Project Report Format

Web Phishing Detection Project

Team ID: PNT2022TMID26927

Team Leader: DARSHAN S

Team member: AAKASH G

Team member: BLESSANCE C

Team member: FARHAAN AHMED S

1. INTRODUCTION

I. Project Overview

The project is a solution to identifying phishing websites from legitimate websites in order to prevent any exploitation of people by those threats. This web phishing detection is built using Machine Learning Modal. The project basically helps users to identify between trustable or cocky websites.

II. Purpose

As remote and hybrid work environments became the new norm, it is essential for a company to make sure that their employees don’t fall for phishing attacks. It is also essential for common internet surfers to identify between legitimate and malicious websites such as phishing websites to prevent identity or financial loss. Number of global phishing sites as of first quarter of 2021 is estimated to be 611,877 by Statista Research Department. It is something fundamental to detect threats before people can be exploited by falling as victim.

2. LITERATURE SURVEY

I. Existing problem

When people are unable to recognise phishing sites, phishing attacks take place. Past anti-phishing research can be divided into four categories: studies to learn why people fall for phishing attacks, strategies for preventing people from falling for phishing attacks, user interfaces to assist people in making better choices when using email and websites, and automated tools to detect phishing. Our research outlines an automated method to identify phishing. The majority of end users typically make their decisions only on how they feel and appear. When a user accesses the internet, all they see is a browser's screen. He or she then works on a web page's command. Most phishing efforts take use of this type of unintended chance provided by the user's lack of care for the back-end procedure.

II. References

- Data Science: Literature Review & State of Art Sanket MantrI(2016)
- Applied Data Science (lessons learned for data driven business) 2019
- Ian Langmore Daniel Krasner - Columbia Applied Data Science

- Introducing Data Science Big Data, Machine Learning, And More, Using Python Tools (Davy Cielen, Arno D.B. Meysman, Mohamed Ali) 2016
- Hands-On Data Science and Python Machine Learning

III. Problem Statement Definition

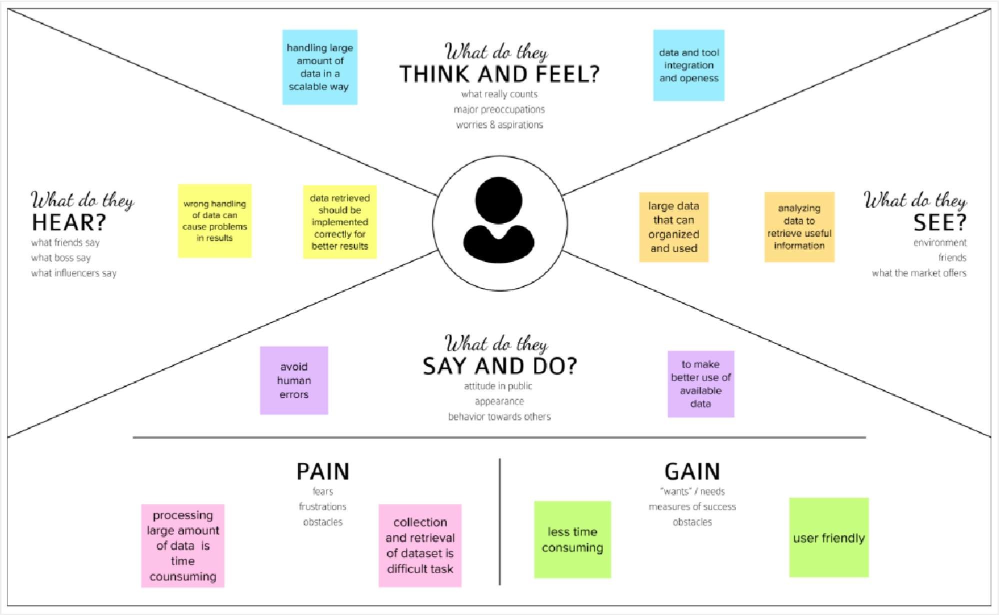
Phishing attacks happen when humans fail to detect phishing sites. Past work in anti-phishing falls into four categories: studies to understand why people fall for phishing attacks, methods for training people not to fall for phishing attacks, user interfaces for helping people make better decisions about rusting email and websites, and automated tools to detect phishing. Our work describes an automated approach to detect phishing. Most of the end user normally takes decision only based on what he/she look and feel. When a user is accessing internet, he/she only see the screen of a browser. He/she then work on the command of a web-page. The user doesn't concern about the back-end process and most phishing attempts get this type of unintentional opportunity given by the user and make them fool. 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.

3. IDEATION & PROPOSED SOLUTION

I. Empathy Map Canvas

An empathy map is a simple, easy-to-digest visual that captures knowledge about a user's behaviours and attitudes.

It is a useful tool to helps teams better understand their users.
Creating an effective solution requires understanding the true problem and the person who is experiencing it. The exercise of creating the map helps participants consider things from the user's perspective along with his or her goals and challenges.



handling large amount of data in a scalable way
What do they THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
data and tool integration and openess
What do they HEAR?
what friends say
what boss say
what influencers say
wrong handling of data can cause problems in results
data retrieved should be implemented correctly for better results
large data that can organized and used
analyzing data to retrieve useful information
What do they SEE?
environment
friends
what the market offers
avoid human errors
What do they SAY AND DO?
attitude in public
appearance
behavior towards others
to make better use of available data
PAIN
fears
frustrations
obstacles
processing large amount of data is time counsuming
collection and retrieval of dataset is difficult task
GAIN
"wants" / needs
measures of success
obstacles
less time consuming
user friendly

II. Ideation & Brainstorming

1

Define your problem statement

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

5 minutes

PROBLEM

How might we find the phishing websites

Key rules of brainstorming

To run an smooth and productive session

Stay in topic.	Encourage wild ideas.
Defer judgment.	Listen to others.
Go for volume.	If possible, be visual.

2

Brainstorm

Write down any ideas that come to mind that address your problem statement.

10 minutes

Darsan

- To avoid public wifi once state phishing can happen
- To block cookies from a suspicious website
- Being careful on websites asking personal information
- Filter phishing emails

Farhaan

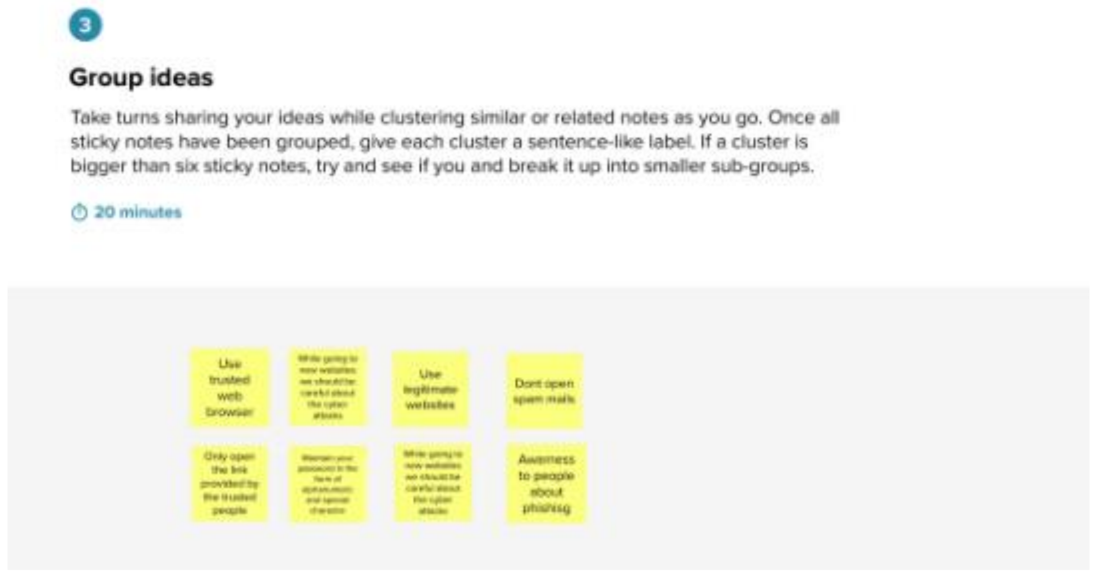
- Dont open spam mails
- Dont attend spam calls
- Dont trust ads
- Use trusted web browser

Blessance

- Change passwords frequently
- Dont open unknown links
- training users to be cautious
- Careful about fraudulent websites

Aakash

- Top resulted websites must be used
- Use indicators for phishing websites
- Use multi login level verification for email
- dont trust blindly



3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. Once all sticky notes have been grouped, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

20 minutes

Use trusted web browser

While going to new websites we should be careful about the cyber attacks

Use legitimate websites

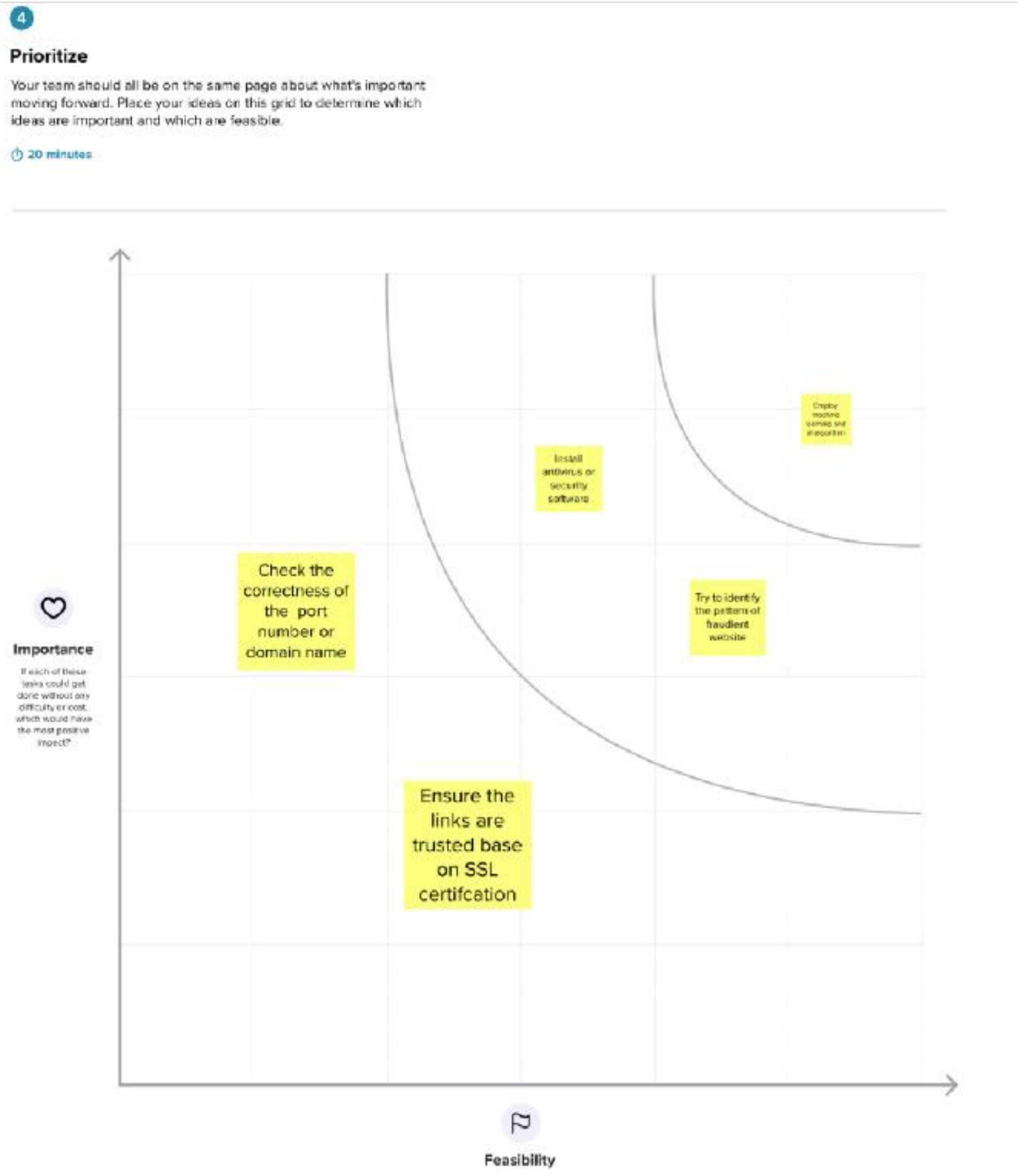
Dont open spam mails

Only open the link provided by the trusted people

Maintain your password in the form of alphanumeric and special character

While going to new websites we should be careful about the cyber attacks

Awarness to people about phishisg



4

Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

20 minutes

Importance

If each of these tasks could get done without any difficulty or cost, which would have the most positive impact?

Feasibility

Check the correctness of the port number or domain name

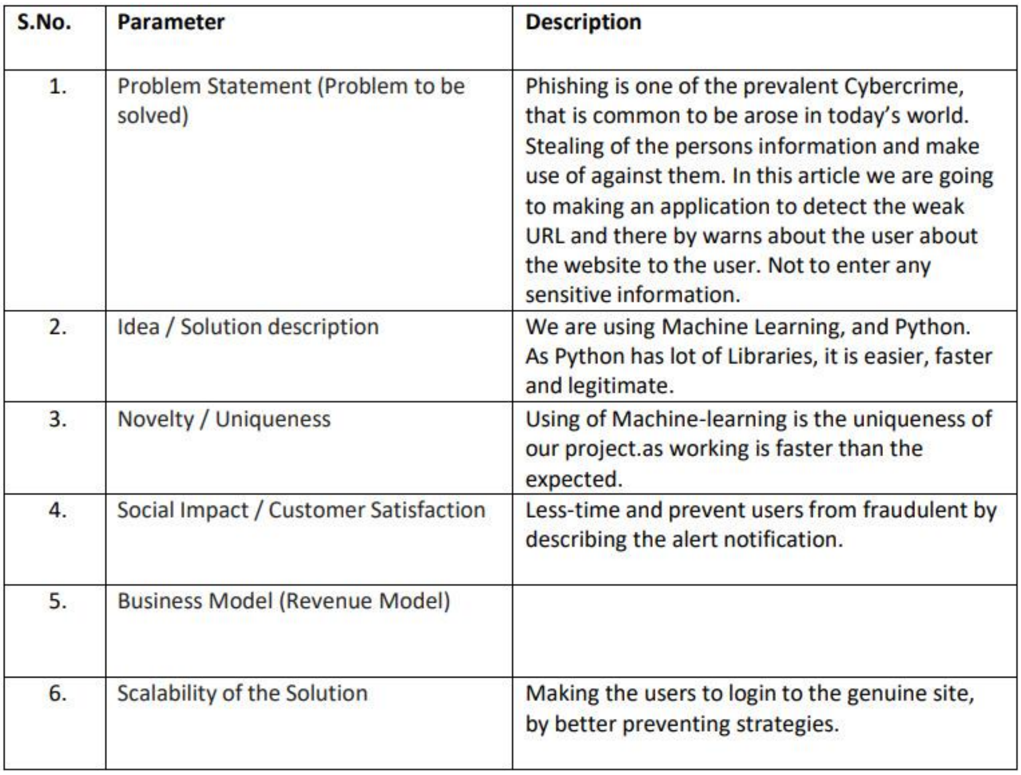
Ensure the links are trusted base on SSL certifcation

Install antivirus or security software

Try to identify the pattern of fraudlent website

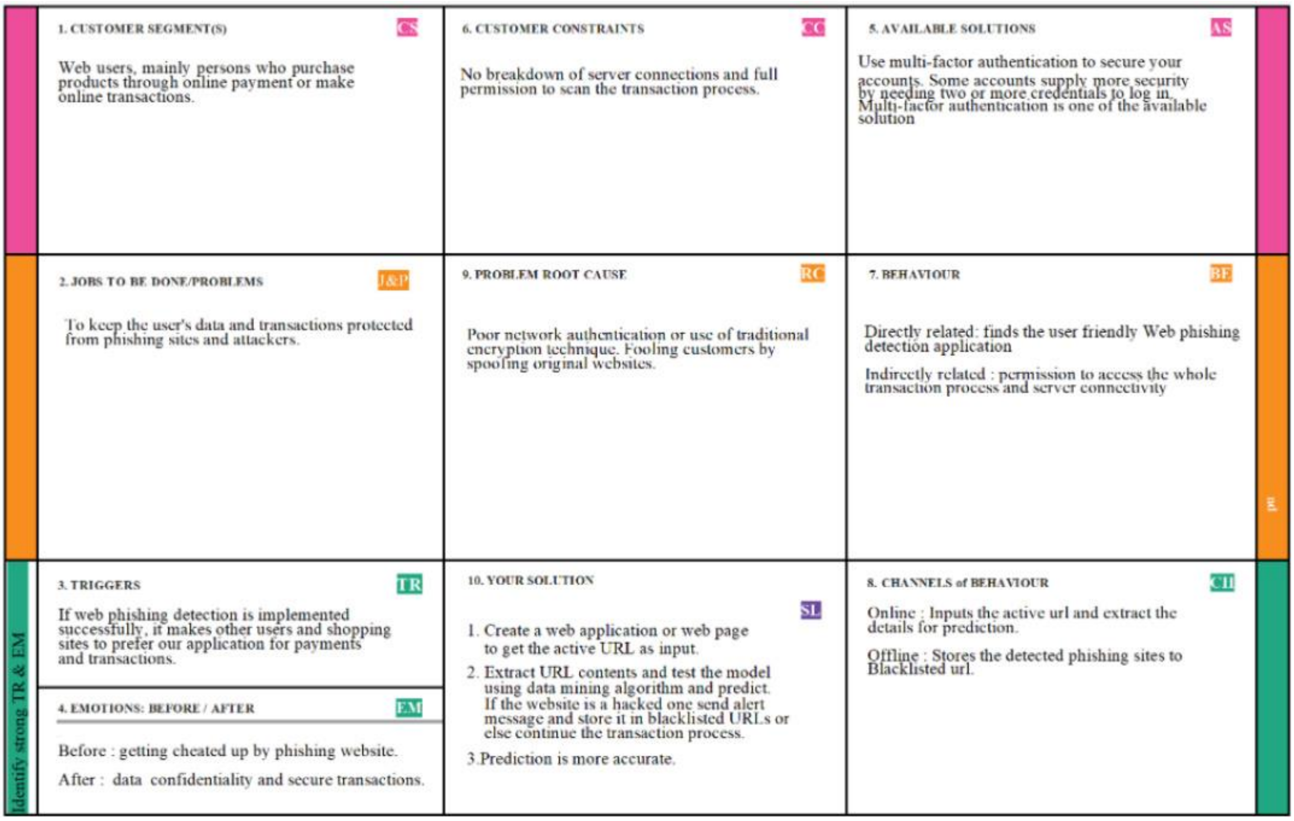
Deploy machine [illegible]

III. Proposed Solution



S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Phishing is one of the prevalent Cybercrime, that is common to be arose in today's world. Stealing of the persons information and make use of against them. In this article we are going to making an application to detect the weak URL and there by warns about the user about the website to the user. Not to enter any sensitive information.
2.	Idea / Solution description	We are using Machine Learning, and Python. As Python has lot of Libraries, it is easier, faster and legitimate.
3.	Novelty / Uniqueness	Using of Machine-learning is the uniqueness of our project.as working is faster than the expected.
4.	Social Impact / Customer Satisfaction	Less-time and prevent users from fraudulent by describing the alert notification.
5.	Business Model (Revenue Model)	
6.	Scalability of the Solution	Making the users to login to the genuine site, by better preventing strategies.

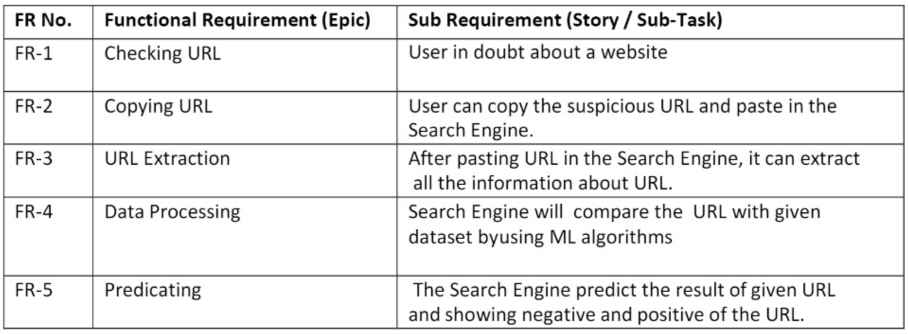
IV. Problem Solution fit



1. CUSTOMER SEGMENT(S) CS Web users, mainly persons who purchase products through online payment or make online transactions.	6. CUSTOMER CONSTRAINTS CC No breakdown of server connections and full permission to scan the transaction process.	5. AVAILABLE SOLUTIONS AS Use multi-factor authentication to secure your accounts. Some accounts supply more security by needing two or more credentials to log in. Multi-factor authentication is one of the available solution
2. JOBS TO BE DONE/PROBLEMS J&P To keep the user's data and transactions protected from phishing sites and attackers.	9. PROBLEM ROOT CAUSE RC Poor network authentication or use of traditional encryption technique. Fooling customers by spoofing original websites.	7. BEHAVIOUR BE Directly related: finds the user friendly Web phishing detection application Indirectly related : permission to access the whole transaction process and server connectivity
3. TRIGGERS TR If web phishing detection is implemented successfully, it makes other users and shopping sites to prefer our application for payments and transactions. 4. EMOTIONS: BEFORE / AFTER EM Before : getting cheated up by phishing website. After : data confidentiality and secure transactions.	10. YOUR SOLUTION SL 1. Create a web application or web page to get the active URL as input. 2. Extract URL contents and test the model using data mining algorithm and predict. If the website is a hacked one send alert message and store it in blacklisted URLs or else continue the transaction process. 3.Prediction is more accurate.	8. CHANNELS of BEHAVIOUR CH Online : Inputs the active url and extract the details for prediction. Offline : Stores the detected phishing sites to Blacklisted url.

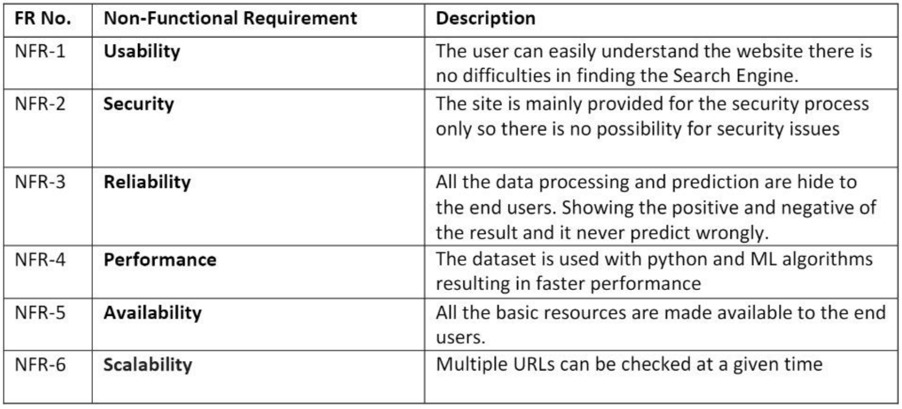
4. REQUIREMENT ANALYSIS

I. Functional requirement



FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Checking URL	User in doubt about a website
FR-2	Copying URL	User can copy the suspicious URL and paste in the Search Engine.
FR-3	URL Extraction	After pasting URL in the Search Engine, it can extract all the information about URL.
FR-4	Data Processing	Search Engine will compare the URL with given dataset byusing ML algorithms
FR-5	Predicating	The Search Engine predict the result of given URL and showing negative and positive of the URL.

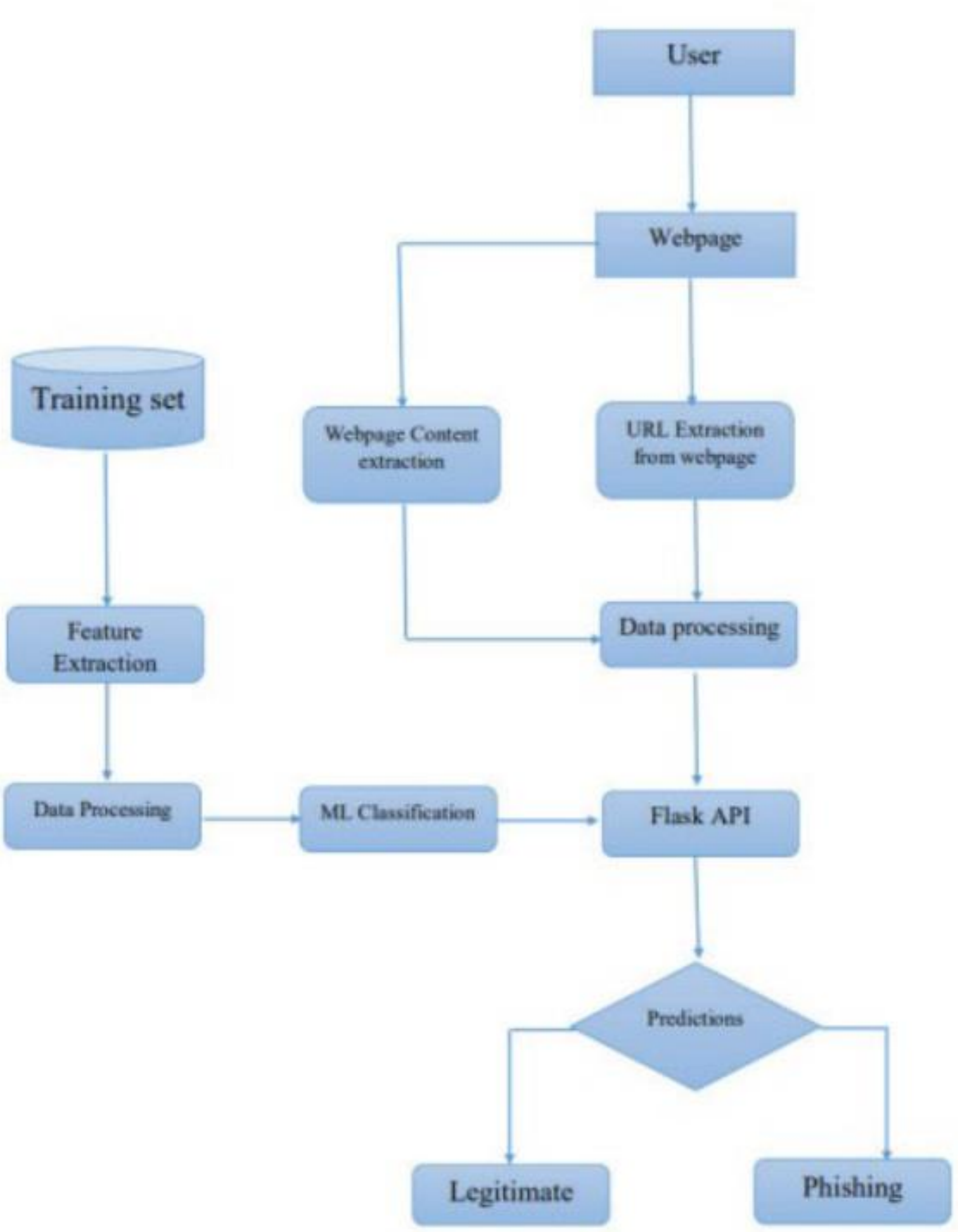
II. Non-Functional requirements



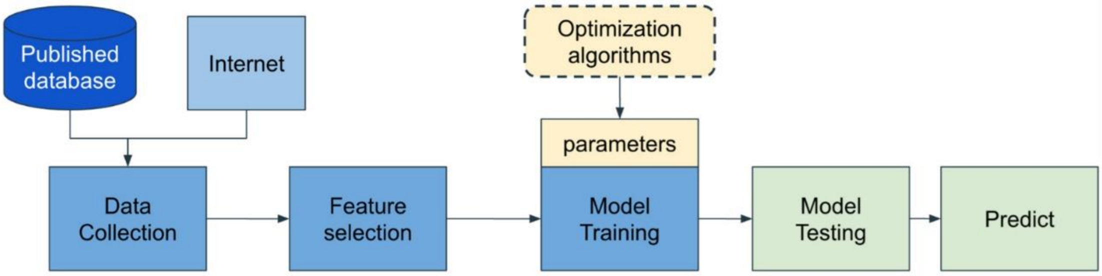
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The user can easily understand the website there is no difficulties in finding the Search Engine.
NFR-2	**Security**	The site is mainly provided for the security process only so there is no possibility for security issues
NFR-3	**Reliability**	All the data processing and prediction are hide to the end users. Showing the positive and negative of the result and it never predict wrongly.
NFR-4	**Performance**	The dataset is used with python and ML algorithms resulting in faster performance
NFR-5	**Availability**	All the basic resources are made available to the end users.
NFR-6	**Scalability**	Multiple URLs can be checked at a given time

5. PROJECT DESIGN

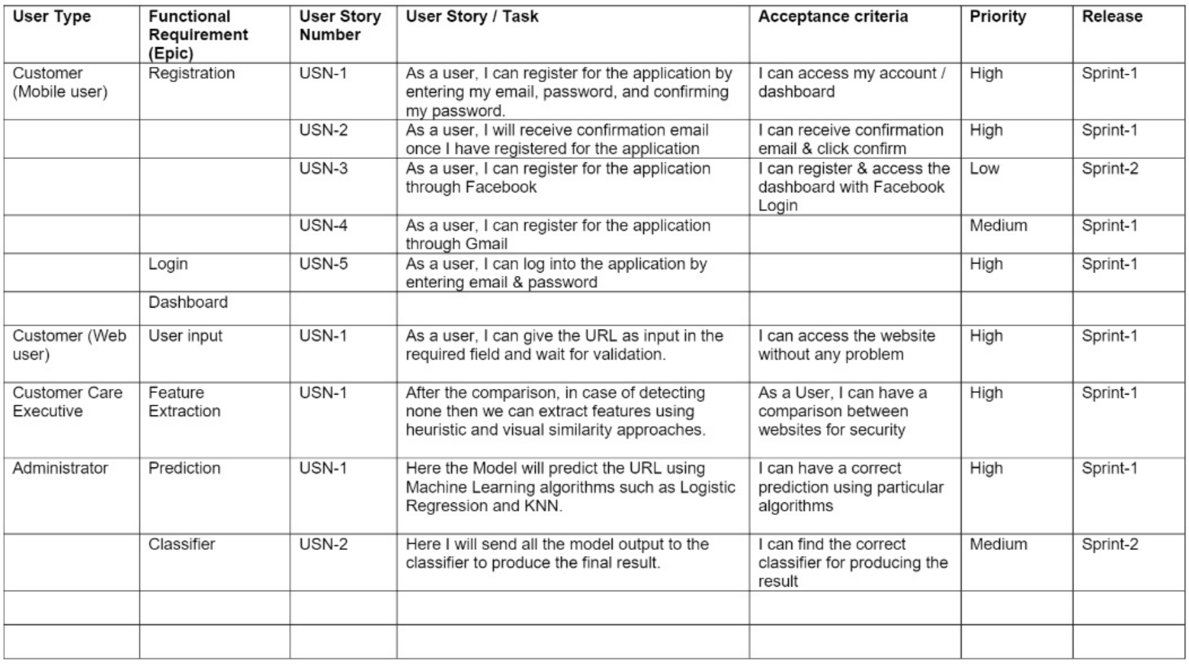
I. Data Flow Diagrams



II. Solution & Technical Architecture



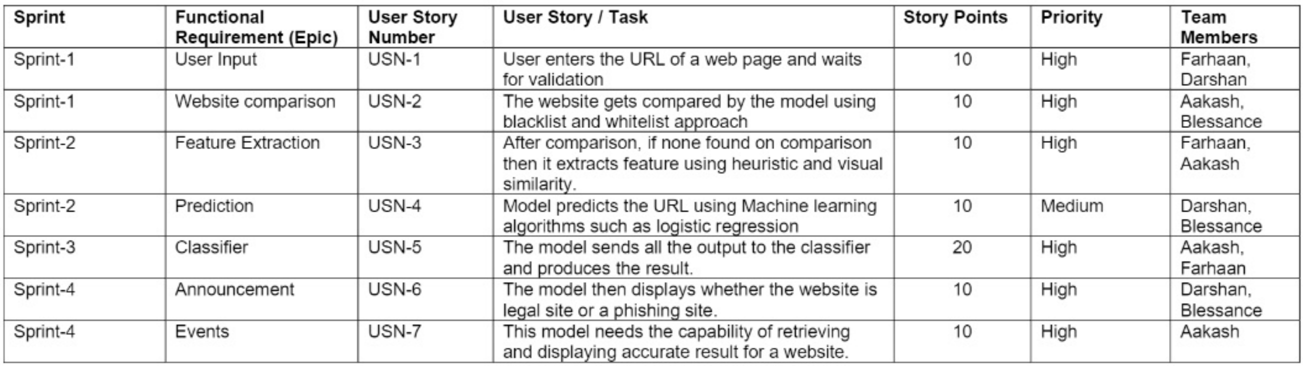
III. User Stories



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 for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register & access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail		Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email & password		High	Sprint-1
	Dashboard					
Customer (Web user)	User input	USN-1	As a user, I can give the URL as input in the required field and wait for validation.	I can access the website without any problem	High	Sprint-1
Customer Care Executive	Feature Extraction	USN-1	After the comparison, in case of detecting none then we can extract features using heuristic and visual similarity approaches.	As a User, I can have a comparison between websites for security	High	Sprint-1
Administrator	Prediction	USN-1	Here the Model will predict the URL using Machine Learning algorithms such as Logistic Regression and KNN.	I can have a correct prediction using particular algorithms	High	Sprint-1
	Classifier	USN-2	Here I will send all the model output to the classifier to produce the final result.	I can find the correct classifier for producing the result	Medium	Sprint-2

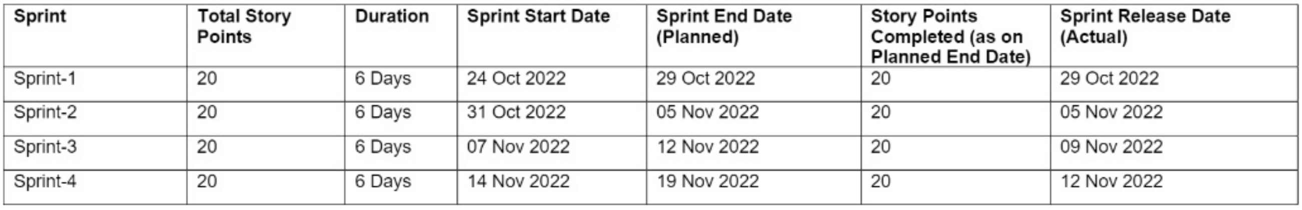
6. PROJECT PLANNING & SCHEDULING

I. Sprint Planning & Estimation



Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	User Input	USN-1	User enters the URL of a web page and waits for validation	10	High	Farhaan, Darshan
Sprint-1	Website comparison	USN-2	The website gets compared by the model using blacklist and whitelist approach	10	High	Aakash, Blessance
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extracts feature using heuristic and visual similarity.	10	High	Farhaan, Aakash
Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic regression	10	Medium	Darshan, Blessance
Sprint-3	Classifier	USN-5	The model sends all the output to the classifier and produces the result.	20	High	Aakash, Farhaan
Sprint-4	Announcement	USN-6	The model then displays whether the website is legal site or a phishing site.	10	High	Darshan, Blessance
Sprint-4	Events	USN-7	This model needs the capability of retrieving and displaying accurate result for a website.	10	High	Aakash

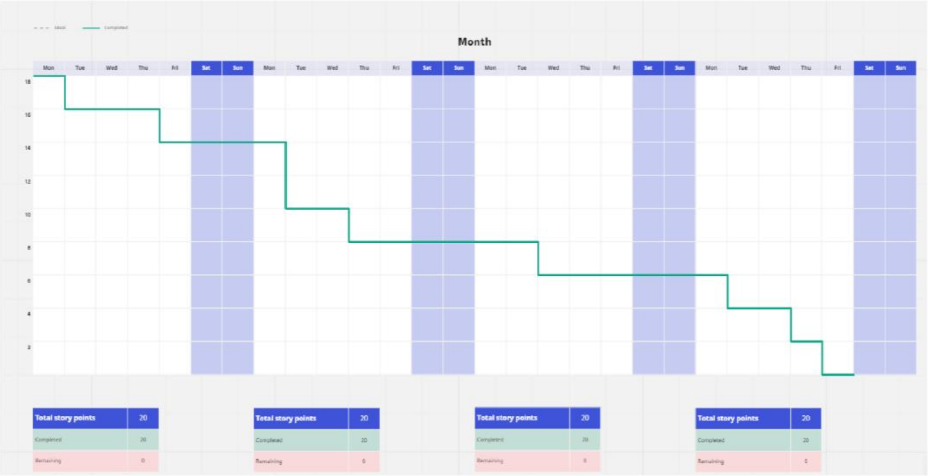
II. Sprint Delivery Schedule

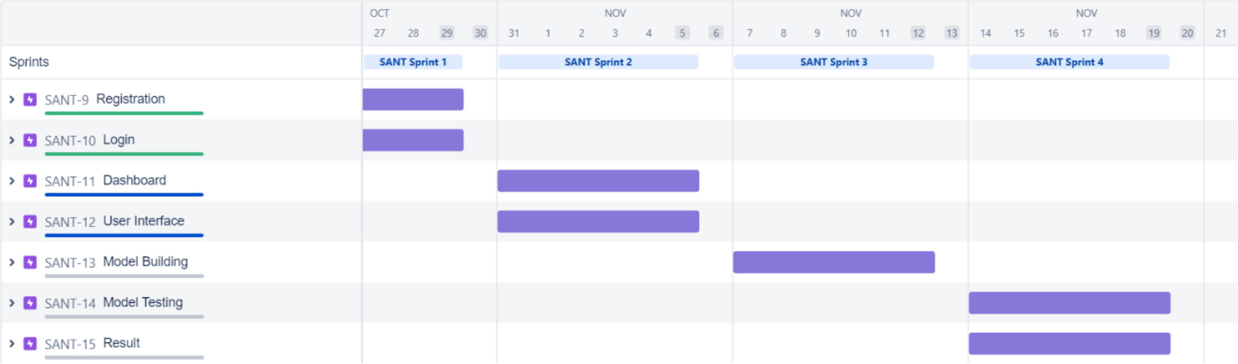


Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	09 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	12 Nov 2022

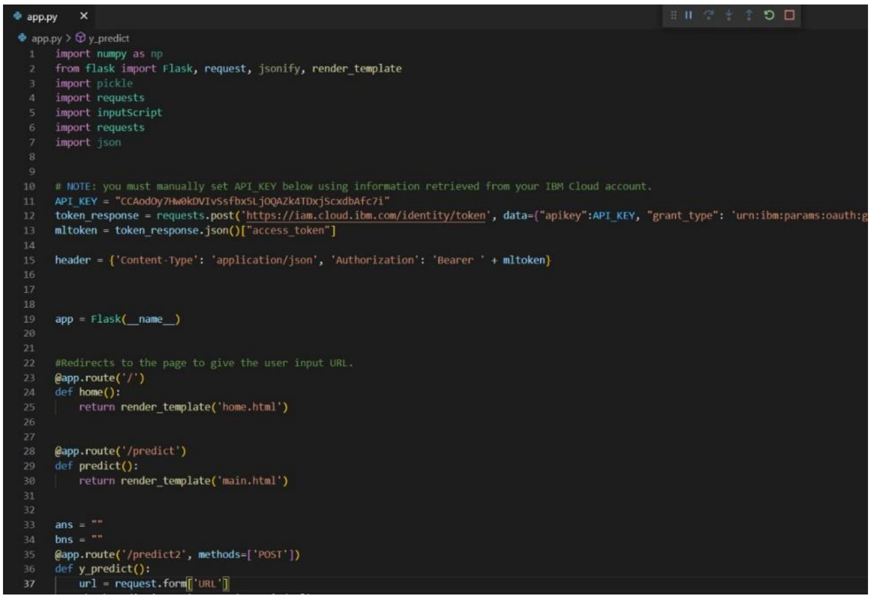
III. Reports from JIRA

Burndown Chart:

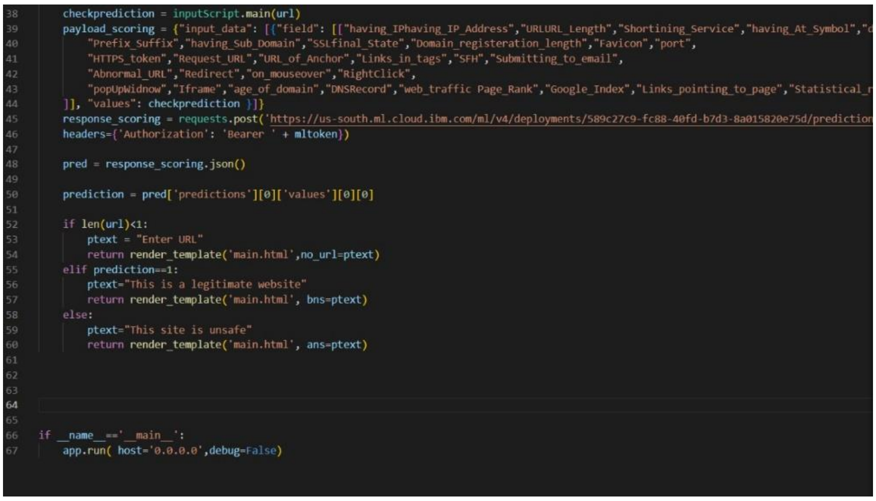




7. CODING & SOLUTIONING



```python
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle
import requests
import inputScript
import requests
import json


# NOTE: you must manually set API_KEY below using information retrieved from your IBM Cloud account.
API_KEY = "CCAodOy7Hw0kDVIvSsfbx5LjOQAZk4TDxjScxdbAfc7i"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token', data={"apikey":API_KEY, "grant_type": 'urn:ibm:params:oauth:g
mltoken = token_response.json()["access_token"]

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' + mltoken}



app = Flask(__name__)


#Redirects to the page to give the user input URL.
@app.route('/')
def home():
    return render_template('home.html')


@app.route('/predict')
def predict():
    return render_template('main.html')


ans = ""
bns = ""
@app.route('/predict2', methods=['POST'])
def y_predict():
    url = request.form['URL']
```



```
    checkprediction = inputScript.main(url)
    payload_scoring = {"input_data": [{"field": [["having_IPhaving_IP_Address","URLURL_Length","Shortining_Service","having_At_Symbol","d
        "Prefix_Suffix","having_Sub_Domain","SSLfinal_State","Domain_registeration_length","Favicon","port",
        "HTTPS_token","Request_URL","URL_of_Anchor","Links_in_tags","SFH","Submitting_to_email",
        "Abnormal_URL","Redirect","on_mouseover","RightClick",
        "popUpWidnow","Iframe","age_of_domain","DNSRecord","web_traffic Page_Rank","Google_Index","Links_pointing_to_page","Statistical_r
    ]], "values": checkprediction }]}
    response_scoring = requests.post('https://us-south.ml.cloud.ibm.com/ml/v4/deployments/589c27c9-fc88-40fd-b7d3-8a015820e75d/prediction
    headers={'Authorization': 'Bearer ' + mltoken})

    pred = response_scoring.json()

    prediction = pred['predictions'][0]['values'][0][0]

    if len(url)<1:
        ptext = "Enter URL"
        return render_template('main.html',no_url=ptext)
    elif prediction==1:
        ptext="This is a legitimate website"
        return render_template('main.html', bns=ptext)
    else:
        ptext="This site is unsafe"
        return render_template('main.html', ans=ptext)





if __name__=='__main__':
    app.run( host='0.0.0.0',debug=False)
```

