An Effective Model of Terminating Phishing Websites And Detection based On Logistic Regression

Shaik Riyaz
UG Student
Dept of computer science and engineering
S.R.M *institute of science and technology*
Kattankulathur, India
sr9087@srmuniv.edu.in

R B Srinivas Raju
UG Student
Dept of computer science and engineering
S.R.M *institute of science and technology*
Kattankulathur, India
rbsraju99@gmail.com

K.R.Jansi
Associate professor
Dept of computer science and engineering
S.R.M *institute of science and technology*
Kattankulathur, India

Abstract—**These days phishing is one of the greatest and the quickest developing risk in light of the fact that phishing assailants will get data which was entered by the client and phishing aggressors will utilize that data without the client information. At present personal information is the most important thing compared to money, So phishing website attackers are focusing on the personal information of the user and they taking advantage when the user enters there personal information in the phishing websites. So the main theme of these projects is to avoid misuse of personal information by phishing attackers. To overcome this problem machine learning algorithm (Logistic regression) is used and provided with the massive dataset. From that dataset, it trains the algorithm and helps in detecting the new web links which are fraudulent links. Attackers disguise their website as legitimate and try to get data from the user for which they make users visit a website and get the personal information that is needed. Before trying to enter into those websites it is important to check whether the given link is good or a phishing link. By checking the link we can save ourselves from the attackers and can keep our data safe.**

KEYWORDS: PHISHING , MACHINE LEARNING , LINEAR REGRESSION, ACCURACY, WEBSITE LINKS.

I. INTRODUCTION

The Internet has become an imperative foundation that carries extraordinary comfort to human culture.

Notwithstanding, the Internet is additionally described by some unavoidable security issues, for example, phishing, malignant programming, and protection exposure, which have just carried genuine dangers to the economy of clients. The APWG describes phishing as a criminal instrument using both social structure and concentrated deception to take singular character data and budgetary record capabilities of buyers. Phishing websites have so many ways to get the client's personal information, Among those ways clicking weblink is one of the dangers and popular ways. In this phishing attack, the attacker sends an attractive link to the client through the mail or any other way. If you open that link and proceed to fill all the sensitive information then phishing attackers take advantage of the received information from a common man is used to taking money from the account or selling information to some third parties which can be used any were. Nowadays there is much software that detects fraudulent ones and blocks. But most people ignore the links whether it is good or bad, So to detect link's we are implementing a machine learning technique that finds out whether the link is good or bad.

In our modern digital life, Phishing attacks are expanding rapidly because phishing attackers are taking advantage of technologies like email, mobiles, and web links. The daily clients may receive many links from email or any other way or even the client may click the link without knowledge about the link that type of links may be or bad, So the main idea of the project is to provide awareness to the client about the link by using the machine learning algorithm (linear regression).

This software is used to find whether the link that leads to a website is good or bad. This software gets the data from the data set and by using logistic regression a machine learning algorithm trains that algorithm. After the algorithm is trained it will be able to find the phishing links. In this process, it breaks the given link accordingly and check the main keywords from it and later check the database which contains good and bad links. When compared if it matches one of those it gives output accordingly. If the output displayed is good we can proceed else quit the process. The list will be updated when it finds a new link. Firstly this algorithm checks whether the link follows the web rules or not. Checking is done by breaking a link into its subdomains which is given in the code of the software. Once the link passes all rules and procedures. It displays the output as good or bad

AI. Proposed Work

A. Web page:

To extract the features of the link first we need a medium to enter the link. This project contains a webpage that is created with HTML and CSS. input, button, h1, a, img, and p are the elements used to make the webpage and By using css style properties like hover, animation, transform, font-style, background and container which makes webpage look premium. After entering the link that webpage directs to python code where it will extract the features using machine learning algorithms

