IJMRSET

e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 7, July 2024

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 7, July 2024 |

| DOI:10.15680/IJMRSET.2024.0707160 |

Spying On Phishers Powered by Machine Learning to Improve Online Security

Ankitha TM, Thanuja J C

Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India

Assistant Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India

ABSTRACT: Phishing sites are fake sites designed to make users input their username and password, credit card number, and other confidential details. The process of phishing detection is initiated by gathering information regarding such URL characteristics as its length, containing keywords, domain age, and usage of the HTTPS protocol. First, the data set is divided into the training data set and the test data set. Once the training is done, the model is applied to the test set as a measure of efficiency. The model accuracy in differentiating phishing from the typical web sites is evaluated using measures such as accuracy, precision, the recall, and F1 measure. Thus, the high score means that the model is highly applicable in detecting genuine phishing attempts, and in turn, shielding the users from future scams. The gradient boosting classifier is built with the training dataset; Gradient boosting learns models iteratively and the next model's learning is aimed at reducing the errors of the previous model. This method is especially useful for analysing sophisticated patterns in obtained data.

KEYWORDS: HTTPs, accuracy, precision, recall, and F1_score, gradient boosting.

I. INTRODUCTION

Phishing is a type of cybercrime whose key objective is to obtain personal information such as credit card numbers, passwords, and other personal information. It includes using fake e-mails, messages, or websites imitating trustworthy sources like banks or government departments. Cyber security is important due to the many activities that are migrating to the internet such as business, learning, and communication among others. The most common tactics used by phishers include getting the users to click on the links, give out personal information, or download infected attachments. Logistic regression, SVM, gradient boosting algorithms could be used to classify the identified URLs as phishing; analysis of their features is possible with the help of machine learning.

Phishing has devastating financial consequences and businesses in the USA are said to be losing as much as $2 billion per year, costs exceeding $5 billion globally. Lack of user awareness is normally the reason as to why most of the phishing attacks are successful. The basic approaches to blacklist are ineffective for identifying new movements and have high false positive indicators. Therefore, more researchers are employing machine learning techniques to enhance the retrieval of phishing emails. Phishing is when phishers set up fake sites that look like the real sites to catch the attention of users in order to steal from them. Pervasive consciousness and coming up with better detection mechanisms form some of the solutions that need to be undertaken to address the problem of phishing.

II.LITERATURE REVIEW

The ML based phishing techniques depend on website functionalities to gather information that can help classify websites for detecting phishing sites [1]. here 7900 malicious URLs are taken from AlexaRank portal and Phishtank dataset using LSTM technique to achieve Accuracy of 95.8 and F1_score of 95.6. we implemented four classifiers decision-tree, Naive-Bayesian, SVM, and neural network to extract features from URLs and classify URLs [2]. the dataset is taken from the UCI's machine learning Repository and these classifiers achieved over 90.39% accuracy but faced limitations with small datasets and discrete features.

we have emploted a combination of Random-Forest algorithm and ReliefF algorithm for feature selection using forward selection approach and achieved remarkable accuracy of 97.63% with10 features and 98.13% accuracy with 48 features [3]. many machine learning-based solutions have not been verified for live browsing environment, and there is lack of analysis of phishing detection [4]. we have used Kaggle and phishtank dataset and published a browser plug-in for quick identification of phishing risks, and we have used RNN_GRU model and achieved 99.18% accuracy but in this model, it is difficult to identify the short phishing URLs.

the two main objectives have been achieved are, first is identifying the best classifiers and second is best feature selection method in order to reduce the dimensionality of the dataset to improve the performance [5]. and FilteredClassifier, J-48, and RandomForest classifiers showed best results and InfoGainAttribute was the best feature selection method and achieved 92.409% accuracy.

We presented the design and evaluation of CANTINA, a novel content-based technique using TF-IDF algorithm for overcoming the page not found problems. and it catch about 90% phishing sites with 1% false positives [6]. we are focusing on detecting phishing websites URLs with domain name features. and developed PhishChecker model, achieving a 94.91% accuracy-rate [7].

we have taken a dataset of a phishing websites named Phishing.csv having 10887 rows and 31 columns. we have employed cross-validation and ranked features using ExtraTreesClassifier [8]. we found out better accuracy with XGBoost classifier and Random-Forest classifier.

PROBLEM STATEMENT

Cyberattacks are increasing more quickly than they usually do. this is a confirmation that measures have not been properly taken as to the measures that should be taken in an attempt to manage the attack. To establish the concept of cyber-attack, the following are some of the examples: Phishing website is amongst one of the most preferred ones. the widespread and frequently utilized type of attack employed by the intruder with the aim of obtaining the user’s personal or sensitive data and financial information through changing the website http addresses and other IP addresses. The sites that are forged to look like a real site are used to frustrate the user that the received website is very genuine and urging a user to click at those websites. Detecting these phishing cyber attackers, the defense of such sites is essential to prevent the attacks and safeguard the user’s information.

The main goal or objective of this project the purpose is creating a machine learning model that can precisely detect the phishing websites. The model should analyze various features of website and classify it as it is either “Phishing” or “legitimate”.

EXISTING SYSTEM

Current methods of detecting a phishing often rely on analyzing textual parameters or an appearance, which, in turn, can be easily fooled. Some methods generate unique signatures out of the readily available components found in actual domain names and the error rates for these techniques can be relatively low. Aaron Blum et al. proposed a system employ confidence-weighted classification, and content-based detection to recognize the new and previously blacklisted phishing domain to provide better security against zero-hour threats than just blacklisting. Heuristic methods are prone to high false positives, and there can be cases in which some phishing attacks may not be detected. Bogus web page developers employ the “iframe” tag to create hidden frames that seem to be part of reputable websites and compel the users to input personal information.

PROPOSED SYSTEM

Within the frame of our project, we are creating a machine learning-based algorithm for scamming website detecting where this model is not trained via the URL list of possibilities, but with some other data set This model is trained using a different data set that provides the 30 properties of the URL.

We use various URL-attributes as input parameters for our model. Out of these, 30 key features are Imagine as the primary factors in making decisions. The designed model protects consumers' information from being misused. We are designing a flask website where user can test their link, where user will give the websites to the model, the model will process the website using ML methods such as gradient boosting algorithm and differentiate the website as legitimate or spoofing website and model will display the outcome, either safe to use or not, displayed on the screen.

III.METHODOLOGY OF PROPOSED SURVEY

DATA COLLECTION

This machine learning research uses a dataset that was sourced from the well-known data science competition and dataset website Kaggle. The provided dataset consists of 11430 URLs with 30 retrieved features. The dataset is meant to act as the industry standard for machine learning-based phishing detection solutions. There are exactly 45% real URLs and 55% phishing URLs in the database. There are thirty features in each instance. Every feature has a rule attached to it. Phishing is defined as follows if the rule is satisfied. The rule is deemed lawful if it is not satisfied. There are three discrete values for the characteristics. if a rule is fully satisfied, it is represented by as 1, if partially satisfied then it is represented as 0, and not satisfied the it is represented as -1. This dataset is utilized for the study implementation since it is the most recent dataset available in the public domain.

The dataset has various factors that are crucial when determining whether a website URL is legitimate or a phishing attempt. The features are as follow

- Address Bar based features
- Abnormal based features
- HTML and JavaScript based features
- Domain based features

