Institute of Engineering & Technology



**MINI PROJECT**

A REPORT ON

**DETECTING-PHISING-WEBSITE**

**Team members** :

AYUSH DWIVEDI

ANUPRIYA RAI

RISHABH TRIPATHI

**PROJECT GUIDE :**

VIVEK KUMAR SIR

Department of Computer Engineering &

Applications

Institute of Engineering & Technology



GLA University

Mathura-281406, INDIA

2019

**ABSTRACT**

Phishing costs Internet users billions of dollars per year. It refers to luring techniques used by identity thieves to fish for personal information in a pond of unsuspecting internet users. Phishers use spoofed e-mail, phishing software to steal personal information and financial account details such as usernames and passwords. This paper deals with methods for detecting phishing web sites by analyzing various features of benign and phishing URLs by Machine learning techniques. We discuss the methods used for detection of phishing websites based on lexical features, host properties and page importance properties. We consider various data mining algorithms for evaluation of the features in order to get a better understanding of the structure of URLs that spread phishing. The fine-tuned parameters are useful in selecting the apt machine learning algorithm for separating the phishing sites from benign sites.

1. **INTRODUCTION**
   1. **GENERAL INTRODUCTION TO THE TOPIC**

Phishing is a criminal mechanism employing both social engineering and technical tricks to steal consumers’ personal identity data and financial account credentials. Social engineering schemes use spoofed e-mails, purporting to be from legitimate business and agencies, designed to lead consumers to counterfeit websites that trick recipients into divulging financial data such as username and passwords. Technical subterfuge schemes install malicious software onto computers, to steal credentials directly, often using systems to intercept consumers’ online account user names and passwords.

The criminals, who want to obtain sensitive data, first create unauthorized replicas of a real website and e-mail, usually from a financial institution or another company that deals with financial information. The e-mail will be created using logos and slogans of a legitimate company. The nature and format of Hypertext Mark-up Language makes it very easy to copy images or even an entire website. While this ease of website creation is one of the reasons that the Internet has grown so rapidly as medium, it also permits the abuse of trademarks, trade names, and other corporate identifiers upon which consumers have come to rely as mechanisms for authentication. Phisher then send the "spoofed" e-mails to as many people as possible in an attempt to lure them in to the scheme. When these e-mails are opened or when a link in the mail is clicked, the consumers are redirected to a spoofed website, appearing to be from the legitimate entity.

The fraudulent e-mail stated that if the receiver did not click on the link and key in his client card number and pass code, access to his account would be blocked. These e-mails were sent within a week of a computer malfunction that prevented customer accounts from being updated. The United States continued to be the top country hosting phishing sites during the third quarter of 2012. This is mainly due to the fact that a large percentage of the world’s Web sites and domain names are hosted in the United States. Financial Services remains to be the most targeted industry sector by Phishers.

* 1. **AREA OF COMPUTERS**

**MACHINE LEARNING-**

Machine learning is an application of artificial intelligence (AI) that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

**PYTHON-**

Python is a general-purpose programming language. Hence, you can use the programming language for developing both desktop and web applications. Also, you can use Python for developing complex scientific and numeric applications. Python is designed with features to facilitate data analysis and visualization.

* 1. **HARDWARE AND SOFTWARE REQUIREMENTS**

**Hardware:**

Computer/Laptop

Processer: i5 or more

8GB RAM

Operating System

Internet

**Software:**

Jupyter Notebook

GitHub

PyCharm

Python

Google colo.

1. **Problem definition**

Phishing is a fraudulent attempt, usually made through email, to steal your personal information. We define a phishing page as any web page that, without permission, alleges to act on behalf of a third party with the intention of confusing viewers into performing an action with which the viewer would only trust a true agent of the third party.

Phishing is a type of computer attack that communicates socially engineered messages to humans via electronic communication channels in order to persuade them to perform certain actions for the attacker’s benefit.

**3-Objectives**

The primary objectives behind phishing attacks, from an attacker’s perspective, are:

Financial Gain: phishers can use stolen banking credentials to their financial benefits.

Identity Hiding: instead of using stolen identities directly, phishers might sell the identities to others whom might be criminals seeking ways to hide their identities and activities

Fame and Notoriety: phishers might attack victims for the sake of peer recognition

1. **Methodology**

(1) Detection Method

User Training Methods - end-users can be educated to better understand the nature of phishing attacks, phishing and non-phishing messages.

Software classification Methods — these mitigation methods aim at classifying phishing and legitimate messages on behalf of the user in an attempt to bridge the gap that is left due to the human error or ignorance.

(2) Offensive Defense method

Offensive defense solutions aim to render phishing cam pains useless for the attackers by disrupting the phishing campaigns.

This is often achieved by flooding phishing web- sites with fake credentials so that the attacker would have a difficult time to find the real credentials.

Example, Bogus Biter : A browser toolbar that submits fake information in HTML forms whenever a phishing website is encountered.

(3)Correction Method

Correction is the act of taking the phishing resources down. This is often achieved by reporting attacks to Service Providers. Phishing campaigns often rely on resources, such as:

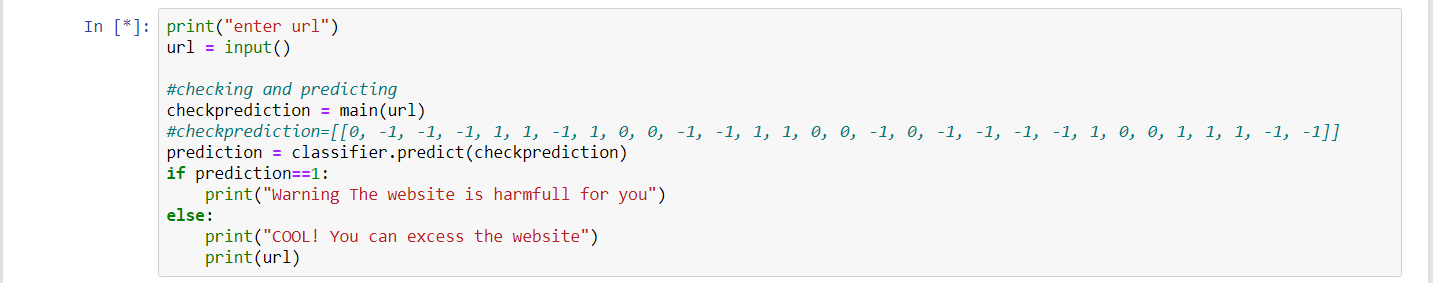
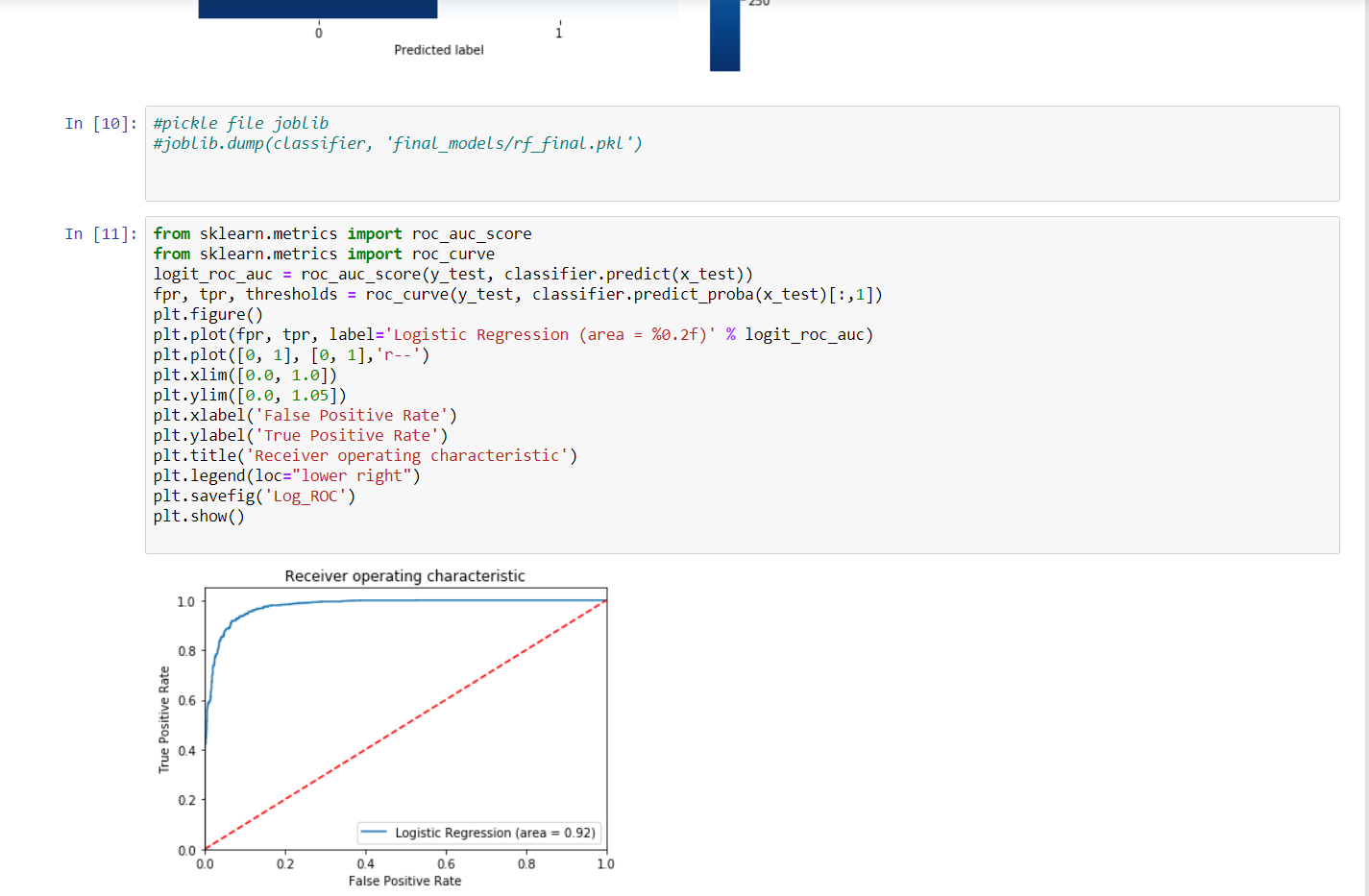
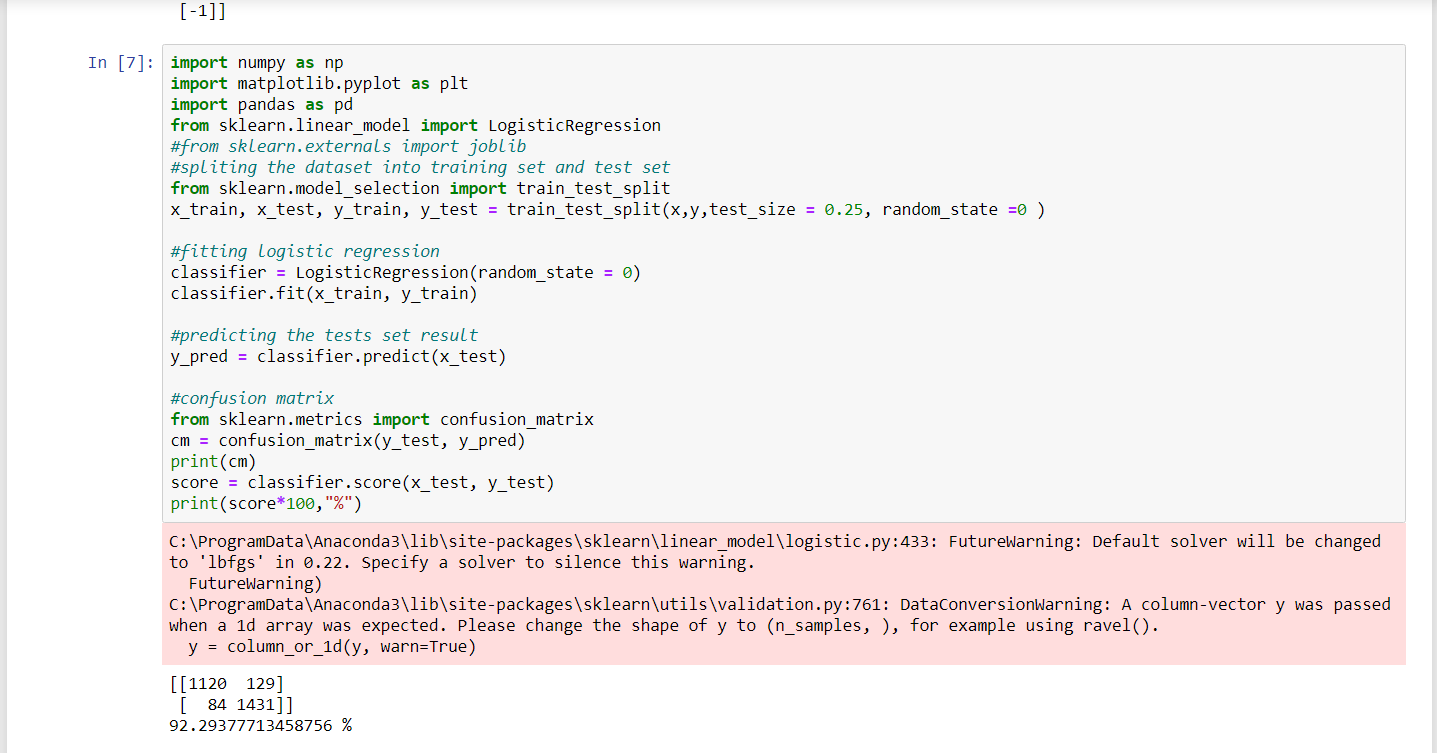
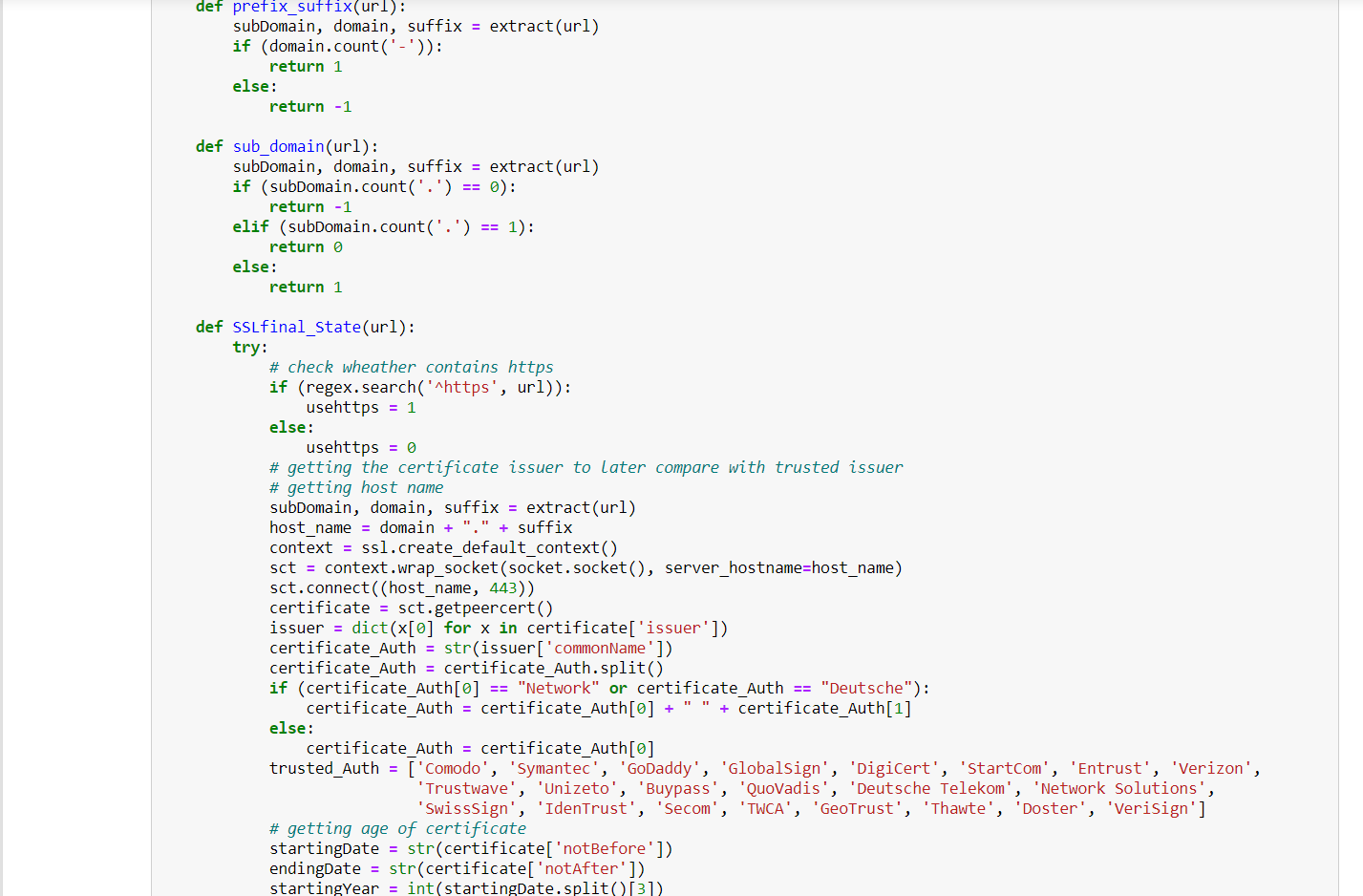
Websites, e-mail messages, Social Networking services

-In order to correct such attempts

• Removal of phishing content from websites, or suspension of hosting services.

• Suspension of email accounts, SMTP relays, VIP services.

• Trace back and shutdown of bot nets.

**5-Implementation details**

1. **Contribution summary**

|  |  |
| --- | --- |
| **NAME** | **CONTRIBUTION** |
| Ayush Dwivedi | Programmer |
| Rishabh Tripathi | GUI Developer |
| Anupriya Rai | Documentation |

1. **Progress till date & the Remaining work**

We have completed the programming part of the project in which we trained only logistic regression model and write the code to create the dataset for prediction we create diff. function to create the dataset like: -

* 1. **Links in <Meta>, <Script> and <Link> tags: -**

Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML document; <Script> tags to create a client side script; and <Link> tags to retrieve other web resources. It is expected that these tags are linked to the same domain of the webpage.

Rule:

IF {% of Links in " < Meta > ", " < Script > " and " < Link>" < 17% → Legitimate

% of Links in < Meta > ", " < Script > " and " < Link>" ≥ 17% And ≤ 81% → Suspicious Otherwise → Phishing}

**7.2 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)”>

Rule: IF {

% of URL Of Anchor < 31% → 𝐿𝑒𝑔𝑖𝑡𝑖𝑚𝑎𝑡𝑒

% of URL Of Anchor ≥ 31% And ≤ 67% → Suspicious

Otherwise → Phishing}

**7.3 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.

Rule: IF {

% of Request URL < 22% → Legitimate

%of Request URL ≥ 22% and 61% → Suspicious

Otherwise → feature = Phishing}

**7.4 Long URL to Hide the Suspicious Part**

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&amp;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.

**7.5 Using the IP Address**

If an IP address is used as an alternative of the domain name in the URL, such as 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 {

If the Domain Part has an IP Address → Phishing

Otherwise → Legitimate}

**Remaining work: -**

We have to make the gui interface of the project and trained 1 more model that is Random Forest to check and compare the accuracy of the dataset and our model right now Logestic regression have the accuracy of 92.7%.

1. **References**

**DATASET: -** <https://archive.ics.uci.edu/ml/machine-learning-databases/00327/>

**Other references are: -**

* github.com
* Manohar Swamynathan (auth.) - Mastering Machine Learning with Python in Six Steps\_ A Practical Implementation Guide to Predictive Data Analytics Using Python (2017, Après)
* Google collaborates.
* [www.kaggale.com](http://www.kaggale.com)