International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | June 2022 www.irjet.net p-ISSN: 2395-0072

Phishing Detection using Decision Tree Model

Aman Ahamed[1], Dr. Ramananda Mallya K[2], Anushri A Shetty[3], Delisha DSouza[4], Ashokkumar Tirumala Gopi[5]

[1,3,4,5] Dept. of Information Science and Engineering, Mangalore Institute of Technology & Engineering, Moodbidri.
[2] Associate Professor, Dept. of Information Science and Engineering, Mangalore Institute of Technology & Engineering, Moodbidri.

Abstract - *In the modern days the security is the main concern in this rapidly evolving world with the technology advancement. There are many of the cases which led to huge number of financial losses by common social attacks. These attacks are the one that made technically or to the targeted device. It's in the form of the virus or Trojan or it may be in the form of a normal website link which we also called as the URL (Uniform Resource Locator).These URLs contains the software or the malicious program which takes out the users all the valuable and more secured and private information (or sensitive data) when this URL is entered by the user in his remote machine. This form of attack is known as Phishing. Normally the user will see the web page appearing as a simple and interactive but in behind it is more and more dangerous one. A fraudulent try made by the attacker in order to steal the users data all the private information like we have username, password, and private details like users financial bank account and details of the users credit card. To avoid these attacks there are many advancements in artificial intelligence and machine learning, which have efficient and more compact techniques to find out the fake URLs. A machine learning model made up of decision tree algorithm is developed which will scan and filtes out the common words and learns the specific features and then it will provide the appropriate result.*

Key Words: Uniform Resource Locator, Decision Tree, Security, Machine Learning

1. INTRODUCTION

Phishing in layman's terms is just giving the user by an attacker the web link or we say it's a programmed URL or abbreviated as Uniform Resource Locator where the term programmed contains the scripts or the virus or malicious infinite time running program or a zombie the process that when invoked runs itself and it will do those tasks or the commands ordered by the attacker.

This URL seems to be the normal one. But the attacker uses this in order to get all the private and confidential information from the user so that there is some benefit enjoyed by the attacker. The domains are more. These attacks majorly occur in the field of online payment sector, web-based email, and in the cases of cloud storage [1]. 78 % of the attacks are made only in the domains like web-based mailing systems in and online payments. The remaining 22 % of the attacks are made for industrial sectors.

The consequences and the results when phishing attacks occur will cause huge financial losses in the case of the banking domain. The current era internet revolution has increasing and the advancement in technologies is also increasingly growing, it has become an attractive place for all potential users. Phishing is normally imitated by mimicking as a trustworthy person or an entity on the Internet which is done by integrating both social engineering and technological tricks.

Lastly, we know that economic and financial helpers such as banks are now becoming more important on the Internet thereby making people's lives in this world easy. Security and the safety of the people against these frauds are mandatory in this digital era. Phishing is a major attack or threat when it comes to securing the website.

There are mainly two types of phishing attacks one is called the Spear phishing, which means targeting the specific and private/public companies and the individual people. The other one is called Clone phishing. This means that this is an attack where the real or the original mail containing an additional attachment or the URL/link is copied to a fresh (new) mail with malicious attachment or URL [2].

2. BACKGROUND

The main goal to achieve successful phishing is the user's data, assets, or private information that is stolen through a fake website [3]. If we detect bad URLs in the early stage this is the best strategy to avoid contact with phishing websites. Phishing websites are to be determined through their basic domains [4].

These are related to the URL that needs to be registered. We will implement machine learning algorithms to classify the data in this case. The basic algorithms used here are as follows. The proposed technique gives 95% accuracy. This mainly depends on the quantity of data set divided into training and testing.

Machine learning implies training the machines to reduce human effort in any domain. Machine learning with a combination of AI (Artificial intelligence) is the most popular thing that is booming. This learning provides some pre-written inbuilt models so that the model can train the data and test the accuracy of the work [5]. It is very highly scalable and has higher computing power. This approach works efficiently in large datasets [6]. This also removes the drawback of the existing approach and can detect zero-day attacks.

Machine Learning-based classifiers are efficient classifiers that achieved an accuracy of more than 99%. Performance depends on the size of training data, feature set, and type of classifier [7]. The limitation of this is it fails to detect when attackers use a compromised domain for hosting their site [8].

Many researchers have performed various analyses on different areas of application [9]. Most research has worked on improving the accuracy of phishing website detection using different classifiers.

