Minor Project (COC3950) Report

on

Phishing Detection

Bachelor of Technology

IN

COMPUTER ENGINEERING

BY

Asif Iqbal
Enrolment number: GI6046

Kirtiratan Sambariya
Enrolment number: GI6089

Under the Guidance of

Mr Mohd Shoaib
Department Of Computer Engineering

Zakir Husain College of Engineering & Technology
Aligarh Muslim University
Aligarh (India)-202002

Nov 2019

Dated.....

Declaration

The work presented in the Minor Project (COC3950) entitled "Phishing Detection using Machine Learning" submitted to the Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University Aligarh, for the award of the degree of Bachelor of Technology in Computer Engineering, during the session 2019-20, is our original work. We have neither plagiarized nor submitted the same work for the award of any degree.

Date: _____

Place *Asif Iqbal*

_____

Kirtiratan Sambariya

Dated**……**

Certificate

This is to certify that the Minor Project (COC3950) Report entitled Phishing Detection, being submitted by Asif Iqbal and Kirtiratan Sambariya in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Engineering, during the session 2019-20, in the Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University Aligarh, is a record of candidates' own work carried out by them under my supervision and guidance.

Mr. Mohd Shoaib
Assistant Professor
Department of Computer Engineering
ZHCET, AMU, Aligarh

Table of Contents

Abstract

Phishing is the fraudulent use of electronic communications to deceive and take advantage of users. Phishing attacks attempt to gain sensitive, confidential information such as usernames, passwords, credit card information, network credentials, and more. This project is based on approach that can detect phishing site by crawling the webpage and extracting features from the webpage . We have collected information from various research papers and used these information to develop a few set of features. On applying different machine learning classifier models to train features and get the maximum accuracy of 91.20% using Random Forest.

Acknowledgements

First of all, we would like to express our sincere thanks and great gratitude to our teacher Mr. Mohd Shoaib, who gave us the golden opportunity to do this wonderful project on the topic Phishing Detection using Machine Learning, he also helped us in doing a lot of research and we learn so many new things. We are really thankful to him.
Secondly we would also like to thank our friends who helped us a lot in finalizing this project within the limited time frame.

Asif Iqbal Kirtiratan Sambariya Date: _____

List of Figures

List of Tables

Chapter 1 : Introduction

Phishing is a type of Internet fraud scam where the scammer sends email messages that appear to be from financial institutions or credit card companies that try to trick recipients into giving private information (i.e., username, password, account number, etc.).

1.1 Motivation

This project deals with the problem of phishing attacks as these attacks are threat to cyber security and these attacks are growing rapidly.

- Phishing attempts grew by 65% in 2017 [1].
- Nearly 1.5 million phishing sites are created each month [1].
- 92% of malware is delivered via email [2].
- In 2017, the average user received an average of 16 phishing emails per month [2].
- 95% of attacks on business networks are the result of successful spear phishing [3].
- The average cost of a phishing attack to a mid-sized company is $1.6 million [1].

1.2 Objectives and Scope

The objective of our project is to make such a model so that the user only needs to feed the URL and model predict whether the URL is legitimate (-1) or phishing (1).

1.2.1 Phishing Impact on Businesses and Prime Targets

- **76% of businesses** reported being a victim of a phishing attack in 2018 [4].
- **Global internet portals** were the most targeted business category in Q3 2018, with **32.27%** of all attacks [5].
- Banks were the second most targeted business category in Q3 2018, with **18%** of all attacks [5].
- **Payment systems** were the third most targeted business category in Q3 2018, with **10%** of all attacks [5].
- **IT companies** were the fourth most targeted business category in Q3 2018, with **7%** of all attacks [5].
- In Q3 2018, **SecureList** registered attacks against **131 universities in 16 countries** worldwide [5].
- **Guatemala** was the country with the highest percentage of users attacked in Q3 2018, with **19%** [5].

Chapter 2 : Related Work

2.1 Various Approaches and their results

The comparison of the proposed approach with the existing approaches is shown below in the table . This comparison is based on the accuracy, search engine independent solution and language independent solution. This search engine based technique believes that legitimate site appears in the top results of search engine. Although only popular sites appear in the top search results. Therefore, we have not considered search engine based feature. Our approach gives the accuracy of 91.20% and language independent.

