IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

Detecting Phishing Website using Machine Learning

Manas Shanul Jagtap[1], Shreyash Sanjay Bhagat[2], Prof. Dr. Ninad More
Department of Information Technology
D Y Patil College of Engineering, Ambi, Pune, India

Abstract: *Phishing can be described as a way by which someone may try to steal some personal and important information like login id's, passwords, and details of credit/debit cards, for wrong reasons, by appearing as a trusted body. Many websites, which look perfectly legitimate to us, can be phishing and could well be the reason for various online frauds. These phishing websites may try to obtain our important information through many ways, for example: phone calls, messages, and pop up windows. So, the need of the hour is to secure information that is sent online and one concrete way of doing so is by countering these phishing attacks. This paper is focused on various Machine Learning algorithms aimed at predicting whether a website is phishing or legitimate. Machine learning solutions are able to detect zero hour phishing attacks and they are better at handling new types of phishing attacks, so they are preferred. In our implementation, we managed an accuracy of 98.4% in prediction a website to be phishing or legitimate.*

Keywords: Phishing, Python, machine learning algorithms

I. INTRODUCTION

Internet has tremendously changed the way we work and communicate with each other. There are applications like email, file transfer, voice communication, You Tube etc. that are available for users to use. But with its humongous success has come its weaknesses and vulnerabilities. The protocols and applications responsible for its success are being exploited by malicious users and hackers for gaining limelight. Phishing websites is one such area where administrators need new techniques and algorithms to protect naïve users from getting exploited. Phishing is an attempt of fraud aimed at stealing our information, which is mostly done by emails. The ideal way to save ourselves from these phishing attacks is by observing such an attack. These phishing emails mostly come from trusted sources and try to retrieve our valuable information, for instance our passwords, bank details or even SSN. Many a times, these attacks come from sites where we have not even made any type of account. The procedure followed by phishers includes us reaching their website through the means of an email. In those emails, they make us click on a certain link that directs us to their websites. Asking for personal information is something that legitimate websites would hardly do. The looks of these phishing websites are quite similar to their respective legitimate ones and the only distinguishing factor is their URLs. Various initiations appearing from social websites, banks and online payment portals are used to deceive users. These phishing emails mostly contain links to websites that are affected with malware. Some of the ways to tackle these phishing attacks include generating awareness among people and training the users.

II. RELATED WORK

In emerging technology industry which deeply influence today's security problems has given a non-ease of mind to some employer and home users. Occurrences that exploit human vulnerabilities have been on the upsurge in recent years. [1] In the dimension of new era there are many security systems being developed to ensure security is given the utmost priority and prevention to be taken from being hacked by those who are involved in cyber-criminal and essential prevention is also taken as high consideration in organization to ensure network security is not being breached. Cyber security employee are currently searching for trustworthy and steady detection techniques for phishing websites detection. [2] Due to wide usage of internet to perform various activities such as online bill payment, banking transaction, online shopping, and, etc. Customer face numerous security threats like cybercrime. There are many cybercrime that are extensively executed for example spam, fraud, cyber terrorisms and phishing. Among this phishing is known as the popular cybercrime today. [3] Phishing has become one amongst the highest 3 most current forms of law-breaking in line with recent reports, and both

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

frequency of events and user susceptibleness has enlarged in recent years, more combination the danger of economic damage. [4]

Phishing is a type of practice done on the Internet whereindividual data are obtained by illegal approaches.[5] It supplyof obtaining sensitive information, as an example, usernames, passwords, and positive identification points of interest, often for malignant reasons, by taking up the looks of an electronic correspondence. Phishing attack will be enforced in varied kindlike Email phishing, web site phishing, spear phishing, Whaling, Tab off his guard, Evil twin phishing etc. [6] Phishing is known as webpage violence. [7] Phishing is often done by email spoofing or texting, and it typically guides user to enter points of interest at a fake web site which look and feel the same. It tries to handle the increasing range of phishing got to be met by clients in awareness and alternative efforts to ascertain protection numerous anti-phishing tools. A number of sites have currently created optional instruments for applications, like maps for redirection but clients ought to not utilize similar passwords anywhere on the net. [8] The primarykey feature is to allow user to inquire whether visited websitesis original or fake. This paper proposes a security tool called asDetecting Phishing Website Using Machine Learning

III. LITERATURE SURVEY

In this section, we review some of the recent existing works that applied some feature selection methods with machine learning techniques to enhance the detection of phishing websites. Generally, the feature selection methods utilized in detecting phishing websites can be categorized into four categories: frequencyanalysis-based feature selection, filter-based feature selection, wrapper-based feature selection, and evolutionary algorithm-based feature selection. Many research works have utilized frequency analysis based feature selection to find significant features to improve the performance of intelligent methods in recognizing the legitimate from phishing websites. In [26], the authors assessed many websites' features using a software tool to compute each feature frequency, which represents the feature importance. In [17], seventeen significant features were identified based on frequency analysis. The selected features were used to train self-structuring neural networks in order to distinguish between phishing websites and legitimate ones. In a similar way to [17], [24] analyzed the frequency of websites' features to select the most popular features of websites. Then, rule-based data mining classification models were trained based on the selected website features to recognize the new phishing websites. Find function was exploited by [25] to investigatethe most substantial features that exist frequently in numerous websites. Neuro-Fuzzy was then trained withthe best five features to detect the phishing websites through an online transaction

Alternatively,several recent texisting works demonstrated that the filter-based feature selection techniques enhanced noticeably the performance of intelligent phishing detection approaches. In [7], the authors exploited both frequency analysis and Chi-Square to select a minimal set of relevant website features from the original features. Based on the selected web site's features, a MCAC (Multi-label Classifier based Associative Classification) model was trained and developed to distinguish the phishing websites from legitimate ones. Information Gain (IG), Chi-square, and Correlation Feature Set were employed by [30] to find the most significant website's features in order to enhance the detection accuracy of phishing websitesfor some rule-based classification machine learning algorithms: C4.5, RIPPER, and PART. In [8], the authors suggested using the IG, Chi-square, and Correlation Features Set (CFS) to reduce the data dimensionality and select the minimal set of important features. Then, four rule-based classification algorithms (OneRule, JRip, Part, and J48) were trained after applying feature selection methods in order tomaximize tdetection rate of phishing emails.

