Volume 6, Issue 3, July 2019 ISSN: 2348-8565 (Online)

International Journal of Modern Engineering and Research Technology

Website: http://www.ijmert.org Email: editor.ijmert@gmail.com

Detection of URL Based Phishing Attacks Using Machine Learning

J Rajaram

Professor

Department of Aeronautical Engineering,

Department of Computer Science and Engineering,

Brilliant Grammar School Educational Society's

Group of Institutions

Hyderabad, (T. S.), INDIA

Email: drjrajaram81@gmail.com

ABSTRACT

A grift attempt to get sensitive and personal information like password, username, and bank details like credit/debit card details by masking as a reliable organization in electronic communication. The phishing website will appear the same as the legitimate website and directs the user to a page to enter personal details of the user on the fake website. Through machine learning algorithms one can improve the accuracy of the prediction. The proposed method predicts the URL based phishing attacks based on features and also gives maximum accuracy. This method uses uniform resource locator (URL) features. We identified features that phishing site URLs contain. The proposed method employs those features for phishing detection. The proposed system predicts the URL based phishing attacks with maximum accuracy. We shall talk about various machine learning, the algorithm which can help in decision making and prediction. We shall use more than one algorithm to get better accuracy of prediction. Different machine learning algorithms are used in the proposed system to detect URL based phishing attacks. The hybrid algorithm approach by combining the algorithms will increase accuracy.

Keywords:— *Phishing, legitimate, URL, feature extraction, machine learning, applications, classification, approach, algorithm.*

I. INTRODUCTION

Phishing imitates the characteristics and features of emails and makes it look the same as the original one. It appears similar to that of the legitimate source. The user thinks that this email has come from a genuine company or an organisation. This makes the user to forcefully visit the phishing website through the links given in the phishing email. These phishing websites are made to mock the appearance of an original organisation website. The phishers force user to fill up the personal information by giving alarming messages or validate account messages etc so that they fill up the required information which can be used by them to misuse it. They make the situation such that the user is not left with any other option but to visit their spoofed website. [8]

Phishing is a cyber crime, the reason behind the phishers doing this crime is that it is very easy to do this, it does not cost anything and it effective. The phishing can easily access the email id of any person it is very easy to find the email id now a day and you can sending an email to anyone is freely available across the world. These

attackers put very less cost and effort to get valuable data quickly and easily. The phishing frauds leads to malware infections, loss of data, identity theft etc. The data in which these cyber criminals are interested is the crucial information of a user like the password, OTP, credit/ debit card numbers CVV, sensitive data related to business, medical data, confidential data etc. Sometimes these criminals also gather information which can give them direct access to the social media account their emails. [3]

A lot of software / approaches and algorithms are used for phishing detection. These are used at academic and commercial organisation levels. A phishing URL and the parallel page have many features which are different from the malignant URL. Let us take an example to hide the original domain name the phishing attacker can select very long and confusing name of the domain. This is very easily visible. Sometimes they use the IP address instead of using the domain name. On the other hand they can also use a shorter domain name which will not be relevant to the original legitimate website. Apart from the URL based feature of phishing detection there are many different features which can also be used for the detection of Phishing websites namely the Domain-Based Features, Page-Based Features and Content-Based Features. [16]

In the training phase, we should use the labelled data in which there are samples such as phish area and legitimate area. If we do this then classification will not be a problem for detecting the phishing domain. To do a working detection model it is very crucial to use data set in the training phase. We should use samples whose classes are known to us, which means the samples whom we label as phishing should be detected only as phishing. Similarly the samples which are labelled as legitimate will be detected as legitimate URL. The dataset to be used for machine learning must actually consist these features. There so many machine learning algorithms and each algorithm has its own working mechanism which we have already seen in the previous chapter. The existing system uses any one of the suitable machine learning algorithms for the detection of phishing URL and predicts its accuracy. The existing system has good accuracy but it is still not the best as phishing attack is a very crucial, we have to find a best solution to eliminate this. In the currently existing system, only one machine learning algorithm is used to predict the accuracy, using only one algorithm is not a good approach to improve the prediction accuracy. Each of the algorithms which explain in the earlier chapter has some disadvantages hence it is not recommended to use one machine learning algorithm to further improve the accuracy. [10]

II. METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict phishing. We shall also explain our proposed methodology to detect phishing website. In section A we shall explain various classifiers and methods which can be used to check the phishing and legitimate website. In section B we shall explain our proposed system.

Machine learning classifiers and methods to detect the phishing website

Detecting and identifying Phishing Websites is really a complex and dynamic problem. Machine learning has been widely used in many areas to create automated solutions. The phishing attacks can be carried out in many ways such as email, website, malware, sms and voice. In this work, we concentrate on detecting website phishing (URL), which is achieved by

making use of the Hybrid Algorithm Approach. Hybrid Algorithm Approach is a mixture of different classifiers working together which gives good prediction rate and improves the accuracy of the system.

Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we can not differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification. Let us discuss each of them in details.[5]

Naive Bayes Classifier:

This classifier can also be known as a Generative Learning Model. The classification here is based on Baye's Theorem, it assumes independent predictors. In simple words, this classifier will assume that the existence of specific features in a class is not related to the existence of any other feature. If there is dependency among the features of each other or on the presence of other features, all of these will be considered as an independent contribution to the probability of the output. This classification algorithm is very much useful to large datasets and is very easy to use. [14]

Random Forest:

This classification algorithm are similar to ensemble learning method of classification. The regression and other tasks, work by building a group of decision trees at training data level and during the output of the class, which could be the mode of classification or prediction regression for individual trees. This classifier accuracy for decision trees practice of over fitting the training data set.[8][14]

Support vector machine (SVM):

This is also one of the classification algorithm which is supervised and is easy to use. It can used for both classification and regression applications, but it is more famous to be used in classification applications. In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the 'n' represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes.

XGBoost:

Recently, the researches have come across an algorithm "XGBoost" and its usage is very useful for machine learning classification. It is very much fast and its performance is better as it is an execution of a boosted decision tree. This classification model is used to improve the performance of the model and also to improve the speed. [21]

Once the model is trained it is very important to evaluate the classifier which we shall use and validate its capability. Now in the above section we have seen all the advantages and disadvantages of all the available classifier. Hence we propose to use more than one classifier that is we can use a combination of two classifiers to improve the accuracy further of prediction. We shall evaluate each of the classifiers and use Naive Bayes and Random forest, by using the combination mentioned in this section we shall improve the accuracy and make it better. After applying the classification the results are generated and the URLs are classified into phishing and legitimate URLs. The Phishing URLs are blacklisted in the database and the

legitimate are white list in the database. [12]

III. Proposed System

The dataset of phishing and legitimate URL's is given to the system which is then pre-processed so that the data is in the useable format for analysis. The features have around 30 characteristics of phishing websites which is used to differentiate it from legitimate ones. Each category has its own characteristics of phishing attributes and values are defined. The specified characteristics are extracted for each URL and valid ranges of inputs are identified. These values are then assigned to each phishing website risk. For each input the values range from 0 to 10, while for output range is from 0 to 100. The phishing attributes values are represented with binary no 0 and 1 which indicates the attribute is present or not.

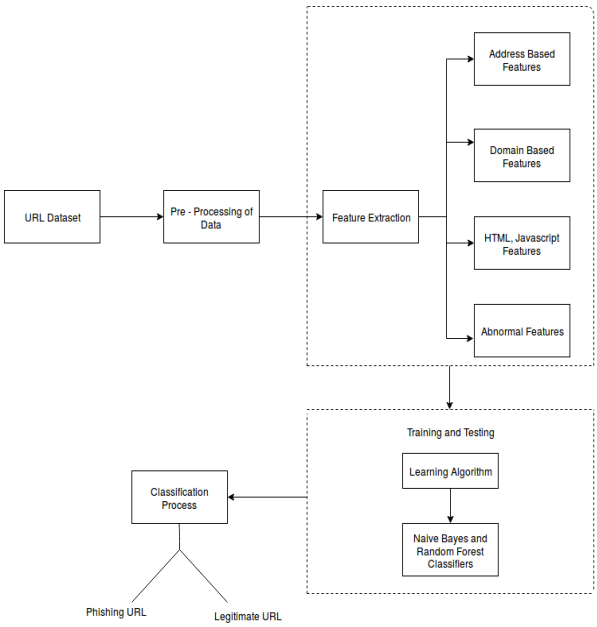

Figure 1: Proposed System block diagram

After this the data is trained we shall apply a relevant machine learning algorithm to the dataset. The machine learning algorithms are already explained in previous section. After this we use a hybrid classification in which we combine two of the classifier namely Naive Bayes and Random forest to predict the accuracy of the detection of the phishing URL, hence we get our desired result. This is also called a hybrid approach to test the data, in this method we propose to use the combination of two classifiers, as mentioned above. We shall then test the data and evaluate the prediction accuracy which shall be more than the existing system. We shall now see the different classifiers and discuss the hybrid combination used for our proposed system.

In the training phase, we should use the labelled data in which there are samples such as phish area and legitimate area. If we do this then classification will not be a problem for detecting the phishing domain. To do a working detection model it is very crucial to use data set in the training phase. We should use samples whose classes are known to us, which means the samples whom we label as phishing should be detected only as phishing. Similarly the samples which are labelled as legitimate will be detected as legitimate URL. The dataset to be used for machine learning must actually consist these features. There so many machine learning algorithms and each algorithm has its own working mechanism which we have already seen in the previous chapter. The existing system uses any one of the suitable machine learning algorithms for the detection of phishing URL and predicts its accuracy. Each of the algorithms which explain in the earlier section has some disadvantages hence it is not recommended to use one machine learning algorithm to detect the phishing website [10]

IV. System Overview

System design is used for understanding the construction of system. We have explained the flow of our system and the software used in the system in this section.

System Flow

The Figure 2 explains the flow chart of the system design, we shall explain each of the components of the flow chart in each section below. To get structured data we do feature generation of the data at the pre-processing stage. We have used techniques like XG Boost, Naive Bayes, SVM, Meta classifiers and stacking classifier to detect the phishing and legitimate websites.

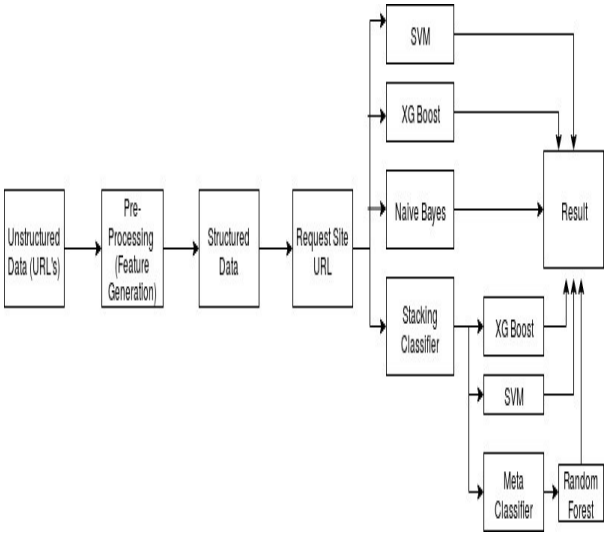

Figure 2. Flow chart of the system

Data set:

The data of urls is obtained from Phish tank website, where Phish tank is an anti-phishing site. It contains 2905 urls which is in unstructured form. Our main objective is to detect whether the url is phishing or legitimate based on the features extracted.

Phish-id	url
4912175	http://www.rollencenter.eu//wells/wells3/index.htm
4912845	https://dice-profit.top/EserviceMain/irs/ir/index.html
4912843	https://glprinters.com/EserviceMain/irs/ir/index.html
4912460	https://3mtoyou.000webhostapp.com/
4912136	https://www.accuweather.com
4912137	https://www.ted.com
4912140	https://www.monster.com

Figure 3.Unstructured Data

In Pre-processing we have done feature extraction where The URLs are transmitted to the feature extractor, which extracts feature values through the predefined URL-based features. The features have assigned binary values 0 and 1 which indicates that feature is present or not as shown in figure below. The extracted feature values are stored as input and passed to the classifiers.

Phish-id	Length of url	http -has	Suspicious-char	Prefix-suffix	dots	slash	Phis-term	Sub-domain	Ip-address
4912175	1	1	1	1	1	1	0	1	0
4912845	1	1	1	1	1	1	1	0	0
4912843	1	1	1	0	0	0	0	0	0
4912460	1	1	1	1	1	1	1	0	0
4912136	0	1	1	0	1	1	0	1	0
4912137	0	1	1	1	1	1	0	0	0
4912140	1	1	1	0	0	1	0	1	0

Figure 4. Structured Data

A structured dataset is given to the classifiers. We use four methods classification namely: XG Boost, SVM, Naive Bayes and stacking classifier for detection of url as phishing or legitimate.
Table 1: URL Features

Sr. No	Feature name	Description
1	IP address	Whether domain is in the form of an IP address
2	Length of URL	Length of URL
3	Suspicious character	Whether URL has _@‘, _//‘
4	Prefix and suffix	Whether URL has _-‘
5	Length of subdomain	Length of subdomain
6	Number of ‘/’	Number of ‘/’ in URL
7	HTTPS protocol	Whether URL use https.
8	Phishing words in URL	Whether url has phishing terms
9	Number of ‘.’	Number of dots ‘.’ in url

Now the classifier will find whether a requested site is a phishing site. When there is a page request, the URL of the requested site is radiated to the feature extractor. It extracts the feature values through the predefined URL-based features. These feature values are act as a input for the classifier. After this we will come to know if the site is phishing or not.

URL Features:

Refering Table 1. above, Features 1 to 4 are associated with suspicious URL patterns and characters. Characters such as _@' and _//' rarely appear in a URL. Feature 5 is known for recognising newly created phishing sites with the proposed methodology. Currently, to prevent a user from identifying that a site is not legitimate, phishing sites typically hide the primary domain; the URLs of these phishing sites have unusually long subdomains.

Feature 8 is another new feature that reflects current phishing trends. This feature includes seven words that are predefined as phishing terms. The seven phishing terms are secure, websrc, ebaysapi, signin, banking, confirm, login. Thus, through experiments, we identified seven new phishing terms and we employ them in our phishing detection technique. We have already discussed the different classifiers in the above sections.

V. IMPLEMENTATION

This section provides knowledge about the implementation environment and throws light on the actual steps for the implementation of dataset to get better accuracy to predict phishing by using different classifiers combination.

Hardware requirements

The following hardware was used for the implementation of the system:

- 4 GB RAM
- 10GB HDD
- Intel 1.66 GHz Processor Pentium 4

Software requirements

The following software was used for the implementationof the system:

- Windows 7
- Python 3.6.0
- Visual Studio Code

Implementation steps

In this section we shall discuss about the actual steps which were implemented while doing the m experiment. We shall explain the stepwise procedure used to analyse the data and to predict the phishing. The system consists of the following main steps, We have used unstructured data which consists only urls. There are 2905 urls obtained from Phishtank website which consists of both phishing and legitimate url where most of urls obtained are phishing.

We have collected unstructured data of urls from Phishtank website.

In preprocessing, feature generation is done where nine features are generated from unstructured data. These features are length of url, url has http, url has suspicious character, prefix/suffix, number. of dots, number of slash, url has phishing term, length of subdomain, url contains ip address.

After this a structured dataset is created in which each feature contains binary value(0, 1) which is then passed to the different classifiers.

Next we train the four different classifiers and compare their performance on the basis of accuracy four classifiers used are XG Boost, SVM, Naive Bayes and Stacking, where stacking uses XG Boost and SVM as

its base classifier and Random Forest as its meta classifier.

Then classifier detects the given url based on the training data that is if the site is phishing it shows a pop-up and if legitimate it opens that page in browser.

We compare the accuracy of different classifiers and found XG Boost and Stacking are the best classifiers which gives the maximum accuracy.

Below are the screen shots for the implementation process.

We have the test screen:

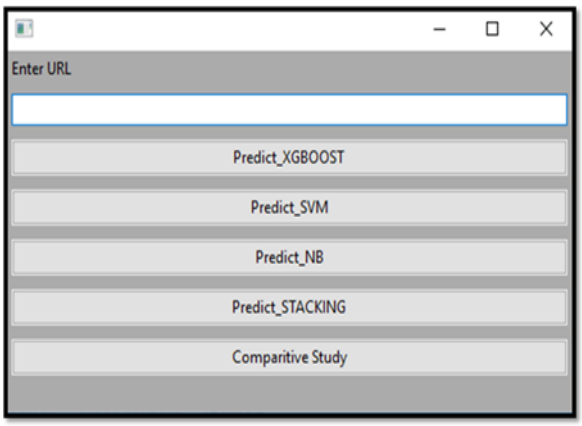

Figure 5: Testing Screen

We will now test the legitimate website by entering the URL on the test screen

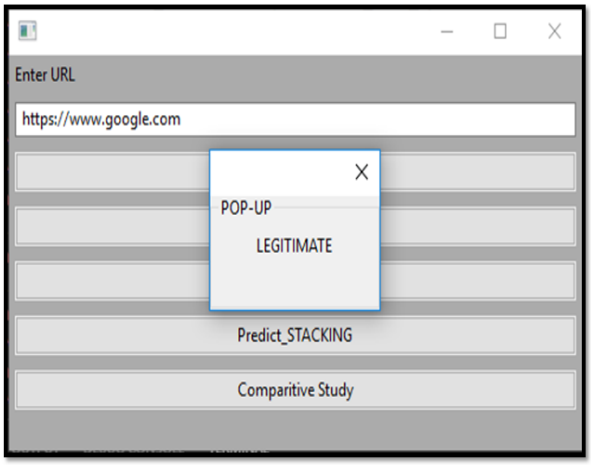

Figure 6. Testing the legitimate site

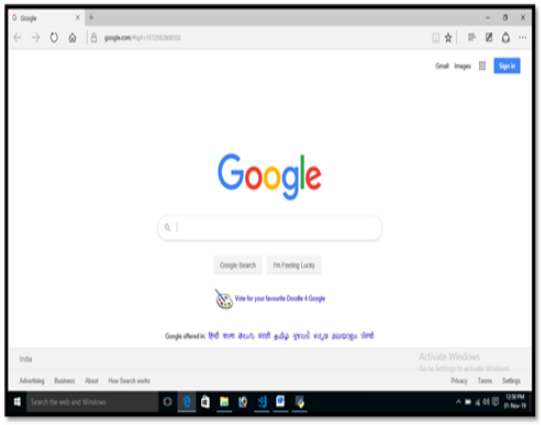

Figure 7. The legitimate site opens up

We will now test the phishing website.

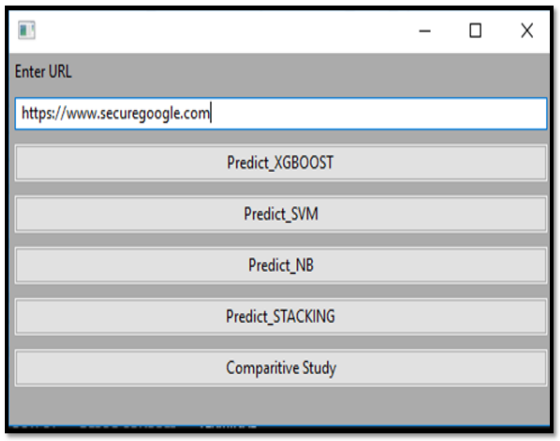

Figure 8. Testing for the phishing site

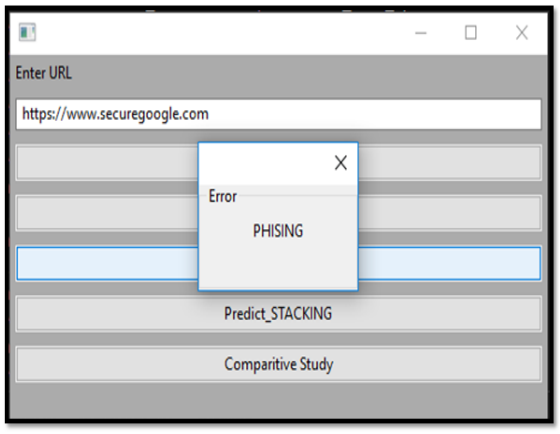

Figure 9. The phishing site

We do this testing by using 4 different techniques of classification. We shall show the screenshot of stacking classifier of all stages and ROC curve for the other 3 methods.

