Journal of Applied Information Science
8 (1&2), 2020, 07-13
http://www.publishingindia.com/jais

Analysis of Phishing Detection Using Logistic Regression and Random Forest

Gokul S.[1] and Nizar Banu P. K.[2*]
[1]Department of Computer Science, CHRIST (Deemed to be University), Bangalore, Karnataka, India.
Email: gokul.s@cs.christuniversity.in
[2]Department of Computer Science, CHRIST (Deemed to be University), Bangalore, Karnataka, India.
Email: nizar.banu@christuniversity.in
*Corresponding Author

Abstract: In the present era of technology, cybercriminals are getting smarter day by day. From the past few years, the cybercrimes have increased to an extent that most of the big companies are finding it difficult to prevent cybercrimes. One such cyber-attack is phishing where the victims are lured in entering their sensitive information like usernames, passwords, bank details, etc. It's very easy for an attacker to get sensitive information through phishing. The attacker should know some information about the victim's profile so that the victims can be easily tricked. A phished URL that the victims receive is very tough to differentiate as looks similar to the original URL. In this paper, we have made use of the information in the URL to determine if the URL is phished or not. So, it is not necessary for the user to enter the website and expose themselves to the malicious code. We have also discussed the metadata that is present in the URL. In this paper, we also make use of metadata to classify a URL. Random forest and logistic regression are the two algorithms used to classify the URL present in the dataset as phished or not phished. After using the classification algorithm on the given datasets, we found that the random forest algorithm has better accuracy in classifying if a URL is legit.

Keywords: Classification, Cyber attack, Logistic regression, Phishing, Random forest, URL phishing.

I. Introduction

In the present world of dominant technology where people are bound to use different applications also increases the cyber-attacks on these technologies. One of the oldest and most famous cyber-attack is Phishing. Phishing is a type of cyber-attack which uses email, a phone call or text message to trick users into entering sensitive credentials like username, passwords, bank details etc [1]. Phishing is also called as a social engineering attack. In this attack, the attacker sends an email to the victim and when the victim clicks on the link a new web page will be opened where the victims are tricked into entering their sensitive information and this information can be used for illegal purposes like identity theft, selling their information in the black market or use for other financial benefits. Most of the common websites that the hackers use for phishing are social media websites or banking websites and after getting the sensitive credentials from the victim the hacker can use it for his own personal benefit [2].

There are many types of phishing attacks [3] like:

- Spear Phishing - It is a type of email spoofing attack where the attacker targets a specific individual or else an organization to penetrate into their system to get their sensitive information.
- Whaling - It is an email spoofing attack that targets higher executives like managers or CEO of the company. So, the email is crafted in such a way that higher officials can easily be targeted.
- Clone Phishing - It is an email spoofing attack in which a legitimate and previously delivered email content and the recipient address containing an attachment is taken and a similar email is cloned except this email contains a malicious attachment.
- Vishing - It is a type of phishing attack that happens through voice calls. Scammers try to convince victims in giving out their sensitive information over the phone.
- Smishing - This is another type of phishing that happens through messaging. In this type of attack, the users are lured into clicking links that come through messages.

In order to perform a phishing attack, we have to follow a few steps, they are given below. The first step in phishing involves getting to know about the victim or the organization. Create a fake page where the victim has to enter credentials. This fake website should look identical to the legitimate website. The link of the website should also look similar to the real link. The last phase consists of sending the link to the victim via email. The email must be crafted in such a way that the victim is easily tricked. After clicking the link in an email, the victim would be redirected to a fake website where they have to enter their sensitive credentials. Fig. 1 shows the steps in a phishing attack.