Table 2-1 Various approaches [8]

Approach	Accuracy	Search Engine Independent	Language Independent	Fast Detection
Pan and Ding	84	Yes	No	No
Zhang 2007	95	No	No	No
Garera 2007	97.3	Yes	Yes	No
Aburous 2010	88.4	Yes	Yes	No
Whittaker	95.92	Yes	No	No
Xiang	95.8	No	No	No
He et al. 2011	96.5	No	No	No
Zhang 2017	97.5	Yes	Yes	No
JainGupta2019	98.42	Yes	Yes	Yes
Our approach	91.20	Yes	Yes	Yes

2.2 Various Types of Phishing Attacks

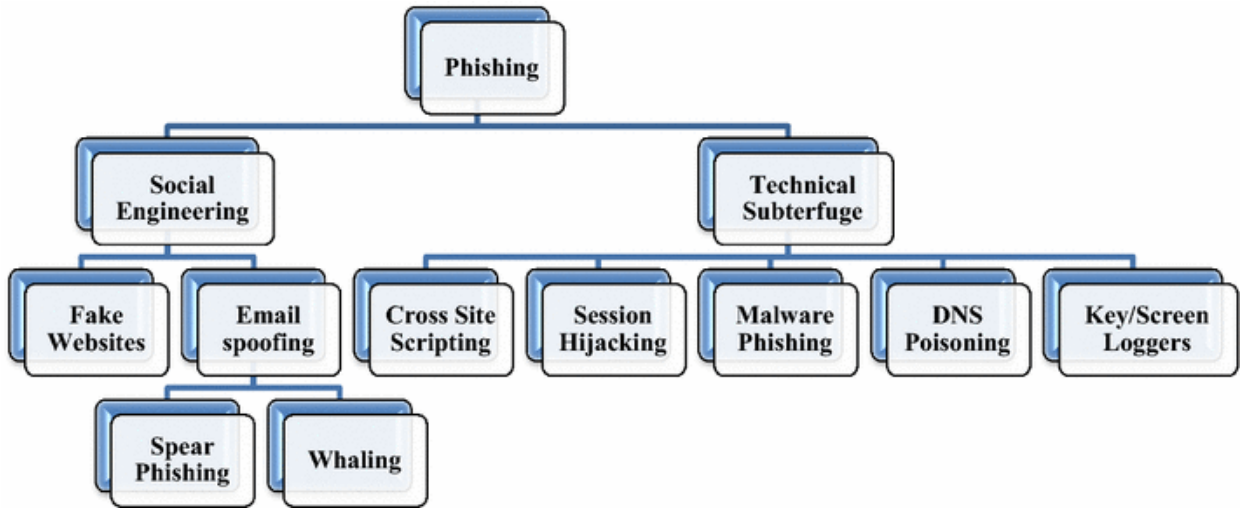


Fig 2.1 Types of Phishing Attacks [12]

Chapter 3 : Proposed Work

3.1 Workflow of Phishing Attack

Phishing attack is performed by taking the advantage of the close resemblance between the fake website and the legitimate ones. Phisher creates the fake websites and sends it to the thousands of users via email or some communication medium. Usually fake messages contains some fear of sense, urgency or offer some money that requires some urgent action like update the PIN etc. As the user updates/fills the information, the phisher gets the details. Phishing is not only done for getting information but sometimes it is done for delivering the malicious software like trojan, ransomware etc. The workflow of phishing process is shown in the figure below .

