WEB PHISHING DETECTION

TEAM ID: PNT2022TMID16807

1.INTRODUCTION

1.1 Project Overview

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. It will lead to information disclosure and property damage. Large organizations may get trapped in different kinds of scams.

1.2 Purpose

Find a way to solve complex problems in a manner that is appropriate for the customer's situation. Utilize existing channels and media to gain more adoption of your solution. q Use the right triggers and messaging to refine your communication and marketing strategy. q Ensure we find the right fit between problem-behavior and your company's needs by resolving frequent annoyances, urgent problems and costly problems. q Understand the existing situation in order to improve it for your target group.

2. LITERATURE SURVEY

[1] Paper Name: 'Phishing Scams Cost American Businesses Half A Billion Dollars A Year'. Author Name : Dr. Gunikhan Sonowal Content:Phishing remains a basic security issue in cyberspace. In phishing, assailants steal sensitive information from victims by providing a fake site which looks like the visual clone of a legitimate site. Phishing shall be handled using various approaches. It is established that single filter methods would be insufficient to detect different categories of phishing attempts.

[2] Paper Name: Phish net: predictive blacklisting to detect phishing attacks. Author Name: Pawan Prakash, Manish Kumar Content: Phish Net is a predictive blacklisting scheme to detect phishing attacks. Traditional blacklist approaches (i.e., exact match with the blacklisted entries) are easy for attackers to evade. Instead, Phish Net uses five heuristics (i.e., top-level domains, IP address, directory structure, query string, brand name) to compute simple combinations of blacklisted sites to discover new phishing sites. Also, it proposes an approximate matching algorithm to determine whether a given URL is a phishing site or not. Phish Net consists of two major components, namely, component I: predicting malicious URLs and component II: approximate matching.

2.1 Existing problem

Phishing Detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating phishing attacks since registering new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database. Furthermore, page content inspection has been used by some strategies to overcome the false negative problems and complement

the vulnerabilities of the stale lists. Moreover, page content inspection algorithms each have different approach to phishing website detection with varying degrees of accuracy. Therefore, ensemble can be seen to be a better solution as it can combine the similarity in accuracy and different error-detection rate properties in selected algorithms.

2.2 References

[1] Dr. Gunikhan Sonowal: 'Phishing Scams Cost American Businesses Half A Billion Dollars A Year'. Forbes, 5 May 2017. Accessed Jan 2018.

[2] Pawan Prakash, Manish Kumar 'Phish net: predictive blacklisting to detect phishing attacks. SANS Institute, 2007. Accessed Jan 2018.

[3] Ramana Rao Kompella, and Minaxi Gupta. 'A machine learning based approach for phishing detection using hyperlinks information' vol.12, no.2, pp.1–27, 2007.

[4] Sahingoz et al 'Phishing websites detection using a novel multipurpose dataset and web technologies features'. vol.55, no.1, pp.74– 81, 2012.

[5] Jian Mao, Wenqian Tian, Pei Li, Tao Wei and Zhenkai Liang 'Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity'. IEEE Access (Volume: 5) 23 August 2017

[6] Saad Al-Ahmadi, Afrah Alotaibi and Omar Alsaleh 'PDGAN: Phishing Detection With Generative Adversarial Networks' IEEE Access (Volume: 10) 18 April 2022.

[7] Dhanalakshmi, R & Prabhu, C & Chellapan, C ' Detection of Phishing Websites and Secure Transactions Detection of Phishing Websites and Secure Transactions'. International Journal Communication & Network Security (IJCNS).

2.3 Problem Statement Definition

Common users who look for information on the web are unsafe on the internet who need a method to ensure the links they click are secure because scams are common and no one should become a victim of web phishing.

Problem Statement(PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	User	Browse the internet	I think my data is leaked	Some illegal websites are using normal website name for stealing my sensitive details such as credit card details, Username, Password, etc.	Unsafe and frustrated while using fake websites.
PS-2	Customer	Get a trusted security service for my organization myself.	I am not sure which is the right tool.	I've already used a service. But my data was stolen.	Confused/Frustrated /Third party intrusion.

