INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 07 ISSUE: 06 | JUNE - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

PHISHING DETECTION USING ML

Nandini Gaikwad
Electronics & Telecommunication Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, India
nandinigaikwad10@gmail.com

Jignesh Shimpi
Electronics & Telecommunication Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, India
jigneshshimpi77@gmail.com

Krishna Gaggad
Electronics & Telecommunication Engineering
JSPM's Rajarshi Shahu College of Engineering
Pune, India
kgaggad@gmail.com

Abstract — Phishing URLs encourage online druggies to pierce fake websites, and reveal their nonpublic information, similar as disbenefit/ credit card figures and other sensitive information and also use the same. These URLs substantially target individualities and/ or associations through attacks by exploiting sins of human's in information security mindfulness. In this work, we introduce a phishing discovery fashion grounded on URL analysis and machine literacy classifiers. The trials were carried out on a dataset that contained 5000 labeled URLs(phishing and licit). This dataset was reused to induce 3 different features that were reduced further to a lower set using different features reduction ways. Random Forest, Decision Tree, XGBoost, Support Vector Machine(SVM) Auto Encoder Neural Network and Multilayer Perceptron classifiers were all estimated, and results show the superiority of XGBoost, which achieved the loftiest delicacy in detecting the anatomized URLs with a rate of 86.8%. Our approach can be incorporated within add- on/ middleware features in Internet cybersurfs for waking online druggies whenever they try to pierce a phishing website using only its URL.

***Keywords*— Neural Network, Random Forest, SVM, GBC, Machine Learning, URL Analysis, Phishing Detection**

I. INTRODUCTION

The use of the Internet has come essential in the ultramodern period and an integral part of technological development, which leads to discoveries and reduction of time, trouble, and costs. [1]With the steady acceleration in information technology, we're no longer vulnerable to being victims of cybercrime. nonetheless, this provides a rich ground for pirating expansion in exploiting the sins to determine private and public interests. Creating a paperless terrain has come a major focus in utmost countries worldwide, adding reliance on these channels.

On the other hand; vulnerable websites may allow fake advertisement exploits under circumstances that enthrall public opinion. [2]This leads the victim to a phishing website. In this environment, individualities' lack of mindfulness in information security plays a crucial part in adding the number of victims of this crime. Although cyber crime doesn't differ much from traditional crimes in terms of its perpetrators' thing, because these crimes are grounded on unlawful targets, cybercrime has come more wide than traditional crimes. It has come a core part in the world of digitization, as multinational crimes within cyberspace

On the other hand; vulnerable websites may allow fake advertisement exploits under circumstances that enthrall public opinion. This leads the victim to a phishing website. In this environment, individualities' lack of mindfulness in information security plays a crucial part in adding the number of victims of this crime.[5] Although cyber crime doesn't differ much from traditional crimes in terms of its perpetrators' thing, because these crimes are grounded on unlawful targets, cybercrime has come more wide than traditional crimes. It has come a core part in the world of digitization, as multinational crimes within cyberspace.

Digital cybercrimes have no limits and are easy to apply. colorful forms of attacks are distributed grounded on their exertion, similar as Phishing, Session Hijacking, Malware/ Trojan, Screenlogger/Keyloggers. etc.(2).[3] Edge security tools are unfit to descry similar attacks since they target end-druggies rather than systems. There have been numerous studies lately trying to come up with a suitable result for detecting phishing URLs. These results can be grouped into generally four groups predefined list, hand- grounded, content based, and machine literacy. Phishing attacks start when the bushwhacker clones content or runner design from licit spots also reconstructs the scamming webpage. Once it's ready, the phisher sends a URL to the victim, soliciting him to fill the fiddle webpage with non public information.[6] Eventually, the phisher steals the victim's information to pierce the original point.

The machine literacy approach substantially depends on the capability to learn the characteristics of the websites that are listed under the phishing order, also enforcing the vaticination capability to distinguish the licit websites from the faked bones in the means of the different machine literacy ways.

A phishing website is a common social engineering system that mimics trusting invariant resource locators(URLs) and webpages. The ideal of this design is to train machine literacy models and deep neural network on the dataset created to find phishing websites. [4]Both phishing and fake URLs of websites are gathered to form a dataset and from them needed URL and website content- grounded features are uprooted. The performance position of each model is measures and compared.

II. METHODOLOGY

A. AIM :-

To develop an Machine Learning URL Phishing Detection application so that common people would not get scammed by Scammers.

B. OBJECTIVES :-

The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared.