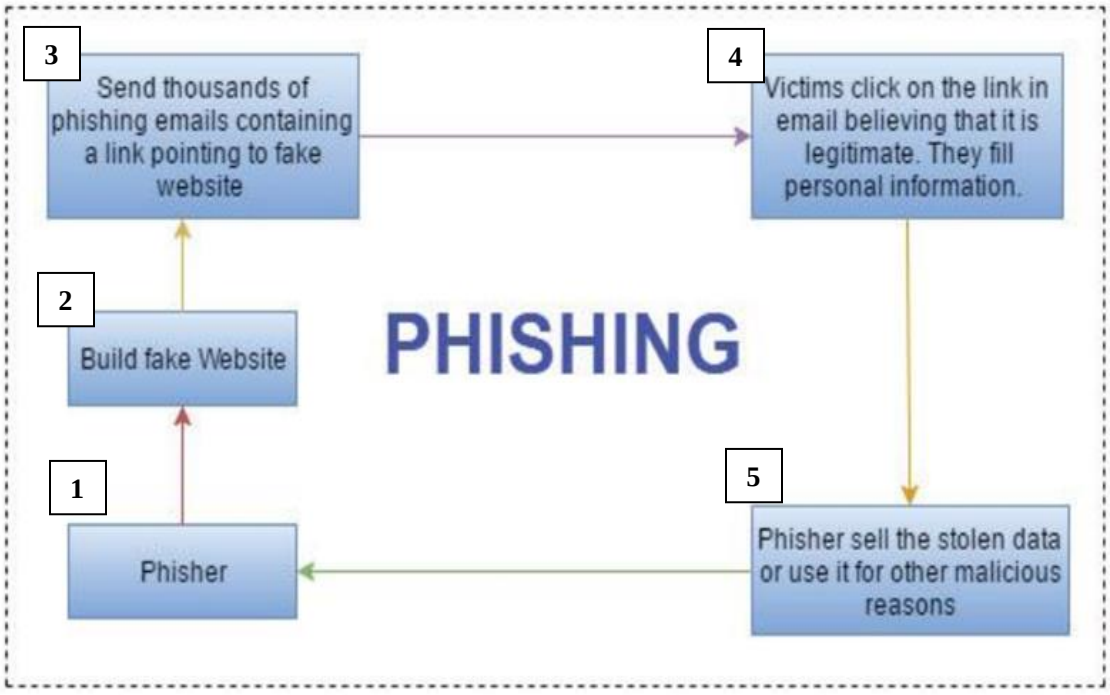


Fig 3.1 Flow of General Phishing Attack [8]

3.2 Methodology (Approach)

3.2.1 Collecting Data

The dataset for the proposed approached is taken from Kaggle [5]. This dataset contains 32 different features out of which we uses 16 best features to train and test the model. The features value are in the form of (-1) means legitimate, (0) means suspicious and (1) means phishing. The web crawler will crawl the input URL and extract those 16 features. The extracted features will feed in the trained model and the model will predict (1) for phishing and (-1) for legitimate. The workflow of the web crawler is shown in the figure 3.2, the procedure for training data is shown in figure 3.3, the project architecture is shown in figure 3.4 and the structure of dataset is shown in figure 3.5.

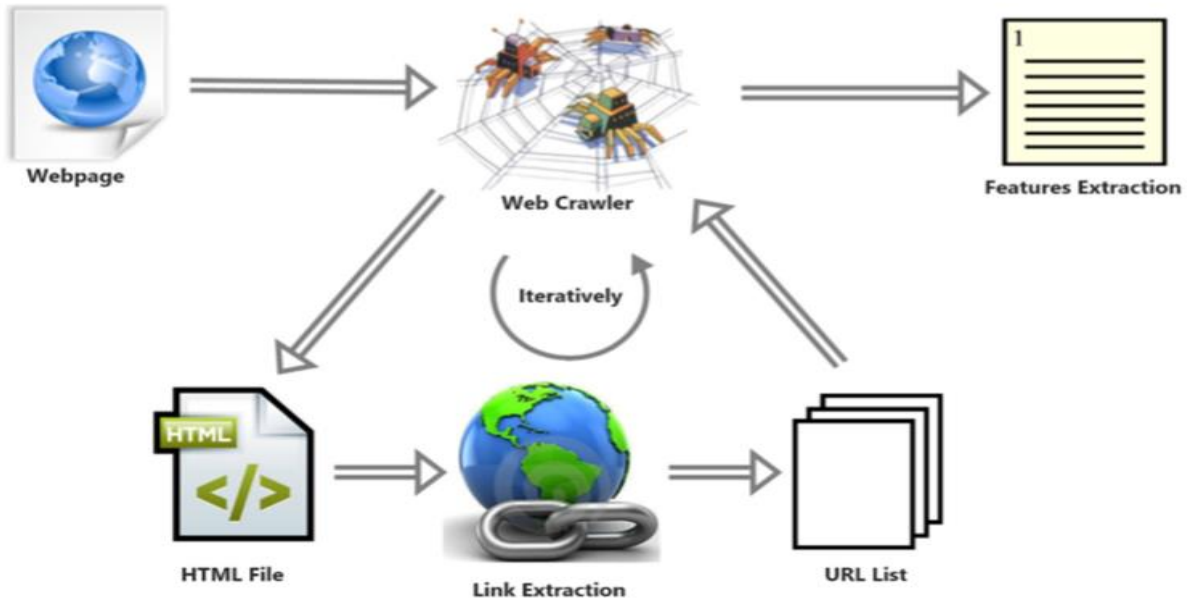


Fig3.2 Web Crawler and Feature Extraction [7]