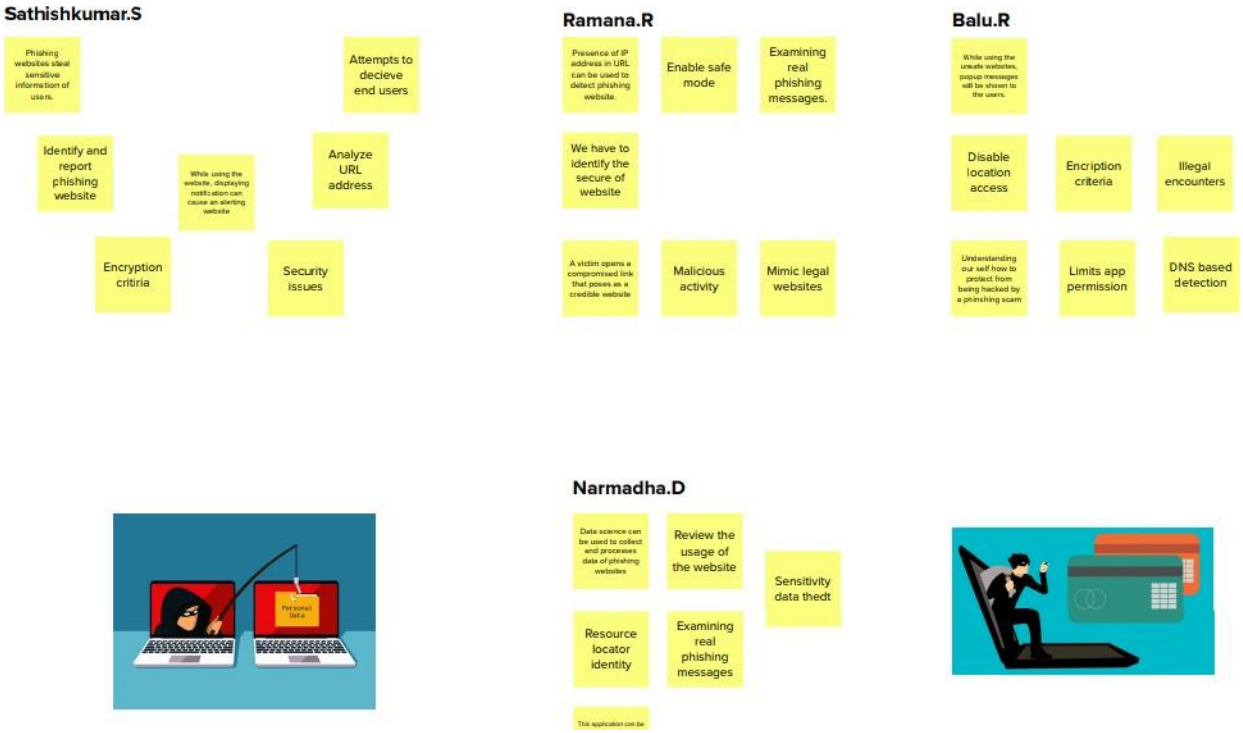
3. IDEATION &PROPOSED SOLUTION

In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and domain identity, and security and

encryption criteria in the final phishing detection rate. Once a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming



3.3 Proposed Solution

In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and domain identity, and security and encryption criteria in the final phishing detection rate. Once a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

3.4 Problem Solution Fit

Essentially, the problem solution fit means that you have found a problem with a customer and that the solution you have realized solves it for them. By Identifying behavioural patterns, it helps enterpreneurs, marketers and corporate innovators identify what and why.

1. CUSTOMER SEGMENT(S)

Users who buy products on-line and make payments via ebanking. Sensitive data will be shared through these kind of websites.

2. JOBS-TO-BE-DONE / PROBLEMS

- Websites with link that contain malware.
- Saying that they've noticed some suspicious activity or log-in attempts.
- Claim there is a problem with your account or your payment information.
- Want you to click on a link to make a payment, but the link has malware.

3. TRIGGERS

- Loss of money
- Loss of intellectual property.
- Damage to reputation
- Disruption of operational activities

4. EMOTIONS: BEFORE / AFTER

- BEFORE:
  - Stressed
  - Fear
  - Frustrated
  - Confused

- AFTER :
  - Confident
  - Safe Peace
  - Happy

5. AVAILABLE SOLUTIONS

- The above solutions check if the website is available in the legitimate websites list, but have property limitations such as exact name and adding items to the list frequently.
- Other ML model solution predictions are based on the content of the URL instead of its properties

6. CUSTOMER CONSTRAINTS

- Not being able to see the main process of the transaction site, they will not be able to know the real nature of the site.
- Sense of insecurity when faced with constraints.
- Not knowing how to protect them and identify malicious

7. BEHAVIOUR

- Users need to be more aware about what information they provide to the sites.
- They should not believe any site they visit even if they look the legitimate ones.

8. CHANNELS OF BEHAVIOUR

8.1 ONLINE

- Enter the URL and predicts the user

8.2 OFFLINE

  - Checks the site already available legitimate sites list.
  - Stores the phishing site to another list.

9. PROBLEM ROOT CAUSE

Attackers keep fooling people by spoofing original sites. They use their knowledge on the domain for cheating and other bad intentions. Common people will not have much knowledge on this domain.

10. YOUR SOLUTION

A deep learning-based framework by implementing it as a browser plug-in capable of determining whether there is a phishing risk in real-time when the user visits a web page and gives a warning message. The real-time prediction includes whitelist filtering, blacklist interception, and machine learning (ML) prediction. To deal with phishing attacks and distinguishing the phishing webpages automatically, Blacklist based detection technique keeps a list of websites' URLs that are categorized as phishing sites. If a web-page requested by a user exists in the formed list, the connection to the queried website is blocked. Machine Learning (ML) based approaches rely on classification algorithms such as Support Vector Machines (SVM) and Decision Trees (DT) to train a model that can later automatically classify the fraudulent websites at run-time without any human.

4. REQUIREMENT ANALYSIS

Requirements analysis, also called requirements engineering, is the process of determining user expectations for a new or modified product. These features, called requirements, must be quantifiable, relevant and detailed. In software engineering, such requirements are often called functional specifications. Requirements analysis is critical to the success or failure of a systems or software project. The requirements should be documented, actionable, measurable, testable, traceable, related to identified business

needs or opportunities, and defined to a level of detail sufficient for system design.

4.1 Functional requirement

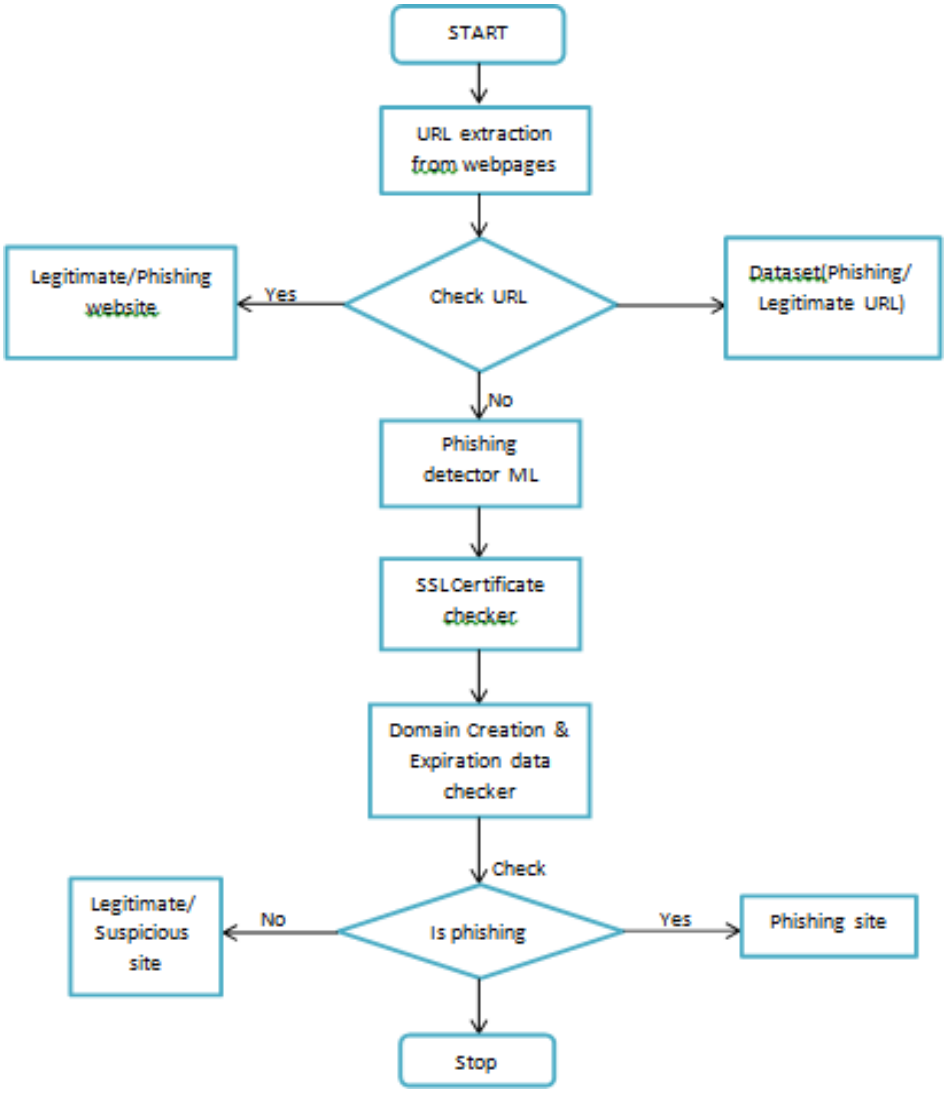
- User raw data
- Website Comparison
- Feature extraction
- Forecast
- Classifier
- Broadcasting
- Events

4.3 Non-Functional requirements

- Speed
- Security
- Portability
- Compatibility
- Capacity
- Reliability
- Environment

5.1 PROJECT DESIGN

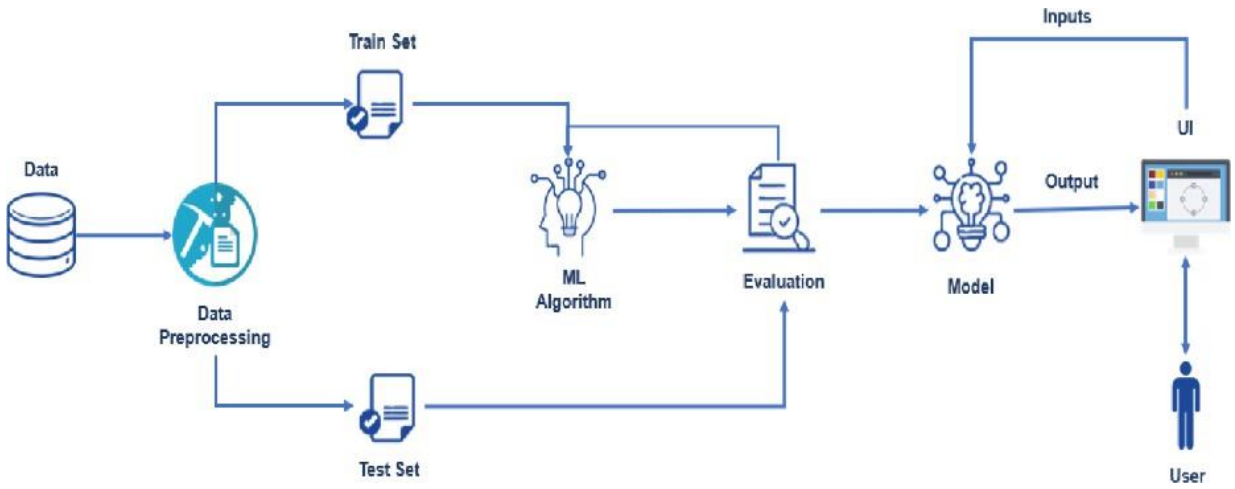
5.1 Data Flow Diagrams



`5.2 Solution & Technical Architecture

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

1. Find the best tech solution to solve existing business problems. Describe the structure, characteristics, behaviour, and other aspects of the software to project stakeholders. Define features, development phases, and solution requirements. Provide specifications according to which the solution is defined, managed, and delivered.



5.3 User Stories

I can look through the homepage's functional resources as a user. As a user, I can get knowledge of the various aspects of web phishing and become informed about scams. I can use the end page's resources to learn more about how it works as a user. As a user, I can

quickly guess the URL to determine whether a website is trustworthy or not As a user, I can quickly guess the URL to determine whether a website is trustworthy or not. As a user, I can provide feedback or contact the administrator for assistance. As administrators we can create user interfaces and maintain the functionality of the website. To make a website more user-friendly, we as administrators can reduce its complexity. As a administrator, you can utilize a variety of ML classifier models to do precise research for URL detection. We can respond to the user feedback for website enhancement as admins.

6. PROJECT PLANNING& SCHEDULING

6.1Sprint Planning& Estimation

Product Backlog, Sprint Schedule, and Estimation (4 Marks)

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Home page	USN-1	I can look through the homepage's functional resources as a user.	10	Low	Sathishkumar.S, Ramana.R
Sprint-1		USN-2	As a user, I can get knowledge of the various aspects of web phishing and become informed about scams.	5	High	Ramana.R, Balu.R
Sprint-2	Final page	USN-3	I can use the end page's resources to learn more about how it works as a user.	15	Low	Sathishkumar.S, Ramana.R, Narmadha.D
Sprint-3	Prediction	USN-4	As a user, I can quickly guess the URL to determine whether a website is trustworthy or not.	10	High	Sathishkumar.S, Ramana.R, Balu.R, Narmadha.D
	Dashboard					
Sprint-4	Chat	USN-5	As a user, I can provide feedback or contact the administrator for assistance.	10	High	Sathishkumar.S, Balu.R, Ramana.R
Sprint-1	Homepage	USN-6	As administrators we can create user interfaces and maintain the functionality of the website.	5	High	Balu.R, Narmadha.D
Sprint-2	Final page	USN-7	To make a website more user-friendly, we as administrators can reduce its complexity.	5	Medium	Sathishkumar.S, Ramana.R
Sprint-3	Prediction	USN-8	As a administrator, you can utilise a variety of ML classifier models to do precise research for URL detection.	10	High	Sathishkumar.S, Ramana.R, Balu.R, Narmadha.D
	Dashboard					
Sprint-4		USN-9	We can respond to the user feedback for website enhancement as admins.	10	Medium	Narmadha.D, Sathishkumar.S

6.2 Sprint DeliverySchedule

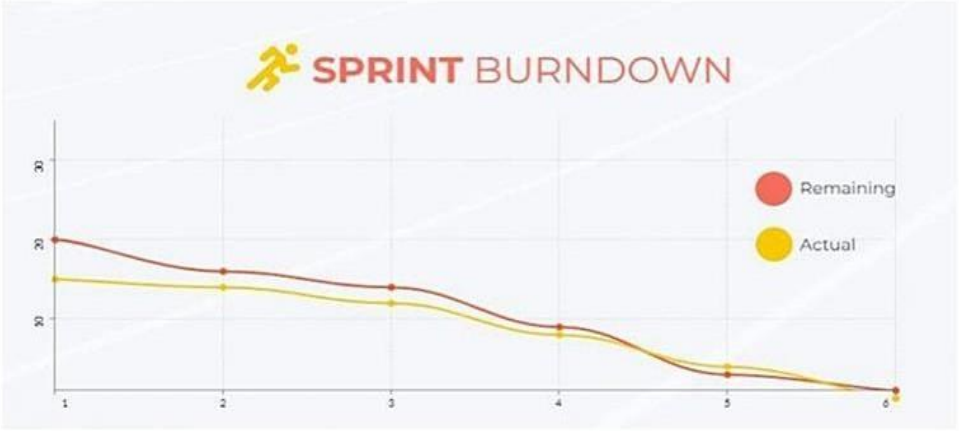
Project Tracker, Velocity & Burndown Chart(4 Marks)

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed(as on Planned End Date)	Sprint Release Date(Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	12 Nov 2022

6.3 Reports from JIRA

Burndown Chart:

A burn down chart is a graphical representation of work left to do versus time. It is often used in agile software development methodologies such as Scrum. However, burn down charts can be applied to any project containing measurable progress over time.



7. CODING &SOLUTIONING

7.1 Feature 1 - FLASK APP

The following is the flask app code and working

```
from flask import Flask, request, render_template
import numpy as np
import pandas as pd
from sklearn import metrics
import warnings
import pickle
warnings.filterwarnings('ignore')
from feature import FeatureExtraction

file = open("model.pkl","rb")
gbc = pickle.load(file)
file.close()


app = Flask(__name__)

@app.route("/", methods=["GET", "POST"])
def index():
 if request.method == "POST":

url = request.form["url"]
obj = FeatureExtraction(url)
x = np.array(obj.getFeaturesList()).reshape(1,30)

y_pred =gbc.predict(x)[0]
#1 is safe
#-1 is unsafe
y_pro_phishing = gbc.predict_proba(x)[0,0]
y_pro_non_phishing = gbc.predict_proba(x)[0,1]
# if(y_pred ==1 ):
pred = "It is {0:.2f} % safe to go ".format(y_pro_phishing*100)
return render_template('index.html',xx =round(y_pro_non_phishing,2),url=url )
return render_template("index.html", xx =-1)

if __name___== "_main_":
app.run(debug=True,port=2002)
```

7.2 Feature 2 - UI

The followingis the UI code for the application.

```
<!DOCTYPE html>
 <html lang="en">
        <head>
                <center>
```

```
            <h1> IBM Project Based Learning </h1>
            </center>
            <meta charset="UTF-8">
            <meta http-equiv="X-UA-Compatible" content="IE=edge">
            <meta name="viewport" content="width=device-width, initial-scale=1.0">
            <meta name="description" content="This website is develop for identify the
            safety of url.">
            <meta name="keywords" content="phishing url,phishing,cyber security,machine
            learning,classifier,python">
            <meta name="author" content="Balajee A V">

            <!-- BootStrap -->
            <link rel="stylesheet"
            href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css"
            integrity="sha384-
            9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1d
            KGj7Sk" crossorigin="anonymous">

            <link href="styles.css" rel="stylesheet">
            <title>URL detection</title>
  </head>

  <body>
   <center>
  <img class="image image-contain" src="https://cdn.activestate.com/wp-
  content/uploads/2021/02/phishing-detection-with-Python.jpg" alt="MDN logo" />
  </center>

  <div class=" container">
<div class="row">
  <div class="form col-md" id="form1">
                    <h2>PHISHING URL DETECTION</h2>
<br>
          <form action="/" method ="post">
                    <input type="text" class="form_input" name ='url' id="url"
                    placeholder="Enter URL" required="" />
                    <label for="url" class="form_label">URL</label>
                          <button class="button" role="button" >Check here</button>
          </form>

  </div>

  <div class="col-md" id="form2">
  <br>
  <h6 class = "right "><a href= {{ url }} target="_blank">{{ url }}</a></h6>
  <br>
  <h3 id="prediction"></h3>
  <button class="button2" id="button2" role="button" onclick="window.open('{{url}}')"
  target="_blank" >Still want to Continue</button>
  <button class="button1" id="button1" role="button"  onclick="window.open('{{url}}')"
  target="_blank">Continue</button>
```

```
        </div>
          </div>
          <br>
          </div>

          <!-- JavaScript -->
          <script src="https://code.jquery.com/jquery-3.5.1.slim.min.js"
          integrity="sha384-
          DfXdz2htPH0lsSSs5nCTpuj/zy4C+OGpamoFVy38MVBnE+IbbVYUew+OrCXaRkfj"
        crossorigin="anonymous"></script>
          <script src="https://cdn.jsdelivr.net/npm/popper.js@1.16.0/dist/umd/popper.min.js"
integrity="sha384-
Q6E9RHvbIyZFJoft+2mJbHaEWldlvI9IOYy5n3zV9zzTtmI3UksdQRVvoxMfooAo"
crossorigin="anonymous"></script>
<script src="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/js/bootstrap.min.js"
        integrity="sha384-
OgVRvuATP1z7JjHLkuOU7Xw704+h835Lr+6QL9UvYjZE3Ipu6Tp75j7Bh/kR0JKI"
crossorigin="anonymous"></script>


<script>
                let x = '{{xx}}';
                let num = x*100;
                if (0<=x && x<0.50){
        num = 100-num;
                }
                let txtx = num.toString();
                if(x<=1 && x>=0.50){
                var label = "Website is "+txtx +"% safe to use...";
                 document.getElementById("prediction").innerHTML = label;
                        document.getElementById("button1").style.display="block";
                }
         else if (0<=x && x<0.50){
                        var label = "Website is "+txtx +"% unsafe to use..."
                document.getElementById("prediction").innerHTML = label ;
                document.getElementById("button2").style.display="block";
                }
          </script>

          </body>
          <footer>
          <center> <p>© 2022 Sathish S</p> </center>
          </footer>
          </html>
```

8. TESTING

8.1 User Acceptance Testing

User Acceptance Testing (UAT) is a type of testing performed by the end user or the client to verify/accept the software system before moving the software application to the production environment. UAT is done in the final phase of testing after functional, integration and system testing is done. The User Acceptance of this product is not surveyed enough to give a solid conclusion. The theoretical and hypothetical acceptance is calculated to be high enough to conclude that this product is usable and valuable.

9. RESULTS

9.1 Performance Metrics

The Performance is the Accuracy of the model trained.The training accuracy of the model is 100%.The testing accuracy of the model is 100%.

10. ADVANTAGES & DISADVANTAGES

- ADVANTAGES
  - Measure the degrees of corporate and employee vulnerability.
  - Eliminate the cyber threat risk level.
  - Increase user alertness to phishing risks.
  - Install a cyber security culture and create cyber security heroes.
- DISADVANTAGES
  - Loss of money
  - Loss of intellectual property
  - Damage to reputational
  - Disruption of operational activities.

11. CONCLUSION

The importance to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective uses of phishing as an attacker's tool. Unfortunately, many of the existing phishing detection tools, especially those that depend on an existing blacklist, suffer limitations such as low detection accuracy and high false alarm that is often caused by either a delay in blacklist update as a result of human verification process involved in classification or perhaps, it can be attributed to human error in classification which may lead to improper classification of the classes. The inconsistent nature of attacks behaviours and continuously changing URL phish patterns require timely updating of the reference model. Therefore, it requires an effective technique to regulate retraining as to enable machine learning algorithm to actively adapt to the changes in phish patterns.Our phishing detection method focused on the learning process. The outcome of the experiment reached over 92% of Accuracy when websites with Logistic Regression are detected.

12. FUTURE SCOPE

In future we would like to enhance the existing model in such a way that consumer feels the same way and other upcoming technologies. Research to improve the accuracy of the system is under progress. Phishing attacks are targeting these users depending on the trikes of social engineering. Despite there are several ways to carry out these attacks, unfortunately the current phishing detection techniques cover some attack vectors like email and fake websites. Phishing is when attackers send malicious emails designed to trick people into falling for a scam. Typically, the intent is to get users to reveal financial information, system credentials or other sensitive data. Phishing has a list of negative effects on a business, including loss of money, loss of intellectual property, damage to reputation, and disruption of operational activities. These effects work together to cause loss of company value, sometimes with irreparable repercussions.

13. APPENDIX Source Code GitHub & Project Demo

Github: https://github.com/IBM-EPBL/IBM-Project-5224-1658751906

Demo: