DEFINITIONS

Different kinds of phishing attacks:

It is possible to use machine learning to understand data and build great data products. The project aims to explore this area by showing a use-case of detecting phishing websites using machine learning. [13] Machine Learning (ML) methods, can also be used in application development for information security. Optimization, classification, prediction and decision support system and great benefits can be provided to the person who isresponsible for information security. [3]

Phishing can be done through email phishing scams and spear phishing hence user should be aware of the consequences and should not give their 100 percent trust on common security application. Machine Learning is one of the efficient techniques to detect phishing as it removes drawback of existing approach. [3]This is a field of artificial intelligence and it has ability to learn without explicitly programmed. Various machine learning techniques are Supervised learning, Unsupervised learning and Reinforcement learning.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001016 |

Machine learning types of machine learning techniques are:
1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning

Machine Learning Algorithms:
1. Extreme Learning Machine (ELM): Extreme
Learning Machine (ELM) is a feed-forward artificial neural network (ANN) model with asingle hidden layer. For the ANN to ensure ahigh-performing learning, parameters such as threshold value, weight and activationfunction must have the appropriate values for the data system to be modelled. In gradient-based learning approaches, all of these parameters are changed iteratively for appropriate values.

2. Random Forest Algorithm:
Random Forest (RF) is an ensemble learning classification and regression method suitable to handle problems involving grouping of data into classes. In RF, prediction is achieved using decision trees. During the training phase, a number of decisions trees are constructed (as defined by the programmer) which are then used for the class prediction; this is achieved by considering the voted classes of all the individual trees and the class with the highest vote is considered to be the output.[10]

3. Decision Tree:
A decision tree looks like a flowchart, where each non-leaf node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. C4.5 is used to construct decisiontree through the learning from class-labelled training tuples; it adopts a greedy, top-downrecursive divide-and-conquer, approach to construct decision trees. Attribute selection is akey problem in constructing decision tree, which determines which attribute to be split (i.e., as anon-leaf node of decision tree). The attribute selection measure provides a ranking for eachattribute according to the training tuples. The attribute with the best score for the measure ischosen to be the splitting attribute. In C4.5, gain ratio is used as the attribute selection measure.[11] determines which attribute to be split (i.e., as anon-leaf node of decision tree). The attribute selection measure provides a ranking for eachattribute according to the training tuples. Theattribute with the best score for the measure is chosen to be the splitting attribute. In C4.5, gainratio is used as the attribute selection measure.[11]

4. SVM (Support Vector Machine):
This technique is used in medical for diagnosis of diseases, text recognition, for classification of image and in the other fields. This will partition the data into two categories using fixed rule, quadratic equation and statistic. Separating hyper plane is used for the binary classification of the data and minimizes the space of the margin on the basis of kernel function. This technique is used to find the best solution of the problem. This technique is fails in analysing thebig data. [2]

III. Conclusion and Future Work

The systems varying from data entry to information processing applications can be made through websites. The entered information can be processed; the processed information can be obtained as output. Nowadays, web sites are used in many fields such as scientific, technical, business, education, economy, etc. Because of this intensive use, it can be also used as a tool by hackers for malicious purposes. One of the malicious purposes emerges as a phishing attack. Contributions of many researches shows different methods, approaches to detect phished URLs and these methodologies have also been implemented. The purpose of the application is to make a classification for the determination of one of the types of attacks that cyber threats called phishing. The system informs user of phishing URLs by prompting of benign URLs even before goes live on those website which ultimately leads to avoidance of a phishing attack. Extreme Learning Machine will be used for this purpose. In this study, we will use a data set from UCI website.

References

[1] Oza Pranali P, Deepak Upadhyay, Review on Phishing Sites Detection Techniques, IJERT, ISSN: 2278-0181, 04, April-2020
[2] Meenu, Sunila godara, Phishing Detection using Machine Learning Techniques, IJEAT, ISSN: 2249 – 8958, 2, December, 2019

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 10, Issue 1, January 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1001016 |

[3] Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, Siri Chandana Guttha, IJITEE, ISSN: 2278-3075, June 2019
[4] Ankit Kumar Jain and B.B. Gupta EURASIP Journal on Information Security (2016) 2016:9
[5] Joby James, Sandhya L., Ciza Thomas, Detection of Phishing URLs Using Machine Learning Techniques, 2013 International Conference on Control Communication and Computing (ICCC), December 2013
[6] Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen, Detecting Phishing Website Using Machine Learning, 2020 16th IEEE International Colloquium on Signal Processing & its Applications, 28-29 Feb. 2020
[7] Suleiman Y. Yerima, Mohammed K. Alzaylaee, High Accuracy Phishing Detection Based on Convolutional Neural Networks, IEEE Xplore
[8] Megha N, K R Remesh Babu, Elizabeth Sherly, An Intelligent System for Phishing Attack Detection and Prevention, IEEE Xplore ISBN: 978-1-7281-1261-9, 2019 IEEE
[9] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani, Detecting Phishing Websites Using Machine Learning 978-1-7281-0108-8/19/ 2019 IEEE
[10]https://www.hindawi.com/journals/jam/2014/425731/(randomforest)
[11]https://pdfs.semanticscholar.org/41ca/257920b5b5e6c1cf4f4417bb85ac5a875935.pdf
[12]https://archive.ics.uci.edu/ml/index.php
[13]https://www.google.com/url?q=https://towardsdatascience.com/phishing-domain-detection-with-ml5be9c99293e5&sa=D&source=hangouts&ust=1604419765459000&usg=AFQjCNEYdPDaUh3C_34k2ofrkdDqq_D

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details