Web Phishing Detection
Project Report Documentation

Team ID: PNT2022TMID45373

Team Members :

Vijayaprabu
Logesh
Huttaratulla
Mani muthu

Table of the content:

1. INTRODUCTION:

1.1 Project Overview

PHISHING is a social engineering attack that aims at exploiting the weakness found in system processes as caused by system users. For example, a system can be technically secure enough against password theft, however unaware end users may leak their passwords if an attacker asked them to update their passwords via a given Hypertext Transfer Protocol (HTTP) link, which ultimately threatens the overall security of the system.

Moreover, technical vulnerabilities can be used by attackers to construct far more persuading socially engineered messages. This makes phishing attacks a layered problem, and an effective mitigation would require addressing issues at the technical and human layers. Since phishing attacks aim at exploiting weaknesses found in humans, it is difficult to mitigate them. For example, as evaluated in, end-users failed to detect 29% of phishing attacks even when trained with the best performing user awareness program. On the other hand, software phishing detection techniques are evaluated against bulk phishing attacks, which makes their performance practically unknown with regards to targeted forms of phishing attacks.

1.2 Purpose

The purpose of phishing detection is detecting phishing domain names. therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us.

2. LITERATURE SURVEY

2.1 Existing problem

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.

2.2 References

[1] A. Y. Fu, L. Wenyin, and X. Deng, "Detecting phishing web pages with visual similarity assessment based on earth mover's distance (emd)," *IEEE Trans. Dependable Secur. Comput.*, vol. 3, no. 4, pp. 301–311, Oct. 2006.

[2] B. Krebs, "HBGary Federal hacked by Anonymous," http://krebsonsecurity.com/2011/02/hbgary-federal-hacked-by-anonymous/, 2011, accessed December 2011.

[3] B. Schneier, "Lockheed Martin hack linked to RSA's SecurID breach," http://www.schneier.com/blog/archives/2011/05/lockheed
martin.html, 2011, accessed December 2011.

[4] C. Whittaker, B. Ryner, and M. Nazif, "Large-scale automatic classification of phishing pages," in *NDSS '10*, 2010.

[5] X. Dong, J. Clark, and J. Jacob, "Modelling user-phishing interaction," in *Human System Interactions, 2008 Conference on*, may 2008, pp. 627-632.

2.3 Problem Statement Definition

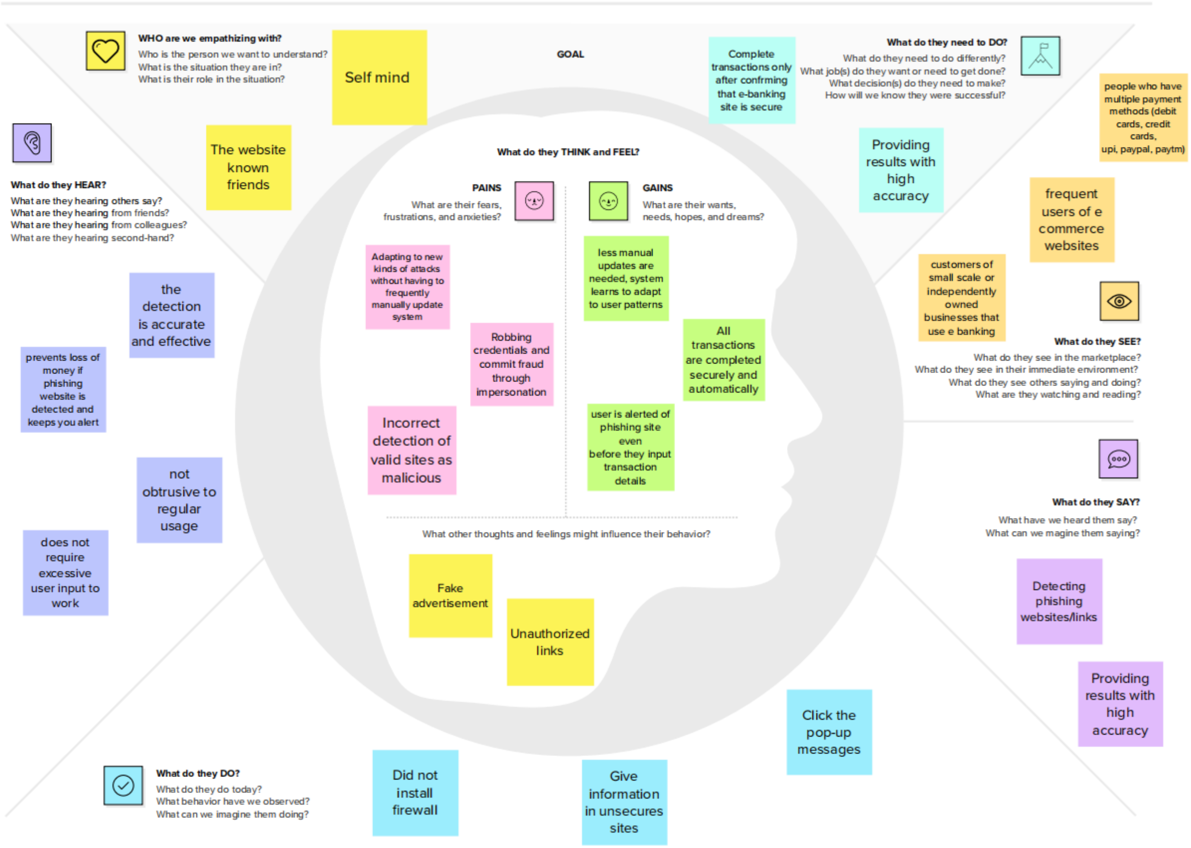
Phishing detection techniques do suffer low detection accuracy and high false alarm especially when novel phishing approaches are introduced. Besides, the most common technique used, blacklist-based method is inefficient in responding to emanating phishing attacks since registering new domain has become 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.

3. **IDEATION & PROPOSED SOLUTION**

3.1 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.