Various classifiers are used and among them is ELM. Among all of these tree-based classifiers, DT, and RF are best to increase the dataset as per THE literature surveys. Therefore, the proposed approach will be phishing website detection using logistic regression [10].

3. METHODOLOGY

In this project, we have first imported a dataset that contains approximately 12000 data in which half of the data is phishing-related data and the rest 50 % of the data is original data. Dataset is divided into training data and testing data.

Using convenient machine learning algorithms such as random forest classifiers and support vector machines are used to classify the data based on extracting its features. The model is a decision tree classifier. The model is trained by giving both the original and phishing link to find out the differences in them so that it will give the correct accuracy when training data is fed to the model.

The front-end design part consists of a simple static page that is written using Hypertext Markup Language. In the design part, we are normally providing the user input to insert the link or the URL which is either a real one or the fake one.

In this one, the design part represents the simple login page. The login page is the one that takes the input as the URL from the user that is processed at the backend. The form is made using the simple HTML and CSS code that consists of a textbox for the input by the user to be entered and a submit button that takes the data to the backend that is written in python.

The URL is the main input to detect whether the website is real or fraudulent. Typically a fraud website's URL differs from the original website's URL. Checking of the website is done by feature extraction, which includes extracting the important characters from the URL. There are mainly four types of features that can be extracted. Address bar features abnormal features, Domain Based features, HTML and Java script based specific best features. The application design front page is shown in Fig.1.

Fig -1: User Interface Design

The format of data containing real and fake links is stored as a CSV file which is shown in Figure 2.

	A	
1	URL	Label
2	nobell.it/70ffb52d079109dca5664cce6f3173	bad
3	www.dghjdgf.com/paypal.co.uk/cycgi-bin/w	bad
4	serviciosbys.com/paypal.cgi.bin.get-into.herf	bad
5	mail.printakid.com/www.online.americanexp	bad
6	thewhiskeydregs.com/wp-content/themes/w	bad
7	smilesvoegol.servebbs.org/voegol.php	bad
8	premierpaymentprocessing.com/includes/bo	bad
9	myxxxcollection.com/v1/js/jih321/bpd.com.(	bad
10	super1000.info/docs	bad
11	horizonsgallery.com/js/bin/ssl1/_id/www.pa	bad
12	phlebolog.com.ua/libraries/joomla/results.pl	bad
13	docs.google.com/spreadsheet/viewform?for	bad
14	www.coincoele.com.br/Scripts/smiles/?pt-br	bad
15	www.henkdeinumboomkwekerij.nl/language	bad
16	perfectsolutionofall.net/wp-content/themes	bad
17	lingshc.com/old_aol.1.3/?Login=&Lis=1(	bad
18	anonymeidentity.net/remax./remax.htm	bad
19	dutchweb.gtphost.com/zimbra/exch/owa/ul	bad
20	www.avedeoiro.com/site/plugins/chase/	bad
21	asladconcentration.com/paplkuk1/webscrcm	bad

Fig -2: The Data Set

The CSV file contains the combination of original URLs and the fake URLs which are extracted from Phish Tank or Kaggle websites. This mainly contains more than 25000 rows and mainly two columns. A first column is named URL and a second column is named label. The label column contains two values namely good and bad. Label good or 0 indicates that the URL is a good URL and the label bad or 1 indicates that the URL is a fake one.

4. RESULTS

Initially, the dataset contains lists of original links and fake links. This data is given as the input to the model called logistic regression. This will classify the data and perform the regression analysis on the data to type the URL as phishing or original.

The Decision Tree model is going to learn from the training data to test the features present in the testing data. The dataset is read through the module called pandas. And the URLs in the dataset are labeled as 0 or 1.

The label 0 represents that the given input link is the original link and the label 1 represents that the input link or the URL which is fed to the machine as the input is the fake one. So, the dataset contains a labeled URL. The URLs which do not have the label either 0 or 1 are removed from the group so that the training will be in an accurate manner.

The proposed model now classifies the data based on the given input and calculates the accuracy or the amount of data that the model has learnt by reading the whole dataset and passing the test data.

Whenever the input is provided the model will yields 95% of the training accuracy and provides the valid results. So, the model is ready to accept the data so that it can go through and iterate each and every data for training. The Chart 1 shows the accuracy of the model.

