International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume 9 Issue 12, Dec 2022 www.irjet.net p-ISSN: 2395-0072

Phishing Website Detection Paradigm using XGBoost

Sumera Anjum[1], T. Uma Devi[2], K.K. Namish[3], B. Vasundhara Devi[4]

[1,2,3]*Student, Dept. of Computer Science and Engineering, Sreenidhi Institute of Science and Technology*
[4] *Assistant Professor, Dept. of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Telangana, India*

Abstract - *One of the largest and most potent cyber hazards today is phishing, which costs thousands of millions of dollars in damages resulting from data breaches that happen every year. Due to the frequent change and short lifespan of phishing websites, several pattern recognition approaches have been explored and developed to address phishing attacks, but none of them are effective in detecting web phishing activities. Among the most pragmatic ways to solve this challenge is with machine learning since it can attain statistics and handle the changing nature of online fraud. In this project, we illustrate using an ensemble machine learning technique, the Extreme Gradient Boosting (XGBoost) Algorithm, to detect malicious URLs with high precision and efficacy using the Uniform Resource Locators. In XGBoost, the target variable yi is predicted using training data xi repeatedly until the model's parameters are improved by merging the trees and boosting. As determined by the confusion matrix created by the XGBoost model's performance, it accurately predicted 7393 positive terms and 7930 negative terms with the set of features identified from the Kaggle dataset. Its merits encompass substantial regularisation capabilities that reduce overfitting, great speed and performance since trees are created in parallel, and flexibility because of costume optimization.*

Key Words: **Phishing, Website, XGBoost, ensemble, Extreme Gradient Boosting, Uniform Resource Locator**

1. INTRODUCTION

Upsurge in web users, phishing threats have grown to be a serious problem. More than 80% of security incidents that have been reported entail phishing attacks. These phishing portals are cyber snoopers attempting to gather data covertly by coercing users into divulging private information like their passwords and credit card details. Attackers generally employ spoofing to lure consumers to malicious websites by mimicking the names and designs of trusted websites like Myntra, Flipkart, Amazon, and Zomato. Hence, it is challenging for the common person to tell them apart from legitimate websites. A Uniform Resource Locator (URL) incorporates different components, including the protocol, domain name, port, path, query, etc. A phishing website's URL may be differentiated from authentic ones by using a few specific characteristics. Although, it may not be always reliable to classify a website simply by looking at the URL. Phishers have employed a variety of sophisticated strategies to trick unsuspecting consumers, including the usage of social engineering techniques and technology to offer carefully designed URLs that lead users to believe that websites are trustworthy. There are several approaches to combat phishing, including technological, educational, and legal means, and numerous research on the subject have been conducted. A credible and plausible solution must be provided to avoid jeopardizing the users' privacy. Since the methodologies from machine learning can identify possible threats by learning provided data and building predictive models, it is a viable field to handle the problem in this case. Single models that effectively process the training data and produce substantially accurate predictions are most commonly implemented. The algorithm predominantly is a collection of Decision Trees, which are used by ensemble machine learning approaches to train several categorization models [4]. The final result is generated through a combining method, such as voting (majority wins), weighted voting (certain classifiers have more authority than others), and averaging the results, as each constituent learning algorithm will have its own separate output [4].

2. LITERATURE REVIEW

In this section, we have articulated several well-known examples because extensive study and research have been done on phishing detection. For detecting attacks, there are several methods and a broad range of data types in academic research and commercial services. URL-based, domain-based, page-based, and content-based features gathered from academic research for phishing domain identification through machine learning approaches [10]. Traditional machine learning techniques like Nave Bayes, Support Vector Machine, and Decision Tree were used in the majority of the research on the topic. Software called "Anti Phishing Simulator" was devised at Firat University to make it easier to identify phishing and spam emails by looking at the email content [3]. As encouraged by Cisco, fog computing makes use of features such as uniform resource locator (URL) and internet activity to identify phishing websites based on a designed neuro-fuzzy framework (dubbed Fi-NFN), and an anti-phishing model was created to transparently monitor and defend fog users from phishing attacks [1]. To some extent, approaches based on visual resemblance can identify phishing websites. The majority of web information is not consistent, though, and when a web page's characteristics change, the approach encounters a detection problem. Blacklisting techniques are the basic and most commonly

used strategies in the business to stop phishing assaults [6]. Checking whether the URL of the matched website is on the blacklist is one of the phishing detection techniques used by Google Safe Browsing [6]. CANTINA is a content-based phishing detection method that was proposed by Zhang et al. The first five phrases based on TF-DF are forwarded to the search unit for comparison with the results returned by the search unit utilizing linkable links in the authors' approach for identifying phishing websites [5].

3. PROPOSED SYSTEM

In this part, the working of the proposed solution is explained in which the data collection and its processing is the first step. The processed data is then used to train the model with an ensemble algorithm. Testing data is used to test the accuracy and precision of the model developed which is displayed through the confusion matrix. In the end, the user can enter any URL to classify it as a phishing website or a legitimate website which is generated as output by the XGBoost paradigm.

3.1 DATASET PRE-PROCESSING

The dataset which is used in this project is obtained from Kaggle. Kaggle provides the public dataset consisting of 71677 unique values. This data is fetched from google's whois API which tells us more about the current status of the URL's registration [2]. The first step following deciding on an algorithm is data collection, often known as the requirements stage. Despite, the fact that this step is only beginning, it is the most important and time-consuming. Because the module's main objective is to learn about and apply cutting-edge technology, this section pays particular focus to this component of the project. From four primary categories, 17 factors are taken out and incorporated into the system. The features are extracted and stored in the CSV file. The resulting CSV file is uploaded to this notebook and stored in the data frame.

3.2 MODEL DEVELOPMENT

It's appropriate to construct the model when the essential data has been obtained and examined. The development of the model's architecture, the creation of orderly yet secure codes, and model training comprise the design portion of the project as it is now being presented. Python is being leveraged throughout the project, thus important libraries that are mostly used for data science are imported, and the scripts are either created from scratch or drawn from the web. Extreme Gradient Boost, often known as XGBoost, is a machine learning technique that employs extreme gradient boosting and is based on Decision Trees. The gradient boosting method was improved by integrating parallel processing, tree pruning, missing value handling, and normalization to get rid of errors and inaccuracies [11]. It's a lethal combination of hardware and software metaheuristics that uses the least amount of processing resources while achieving better significant results. The fundamental purpose of this work is to establish certain dataset parameters that the model will use in the future to determine whether a URL is genuine or not. Here, each parameter transforms into a tree and increases the deciding factor [11]. Although these trees might not perform as well as anticipated, by merging these trees and boosting them, the prediction might noticeably enhance. In XGBoost, the target variable yi is predicted using training data xi repeatedly until the model's parameters are improved.

3.3 PHISHING WEBSITE DETECTION

The developed paradigm is saved and tested for accuracy with the testing data. This paradigm can be used in real-time to classify the URLs into legitimate or phishing, given by the user as input.

4. RESULTS

The output screenshots display the user inputs classification and the confusion matrix shows the performance of the XGBoost phishing website identification model.

Fig -1: Phishing website detection output 1

Fig -2: Phishing website detection output 2

```
---------- XGBoost ----------

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.85      0.83      9304
           1       0.84      0.79      0.82      9311

    accuracy                           0.82     18615
   macro avg       0.82      0.82      0.82     18615
weighted avg       0.82      0.82      0.82     18615


Confusion matrix:
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay object at 0x0000027976E1DD60>
```

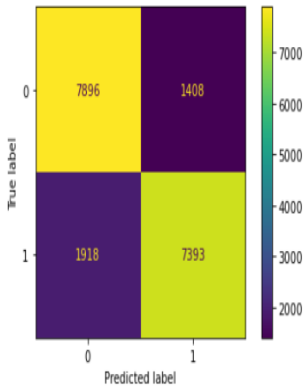
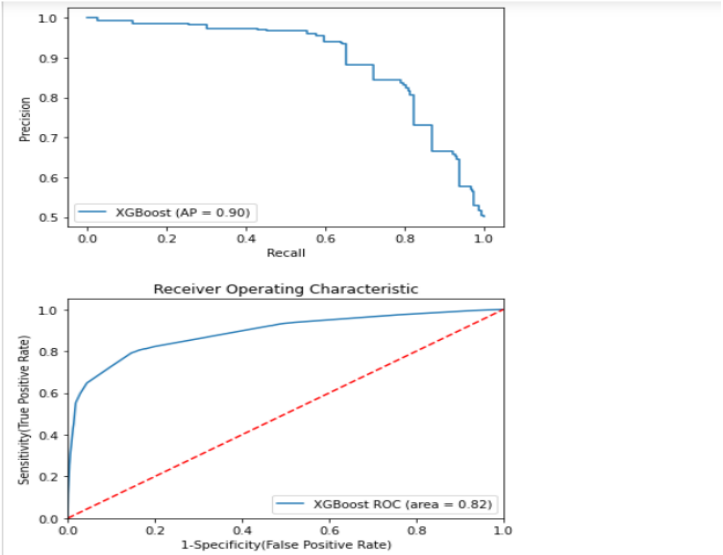


Fig -3: Confusion matrix of the developed XGBoost model



	ML Model	Train Accuracy	Test Accuracy	Kappa Score	Training Time	Testing Time
0	XGBoost	0.821872	0.822777	0.645579	0.753585	0.013743

Fig -4: Precision graph for developed XGBoost model

5. CONCLUSIONS

Today there are more uncontrolled websites than ever before due to a mammoth increase in internet users. Phishing variegates over time since fraudulent websites are often updated and do not last forever. With the use of the Ensemble Algorithm XGBoost and a feature set well stipulated, phishing detection using website URLs is predicted to generate highly accurate results with a reasonable bias-variance trade-off in a robust and efficient manner. According to the above models' assertions, XGBoost Classifier has the finest model performance at 86.4%. The Python pickle module has been used to retain this model as the regression design and demonstrates how reliable and accurate the model is at intercepting web phishing.

REFERENCES

[1] Chuan Pham, Luong A. T. Nguyenz, Nguyen H. Tran, Eui-Nam Huh, Choong Seon Hong, "Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks", IEEE Transactions on Network and Service Management, 2018

[2] Aman Nagariya; https://www.kaggle.com/aman9d/phishing-dataR.

[3] M. Baykara, Z. Z. Gürelr, 6th International Symposium on Digital Forensic and Security, 1 (2018)

[4] Dharani M, Soumya Badkul, Kimaya Gharat, Amarsinh Vidhate, and Dhanashri Bhosale, "Detection of Phishing Websites Using Ensemble Machine Learning Approach", Mar 2021

[5] Zhang, Y.; Hong, J.I.; Cranor, L.F. Cantina: A content-based approach to detecting phishing web sites. In Proceedings of the 16th International Conference on World Wide Web, Banff, AB, Canada, 8–12 May 2007; pp. 639–648

[6] Jain, A.K.; Gupta, B. Comparative analysis of features-based machine learning approaches for phishing detection. In Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 16–18 March 2016; pp. 2125–2130

[7] Lin, Y.; Liu, R.; Divakaran, D.M.; Ng, J.Y.; Chan, Q.Z.; Lu, Y.; Si, Y.; Zhang, F.; Dong, J.S. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages. In Proceedings of the 30th {USENIX} Security Symposium ({USENIX} Security 21), Virtual Event, 11–13 August 2021.

[8] Jiaqi Gu; Hui Xu; An Ensemble Method for Phishing Websites Detection Based on XGBoost, 15 March 2022

[9] Musa Hajara; A.Y. Gital;Fatima Umar Zambuk; Jamilu Usman Waziri; A comparative analysis of phishing website detection using XGBOOST algorithm; March 2019

[10] Ebubekir Büber; "Phishing URL Detection with ML"; https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5, Feb 2019

[11] Nishant Nityanand Naik; "Modelling Enhanced Phishing detection using XGBoost"; https://norma.ncirl.ie/5512/1/nishantnityanandnaik.pdf, Aug 2021

[12] Ali Ahmad Aminu;Abdulrahman Abdulkarim; Amatullah Yahaya Aliyu; Muhammad Aliyu; Abdulkadir Maigari Turaki; "Detection of Phishing WebsitesUsing Random Forest and XGBOOST Algorithms"; http://www.smrpi.com/images/journals/IJPAS/20.pdf; Sep 2019

[13] Ali Aljofey, Qingshan Jiang, Abdur Rasool, Hui Chen, Wenyin Liu, Qiang Qu & Yang Wang; "An effective detection approach for phishing websites using URL and HTML features"; https://www.nature.com/articles/s41598-022-10841-5; May 2022