URL Phishing Detection System

For

Networks Threats and Attacks Laboratory

BE Computer Engineering

By

Zahan Shahana 60004150113
Ashish Jain 60004150031

Faculty-In-Charge
Prof. Pranit Bari

Name of the Group Members with SAP Id with Batch No:

Zahan Shahana	60004150113 A2 Batch
Ashish Jain	60004150031 A2 Batch

CLASS with Division:

BE A Division

Name of the Supervisor/Guide:

Prof. Pranit Bari

Title of Report:

URL Phishing Detection System

Field of Project: Networking

Area of the project: Mitigation of Networking Threats

Abstract-

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels. Visual similarity based phishing detection techniques utilise the feature set like text content, text format, HTML tags, Cascading Style Sheet (CSS), image, and so forth, to make the decision. These approaches compare the suspicious website with the corresponding legitimate website by using various features and if the similarity is greater than the predefined threshold value then it is declared phishing.

Detection of phishing websites is a really important safety measure for most of the online platforms. So, as to save a platform with malicious requests from such websites, it is important to have a robust phishing detection system in place.

We try to build such a model a machine learning model that, given a URL it identifies whether it is a Phishing URL or not based on a pre-trained Random Forests Model.

Background Work/Prior History-

First phishing attack was observed on America online network systems in the early 1990s where many fraudulent users registered on America online network systems website with fake credit card details. America online network systems passed these fake accounts with a simple validity test without verifying the legitimacy of the credit card. After activation of the fake account, attackers accessed the resources of America online system. At the time of billing, America online network systems determined that the accounts were fraudulent, and associated credit cards were also not valid; therefore, America online network systems ceased these accounts immediately. After this incident, America online network systems took measures to prevent this type of attack by verifying the authenticity of credit card and associated billing identity, which also enabled the attackers to change their way of obtaining America online network systems accounts.

Instead of creating a fake account, attackers would steal the personal information of registered America online network systems user. Attackers contacted registered America online network systems users through instant messenger or e-mail and asked them to verify the password for security purposes. E-mail and instant messages appeared to come from an America online network systems employee. Many users provided their passwords and other personal information to the attackers. The attackers then used the variously billed portions of America online website on behalf of a legitimate user. Moreover, an attacker no longer restricts themselves to masquerading America online website but actively masquerade a large number of financial and electronic commerce websites.

Description of the Project Work-

i. Introduction-

Phishing is a crime in which a perpetrator sends the fake e-mail, which appears to come from popular and trusted brand or organization, asking to input personal credential like bank password, username, phone number, address, credit card details, and so forth. The fake e-mails often look amazingly legitimate, and even the website where the Internet user is asked to input personal information also looks similar to legitimate one. Phishing messages propagate over e-mail, SMS, instant messengers, social networking sites, VoIP, and so forth, but e-mail is the popular way to perform this attack and 65% of the total phishing attack is achieved by visiting the hyperlink attached to the e-mail. Moreover, spear phishing attack is becoming popular nowadays. Business e-mail compromise is observed as a major Internet threat in 2015.

In BEC, the intruder uses spear phishing methods to fool organizations and Internet persons. More sophisticated spear phishing attacks targeted particular individual or groups within the organization. Phishing is metaphorically similar to fishing in the water, but instead of trying to catch a fish, attackers try to steal consumer's personal information. When a user opens a fake webpage and enters the username and protected password, the credentials of the user are acquired by the attacker which can be used for malicious purposes. Phishing websites look very similar in appearance to their corresponding legitimate websites to attract large number of Internet users. Recent developments in phishing detection have led to the growth of numerous new visual similarity based approaches. Visual similarity based approaches compare the visual appearance of the suspicious website to its corresponding legitimate website by using various parameters.

ii. Aims and Objectives-

According to Internet world stats, total numbers of Internet users worldwide are 2.97 billion in 2014; that is, more than 38% of the world population uses Internet. Hackers take advantage of the insecure Internet system and can fool unaware users to fall for phishing scams. Phishing e-mail is used to defraud both individuals and financial organizations on the Internet. The Anti-Phishing Working Group (APWG) is an international consortium which is dedicated to promoting research, education, and law enforcement to eliminate online fraud and cyber-crime. Our main aim is to reduce the possibility of an attack by preventing the user from clicking the link and detecting the attack from the URL itself.

Our main Objectives are as follows-

- Find out malicious URLs even before a user clicks on them so that a lot of time of the user is saved.
- Further reduce the risk from phishing be not allowing the user to click on the website if found malicious in nature.
- In addition, we provide several issues and challenges in detection of phishing attacks.

iii. Platform Used-

The entire project is coded in the Python3 Programming language.
MacOs is the operating system used.
Main python libraries used are-
Sci-kit learn, BeautifulSoup, tdlextract and pythonssl.
The dataset is downloaded from UCI machine learning repository. The dataset contains 31 columns, with 30 features and 1 target. The dataset has 2456 observations.

iv. Working-

First a python script will detect if the following features are present or not. The presence of some features indicate phishing while the absence of some features indicate phishing.

- Using the IP Address
- Long URL to Hide the Suspicious Part
- Using URL Shortening Services “TinyURL”
- URL’s having “@” Symbol
- Redirecting using “//”
- Adding Prefix or Suffix Separated by (-) to the Domain
- Sub Domain and Multi Sub Domains
- HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
- Domain Registration Length
- Favicon
- Using Non-Standard Port
- The Existence of “HTTPS” Token in the Domain Part of the URL

The following abnormalities in the URL are also detected-

- URL of Anchor
- Links in <Meta>, <Script> and <Link> tags
- Server Form Handler (SFH)
- Submitting Information to Email
- Abnormal URL
- Website Forwarding
- Status Bar Customization
- Disabling Right Click

After all the features are detected, a input vector is created that in which a 1 indicates that the feature is present and a 0 indicates that the feature is absent.

Model Training-

To fit the models over the dataset the dataset is split into training and testing sets. The split ratio is 75-25. Where in 75% accounts to training set.

Now the training set is used to train the classifier. The classifiers chosen are:

- Logistic Regression
- Random Forest Classification
- Support Vector Machine

We will see which one fits best in our dataset.

1.Logistic Regression

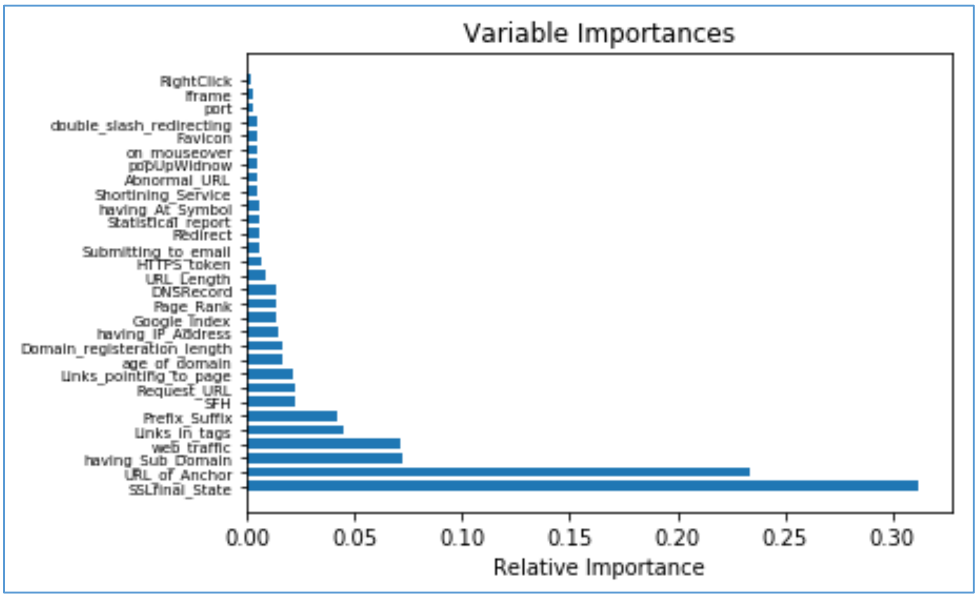
Fitting logistic regression and creating confusion matrix of predicted values and real values I was able to get 92.3 accuracy. Which was good for a logistic regression model.

2.Support Vector Machine

Support vector machine with a rbf kernel and using gridsearchcv to predict best parameters for svm was a really good choice, and fitting the model with predicted best parameters I was able to get 96.47 accuracy which is pretty good.