C. BLOCK DIAGRAMS :-

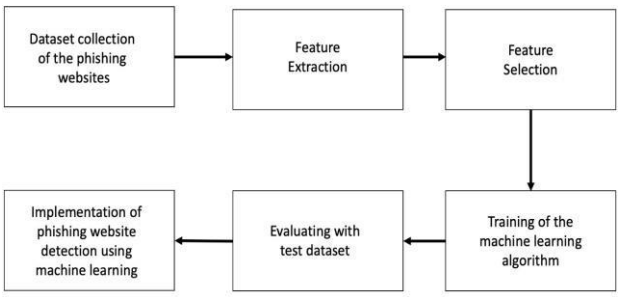

FIG-1 Phishing Detection Block Diagram

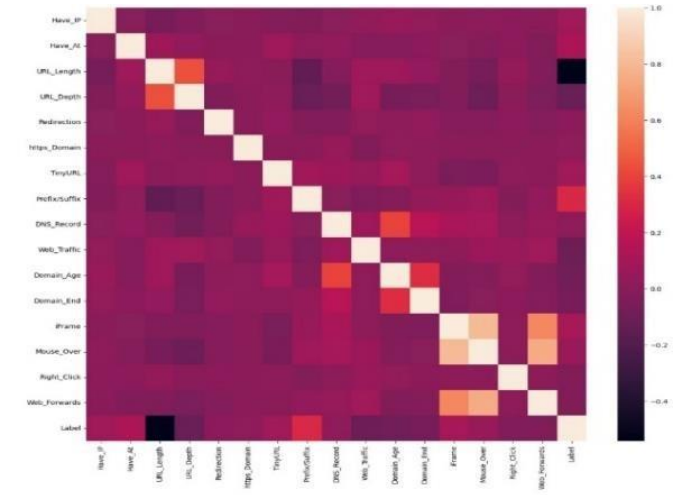

FIG-2 Correlation HeatMap

Fig1 describes the high- position methodology and the following way in this work. Also, for each step in the figure, the following subsections explain the details.

A. Dataset

Originally, we used the dataset available from open source platforms. The dataset contains only labelled URLs that are collected from different sources. It firstly contains 5000 URLs that are undressed. either, the markers are decoded with 0 and 1 that represent the " Not-vicious " and " vicious " classes independently.

B. Data Preprocessing

Data preprocessing starts with checking the indistinguishable URLs listed in the dataset. Many indistinguishable records were set up and dropped. Next, a check for missing values was conducted, but no missed values were set up to handle.

C. Feature Extraction

The way followed in negotiating the given URLs dataset's features substantially depend on assaying the URLs lexically. In other words, the URLs parsed and anatomized to induce farther features that reflect the nature of the URLs contained in the given dataset.

In this step, features are uprooted from the URLs dataset. The uprooted features are distributed into:

1. **Address Bar Grounded Features**:

Numerous features can be uprooted that can be considered as address bar base features. Out of them, below mentioned were considered for this design.

- Domain of URL
- IP Address in URL
- "@" Symbol in URL
- Length of URL
- Depth of URL
- Redirection"//" in URL
- " http/ https" in Domain name using URL Shortening Services " TinyURL "
- Prefix or Suffix"" in Domain.

2. Domain Based Features:

Multitudinous features can be pulled that come under this order. Out of them, below mentioned were considered for this design.

- DNS Record
- Website Traffic
- Age of Domain
- End Period of Domain

3. HTML and JavaScript based Features

Multitudinous features can be pulled that come under this order. Out of them, below mentioned were considered for this design.

- IFrame Redirection
- Status Bar Customization
- Disabling right click website forwarding

D. Machine Learning Models & Training

From the dataset over, it's clear that this is a supervised machine learning task. There are two major types of supervised machine knowledge problems, called type and regression.

The input URL is classified as phishing(1) or legit(0) as this data set comes under type problem. The supervised machine knowledge models(type) considered to train the dataset in this tablet are:

- Decision Tree
- Random Forest
- Multilayer Perceptrons
- XGBoost
- Autoencoder Neural Network
- Support Vector Machines

III.CONCLUSION

We achieved 86.8% discovery delicacy using XGBoost algorithm. Also affect shows that classifiers give better performance when we used more data as training data XGBoost (Extreme Gradient Boosting) XGBoost is an open- source software library that implements optimized distributed grade boosting machine knowledge algorithms under the Gradient Boost frame.

In this paper, the URLs' verbal analysis approach was applied, where different machine knowledge models were trained and tested on several features sets. This approach seems to give precious benefits for phishing attacks discovery and prevention, and only depends on the URLs included in the web requests' heads. Also, machine knowledge ways also prove their robustness in the security sphere, as reflected by the results illustrated in this work.

In the future, the work shall continue to dig deeper into involving the web content features along with the real time learning capabilities. This would help further in designing strong precautionary security appliances that can learn and work simultaneously.

ML Model	Train Accuracy	Test Accuracy
XGBoost	0.868	0.857
Multilayer Perceptrons	0.866	0.854
AutoEncoder	0.810	0.810
Random Forest	0.820	0.809
Decision Tree	0.816	0.803
SVM	0.806	0.786

VI .ACKNOWLEDGMENT

We wish to express our sincere thanks to Dr. R.K. Jain, Principal, Rajarshi Shahu College of Engineering, Tathawade, for providing us all the necessary facilities.

We would like to place on record our deep sense of gratitude to Dr. S.C. Wagaj, HOD, Department of Electronics & Telecommunication Engineering, for his stimulating guidance and continuous encouragement.

We also wish to extend our thanks to Dr. Aditi Vedalankar Project coordinator for her supervision throughout the course of present work.

We are extremely thankful to Mrs. Asmita Shirke for her insightful comments and constructive suggestions to improve the quality of project work.

Lastly we are thankful to Teaching and non-teaching faculty of the Department for their continuous co-operation.

V.REFERENCES

1. https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5
2. infosecinstitute
3. Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013
4. Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016
5. dataaspirant
6.kdnuggets
7.phishtank
8. https://www.activestate.com/blog/phishing-url-detection-with-python-and-ml/
9.https://www.ijert.org/phishing-detection-using-machine-learning-based-url-analysis-a-survey
10.https://nevonprojects.com/detecting-phishing-websites-using-machine-learning/
11.https://www.checkpoint.com
12.https://www.imperva.com/learn/application-security/phishing-attack-scam 13.https://www.kaggle.com/datasets/