International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

PHISHING WEBSITE DETECTION USING MACHINE LEARNING

Vinay Kumar Kureel[1], Sachin Maurya[1], Amanullah Shaikh[1], Shivam Tiwari[1], Sampada Nagmote[2]

[1]Computer Engineering Department, Student, Shree LR Tiwari College of Engineering, Mira Road, Mumbai, Maharashtra, India 40107
[2]Computer Engineering Department, Assistant Professor, Shree LR Tiwari College of Engineering, Mira Road, Mumbai, Maharashtra, India 40107

A B S T R A C T

In recent years the world is rapidly transformed into digital due to this many cyber frauds are emerging in order to harm the life of people. Phishing is one of the cybercrimes where innocent users are attack in order to steal their sensitive information like bank credentials, passwords, etc. In this research we present amachine learning based approach to detect the phishing websites. The proposed research will compare the different classification model such as Support Vector Machine, Random Forest, Decision Tree, Gradient Boosting Classifier. which has highest accuracy of differentiate between legitimate sites from phishing sites. The main of the research is to detect the phishing sites and best ml algorithm which has highest rates accuracy, false positive and negative rate.

Keywords*: Phishing, Machine Learning, Random Forest, Support Vector Machine, Decision Tree, Gradient Boosting Classifier*

1. INTRODUCTION

In today's world, Phishing has become a concern for security researchers because it is not difficult to create the fake or phishing website which looks as same as legitimate website. Experts can identify these fake websites but not all the users who are using them can identify between the fake website and legitimate website and such users are get attack by the attackers. Main aim of the hacker is to steal their sensitive information such as banks account credentials. Phishing is the social engineering attack which is commonly used against the innocent users. Through such attacks the attacker targets naive online users by tricking them into revealing confidential information with the purpose of using if fraudulently methods to prevent this attack is blacklisted URLs but it has the major drawback which helps the attacker to evade this security. Another method is to have the general awareness of phishing sites but no one remember which is legal website. In order to avoid getting phished is to detect them in their early appearance using machine learning algorithms. Using this method, algorithm will analyze the various phishing and legitimate URLs to detect which is not safe to use.

2. LITERATURE REVIEW

Numerous researches concerning phishing and phishing website detection framework are carried out previously.

Manish Jain, Kanishk Rattan, Divya Sharma, Kriti Goel, Nidhi Gupta proposed a phishing website detection framework by utilizing machine learning algorithms such as Naive Bayes Classifier, Random Forest and Support Vector Machine. Among all the algorithms

random forest gives the most precision and this framework uses Address bases, Domain based, HTML JS based features to detect the legitimacy of the website.

Suleiman Y. Yerima, Mohammmed K. Alzaylae proposed a phishing website detection framework by using Deep Learning Approach and they used CNN model to achieve high accuracy. They only used URL based feature to detect the phishing website, it has 30 attributes of urls. This has better F1 score thanany other approach.

Weiwei Zhuang, Qingshan Jiang, Tengke Xiong proposed an intelligent anti-phishing strategy model for phishing website detection. It uses URL heuristic detection module.It has categorization module. It categorizes phishing such as banking, lottery, etc. It uses hierarchical clustering algorithm for phishing categorization.

Rishikesh Mahajan, Irfan Siddavatam developed Phishing website Detection system using machine learning algorithms such as decision tree, random forest and support vector machine where random forest gave the best accuracy.

3. DATASET

URLS of the websites are collected from phishtank.com. The dataset Consists of total 11504 URLs out of which 6158 are phishing URLs and 4896 are legitimate websites. Phishing websites are labelled as '1' and legitimate websites as '-1'.

4. FEATURE EXTRACTION

To extract the features of the URLs we have implemented the python program. The features that we have extracted are mentioned below:-

1. Address Bar based features

Using the IP Address

8`5If The Domain Part has an IP Address → Phishing

Rule: IF{

Otherwise → Legitimate

Long URL to Hide the Suspicious Part

$URL\ length < 54 \rightarrow feature$ = Legitimate

Rule: IF{ $else\ if\ URL\ length \geq 54\ and \leq 75 \rightarrow feature = Suspicious$

$otherwise \rightarrow feature$ = Phishing

Using URL Shortening Services "TinyURL"

Rule: IF{ TinyURL → Phishing Otherwise → Legitimate

URL's having "@" Symbol

Rule: IF {Url Having @ Symbol → Phishing

Otherwise → Legitimate

Redirecting using "//"

ThePosition of the Last Occurrence of "//" in the URL > 7 → Phishing

Rule: IF {

Otherwise → Legitimate

Adding Prefix or Suffix Separated by (-) to the Domain

Rule: IF {Domain Name Part Includes (−) Symbol → Phishing

Otherwise → Legitimate

Sub Domain and Multi Sub Domains

Dots In Domain Part = 1 → Legitimate

Rule: IF {Dots In Domain Part = 2 → Suspicious

Otherwise → Phishing

HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

Use https and Issuer Is Trusted and Age of Certificate ≥ 1 Years → Legitimate

Rule: IF{

Using https and Issuer Is Not Trusted → Suspicious Otherwise → Phishing

Domain Registration Length

Domains Expires on ≤ 1 years → Phishing

Rule: IF{

Favicon

Otherwise → Legitimate

Favicon Loaded From External Domain → Phishing

Rule: IF{

Otherwise → Legitimate

Using Non-Standard Port

Port # is of the Preffered Status → Phishing

Rule: IF{

Otherwise → Legitimate

The Existence of "HTTPS" Token in the Domain Part of the URL

Using HTTP Token in Domain Part of The URL → Phishing

Rule: IF{

Otherwise → Legitimate

2. Abnormal Based feature

URL of Anchor

% of URL Of Anchor < 31% → $Legitimate$

Rule: IF{% of URL Of Anchor ≥ 31% And ≤ 67% → Suspicious

Otherwise → Phishing

Links in <Meta>, <Script> and <Link> tags

% of Links in " < Meta > ", " < Script > " and " < Link>" < 17% → Legitimate

Rule: IF{% of Links in < Meta > ", " < Script > " and " < Link>" ≥ 17% And ≤ 81% → Suspicious

Otherwise → Phishing

Request URL

% of Request URL < 22% → Legitimate

Rule: IF {%of Request URL ≥ 22% and 61% → Suspicious Otherwise → feature = Phishing

Server Form Handler (SFH)

SFH is "about: blank" Or Is Empty → Phishing

Rule: IF{ SFH Refers To A Different Domain → Suspicious

Otherwise → Legitimate

Submiting Information to Email

Using "mail()" or "mailto:" Function to Submit User Information → Phishing

Rule: IF{

Otherwise → Legitimate

Abnormal URL

Rule: IF {The Host Name Is Not Included In URL → Phishing

Otherwise → Legitimate

3. HTML and JavaScript based Features

Website Forwarding

Rule: IF {

Otherwise → Phishing

ofRedirect Page ≤ 1 → Legitimate

of Redirect Page ≥ 2 And < 4 → Suspicious

Status Bar Customization

onMouseOver Changes Status Bar → Phishing

Rule: IF{

It Does′t Change Status Bar → Legitimate

Disabling Right Click

Right Click Disabled → Phishing

Rule: IF{

Otherwise → Legitimate

Using Pop-up Window

Rule: IF {Popoup Window Contains Text Fields → Phishing

Otherwise → Legitimate

IFrame Redirection

Rule: IF {Using iframe → Phishing Otherwise → Legitimate

4. Domain based Features

Age of Domain

Rule: IF {Age Of Domain ≥ 6 months → Legitimate

Otherwise → Phishing

DNS Record

no DNS Record For The Domain → Phishing

Rule: IF{

Otherwise → Legitimate

Website Traffic

Website Rank < 100,000 → Legitimate

Rule: IF{ Website Rank > 100,000 → Suspicious

Otherwise → Phish

PageRank

PageRank < 0.2 → Phishing

Rule: IF{

Otherwise → Legitimate

Google Index

Webpage Indexed by Google → Legitimate

Rule: IF{

Otherwise → Phishing

Number of Links Pointing to Page

Of Link Pointing to The Webpage = 0 → Phishing

Rule: IF{

Otherwise → Legitimate

5. MACHINE LEARNING ALGORITHMS

The various Machine learning Algorithm for implementation are mentioned below:-

1. Decision Tree

Decision Tree is the machine learning algorithm which is most commonly widely used in the technology used for classification it is the graphical representation of all the possible solution to decision. it is tree-structed based classifier, where features of a dataset are represented by internal nodes and branches is used for representing the decision rules and outcome is represented by the leaf. the algorithm starts from root node for predicting the class. it compares the root values with dataset attribute and jumped to the next node based on the comparison.

2. Support Vector Machine

Support Vector Machine is the supervised learning algorithm which is used for classification. The main goal of SVM is to create the decision boundary that can segregate n-dimensional space in to classes where we can easily put the data in the right category in future , the decision boundary is known as hyperplane. SVM picks the extreme vectors which will help in creating the hyperplane. these extremes points are known as support vectors.

3. Random Forest

Random Forest is the popular supervised learning algorithm which is used for classification. It is based on decision tree algorithm. In Random Forest, the forest of decision tree is created. higher the number of trees will give the highest accuracy for detection. it creates the decision tree based on data samples and predication from the each of tree is gathered. From the gathered predictions it selects the best the solution by voting.

4. Gradient Boosting Classifier

Gradient Boosting Classifier is the group of many ml algorithms which combines many weak learning models together to form a strong predictive model. It is used to decrease the Bias Error. The base estimator for this algorithm is fixed (Decision_Stump). The cost function when GBC is used as classifier is Log loss.

6. IMPLEMENTATION AND RESULT

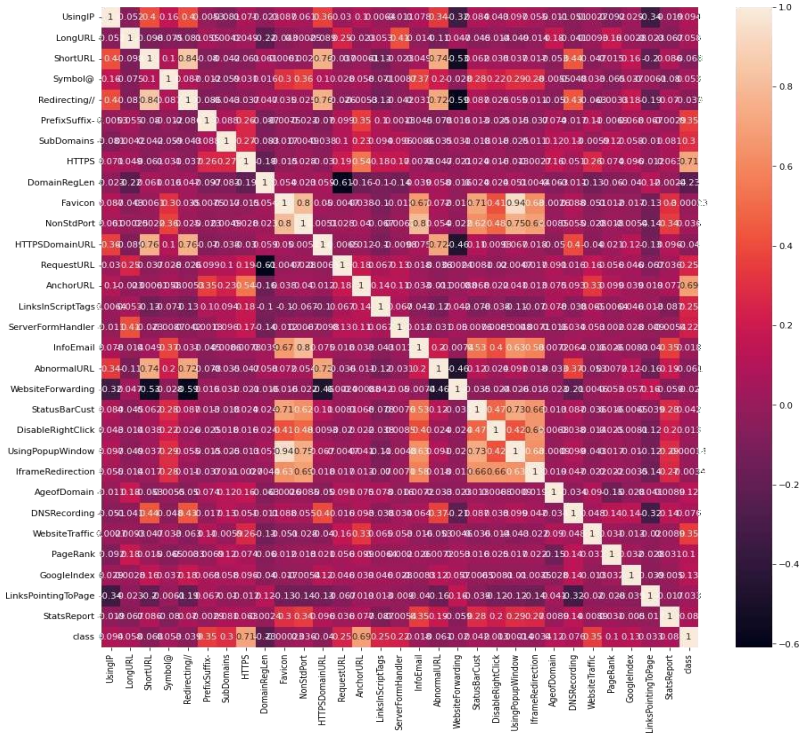

Fig 1. Correlation Heatmap

30 features are extracted from the URLs. Correlation heatmap is generated to get the statistical graph of linear relationship between attributes of dataset. Seaborn library is used draw to the heatmap. Sklearn library is used to split the dataset into train and test sets. Dataset is splitted into training set and testing set in 80:20 ratio. SciKit-Learn library has used to import machine learning algorithms. Each Classifier algorithm is trained using training and testing sets to evaluate the performance of the classifier. For each classifier accuracy score, f1 score, recall and precision is calculated. Performance of a Classifier is evaluated by accuracy score, f1, recall and precision.

Result shows that GBC gives the better detection accuracy with lowest false negative rate than RF,SVM, DF. GBC has accuracy of 0.974, F1 score is 0.977, Recall score is 0.994 and precision is 0.986. By comparing and analysing we find that Gradient Boosting Classifier model outperforms ever other model in almost all the metrics.

ML Model	Accuracy	f1_score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Random Forest	0.967	0.970	0.993	0.991
Support Vector Machine	0.964	0.968	0.980	0.965
Decision Tree	0.958	0.962	0.991	0.993

Fig 2. Performance Table

7. CONCLUSION AND FUTURE WORK

This study focuses on comparative research to detect legitimacy of the websites, dataset is taken from the Kaggle website and the model can be deployed using the web application which allows the user to easily detect the legitimacy of the website. The accuracy of Random Forest on test dataset was 0.967 with f1_score of 0.970 and the accuracy of Decision Tree on test dataset was 0.958 with f1_score of 0.962 and the accuracy of SVM on test dataset was 0.964 with f1_score of 0.968 and the accuracy of Gradient Boosting Classifier on test dataset was 0.974 with f1_score of 0.977.

In future we can develop an application for mobile phones and browser extension so it will automatically detect the legitimacy of the websites and warn the user if website is suspicious.

REFERENCE

[1] Manish Jain, Kanishk Rattan, Divya Sharma, Kriti Goel, Nidhi Gupta "Efficient Phishing Website Detection using Machine Learning Algorithm",2020, DOI:10.22214/ijraset.2021.33957

[2] S. Y. Yerima and M. K. Alzaylaee, "High Accuracy Phishing Detection Based on Convolutional Neural Networks," 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS), 2020, pp. 1-6, doi: 10.1109/ICCAIS48893.2020.9096869.

[3] W. Zhuang, Q. Jiang and T. Xiong, "An Intelligent Anti-phishing Strategy Model for Phishing Website Detection," 2012 32nd International Conference on Distributed Computing Systems Workshops, 2012, pp. 51-56, doi: 10.1109/ICDCSW.2012.66.

[4] Rishikesh Mahajan and Irfan Siddavatam. Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications 181(23):45-47, October 2018.

[5] www.phistank.com

[6] www.kaggle.com