International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 1, ISSN No. 2455-2143, Pages 133-137
Published Online May 2021 in IJEAST (http://www.ijeast.com)

COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR PHISHING WEBSITE DETECTION

Anuraag Velamati
School of Computer Science and Engineering
Vellore Institute of Technology,
Vellore, India

Abstract-Phishing is the most commonly approached cyber-attack in this modern era. Through such attacks, the phisher will target the innocent users by tricking them into revealing their secure and personal information, with the purpose of using it fraudulently. In order to avoid getting phished, users should have awareness of phishing websites, have a blacklist of phishing websites which requires the knowledge of website being detected as phishing.

Keywords: **Phishing website, machine learning, URL, data.**

I. INTRODUCTION

A. Relevance to the practical field:

As cyber-crime has been on a prowl all over the internet in the recent times [35]. My research would help a user in detecting a phishing website which would further help him protecting his personal details from getting exposed [2]. As it was a website the user can easily access our approach. The main objective of this paper was to detect them in their early stage, using both machine learning and deep learning [1]. Of the above three, the machine learning based method is proven to be most effective than the other methods. Even then, online users are still being trapped into revealing sensitive information in phishing websites[34].

B. Importance of the proposed methodology:

For our project,5000 URLs of both the legitimate and the phishing URL's are randomly picked and trained using machine learning algorithms which provides the more accurate information about the phishing URL. It is independent on the third party [3]. This model is cheap and easily accessible to anyone.

It works on the real time environment as it is trained based on several feature selection so even if new phishing website created it can detect up-to some extent.

II. BACKGROUND STUDY

A. Background Study-

Coming to the Background study of the project we had to refer to many websites, tutorials etc. to concluding on flags that have to be used for the data preprocessing in addition to this multiple books and videos have been referred for understanding and deciding not he machine learning algorithms that have to be used for our project [4]. For the real-time integration of the project the official documentation of the frame work has been read in order to understand it's functionality.

B. Limitations of Previous study-

Constructing a dataset for the anti-phishing system is a trivial issue [5]. There are some web-based services, which give URLs of the phishing web pages. However, they share a limited amount of data in their web pages. The existing models are not cost effective and require the good configuration of the device to run the model [28]. Since creating a web page is usually considered as an easy and cheap task, for phishing the users, an attacker can quickly create a webpage that is fraudulent, which will have a very small lifespan. Therefore, the detection of these phishing websites at their early stages itself is very important for saving the information from getting stolen from a user [6]. But this thing lacks in the current existing model.

C. Objectives of proposed methodology-

One of the most common social engineering methods that are on a prowl in this century is a phishing website.

A phishing website is basically a website, that mimics trustful uniform resource locators (URLs) and webpages [7]. The objective of our paper is to test and compare some of the most renowned machine learning algorithms and usage of neural networks on the dataset that was created by us, in order to detect the phishing websites [8]. The dataset has been prepared by looking at multiple sources which stated about the multiple flags that we must place in order to decide a website as phishing website[9].

The accuracies achieved by each model will be measured and a table will be provided stating their respective training and testing accuracies. In orders increase the functionality of the project the best classifier is saved [29]. The best classifier model saved and the website has been developed and linked to this model using flask, in order to improve their real time approach inaccessibility [10].

D. Exact definition of the problem:

As of now many phishing website detectors has come in to the technology but their efficiencies are not high and accurate [30]. The algorithms usage and data preprocessing would have been major issues during their respective implementations of the projects [11]. In our implementation we have done our own data preprocessing that we have learnt from our research and performed the comparative analysis using appropriate relatable algorithms [12]. After the data preprocessing is done and they are linked with the relatable machine learning algorithms, their accuracy scores are evaluated and compared how they performed when compared to the other algorithms [31].

III. METHODOLOGY

A. Methodology adopted:

Data Collection : Legitimate URLs are collected from the dataset provided by the University of New Brunswick [13]. Phishing URLs are collected from ‘PhishTank’. The main benefit of using ‘PhishTank’ is that the format of dataset availability is vast.

Feature Selection: After the Data collection feature selection process has to be done. For the feature selection, multiple categories of features are taken into consideration. The main features that are taken into consideration are the address Bar based Features, Domain Based Features and webpage dependent features [14].