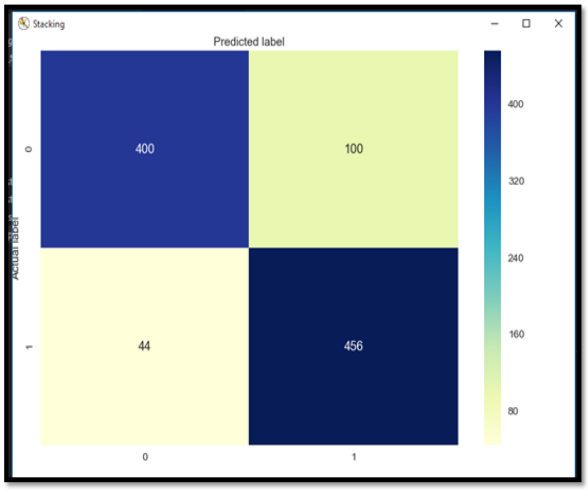

Figure 10. Confusion matrix of stacking classifier

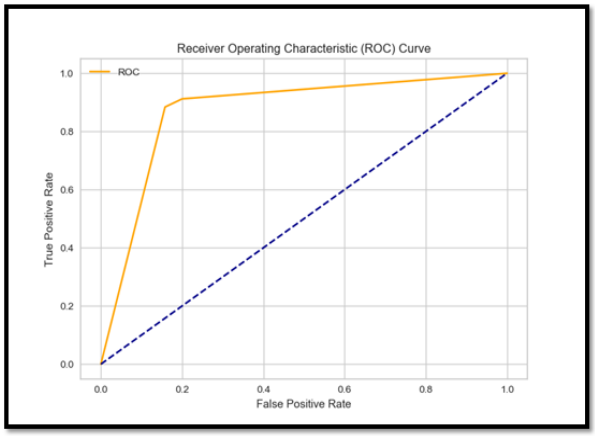

Figure 11. ROC curve of stacking classifier

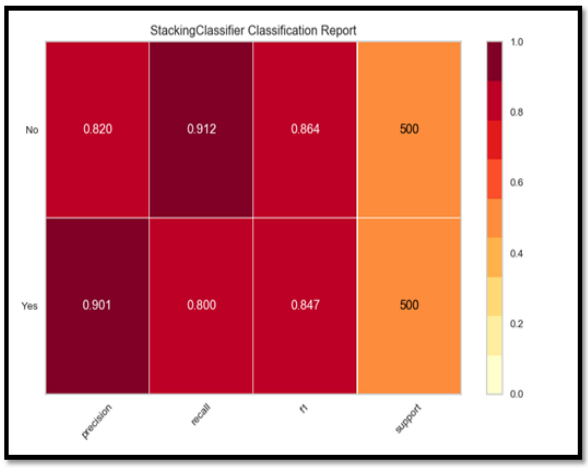

Figure 12. Classification of the stacking classifier

We have used similar steps and got ROC curves for XGBoost, SVM and Naive Bayes classifier. See the below screenshots for the same.

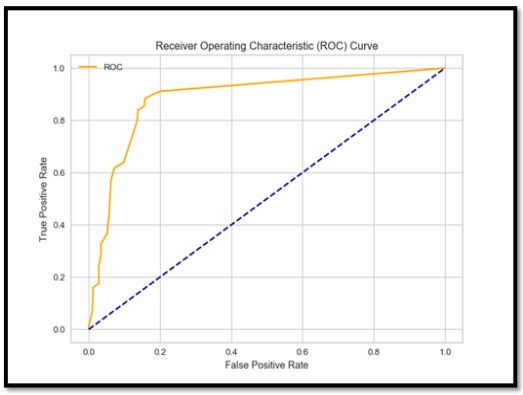

Figure 13. ROC curve for SVM

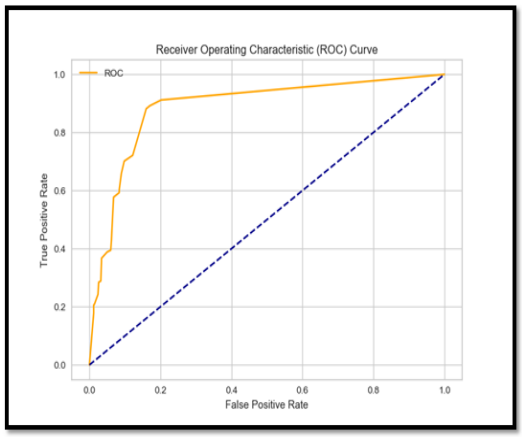

Figure 14. ROC curve for Naive Bayes

VI. Observation And Result

Observation

As discussed in the earlier sections, we have used four different classifiers to predict and detect if the website is phishing or legitimate. Comparisons of these classifiers have been shown below in the accuracy table.

Table 2: Observation Table

Classifiers	Precision	Recall	F1	AUC	Accuracy(%)
XG Boost	0.90	0.80	0.85	0.90	85.5
SVM	0.88	0.84	0.86	0.89	86.3
Naive Bayes	0.75	0.90	0.82	0.89	80.2
Stacking (XGBoost,SVM,Random Forest)	0.90	0.80	0.85	0.87	85.6

Result

We have got the desired results of testing the site is phishing or not by using four different classifiers. Refer the graph below for the exact results. Refer the graphs in Figure 15 and Figure 16 for the results. In the graph, shown in Figure 15 shows the AUC, precision, recall and the F1 score obtained by using different classifiers. The graph shown in Figure 16. explains about the accuracy obtained by using different classifiers in the histogram graphical representation.

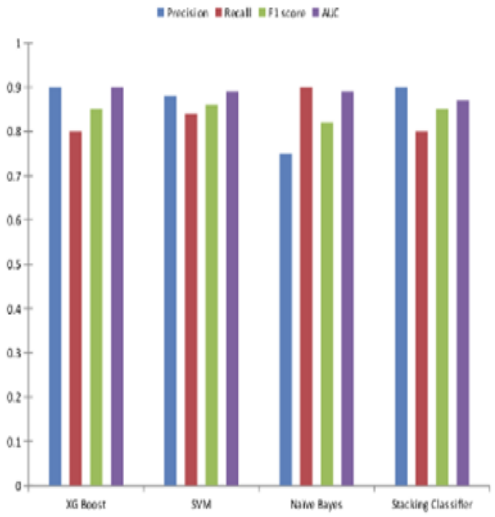

Figure 15. Graph of AUC, Precision, Recall and F1score

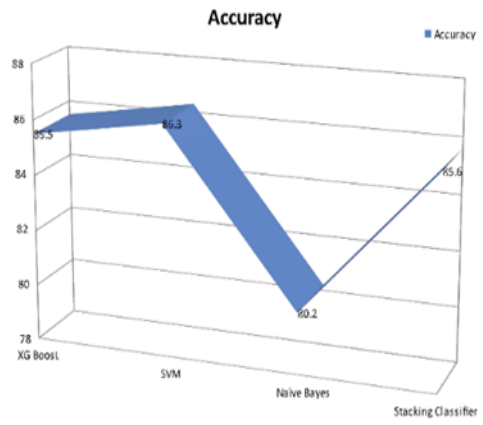

Figure 16. Results

VII. Conclusion And Future Scope

Conclusion

It is found that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing beside, but after our survey that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also we generated classifiers through several machine learning algorithms.We have found that our system provides us with 85.5 % of accuracy for XG Boost Classifier, 86.3% accuracy for SVM Classifier, 80.2 % accuracy for Naïve Bayes Classifier and finally 85.6 percentage of accuracy when using Stacking Classifier. Hence we found that the best among all the above classifiers is SVM and Stacking Classifier which shows maximum accuracy. The proposed technique is much more secured as it detects new and previous phishing sites.

Future scope

In future if we get structured dataset of phishing we can perform phishing detection much more faster than any other technique. In future we can use a combination of any other two or more classifier to get maximum accuracy. We also plan to explore various phishing techniques that uses Lexical features, Network based features, Content based features, Webpage based features and HTML and JavaScript

features of web pages which can improve the performance of the system. In particular, we extract features from URLs and pass it through the various classifiers.

REFERENCES:

[1] Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.

[2] Rao, R. S., & Pais, A. R. (2019). Jail-Phish: An improved search engine based phishing detection system. Computers & Security, 83, 246-267.

[3] Ding, Y., Luktarhan, N., Li, K., & Slamu, W. (2019). A keyword-based combination approach for detecting phishing webpages. computers & security, 84, 256-275.

[4] Marchal, S., Saari, K., Singh, N., & Asokan, N. (2016, June). Know your phish: Novel techniques for detecting phishing sites and their targets. In 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) (pp. 323-333). IEEE.

[5] Shekokar, N. M., Shah, C., Mahajan, M., & Rachh, S. (2015). An ideal approach for detection and prevention of phishing attacks. Procedia Computer Science, 49, 82-91.

[6] Rathod, J., & Nandy, D. Anti-Phishing Technique to Detect URL Obfuscation.

[7] Hodžić, A., Kevrić, J., & Karadag, A. (2016). Comparison of machine learning techniques in phishing website classification. In International Conference on Economic and Social Studies (ICESoS'16) (pp. 249-256).

[8] Pujara, P., & Chaudhari, M. B. (2018). Phishing Website Detection using Machine Learning: A Review.

[9] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine leaning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.

[10] Lakshmi, V. S., & Vijaya, M. S. (2012). Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering, 30, 798-805.

[11] Jain, A. K., & Gupta, B. B. (2018). PHISH-SAFE: URL features-based phishing detection system using machine learning. In Cyber Security (pp. 467-474). Springer, Singapore.

[12] Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications, 42(3), 1166-1177.

[13] Mao, J., Bian, J., Tian, W., Zhu, S., Wei, T., Li, A., & Liang, Z. (2019). Phishing page detection via learning classifiers from page layout feature. EURASIP Journal on Wireless Communications and Networking, 2019(1), 43.

[14] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012, December). An

assessment of features related to phishing websites using an automated technique. In 2012 International Conference for Internet Technology and Secured Transactions (pp. 492-497). IEEE.

[15] https://www.researchgate.net/publication/226420039-Detection-of-Phishing-Attacks-A-Machine-Learning-Approac

[16] https://www.proofpoint.com/us/threat-reference/phishin

[17] https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e

[18] https://en.wikipedia.org/wiki/Phishin

[19] https://www.techrepublic.com/article/how-to-tackle-phishing-with-machine-learning

[20] https://www.irjet.net/archives/V5/i3/IRJET-V5I3580.pd

[21] https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/beginners-tutorial-on-xgboost-parameter-tuning-r/tutorial

[22] https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-pytho

[23] He, M., Horng, S. J., Fan, P., Khan, M. K., Run, R. S., Lai, J. L., & Sutanto, A. (2011). An efficient phishing webpage detector. Expert systems with applications, 38(10), 12018-12027.

[24] Le, A., Markopoulou, A., & Faloutsos, M. (2011, April). Phishdef: Url names say it all. In 2011 Proceedings IEEE INFOCOM (pp. 191-195). IEEE.

[25] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357.

[26] Tewari, A., Jain, A. K., & Gupta, B. B. (2016). Recent survey of various defense mechanisms against phishing attacks. Journal of Information Privacy and Security, 12(1), 3-13.

[27] Jain, A. K., & Gupta, B. B. (2016, March). Comparative analysis of features based machine learning approaches for phishing detection. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 2125-2130). IEEE.

[28] Yuan, H., Chen, X., Li, Y., Yang, Z., & Liu, W. (2018, August). Detecting Phishing Websites and Targets Based on URLs and Webpage Links. In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 3669-3674). IEEE.

[29] Nguyen, L. A. T., To, B. L., Nguyen, H. K., & Nguyen, M. H. (2013, October). Detecting phishing web sites: A heuristic URL-based approach. In 2013 International Conference on Advanced Technologies for Communications (ATC 2013) (pp. 597-602). IEEE.

* * * * *