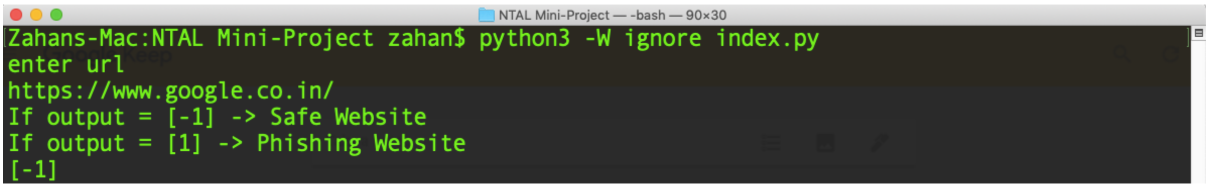
3.Random Forest Classification

Next model I wanted to try was random forest and I will also get features importances using it, again using gridsearchcv to get best parameters and fitting best parameters to it I got very good accuracy 97.26. Ultimately, Random forest was giving very good accuracy. We can also try artificial neural network to get a improved accuracy.

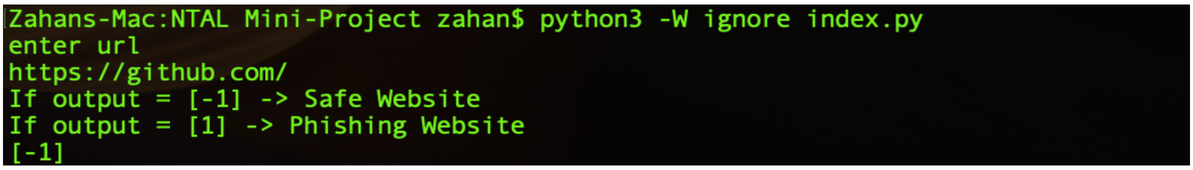


v. Screenshots-

Classifying Safe Examples-

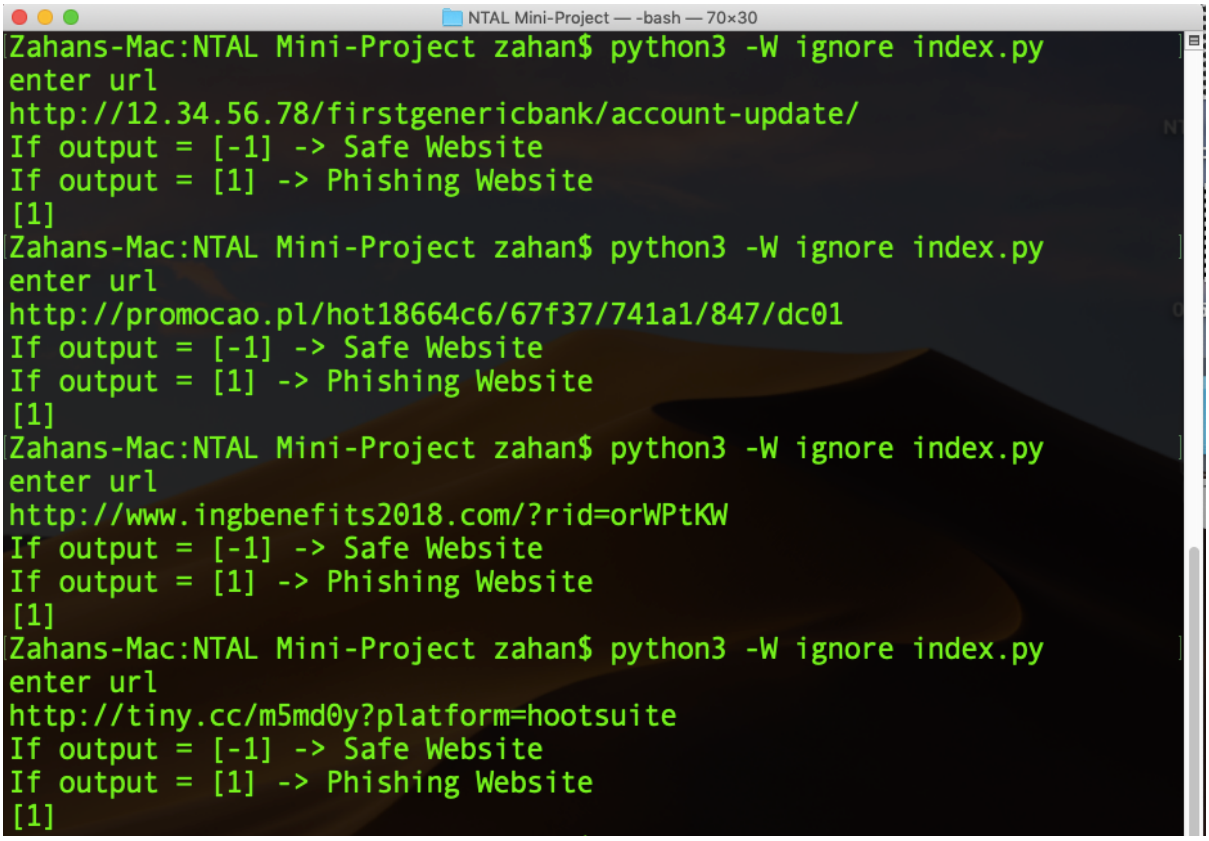


```
Zahans-Mac:NTAL Mini-Project zahan$ python3 -W ignore index.py
enter url
https://www.google.co.in/
If output = [-1] -> Safe Website
If output = [1] -> Phishing Website
[-1]
```



```
Zahans-Mac:NTAL Mini-Project zahan$ python3 -W ignore index.py
enter url
https://github.com/
If output = [-1] -> Safe Website
If output = [1] -> Phishing Website
[-1]
```

Phishing Website Classification-



```
Zahans-Mac:NTAL Mini-Project zahan$ python3 -W ignore index.py
enter url
http://12.34.56.78/firstgenericbank/account-update/
If output = [-1] -> Safe Website
If output = [1] -> Phishing Website
[1]
Zahans-Mac:NTAL Mini-Project zahan$ python3 -W ignore index.py
enter url
http://promocao.pl/hot18664c6/67f37/741a1/847/dc01
If output = [-1] -> Safe Website
If output = [1] -> Phishing Website
[1]
Zahans-Mac:NTAL Mini-Project zahan$ python3 -W ignore index.py
enter url
http://www.ingbenefits2018.com/?rid=orWPtKW
If output = [-1] -> Safe Website
If output = [1] -> Phishing Website
[1]
Zahans-Mac:NTAL Mini-Project zahan$ python3 -W ignore index.py
enter url
http://tiny.cc/m5md0y?platform=hootsuite
If output = [-1] -> Safe Website
If output = [1] -> Phishing Website
[1]
```

vi. Results-

We have achieved an accuracy of 94.37% in classifying whether the given URL is safe of unsafe.

vii. Conclusion-

Various types of anti-phishing techniques based on visual similarity approach have been given in the literature. However, still there is no single technique that can detect all types of phishing attacks (i.e., zero-hour phishing attack, embedded objects, DNS poisoning, etc.). Day by day phishing attack is increasing continuously and becomes the most popular e-crime. Consistently, when researchers design a new technique to control phishing attack, attackers change their way to perform attack or exploit the vulnerability in the solution. Hence, there is the tight race between attackers and anti-phishing developers.

References

1. M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: a literature survey,” IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 2091–2121, 2013.

2. R. Islam and J. Abawajy, “A multi-tier phishing detection and filtering approach,” Journal of Network and Computer Applications, vol. 36, no. 1, pp. 324–335, 2013.

3. A. K. Jain and B. B. Gupta, “Comparative analysis of features based machine learning approaches for phishing detection,” in Proceedings of the 10th INDIA-COM, New Delhi, India, 2016.

4. G. Weaver, A. Furr, and R. Norton, Deception of Phishing: Studying the Techniques of Social Engineering by Analyzing Modern-Day Phishing Attacks on Universities, 2016.

5. Kaspersky Lab, “Spam in January 2012 love, politics and sport,” 2013.

6. APWG Q1-Q3 Report, 2015

7. B. Parmar, “Protecting against spear-phishing,” Computer Fraud & Security, vol. 2012, no. 1, pp. 8–11, 2012.