Detecting E-Commerce Phishing Website by Data Mining

Rishabh Jain [1] and Narinder Kaur [2]

[1]Department of I.T., Maharaja Agrasen Institute of Technology, GGSIPU, New Delhi, India

Abstract: A fraud attempt is made to get sensitive and personal information like password, username, and bank details like credit/debit card details by looking as a reliable organization in electronic communication. The phishing website will appear the same as the legitimate website and directs the user to a page to enter personal details of the user on the malicious website. Through machine learning algorithms one can improve the accuracy of the prediction of such phishing websites. The proposed method predicts the URL based phishing websites based on features defined by W3C and also gives us the maximum accuracy. This method uses URL features. We identified features that phishing site URLs contain. The given method employs those features for phishing detection.

Keywords:- Phishing, Algorithm, Legitimate, Prediction.

DOI of the Article: https://doi.org/10.46501/IJMTST0706013

Available online at: http://www.ijmtst.com/vol7issue06.html

As per UGC guidelines an electronic bar code is provided to seure your paper

To Cite this Article: Rishabh Jain; Narinder Kaur. Detecting E-Commerce Phishing Website by Data Mining. *International Journal for Modern Trends in Sceicen and Technology* 2021, *7*, 0706087, pp. 75-77, doi: 10.46501/IJMTST0706013

Article Info. Received: 29 April 2021; Accepted: 2 June 2021; Published: 8 June 2021

INTRODUCTION

Phishing is a cyber crime and the reason behind the phishers doing this crime is that it is very easy to do this, it does not cost anything and it is very effective. The phishers can easily access the email id of any person it is very easy to find the email id now a days and you can send an email to anyone for free across the world. These phishers put very small cost and efforts to get valuable data quickly and easily. The phishing frauds leads to malware infections, loss of data, identity theft, money theft etc. The data in which these cyber criminals are interested is the crucial information of a user like the account password, OTP, credit/ debit card numbers CVV, sensitive data related to business, medical data, confidential military data etc. Sometimes these criminals also gather information which can give them direct access to the social media accounts of users to their emails.

DATASET

The data of urls is obtained from Phish tank website ,where Phishtank is an anti-phishing site.It contains 2905 urls which is in unstructured form. Our main objective is to detect whether the url is phishing or legitimate.This dataset contains few website links (Some of them are legitimate websites and some are fake websites).Pre-Processing this data before building a model and also Extracting the features from the data based on specified conditions. We need to split the data according to features of the URL.

MODEL USED

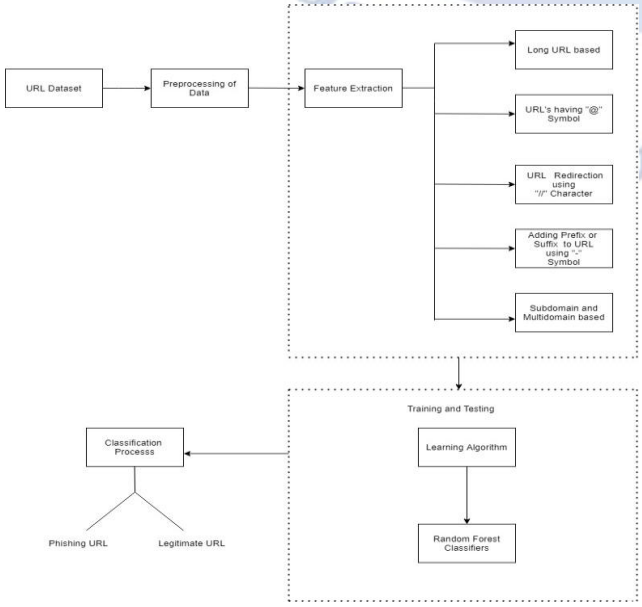


Fig. 1: Proposed System block diagram

We have collected unstructured data of urls from Phishtank website.In preprocessing ,feature generation is done where 5 features are generated from unstructured data. These features are length of url, has prefix/suffix, number of dots, number of slash, length of sub domain.After this a structured dataset is created in which each feature contains binary value(0,1) which is then passed to the different classifiers.Next we train the Random Forest classifier and compare their performance on the basis of accuracy.Then classifier detects the given url based on the training data that is if the site is phishing it shows a spoofed website as yes or no.