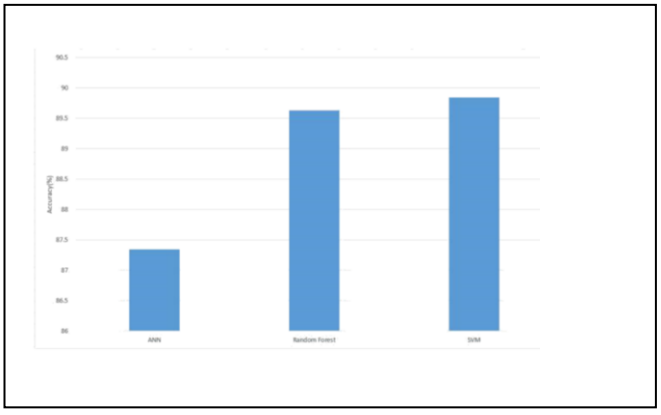

Chart -1: The Accuracy of the Model

The above chart shows the training accuracy of the models and the best fit model is chosen to be random forest as it gives the highest accuracy rate in classification of the data frequency.

In our project there is only one message that shows whether a link is real one or the fake one. Display the appropriate results after performing the tasks on the backend when the input is fed into the model. The User Interface Output of the model is shown in Figure 3.

Figure -3: User Interface Output

5. CONCLUSIONS

In this part how to avoid common types of phishing attacks is explained. First of all, proper education awareness is needed. Those people who are using the internet worldwide have to be provided with some basic knowledge about all the security measures and the alerts which are mainly given by the experts.

Every user around the world should know not to blindly follow and click on the links to those specific websites where they enter their sensitive information like username and password.

It is very necessary to check the URL or the link before entering that website. In the Future System can upgrade itself automatically in order to Detect the web page and the performance of the running Application with the current working web browser.

In this project, we implemented the classifier such as the decision tree. This classifier is used to detect phishing URLs. In detecting phishing URLs, there are two steps. The first step is to the extraction of a specific set of features from the URLs and the second step is classification of URLs using the model developed with the help of the training set data.

This project uses the data set that provided the extracted features. One of the main concerns in the decision tree classifiers is over fitting. Generally, the decision tree classifies the training set data very well but gives poor

results with a testing dataset. It is required to match the algorithmic decision tree to work better with testing data.

The algorithmic decision tree provides the highest classification accuracy of 95 percent with more features in the data set. In addition to that better accuracy may be improved through the ensembling of trees.

REFERENCES

[1] Das, Avisha, "SoK: a comprehensive reexamination of phishing research from the security perspective," IEEE Communications Surveys & Tutorials, Volume 22, Issue 1, 2019.

[2] J. Ma, S. S. Savage, G. M. Voelker, "Learning to detect maliciously URLs," ACM Transactions on Intelligent Systems and Technology, Volume 2, Issue 9, 2011.

[3] S. Purkait, "Phishing countermeasures and their effectiveness–literature review," Information Management & Computer Security, Volume 20, Issue 5, pp. 382–420, 2012.

[4] N. Abdelhamid, A. Ayesh, F. Thabtah, "Phishing Detection based Associative Classification," Data Mining. Expert Systems with Applications Volume 41, pp 5948-5959, 2014.

[5] Tan CL, Chiew KL, Wong K, "PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder," Decision Support Systems, Volume 88, pp 18–27, 2016.

[6] Almseidin M, Zuraiq AA, Al-kasassbeh M, Alnidami N, "Phishing detection based on machine learning and feature selection methods," International journal of interactive mobile technology, Volume 13, Issue 12, pp. 171–183, 2019.

[7] Zamir A, Khan HU, Iqbal T, Yousaf N, Aslam F, "Phishing web site detection using diverse machine learning algorithms," *The Electronic Library*, Volume.38, Issue.1, pp. 65–80, 2019.

[8] Ramananda Mallya K, and B. Srinivasan, "Usable authentication for cloud based mobile learning in engineering education," International Journal of Civil Engineering and technology, Volume 10, Issue 4, pp. 209-218, 2019.

[9] Ramananda Mallya K, and B. Srinivasan, "Secure Architecture for Cloud based Mobile Learning," International Research Journal of Engineering and technology, Volume 6, Issue 7, Pages 1775-1779, 2019.

[10] Sahingoz OK, Buber E, Demir O, Diri B, "Machine learning based phishing detection from URLs," Expert System Application, Volume 117, pp. 345–357, 2019.