8. TESTING

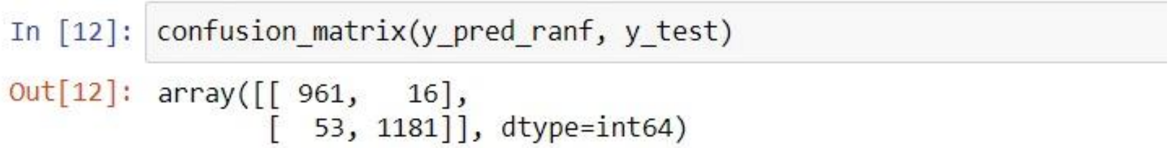
I. Test Cases

- www.youtube.com – Predicted as legitimate website
- www.miniclip.com – Predicted as legitimate website
- https://lbancalnterbank-porinternelt.financiatupres-tamoperu.top/ - Predicted as phishing website

9. RESULTS

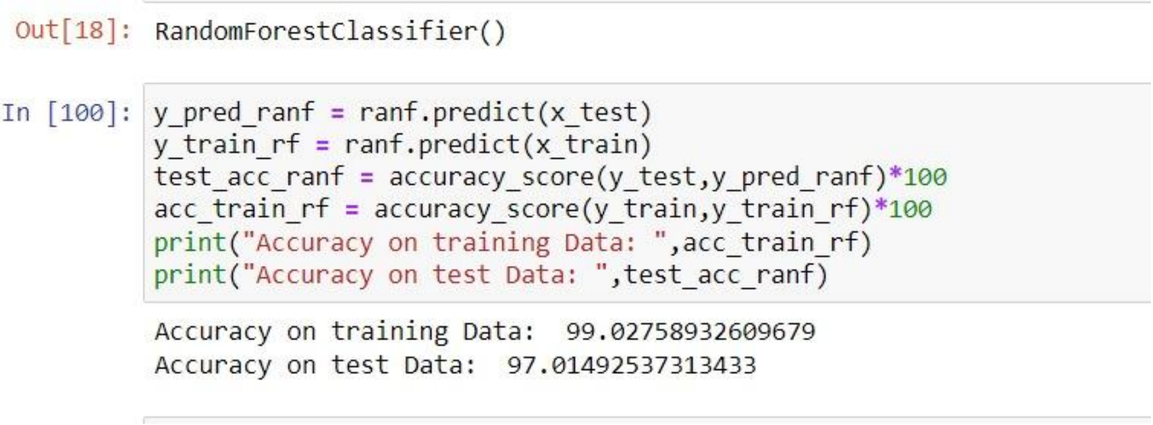
I. Performance Metrics

Confusion Matrix



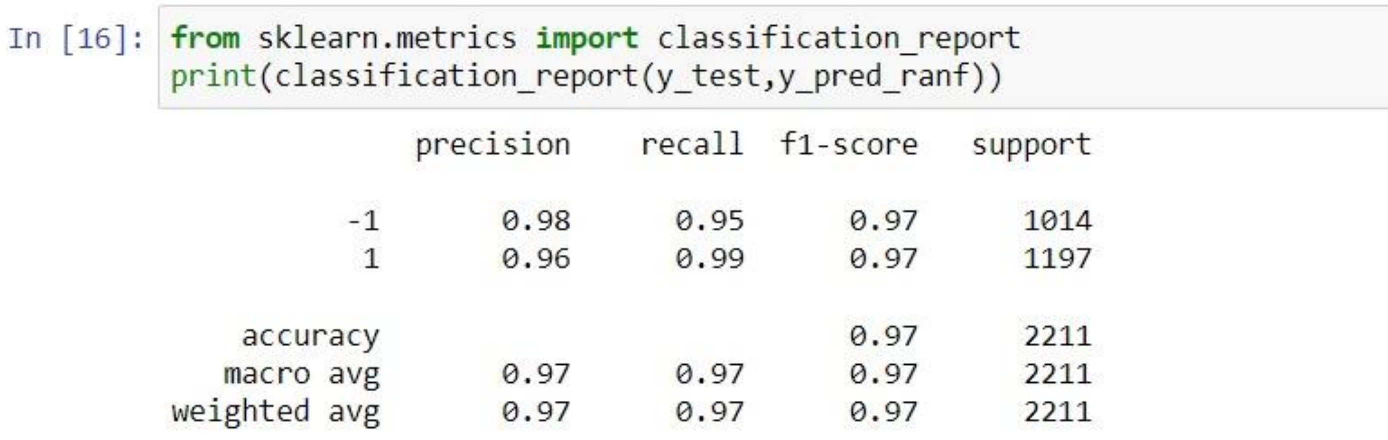
```
In [12]: confusion_matrix(y_pred_ranf, y_test)

Out[12]: array([[ 961,   16],
                [  53, 1181]], dtype=int64)
```

Accuracy Score



```
Out[18]: RandomForestClassifier()

In [100]: y_pred_ranf = ranf.predict(x_test)
          y_train_rf = ranf.predict(x_train)
          test_acc_ranf = accuracy_score(y_test,y_pred_ranf)*100
          acc_train_rf = accuracy_score(y_train,y_train_rf)*100
          print("Accuracy on training Data: ",acc_train_rf)
          print("Accuracy on test Data: ",test_acc_ranf)

          Accuracy on training Data:  99.02758932609679
          Accuracy on test Data:  97.01492537313433
```

Classification Report



```
In [16]: from sklearn.metrics import classification_report
         print(classification_report(y_test,y_pred_ranf))
```

```
              precision    recall  f1-score   support

          -1       0.98      0.95      0.97      1014
           1       0.96      0.99      0.97      1197

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

10. ADVANTAGES & DISADVANTAGES

I. Disadvantages

- Sophistically created phishing websites can show themselves as legitimate websites and bypass detection
-

II. Advantages

- Detects malicious websites
- Prevents users from exposing their credentials to malicious websites
- Helps Non-IT people to stay away from these malicious threats
- Helps elderly people to identify whether a website is trustworthy or not before making any transactions or entering their card details

11. CONCLUSION

The phishing detector we have developed using machine learning modal predicted websites as legitimate or as malicious websites. This way, the existing problem can be overcome by prevention of visiting or exposing credentials to the threats.

12. FUTURE SCOPE

- Detecting well-crafted malicious phishing websites
- Introducing a feature where users can store previously encountered phishing websites in a database which will be linked with their account and used for ready reference in future

13. APPENDIX

GitHub Link - https://github.com/IBM-EPBL/IBM-Project-1952-1658421175

Project Demo Link - https://youtu.be/nDda28ER9nk