International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Machine Learning Method to Detect the Phishing Websites

Mohammad Musaddiq Affan. M[1]
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Mohammed Zafeer. K. A[2]
Computer Science and Engineering
SRM Institute of Science and Technology
Chenai, India

Nihal Ahmed Tayeb. K[3]
Computer Science and Engineering
SRM Institute of Science and Technology
Chennai, India

Abstract - *Phishing continues to be one of the most commonly committed crimes in whole world since people give away their personal details or any confidential information online unknowingly. The malicious websites takes advantage of such person by undergoing into this crime to obtain their personal details like username, password, bank details, etc and use it either to steal their secret information or exploit their financial situation. Today, technology is growing rapidly, we try to safeguard people by such crimes and educate them more about such events. As the technology is booming, there are various tools to detect and analyze such attacks or crimes. We have used Machine Learning which is a powerful tool to fight against such phishing and related crimes. Alongside random forest and logistic regression algorithms are also used to obtain better results on detecting the illegitimate sites.*

Keywords: phishing, phishing websites, detecting phishing

1. INTRODUCTION

Phishing is a cybercrime which involves the act of obtaining essential information of a user by showing themselves as promising, the target is contacted via email or phone which they trick them into thinking they are a legitimate website or any related to their welfare and obtain their essential information such as bank details, username, passwords. Today, websites creation became normal and programmers are creating so professionally that a common man cannot differentiate and that is being exploited. The content and data is made so professional that it looks like a promising website which helped these attackers to obtain the essential and secret data from the user such as bank details, username, password, account details, etc. The attackers even take similar steps as the real websites would do to prevent them from being held as the illegal user such as making the user to answer certain question or not showing any signs of illegal activity. A lot of efforts and work has been put to prevent such crimes, these attacks shall be prevented by detecting the websites and educating the user about such illegality. Various algorithms have been developed which would detect such websites. In this paper those methods are discussed.

2. LITERATURE SURVEY.

A. Large Scale automatic classification of phishing pages, [1] (Colin Wh, Brian Ryner, Marria Nazif : 2010)

In this paper, the detection of phishing websites by feature extraction of URL by examining lexical features, this method is used to identify if an URL is real or not by analyzing the characteristic features of a particular URL which gives an output of about 90 % but does need to be provided as the content feature analysis.

B. Phool proof phishing prevention, [2] (Bryan Parno, Cynthia Kuo, Adrian Perrig : 2006)

In this paper the author proposed a separate trusted system to perform mutual authentication instead of a method to eliminate any false intervention but this system does not provide relevance with the e attacks that might happen with more developed terms of usage and techniques.

C. The battle against phishing : Dynamic security skins, [3] (Rachna Dhamija, J Doug Tygar : 2005)

In this paper the author provides a novel approach to tackle phishing as he introduces a browser extension which would provide a dedicated username and password and also the remote server to generate a unique abstract image or generating a “skin” and by entering the image details and verify the skin on the website provides authentication.

D. A framework for detection and measurement of phishing attacks, [4] (Sujata garera, Niels Provos, Monica Chew : 2007)

In this paper the author suggests that rather than warning the user on such ongoing attacks of phishing educating the users about the attacks which would provide the user the knowledge to distinguish between such websites but here the attacker develops methods of attacks so user needs to keep up with such ideologies and knowledge to prevent the attacks.

E. Behavioral response to phishing risk, [5] (Julie S, Mandy Holbrook: 2007)

In this paper the author gives an approach with algorithms such as support vector machines (SVM) classification and regression trees (CART) and logistic regression (LR) for predicting phishing emails.

3. CHALLENGES.

Phishing attacks which involves in identifying the real websites becomes difficult to identify major and minor such issues,

The challenges include how to educate the users in identifying such websites or emails to be aware of such attacks. Corporations should focus on trying to obtain minimal data about an user since the user is also unaware of such question or data required by the organizations and try to use multi factor authentication process which involves cost. There also exist that websites are creating in similar reference to the real websites in which the attackers et paid to do so and there are not much efforts taken to prevent this such that they are paid more for performing such actions and also they keep themselves in an anonymous state and also regarding the website accountability user is at a speed to complete the action which makes him or her vulnerable as the user does not wait to read the policies and any errors within the specified document or subject it makes it more easier for the attacker to exploit the users.

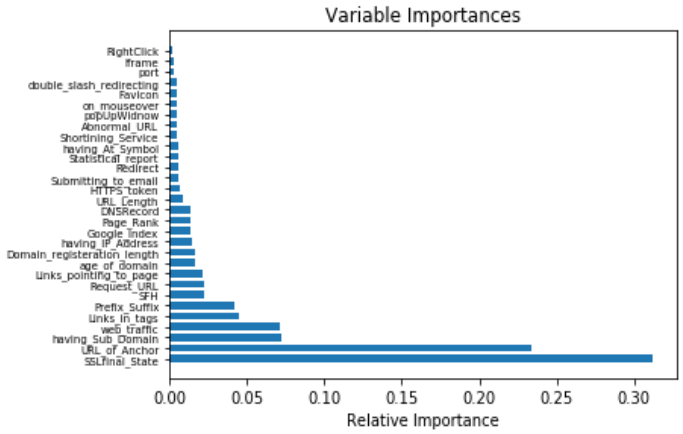
4. RELATED WORKS.

In the previous system, the result occurred in the form of binary number such they used only logistic regression classification technique which is used to determine whether the independent variable has a dependent binary variable, due to which for a legit URL it showed as 1 and for a non-legit URL it showed as 0 which is only a 92% accurate result. Here we added two more machine learning features including logistic regression technique known as Random forest and Support Vector Machine (SVM) technique that shows 99% accurate results by providing a user a message of "This is a legit URL" for true websites and "This is not a legit URL" for false websites.

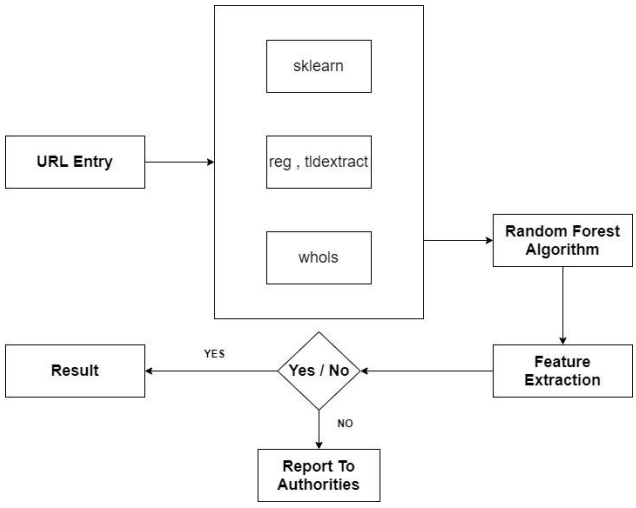
There is an increase in web development which means the attack rate and methodologies also increase herby paving way to newer methods for the attacker to exploit the user. Now since the user is not aware of the oncoming techniques they are easily exploited and can't withhold anyone for the crime committed and there exist this issue in educating the people about these type of attacks in which the user should concentrate on and also since most of the attackers are anonymous their locations are difficult to identify.

5. PROPOSED METHODOLOGIES.

The existing model tries to evaluate a website based on the content rather than the classification the features or URL this system attempts to classify whether a site is legit or not based on the URL extraction features and since URLS don't share the same features and depends upon the site this tries to look at the symbols and type of address used and would try to check the secure links, if possible would be able to clear out the non-legit sites with better accuracy.



A. SYSTEM ARCHITECTURE.



6. CONCLUSION

In this paper, we describe that our system detects the malicious or phished URLs by letting the user to check each website which will give an output by a message that "This is a legit URL" or "This is not a legit URL", if it is malicious or phished. By using various Machine Learning techniques such as logistic regression, Random Forest algorithm, we believe that our system investigates each URL in such a way that it shows 98% accurate results. This system examines only the developed websites that are available on the internet. However, further research is needed to make our system examine the partially

developed URLs in order to make a loss for the phishers and reduce the internet crime.

7. REFERENCES

[1] Large Scale automatic classification of phishing pages, Colin Wh, Brian Ryner, Marria Nazif, 2010

[2] Phool proof phishing prevention, Bryan Parno, CynthiaKuo, Adrian Perrig International conference on financial cryptography and data security 2006

[3] The battle against phishing: Dynamic security skins, Rachna Dhamija, J Doug Tygar Proceedings of the 2005 Symposium on Usable privacy and security 77- 88, 2005

[4]A framework for detection and measurement of phishing attacks, Sujata garera, Niels Provos, Monica Chew, Aviel D Rubin Proceedings of 2007 ACM workshop on recurring malcode 2007

[5] Behavioral response to phishing risk, Julie S, Mandy Holbrook, Lorrie Cranor Proceedings of anti phishing working groups, 2007

[6] A comparison of machine learning techniques for phishing detection, saeed Abu, Dario Nappa, Xinlei Wang, Suku, Proceedings of the anti phishing working groups, 2007

[7] J. Bem, G. Harik, J. Levenberg, N. Shazeer, and S. Tong. Large scale machine learning and methods. US Patent 7222127, 2007

[8] L. Breiman. Random Forests. Machine Learning, 45(1):5– 32, 2001

[9] Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "An examination of machine learning systems for phishing recognition." In Proceedings of the counter phishing working gatherings second yearly eCrime specialists summit,ACM, pp. 60-69, 2007

[10] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker. Beyond Blacklists: Learning to Detect Malicious Websites from Suspicious URLs. In KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1245–1254, June–July 2009.