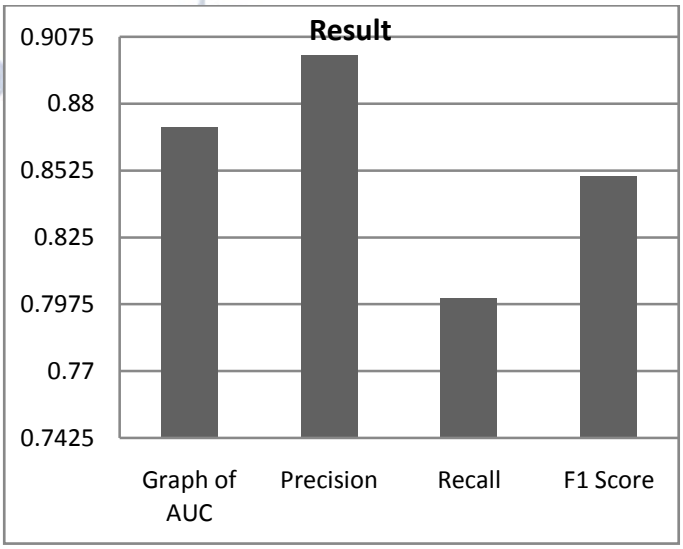
OBSERVATIONS AND RESULTS

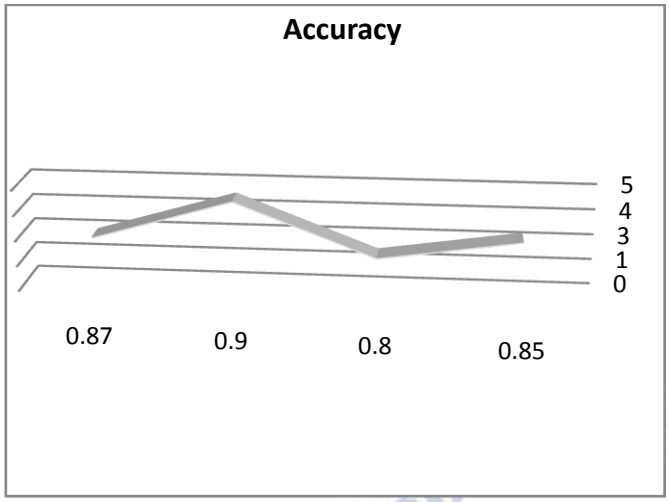
As discussed in the earlier sections, we have used one classifier to predict and detect if the website is phishing or legitimate.

Classifiers	Precision	Recall	F1	AUC	Accuracy (%)
Random Forest Classifier	0.90	0.80	0.85	0.87	85.6

RESULTS

We have got the desired results of testing the site is phishing or not by using random classifiers. Refer the graph below for the exact results. In the graph shown, shows the AUC, precision, recall and the F1 score obtained by using classifier.





The graph shown, explains about the accuracy obtained by using different classifiers in the histogram graphical representation

CONCLUSION

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through such phishing websites, it becomes more critical to take care of this issue. This problem can be resolved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing website, but after our survey it is found that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also we generated classifiers through various machine learning algorithms. We have found that our system provides us with 85.6 % of accuracy for Random Forest Classifier. The proposed technique is much more reliable as it detects new and previous phishing sites.

REFERENCES

1. Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.
2. Rao, R. S., & Pais, A. R. (2019). Jail-Phish: An improved search engine based phishing detection system. Computers & Security, 83, 246-267.
3. Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages. computers & security, 84, 256-275.
4. Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016, June). Know your phish: Novel techniques for detecting phishing sites and their targets. In 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) (pp. 323-333). IEEE.
5. Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.
6. Rathod, J., & Nandy, D. Anti-Phishing Technique to Detect URL Obfuscation.
7. Hodžić, A., Kevrić, J., & Karadag, A. (2016). Comparison of machine learning techniques in phishing website classification. In International Conference on Economic and Social Studies (ICESoS'16) (pp. 249-256).
8. Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.