Machine Learning Models Training: This comparative analysis comes under the taskof classification problem, as the output label is classified as phishing (1) or legitimate (0)[15]. Thus, taking this into consideration, the machine learning models that are considered for this task are Random Forest, Decision Tree, XGBoost, Multilayer perceptron, Support Vector Machines and autoencoder neural networks [16].

Model Evaluation: The models are evaluated, and the considered metric is accuracy and the F1 score [26]. All the above ML algorithm accuracy on the data set will be calculated and the highest will be stored and will selected for the further deployment of that model. Real Time Integration: The best Classifier is saved and linked to the developed website for a client to use our product in real time [17].

B. Data Collection:

Legitimate URLs are collected from the dataset provided by University of New Brunswick. From the total dataset, 5000 URLs are randomly picked [18].

The Phishing URLs are downloaded from a website called Phishtank. This service provides phishing URLs in different formats such as csv, json etc. The best part about this service is that the data that is being provided gets updated hourly [19]. This dataset is available to public; thus, it was used for this research. Form the obtained phishing URL dataset’s collection, 5000 URLs are randomly picked. After the URL data is selected the dataset is created using feature selection [32].

After the Data collection feature selection process has to be done. The following features are selected for the feature selection flags [20].

HTML & Javascript based Feature Address Bar based Features considered are ‘Domian’ of URL, Redirection ‘//’ in URL, IP Address in URL, ‘http/https’ in Domain name, ‘@’ Symbol in URL,Using URL Shortening Service , Length of URL , Prefix or Suffix "- " in Domain ,Depth of URL [21].

Domain based Features considered are, DNS Record, Age of Domain, Website Traffic, End Period of Domain HTML and JavaScript based Features considered are, Iframe Redirection, Disabling Right Click, Status Bar Customization, Website Forwarding [22].

After extracting all the features, 17 independent variables were collected for the utilization. The distribution has been provided in the below [27]. The distributions state about the major numerical value presence of the data in every separate individual variable.

The visualization is developed depending on their frequency of variables in the data. For the convenience of easier understanding, we have provided a histogram as a method for visualizing the data [23].

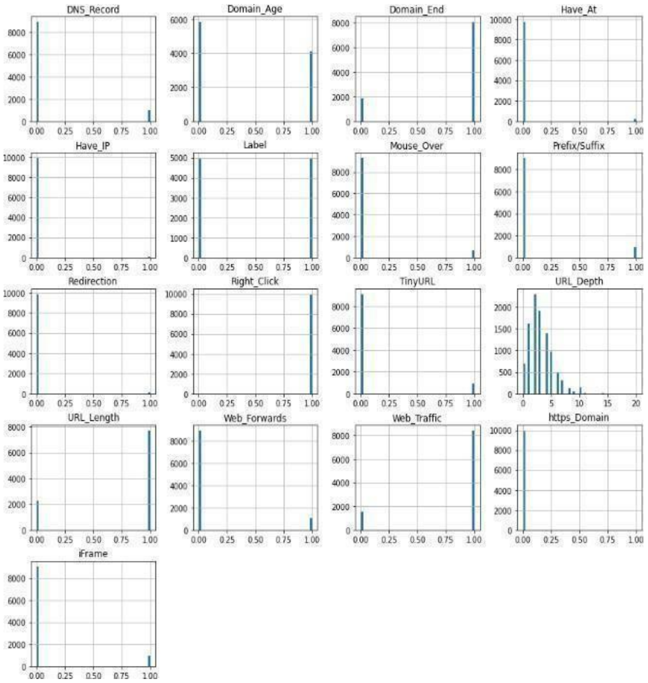

Fig. 1 – Feature Distribution of the dataset

Analytical computation and tools used: For the implementation and execution of the project multiple software and analytical tools.
Software Tools: VS code, Anaconda prompt,

Anaconda Navigator, Jupyter NoteBook.

Analytical Tools: Numpy, Pandas, Tensorflow , Sci-kit learn

, Keras, Matplot- Lib.