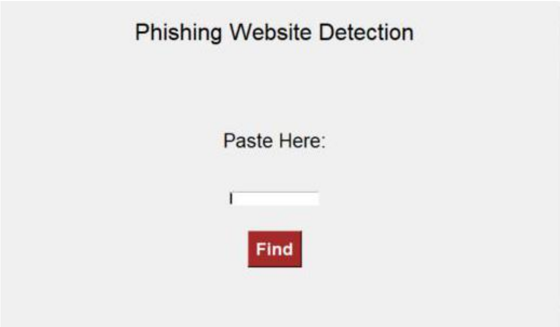


Fig 2.1 webpage

B. Feature extraction:

In order to reach a website on the internet, we use a URL like a medium. Phishing links are generally disguised as legitimate by doing this attacks, attacker can make a user click on their web links which are pretending to be the original ones. These phishing links have few identifiable features. As listed in the above fig 2 .these features are divide into four classes they are 1) Address bar

2) Abnormal Features

3) Domain features

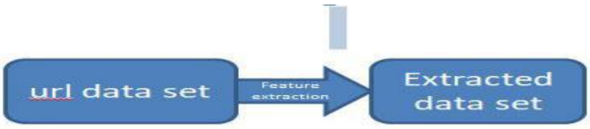
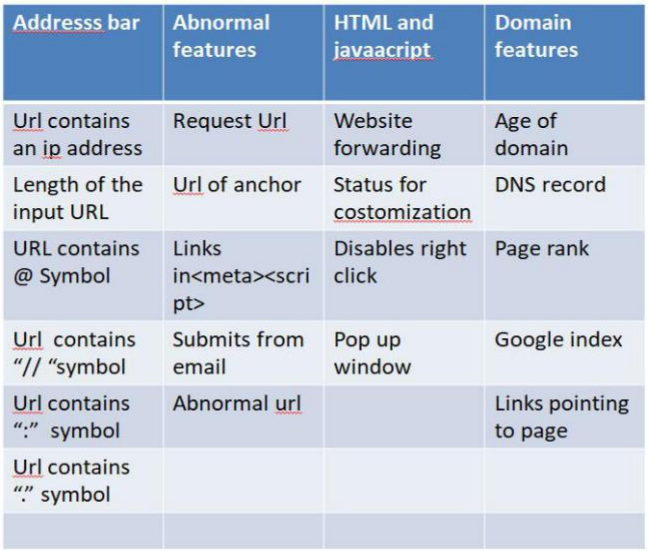


Fig 2.2:feature extraction



Addresss bar	Abnormal features	HTML and javaacript	Domain features
Url contains an ip address	Request Url	Website forwarding	Age of domain
Length of the input URL	Url of anchor	Status for costomization	DNS record
URL contains @ Symbol	Links in<meta><script>	Disables right click	Page rank
Url contains “// “symbol	Submits from email	Pop up window	Google index
Url contains “:” symbol	Abnormal url		Links pointing to page
Url contains “.” symbol			

Fig 2.3:sensitive features of the phishing links

1. Address bar features:

In this IP address special characters like @,"",/,.,_,- etc.. and the length of the input URLs are needed to specified to detect phishing attacks accurately. It also has a length of the URL’s, Short URL”s existence. We also check the features like Domain registration length. We check all these features and see whether the given link contains all these or not if the criteria is satisfied then only it will go to the other one.

2.Abnormal Features:

Abnormal features can be checked from the requested URL and the link tags like <META>,<Script>,<Link>, The S F H status information. Checking from the handler. Check Whether the web page having the link that is going to submit the data entered to any mail. Lastly, say it as an Abnormal URL or not.

3.Domain features:

This section of the code generally checks the age of the domain because that phishing websites have a very short time. It checks the DNS record Web Traffic and page rank as the phishing pages have very few visitors it will have the index of 0 or 1 not more than that. It also checks the google index for that and the number of pages pointing to that page all this information of the features is saved in the Statistical report.

III. RELATED WORK

1. A Systematic Review of Software Based Web Phishing Detection

Phishing is one of the forms of attack that that is growing in day to day life. To acquire the personal and confidential information from the users and use it against them. In the past six years is around 36.29% of the phishing websites are detected.

In those 97.36% was found in last two years. This is due to increase in usage of the internet. This brings the biggest task for the cyber security. To find new methods to stop them completely which is not possible but we can stop them for few years.

By finding new methodology to detect those URLs and block them. The existing methodologies differ in terms of the method and the process.

