URL BASED DETECTION OF PHISHING SITES USING RANDOM FOREST CLASSIFIER

Vijey Nathan V
Department of Computer Science and Engineering
SRM TRP Engineering College, Trichy, Tamilnadu, India

Siddeshwar S, Kaviarasan V, Vikash R, Rishikrishna J
Department of Computer Science and Engineering
SRM TRP Engineering College, Trichy, Tamilnadu, India

Abstract - Phishing is a type of social engineering attack often used to get user data, including login details and credit card information. It occurs when an attacker, playacting as a legitimate entity, dupes a victim into opening an email, instant message, or text message. Phishing will direct the user to visit where their details such as password, credit card number, social security or bank account details can be theft by the attacker. The attacker may create a will like a legitimate page which plays a major role to steal the information of an individual. We are developing a website that can identify a website that is legitimate or not, which is developed by using machine learning. We are going to use 30 features that help to determine a site is legitimate or a phishing site. This model will ensure more security regarding internet scams and password-stealing purposes. We hope that this model will help us to fulfil the expectation that exists for a long time.

Keywords – Phishing, Machine learning, Random forest, Classification.

I. INTRODUCTION

Phishing is the fraudulent use of electronic communications to defraud and take benefits of users. Phishing attacks attempt to gain sensitive, private information such as login details, credit card details, network credentials, and more. By masquerading as a genuine individual or institution by phone or email, attackers use social engineering to manipulate victims into performing specific actions—like clicking on a malicious link or attachment—or wilfully divulging confidential information.

EFFECTS OF PHISHING ATTACKS:

Both individuals and organizations are at danger; almost any kind of personal or organizational data may be valuable, whether it be to commit scam or access an institution's network. Additionally, some phishing frauds can target institution data in order to support intelligence efforts or state-backed spying on opposition groups. There are 3 types of Phishing attacks

- Spear Fishing
- Clone Phishing
- Whaling
SPEAR FISHING:
Spear phishing email messages won't look as odd as more general phishing attempts. Attackers will often gather details about their targets to fill emails with a more authentic context. Some attackers even steal business email communications and create highly customized messages.

CLONE PHISHING:
Attackers are able to view original, previously delivered email messages, make a nearly exact copy of it or "clone" and then change an attachment or link to something malicious.

WHALING:
Whaling specifically targets high profile and/or senior administrators in an institution. The content of a whaling attempt will often present as a legal information or other high-level executive business.

II. PROPOSED ALGORITHM

2.1 Random forest algorithm –
The model improves the existing model by adding features to the training model by 10 attributes that improve the accuracy of the model. The potency of the model is improved by,

- Feature choice
- Algorithm standardization
- Adding a lot of data

![Variable importance chart]
Within the projected model 30 features were thought of to seek out whether or not the site is legitimate or not, that got an accuracy of 97.2%. The projected is made victimization Random Forest Classifier that yields higher accuracy compared to different classification algorithms.

The 30 features are:

- URL Length
- Shortening service
- Having @ Symbol
- Double slash redirecting
- Prefix Suffix
- Having Sub Domain
- SSL final State
- Domain registration length
- Favicon
- Port
- HTTPS token
- Request URL
- URL of Anchor
- Links in tags
- SFH
- Submitting to email
- Abnormal URL
- Redirect
- On mouse-over
- Right Click
- Pop-up window
- I-frame
- Age of domain
- DNS Record
- Web traffic
- Page Rank
- Google Index
- Links pointing to page
- Statistical report

Advantages: The proposed model uses 30 features to detect whether the website is legitimate or not which improves the accuracy and reliability of the model.

2.2. Dataset collection –
In this module, the dataset for training the model is chosen. The dataset is selected in a CSV (comma-separated value) which has 30 unique features and one target. The dataset does not have any Duplicate and Irrelevant observations, which is suitable for training the machine learning model. The data set has the following features having_IP_Address,URL_Length,Shortining_Service,having_At_Symbol,double_slash_redirecting,Prefix_Suffix,having_Sub_Domain,SSLfinal_State,Domain_registration_length,Favicon,port,HTTPS_token,Request_URL,URL_of_A...
In this module, the machine learning model is trained with the Random Forest algorithm using the collected dataset. The random forest model creates a decision tree on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. The trained model is then serialized using the pickled model.

2.4. Feature extraction –
In this module, the user given input is accepted using a python program and the features from the URL are extracted using python libraries. This module returns a set of values that are passed to the pickled model for predicting the result. The trained model will return any one of the following results \{1, 0, -1\} where,

- 1 stands for a Phishing site,
- 0 stands for a Suspicious site,
- -1 stands for a Legitimate site.
III. EXPERIMENT AND RESULT

When the user enters a URL, the extension takes the URL using the GET method and passes the same to the python code then extracts all the features from the URL and forms an array. We then test this on the trained classifier of Random Forest.

The Proposed model gives 97.23% accuracy. Using the Random Forest Classifier.

![Output 1](image1.png)

**Figure 4. Output 1**

![Output 2](image2.png)

**Figure 5. Output 2**

IV. CONCLUSION

Since the future is moving towards automation, our model may be the first step towards achieving that goal. Now, the proposed model will help in finding the phishing site. It also helps to safeguard the users from spam links. If the user feels suspicious about a specific link they can use our model for finding whether it is a legitimate or phishing site. The Proposed model gives 97.23% accuracy. We still need time to improve its accuracy by taking in account of more features and more instances. So this model can be used commercially for internet security purpose. Thus this model helps us to reduce the fraud happening in the internet and bridges the gap between the human and the internet.
REFERENCES