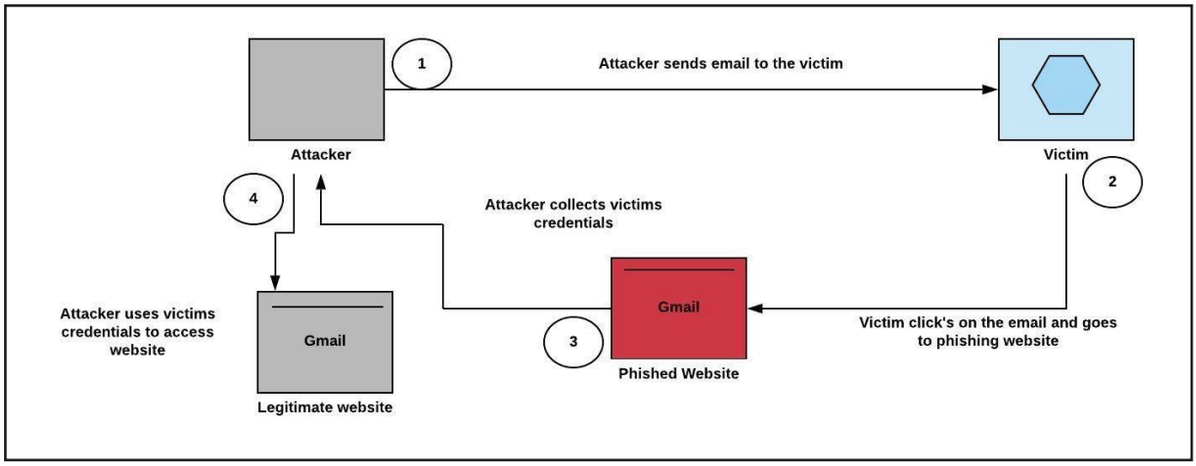

Fig. 1: Process of Phishing

The Present Statistics of Phishing

Verizon's data breach investigation report shows that in 2018, more than one-third of the data breach happened because of phishing. A cybersecurity platform Avanan, reports that Microsoft and Amazon are the most popular brands that are used by the attackers. In 2011 the United States computer emergency readiness team reports the incident that happened in federal, state and local government agencies shows that more than 51% of the attacks happened through phishing [4].

Fig. 2 shows the statistics of different ways of phishing attacks in which victims are lured into clicking onto the links.

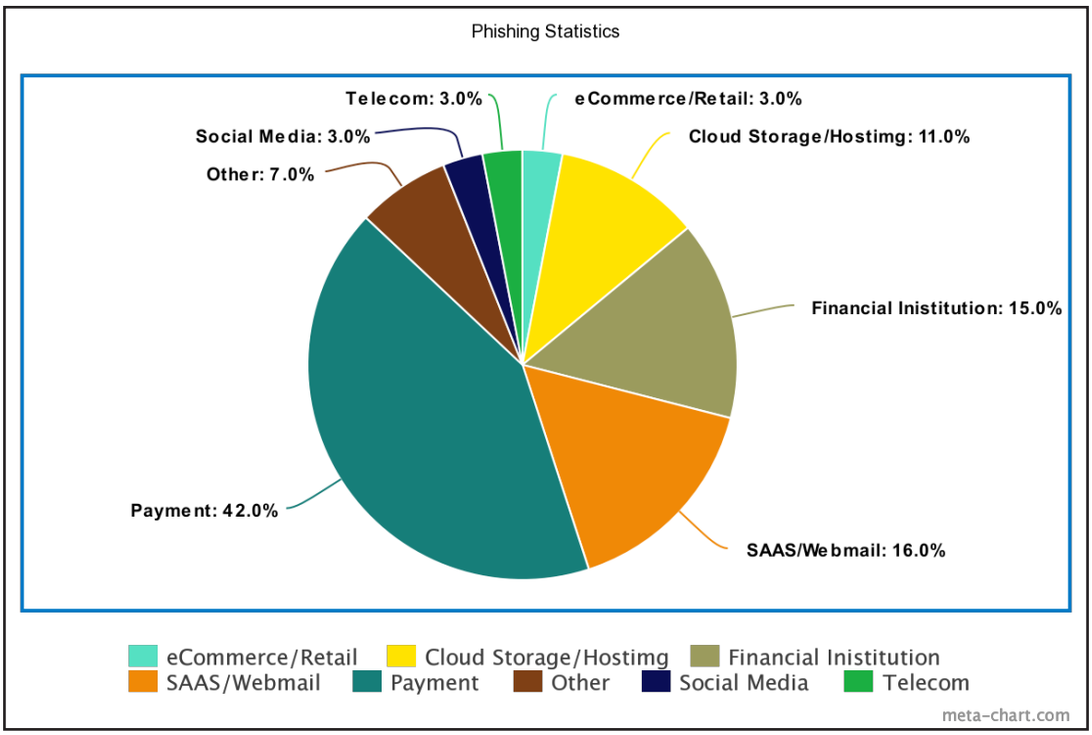

Fig. 2: Statistics of Phishing

We can classify if a website is phished or not phished by seeing the contents of the website and also by using the page rank algorithm. We can also classify if a website is phished by taking the metadata of the URL [5].

Fig. 3 shows a real web page of a popular mailing service Gmail. Fig. 4 shows a phished mail. Gmail shows a warning where a victim receives a phished mail. But mostly the email is crafted in such a way that even the Gmail filters are unable

to distinguish if the email is phished or not. Fig. 5 shows a phished page that opens after clicking the link that is present in the phished mail.

In this paper, we have used metadata to classify if a URL is phished or not. A URL contains many attributes like URL length, having a symbol, double slash redirecting, having IP address etc. This information can be used to classify if a URL is phished or not without even entering the website. After getting all the required information from a number of URLs then we could use a classifier to test if the URL is phished or not. We have chosen the Logistic Regression and Random Forest Algorithm to classify if a URL is phished or not.

The remaining paper is discussed as follows. Section II discusses related work that is already there in this field. Section III presents the proposed methodology. This section deals with the datasets we have used and also describes the dataset. Further, we also discuss the machine learning algorithm we have used to classify the URL. Section IV discusses the accuracy of detecting the phished URL that we get after inserting the test dataset into the model.