IV. RESULTS

A. Results:

After the training and testing of the machine learning algorithms the comparison between the algorithms has been made by gating their accuracy scores on both the training and testing data in to considerations[24]. The accuracy scores have been shown of the model have been presented below.

B. Interpretation of the results:

According to our project we thought of detecting a phishing website using machine learning models XGBOOST,
Random forest, decision Tree, Autoencoder model, SVM, Multilayer Perceptrons and the machine models have been trained using dataset [25]. Our aim was to achieve a model with an efficient accuracy which we have achieved successfully after the implementation of the project.

C. Inferences from the results:

The inference for our project would be the accuracy table which we have developed at the end of our implementation. This result will help us in understanding the most prominent algorithm that can be used for the task of detection.

ML MODEL	TRAIN ACCURACY	TEST ACCURACY
XGBoost Classifier	0.868	0.851
Multilayer Perceptrons	0.866	0.85
AutoEncoder	0.817	0.818
Random Forest	0.826	0.807
Decision Tree	0.817	0.797
Support Vector Machines	0.805	0.792

Table 1 -Accuracy Scores of models

V. SUMMARY

A. Summary:

After our research, implementation and execution we have come to the conclusion that there are quite a few independent variables in order to classify whether a website is legitimate or not. after the implementation of the machine learning models we have also inferred from the results that the working of machine learning algorithm will be affected widely in it's working depending on the data set(the number of independent and dependent variables).

Out of all of the machine learning models used XGBOOST classifier worked very efficiently compared to the other's that have been considered for our comparative analysis. The website that was developed as a part of a project was also working very effectively in detecting the fishing website URL's.

B. Conclusion:

Through the development of this research not only have I learnt about Machine learning but also the leading frame works which are used in the industry currently. This research is also helpful for me in my future researches continuing in this domain.

C. Scope for future study:

As we have already a real time implementation in our project the scope for future work for our project would be creating a GUI or web extension which would help our user if he accesses any phishing websites by any chance [33]. The efficient of our product can be increased drastically provided your given access to the current fishing website data collection. As cyber- crime is a very prominent in our generation. The scope of future work for this project is perennial.

VI. REFERENCE

[1] Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning algorithms for phishing website detection. In Proceedings of the anti-phishing working groups 2nd annual ecrime researchers summit, eCrime '07, ACM, New York, NY, USA (pp. 60–69).

[2] Abdelhamid N, Ayesh A, Thabtah F (2014) Phishing detection based associative classification data mining. Expert Syst Appl 41:5948- 5959.

[3] Buber, E., Diri, B., & Sahingoz, O. K. (2017). NLP based phishing attack detection from URLs. In International Conference on Intelligent Systems Design and Applications (pp. 608-618).

[4] Babagoli, M., Aghababa, M. P., & Solouk, V. (2018). Heuristic nonlinear regression strategy for detecting phishing websites Soft Computing(pp.1-13)

[5] Buber, E., Diri, B., & Sahingoz, O. K. (2017). Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) (pp. 337– 342). 28.

[6] In A. Abraham, P. K. Muhuri, A. K. Muda, & N. Gandhi (2019), Intelligent systems design and Applications, springer.

[7] Cao, Y., Han, W., & Le, Y. (2008). Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM workshop on Digital identity management (pp. 51-60).

[9] Barraclough, P. A., Hossain, M. A., Tahir, M. A., Sexton, G., & Aslam, N. (2013). Intelligent phishing detection and protection scheme for online transactions. *Expert Systems with Applications*, *40*(11), 4697- 4706.

[10] He, M., Horng, S. J., Fan, P., Khan, M. K., Run, R. S., Lai, J. L., ... & Sutanto, A. (2011). An efficient phishing webpage detector. *Expert systems with applications*, *38*(10), 12018- 12027.

[11] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review*, *17*, 1-24.

[12] Hadi, W. E., Aburub, F., & Alhawari, S. (2016). A new fast associative classification algorithm for detecting phishing websites. *Applied Soft Computing*, *48*, 729-734.

[13] Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Predicting phishing websites using classification mining techniques with experimental case studies. In *2010 Seventh International Conference on Information Technology: New Generations* (pp. 176-181). IEEE.

[14] Abdelhamid, N. (2015). Multi-label rules for phishing classification. *Applied Computing and Informatics*, *11*(1), 29-46.

[15] Mohammad, R., McCluskey, T. L., & Thabtah, F. A. (2013). Predicting phishing websites using neural network trained with back-propagation. World Congress in Computer Science, Computer Engineering, and Applied Computing.

[16] Gowtham, R., & Krishnamurthi, I. (2014). A comprehensive and efficacious architecture for detecting phishing webpages. *Computers & Security*, *40*, 23-37.

[17] Pan, Y., & Ding, X. (2006). Anomaly based web phishing pagedetection. In *2006 22nd Annual Computer Security Applications Conference (ACSAC'06)* (pp. 381-392).

[18]Ramesh, G., Krishnamurthi, I., & Kumar, K. S. S. (2014). An efficacious method for detecting phishing webpages through target domain identification. *Decision Support Systems*, *61*, 12-22.

[19] Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2008). An evaluation of machine learning-based methods for detection of phishing sites. In *International Conference on Neural Information Processing* (pp. 539-546).

[20] Khonji, M., Iraqi, Y., & Jones, A.(2011). Lexical URL analysis for discriminating phishing and legitimate websites.

[21] Bergholz, A., Chang, J. H., Paass, G., Reichartz, F., & Strobel, S. (2008). Improved Phishing Detection using Model- Based Features. In *CEAS*.

[22] Gansterer, W. N., & Pölz, D. (2009). E-mail classification for phishing defense. In *European Conference on Information Retrieval* (pp.

449-460).

[23] Zhang, D., Yan, Z., Jiang, H., & Kim,T. (2014). A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. *Information & Management, 51*(7), 845-853.

[24] Feroz, M. N., & Mengel, S. (2015). Phishing URL detection using URL ranking. In *2015 ieee international congress on big data* (pp. 635-638). IEEE.

[25] Chen, J., & Guo, C. (2006). Online detection and prevention of phishing attacks. In *2006 First International Conference on Communications and Networking in China* (pp. 1-7).

[26] Olivo, C. K., Santin, A. O., & Oliveira, L. S. (2013). Obtaining the threat model for e-mail phishing. *Applied soft computing, 13*(12), 4841-4848.

[28] Khonji, M., Jones, A., & Iraqi, Y.(2011). A study of feature subset evaluators and feature subset searching

[29] Afroz, S., & Greenstadt, R. (2011). Phishzoo: Detecting phishing websites by looking at them. In *2011 IEEE fifth international conference on semantic computing* (pp. 368- 375).

[30] Blum, A., Wardman, B., Solorio, T., & Warner, G. (2010). Lexical feature based phishing URL detection using online learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security* (pp. 54-60).

[31] Medvet, E., Kirda, E., & Kruegel, C. (2008). Visual-similarity-based phishing detection. In *Proceedings of the 4th international conference on Security and privacy in communication netowrks* (pp. 1-6).

[32] Miyamoto, D., Hazeyama, H., & Kadobayashi, Y. (2007). A proposal of the AdaBoost-based detection of phishing sites. In *Proceedings of the joint workshop on information security*.

[33] Prakash, P., Kumar, M., Kompella, R. R., & Gupta, M. (2010, March). Phishnet: predictive blacklisting to detect phishing attacks. In *2010 Proceedings IEEE INFOCOM* (pp. 1-5).

[34] Wardman, B., & Warner, G. (2008). Automating phishing website identification through deep MD5 matching. In *2008 eCrime Researchers Summit* (pp. 1-7).

[35] Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications, 31*(8), 3851- 3873.

[36] aberi, A., Vahidi, M., & Bidgoli, B. M. (2007). Learn to detect phishing scams using learning and ensemble? methods. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops* (pp. 311-314).

[37] Xiang, G., & Hong, J. I. (2009). A hybrid phish detection approach by identity discovery and keywords retrieval. In *Proceedings of the 18th international conference on World wide web* (pp. 571- 580)

[38] Ludl, C., McAllister, S., Kirda, E., & Kruegel, C. (2007). On the effectiveness of techniques to detect phishing sites. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment* (pp. 20-39)