International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1S, October 2019

Phishing Scam Detection using Machine Learning

Elakya. R, Badri Narayan Mohan, Vijay Kishore. V, Keerthivasan R, Naresh Kumar Solanki

Abstract- As a wrongdoing of utilizing specialized intends to take sensitive data of clients and users in the internet, phishing is as of now an advanced risk confronting the Internet, and misfortunes due to phishing are developing consistently. Recognition of these phishing scams is a very testing issue on the grounds that phishing is predominantly a semantics based assault, which particularly manhandles human vulnerabilities, anyway not system or framework vulnerabilities. Phishing costs. As a product discovery plot, two primary methodologies are generally utilized: blacklists/whitelists and machine learning approaches. Every phishing technique has different parameters and type of attack. Using decision tree algorithm we find out whether the attack is legitimate or a scam. We measure this by grouping them with diverse parameters and features, thereby assisting the machine learning algorithm to edify.

Indexterms: Decision Tree Algorithm, Machine Learning, Phishing scam, Client sensitive information.

I. INTRODUCTION

The internet has become a crucial and indispensable infrastructure for the human society which has helped both individuals and corporations over the years and has given a platform for worldwide connectivity. However, it also has its fair share of drawbacks especially when it comes to security. One of those security threats comes in the form of Phishing. Phishing is a technique which employs technical tactics and social engineering to lure gullible people into leaking personal and valuable data and information. Phishers have multiple methods in their disposal to steal sensitive information. One such form of phishing is achieved by creating replicas of real websites which are designed in such a way that users are led to fraudulent websites where unsuspecting users release credible values such as atm card values, pins and many important data.

Manuscript published on 30 October 2019.
***** Correspondence Author (s)
Mrs. R.Elakya, Assistant Professor, SRM Institute of Science and Technology (formerly known as SRM University), Ramapuram
Mr. Badri Narayan Mohan, Computer Science student, SRM Institute of Science and Technology, Ramapuram.
Mr.Vijay Kishore, Computer Science student, SRM Institute of Science and Technology, Ramapuram.
Mr. Keerthivasan, Computer Science student, SRM Institute of Science and Technology, Ramapuram.
Mr. Naresh Solanki, Computer Science student from SRM Institute of Science and Technology, Ramapuram.

Phishers also create spoofed e-mails disguising to be from legitimate corporations which tricks recipients into believing such e-mails are from those legitimate corporations and buy into the contents of such e-mails which slyly demand users for information such as username, user id, and passwords for accounts commonly held in social media and e-commerce websites among others. Such e-mails also lure people into phoney schemes. The main reason why consumers of the internet buy into such phishing methods is because of how phishers abuse everything right from logos and slogans to trademarks among many such corporate identifiers which makes the fake websites and e-mails dangerously similar and bear resemblance to their original and legitimate counterparts. In the United States solidly it cost 71 billion dollars in harm due to these scams and thefts that happen over the internet. Hence, phishing continues to be one of the briskly growing identity theft scams on the internet.

Blacklisting is a process commonly used by many web browsers and is used to warn users about potentially dangerous web pages that are included in their blacklist listings. However such listings don't include previously unseen URLs since it is non-trivial to decide if an unseen URL is malicious. Hence, phishing detection faces challenges such as real time detection which is not possible with blacklisting as it is impossible to have an exhaustive list of phishing websites. Speculation ability is another test looked by phishing as assailants are constantly patching up their techniques to set up flourishing framework so as to help continuous phishing movement. One of the important blocks of such an infrastructure is botnet, which is used to generate automated phishing emails and also anchor phishing sites. A recent study by the APWG supports the fact that there could be more sophisticated schemes and infrastructures used by attackers to exploit the ever expanding volume of popular brands. To summarize, we are in an urgent need of a reliable phishing detection system which can potentially assure almost-perfect accuracy in an internet environment where the amount of attackers and phishing activity continues to expand and grow.

