International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

DETECTION OF PHISHING URL USING MACHINE LEARNING ALGORITHM

R.M.Suruthi[1], K.Saranya[2], V.Tamilarasi[3], Chetana Chakma[4] , Huidrom Rohit Singh[5]

[1]*Assistant Professor, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India*

[2-5]*Student Pursuing Bachelor of Engineering, Department of Computer Science and Engineering, Coimbatore Institute of Technology, Coimbatore, India*

Abstract - *Phishing is a technique used by attackers to steal personal information from the users. Phishers attain it by creating spoofed web pages or emails which when utilized by individuals, information such as their login details, credit card or debit card details and many confidential information were getting stolen by attackers without their knowledge .As the users could not differentiate between these legitimate and spoofed web pages they trust them easily and enter their personal details. Each and every hour many of the user's personal information were being stolen by the attackers. This project aims to aid this issue by detecting phishing URLs using machine learning algorithms. This model uses stochastic gradient descent and also Support vector machine algorithm to classify the URLs into legitimate and phishing. It works in real time in the back end of a website which collects all the URLs in that website , perform feature extraction on that URL and thus identify whether it is phishing URL or not and finally displaying an alert message below that specific phishing URL in that particular website.*

Key Words: **Feature extraction, Extra tree classifier, Support Vector Machine, Stochastic Gradient Descent.**

1. INTRODUCTION

Now days, Phishing becomes an important area of concern for security researchers because it is so easier to create the fake website which looks so close to legitimate website. Experts can identify the fake website easily but not all the users can identify those fake websites or fake emails and such user becomes the victim of phishing attack. The main aim of the attacker is to steal bank account credentials. Over millions of online users are becoming the victim of this attack every day. Previously security researchers used blacklist method to identify the phishing website. But due to growth of many phishing websites it is impossible to store all those information efficiently in blacklist method. So many security researchers now focused on machine learning technique to solve this problem in efficient way. Machine learning consists vast of algorithms which requires past data to make a prediction on future data. Attackers now a day using many new patterns and techniques in order to steal data .So in order to learn about these patterns huge number of past dataset needed. So machine learning technique is incorporated in order to efficiently learn about these patterns. In this project machine learning algorithm such as Stochastic Gradient descent and Support vector machine algorithms were used to analyze various phishing and legitimate URLs datasets and their features to accurately detect the phishing website.

2. RELATED WORK

Jain Mao et al [1] have proposed a system which detects the phishing using page component similarity which analyzes URL tokens to increase prediction accuracy. Phishing pages always keep its CSS style similar to their target pages. Based on the observation, detection of phishing pages is made by comparing all CSS rules of two web pages. It uses Phishing- Alert as an extension to the Google Chrome browser and demonstrated its effectiveness in detection using real-world phishing samples.

Immadisetti Naga et al [2] have proposed a system based on how a machine can able to judge the URLs based upon the given feature set. In their work the feature set was proposed which can able to classify the URLs. Their future work is to fine tuning the machine learning algorithm that will produce the better result by utilizing the given feature set. Handling of huge data sets which is a tedious process is a main disadvantage in their model.

Hemali Sampath [3] has a proposed system that implements both algorithms which is Classification and Association that optimizes the system. It uses two algorithms with WHOIS protocol and thus managed to reduce the error rate by 30% although there does not exist a model which can detect the entire phishing website but these methods still leads to the creation of efficient way to detect the phishing website.

Giovanni Armano [4] proposed a use of add on in the browser which is Real-Time Client-Side Phishing Prevention. It uses details extracted from website visited by the user to detect if it is a phish and warn the user. A warning message is displayed in the foreground while the background displays the phishing web page darkened by a black semi-transparent layer preventing interactions with the website.

3. DATASETS

Datasets plays a major role in Machine learning. Perhaps the most serious issue that everyone will experience while doing projects on machine learning was other than building a model for a particular issue it is having a decent data set that appropriately identifies with the current issues we were facing. There are two ways to gather data:

- Collecting it from a primary source our self.
- Collecting data from a secondary source, where we will be re-using data that has already been collected by different sources.

In this project, two types of datasets were used; one is already in the preprocessed state that consists of nearly 30 features extracted. Another type consists of datasets collected from Phish tank which is an open repository that consists of list of phishing URL datasets.

4. FEATURE EXTRACTION

Every URL consists of several features which will be used to discriminate a legitimate URL from a phishing URL. Features extracted from the URL includes Lexical based feature, Host based feature, Abnormal based feature and HTML based feature. To extract features from URL we have implemented python program. Below are the features that we have extracted from URL.

4.1 Presence of IP address in URL

If IP address is present in URL as an alternate for domain name then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to direct to a webpage. Use of IP address in URL may indicate that the attacker is trying to steal information.

4.2 URL length to hide the suspicious part

Most of the phishing URL length will be greater than benign URL. If the average length of the URL is found to be greater than or equal to 54 characters then the URL classified as phishing. Then the value is made to 1 else to 0.

4.3. URL shortening Services 'Tiny URL'

Tiny URL service allows attackers to hide long phishing URL by making it short. The goal is to redirect users to phishing websites. If the URL is using shortening services (like bit.ly, t.co, lnkd.in, goo.gl, cutt.us etc.) then feature is set to 1 else to 0.

4.4. Presence of "@" symbol

If @ symbol is present in URL, then the feature is set to 1 else to 0. Phishers add symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol.

4.4. Prefix or Suffix separated by (-) to domain

If domain name separated by dash (-) symbol then feature is set 1 else to 0. The dash (-) symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that the users feel that they are dealing with a legitimate webpage.

4.5. Sub domain and Multi Sub domain

If dots in domain part is equal to 1 then it is legitimate, if number of dots in domain part is 2 then it is suspicious else it is confirmed as phishing.

4.6. Presence of "https" token in URL

If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to trick the users.

4.7. Domain Registration Length

If domain expires before 1 year then it is set to 1 else to 0.Based on the fact that a phishing website lives for a short period of time, we believe that legitimate and trustworthy domains are regularly paid for several years in advance. So in our dataset, we find the longest fraudulent domains have been used for one year only.

4.8. Favicon

A favicon is a graphic image associated with a specific web page. If favicon loaded from external domain then it is phishing otherwise it is legitimate.

4.9. Presence of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL".

- If % of URL of Anchor <31% then it is legitimate.
- If % of URL of Anchor >31% and < 67% it is suspicious otherwise it is phishing.

4.10. Server Form Handler (SFH)

If SFHs contain an empty string or "about: blank" it is phishing, if SFHs refers to different domain name it is suspicious otherwise it is legitimate.

4.11. Request URL

The request URL examines whether the external objects contain within a webpage such as images, videos and sounds are loaded from another domain. If % of Request URL is <22% then it is legitimate. If % of Request URL >22% and 61% then it is suspicious otherwise it is phishing.

4.12. Redirecting using"//"

The existence of double slash "//" within the URL path means that the user will be redirected to another website. If the URL starts with "HTTP" then the "//" should appear in sixth position, if it starts with "HTTPS" then the "//" should appear in seventh position.

4.13. Presence of Port number

If the port number is equal to 80-http, 443-https, 21-ftp then it can be legitimate URL, if it does not fall under this category then it can be suspicious

5. ALGORITHM

5.1 Extra tree classifier

As the numbers of features are very large it is necessary to select important features from them in order to work efficiently with the dataset. In this model extra tree classifier algorithm is used in order to select necessary features from the extracted features. This algorithm is given with the input of all the features extracted from the feature extraction step. The Extra Trees Forest consists of decision tree which is constructed from the original training sample. Each decision tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria.

Entropy(S) = $\Sigma - pi \log 2(pi)$

c - Number of unique class labels

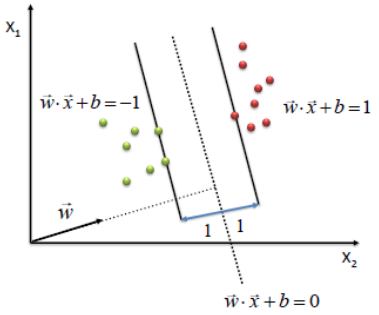
i - Output label

Formula for information gain is

Gain(S,A)=Entropy(S)-$\Sigma veValues(A)|Sv|/|S|Entropy(Sv)$

5.2 Support vector machine

Support Vector Machine (SVM) is to find hyper plane in an N-dimensional space (N-the number of features) that distinctly classifies the data points. We need to maximize the width of the margin (w) to define an optimal hyper.



Max 2/||W||

s.t.
(w.x+b)>= 1∀xofclass1
(w.x+b)<= -1 ∀xofclass2
w- Width of the margin

Fig.1.SVM graph

5.3. Stochastic gradient descent

Gradient Descent algorithm can be described as an iterative method which is used to find the values of the parameters of a function that minimizes the cost function as much as possible. The word '*stochastic* 'means a process that is linked with a random probability. In this algorithm we find out the gradient of the cost function of a single example at each iteration instead of the sum of the gradient of the cost function of all the examples. Stochastic gradient descent algorithm updates the parameters θ of the objective J (θ) as,

θ=θ−α∇θ E [J (θ)]

The new update is given by,

θ=θ−α∇θ J (θ; x (i), y (i)) Where, ((x (i), y (i)) - training sets

6. PROPOSED SYSTEM

The main objective of the project is to detect phishing URL using method like support vector machine or stochastic gradient descent by comparing which is more appropriate to the system model. If the datasets

are huge, high accuracy can be reached using SGD. First the data is preprocessed and we extracted the required features from the single string which is a URL in our case. The next step is to group the URL into phishing or legitimate. The features we extracted includes lexical, host based, abnormal based. Our model works in back-end of the webpage and thus detecting a phishing URL and giving as alert message. The proposed system relay on pycharm for implementing SVM and SGD algorithm.