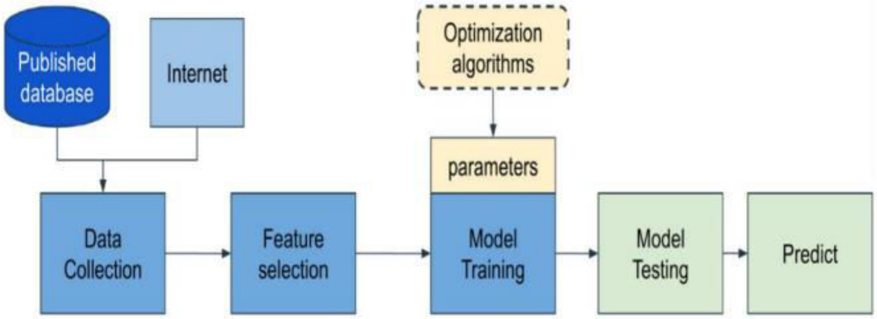


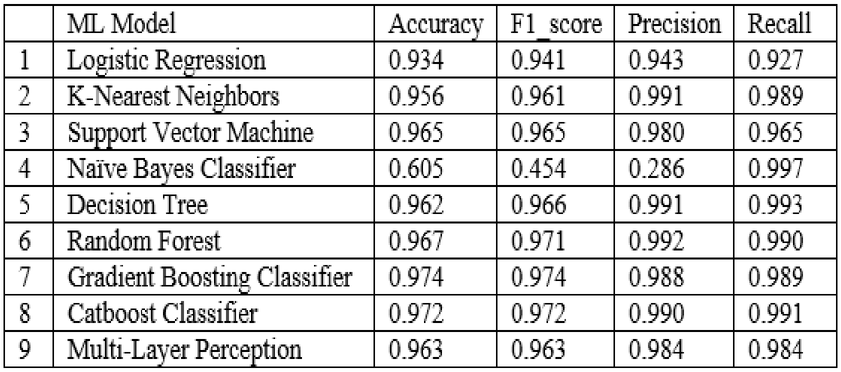
Fig 1: Block Diagram

DATA PREPROCESSING

In machine learning, there are several steps of preprocessing that are carried out in order to detect phishing websites. First of all, the set of URLs is gathered from a trustworthy source such as Kaggle, and the collected set should contain both the legitimate and phishing URLs. It involves cleaning the data by eliminating the irrelevant records and dealing with the missing records. and Extract characteristics like length of the URL, age of domain, use of special characters, subdomains, SSL certificate validity, and presence of certain terms in the URLs. then Split the data set into training data set and testing data set in to 80:20 ratio. Different classifier methods are used to differentiate between phishing and legitimate websites.

When comparing different machine learning classifiers, it is crucial to assess their performance using a variety of metrics to gain a comprehensive understanding of their effectiveness. Four key metrics for evaluation are accuracy, f1_score, Precision, and recall. The criteria for selecting the best model should be defined, taking into account the specific nature of the problem. This might involve prioritizing a particular metric, depending on whether accuracy, precision, or recall is of greater importance.

Table 1: comparing the models



	ML Model	Accuracy	F1_score	Precision	Recall
1	Logistic Regression	0.934	0.941	0.943	0.927
2	K-Nearest Neighbors	0.956	0.961	0.991	0.989
3	Support Vector Machine	0.965	0.965	0.980	0.965
4	Naïve Bayes Classifier	0.605	0.454	0.286	0.997
5	Decision Tree	0.962	0.966	0.991	0.993
6	Random Forest	0.967	0.971	0.992	0.990
7	Gradient Boosting Classifier	0.974	0.974	0.988	0.989
8	Catboost Classifier	0.972	0.972	0.990	0.991
9	Multi-Layer Perception	0.963	0.963	0.984	0.984

IV.RESULT

From the above Table1, we can analyze that Gradient Boosting Classifier Algorithm is giving highest accuracy rate, so it is considered as the best model among them as it gave high accuracy rate of 97.4%, and it is selected as the final model. The accuracy graph of the gradient boosting is shown below.

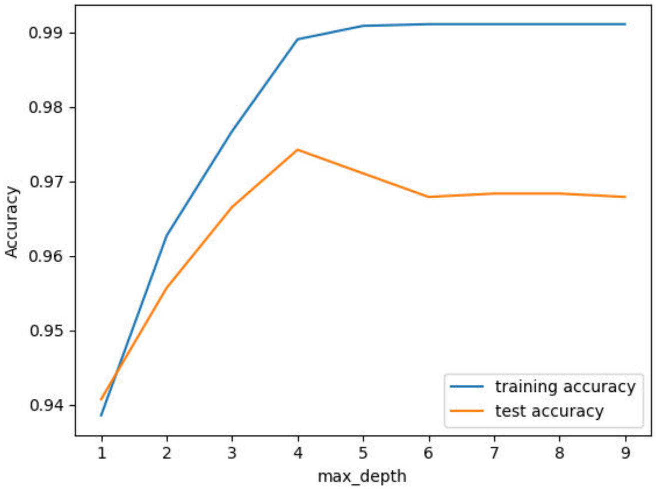


Fig 2: accuracy graph of gradient boosting algorithm

IV.CONCLUSION AND FUTURE WORK

Phishing attacks are a serious and severe threat to modern online security since they try to steal users' personal information and highly sensitive material. It is critical that we identify fake sites in order to prevent people from clicking on phishing links and losing confidential data. Blacklist-based solutions and other traditional techniques for spotting phishing websites typically don't work since they can't find new, unknown phishing sites.

Data-driven models with machine learning approaches improve the accuracy of phishing website identification. This makes it possible for us to identify phishing websites using a range of machine learning (ML) algorithms. Our research indicates that using a half breed technique will be considerably more effective in predicting and improving the accuracy of phishing website detection than using classifier methods alone, which is what existing systems utilize to categorize phishing websites. Since the accuracy of the existing systems is low, we proposed a model that uses URL features to predict phishing websites. We also generated a classifier through multiple machine learning approaches, and since the gradient boosting classifier approach yields high accuracy, the model is trained using this approach. We have obtained the intended outcomes of testing the site is spoofed or not.

Future Study, it is still possible to boost the model even further by using ensemble models to enhance the quality of the information and integration procedures to raise the score. Collective methods are one of the many techniques used in machine learning. They include integrating several basic models to create an ideal prediction model. Research efforts in the future should concentrate on combining different classifiers that have been trained on different aspects of the same training set to create a single classifier that may be able to produce a prediction that is more resilient than any other single classifier when compared to other classifiers working independently.

REFERENCES

[1]. K. Dutta, "Detecting phishing websites using machine learning technique", *PLOS ONE*, vol. 16, no. 10, pp. e0258361, Oct. 2021

[2]. AD Kulkarni, LL Brown III: “Phishing websites detection using Machine learnings” 2019•scholarworks.uttyler.edu

[3]. Anirudha J and Tanuja P: “Phishing Attack Detection using Features Selection Techniques”, Proceedings of International conference on communication and Information Processing (ICCIP), 2019

[4]. L Tang, QH Mahmoud: “A deep learning-based framework for phishing website detection” IEEE Access, 2021

[5]. R. Alazaidah, A. Al-Shaikh, M. R. AL-Mousa, H. Khafajah, G. Samara, M. Alzyoud, N. Al-Shanableh, and S. Almatarneh: “Websites phishing detection using machine learning technique Journal of …, 2024

[6]. Zhang, Y., Hong, J. I., & Cranor, L. F: CANTINA: “a content-based approach to detecting phishing websites.” In Proceeding of the 16th International Conference on World Wide Web.

[7]. R. Kiruthiga, D. Akila: “Phishing websites detection using machine” learning International Journal of Recent Technologies …, 2019

[8]. MSS Chaudhari, SN Gujar, F Jummani: “Detection of Phishing Web as an Attack: A Comprehensive Analysis of Machine Learning Algorithms on Phishing Dataset”

[9]. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[10]. SA Khan, W Khan, A Hussain-Phishing attacks and websites classification using machine learning and multiple datasets (a comparative analysis) Conference, ICIC 2020, Bari, Italy, October …, 2020

[11]. Sharma, Ushamary and Ghisingh, Seema and Ramdinmawii, Esther, “A Study on the Cyber - Crime and Cyber Criminals: A Global Problem,” International Journal of Web Technology, vol 03, pp. 172-179, June 2014.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com