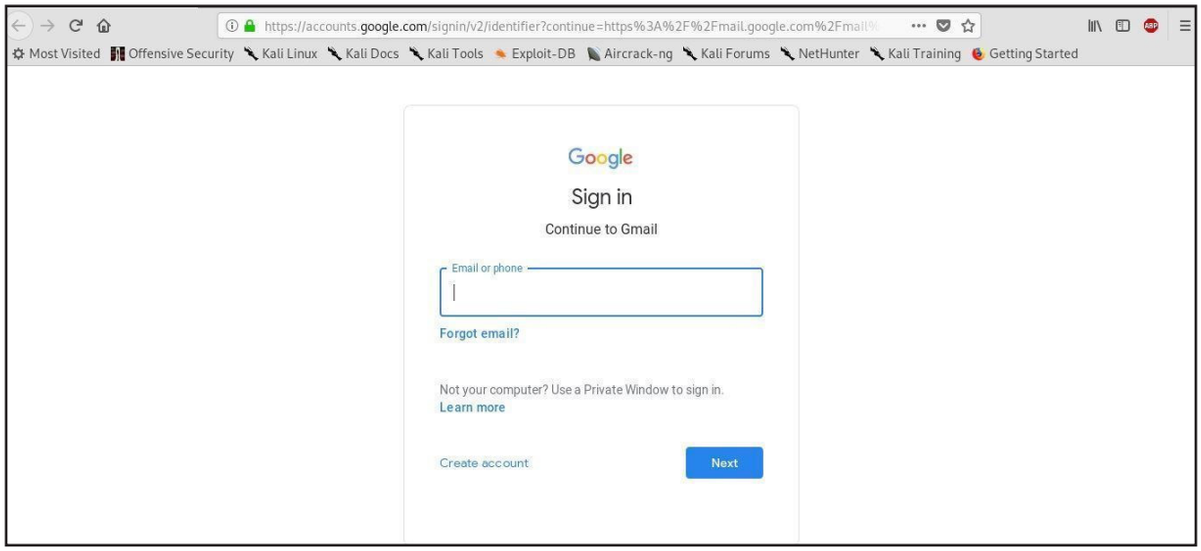

Fig. 3: Original Gmail Login Page

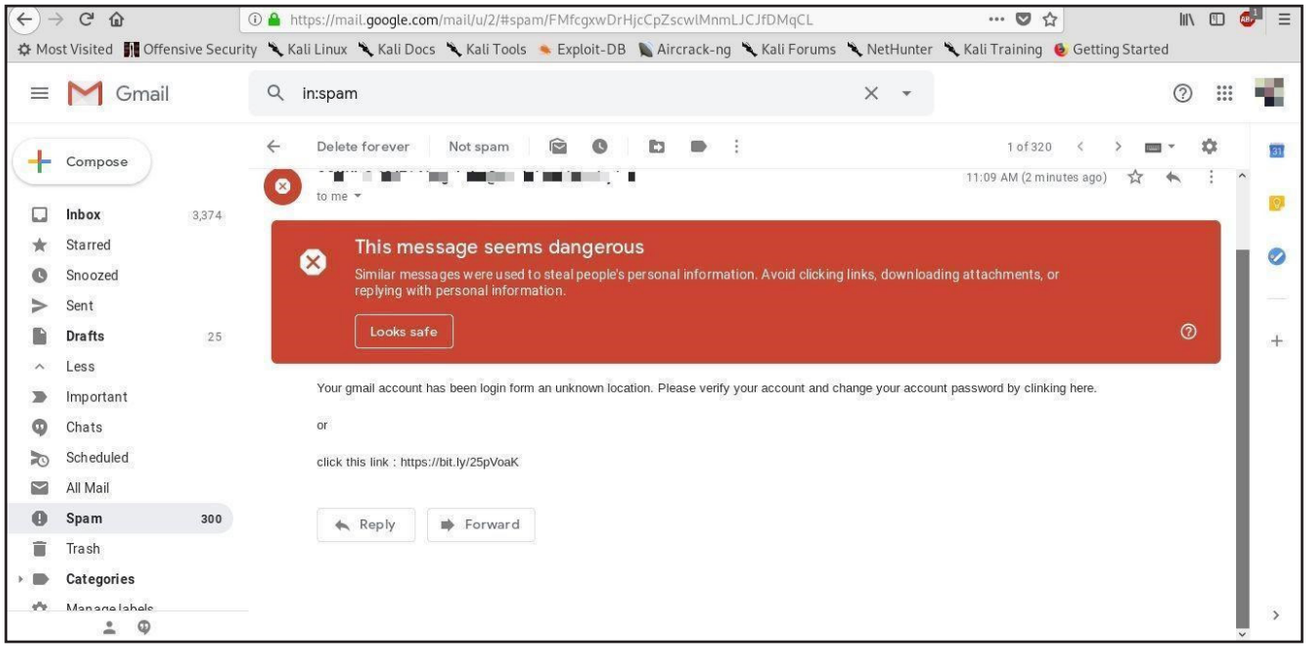

Fig. 4: Gmail Showing Warning

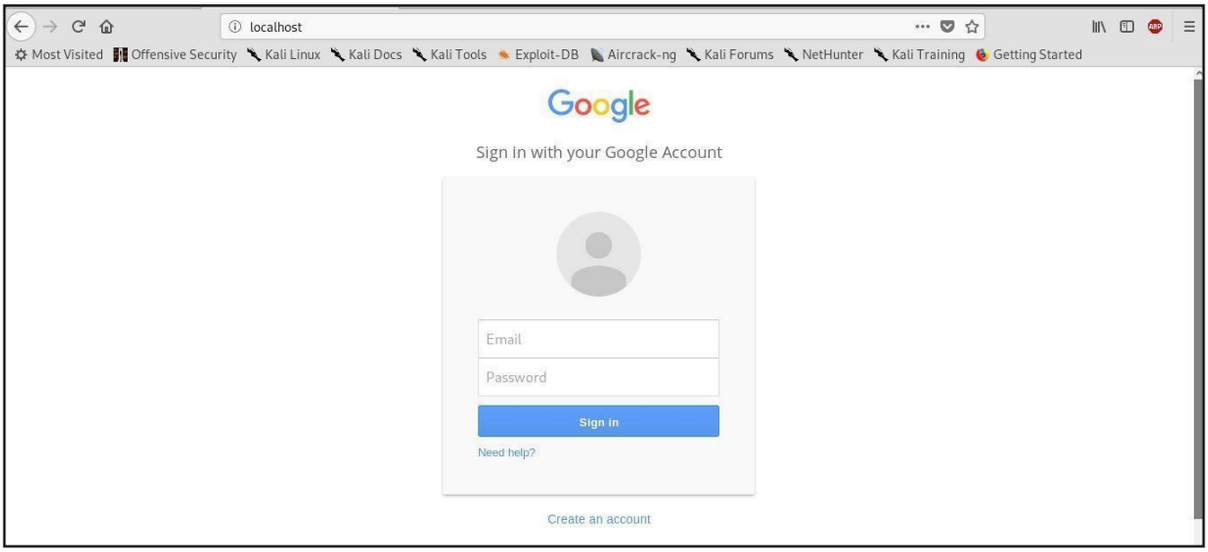

Fig. 5: Phished Gmail Page

II. Literature Review

In [1] they have discussed the different phishing attacks and anti-phishing tools. They also tell about the four-stage life cycle that are:

- Planning
- Setup
- Attack
- Collection

The different types of phishing attacks presented are:

- Deceptive Phishing - Victims are lured into clicking the links in the mail and enter sensitive credentials.
- Malware Based Phishing - This type of phishing contains sending malware which may be a keylogger or else screen logger. This type of attack can also lead to data theft.
- Other - Other types of phishing are Domain Name Server-based phishing, a man in the middle phishing and fake websites.

The anti-phishing tools that they have used in [1] are mail-Secure, Security ToolBar, ESET Security. This paper also gives more awareness to users about the present phishing problems and gives solutions.

Authors Ram B. Basnet and Andrew H. Sung [5] introduce a new technique of website phishing in which they proposed a technique to automatically detect phishing URLs by mining and extracting Metadata from the URL. In [5] they have used Logistic Regression classifier and it produces an overall accuracy of 97.2% in detecting if a URL is phished or not. The false-positive and false-negative rates are less than 1%. Even though this system had some challenges, in a real-world application mining of Meta information from the URL can be used to detect dangerous URLs.

S. Jagadeesa has compared two machine learning algorithms that are SVM and Random Forest algorithm by finding accuracy. From his work, we could see that Random Forest has much better accuracy compared to SVM in detecting a phishing URL. This helps users from entering the website and give out sensitive information. In this, the author has taken two different datasets and used those datasets in the classifier algorithms [6].

In [4] authors Himani Thakur and Dr. Supreet Kaur have addressed two major concerns. First, they discuss the history and motivation of the attackers. They also discuss the various types of phishing attacks that are presently happening. Secondly, they address the various methods that can be used to detect and defend against phishing attacks. The defence mechanism that is discussed is data mining, heuristics and blacklisting. Data mining and heuristics are better in detecting a phishing email compared to blasting. They also present about educating the users about phishing attacks so that it will be much easier to detect a phished email.

In [7] the authors discussed how phishing attacks happen and how to defend it at the entry point. They also elaborate upon the use of NM cache poisoning that is used for spear phishing. They have used 3-way DNS function exploiters that are server-client, client-client and server-server. Scanning of attachments further increased the detection of phishing emails. On average it was able to detect 83% of the phished website and 94% of the phished URLs.

In this method [8] the authors address different phishing attacks that are targeted at the end-user. Based on this they have come up with a method that can report, block, warn and educate users on detecting a phished email in a corporate environment.

They have created an environment in the mail called Webmail UI, which shows a message if the user has received a phished email. It also contains a button to report the email as phishing. This is helpful for the user so that they don't receive email from this particular sender again.

In this paper, the authors J. Ingguo, W. Ang, Tejaswini Herath, Rui C. Hen, Arun Vishwanath, and H. Raghav Rao [9] has discussed increasing the security in email service. They mainly address how a victim is disillusioned in clicking a phishing URL. They show that the two main features of a victim clicking a phished email are visceral triggers and deception indicators.

In [10] authors have developed a new cloud-based module called Cloud-Threat Inspection Appliance (CIA) that can defend against spear phishing. The main advantage of the CIA is that it uses a behavior-based detection technique rather than signature-based solutions. In this, they use hypervisor monitor for detection that is present in the hypervisor kernel. The hypervisor is a hardware-based virtualization technique that is used for detection of spear phished emails. This module cannot be evaded as it is concealed in a virtual kernel. This also contains an inbuilt mechanism to scan PDFs. The CIA mechanism was able to filter 77% of PDF before it goes to the hypervisor mechanism for deeper analysis.

III. Proposed Method

In the given analysis, we have taken two different datasets to find a suitable model. Both the dataset is taken from the UCI repository [11]. Phishing dataset one contains 2456 number of instances and 30 attributes. This dataset contains both phishing and non-phishing URL links. This dataset contains some essential features that were required to classify [12]. Some of the features that were used are page rank, age of the domain. The other important features that were used are a number of slashes, contains IP address, URL length, having any special symbols like '@', etc.

The Phishing dataset two also contains 1353 number of instances and 10 attributes. Most of the time phishing websites are mostly collected from the Phishtank data archive. This is a platform where users can submit, verify and track phishing data. This dataset is divided into three parts mainly legitimate, phished and suspicious. Out of 1353 URL collected 548 are legitimate, 702 are phished and 103 is suspicious. The important features in this dataset are 'age of domain' etc.

Table I gives the link to download the dataset and the description of the dataset.

Table II shows all the important features with the description that both datasets contain. These are some of the few important features taken to determine if a website is phished or not phished.

When we get the dataset, we first slice it. We divide the dataset into train dataset and test dataset. We train our model using the training dataset until we get the optimal result. We only use the test dataset once after we train our model precisely.

Table I: Dataset Link and Description

Dataset	Link to Download	Description
Phishing dataset 1	https://archive.ics.uci.edu/ml/dat asets/phishing+websites	This dataset contains important features that are required in predicting a phishing website.
Phishing dataset 2	https://archive.ics.uci.edu/ml/dat asets/Website+Phishing	Phishing attacks usually happen on e-commerce and e-banking websites. Phishing websites are collected from PhishTank data archive which is a free community where users can submit, verify and keep track of different phishing websites. The legit URLs were collected from Yahoo.

Table II: Important Features in Datasets

Dataset	Features	Description
Phishing dataset 1	Having an IP Address	URL contains IP address instead of the domain name.
	URL Length	The long URL address to hide the suspicious part in the URL.
	Shortening Services	Using shortening service like Tiny URL to redirect victims to a fake website.
	Having @ symbol	The browser ignores anything preceding @ symbol.
	Double Slash Redirecting	Presence of // in URL forces victims to redirect to a fake website.
Phishing dataset 2	Pop Up Window	Asking users to enter sensitive information in a popup window.
	Age of the Domain	By seeing the age of the domain to determine if the website is phished or not phished.
	Website Traffic	Phished websites will have less traffic as compared to the original website.

We use two classification models on these datasets. The algorithms that we use are Logistic regression and Random Forest [13].

A. Logistic Regression

Logistic regression comes under a supervised machine learning algorithm. Logistic regression helps in finding the relationship

between the categorical dependent variable and one or more independent variable by approximating the probability using the logistic function. The logistic function is given by the formula $\phi(z) = 1/1 + e{-}z$. This function is also called a Sigmoid function. The output of this function is an S-shaped curve that is called a Sigmoid curve. This function always produces a value that lies between 0 and 1 because z tends towards 0 to 1.

There are many times where we have to predict values that are between 0 and 1 and not exactly 0 or 1. So, Logistic regression is a technique that is used to predict values that are between 0 and 1. Logistic regression is used mainly when the dependent variable is categorical.

There are basically three types of Logistic regression:

- Binary Logistic Regression: The output of this will have only two possible outcomes.
- Multinomial Logistic Regression: This has three or more categories without ordering.
- Ordinal Logistic Regression: This also contains three or more categories with ordering.

We usually use Binary Logistic Regression for classifying if the email is phished or not because it only contains two options. So, when this model infers a value of 0.8423 on a particular email, then it implies there is a probability of 84.23% that the email is phished and the rest 15.77% means that the email is not phished.

B. Random Forest

In machine learning or data mining, random forest algorithms play a very vital role in regression and classification. To increase the predictive performance random forest uses multiple learning algorithms. This is the reason it is called ensemble learning method. A random forest works by creating a number of decision trees based on a random selection of data and random selection of variables. As the tree is based on a random selection of data and variables and there are many such random trees, it is called random forest. These random trees give the final output.

The random forest takes random data from the datasets and uses different features to create the trees. Thus, all the trees are made randomly using random features. The main advantage of the random forest is that it limits the overfitting of data. Each individual tree in the random forest gives out a class prediction and the class with the greatest number of votes will be taken as the model class. Most of the time the trees can provide the correct decision of class for the data. But at times the tree may also make mistakes by taking wrong features while classification. The output class of all the decision trees are taken and the output with most number occurrence will be taken as the final prediction value.

IV. Result and Analysis

The dataset used is divided in the ratio of 70:30 as training dataset and testing dataset. The training dataset was only used for training the model. After tuning the model to find the best parameters, only then we used the test dataset to get the results.

To increase the accuracy of the trained model repeated cross-validation was used for 10 folds. To find the best parameter grid search and cross-validation was used. This showed the most reliable parameter which gave the best results in the training dataset.

First, logistic regression algorithm was used on the Phishing dataset one. Grid search was used to tune the algorithm to find the best accuracy. The most suitable value that was found after the grid search was 5. Using this value, we got the best accuracy as 93.10% for the training dataset. When this was used for the test dataset, we got the highest accuracy of 92.45%.

Random forest was used for Phishing dataset one again. The Grid search was again used to get the most reliable parameters and found it to be 10. Then this was applied to the training dataset and the accuracy was found to be 96.28% and further this was applied to the testing dataset. The accuracy was found to be 95.36%. Thus, the random forest gave much more accuracy than logistics regression in detecting if a URL is phished or not.

The results on the Phishing dataset one is summarized in Table III.

Table III: Accuracy for Dataset 1

Algorithm	Accuracy on Training Dataset (%)	Accuracy on the Testing Dataset (%)
Logistic Regression	93.10	92.45
Random Forest	96.28	95.36

A similar process was applied to the Phishing dataset two. We first applied this dataset on logistic regression. After doing the grid search, the most reliable parameter that we found was 5. Using this parameter, the highest accuracy obtained is 83.84% for the training dataset. Then this was used on the testing dataset and got an accuracy of 82.38%.

Random forest classifier was used on Phishing dataset two. Again, grid search was used on this to get the most reliable parameter as 5. Using this we get an accuracy of 99.23% for the training dataset. The same model was used again for the testing dataset and got an accuracy of 87.45%. From this, we know that the random forest algorithm is better in detecting if a URL is phished or not.

The results of the Phishing dataset two are summarized in Table IV.

Table IV: Accuracy for Dataset 2

Algorithm	Accuracy on Training Dataset (%)	Accuracy on the Testing Dataset (%)
Logistic Regression	83.84	82.38
Random Forest	99.23	87.45

V. Conclusion

Phishing is one of the most common types of attack that is seen at present. URL phishing analysis is required to check if a URL is phished or not phished. Email is one of the most common mediums of sending a phished URL. URL phishing analysis warns the users if there is a phishing URL. This helps the user from not entering the fake website where the users are exposed to malicious code and giving out their sensitive information like password, bank details etc. The URL contains many features; it must be selected correctly to get the best results.

In our work, we have made use of two standard classifiers, namely logistic regression and random forest algorithm. We have compared the two algorithms to find the best accuracy. The Phishing dataset one contains 2456 number of instances and 30 attributes and the Phishing dataset two contains 1353 number of instances and 10 attributes. When we tried the above two classifiers random forest outplayed logistic regression by giving better accuracy in detecting a phished URL. For Logistic Regression we got an overall accuracy of 92.45% and 82.38% and Random Forest gave us an overall accuracy of 95.36% and 87.45% while using the two datasets while classifying. Thus, from this, we know that the random forest is better in finding if a URL is phishing or not.

References

[1] T. Dakpa, and P. Augustine, "Study of phishing attacks and preventions," *International Journal of Computer Applications*, vol. 163, no. 2, pp. 5-8, April 2017.

[2] R. G. Brody, E. V. Mulig, and V. Kimball, "Phishing, pharming and identity theft," *Academy of Accounting and Financial Studies Journal*, vol. 11, no. 3, pp. 43-56, 2007.

[3] A. Mahalakshmi, N. S. Goud, and G. V. Murthy, "A survey on phishing and it's detection techniques based on support vector method (SVM) and software defined networking (SDN)," *International Journal of Engineering and Advanced Technology*, vol. 8, no. 2s, pp. 498-503, December 2018.

[4] H. Thakur, and S. Kaur, "A survey paper on phishing detection," *International Journal of Advanced Research in Computer Science*, vol. 7, no. 4, pp. 64-68, January 2017.

[5] R. B. Basnet, and A. H. Sung, "Mining web to detect phishing URLs," *2012 11th International Conference on Machine Learning and Applications*, Boca Raton, FL, 2012.

[6] S. Jagadeesan, A. Chaturvedi, and S. Kumar, "URL phishing analysis using random forest," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 20, pp. 4159-4163, 2018.

[7] D. N. Pande, and P. S. Voditel, "Spear phishing: Diagnosing attack paradigm," *International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, pp. 2720-2724, 2017.

[8] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, "Intelligent phishing website detection using random forest classifier," *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, Ras Al Khaimah, pp. 1-5, 2017.

[9] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email," *IEEE Transactions on Professional Communication*, vol. 55, no. 4, pp. 345-362, December 2012.

[10] C. Lin, C. Tien, C. Chen, C. Tien, and H. Pao, "Efficient spear-phishing threat detection using hypervisor monitor," *International Carnahan Conference on Security Technology (ICCST)*, Taipei, pp. 299-303, 2015.

[11] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml

[12] J. C. S. Fatt, C. K. Leng, and S. S. Nah, "Phishdentity: Leverage website favicon to offset polymorphic phishing website," *2014 Ninth International Conference on Availability, Reliability and Security,* Fribourg, 2014.

[13] T. Ayodele, "Types of machine learning algorithms," *New Advances in Machine Learning*, 2010.

[14] M. Khonji, Y. Iraqi, and A. Jones, "Phishing detection: A literature survey," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 2091-2121, Fourth Quarter 2013.

[15] N. Stembert, A. Padmos, M. S. Bargh, S. Choenni, and F. Jansen, "A study of preventing email (Spear) phishing by enabling human intelligence," *2015 European Intelligence and Security Informatics Conference, Manchester*, 2015.

[16] http://dataaspirant.com/2017/05/22/random-forest-algorithm -machine-learing/