II. HISTORY OF PHISHING

The possibility of phishing has been here for over 2 decades and can be followed back to the 1990s by AOL (America Online). A group of programmers gathered together and formed a group by the name 'warez network' who can be considered as the first set of "phishers". During the initial stages of phishing, a generator was made to generate random

charge card numbers which would be later used to create counterfeit accounts on AOL. When they had the capacity to coordinate a certified card, they made records and spammed others in AOL's society where individuals were there to take the bait. Around 1995, AOL had the ability to stop such irregular charge card generators, but the warez community moved on to other techniques and started to disguise as AOL representatives and requesting users through AOL messenger for private data.

2.2 A Switch to Email

Internet users started becoming more aware about such malicious activity over time and this forced phishers to move on to emails which during the time was extremely difficult to make, shabby to convey and was never good enough to capture individuals as they were ineffectively built and loaded with various syntactic mistakes. This forced them to reform their techniques and they rapidly changed them to get more modern. In late 2003, phishers started disguising like mainstream organizations such as hurray billing.com and ebay-fulfillment.com. This time they employed refined techniques such as increasingly genuine looking mails which easily lured gullible people into believing they were genuine. In October 2003, PayPal clients were hit by the Mimail infection by means of utilizing pop up windows resembling that of PayPal and made users give away their client/secret phrases which immediately went to the programmers. In today's climate phishers have vastly changed strategies and such malicious attackers are increasingly becoming difficult to trace and have developed many approaches to easily pickup trust from individuals.

III. EXISITING SYSTEM

A. E-Mail Phishing Scams

Email phishing is a numbers amusement. An aggressor passing on an extensive number of beguiling messages can net significant information and sums of money, paying little respect to whether only somewhat dimension of recipients falls for the trap. As saw above, there are a couple of techniques aggressors use to fabricate their success rates.

For one, they will put everything on the line in planning phishing messages to imitate real messages from a caricature association. Utilizing a similar stating, typefaces, logos, and marks influences the messages to seem real. Moreover, aggressors will ordinarily attempt to push clients enthusiastically by making a feeling of direness. For instance, as recently appeared, an email could compromise account termination and spot the beneficiary on a clock. Applying such weight makes the client be not so much determined but rather more inclined to mistake. In conclusion, connects inside messages look like their real partners, yet regularly have an incorrectly spelled area name or additional subdomains. In the above urls, the avitahr.in/careers URL was changed to avitahr.inrenewal.com. Similarities between the two tends to offer the impression of a protected connection, making the beneficiary less mindful that an assault is occurring.

B. Spear Phishing

Spear phishing concentrates on a isolated individual or endeavour..

An assault may happen as pursues:

A culprit inquiry about names of workers inside an association's advertising office and accesses the most recent venture solicitations. Acting like the showcasing chief, the assailant messages a departmental undertaking supervisor (PM) utilizing a headline that peruses, Updated receipt for Q3 crusades. The content, style, and included logo copy the association's standard email layout. A connection in the email side-tracks to a secret phrase ensured inside report, which is in fact a satirize adaptation of a stolen receipt. The PM is asked for to sign in to see the report. The assailant takes his accreditations, increasing full access to delicate regions inside the association's system. By giving an aggressor legitimate login accreditation, skewer phishing is a powerful strategy for executing the principal phase of an APT.

C. Cybersquatting

Cybersquatting, is selecting, managing in, or using a space name with deceitfulness objective to profit by the unselfishness of a trademark having a spot with someone else. The cybersquatter may offer pitching the domain to an individual or organization who claims a trademark contained inside the name at an expanded cost or may utilize it for false purposes, for example, phishing. For instance, the name of your organization is "Avita HR solutions" and you register as avitahrsolutions.com. At that point phishers can enroll avitahrsolutions.net, abcompany.org, abcompany.biz and they can utilize it for fake reason.

D. Typosquatting

Typosquatting, likewise called URL seizing/hijacking, is a type of cybersquatting which depends on oversights, for example, typographical mistakes made by Internet clients while contributing a site address into an internet browser or dependent on typographical blunders that are difficult to see while a quick glance. URLs which are made with Typosquatting resembles to be a confided URL. A client may coincidentally enter a wrong site address or snap a connection which resembles a confided in space, and along these lines, they may visit an elective site claimed by a phisher.

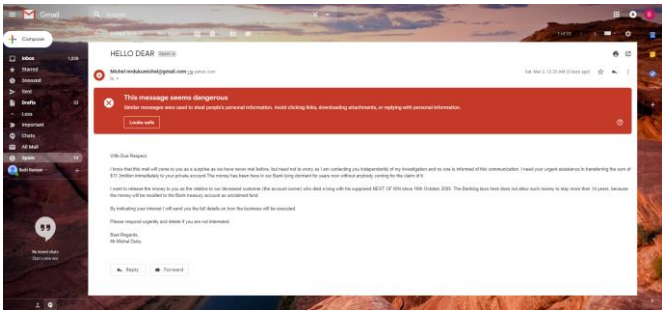

Figure 1: Example of E-Mail Phishing

IV. PREVENTING PHISHING ATTACKS

Phishers constantly come up with new methods to lure people into their phishing environment but there are certain steps one can take to protect themselves and their organization:

• Many sites expects their customers to enter login information while the customer's image is displayed in the login page. This infrastructure is prone to security threats. One practice to periodically change passwords from time to time assures security. Another framework which can be utilized is a CAPTCHA enabling users to go through an extra step to access information which increases security.
To avoid spam, spam channels can be used. These channels make use of various parameters such as source, product used and presence of the message to decide if messages are spam. Program settings should be changed in order to keep illegitimate sites from opening. Thus the program settings should only allow access to legitimate sites.
• Banks and other money related organizations can anticipate attackers by using observing frameworks. Cases of phishing when observed can be reported to these organizations where lawful action can be taken on such attackers.
• If a URL link is present in an email, a quick look at the URL will help you decide if a website is legitimate as sites that begin with "https" have SSL certification which represents the site being safe.
In the end all locales will be required to have a legitimate SSL.

Sometimes, spam channels can also classify messages from genuine sources as spam so it is uncertain if all the blocked sources are entirely spam. Thus it is essential to keep changing the program settings from time to time in order to make up for the inaccuracy these spam channels offer. These settings keep a record of illegitimate sites and when such sites are accessed, the spam channel intercepts and gives warning messages or blocks such locations. This implies that it is necessary for the settings of the program to be able to identify trustworthy sites accurately and allow access to them. Numerous sites expect clients to enter login data while the client picture is shown. This sort of framework might be available to security assaults. One way to deal with assurance security is to change passwords constantly, and never use a comparable mystery state for different record. Legitimate websites can use CAPTCHA to increase security for users. This way phishing sites can be prevented as users have to go through an extra layer of security check to access information. Changes in browsing habits are required to prevent phishing. If verification is required, always contact the company personally before entering any details online. When received an email with links contained, try analysing what the link maybe. One way is to hover over the URL and see what it maybe. Usually, all secure websites have certifications from the world wide web consortium with a recognizable Secure Socket Layer(SSL) where the URL begins with "https". Generally all online tools and sites require this certification for security purposes.

V.PHISHING FEATURES CLASSIFICATION

Alongside URL based classifications, different types of features are used for Machine Learning calculations to extensively check for phishing environments. These features used for phishing recognition is given below:

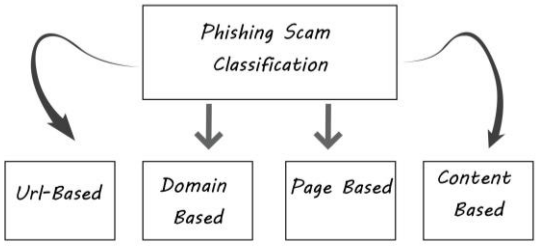

***Figure 2**: Phishing Classification*

VI. DISADVANTAGES OF EXISTING SYSTEM

• Phishing destinations are likewise developing quickly in amount and unpredictability
• it is hard for clients to observe if an approaching connection is authentic or not.
• Numerous phishing sites are made on well-known sites, for example, online journals or Google destinations, where the positioning highlights are not helpful for phishing distinguishing proof.
• New URLs have low ranking values that are like phishing URLs.

VII. PROPOSED SYSTEM

Distinguishing Phishing Domains is an classification issue, so it suggests we need marked datasets which have tests as phish areas and real spaces in the preparation stage.The dataset which will be utilized in the preparation stage is a vital point to manufacture fruitful discovery mechanism. We need to utilize tests whose classes are correctly known. So it implies, the examples which are named as phishing must be completely identified as phishing. Similarly, the examples which are marked as real should be completely distinguished as genuine. Something else, the framework won't work accurately on the off chance that we use tests that we don't know about. Hence, we use a machine learning algorithm called Decision Tree algorithm. Based on the raw data acquired from reputed sources like Alexa, Phishtank and other data resources, we create datasets consisting of features which are checked one by one in the decision tree. Generating a tree is the main structure of the detection mechanism. In the tree, yellow and elliptical shaped cells represent the features from the feature sets and are called nodes whereas green and rectangular shaped cells represent classes and are called leaves. When an example arrives, it goes through the process of getting its features checked starting from its length. Once the journey of checking is complete, it will be clear which class the sample belongs. The crucial question is which feature will now be observed as the root node and which features must come after the root. Choosing features intelligently affects the efficiency and success rate of the algorithm. The decision tree algorithm uses a features called info gain measure which stipulates how well a given feature distinguishes from the training features according to their target stratification which implies how easy it is to distinguish the features. Here is the equation used to determine information gain measure:

$$IG(Q,A) = Entropy(Q) - \Sigma((|Q_v|)/(|Q|)).Entropy(Q_v)$$

High gain score implies the feature has a high distinguishable ability. Thus the feature with highest gain score is selected as the root. Another variable called entropy is used to regulate the purity of an arbitrary collection of samples 'Q'. Here is the equation used to determine the Entropy:

$$H(Q) \equiv \sum_{i=1}^{n} p_i \ -\log_2 p_i$$

Original entropy is constant while relative entropy is a dynamic parameter which keeps changing. A low relative entropy means the purity is high and a high relative entropy means the purity is low. Thus, our objective is to ensure we have a higher purity down the tree because higher purity implies high success rate.

The dataset acquired is divided into two parts by feature values. The length feature is used to divide the samples used. Here we have used 18 samples in which the '+' sign indicates phishing class and '-' sign indicates legitimate class. Nine of them are on the left and the remaining nine are on the right. After the calculations using the equations above, we observe that the right part has high purity which implies low Entropy value and the left part has low purity which implies high Entropy value. Thus, with the original Entropy value and the Relative entropy values we can calculate the information gain score.

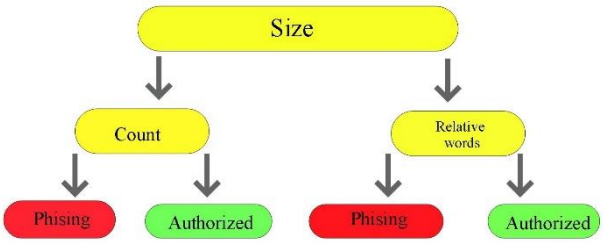

Figure 3: *Decision Tree*

These information scores are used by the decision tree algorithm to determine the features with maximum gain scores and used as nodes to construct the tree in which leaves are changed as nodes to represent each unique feature. As the levels of the tree increase, the leaves will have higher purity. When the tree is big enough, the training process for the machine learning algorithm is finished and now can be used to determine. This tree now contains the most distinguishing features required for the detection. Thus, in order to have a higher success rate of detection the tree which we construct must have a training dataset containing a wide variety of samples and sources of data. The objective is to attain generalization of system success in order to assure successful detection of phishing websites based on real world data.

VIII. CONCLUSION

Thus we have devised a rigid mechanism containing a rigorous machine learning algorithm to distinguish the features which are used to determine if an infrastructure is a legitimate infrastructure or a phishing infrastructure with a steep accuracy rate. This mechanism takes in newest parameters to device a method to protect people from getting phished. This method uses a scoring technique to distinguish whether its phishing or legitimate. With proper training set, the comparison set can be easily distinguished which makes this a really powerful tool to use.

IX. FUTURE ENHANCEMENTS

With enhancements in the training datasets used and tools we will be able to create more efficient machine learning algorithms which can guarantee better success rates in real world detection of phishing infrastructures.

ACKNOWLEDGMENT

The Scholars would like to extend its gratitude to Mrs.R.Elakya for assisting and guiding us through the project. The concepts presented are thought out while implementing the project.

REFERENCES

1. Ebubekir Buber, Önder Demir and Ozgur Koray Sahingoz, Istanbul, Turkey; Feature Selections for the Machine Learning based Detection of Phishing Websites. (2017)
2. Peng Yang, Guangzhen Zhao and Peng Zeng; Phishing Website Detection based on Multidimensional Features driven by Deep Learning, Southeast University, Nanjing 211189, China. (2019)
3. Joby James, Sandhya L and Ciza Thomas, Detection of Phishing URLs Using Machine LearningTechniques, SCT College of Engineering. (2013)
4. F.C. Dalgic, A.S. Bozkir and M. Aydos; Phish-IRIS: A New Approach for Vision Based Brand Prediction ofPhishing Web Pages via Compact Visual Descriptors, Hacettepe University, Turkey. (2018)
5. Sreyasee Das Bhattacharjee, Ashit Talukder, Ehab Al-Shaer, Pratik Doshi; Prioritized Active Learning for Malicious URL Detection using Weighted Text-Based Features, University of North Carolina, Charlotte. (2017)
6. Ram Basnet, Srinivas Mukkamala and Andrew H. Sung; Detection of Phishing Attacks: A Machine Learning Approach. (2008)
7. Murat Karabatak and Twana Mustafa; Performance Comparison of Classifiers on ReducedPhishing Website Dataset, Firat University Elazig, Turkey. (2018)
8. Wesam Fadheel, Mohamed Abusharkh and Ikhlas Abdel-Qader On Feature Selection for the Prediction of Phishing Websites. (2017)
9. Yasin Sönmez, Türker Tuncer, Hüseyin Gökal, Engin Avc; Phishing Web Sites Features Classification Based on Extreme Learning Machine. (2018)
10. Anu Vazhayil, Vinayakumar R and Soman KP; Comparative Study of The DetectionOf Malicious URLS Using Shallow and Deep Networks. (2018)

AUTHORS PROFILE

Mrs. R.Elakya is an Assistant Professor, SRM Institute of Science and Technology (formerly known as SRM University), Ramapuram Campus from June 19, 2017. She has achieved achieved 100 percent results in various subjects. She is driven towards wireless networks and Computer Networks.

Mr. Badri Narayan Mohan is a Computer Science student from SRM Institute of Science and Technology, Ramapuram. He has a driven passion towards Machine learning with game development and is pursuing his Masters at Carnegie Mellon University in Entertainment Technology.

Mr.Vijay Kishore is a Computer Science student from SRM Institute of Science and Technology, Ramapuram. He is driven with passion towards music production and computer science

Mr. Keerthivasan is a Computer Science student from SRM Institute of Science and Technology, Ramapuram. He is driven towards Cybersecurity and Information technology with computer Science.

Mr. Naresh Solanki is a Computer Science student from SRM Institute of Science and Technology, Ramapuram. He is driven towards data science and machine learning.