2. Component Page Similarity Phishing Detection

Social networking sites are the most rapidly increasing sites which are used by the most of the users .These site will contain all the personal information of the user. There are many of those kinds in the market .

So there should be most research in this particular area where you can find more users increasing daily. There are few who attack these platforms to get the data from it. Few will make a new social website to acquire the data which exactly looks the same.

Those pretend to be the same website and few may get fooled by these website and give their information sensitive data in it. In order to find these websites we check the webpage similarity components in the actual webpage and find weather the page is trusted or not.

3. Intelligent rule-based phishing websites classification

Phishing is said to be taking the sensitive information from the user and uses it against him. There are Different types of phishing attacks. No one phishing attack is similar to the other. But the process of phishing could be classified.

It is like attacks can be classified by the means of mode of acquiring the data. One of them is like through links via email. The other like creating a fake website and taking the information and there are few in which with one click on the link and the control would be lost and the entire control will be in the attacker hand. So we cannot have a same solution to every attack. But we can have a same process by classifying them. So that approach to it would be easier.

This project is all about classifying the attack by the mode of process to acquire the data..

4. Detecting phishing attacks from their network performance characteristics

network based phishing which is done to slowdown the certain server so that more traffic in server can occur which gives a chance to the attacker to get into someone's information this will happen the most this can also could be done by phishing it's is like a bug in the link by which on clicking it does nothing but gets into your device and have the control over the network and get the data.

so in order to find such links we have the four lever based checker in which the process will undergo into these steps to get the output from it.These will take the things from the link and help us to find the phishing link.

5. Study of Phishing and Malware Warnings.

However the phishing attack may happen only if we allow to do it .Our computer is good enough to block certain links that are harmful to the pc. At the end it is the hands of the user so this methodology will look into human actions while we clicking those links and in the same them find the phishing links. This study will help us to find the same kind of websites or links so that we can find them easily and shut them down and save people from getting attacked.

The implementation process consists of the following types:

- Pre processing and data collection
- Data selection

- Data transformation
- Data selection
- Data sets
- Feed into algorithm

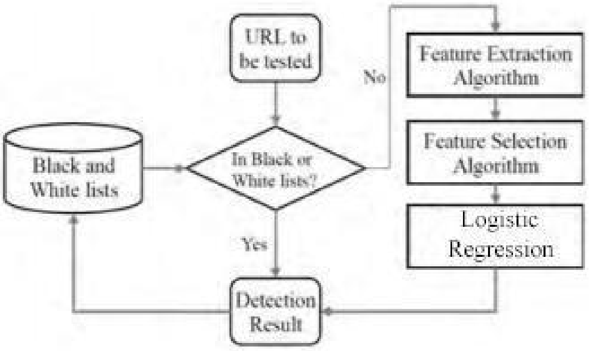


Fig 3.1 flowchart of the working process

A. Pre processing and data collection

Data Collection is one of the most important tasks in building a machine learning model. It is the gathering of task-related information based on some targeted variables to analyze and produce some valuable outcomes. However, some e data may be noisy, i.e. may contain inaccurate values, incomplete values or incorrect values.Hence, it is must to process the data before analyzing it and coming to the results. Data pre-processing can be done by data cleaning, data transformation, data selection.

B.Data cleaning

Data cleaning is the process in which we get noisy data. This noisy data in the sense which is not fully completed and has some missing values that will be written fully. This will be taken and the entire data will be read and will take the possibility of clean data so that the data can go to the next process.

C.Data transformation

Data transformation may include smoothing, aggregation, generalization, transformation which improves the quality of the data.

D.Data selection

Data selection includes some methods or functions which allow us to select the useful data for our system. Data input: After finding the best algorithm we are used that algorithm for finding the phishing website. Then we are going to give input to the algorithm and we will find the output based on the output.

E.Data sets:

A collection of data of a particular instance in order to work on the machine learning we need those sets for many purposes Training Data set: A data set that To train our model, we feed in our machine learning algorithm.

This is the data type used to provide an unbiased evaluation of the final that is completed and fit on the training data set that data is used for the learning purpose.

F. Feed into the algorithm:

Before getting to know about the logistic regression we should know what is regression first.

Regression: It is a predicting model. It predicts the relationship of two values by using the probability

Dependent:-What is the sale on Feb 14. Her sale is the dependent variable.

Independent:-i)Number of product sales on Feb 14.
ii)quantity(predector).

Logistic Regression:-

Logistic regression gives the result in 0’s and 1’s with this values it predicts the ouput of the given data like High(or)Low etc..

Normally linear regression has a range from 0-∞and -∞ to +∞.

In logistic regression is using sigmoids function so linear line ihas to be clipped to 0 and 1.By sigmoid with this the threshold value indicate the probability of 0or 1 or good or bad.

Equation:

These equation are derived from the Straight line equation.

$$y\ m1x1\ \ c()$$

for the logistic regression range 0→∞ to convert it to 0→1 we use

equation:- $\frac{y}{1-v}$ if y=0 then equation become 0.so y value ranges between y=1 → ∞

for transforming it further to get between -∞ → +∞ use

$$\log\{\frac{y}{1-y}\}$$

final logitic regression equation .

1)collection of data

2)Analysing data

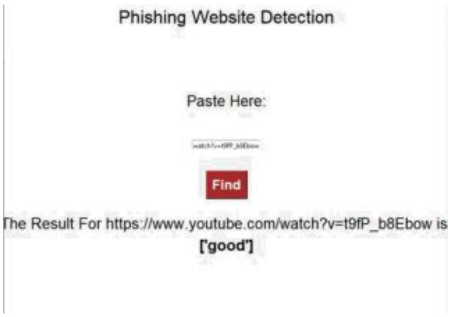
3)Data wrangling

4)Test and train

IV. RESULTS DISCUSSION

Based on that data set we can get the result used our decision tree and logistic regression algorithm to predict the result.

Here we can also find out the accuracy rate of the algorithm. It will be helpful for finding phishing website whether it is good or bad



Phishing Website Detection

Paste Here:

The Result For https://www.youtube.com/watch?v=t9fP_b8Ebow is ['good']

Fig: 5.1 sample output

V. CONCLUSION

Phishing is an unspeakable risk in the web field. In this episode, the basic man inputs individual data to a bogus site which looks equivalent to a typical site. We had done a survey on phishing methods based on visual comparison. This provided a good understanding of the attack and many solutions. Various approaches have conversed in this paper for the detection of phishing; however, most of the methods still have boundaries like accuracy, failing to distinguish objects, and so forth.But In this paper Logistic regression technique finds the accuracy as it is an open challenge in this phishing field the accuracy obtained is 95% percent and it may further increase based on the research.

~!

VI. FUTURE WORK

In the future, The web page is going connect with sign in and sign up page which makes increasing in the security of the web page and clients . If we find any better algorithms to increase the accuracy or for fetching data then replace that algorithm with an old algorithm, so that the speed of the web page will increase.

VII. REFERENCES

[1] (2016). PhishMe Q1 2016 Malware Review. [Online].Available:https://phishme.com/project/phishme-q1-2016-malware-review/

[2] A. Belabed, E. Aimeur, and A. Chikh, ‘‘A personalized whitelist approach for phishing webpage detection,’’ in Proc. 7th Int. Conf. Availability, Rel. Security (ARES), Aug. 2012, pp. 249–254.

[3] Y. Cao, W. Han, and Y. Le, ‘‘Anti-phishing based on automated individual white-list,’’ in Proc. 4th ACM Workshop Digit. Identity Manage., 2008, pp. 51–60.

[4] T.-C. Chen, S. Dick, and J. Miller, "Visually detecting identical Web pages: Application for phishing detection," Internet Technol. vol. 10, no. 2, pp. 1–38, May 2010.

[5] N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, and J. C. Mitchell, "Clientside protection against Web-based identity theft.'11th Annu. Netw. Distrib. Syst. Security Symp. (NDSS), 2004, pp. 1–16

[6] C. Inc. (Aug. 2016). Couldmark Toolbar. [Online]. Available: http://www.cloudmark.com/desktop/ie-toolbar