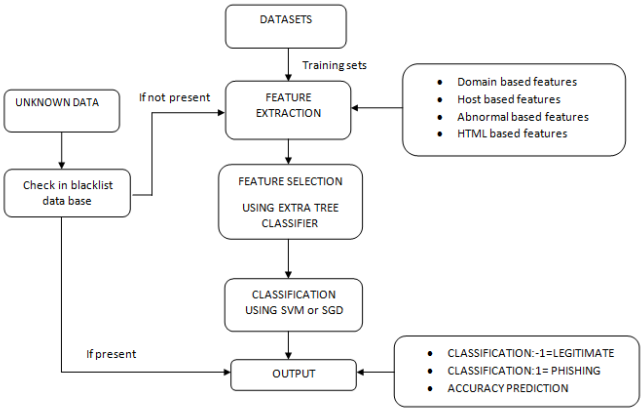


Fig.2.Flow chart

7. RESULT AND DISCUSSION

7.1. Feature Selection

For feature selection extra tree classifier algorithm is used for selecting particular features which will increase the efficiency of the model. The bar graph representation of this algorithm is given below.

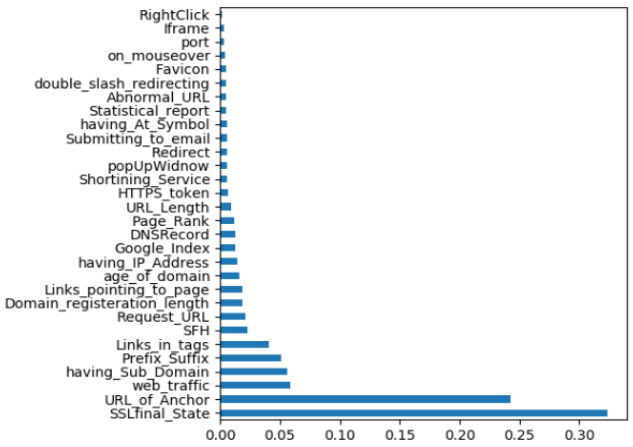


Chart-1: Feature selection bar graph

From these result 14 features were selected which contribute more for the classification of URLs and thus help in increasing the accuracy of the model.

7.2. Feature Extraction

In this project 14 features were extracted which includes having IP_Address,URL_Length,Shortening Service,Prefix_Suffix,SFH,having_At_symbol,having_Sub _Domain,webtraffic,SSLfinal_State,Domain_registration _length,port,HTTPS_token,Request_URL.
Each feature extracted will have the values [1, 0,-1] where,

- 1 indicates that the given URL is phishing
- 0 indicates that the given URL is legitimate
- -1 indicates that the given URL is suspicious

7.3. Results from Support vector machine

Support vector machine is used to classify the URL into Phishing or Legitimate URL .Features extracted is given as an input for this algorithm. The model is trained and it is tested with the given datasets.

Out of 2200 input datasets this model can able to classify **2004 URLs** correctly using this SVM algorithm. **Accuracy of 0.906377** can be reached using this algorithm.

7.4. Results from Stochastic Gradient Descent

Stochastic Gradient Descent is used to classify the URL into Phishing or Legitimate URL .Features extracted is given as an input for stochastic gradient descent algorithm. The model is trained and it is tested with the given datasets. This algorithm takes a single sample for one iteration at a time.

Out of 2200 input datasets this model can able to classify **2048 URLs** correctly using this SGD algorithm. **Accuracy of 0.926277** can be reached using this algorithm.

8. CONCLUSION

Phishing is one of the major areas of concern in security nowadays. Day to day millions of data were getting phished. To protect end users from visiting these sites, this model identifies phishing URL by analyzing their lexical and host based features and warning them about these sites. A particular challenge in this domain is that criminals are constantly making new strategies to counter our defense measures. To succeed this we need algorithms that continually adapt to new examples and features of phishing URLs. This model uses online machine learning algorithm which

provides better results compared to batch learning mechanisms when huge number of datasets were involved. Going forward we are interested in various aspects of online learning and collecting data to understand the new trends in phishing activities such as fast changing DNS servers.

REFERENCES

[1] Jain Mao, Wenqian Tian and Zhenkai Liang "Phishing-Alarm: Robust and Efficient Phishing Detection "August 23, 2017

[2] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma "Detection of Malicious URLs using Machine Learning Techniques" Volume-8 Issue-4S2 March, 2019.

[3] Hemali Sampat, Manisha Saharkar, Ajay Pandey , Hezal " Detection of Phishing Website Using Machine Learning" Volume 66, Issue 10. 15 pages year 2017

[4] Giovanni Armano, Samuel Marchal and N.Asokan "Real-Time Client-Side Phishing Prevention Add-on "Volume 66, year 2017.