IV. PROPOSED SYSTEM

Address Bar based Features

Using the IP Address

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 information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

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/ab51e2e319e51502f416dbe46b773a5e/? cmd=_home&amp %3Bdispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8 dd4105e8"
To ensure 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 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 dataset size

Using URL Shortening Services "TinyURL"

URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".

V. METHADOLOGY

Writing a review is the most critical advance in the programming improvement process. Before building up the instrument it is important to decide the time factor, economy and friends quality. When these things are fulfilled,at that point following stages is to figure out which working framework and dialect can be utilized for building up the instrument. When the developers begin fabricating the instrument the software engineers require part of outside help. This help can be gotten from senior software engineers, from book or from sites. Before building the framework the above thought are considered for building up the proposed framework**.**

Benefits of Machine learning:

- Simplifies Product Marketing and Assists in Accurate Sales Forecasts.
- Utilization and efficiency improvement

Hardware Resource Requirement

1	Hardware	any processor above 500 Mhz
2	Speed	2.80 GHz
3	RAM	4GB
4	Hard Disk	1 tb
5	Floppy Drive	1.44 MB
6	Key Board	Standard Windows Keyboard

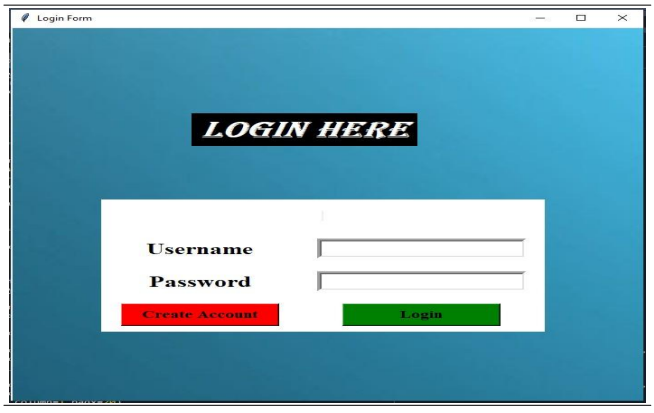
Software Resource Requirement

1	Operating System	Windows 7 and above
2	Technology	machine learning
3	Web Technologies	Html, JavaScript, CSS
4	IDE	anaconda 3.5.0.3
5	Web Server	any
6	platform	Spyder

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

VI. RESULTS

Fig . Main Interface



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

Fig. Phishing Detection Interface

VII. APPLICATIONS

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

Outcome

Our model solves these problems by automating the process of structuring the network and shows high acceptances for noisy data,fault tolerance and high prediction accuracy .Several experiments were conducted in our research and the number of epochs differs in each experiment.

VIII. CONCLUSION

It is outstanding that a decent enemy of phishing apparatus ought to anticipate the phishing assaults in a decent timescale. We accept that the accessibility of a decent enemy of phishing device at a decent time scale is additionally imperative to build the extent of anticipating phishing sites. This apparatus ought to be improved continually through consistent retraining. As a matter of fact, the accessibility of crisp and cutting-edge preparing dataset which may gained utilizing our very own device will help us to retrain our model consistently and handle any adjustments in the highlights, which are influential in deciding the site class. Albeit neural system demonstrates its capacity to tackle a wide assortment of classification issues, the procedure of finding the ideal structure is very difficult, and much of the time, this structure is controlled by experimentation. Our model takes care of this issue via computerizing the way toward organizing a neural system conspire; hence, on the off chance that we construct an enemy of phishing model and for any reasons we have to refresh it, at that point our model will encourage this procedure, that is, since our model will mechanize the organizing procedure and will request scarcely any client defined parameters.

REFERENCES

[1] “WC-PAD: Web Crawling based Phishing Attack Detection” Nathezhtha.T, Sangeetha.D,Vaidehi.V
[2] “Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture” Ivan Ortiz-Garces, Roberto O. Andrade, and Maria Cazares R. M. Mohammad, F. Thabtah, and L. McCluskey, ‘‘Tutorial and critical analysis of phishing Websites methods,’’ Comput. Sci. Rev., vol. 17, pp. 1–24, Aug. 2015.
[3] “A Methodical Overview on Phishing Detection along with an OrganizedWay to Construct an Anti-Phishing Framework” Srushti Patil, Sudhir Dhage
[4] “A survey of the QR code phishing: the current attacks and countermeasures” Kelvin S. C. Yong, Kang Leng Chiew and Choon Lin Tan
[5] “Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection” Mahdieh Zabihimayvan and Derek Doran
[6] N. Goel, A. Sharma, and S. Goswami, “A way to secure a qr code: Sqr,” in 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE, 2017, pp. 494–497

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 15, May 2023

[7] V. Mavroeidis and M. Nicho, "Quick Response Code Secure: A Cryptographically Secure Anti-Phishing Tool for QR Code Attacks," in International Conference on Mathematical Methods, Models.
[8] K. Krombholz, P. Fr ¨uhwirt, T. Rieder, I. Kapsalis, J. Ullrich, and E. Weippl, "QR Code Security–How Secure and Usable Apps Can Protect Users Against Malicious QR Codes," in 2015 10th International Conference on Availability, Reliability and Security. IEEE, 2015, pp. 230–237