2022 IEEE 7th International conference for Convergence in Technology (I2CT)
Pune, India. Apr 07-09, 2022

Phishing Website Detection Using Machine Learning

Adarsh Mandadi
Dept of Computer Science and Engineering
Vel Tech Rangarajan Dr Sagunthala R & D institute of Science and Technology
Chennai, India
adarshmandadi@gmail.com

Saikiran Boppana
Dept of Computer Science and Engineering
Vel Tech Rangarajan Dr Sagunthala R & D institute of Science and Technology
Chennai, India
saikiranboppana888@gmail.com

Vishnu Ravella
Dept of Computer Science and Engineering
Vel Tech Rangarajan Dr Sagunthala R & D institute of Science and Technology
Chennai, India
vishnuravella3@gmail.com

Prof. Dr R Kavitha
Dept of Computer Science and Engineering
Vel Tech Rangarajan Dr Sagunthala R & D institute of Science and Technology
Chennai, India
rkavitha@veltech.edu.in

2022 IEEE 7th International conference for Convergence in Technology (I2CT) | 978-1-6654-2168-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CT54291.2022.9824801

***Abstract*—Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach deals with the Random Forest, Decision Tree classifiers. The proposed approach effectively classified the Phishing and Legitimate URLs with an accuracy of 87.0% and 82.4% for Random Forest and decision tree classifiers respectively.**

Keywords— Phishing, Legitimate, Classification, Random Forest, Decision Tree

I. INTRODUCTION

In today's increasingly technological world, so much of what we do is managed online, whether it is for work or pleasure. This surge in online engagement has resulted in a tremendous spike in cyber-crime. Phishing has been the most powerful and harmful of all cyber-attacks. Phishing has been a critical security problem that has resulted in significant losses for both businesses and customers. Because of a lack of adequate identification techniques and protective methods, phishing attempts are becoming more common by the day. To ensure that internet users are protected from phishing assaults, a thorough and effective detection technique should be devised so that users information won't be compromised. Machine learning techniques may be accustomed to detecting this sort of attack. Phishing can be a cyber-crime at some point of which a goal or objectives are contacted through email, phone or textual content message through a person posing as a valid organization to trap people into presenting touchy records like personally identifiable information, banking and credit card details, and passwords. the overall method to detect phishing websites by updating blacklisted URLs, Internet Protocol to the antivirus database which is additionally referred to as the blacklist method. The major disadvantage of this approach is that it cannot detect zero-hour phishing attacks. Characteristics observed in phishing attempts are included in a heuristic-based detection system that detects zero-hour phishing attacks, but the characteristics aren't sure to always exist in such attacks and the false-positive rate in detection is incredibly high to beat this, we are using machine learning technology. Machine learning technology consists of many algorithms which need past data to form a choice or prediction on future data. With the assistance of this, algorithms will examine diverse phishing and valid URLs and their features to correctly hit upon the phishing websites along with zero-hour phishing websites.

II. RELATED WORK

[1] Attempting to obtain personal information through unlawful means has become more popular in recent years. We'll need some sort of way to notify the user in advance. This system is proposed based on the black listing method. This consists of mainly two modules Admin and User. Admin can filter those URLs and copy those in rows blacklisted URLs are classified as 1 and Legitimate URLs as 0. In the user, module user can see the blacklisted URLs as red colour and legitimate as a white colour and when the user clicks on the URLs it shows the popup not to click the URL and if the URL is legitimate then it redirects the website and also the result will be sent to the e-mail also.

[2] Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. The proposed method is an intelligent model based on the extreme learning machine. The algorithms used are Artificial Neural Network, Naive Bayes and Extreme learning Machines. The proposed work is classified into two parts firstly, the details of the data set are described which is used in the model and Secondly, rules of the features used and use of k-fold cross-validation in selecting the features with the help of this a total of 30 features are selected then data set is divided and model is trained.

[3] Phishing has cost Internet users a lot of money. It refers to exploiting a user's vulnerability, which is vulnerable to such attacks. Create a chrome extension to analyse all

"HTTP" traffic from end-user systems, and check the domain from each URL to the white-list of trustworthy domains and the black-list of illegal domains. Web scraping would be used to gather data for both lists, which would then be saved on the server. If the URL is located on the white-list, it is considered authentic; else, the entire website is analysed using several criteria. We looked at the length of the URL, the number of @ symbols in the URL, the website protocol, the number of hyphens (-) in the URL, whether the URL uses a direct IP address, and the number of dots in the URL. On the data that has been collected, classification techniques such as decision trees and random forests will be applied, and a score will be computed. If the score is higher than the threshold, the URL is flagged as phishing and blocked.

[4] Phishing is a serious security issue that includes impersonating legitimate websites to obtain personal data from online users. Firstly, it recognizes and examines attributes in phishing sites. Then, it suggests several new features and integrates them into an existing method to increase the overall detection performance. It starts by analysing and looking for unusual attributes of a phishing site. The unusual attributes that normally appear on phishing websites include abnormal symbols in the URL and some uneven HTML form and title elements. So, withdrawing features from these attributes will increase phishing detection ability.

III. PROPOSED APPROACH

The Proposed Approach mainly deals with two modules which are Feature Extraction of the URLs and Training Machine Learning Algorithms.

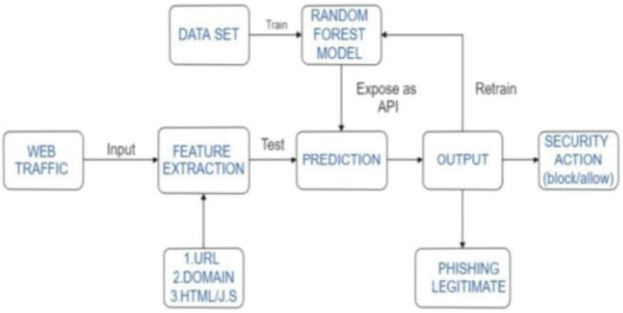

Fig. 1. Architecture

Figure 1 represents the architecture diagram. In this, the first algorithm is trained with base data set which is used as training data and the data which is taken from the web traffic acts as input for the feature extraction which is done mainly on three types of features URL based, domain-based, Html/JS-based features and this feature extracted data acts as testing data and this machine learning model is exposed to API and the prediction will be done and output is generated as phishing or legitimate. If it is phishing then we should block the website and if the output is legitimate then we should allow it.

A. Data Set Description

The Data set contains 10,000 URLs which is 5000 Phishing and 5000 Legitimate. Phishing URLs are collected from the PishiTank website. All the Phishing URLs are labelled as '1' and all the Legitimate URLs are labelled as '0'.

B. Feature Extraction

The Features of the data set has been extracted by using python programming language. This model is trained based on three kinds of features those are:

1. Domain-Based Features
    - DNS Record
    - Website Traffic
    - Age of Domain
    - End Period of Domain
2. HTML and JavaScript Based Features
    - IFrame Redirection
    - Status Bar Customization
    - Disabling Right Click
    - Website Forwarding
3. Address Bar Based Features
    - Domain
    - IP Address
    - "@" Symbol
    - Length
    - Depth
    - Redirection "//"
    - "HTTP/HTTPS" in Domain name
    - Using URL Shortening Services"Tiny URL"
    - Prefix or Suffix "-" in Domain

There are total of 17 features taken and the data set is shuffled and this data set is used to train the model.

In Fig.2 we can see the histogram Visualization of the data set and also how the data is distributed of all the 17 features.

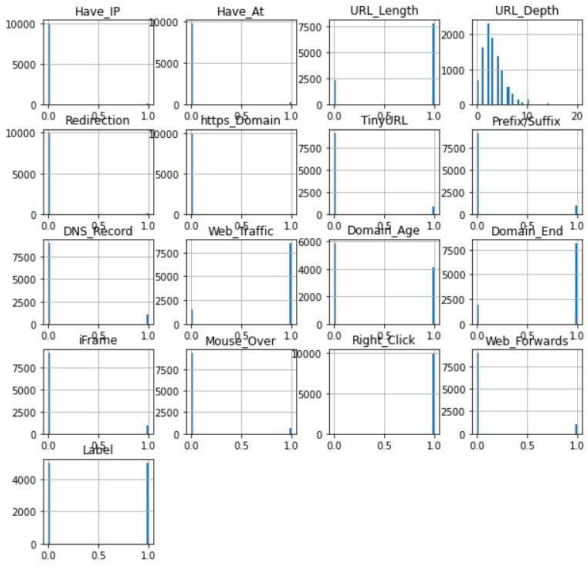

Fig. 2. Data Distribution

Fig. 3 represents a correlation heat map that shows how the features of the data set are related to each other.

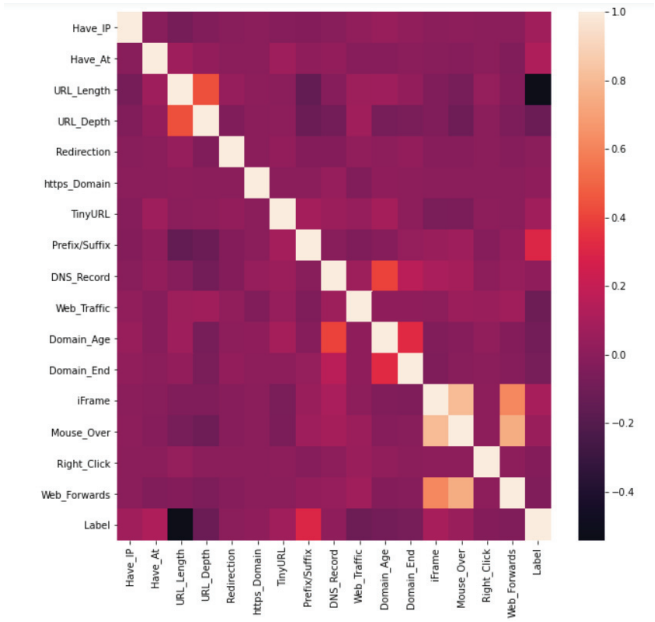

Fig. 3. Correlation Heat Map

C. Methodology

The proposed approach deals with supervised machine learning algorithms. Since it is a classification problem, we can use classifiers like Random Forest and Decision Tree.

1) Decision Tree Algorithm:

A decision tree is a Supervised machine learning technique that can be used for both regression and classification problems. A decision tree is a flow chart like tree structure, by sorting them based on the attribute value it classifies instances. At the ending nodes of a decision tree, a decision is made they are leaves of the tree. Each node in the tree represents a feature in a classification instance. The decisions or tests are made based on the characteristics of the given data collection. Every branch denotes an output of the test, each leaf node holds the label of class. Instances are classified from starting based on the value of a feature. For the classification of the data set, it generates the rule.

2) Random Forest Algorithm:

Random Forest is based on ensemble learning, which is a method for solving a complex problem and improving the performance of the model by mixing multiple classifiers. It's a classifier that uses numerous decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy. Instead of relying on a single decision tree, the random forest collects the forecasts from every tree and predicts the output data based on the majority of votes of predictions. The more trees in a random forest, the greater the accuracy and the less chance of overfitting.

IV. Implementation and Results

The feature Extracted data set is taken and it is divided into training and testing data with a ratio of 80-20 this training data is used to train the Random Forest and Decision Tree algorithm and testing data is used to test and find the accuracy of the algorithm.

	precision	recall	f1-score	support
0	0.73	0.98	0.84	945
1	0.98	0.67	0.80	1055
accuracy			0.82	2000
macro avg	0.85	0.83	0.82	2000
weighted avg	0.86	0.82	0.82	2000

FIG.4. DECISION TREE METRICS

	precision	recall	f1-score	support
0	0.80	0.93	0.86	945
1	0.93	0.79	0.85	1055
accuracy			0.86	2000
macro avg	0.86	0.86	0.86	2000
weighted avg	0.87	0.86	0.86	2000

Fig. 4. Random Forest Metrics

The above figures represent the classification metrics of both the algorithms which is used to find the accuracy of the model.

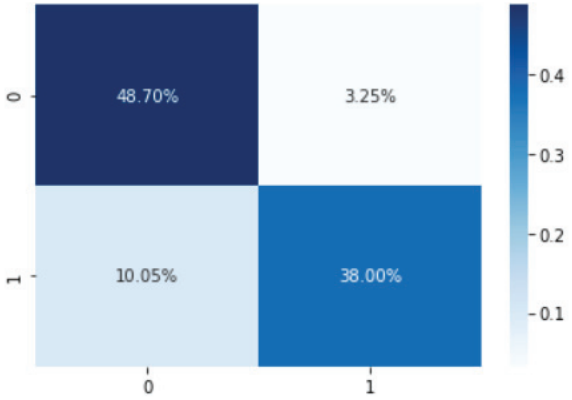

Fig. 5. Confusion Matrix

Figure 5 represents the confusion matrix for the random forest algorithm.

V. MODEL DEPLOYMENT

From the above classification report and the confusion matrix, it is clearly shown that random forest is having a high accuracy than a decision tree so, this Random Forest algorithm is stored by using pickle for deployment. The model is deployed as a webpage with the help of FlaskAPI. This webpage contains a textbox and a submit button which is developed using Hyper Text Markup Language (HTML).

Whenever we enter a URL and click on a submit button then this URL will be processed by the model and returns a value as a binary that is 0 or 1. If the returned value is '0' then the output is displayed as "Legitimate" and if the returned value is other than '0' the output is displayed as "Phishing".

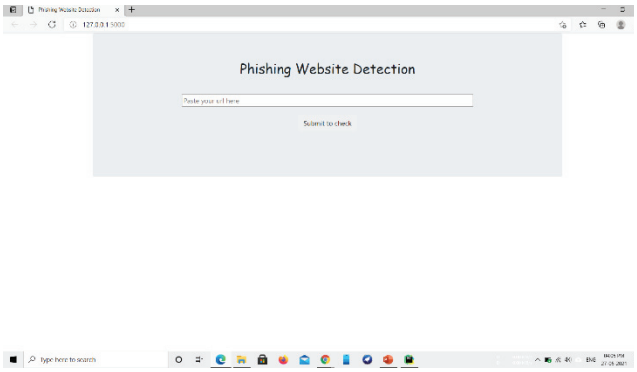

Fig. 6. Webpage

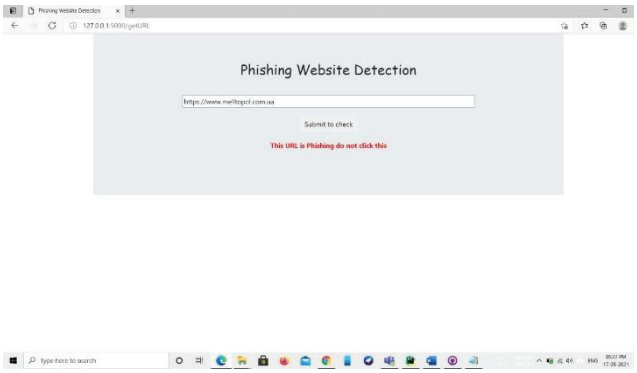

Fig. 7. Phishing

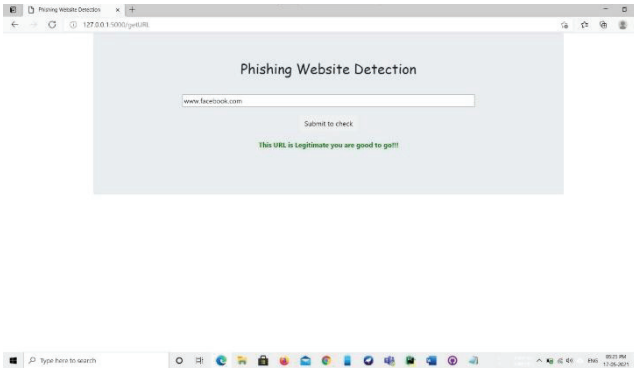

Fig. 8. Legitimate

VI. CONCLUSION

This paper helps to develop a model by using Machine Learning which is used to detect the phishing URL's and warn the user in advance. The features of the URL is extracted which is entered by the user in the respective field and this acts as input data for the Machine learning model. The model process this and gives the output as to whether it is phishing or legitimate. The algorithms that are used to build this model is Random forest and Decision Tree. After training the accuracy of Random forest is 87.0% and the accuracy of the Decision tree is 82.4%.

REFERENCES

[1] Mohammed Hazim Alkawaz, Stephanie Joanne Steven, Asif Iqbal Hajamydeen, "Detecting Phishing Website Using Machine Learning", 16th IEEE International Colloquium on Signal Processing its Applications (CSPA 2020), 28-29 Feb. 2020.

[2] Sandeep Kumar Satapathy, Shruti Mishra, Pradeep Kumar Mallick, Lavanya Badiginchala, Ravali Reddy Gudur, Siri Chandana Guttha,"Classification of Features for detecting Phishing Web Sites based on Machine Learning Techniques",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-8S2, June 2019.

[3] Vaibhav Patil,Prof. S. P. Godse,"Detection and Prevention of Phishing Websites using Machine Learning Approach", IEEE, 978-1-5386-5257-2018.

[4] Nandhini.S,Dr.V.Vasanthi ,"Extraction of Features and Classification on Phishing Websites using Web Mining Techniques",Volume 5, Issue 4 — ISSN: 2321- 9939, 2017.

[5] Sagar Patil, Yogesh Shetye, Nilesh Shendage "Detecting Phishing Websites" ,International Research Journal of Engineering and Technology,Volume: 07 Issue: 02 — Feb 2020.

[6] R. Kiruthiga, D. Akila, "Phishing Websites Detection Using Machine Learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-8,Issue-2S11, September 2019.

[7] Mahajan Mayuri Vilas ,Kakade Prachi Ghansham ,Sawant Purva Jaypralash, Pawar Shila , "Detection of Phishing Website Using Machine Learning Approach", "International Conference on Electrical,Electronics, Communication, Computer Technologies and Optimization Techniques(ICEECCOT)" 2019.

[8] Ankit Kumar Jain, B B Guptha, "A Machine Learning based approach for phishing detection using Hyper links information", Part of Springer Nature 2018.