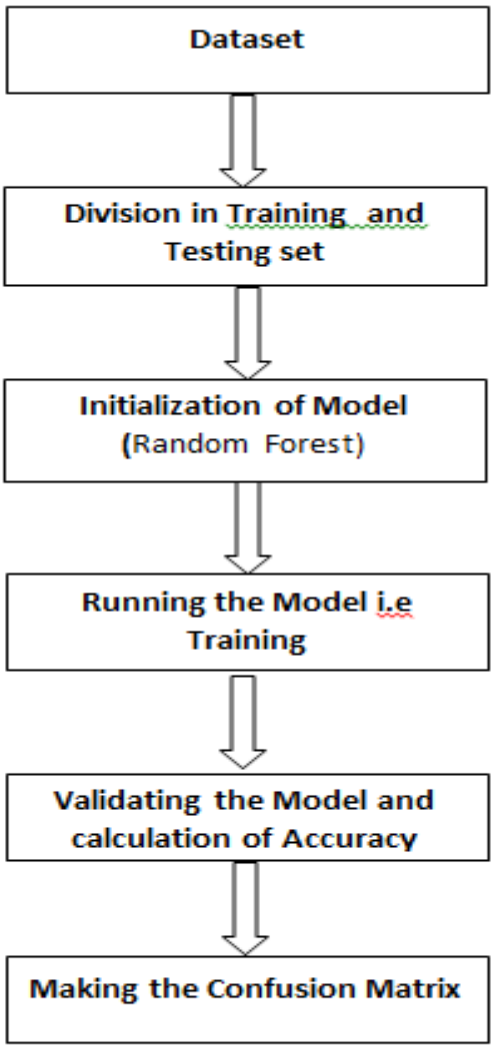


Fig 3.3 Procedure for Training Model

3.2.2 Features [6]

There were many features in the dataset but we select the best 16 features out of these and train the classifier model using those features. We use 80% of the dataset for the training purpose . The features that were used are explain below.

a. Using IP address

If an IP address is used as an alternative of the domain name in the URL, such "http://125.98.3.123/fake.html", users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link "http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html".

Rule: If The Domain Part has an IP Address → Phishing
Otherwise→ Legitimate

b. Long URL

Phishers can use long URL to hide the doubtful part in the address bar. For example:
http://federmacedoadv.com.br/3f/aze/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.website.html

To ensure accuracy of our study, we calculated the length of URLs in the dataset and produced an average URL length. The results showed that if the length of the URL is greater than or equal 54 characters then the URL classified as phishing. By reviewing our dataset we were able to find 1220 URLs lengths equals to 54 or more which constitute 48.8% of the total dataset size.

If $URL\ length$<54 Legitimate
$else\ if\ URL\ length \geq 54\ and \leq 75$ $Suspicious$
$otherwise$ Phishing

c. URL shortening service

URL shortening is a method on the "World Wide Web" in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an "HTTP Redirect" on a domain name that is short, which links to the webpage that has a long URL. For example, the URL "http://portal.hud.ac.uk/" can be shortened to "bit.ly/19DXSk4".
TinyURL → Phishing
Otherwise→ Legitimate

d. Using symbols

URL containing symbols like @, -, # to make the target person not to focus on the link.
e.g. www.face-book.com
www.am@zon.com
if URL contains symbol => Phishing
otherwise => legitimate

e. Redirection Using //

The existence of "//" within the URL path means that the user will be redirected to another website. An example of such URL's is: "http://www.legitimate.com//http://www.phishing.com". We examine the location where the "//" appears. We find that if the URL starts with "HTTP", that means the "//" should appear in the sixth position. However, if the URL employs "HTTPS" then the "//" should appear in seventh position.
The Position of the Last Occurrence of "//" in the URL > 7→ Phishing
Otherwise→ Legitimate

f. Domain and Multi Sub-domain

Let us assume we have the following link: http://www.hud.ac.uk/students/. A domain name might include the country-code top-level domains (ccTLD), which in our example is “uk”. The “ac” part is shorthand for “academic”, the combined “ac.uk” is called a second-level domain (SLD) and “hud” is the actual name of the domain. To produce a rule for extracting this feature, we firstly have to omit the (www.) from the URL which is in fact a sub domain in itself. Then, we have to remove the (ccTLD) if it exists. Finally, we count the remaining dots. If the number of dots is greater than one, then the URL is classified as “Suspicious” since it has one sub domain. However, if the dots are greater than two, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.
Dots In Domain Part=1 → Legitimate
Dots In Domain Part=2 → Suspicious
Otherwise→ Phishing

g. “HTTPS” Token in the Domain Part

The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-webapps-mpp-home.soft-hair.com/.
Using HTTP Token in Domain Part of The URL→ Phishing
Otherwise - Legitimate

h. Using Non-Standard Port

This feature is useful in validating if a particular service (e.g. HTTP) is up or down on a specific server. In the aim of controlling intrusions, it is much better to merely open ports that you need. Several firewalls, Proxy and Network Address Translation (NAT) servers will, by default, block all or most of the ports and only open the ones selected. If all ports are open, phishers can run almost any service they want and as a result, user information is threatened. The most important ports and their preferred status are shown in Table 3.1.

Table 3-1 Standard Ports..

PORT	Service	Meaning	Preferred Status
21	FTP	Transfer files from one host to another	Close
22	SSH	Secure File Transfer Protocol	Close
23	Telnet	provide a bidirectional interactive text-oriented communication	Close
80	HTTP	Hyper test transfer protocol	Open
443	HTTPS	Hypertext transfer protocol secured	Open
445	SMB	Providing shared access to files, printers, serial ports	Close
1433	MSSQL	Store and retrieve data as requested by other software applications	Close
1521	ORACLE	Access oracle database from web.	Close
3306	MySQL	Access MySQL database from web.	Close
3389	Remote Desktop	allow remote access and remote collaboration	Close

If Port # is of the Preferred Status→ Phishing Otherwise→ Legitimate

i. Request URL

Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

% of Request URL <22% → Legitimate
%of Request URL≥22% and 61%→ Suspicious
Otherwise→ feature=Phishing

j. URL of Anchor

An anchor is an element defined by the <a> tag. This feature is treated exactly as "Request URL". However, for this feature we examine:
1. If the <a> tags and the website have different domain names. This is similar to request URL feature.
2. If the anchor does not link to any webpage, e.g.:
A. <a href="#">
B. <a href="#content">
C. <a href="#skip">
D. <a href="JavaScript ::void(0)">

% of URL Of Anchor <31% → $Legitimate$
% of URL Of Anchor ≥31% And≤67% → Suspicious
Otherwise→ Phishing

k. Server From Handler

SFHs that contain an empty string or "about:blank" are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

SFH is "about: blank" Or Is Empty → Phishing
SFH Refers To A Different Domain→ Suspicious
Otherwise → Legitimate

l. Submitting Information to Email

Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user's information to his personal email. To that end, a server-side script language might be used such as "mail()" function in PHP. One more client-side function that might be used for this purpose is the "mailto:" function.

Using "mail()" or "mailto:" Function to Submit User Information→ Phishing
Otherwise → Legitimate

m. Website Forwarding

The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times.
Number of Redirect Page≤1 → Legitimate
Number of Redirect Page≥2 And<4→ Suspicious
Otherwise → Phishing

n. Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). By reviewing our dataset, we find that in worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as "Phishing". Otherwise, it is classified as "Suspicious".
Website Rank<100,000 → Legitimate
Website Rank>100,000 →Suspicious
Otherwise → Phishing

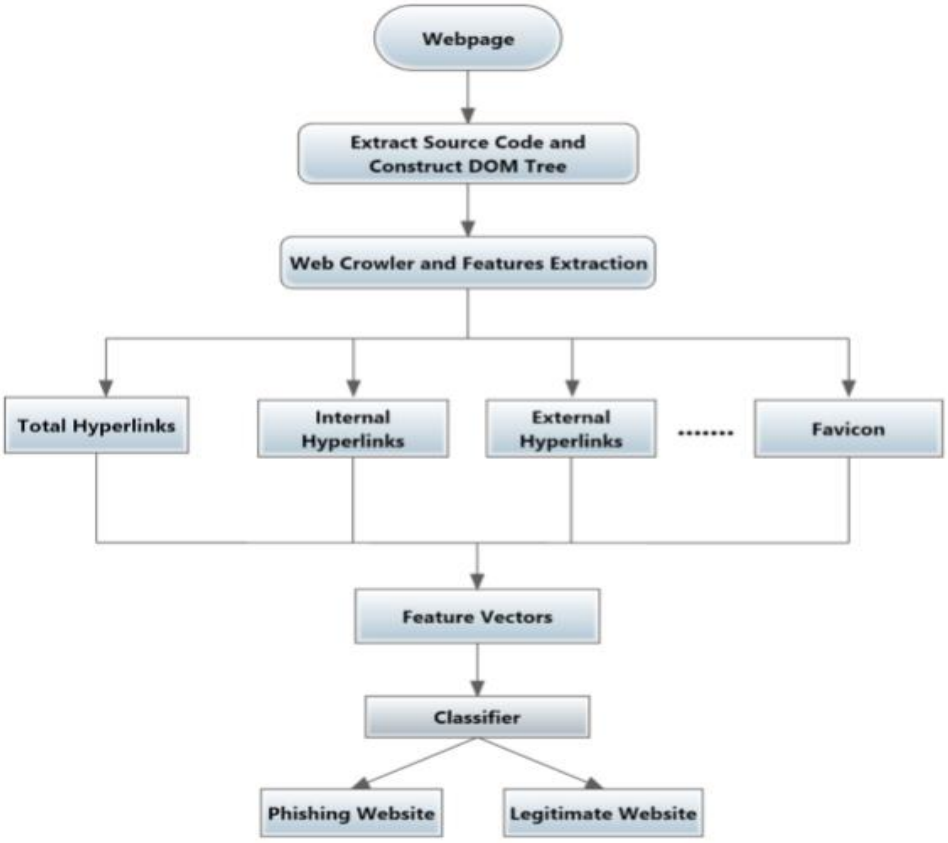
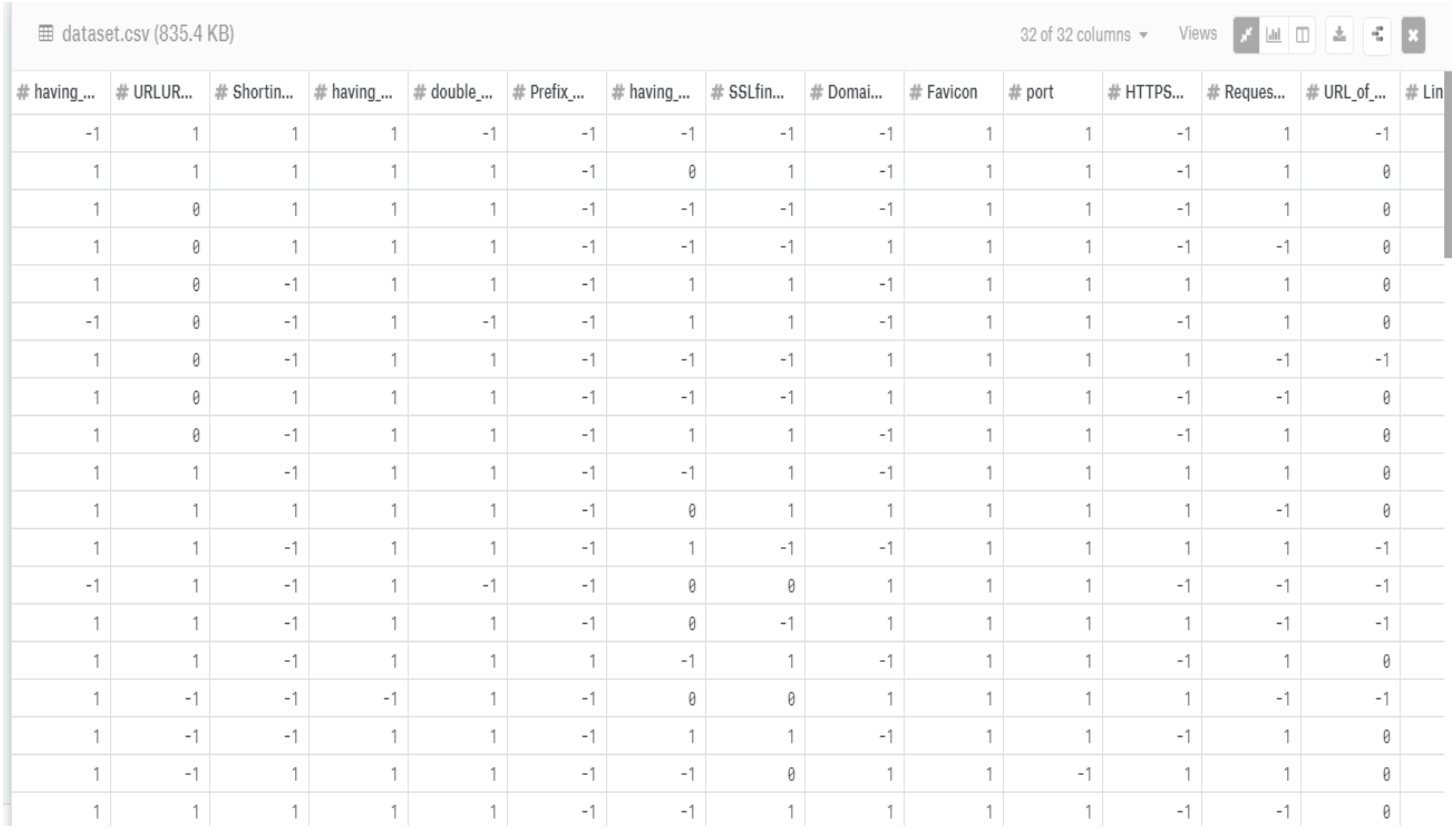


Fig 3.4 Our project Architecture [7]



dataset.csv (835.4 KB) — 32 of 32 columns — Views

# having_...	# URLUR...	# Shortin...	# having_...	# double_...	# Prefix_...	# having_...	# SSLfin...	# Domai...	# Favicon	# port	# HTTPS...	# Reques...	# URL_of_...	# Lin
-1	1	1	1	-1	-1	-1	-1	-1	1	1	-1	1	-1	
1	1	1	1	1	-1	0	1	-1	1	1	-1	1	0	
1	0	1	1	1	-1	-1	-1	-1	1	1	-1	1	0	
1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	
1	0	-1	1	1	-1	1	1	-1	1	1	1	1	0	
-1	0	-1	1	-1	-1	1	1	-1	1	1	-1	1	0	
1	0	-1	1	1	-1	-1	-1	1	1	1	1	-1	-1	
1	0	1	1	1	-1	-1	-1	1	1	1	-1	-1	0	
1	0	-1	1	1	-1	1	1	-1	1	1	-1	1	0	
1	1	-1	1	1	-1	-1	1	-1	1	1	1	1	0	
1	1	1	1	1	-1	0	1	1	1	1	1	-1	0	
1	1	-1	1	1	-1	1	-1	-1	1	1	1	1	-1	
-1	1	-1	1	-1	-1	0	0	1	1	1	-1	-1	-1	
1	1	-1	1	1	-1	0	-1	1	1	1	1	-1	-1	
1	1	-1	1	1	1	-1	1	-1	1	1	-1	1	0	
1	-1	-1	-1	1	-1	0	0	1	1	1	1	-1	-1	
1	-1	-1	1	1	-1	1	1	-1	1	1	-1	1	0	
1	-1	1	1	1	-1	-1	0	1	1	-1	1	1	0	
1	1	1	1	1	-1	-1	1	1	1	1	-1	-1	0	

Fig 3.5 Dataset [9]

3.2.3 Model Selection

On implementing these features on different classifying models(Logistic Regression, K-NN, SVM, K-SVM, Naïve Bayes, Decision Tree and Random Forest) on the training set, different range of accuracy where found. The maximum accuracy is found to be 91.20% on using the Random Forest classifier model. So, we use Random Forest Model for testing the input URL.

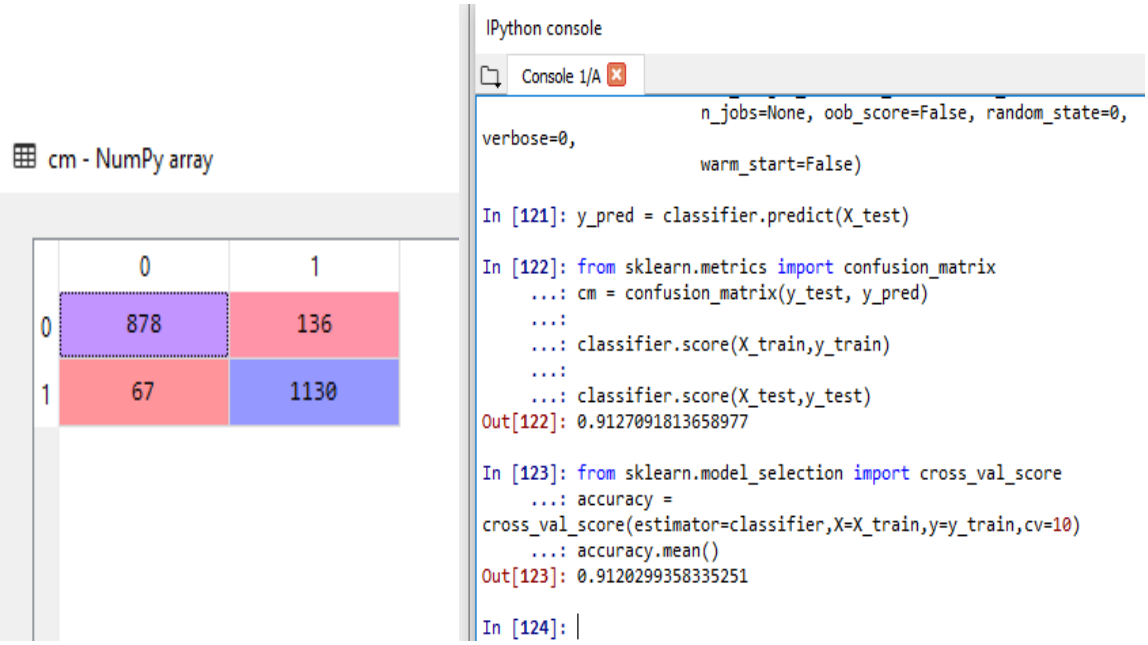
Table 3-2. Accuracy with different models

Models	**Accuracy(%)**
Logistic Regression	87.20
K-Nearest Neighbors	88.30
Support Vector Machines(SVM)	87.36
Kernel SVM	89.15
Naïve Bayes	57.55
Decision Tree	90.98
Random Forest	91.20

Chapter 4 : Results and Conclusions

4.1 Results

The accuracy found using Random Forest classifier is shown in figure 3.5(b) and the Confusion Matrix associated with it is shown in figure 3.5(a).



(a) Confusion Matrix (b) Accuracy

Fig 3.6

4.2 Conclusions and Future Work

The experimental results showed that proposed method is very efficient in classification of phishing websites as it has 91.20% overall accuracy.

The accuracy of our approach may be improved by adding certain more features. However, extracting other features from the third party will increase the running time complexity of the scheme.

Nowadays, Mobile devices are more popular and seem to be a perfect target for malicious attacks like mobile phishing. Therefore, detecting the phishing websites in the mobile environment is a challenge for further research and development. This model can be extended as a Chrome Extension so that phishing attack can be prevented in real time.

References

[1] DashLane, https://blog.dashlane.com/phishing-statistics/ .(last accessed on 01/09/2019)

[2] AlertLogic, https://blog.alertlogic.com/must-know-phishing-statistics-2018/.(last accessed on 01/09/2019)

[3] ExplainHowNow, https://www.explainhownow.com/2019/social-engineering/.(last accessed on 01/09/2019)

[4] WombatSecurity, https://www.wombatsecurity.com/state-of-the-phish.(last accessed on 01/09/2019)

[5] SecureList, https://securelist.com/spam-and-phishing-in-q3-2018/88686/(last accessed on 01/09/2019)

[6] Rami M Mohammad, Fadi Tabatah , Lee McCluskey – "Phishing Websites Features" http://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf . (last accessed on 08/09/2019)

[7] Ankit Kumar Jain, BB Gupta, "A machine learning based approach for phishing detection using hyperlinks information", Journal of Ambient Intelligence and Humanized Computing, vol.10, pages 2015–2028, 2019.

[8] Hemali Sampat1, Manisha Saharkar2, Ajay Pandey 3, Hezal Lopes4 - International Research Journal of Engineering and Technology (IRJET) -" Detection of Phishing Website Using Machine Learning" https://irjet.net/archives/V5/i3/IRJET-V5I3580.pdf. (last accessed on 22/09/2019)

[9] Dataset - https://www.kaggle.com/akashkr/phishing-website-dataset .(last accessed on 01/09/2019)

[10] Stackoverflow for solving raised errors – https://www.stackoverflow.com

[11] Udemy Course – Machine Learning A-Z in R and Python

[12] TypesOfPhishingAttack,https://www.google.com/search?rlz=1C1CHBD_enIN784IN784&tbm=isch&q=types+of+phishing+attacks&chips=q:types+of+phishing+attacks,online_chips:scientific+diagram&usg=AI4_-kQ1B1ST9J02xWZxTrlEt0Q-vOfyEw&sa=X&ved=0ahUKEwiRqbG76YDmAhXZwjgGHZbMBecQ4lYIKigA&biw=1366&bih=657&dpr=1#imgdii=hgicWbYHQzXuYM:&imgrc=Np8wmyHm6obMZM: