WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phishing Detection using Machine Learning Technique: Random Forest

Prof.SayaliSabale, GlaraNadar, SabihaNaikawadi, SayaliPatil, Richa Patel

Computer Department, GenbaSopanraoMoze College,Balewadi Pune

ABSTRACT:

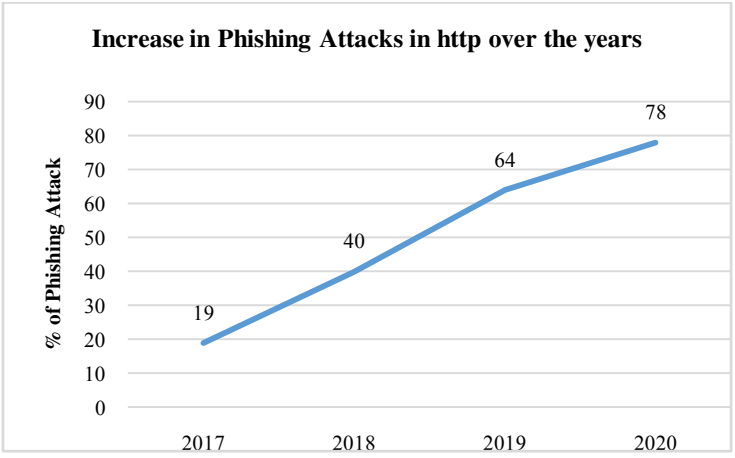
Phishing Attack can be described as a cyber-crime in which the phisher creates a fake page that seems to be from legitimate source through emails or e-messages. In this paper we will be discussing the requirement for phishing detection, an epitome on machine learning, the methods used to achieve phishing detection, particularly Random Forest algorithm using machine learning.

Keywords: Phishing, Phishing websites, Machine Learning, Anti- Phishing, Phishing Attack, Security And Privacy, Classification, Phishing Approaches.

I.INTRODUTION

Nowadays, with increasing online dependency, there is also an increase in cyber-crimes. Cyber-crimes may be hacking, malware, privacy threat, phishing attack, etc. In Phishing Attack the phisher tricks the user to believe that the website they are in is a legitimate one and proceeds to steals important details like credentials, personal data, etc.

In recent years, a huge increase in phishing sites is found.As of 2020, 6.95 million new phishing and scam pages were created. Http websites are considered to be safe website that are free from threat but now most of the phishing attackers manage to insert their scam pages in http too.



Most attacked phishing website is financial institutions that is followed by social media and webmails with the least attacked kind of institution being cryptocurrency. In United States businesses, there is a loss of more than US$2billion every year due to their clients becoming victim to phishing [1].Users are usually unaware of these attacks which results in the attackers to exploit user's weaknesses, due to which it is more difficult to reduce such attacks [2].To prevent such violations in terms of currency and personal information, such attacks are to be detected in first place.

To stop such attacks from occurring, it needs to be detected in first place.For detection of such threats, machine learning based classification algorithms (Random forest,KNN,Naive Bayes, etc.) along withpython library files (scikit,scipy, mahotas)can be used.

Machine Learning

Machine learning is when machine uses certain algorithms to learn and understand data inorder to make predictions or take decision when required.

Dataset i.e. set of data to be analysed can be trained by different machine learning algorithms which simply means that the machine learns and understands the given dataset. Whenever or wherever we need such machine learning process, it can be fed to the required algorithm and data in advance along with the rules and limitations for analysis of the pattern [3].

II.LITERATURE SURVEY

In[4] an effective "Anti phishing simulator" is developed which uses Bayesian network classifier to detect phishing emails by checking if the email url is present in the phishing URL database and if true adds the Url in spam folder.

To overcome the difficulties in hardware based phishing detection a software is introduced in [5]which works in 3 stages

- DNS (domain name system) blacklist: Lists the phishing URLs and WC-PAD checks if the IP address of web address matches the one in list and if matched alerts the user.
- Web crawler: Extracts few features and crawls to interconnected links and pages.
- Heuristic analysis: From the web crawler certain extracted features namely URL, web content, web traffic is given as input and heuristic analysis identifies the phishing websites.

In [6], a system for detecting phishing URLs is constructed which isknown as Multi-label Classifier based on Associative Classification (MCAC). The MCAC which is a rule-based algorithm, here from the phishing data set various label rules are extracted sixteen features were used by them and sorted Uniform resource locators into three classes:- Phishing, legitimate and suspicious.

The client or customer system not able to identify and remove the malicious Uniform Resource Locator can put the lawful or legal customers in unsafe condition. In this paper [7], they have developeda newclassification method to solve the problem that is faced by the previous mechanisms in the detecting the malicious URLs. Here, the classification models are constructed using machine learning method. This method directs the syntactical nature of the URL it also directs the lexical and semantic meaning of these dynamically changing URLs. Few of the previous filtering mechanisms that detected the malicious URLs includes Black-Listing, Heuristic Classification etc. These typical mechanisms depended on URL syntax matching and keyword matching. Hence these cannot effectively manage with the webaccess techniques and consistently developing technologies. Also, these approaches do no detect the modern URLs like short URLs, dark web URLs correctly. To overcome this new method proposed.

For classifying phishing URLs, Hadimade use of the Fast-Associative Classification Algorithm (FACA) [8]. They investigated a data set containing of 11,055 websites including two classes, legitimate and phishing. FACA works by constructing a model for classification and finding out all frequent rule item sets. The data set contained 30 features. For FACA, there was usage of minimal support and the minimum confidence threshold values as 2% and 50% percent, respectively.

An initial study was carried out to analyse several URLs used in phishing attacks, which encouraged the authors to propose a set of eighteen URL features [9]. These features were classified into four groups that is

- Whitelist-based
- Reputation-based
- Obfuscation-based
- Sensitive word-based

This experiment was performed using a logistic regression classifier, which gives an accuracy of 97.31%.

In [10], 48 new features concentrating on specific source codes in the HTML content, and another ten features that were based upon common URL features from pre-existing works is explored. Here, SVM classifier calculated the two feature classification. Results suggest that the execution of URL features and HTML content features are comparable and differing both in true positives and true negatives by less than 1%.

Researchers all around the world has given significant and successful result to detect malicious URLs. There is a lot of issues due to the evolving nature of cyber-attacks. This paper [11] mentions effective hybrid methodology as well as new features to handle with this problem. For calculating the results they have used supervised decision tree.Experiments are performed on balanced dataset.The demonstration results show that new features achieved 98- 99% detection accuracy when all the decision tree learning classifiers worked efficiently in their labelled dataset. Using majority voting technique they have achieved 99.30% detection accuracy with very low FPR and FNR .It is more favourable than the well-known anti-virus and anti-malware solutions.

By changing the messages a little, phishers seem to pass the emails through filters and to overcome this problemSAFEPC (Semi-Automated Feature generation for Phish Classification) [12] operating on a large collection of phishing and legitimate emails is designed, which is a system that extracts some high level features so as to reduce the common phishing email detection strategies. For execution of this classifier they aggregated huge set of phishing emails from central IT oranizations. It does far better than spam assassin which is a common filtering tool.

A new way to analyze phishing attacks by investigating Effectiveness of phishing attack and change in dynamics [13]. For performance of methods of ideology of wavelet coherence and chaos theory new tools are used. Real data related to phishing attacks are usedfor analysis purpose.

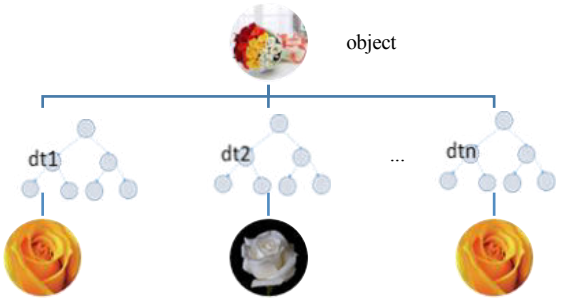
III.RANDOM FOREST ALGORITHM

Random Forest Algorithm is a machine learning algorithm also called bootstrap aggregation solely used for regression as well as classification problems and it comprises of huge number of decision trees that executes in parallel for more accuracy than the other machine learning algorithm i.e. Decision tree.Based on bootstrap methods, decision trees are created, this method randomly selects features and samples of dataset with replacement to construct a tree [14].Decision Trees also called classification trees which is composed of class variable on its leaf nodes and the dependent variables reside on the intermediate nodes. Class variables can be also termed as decision variable or predictor variable since they do the decision or prediction [15]. This algorithm also outperforms the inconvenience of high variance, which is found in decision tree algorithm. Random forest algorithm operates

in the following steps:

1. An object is identified by number of decision trees, say dt1, dt2, dt3…..dtn.
2. Consider the object to be a collection of flowers of various colours (red, yellow, white).
3. In this case if d1 and d3 says that the object is yellow in colour and d2 says it's white in colour then the output of the random forest algorithm will be yellow colour.

In simple words, Random forest algorithm works on majority of voting from the number of decision trees present.



Therefore, the Algorithm will select yellow flower with majority of interpretation.

CONCLUSION

This paper provides information onphishing website detection using machine learning algorithm. The proposed technique uses Random Forest Classifier which will give higher accuracy than the other techniques.

REFERENCES

[1] Gunter Ollmann, "The Phishing Guide Understanding & Preventing Phishing Attacks", IBMInternet Security Systems, 2007.
[2] AtharvaDeashpande, OmkarPedamkar, Nachiket Chaudhary, Dr.SwapnaBorde, Detection of Phishing Webiteusing Machine Learning Algorithms, 2021
[3] Jino S Ganesh, NiranjanSwarup.V, MadhanKumar.R, Harinisree.A, P.G Scholars,
[4] MuhammetBaykara,ZahitZiyaGürel,Detection of Phishing attacks, 2018Machine Learning Based Malicious Website Detection.
[5] Nathezhtha,Sangeetha,Vaidehi, WC-PAD:Web Crawling based Phishing Attack Detection' , 2019
[6] NedaAbdelhamid, Aladdin Ayesh, FadiThabtah,Phishing detection based associative classification data mining, 2014
[7] Immadisetti Naga VenkataDurga Naveen, Manamohana K, RohitVerma,Detection of Malicious URLs using Machine Learning Techniques, 2019
[8] Wa'elHadi, Faisal Aburub, SamerAlhawari,A new fast associative classification algorithm for detecting phishing websites, 2016
[9] Sujata Garera, Niels Provos, Monica Chew, Aviel D Rubin, A framework for detection and measurement of phishing attacks, 2007
[10] Hiba Zuhair, Ali Selamat, MazleenaSalleh,Feature selection for phishing detection:a review of research, 2016
[11] Dharmaraj R. Patil, Malicious URLs Detection Using
Decision Tree Classifiers and Majority Voting Technique, 2018.
[12] Christopher N. Gutierrez†, Taegyu Kim†, Raffaele Della Corte, Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks.
[13] VyacheslavLyashenko, Oleg Kobylin, MykytaMinenko, Tools for Investigating the Phishing Attacks Dynamics.
[14] RishikeshMahajan,IrfanSiddavalam, Phishing Webite Detection using Machine Learning Algorithms, 2018
[15] Jitendra Kumar Jaiswal, Rita Samikannu,Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression