Project Report

WEB PHISHING DETECTION

Submitted by:

Team ID: PNT2022TMID07412

1.VIJAYALAKSHMI K

2.SHAFEERA Z

3.SHARAN M

4.VIGNESH BUBU

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Project Overview:

Phishing is a fraudulent technique that is used over the Internet to deceive users with the goal of extracting their personal information such as username,passwords, credit card, and bank account information. The key to phishing is deception. Phishing uses email spoofing as its initial medium for deceptive communication followed by spoofed websites to obtain the needed information from the victims. Phishing was discovered in 1996, and today, it is one of the most severe cybercrimes faced by the Internet users. Researchers are working on the prevention, detection, and education of phishing attacks, but to date, there is no complete and accurate solution for thwarting them. This paper studies, analyzes, and classifies the most significant and novel strategies proposed in the area of phished website detection, and outlines their advantages and drawbacks. Furthermore, a detailed analysis of the latest schemes proposed by researchers in various subcategories is provided. The paper identifies advantages, drawbacks, and research gaps in the area of phishing website detection that can be worked upon in future research and developments. The analysis given in this paper will help academia and industries to identify the best anti-phishing technique.

1.2 Purpose:

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. This paper mainly focuses on applying a deep learning framework to detect phishing websites. This paper first designs two types of features for web phishing: original features and interaction features. A detection model based on Deep Belief Networks (DBN) is then presented. The test using real IP flows from ISP (Internet Service Provider) shows that the detecting model based on DBN can achieve an approximately 90% true positive rate and 0.6% false positive rate.

2. LITERATURE SURVEY

2.1 Existing problem:

Web service is a communication protocol and software between two electronic devices over the Internet [1]. Web services extends the World Wide web infrastructure to provide the methods for an electronic device to connect to other electronic devices [2]. Web services are built on top of open communication protocols such as TCP/IP, HTTP, Java, HTML, and XML. Web service is one of the greatest inventions of mankind so far, and it is also the most profound manifestation of computer influence on human beings [3].With the rapid development of the Internet and the increasing popularity of electronic payment in web service, Internet fraud and web security have gradually been the main concern of the public [4]. Web Phishing is a way of such fraud, which uses social engineering technique through short messages, emails, and WeChat [5] to induce users to visit fake websites to get sensitive information like their private account, token for payment, credit card information, and so on. The first phishing attack on AOL (America Online) can be traced back to early 1995 [6]. A phisher successfully obtained AOL users personal information. It may lead to not only the abuse of credit card information, but also an attack on the online payment system entirely feasible.

2.2 References:

1. https://en.wikipedia.org/wiki/Web_service.

2.O. Adam, Y. C. Lee, and A. Y. Zomaya, “Stochastic resource provisioning for containerized multi-tier web services in clouds,” *IEEE Transactions on Parallel and*

Distributed Systems, vol. 28, no. 7, pp. 2060–2073, 2017.
View at: Publisher Site | Google Scholar

3.T. Bujlow, V. Carela-Espanol, J. Sole-Pareta, and P. Barlet-Ros, “A survey on web tracking: Mechanisms, implications, and defenses,” *Proceedings of the IEEE*, vol. 105, no. 8, pp. 1476–1510, 2017.
View at: Publisher Site | Google Scholar

4. H.-C. Huang, Z.-K. Zhang, H.-W. Cheng, and S. W. Shieh, “Web application security: Threats, countermeasures, and pitfalls,” *The Computer Journal*, vol. 50, no. 6, pp. 81–85, 2017.
View at: Publisher Site | Google Scholar

5. https://en.wikipedia.org/wiki/WeChat.
K. Rekouche, *Early phishing*, 2011.

6.http://www.antiphishing.org/.
Microsoft, “20% Indians are victims of online phishing attacks: Microsoft,” *IANS*, 2014

7.http://news.biharprabha.com/.
View at: Google Scholar

8. L. Wu, X. Du, and J. Wu, “Effective defense schemes for phishing attacks on mobile computing platforms,” *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6678–6691, 2016.
View at: Publisher Site | Google Scholar

9. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “PhishNet: Predictive blacklisting to detect phishing attacks,” in *Proceedings of the 2017 IEEE Conference on Computer Communications (IEEE INFOCOM 2010)*, San Diego, USA, March 2010.
View at: Google Scholar

2.3 Problem Statement Definition:

Phishing is a major problem, which uses both social engineering and technical deception to get users' important information such as financial data, emails, and other private information. Phishing exploits human vulnerabilities; therefore, most protection protocols cannot prevent the whole phishing attacks.A measurement for phishing detection is the number of suspicious e-mails reported to the security team. This measurement is designed to evaluate the number of employees who followed the proper procedure for reporting suspicious messages. The only problem is that the fake email is directing you to a fake site, where the information you enter will be used to commit identity theft, fraud and other crimes.Phishing works by sending messages that look like they are from a legitimate company or website. The message will usually contain a link that takes the user to a fake website that looks like the real thing. The user is then asked to enter personal information, such as their credit card number.

Three common characteristics of a phishing website:

- Uses genuine-looking images.
- Uses authentic logos from a well-known company.
- Attempt to collect personal or financial information.

3. IDEATION & PROPOSED SOLUTION:

3.1 Empathy Map Canvas:

If you haven’t yet defined your ideal customer, you can start by using a Customer Persona template. It’s important to have basic information about your ideal customer in order to complete a Customer Empathy Map. You could also use a customer journey map in order to improve customer experience across various touchpoints.

1. While you can complete it alone, you'll get much deeper insights if you can do it as a collaborative activity. That way everyone involved will benefit from understanding your customer better. Book a meeting slot (we recommend an hour), then invite team members from different departments by sending them a link to the board.
2. Go through each of the six sections and encourage everyone to add at least one sticky note in each quadrant based on their customer knowledge or experience.
3. Return to each section and delve deeper or consolidate thoughts until you have a succinct collection of customer traits in each section.
4. Once you've finished, you can share the completed board by downloading it as an image or a PDF.

A Customer Empathy Map is a tool used when collecting data about customers to better understand your target customer base. They allow you to visualize customer needs, condense customer data into a clear, simple chart, and help you see what customers want — not what you think they want. By following this map, you can systematically find answers, without playing a guessing game.

When we look at empathy from a marketing perspective, we're talking about putting ourselves into our customers shoes, to be able to understand their needs and wants better. And thus, deliver a product or service that not only meets but exceeds their expectations!

There are six key steps in a Customer Empathy Map that will allow you to collect important information about your ideal customer to be able to really understand them. The six different components you'll consider are:

1. What the customer thinks and feels
2. What the customer hears
3. What the customer sees
4. What the customer says and does
5. The customer's pains
6. The customer's gains

So let's jump right in, and find out more about the Customer Empathy Map.

3.2 Ideation & Brainstorming:

In the Ideation stage, design thinkers spark off ideas — in the form of questions and solutions — through creative and curious activities such as Brainstorms and Worst Possible Idea. In this article, we'll introduce you to some of the best Ideation methods and guidelines that help facilitate successful Ideation sessions and encourage active participation from members.

When facilitated in a successful way, Ideation is an exciting process. The goal is to generate a large number of ideas — ideas that potentially inspire newer, better ideas — that the team can then cut down into the best, most practical and innovative ones.

"Ideation is the mode of the design process in which you concentrate on idea generation. Mentally it represents a process of "going wide" in terms of concepts and outcomes. Ideation provides both the fuel and also the source material for building prototypes and getting innovative solutions into the hands of your users."
– d.school, An Introduction to Design Thinking PROCESS GUIDE

The main aim of the Ideation stage is to use creativity and innovation in order to develop solutions. By expanding the solution space, the design team will be able to look beyond the usual methods of solving problems in order to find better, more elegant, and satisfying solutions to problems that affect a user's experience of a product.

3.3 Proposed Solution:

List-based phishing detection methods use either whitelist or blacklist-based technique. A blacklist contains a list of suspicious domains, URLs, and IP addresses, which are used to validate if a URL is fraudulent.

- Small Business Guide: Cyber Security.
- Step 1 - Backing up your data.
- Step 2 - Protecting your organisation from malware.
- Step 3 - Keeping your smartphones (and tablets) safe.
- Step 4 - Using passwords to protect your data.

- Step 5 - Avoiding phishing attacks.
- Actions to take.
- Resources.

Microsoft Exchange Online Protection (EOP) offers enterprise-class reliability and protection against spam and malware, while maintaining access to email during and after emergencies.

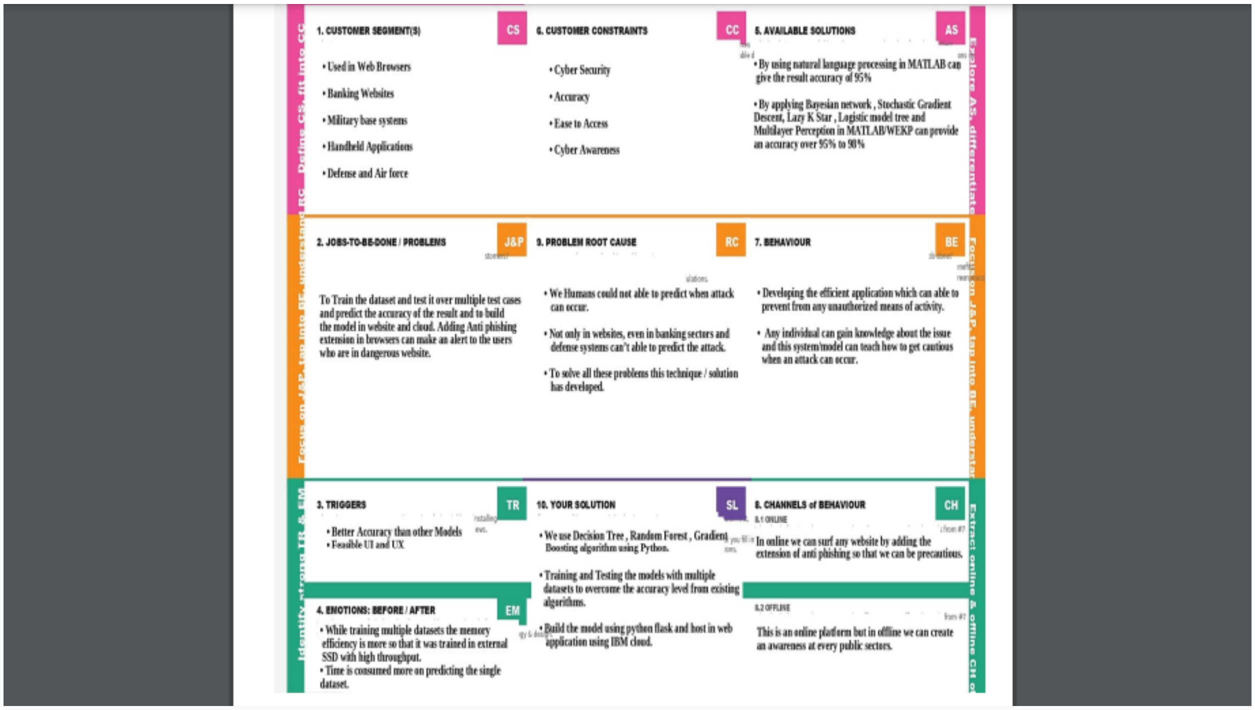
Google Chrome

Results evaluation showed that **Google Chrome** provides best security against phishing websites.

Four Ways To Protect Yourself From Phishing

- Protect your computer by using security software. ...
- Protect your cell phone by setting software to update automatically. ...
- Protect your accounts by using multi-factor authentication. ...
- Protect your data by backing it up.

3.4 Problem Solution fit:



CS
CC
AS
J&P
RC
BE
TR
SL
CH
EM
Define CS, fit into CC
Focus on J&P, tap into BE, understand RC
Identify strong TR & EM
Explore AS, differentiate
Focus on J&P, tap into BE, understand
Extract online & offline CH of

1. CUSTOMER SEGMENT(S)

- Used in Web Browsers
- Banking Websites
- Military base systems
- Handheld Applications
- Defense and Air force

6. CUSTOMER CONSTRAINTS

- Cyber Security
- Accuracy
- Ease to Access
- Cyber Awareness

5. AVAILABLE SOLUTIONS

- By using natural language processing in MATLAB can give the result accuracy of 95%
- By applying Bayesian network , Stochastic Gradient Descent, Lazy K Star , Logistic model tree and Multilayer Perception in MATLAB/WEKP can provide an accuracy over 95% to 98%

2. JOBS-TO-BE-DONE / PROBLEMS

To Train the dataset and test it over multiple test cases and predict the accuracy of the result and to build the model in website and cloud. Adding Anti phishing extension in browsers can make an alert to the users who are in dangerous website.

9. PROBLEM ROOT CAUSE

- We Humans could not able to predict when attack can occur.
- Not only in websites, even in banking sectors and defense systems can't able to predict the attack.
- To solve all these problems this technique / solution has developed.

7. BEHAVIOUR

- Developing the efficient application which can able to prevent from any unauthorized means of activity.
- Any individual can gain knowledge about the issue and this system/model can teach how to get cautious when an attack can occur.

3. TRIGGERS

- Better Accuracy than other Models
- Feasible UI and UX

4. EMOTIONS: BEFORE / AFTER

- While training multiple datasets the memory efficiency is more so that it was trained in external SSD with high throughput.
- Time is consumed more on predicting the single dataset.

10. YOUR SOLUTION

- We use Decision Tree , Random Forest , Gradient Boosting algorithm using Python.
- Training and Testing the models with multiple datasets to overcome the accuracy level from existing algorithms.
- Build the model using python flask and host in web application using IBM cloud.

8. CHANNELS of BEHAVIOUR

8.1 ONLINE

In online we can surf any website by adding the extension of anti phishing so that we can be precautious.

8.2 OFFLINE

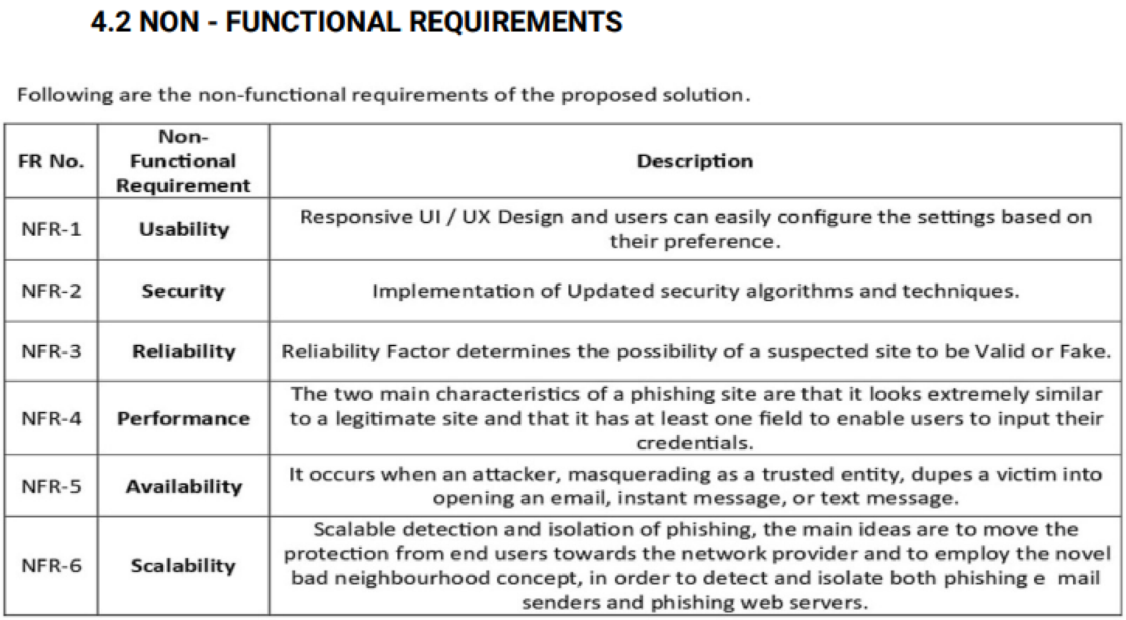
This is an online platform but in offline we can create an awareness at every public sectors.

4.REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration through LinkedIN
FR-2	User Confirmation	Confirmation via Email Confirmation via OTP
FR-3	User Authentication	Confirmation of Google Firebase
FR-4	User Security	Strong Passwords , 2FA and FIDO2.0 Webaucn
FR-5	User Performance	Usage of Legitimate websites, Optimize Network Traffic

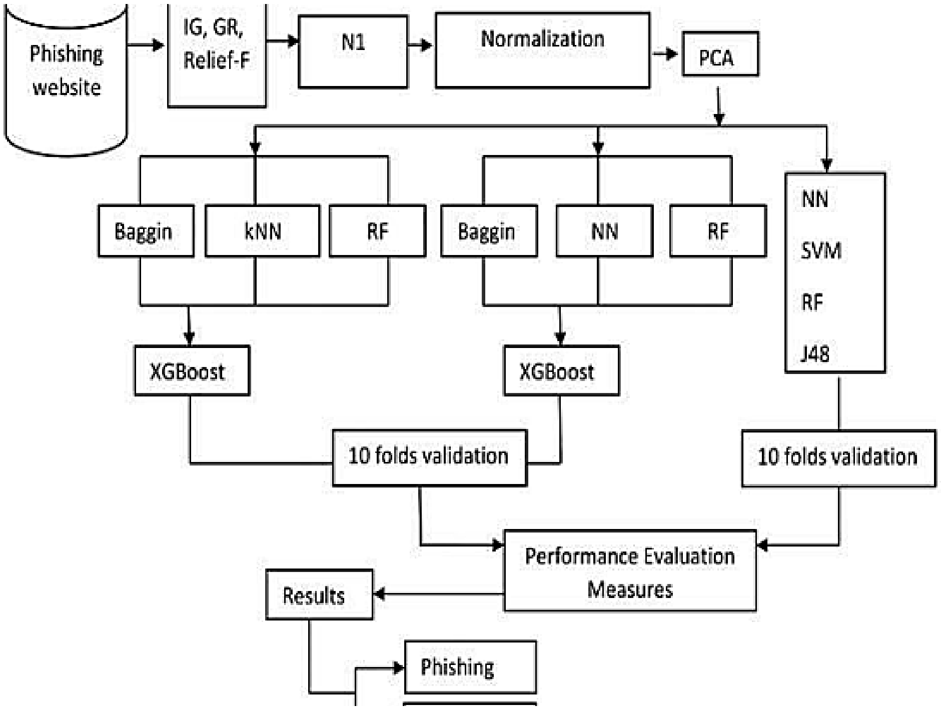


4.2 NON - FUNCTIONAL REQUIREMENTS

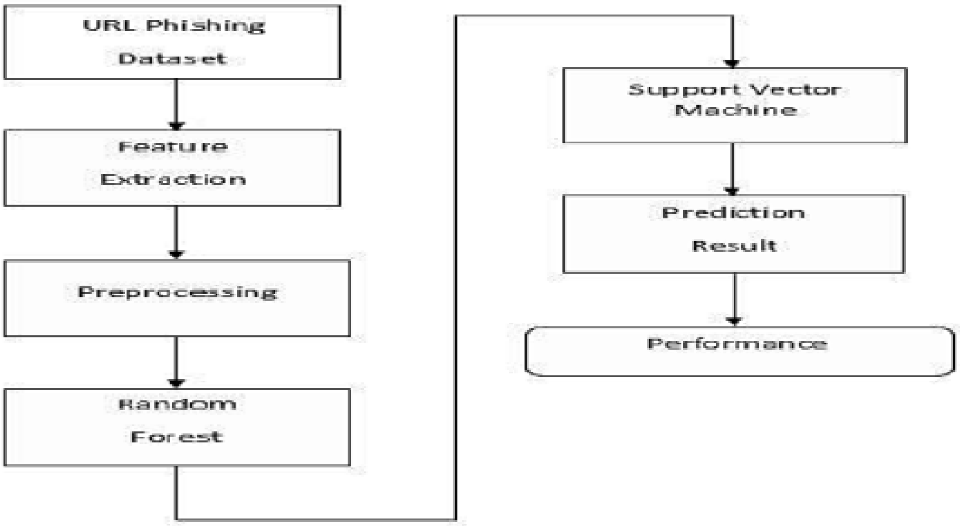
Following are the non-functional requirements of the proposed solution.

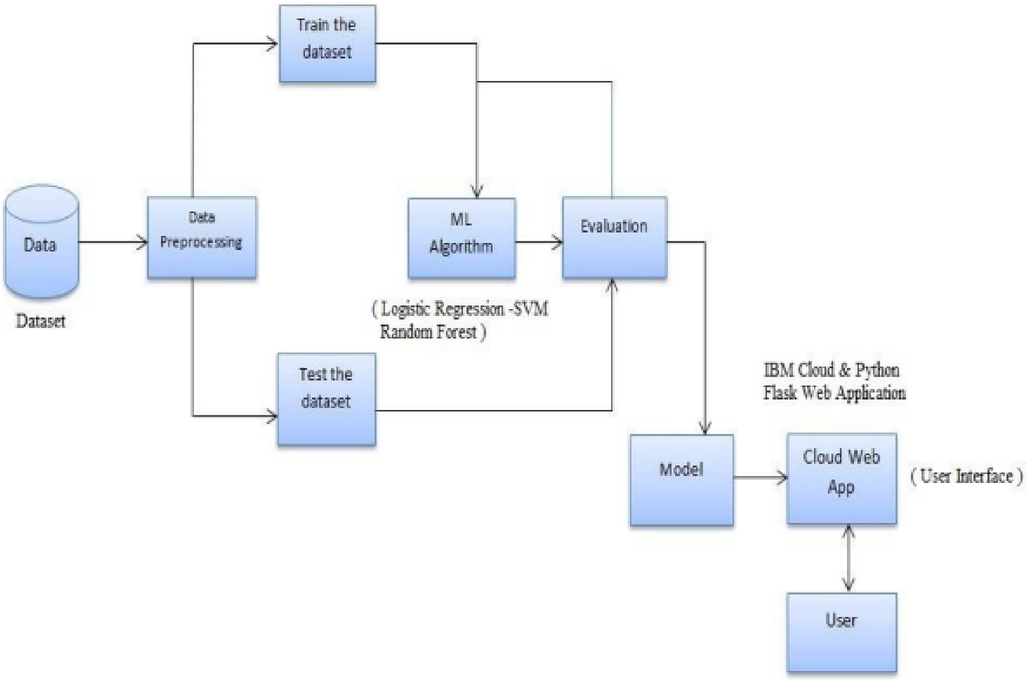
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Responsive UI / UX Design and users can easily configure the settings based on their preference.
NFR-2	**Security**	Implementation of Updated security algorithms and techniques.
NFR-3	**Reliability**	Reliability Factor determines the possibility of a suspected site to be Valid or Fake.
NFR-4	**Performance**	The two main characteristics of a phishing site are that it looks extremely similar to a legitimate site and that it has at least one field to enable users to input their credentials.
NFR-5	**Availability**	It occurs when an attacker, masquerading as a trusted entity, dupes a victim into opening an email, instant message, or text message.
NFR-6	**Scalability**	Scalable detection and isolation of phishing, the main ideas are to move the protection from end users towards the network provider and to employ the novel bad neighbourhood concept, in order to detect and isolate both phishing e mail senders and phishing web servers.

5.PROJECT DESIGN

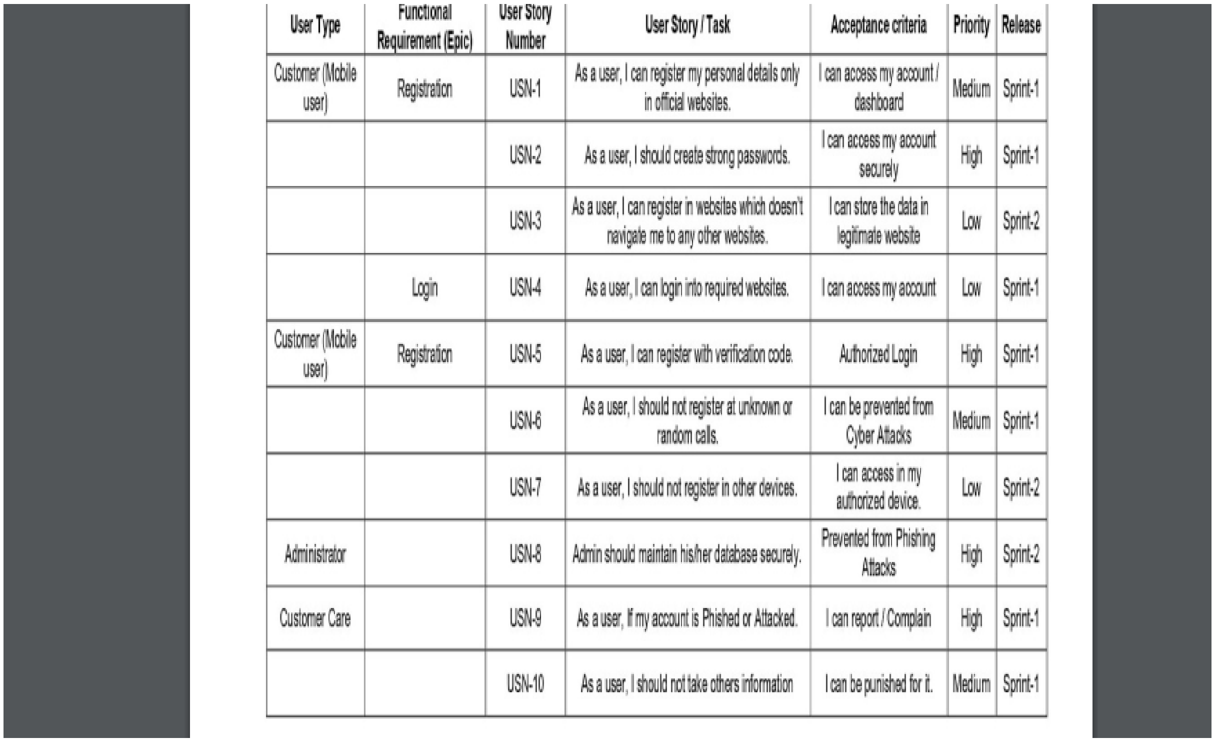


5.2 SOLUTION AND TEECHNICAL ARCHITECTURE





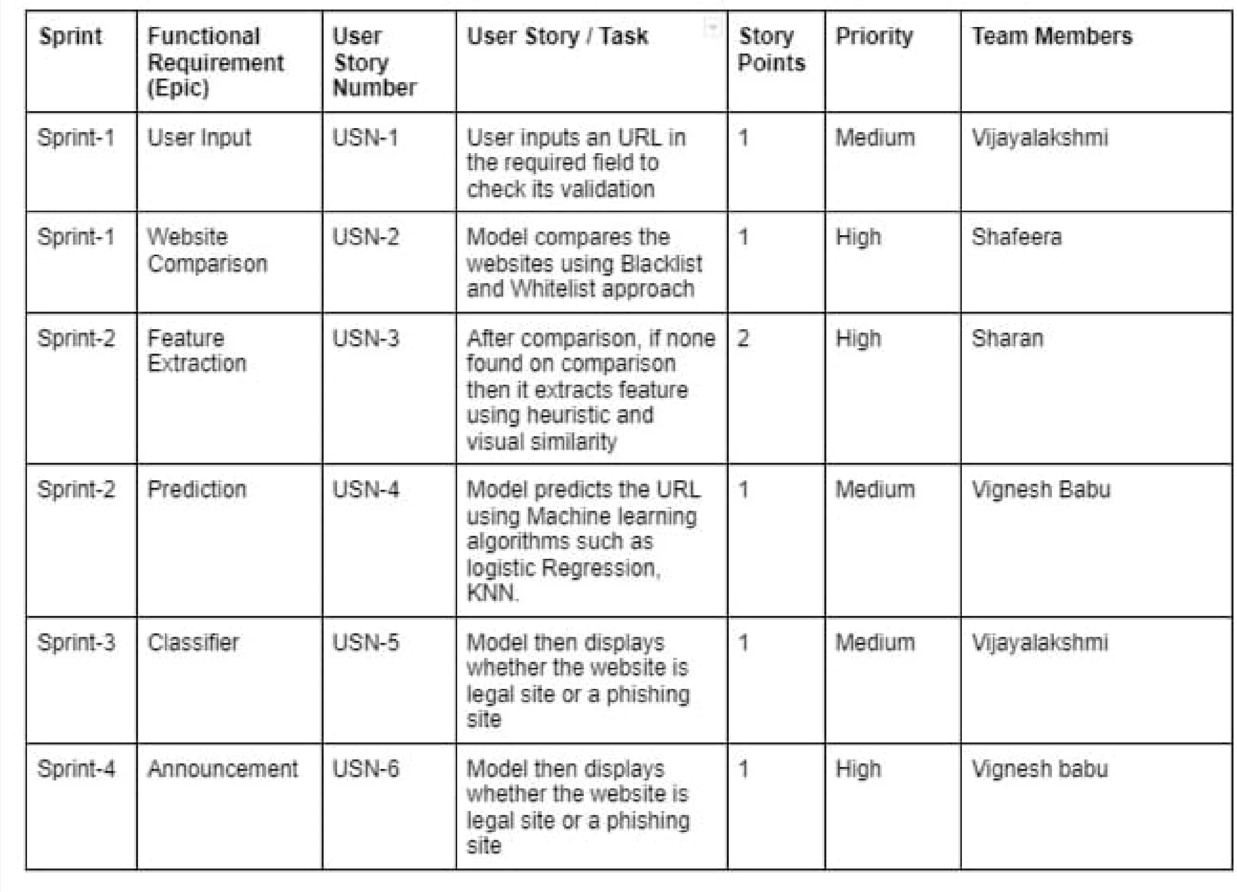
5.3 USER STORIES



User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register my personal details only in official websites.	I can access my account / dashboard	Medium	Sprint-1
		USN-2	As a user, I should create strong passwords.	I can access my account securely	High	Sprint-1
		USN-3	As a user, I can register in websites which doesn't navigate me to any other websites.	I can store the data in legitimate website	Low	Sprint-2
	Login	USN-4	As a user, I can login into required websites.	I can access my account	Low	Sprint-1
Customer (Mobile user)	Registration	USN-5	As a user, I can register with verification code.	Authorized Login	High	Sprint-1
		USN-6	As a user, I should not register at unknown or random calls.	I can be prevented from Cyber Attacks	Medium	Sprint-1
		USN-7	As a user, I should not register in other devices.	I can access in my authorized device.	Low	Sprint-2
Administrator		USN-8	Admin should maintain his/her database securely.	Prevented from Phishing Attacks	High	Sprint-2
Customer Care		USN-9	As a user, If my account is Phished or Attacked.	I can report / Complain	High	Sprint-1
		USN-10	As a user, I should not take others information	I can be punished for it.	Medium	Sprint-1

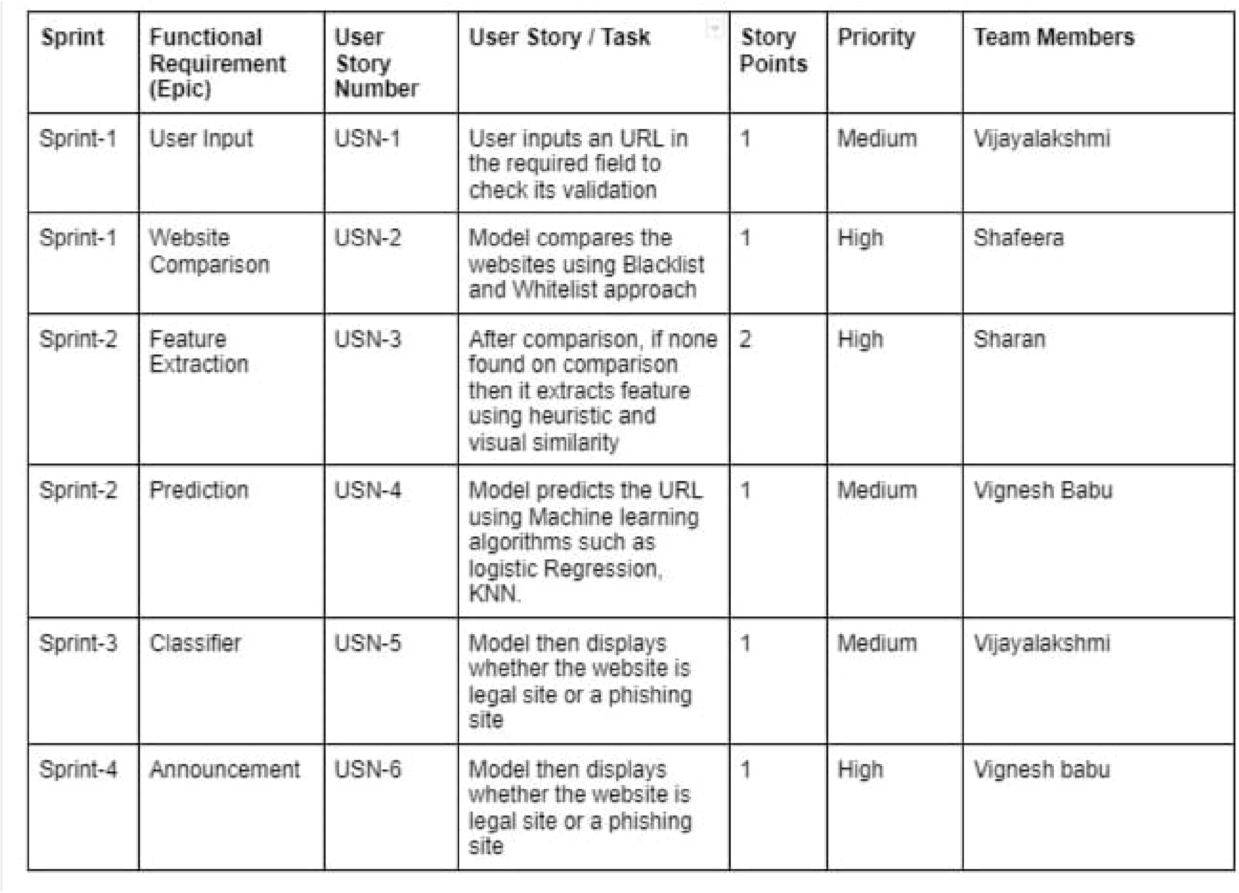
6.PROJECT PLANNING AND SCHEDULING

6.1 SPRINT PLANNING AND EXECUTION



Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	User Input	USN-1	User inputs an URL in the required field to check its validation	1	Medium	Vijayalakshmi
Sprint-1	Website Comparison	USN-2	Model compares the websites using Blacklist and Whitelist approach	1	High	Shafeera
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extracts feature using heuristic and visual similarity	2	High	Sharan
Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic Regression, KNN.	1	Medium	Vignesh Babu
Sprint-3	Classifier	USN-5	Model then displays whether the website is legal site or a phishing site	1	Medium	Vijayalakshmi
Sprint-4	Announcement	USN-6	Model then displays whether the website is legal site or a phishing site	1	High	Vignesh babu

6.2 SPRINT DELIVERY SCHEDULE



Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	User Input	USN-1	User inputs an URL in the required field to check its validation	1	Medium	Vijayalakshmi
Sprint-1	Website Comparison	USN-2	Model compares the websites using Blacklist and Whitelist approach	1	High	Shafeera
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extracts feature using heuristic and visual similarity	2	High	Sharan
Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic Regression, KNN.	1	Medium	Vignesh Babu
Sprint-3	Classifier	USN-5	Model then displays whether the website is legal site or a phishing site	1	Medium	Vijayalakshmi
Sprint-4	Announcement	USN-6	Model then displays whether the website is legal site or a phishing site	1	High	Vignesh babu

7.CODING & SOLUTIONING

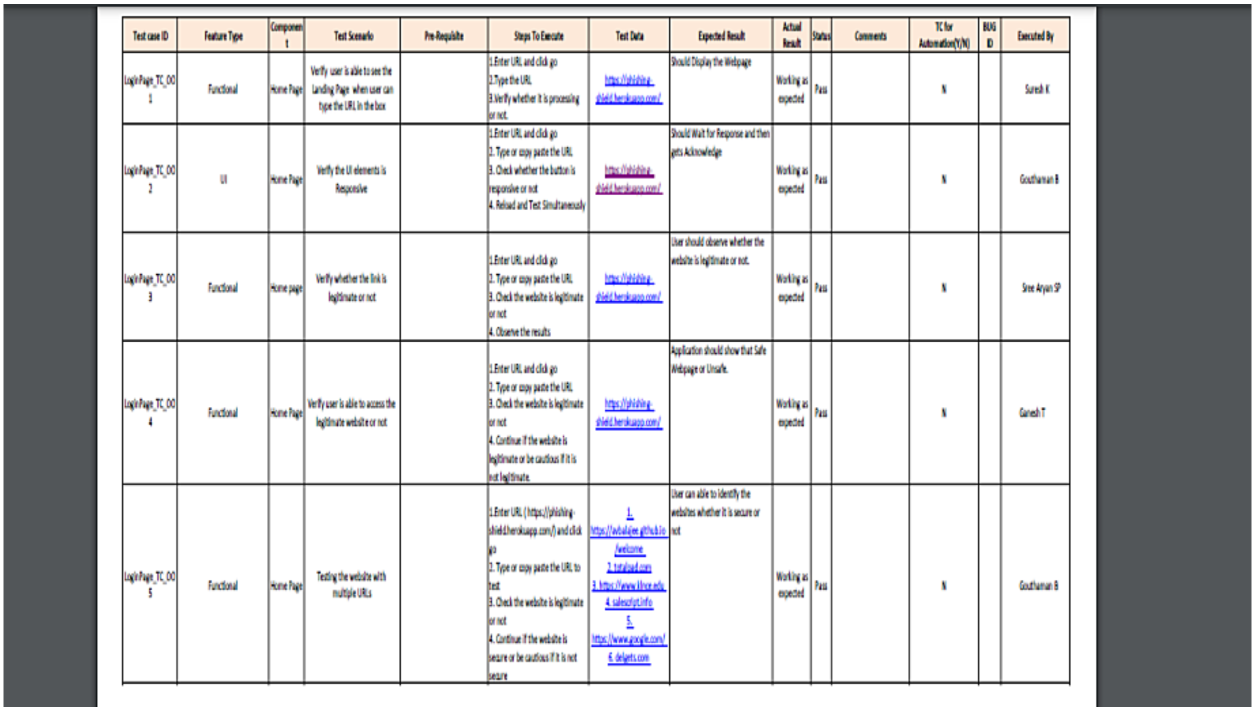
```
<!DOCTYPE html>
<html>
<head>
<meta name="viewport" content="width=device-width, initial-scale=1">
<style>
body {
font-family: Arial, Helvetica, sans-serif;
background-color: black;
}
* {
box-sizing: border-box;
}
/* Add padding to containers */
```

```
.container {
padding: 16px;
background-color: white;
}
/* Full-width input fields */
input[type=text], input[type=password] {
width: 100%;
padding: 15px;
margin: 5px 0 22px 0;
display: inline-block;
border: none;
background: #f1f1f1;
}
input[type=text]:focus, input[type=password]:focus {
background-color: #ddd;
outline: none;
}
/* Overwrite default styles of hr */
hr {
border: 1px solid #f1f1f1;
margin-bottom: 25px;
}
/* Set a style for the submit button */
.registerbtn {
background-color: #04AA6D;
color: white;
padding: 16px 20px;
margin: 8px 0;
border: none;
cursor: pointer;
width: 100%;
opacity: 0.9;
}
.registerbtn:hover {
opacity: 1;
}
/* Add a blue text color to links */
a {
color: dodgerblue;
}
/* Set a grey background color and center the text of the "sign in" section */
.signin {
background-color: #f1f1f1;
text-align: center;
}
</style>
</head>
<body>
<form action="success.html" method="POST">
<div class="container">
<h1>Register</h1>
<p>Please fill in this form to create an account.</p>
<hr>
<label for="email"><b>Email</b></label>
<input type="text" placeholder="Enter Email" name="email" id="email"
required>
<label for="psw"><b>Password</b></label>
<input type="password" placeholder="Enter Password" name="psw"
id="psw" required>
<button type="submit" class="registerbtn">Register</button>
</div>
</form>
</body>
</html>
```

```
<!DOCTYPE html>
<html>
<head>
<style>
#rcorners1 {
border-radius: 25px;
background: #87CEEB;
padding: 200px;
width: 200px;
height: 150px;
}
div {
width: 100px;
height: 100px;
background-color: red;
position: relative;
animation-name: example;
animation-duration: 4s;
animation-iteration-count: infinite;
}
@keyframes example {
0% {background-color:#6699ff; left:0px; top:0px;}
25% {background-color: #0066ff; left:200px; top:0px;}
50% {background-color:#0000ff; left:200px; top:200px;}
75% {background-color:#000099; left:0px; top:200px;}
100% {background-color:#0000cc; left:0px; top:0px;}
}
</style>
</head>
<body>
<center>
<div id="rcorners1">
<h2><b>REGISTRATION SUCCESSFULL!!!</b></h2>
</div>
</center>
</body>
</html>
```

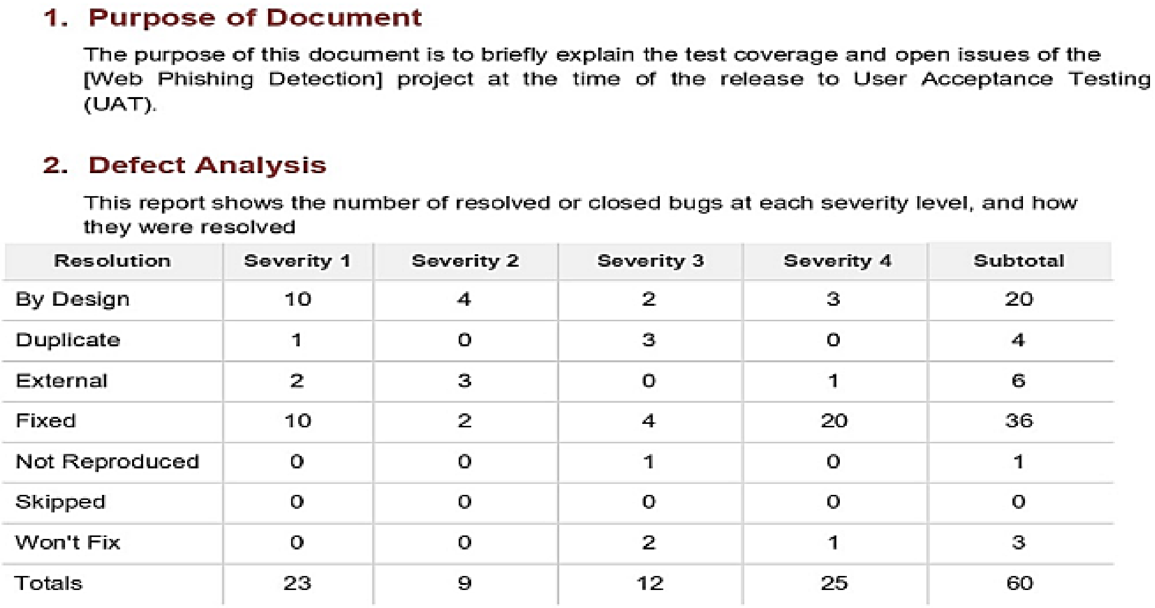
8.TESTING

8.1 TEST CASES



Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Test Data	Expected Result	Actual Result	Status	Comments	TC for Automation(Y/N)	BUG ID	Executed By
LoginPage_TC_001	Functional	Home Page	Verify user is able to see the Landing Page when user can type the URL in the box		1.Enter URL and click go 2.Type the URL 3.Verify whether it is processing or not.	https://phishing-shield.herokuapp.com/	Should Display the Webpage	Working as expected	Pass		N		Suresh K
LoginPage_TC_002	UI	Home Page	Verify the UI elements is Responsive		1.Enter URL and click go 2. Type or copy paste the URL 3. Check whether the button is responsive or not 4. Reload and Test Simultaneously	https://phishing-shield.herokuapp.com/	Should Wait for Response and then gets Acknowledge	Working as expected	Pass		N		Gouthaman B
LoginPage_TC_003	Functional	Home page	Verify whether the link is legitimate or not		1.Enter URL and click go 2. Type or copy paste the URL 3. Check the website is legitimate or not 4. Observe the results	https://phishing-shield.herokuapp.com/	User should observe whether the website is legitimate or not.	Working as expected	Pass		N		Sree Aryan SP
LoginPage_TC_004	Functional	Home Page	Verify user is able to access the legitimate website or not		1.Enter URL and click go 2. Type or copy paste the URL 3. Check the website is legitimate or not 4. Continue if the website is legitimate or be cautious if it is not legitimate.	https://phishing-shield.herokuapp.com/	Application should show that Safe Webpage or Unsafe.	Working as expected	Pass		N		Ganesh T
LoginPage_TC_005	Functional	Home Page	Testing the website with multiple URLs		1.Enter URL (https://phishing-shield.herokuapp.com/) and click go 2. Type or copy paste the URL to test 3. Check the website is legitimate or not 4. Continue if the website is secure or be cautious if it is not secure	1. https://avbalajee.github.io/welcome 2. totalad.com 3. https://www.kinox.edu 4. salescript.info 5. https://www.google.com/ 6. delgets.com	User can able to identify the websites whether it is secure or not	Working as expected	Pass		N		Gouthaman B

8.2 USER ACCEPTANCE TESTING



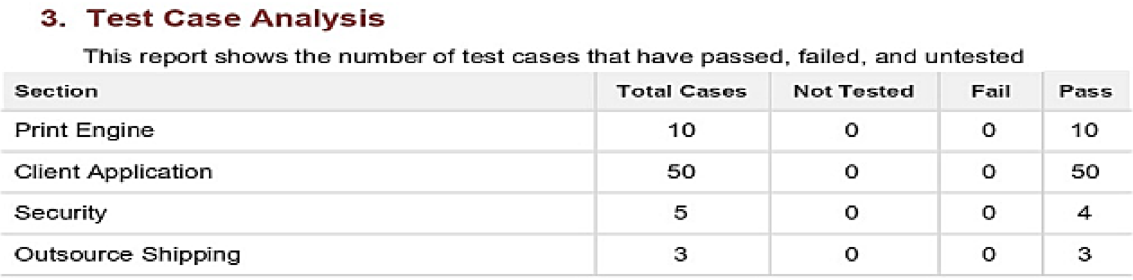
1. Purpose of Document

The purpose of this document is to briefly explain the test coverage and open issues of the [Web Phishing Detection] project at the time of the release to User Acceptance Testing (UAT).

2. Defect Analysis

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

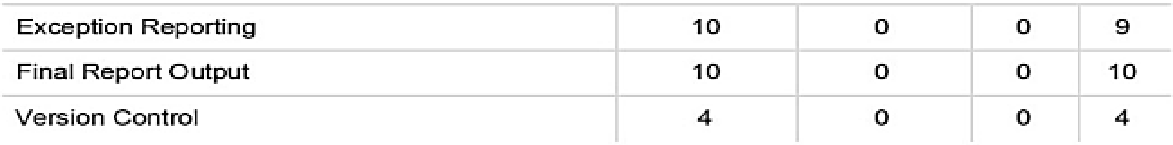
Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	10	2	4	20	36
Not Reproduced	0	0	1	0	1
Skipped	0	0	0	0	0
Won't Fix	0	0	2	1	3
Totals	23	9	12	25	60



3. Test Case Analysis

This report shows the number of test cases that have passed, failed, and untested

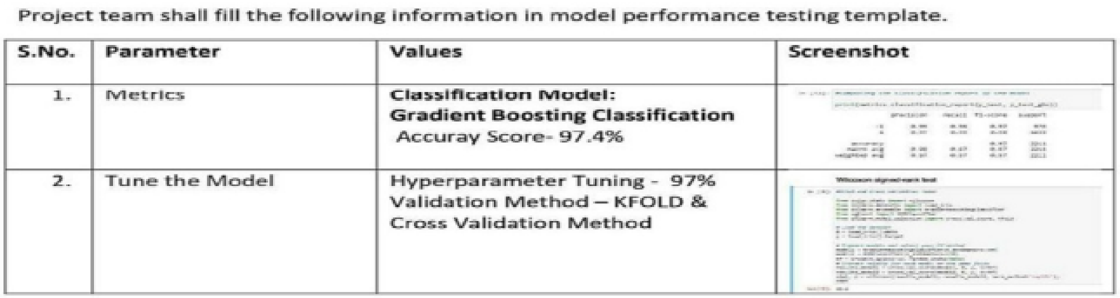
Section	Total Cases	Not Tested	Fail	Pass
Print Engine	10	0	0	10
Client Application	50	0	0	50
Security	5	0	0	4
Outsource Shipping	3	0	0	3



Exception Reporting	10	0	0	9
Final Report Output	10	0	0	10
Version Control	4	0	0	4

9. RESULTS

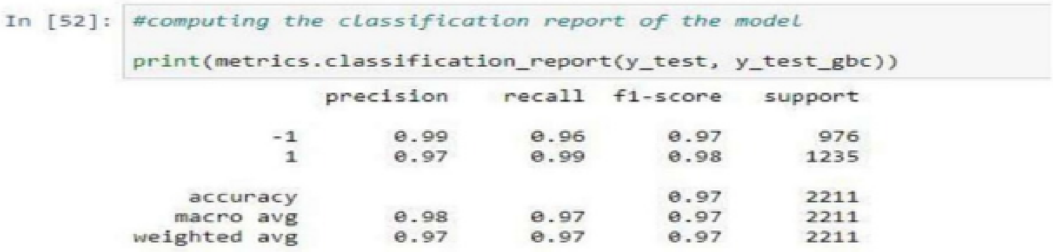
9.1 PERFORMANCE METRICS



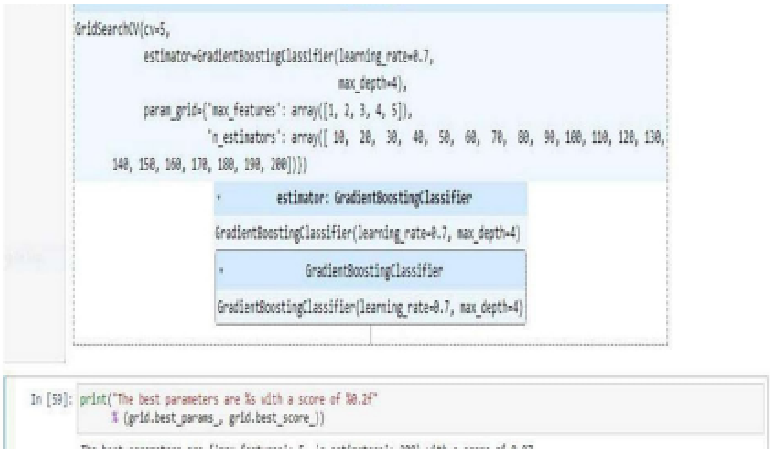
Project team shall fill the following information in model performance testing template.

S.No.	Parameter	Values	Screenshot
1.	Metrics	**Classification Model:** **Gradient Boosting Classification** Accuray Score- 97.4%	
2.	Tune the Model	Hyperparameter Tuning - 97% Validation Method – KFOLD & Cross Validation Method	

1. METRICS:
CLASSIFICATION REPORT:



```
In [52]: #computing the classification report of the model

         print(metrics.classification_report(y_test, y_test_gbc))

                      precision    recall  f1-score   support

                -1         0.99      0.96      0.97       976
                 1         0.97      0.99      0.98      1235

          accuracy                               0.97      2211
         macro avg         0.98      0.97      0.97      2211
      weighted avg         0.97      0.97      0.97      2211
```

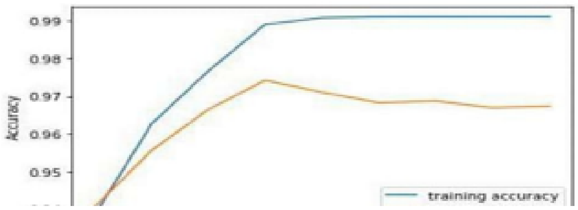


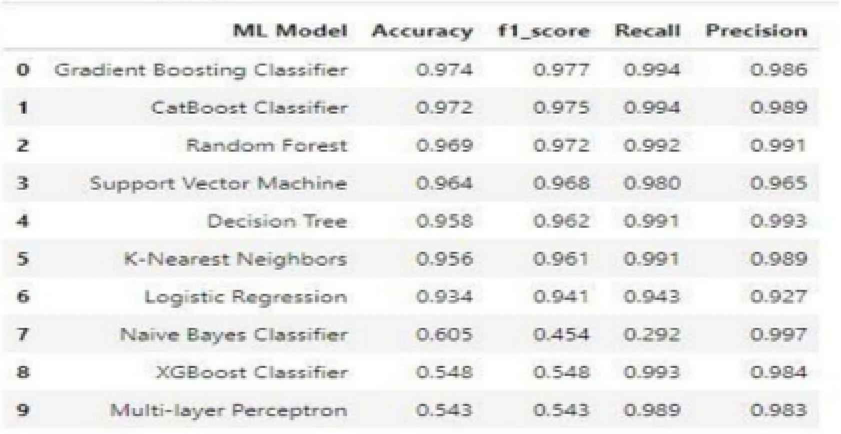
```
GridSearchCV(cv=5,
             estimator=GradientBoostingClassifier(learning_rate=0.7,
                                                  max_depth=4),
             param_grid={'max_features': array([1, 2, 3, 4, 5]),
                         'n_estimators': array([ 10,  20,  30,  40,  50,  60,  70,  80,  90, 100, 110, 120, 130,
       140, 150, 160, 170, 180, 190, 200])})
```

estimator: GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.7, max_depth=4)
GradientBoostingClassifier
GradientBoostingClassifier(learning_rate=0.7, max_depth=4)

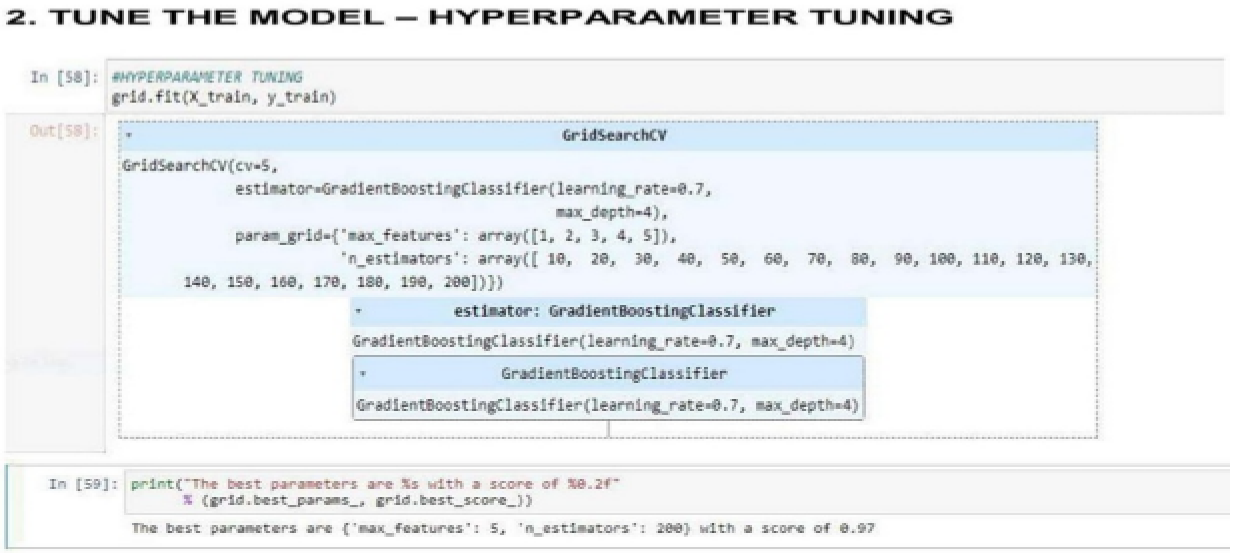
```
In [59]: print("The best parameters are %s with a score of %0.2f"
               % (grid.best_params_, grid.best_score_))

         The best parameters are {'max_features': 5, 'n_estimators': 200} with a score of 0.97
```

PERFORMANCE :





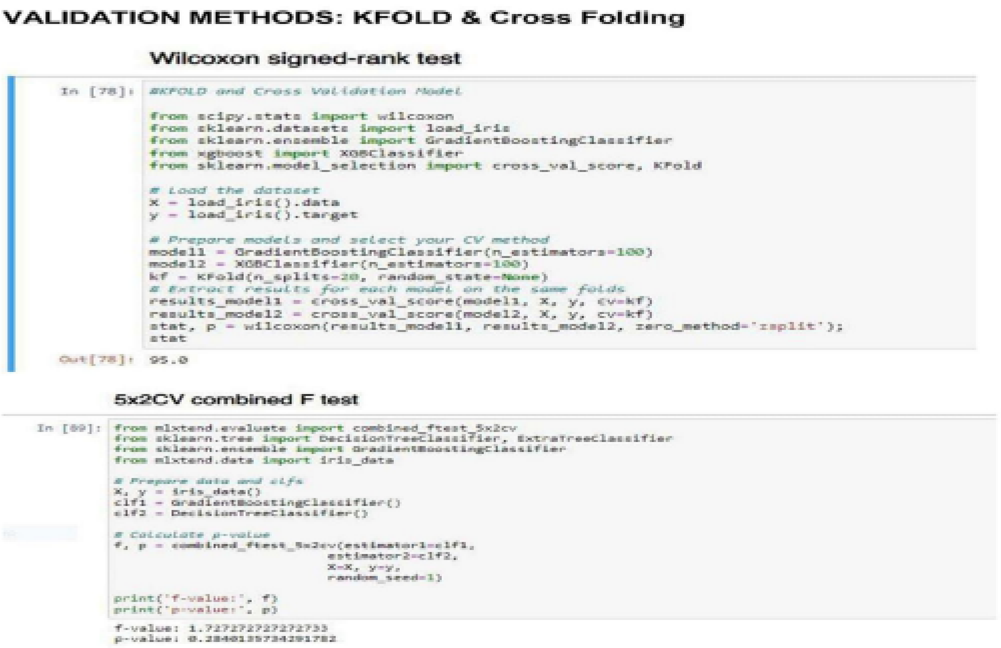
	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	CatBoost Classifier	0.972	0.975	0.994	0.989
2	Random Forest	0.969	0.972	0.992	0.991
3	Support Vector Machine	0.964	0.968	0.980	0.965
4	Decision Tree	0.958	0.962	0.991	0.993
5	K-Nearest Neighbors	0.956	0.961	0.991	0.989
6	Logistic Regression	0.934	0.941	0.943	0.927
7	Naive Bayes Classifier	0.605	0.454	0.292	0.997
8	XGBoost Classifier	0.548	0.548	0.993	0.984
9	Multi-layer Perceptron	0.543	0.543	0.989	0.983



2. TUNE THE MODEL – HYPERPARAMETER TUNING

```
In [58]: #HYPERPARAMETER TUNING
         grid.fit(X_train, y_train)

Out[58]:
GridSearchCV

GridSearchCV(cv=5,
             estimator=GradientBoostingClassifier(learning_rate=0.7,
                                                  max_depth=4),
             param_grid={'max_features': array([1, 2, 3, 4, 5]),
                         'n_estimators': array([ 10,  20,  30,  40,  50,  60,  70,  80,  90, 100, 110, 120, 130,
       140, 150, 160, 170, 180, 190, 200])})

         estimator: GradientBoostingClassifier
         GradientBoostingClassifier(learning_rate=0.7, max_depth=4)

         GradientBoostingClassifier
         GradientBoostingClassifier(learning_rate=0.7, max_depth=4)
```

```
In [59]: print("The best parameters are %s with a score of %0.2f"
               % (grid.best_params_, grid.best_score_))

         The best parameters are {'max_features': 5, 'n_estimators': 200} with a score of 0.97
```



VALIDATION METHODS: KFOLD & Cross Folding

Wilcoxon signed-rank test

```
In [78]: #KFOLD and Cross Validation Model

         from scipy.stats import wilcoxon
         from sklearn.datasets import load_iris
         from sklearn.ensemble import GradientBoostingClassifier
         from xgboost import XGBClassifier
         from sklearn.model_selection import cross_val_score, KFold

         # Load the dataset
         X = load_iris().data
         y = load_iris().target

         # Prepare models and select your CV method
         model1 = GradientBoostingClassifier(n_estimators=100)
         model2 = XGBClassifier(n_estimators=100)
         kf = KFold(n_splits=20, random_state=None)
         # Extract results for each model on the same folds
         results_model1 = cross_val_score(model1, X, y, cv=kf)
         results_model2 = cross_val_score(model2, X, y, cv=kf)
         stat, p = wilcoxon(results_model1, results_model2, zero_method='zsplit');
         stat

Out[78]: 95.0
```

5x2CV combined F test

```
In [89]: from mlxtend.evaluate import combined_ftest_5x2cv
         from sklearn.tree import DecisionTreeClassifier, ExtraTreeClassifier
         from sklearn.ensemble import GradientBoostingClassifier
         from mlxtend.data import iris_data

         # Prepare data and clfs
         X, y = iris_data()
         clf1 = GradientBoostingClassifier()
         clf2 = DecisionTreeClassifier()

         # Calculate p-value
         f, p = combined_ftest_5x2cv(estimator1=clf1,
                                     estimator2=clf2,
                                     X=X, y=y,
                                     random_seed=1)

         print('f-value:', f)
         print('p-value:', p)

         f-value: 1.7272727272727273
         p-value: 0.2840135734291782
```

10.ADVANTANGES & DISADVANTAGES ADVANTAGES

- It takes the Load off the Security team
- Improve on Inefficiencies of SEG and Phishing Awareness Training
- Password Management made Easy

DISADVANTAGES

- Low Detection Accuracy
- False Alarm

11.CONCLUSION

This paper presented an intelligent phishing detection and protection scheme by employing a new approach using the integrated features of images, frames and text of phishing websites. An efficient ANFIS algorithm was developed, tested and verified for phishing website detection and protection based on the schemes proposed in Aburrous et al. (2010) and Barraclough and Sexton (2015). A set of experiments was performed using 13,000 available datasets. The approach showed an accuracy of 98.3%, which so far, is the best-integrated solutions for web-phishing detection and protection. The primary contribution of this study is the integration of hybrid features that have been extracted from text, images and frames and that are then used to develop a robust ANFIS solution. Future work will include using another algorithm like deep-learning for phishing web page detection and compare the effectiveness with the current result. More also, a web browser plug-in will be developed based on an efficient algorithm to detect phishing website and thus protect users in real time.

12.FUTURE SCOPE

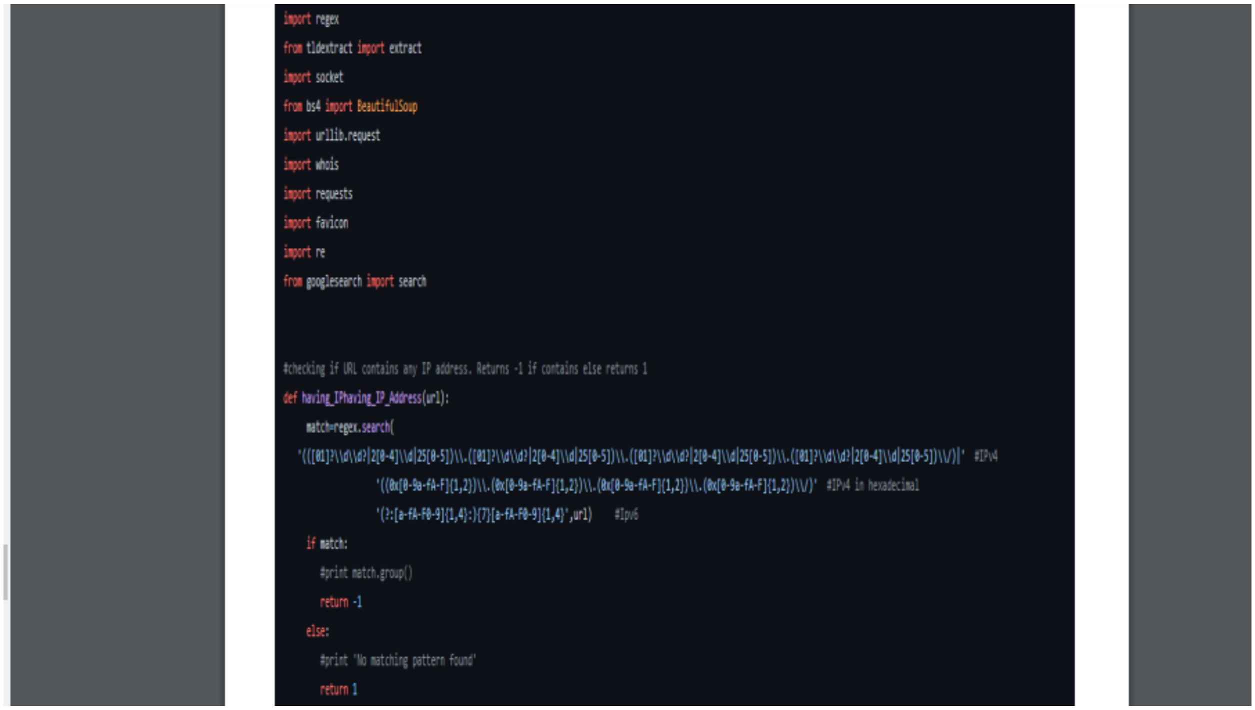
In future if we get structured dataset of phishing we can perform phishing detection much more faster than any other technique.In future we can use a combination of any other two or more classifier to get maximum accuracy. We also plan to explore various phishing techniques that uses Lexical features, Network based features,Content based features, Webpage based features and HTML and JavaScript features of web pages which can improve the performance of the system. In particular, we extract features from URLs and pass it through the various classifiers.

We have seen that existing system gives less accuracy so we proposed a new phishing method that employs URL based features and also we generated classifiers through several machine learning Future work will consist in releasing components of the tools as an add-on for a Web browser such as Mozilla Firefox. In addition, the technique proposed, which is complementary to that introduced in this paper, will be merged to create a phishing detection system with a larger scope of action. We also plan to release the analytics related part in a larger Big Data security analytics stack, which is under current development in our lab.

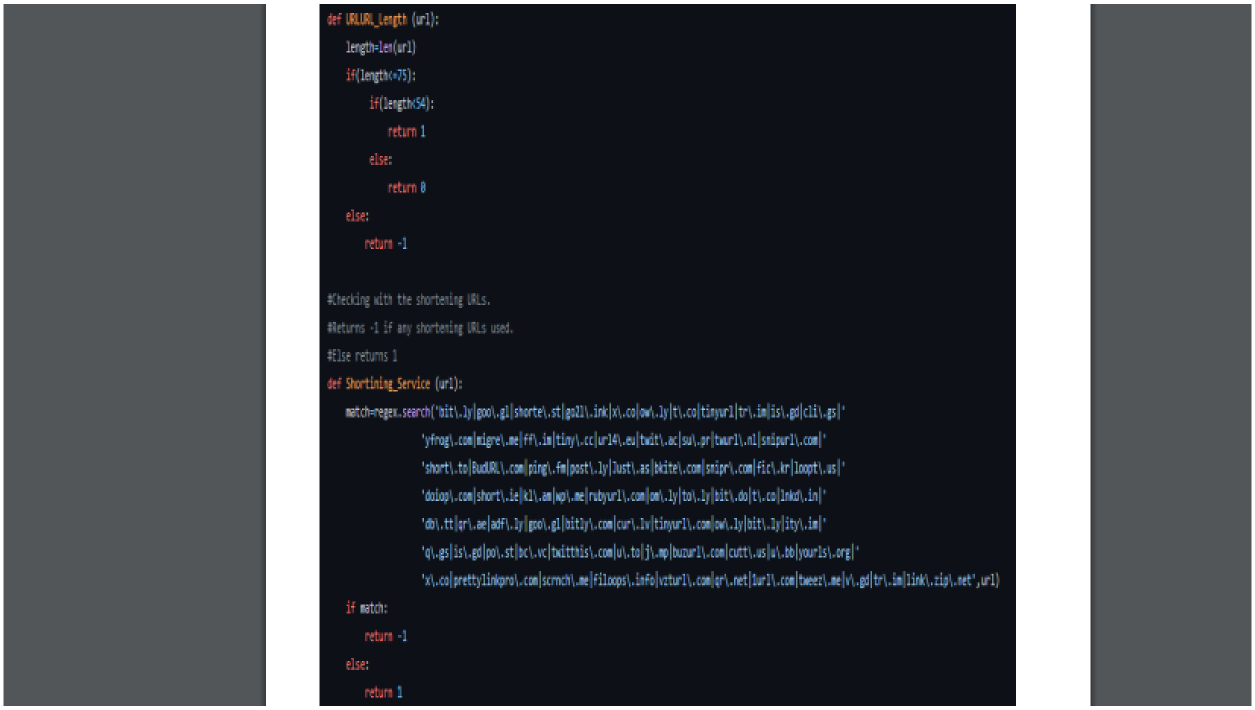
13. APPENDIX

SOURCE CODE

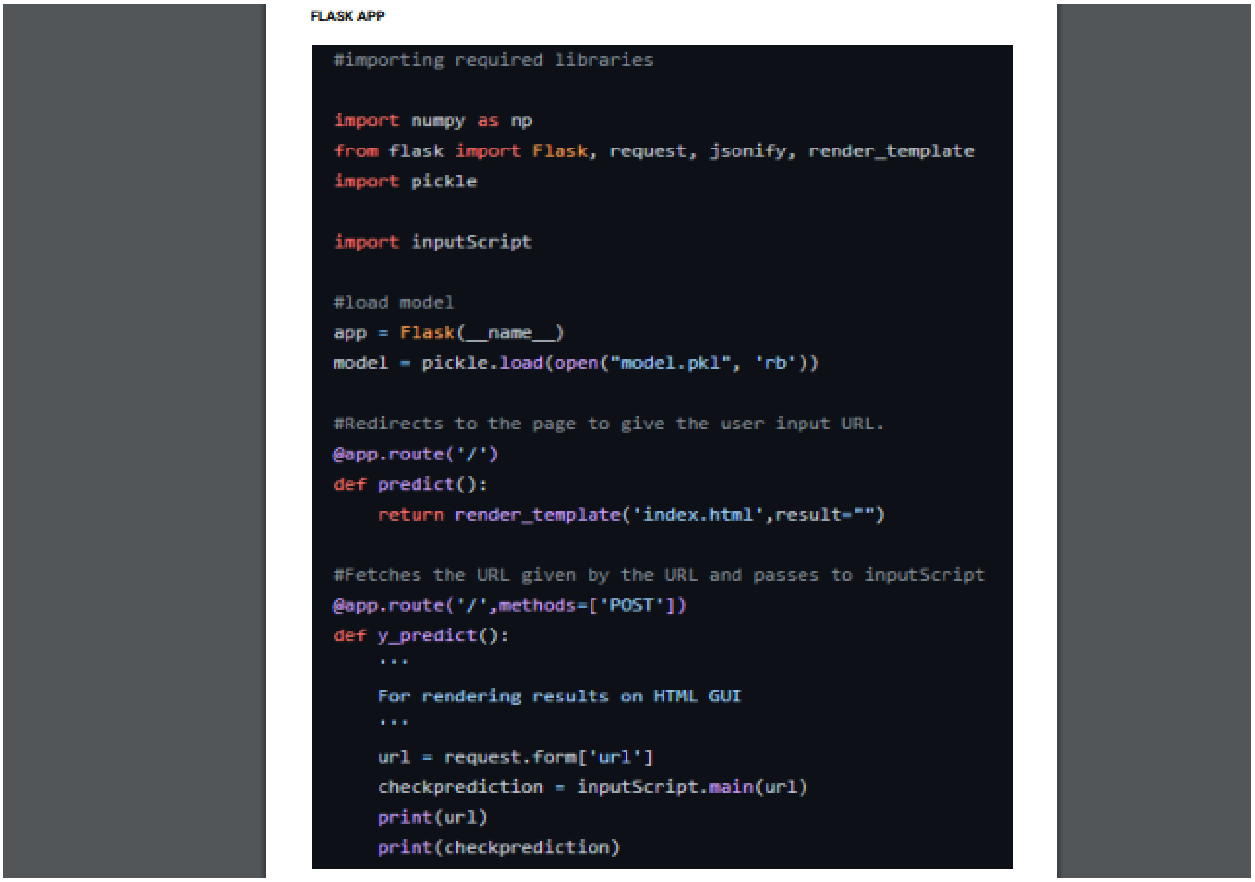
MODEL CREATION



```
import regex
from tldextract import extract
import socket
from bs4 import BeautifulSoup
import urllib.request
import whois
import requests
import favicon
import re
from googlesearch import search



#checking if URL contains any IP address. Returns -1 if contains else returns 1
def having_IPhaving_IP_Address(url):
    match=regex.search(
  '(([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\.([01]?\\d\\d?|2[0-4]\\d|25[0-5])\\/)|'  #IPv4
                    '((0x[0-9a-fA-F]{1,2})\\.(0x[0-9a-fA-F]{1,2})\\.(0x[0-9a-fA-F]{1,2})\\.(0x[0-9a-fA-F]{1,2})\\/)'  #IPv4 in hexadecimal
                    '(?:[a-fA-F0-9]{1,4}:){7}[a-fA-F0-9]{1,4}',url)    #Ipv6
    if match:
      #print match.group()
      return -1
    else:
      #print 'No matching pattern found'
      return 1
```



```
def URLURL_Length (url):
    length=len(url)
    if(length<=75):
        if(length<54):
            return 1
        else:
            return 0
    else:
        return -1

#Checking with the shortening URLs.
#Returns -1 if any shortening URLs used.
#Else returns 1
def Shortining_Service (url):
    match=regex.search('bit\.ly|goo\.gl|shorte\.st|go2l\.ink|x\.co|ow\.ly|t\.co|tinyurl|tr\.im|is\.gd|cli\.gs|'
                    'yfrog\.com|migre\.me|ff\.im|tiny\.cc|url4\.eu|twit\.ac|su\.pr|twurl\.nl|snipurl\.com|'
                    'short\.to|BudURL\.com|ping\.fm|post\.ly|Just\.as|bkite\.com|snipr\.com|fic\.kr|loopt\.us|'
                    'doiop\.com|short\.ie|kl\.am|wp\.me|rubyurl\.com|om\.ly|to\.ly|bit\.do|t\.co|lnkd\.in|'
                    'db\.tt|qr\.ae|adf\.ly|goo\.gl|bitly\.com|cur\.lv|tinyurl\.com|ow\.ly|bit\.ly|ity\.im|'
                    'q\.gs|is\.gd|po\.st|bc\.vc|twitthis\.com|u\.to|j\.mp|buzurl\.com|cutt\.us|u\.bb|yourls\.org|'
                    'x\.co|prettylinkpro\.com|scrnch\.me|filoops\.info|vzturl\.com|qr\.net|1url\.com|tweez\.me|v\.gd|tr\.im|link\.zip\.net',url)
    if match:
        return -1
    else:
        return 1
```



FLASK APP

```
#importing required libraries

import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle

import inputScript

#load model
app = Flask(__name__)
model = pickle.load(open("model.pkl", 'rb'))

#Redirects to the page to give the user input URL.
@app.route('/')
def predict():
    return render_template('index.html',result="")

#Fetches the URL given by the URL and passes to inputScript
@app.route('/',methods=['POST'])
def y_predict():
    '''
    For rendering results on HTML GUI
    '''
    url = request.form['url']
    checkprediction = inputScript.main(url)
    print(url)
    print(checkprediction)
```

GitHub:

https://github.com/IBM-EPBL/IBM-Project-28852-1660117443

DEMO video

https://web-phishing-detection.herokuapp.com/