International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Phishing Detection Using Machine Learning Based on URL's

Ritika Verma[1], Muskaan Singh[2], Aarti Goswami[3]

[1,2,3]*Dept. of CSE Engineering, MIET college, UP, India*

Abstract- *Phishing is described as gaining private information from a user via hacking into an affiliate's website. To combat phishing, a variety of strategies have been offered. This menace, however, cannot be eliminated by a single miraculous bullet. Data mining is an effective method for detecting phishing assaults. An intelligent approach to identifying phishing attempts is shown in this article. We employ a variety of data mining techniques to classify websites as real or fraudulent. To construct an accurate intelligent phishing analysis system, many categories are employed. The performance of* ***data mining*** *approaches was assessed using classification accuracy,* ***ROC*** *(area under receiver) operating characteristic (AUC) curves, and F-size. The results demonstrate that Random Forest performs the best among the categorization algorithms, with a* ***97.36*** *percent accuracy rate.* ***Random forest algorithm*** *is very fast and can handle various phishinganalysis sites.*

1. Introduction

A phishing URL is created to get the personal data of the user, such as usernames and passwords, or to attack or send some malicious data to the user's system. Ideally, the attacker manipulates the user to click the links and get sensitive information. Phishers can clone the legitimate link data to trick the user into filling in the sensitive information. Phishing links can be used to get a user's confidential details also. This is a very difficult condition for users. The phisher can misuse it for personal gain. According to a survey, phishing attacks are increasing day by day. Therefore, a lot of effort has been made in this area to minimize these phishing attacks. By viewing the content of a website or web page, or using URL metadata, we can determine if the site is a phishing site or not. In our project, we deal with website URL metadata, whether it is a phishing site or not. By using metadata in the URL, we no longer need to attempt phishing websites or download any of their content, making it much more secure access.

We can look into certain parts of the URL, such as the number of slashes, keywords in part of the URL path, etc. After getting the necessary information, we only need information data about a series of URLs to be classified using some algorithms. In our project, we used the Support Vector Machine (SVM) and Random Forest algorithms.

2. Methodology

We used machine learning in our project to deal with the phishing attack problem. Since we have a large amount of data about phishing attack patterns, it can be a good application of the machinelearning approach. Our idea is to use basic **ML algorithms** on the pre-defined dataset to deal withphishing detection in real-time. Since we aimed to deploy the model in real-time we decided to create a **web extension** with the help of JavaScript. Also, we deployed the ML model into a **chatbot** built using python so that, a person can also detect phishing by sending a URL to the bot.

For building the model that can be deployed in real-time, we focused on three parameters; first of all, to choose a dataset and train the model in such a way that its accuracy must be high so that end-user won't get false results. Secondly, since we want real-time protection we need to choosethe algorithm such that it won't take a longer time to execute and the user must not wait for a longer time for getting the result. And lastly, our dataset must have false positive (the website thatseems like phishing but they aren't) and true positive (really phishing sites) URL data in it, so that the ML model can be trained very well and end-user can get true results.

Fig 1: Phishing Detection

2.1 DATASET

To train the ML model, we used the dataset from the "UCI Machine Learning Repository", named 'Phishing Website Dataset'. The dataset has 11,055 URLs in it. Out of 11,055 URLs, 6157 are phishing URLs, and the rest (4898) are legitimate URLs. Each data contains 30 features in it, these features are nothing but the rule which defined the website as a phishing website or legitimate URL. Each feature has 3 different values, '1' if the rule is satisfied, '-1' if the rule is not satisfied, and'0' if it is partially satisfied.

- Our model works with the following 30 rules:

- having_IP_Address
- URL_Length
- Shortining_Services
- having_At_Symbol
- double_slash_redirecting
- Prefix_Suffix
- having_Sub_Domain
- SSLfinal_State
- Domain_registeration_length
- Favicon
- port
- HTTPS_token
- Request_URL
- URL_of_Anchor
- Links_in_tags
- SFH
- Submitting_to_email
- Abnormal_URL
- Redirect
- on_mouseover
- RightClick
- popUpWidnow
- Iframe
- age_of_domain
- DNSRecord
- web_traffic
- Page_Rank
- Google_Index
- Links_pointing_to_page
- Statistical_report
- Result

These features can be classified into 4 different categories:

- URL based rules
- Abnormal based rule

- HTML and programming based rule
- Domain features

URL based features

- Using IP address: If the given URL contains suspicious IP, we can consider it as a phishing URL eg: http:125.98.3.123fake.html or ttp:x58.0xCC.0xCA.0x622paypal.caindex.html
- Long URL: according to the previous observance, it's found that generally, a URL with a long length hides something suspicious in it.
- Use of URL shortening services: Generally attackers hide long URLs behind the short URL by using a shortening tool. A web page that uses a URL shortening service such as Tiny URL is highly suspicious and is likely to be a phishing attempt.
- Use of "@" symbol: The "@" symbol is a reserved keyword in Web standards. So the presence of "@" in a URL is suspicious.
- Redirection with "//": The presence of "//" in the URL path indicates the page will be redirected to another page. If the "//" is placed after 7th place then it is a phishing site.
- Adding prefix or suffix: attackers adds prefix or suffix to legitimate domain with"-" very suspicious symbol Eg: www.a-paypal.com
- Subdomains: If URL has a high value of dots in the domain, it's considered phishing.

Abnormal based rule

- Request URL: Geniune websites load their objects like images, animations, files, etc. from external sources, which be accessed by a request URL that shares the same domain as the web page URL. But if the request URL has a different domainthan the URL, it's considered phishing.
- Anchor tag and links inside them: If the domain in the anchor tag's links matcheswith the domain of the URL given, it's genuine.
- Server Form Handler (SFH): When a form is submitted, some valid action must be taken. So if the action handler of a form is empty or "about: blank" or if the domain of the action URL is Unique from the domain of the main URL, then it istaken as a phishing site.

HTML and Programming based features

- Status bar customization: attackers can alter the status bar using codes to showa legitimate URL. By
- Analyzing the events in the web page we can detect if such a modification hasoccurred.
- Disabling right-click option: Phishers can turn off the right-click option to resistthe user to inspect code.
- Using pop-up window: genuine URLs rarely take user information in the popup window, whereas phishing sites generally use pop-up windows to get user info.
- I-frame redirection: attackers also use I-frame tags with invisible borders to getuser info and redirect to the Real site.

Domain Features

- Domain age: generally the phishing URLs are not much older, so we candetermine the age of the domain check if it's genuine or not.
- Web Traffic: we can check the traffic flow in the URL, as the phishing URLs don'thave many users.
- Page Rank: we can find the page rank of URL on a search engine, if the page rankis too low, it's considered phishing.

3. ML Implementation

We've trained and evaluated supervised ML algorithms with our dataset. We tried training with four different algorithms and tested the accuracy score of each algorithm and choose the best algorithm for our model. We split the dataset into 7:3, 70% for training and 30% for testing. The results of each algorithm are mentioned below:

- KNN_Accuracy: 0.93
- Naive_Accuracy: 0.588
- DecisionTree_Accuracy: 0.95
- SVM_Accuracy: 0.92

Since the decision tree got the highest accuracy score, we chose it as our final model

4. Prediction

Taking URL as an input and loading the model, & using the prediction function.

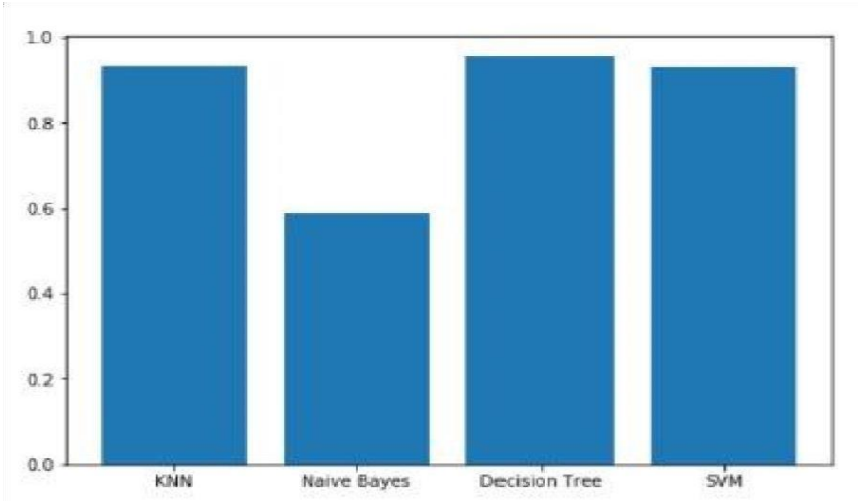

Fig 2: Accuracy Graph

CONCLUSION

Thus concluding the research we have seen how the phishing attack is a very important concern incyber security. We also reviewed some old methods for phishing detection and their drawbacks. We trained and tested four different supervised ml algorithms using a dataset from 'The UCI machine learning repository and showed their results. We selected the Decision tree algorithm asbest based on its speed and accuracy. Our proposed tool can easily be deployed and used in real-time for the detection of phishing URLs and preventing **cyber attacks**.

Future Scope

The existing project can achieve more security and we can extend it to detect phishing with higheraccuracy in the future, we can create an app filter that can scan app data before installation. We can check the app Data like the permission that the app is asking for and reviews of users to see whether the app is good or not, also according to search majority of the phishing links are sharedvia emails to the target user. So it is necessary to add a filter over emails also. Hence we can add filter phishing detection filter also and check all emails before clicking any links in it.

REFERENCES

[1] Microsoft, Microsoft Consumer safety report. Available at https://news.microsoft.com/ensg/2014/02/11/microsoft-consumersafety -index- revealsimpact -of -poor- online- safety behaviours -in Singapore/ sm. 001xdu50tlxsej 410r11kqvksu4nz

[2] Internal Revenue Service, IRS E-mail Schemes. Available at https://www.irs.gov/uac/newsroom/consumers-warned-of-new-surge in-IRS-email-schemes during-2016-tax-season-tax-industry-also-targeted

[3] Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S. (2007), A comparison of machine learning techniques for phishing detection. Proceedings of the Anti-phishing Working Groups 2nd Annual E-Crime Researchers Summit on-E-Crime '07. doi:10.1145/1299015.1299021

[4] E., B., K., T. (2015)., Phishing URL Detection: A Machine Learning and Web Mining-based Approach. International Journal of Computer Applications,123(13), 46-50. doi:10.5120/ijca2015905665

[5] Wang Wei-Hong, L V Yin-Jun, CHEN Hui-Bing, FANG Zhao-Lin., A Static Malicious Java-scriptDetection Using SVM, In Proceedings of the 2nd International Conference on Computer Science and Electrical Engineering(ICCSEE 2013)

[6] Ningxia Zhang, Yongqing Yuan, Phishing Detection Using Neural Network, In Proceedings of International Conference on Neural Information Processing, pp. 714–719. Springer, Heidelberg (2004)