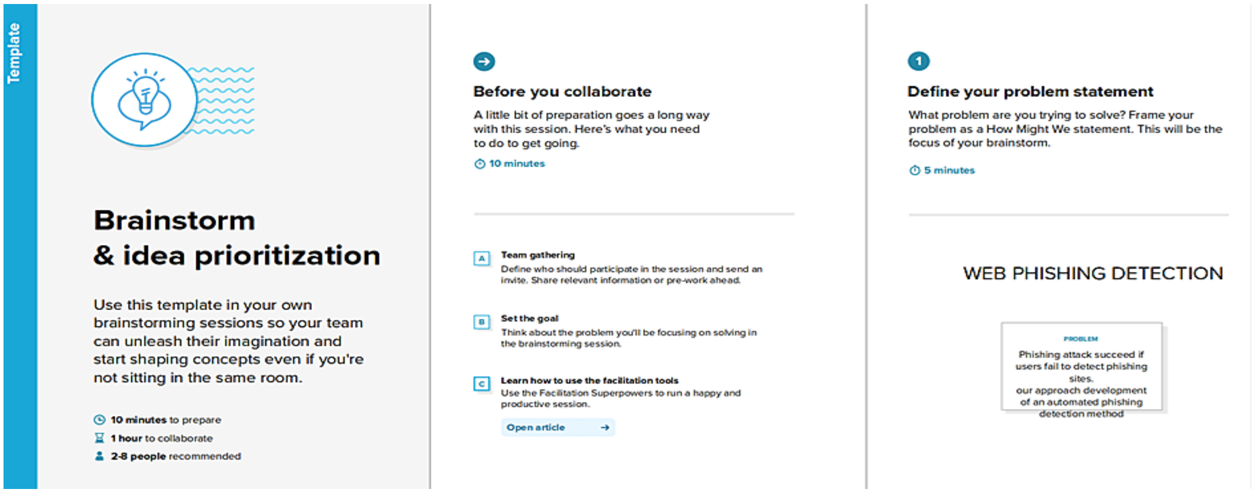
WEB PHISHING DETECTION


WHO are we empathizing with?
Who is the person we want to understand?
What is the situation they are in?
What is their role in the situation?
Self mind
GOAL
Complete transactions only after confrming that e-banking site is secure
What do they need to DO?
What do they need to do differently?
What job(s) do they want or need to get done?
What decision(s) do they need to make?
How will we know they were successful?
people who have multiple payment methods (debit cards, credit cards, upi, paypal, paytm)
Providing results with high accuracy
frequent users of e commerce websites
customers of small scale or independently owned businesses that use e banking
What do they SEE?
What do they see in the marketplace?
What do they see in their immediate environment?
What do they see others saying and doing?
What are they watching and reading?
What do they SAY?
What have we heard them say?
What can we magine them saying?
Detecting phishing websites/links
Providing results with high accuracy
What do they HEAR?
What are they hearing others say?
What are they hearing from friends?
What are they hearing from colleagues?
What are they hearing second-hand?
The website known friends
the detection is accurate and effective
prevents loss of money if phishing website is detected and keeps you alert
not obtrusive to regular usage
does not require excessive user input to work
What do they THINK and FEEL?
PAINS
What are their fears, frustrations, and anxieties?
Adapting to new kinds of attacks without having to frequently manually update system
Robbing credentials and commit fraud through impersonation
Incorrect detection of valid sites as malicious
GAINS
What are their wants, needs, hopes, and dreams?
less manual updates are needed, system learns to adapt to user patterns
All transactions are completed securely and automatically
user is alerted of phishing site even before they input transaction details
What other thoughts and feelings might influence their behavior?
Fake advertisement
Unauthorized links
What do they DO?
What do they do today?
What behavior have we observed?
What can we imagine them doing?
Did not install firewall
Give information in unsecures sites
Click the pop-up messages

3.2 Ideation & Brainstorming

Brainstorming provides a free and open environment that encourages everyone within a team to participate in the creative thinking process that leads to problem solving. Prioritizing volume over value, out-of-the-box ideas are welcome and built upon, and all participants are encouraged to collaborate, helping each other develop a rich amount of creative solutions.

Step-1: Team Gathering, Collaboration and Select the Problem Statement



Template

Brainstorm
& idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

10 minutes to prepare
1 hour to collaborate
2-8 people recommended

Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

10 minutes

A Team gathering
Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.

B Set the goal
Think about the problem you'll be focusing on solving in the brainstorming session.

C Learn how to use the facilitation tools
Use the Facilitation Superpowers to run a happy and productive session.

Open article →

1 Define your problem statement

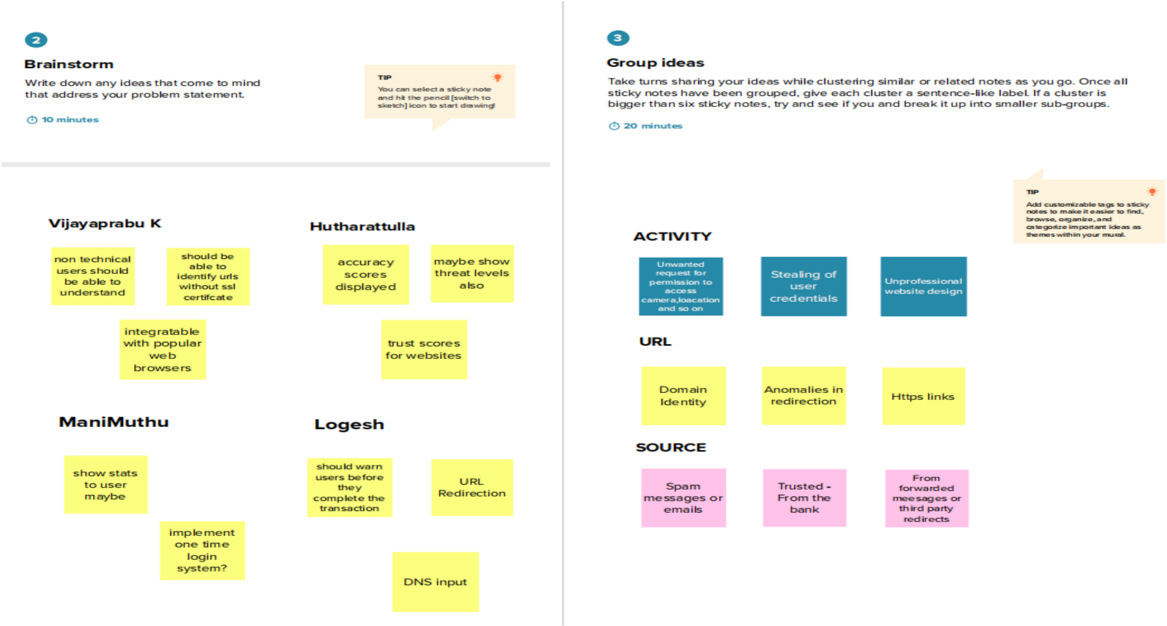
What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

5 minutes

WEB PHISHING DETECTION

PROBLEM
Phishing attack succeed if users fail to detect phishing sites.
our approach development of an automated phishing detection method

Step-2: Brainstorm, Idea Listing and Grouping



2 Brainstorm

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

10 minutes

TIP
You can select a sticky note and hit the pencil [switch to sketch] icon to start drawing!

Vijayaprabu K
- non technical users should be able to understand
- should be able to identify urls without ssl certifcate
- integratable with popular web browsers

Hutharattulla
- accuracy scores displayed
- maybe show threat levels also
- trust scores for websites

ManiMuthu
- show stats to user maybe
- implement one time login system?

Logesh
- should warn users before they complete the transaction
- URL Redirection
- DNS input

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

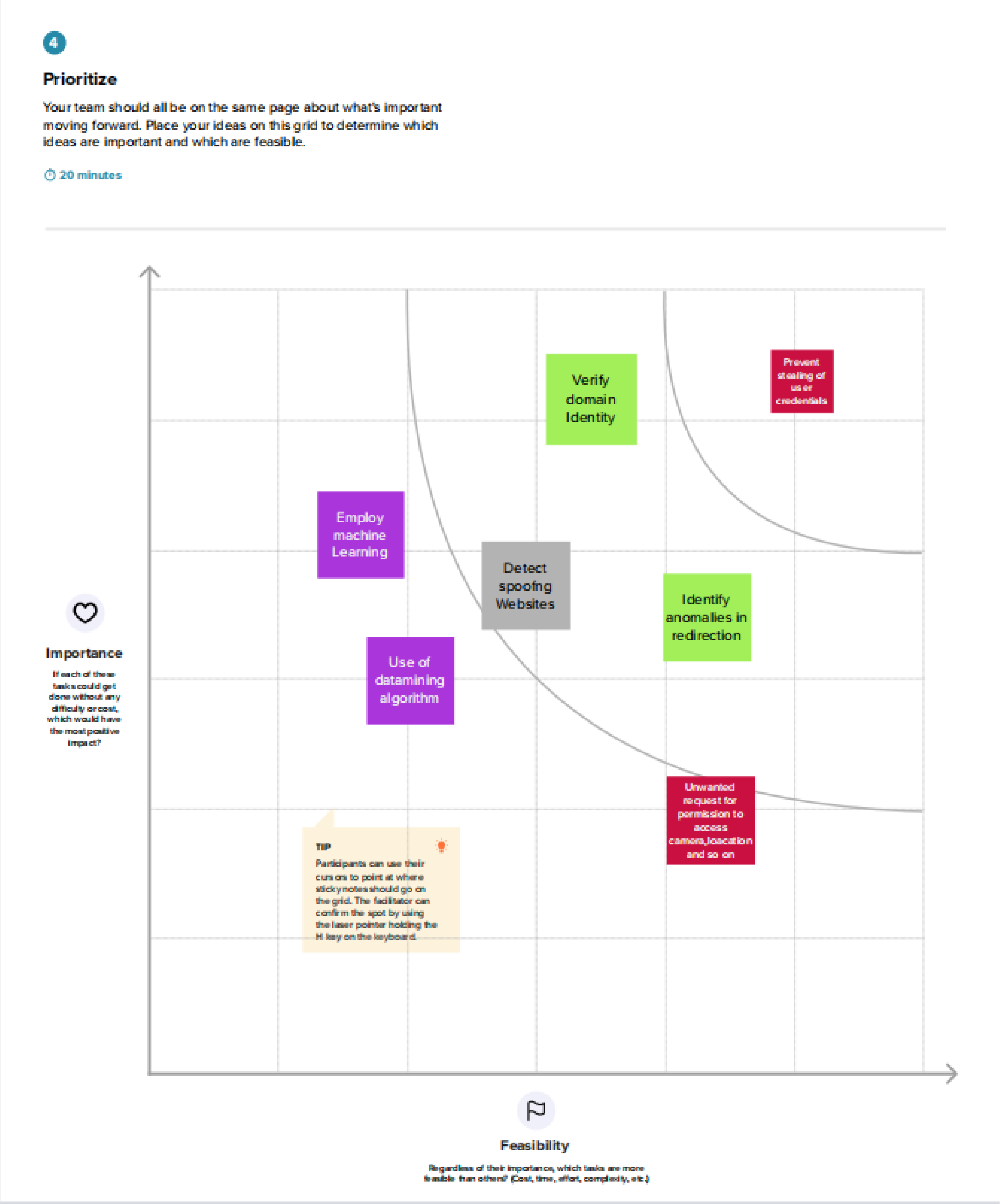
TIP
Add customizable tags to sticky notes to make it easier to find, browse, organize, and categorize important ideas as themes within your mural.

ACTIVITY
- Unwanted request for permission to access camera,loacation and so on
- Stealing of user credentials
- Unprofessional website design

URL
- Domain Identity
- Anomalies in redirection
- Https links

SOURCE
- Spam messages or emails
- Trusted - From the bank
- From forwarded messages or third party redirects

Step-3: Idea Prioritization



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

Verify domain Identity
Prevent stealing of user credentials
Employ machine Learning
Detect spoofng Websites
Identify anomalies in redirection
Use of datamining algorithm
Unwanted request for permission to access camera,loacation and so on

Importance

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

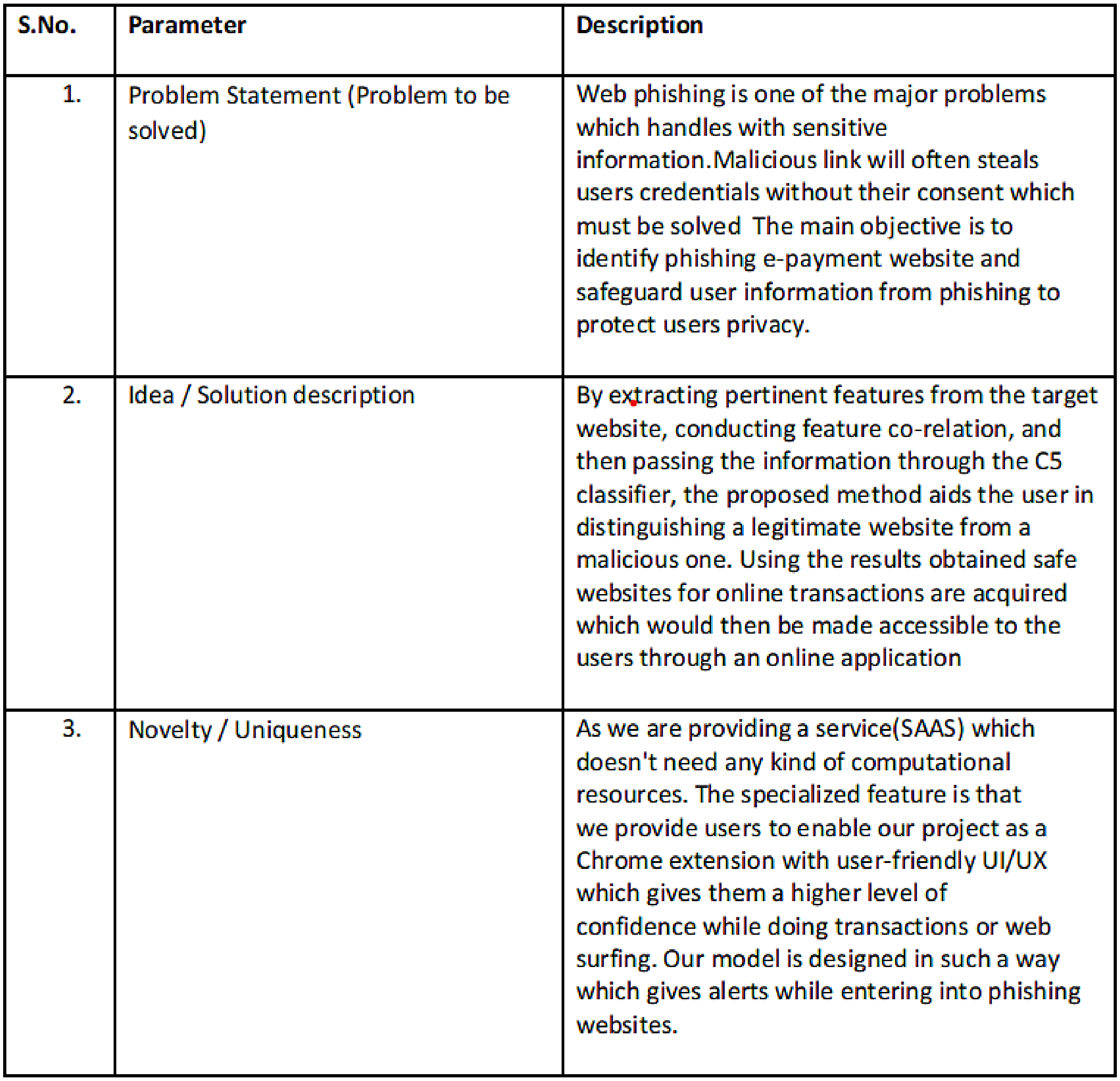
TIP

Participants can use their cursors to point at where sticky notes should go on the grid. The facilitator can confirm the spot by using the laser pointer holding the H key on the keyboard.

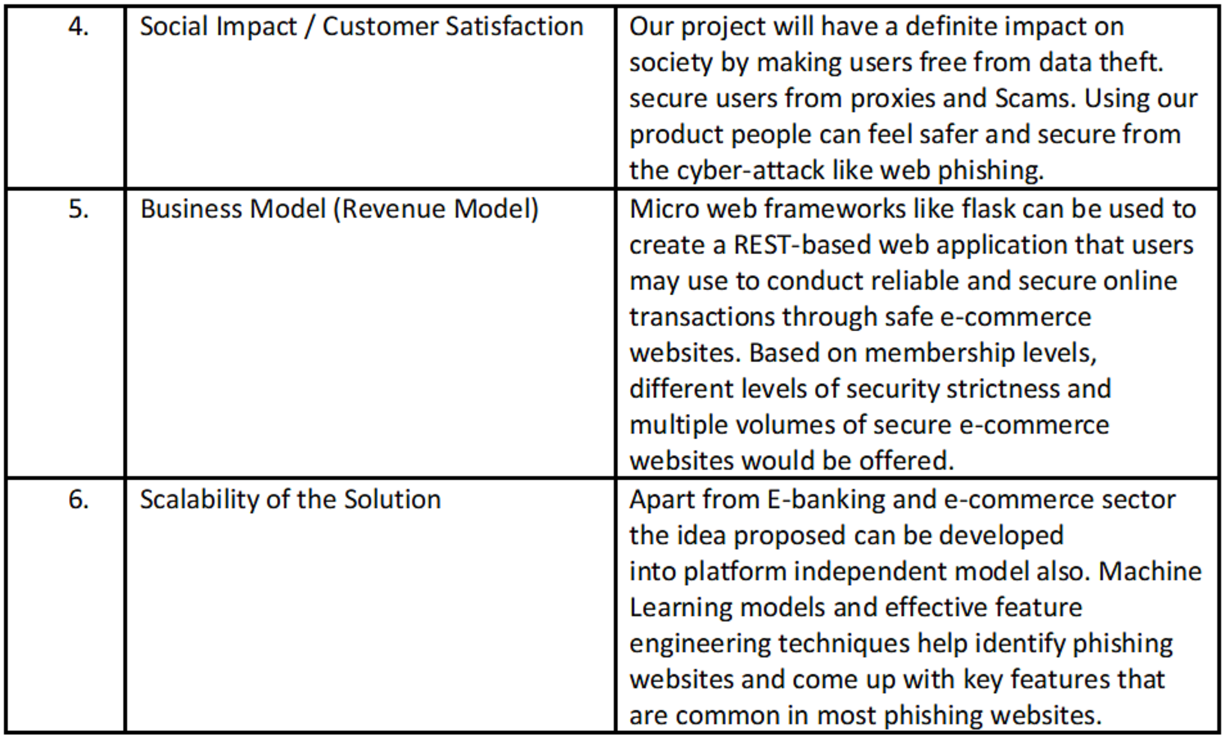
Feasibility

Regardless of their importance, which tasks are more feasible than others? (Cost, time, effort, complexity, etc.)

3.3 Proposed Solution



S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Web phishing is one of the major problems which handles with sensitive information.Malicious link will often steals users credentials without their consent which must be solved The main objective is to identify phishing e-payment website and safeguard user information from phishing to protect users privacy.
2.	Idea / Solution description	By extracting pertinent features from the target website, conducting feature co-relation, and then passing the information through the C5 classifier, the proposed method aids the user in distinguishing a legitimate website from a malicious one. Using the results obtained safe websites for online transactions are acquired which would then be made accessible to the users through an online application
3.	Novelty / Uniqueness	As we are providing a service(SAAS) which doesn't need any kind of computational resources. The specialized feature is that we provide users to enable our project as a Chrome extension with user-friendly UI/UX which gives them a higher level of confidence while doing transactions or web surfing. Our model is designed in such a way which gives alerts while entering into phishing websites.



4.	Social Impact / Customer Satisfaction	Our project will have a definite impact on society by making users free from data theft. secure users from proxies and Scams. Using our product people can feel safer and secure from the cyber-attack like web phishing.
5.	Business Model (Revenue Model)	Micro web frameworks like flask can be used to create a REST-based web application that users may use to conduct reliable and secure online transactions through safe e-commerce websites. Based on membership levels, different levels of security strictness and multiple volumes of secure e-commerce websites would be offered.
6.	Scalability of the Solution	Apart from E-banking and e-commerce sector the idea proposed can be developed into platform independent model also. Machine Learning models and effective feature engineering techniques help identify phishing websites and come up with key features that are common in most phishing websites.

3.4 Problem Solution fit

The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you have realized for it actually solves the customer's problem. It helps entrepreneurs, marketers and corporate innovators identify behavioural patterns and recognize what would work and why

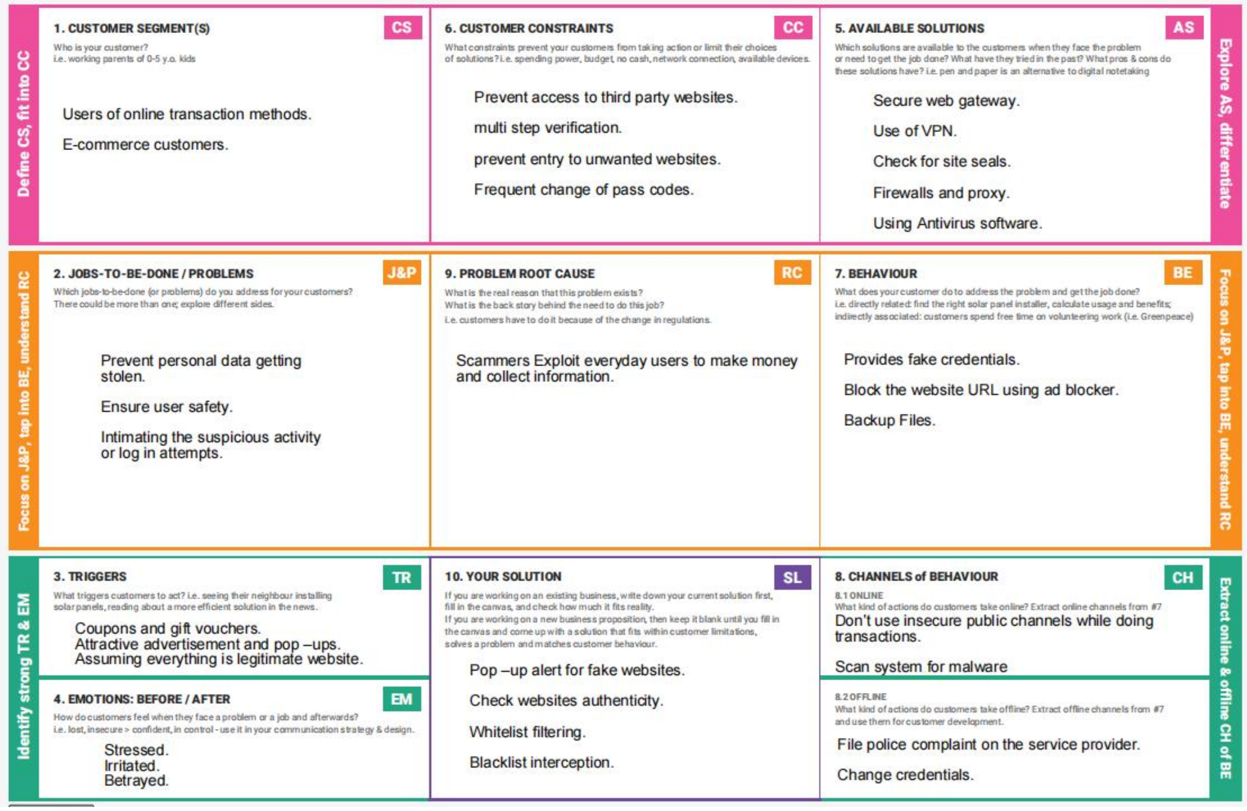
Purpose:

A. Solve complex problems in a way that fits the state of your customers.

B. Succeed faster and increase your solution adoption by tapping into existing mediums
and channels of behaviour.

C. Sharpen your communication and marketing strategy with the right triggers and messaging.

D. Increase touch-points with your company by finding the right problem-behaviour fit and
building trust by solving frequent annoyances, or urgent or costly problems.



Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Who is your customer?
i.e. working parents of 0-5 y.o. kids

Users of online transaction methods.

E-commerce customers.

6. CUSTOMER CONSTRAINTS CC

What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available devices.

Prevent access to third party websites.

multi step verification.

prevent entry to unwanted websites.

Frequent change of pass codes.

5. AVAILABLE SOLUTIONS AS

Which solutions are available to the customers when they face the problem or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking

Secure web gateway.

Use of VPN.

Check for site seals.

Firewalls and proxy.

Using Antivirus software.

Explore AS, differentiate

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS J&P

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

Prevent personal data getting stolen.

Ensure user safety.

Intimating the suspicious activity or log in attempts.

9. PROBLEM ROOT CAUSE RC

What is the real reason that this problem exists?
What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in regulations.

Scammers Exploit everyday users to make money and collect information.

7. BEHAVIOUR BE

What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)

Provides fake credentials.

Block the website URL using ad blocker.

Backup Files.

Focus on J&P, tap into BE, understand RC

Identify strong TR & EM

3. TRIGGERS TR

What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.

Coupons and gift vouchers.
Attractive advertisement and pop –ups.
Assuming everything is legitimate website.

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards?
i.e. lost, insecure > confident, in control - use it in your communication strategy & design.

Stressed.
Irritated.
Betrayed.

10. YOUR SOLUTION SL

If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.
If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.

Pop –up alert for fake websites.

Check websites authenticity.

Whitelist filtering.

Blacklist interception.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE
What kind of actions do customers take online? Extract online channels from #7

Don't use insecure public channels while doing transactions.

Scan system for malware

8.2 OFFLINE
What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.

File police complaint on the service provider.

Change credentials.

Extract online & offline CH of BE

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

A function of software system is defined in functional requirement and the behavior of the system is evaluated when presented with specific inputs or conditions which may include calculations, data manipulation and processing and other specific functionality.

- Our systemshould be able toload air quality data and preprocess data.
- It shouldbe able to analyzethe air quality data.
- It shouldbe able to group data based on hidden patterns.
- It shouldbe able to assigna label based on its data groups.

- It shouldbe able to splitdata into trainsetand testset.
- It shouldbe able to train model using trainset.
- It must validate trainedmodel using testset.
- It shouldbe able to display the trained model accuracy.
- It shouldbe able to accurately predictthe air quality on unseendata.

4.2 Non-Functional requirements

Nonfunctional requirements describe how a system must behave and establish constraints of its functionality. This type of requirements is also known as the system’s *qualityattributes*. Attributes such as performance, security, usability, compatibility are not the feature of the system, they are a required characteristic. They are "developing" properties that emerge from the whole arrangement and hence we can't compose a particular line of code to execute them. Any attributes required by the customer are described by the specification. We must include only those requirements that are appropriate for our project. Some Non-Functional Requirements are as follows:

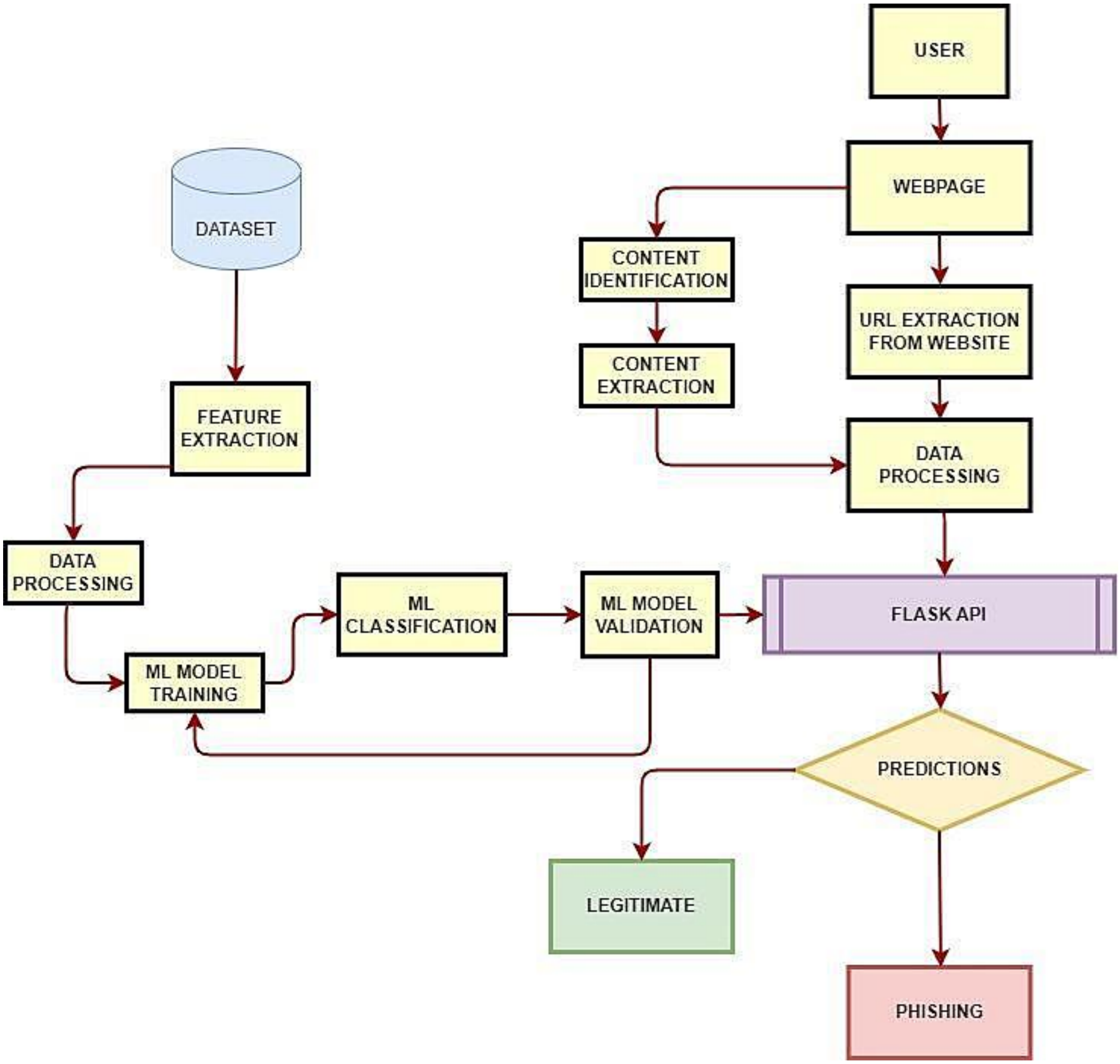
- Reliability
- Maintainability
- Performance

5. PROJECT DESIGN

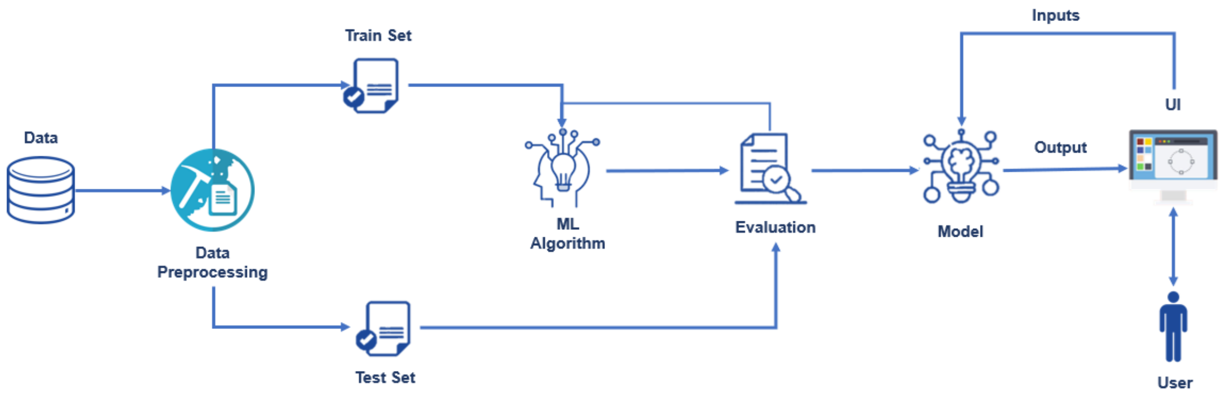
5.1 Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the

right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where
data is stored.



5.2 Solution & Technical Architecture

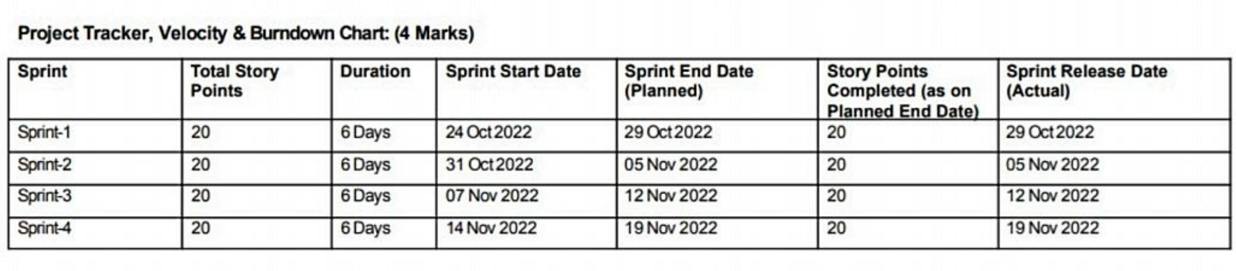


5.3 User Stories

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 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 other platforms.	I can register & access the dashboard through other login platforms	Low	Sprint-2
	Login	USN-4	As a user, I can log into the application by entering email & password	I can login using my respective credentials	High	Sprint-1
	Dashboard	USN-5	As a user, I can navigate the intuitive dashboard to complete the task	Intuitive and easy to use dashboard	High	Sprint-1
Customer (Web user)	Login & Dashboard	USN-6	As a user, I can navigate the application as I did using my mobile. I am able to access the same resources	I can login to the application and access the resources using the dashboard	High	Sprint-1
Customer Care Executive	Login	CCE-1	As a CCE, I can login to the application using the respective credentials and I can interact with the users	I can login using my credentials	High	Sprint-1
	Dashboard	CCE-2	As a CCE, I can view all user queries and respond appropriately	Dashboard displays all the queries and offers response capabilities	High	Sprint-2
Administrator	Login & Dashboard	A-1	As an admin, I can login and manage the activities	I can login and interact with the application's features	High	Sprint-1

6. PROJECT PLANNING & SCHEDULING

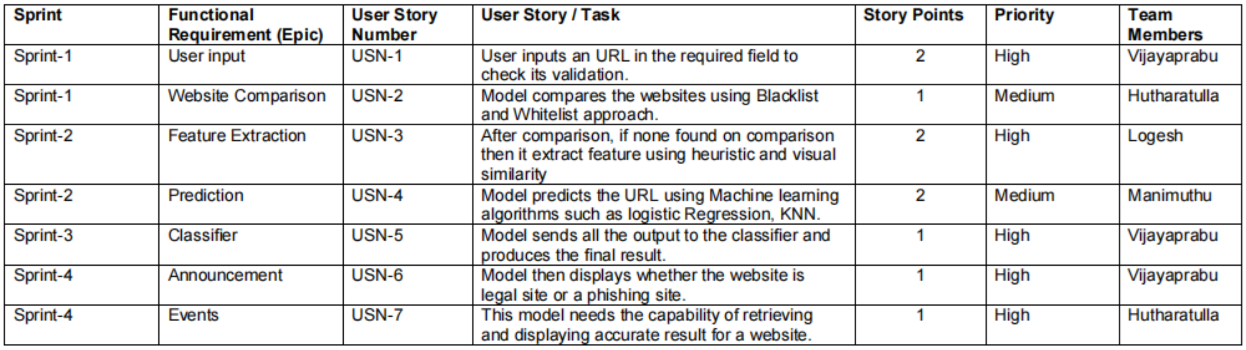
6.1 Sprint Planning & Estimation



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	19 Nov 2022

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.	2	High	Vijayaprabu
Sprint-1	Website Comparison	USN-2	Model compares the websites using Blacklist and Whitelist approach.	1	Medium	Hutharatulla
Sprint-2	Feature Extraction	USN-3	After comparison, if none found on comparison then it extract feature using heuristic and visual similarity	2	High	Logesh
Sprint-2	Prediction	USN-4	Model predicts the URL using Machine learning algorithms such as logistic Regression, KNN.	2	Medium	Manimuthu
Sprint-3	Classifier	USN-5	Model sends all the output to the classifier and produces the final result.	1	High	Vijayaprabu
Sprint-4	Announcement	USN-6	Model then displays whether the website is legal site or a phishing site.	1	High	Vijayaprabu
Sprint-4	Events	USN-7	This model needs the capability of retrieving and displaying accurate result for a website.	1	High	Hutharatulla

7. CODING & SOLUTIONING

7.1 Feature 1

* Login page is used with html code.

* designed index page using css.

* Using flask for app creation.

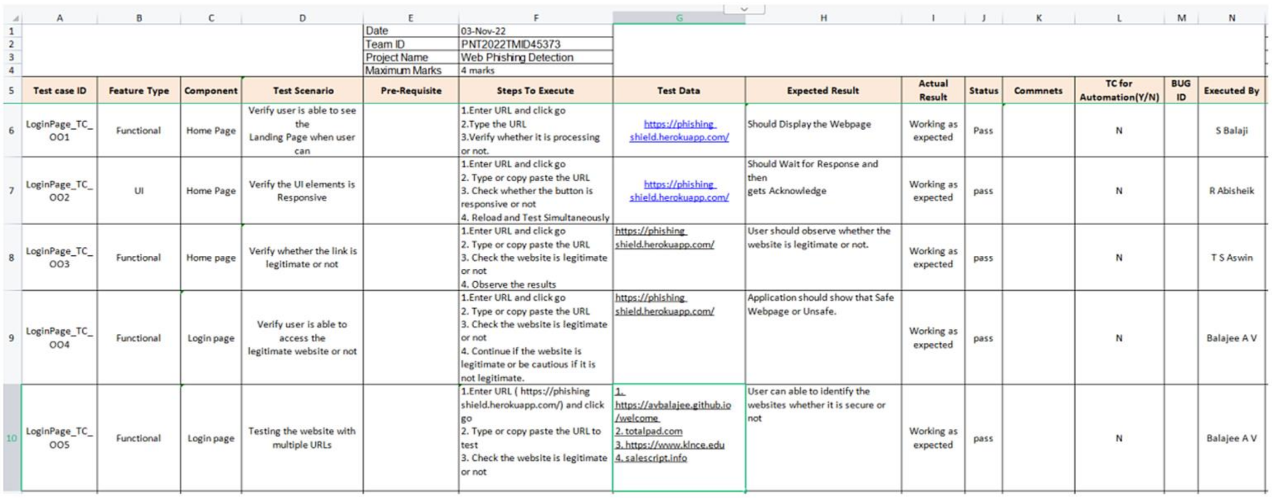
7.2 Feature 2

* User can paste link in the url detecter.

* Then the user can identify the url is safe or not.

8. TESTING

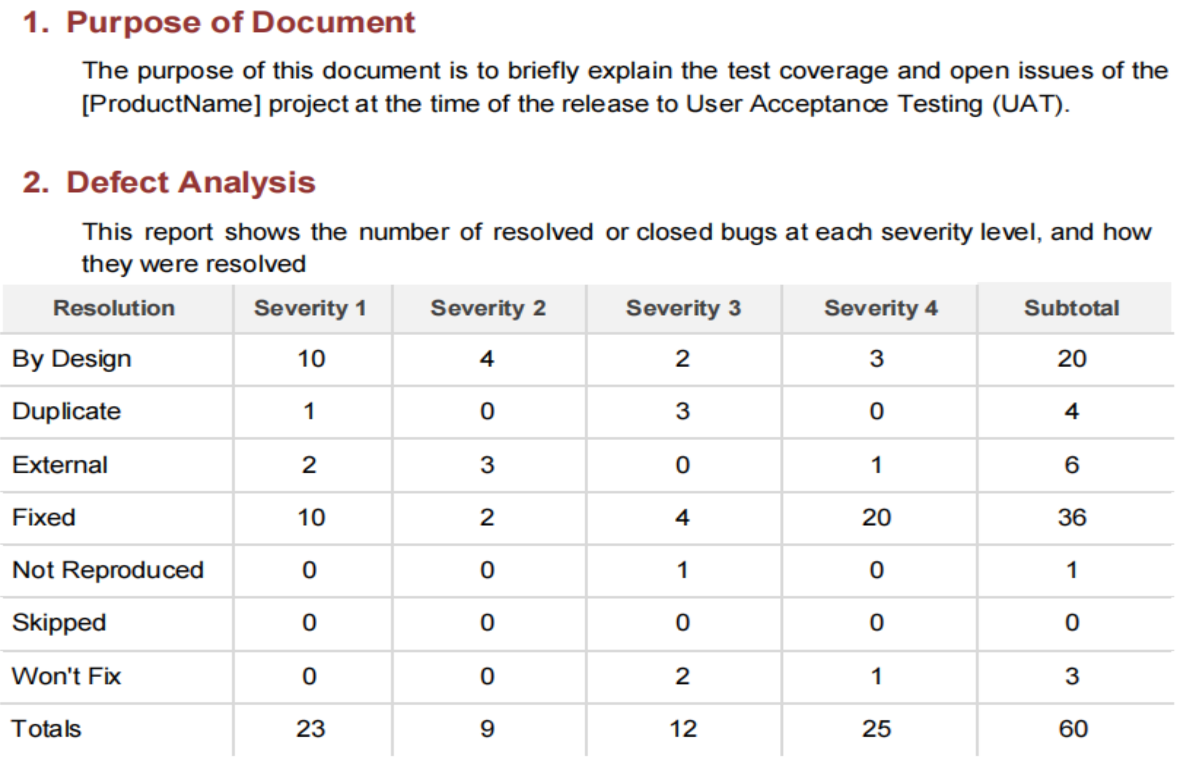
8.1 Test Cases



Date	03-Nov-22
Team ID	PNT2022TMID45373
Project Name	Web Phishing Detection
Maximum Marks	4 marks

	Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Test Data	Expected Result	Actual Result	Status	Commnets	TC for Automation(Y/N)	BUG ID	Executed By
6	LoginPage_TC_OO1	Functional	Home Page	Verify user is able to see the Landing Page when user can		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		S Balaji
7	LoginPage_TC_OO2	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		R Abisheik
8	LoginPage_TC_OO3	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		T S Aswin
9	LoginPage_TC_OO4	Functional	Login 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		Balajee A V
10	LoginPage_TC_OO5	Functional	Login 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	1. https://avbalajee.github.io/welcome 2. totalpad.com 3. https://www.klnce.edu 4. salescript.info	User can able to identify the websites whether it is secure or not	Working as expected	pass		N		Balajee A V

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 [ProductName] 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

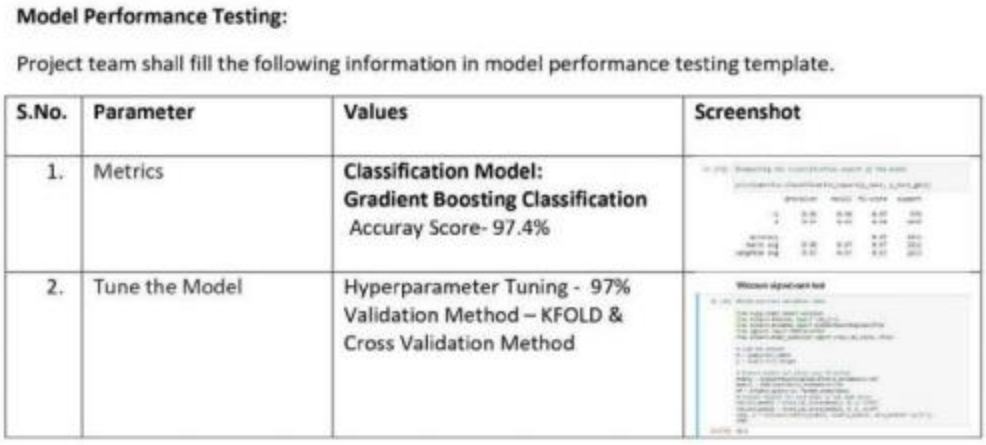
3. Test Case Analysis

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

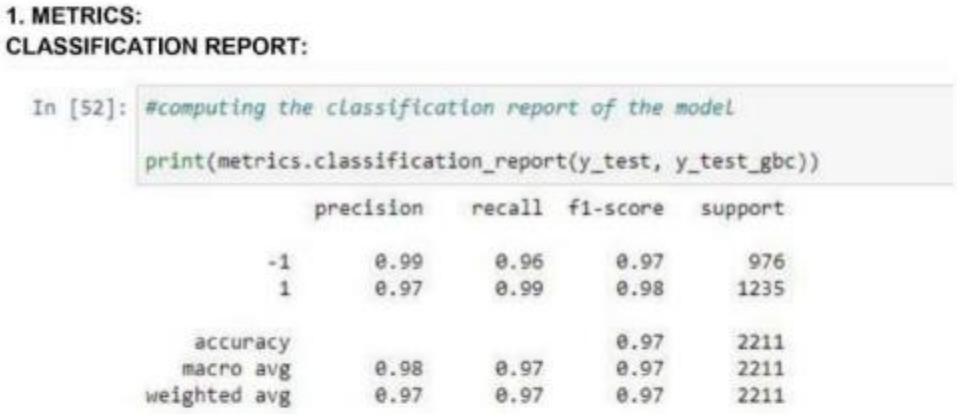
9.1 Performance Metrics



Model Performance Testing:

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

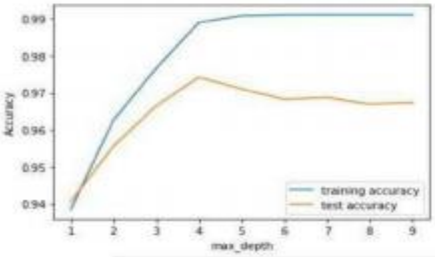
S.No.	Parameter	Values	Screenshot
1.	Metrics	**Classification Model:** **Gradient Boosting Classification** Accuray Score- 97.4%	
2.	Tune the Model	Hyperparameter Tuning - 97% Validation Method – KFOLD & Cross Validation Method	

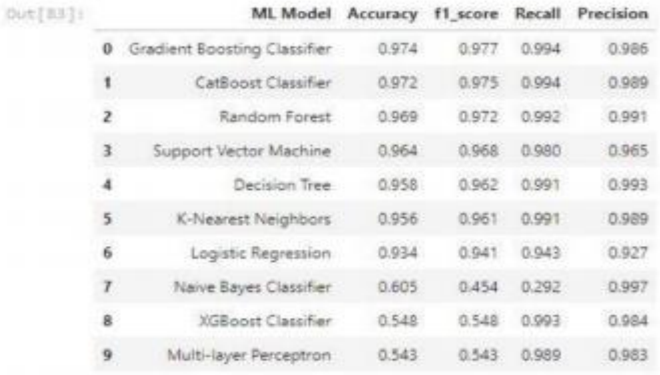


1. METRICS:
CLASSIFICATION REPORT:

```
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
```

PERFORMANCE :

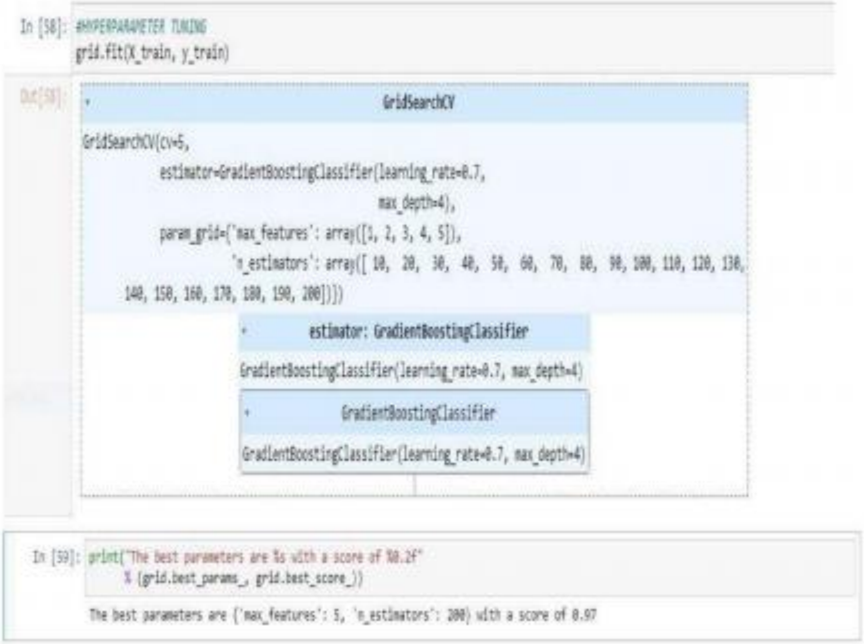




Out[83]:

	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

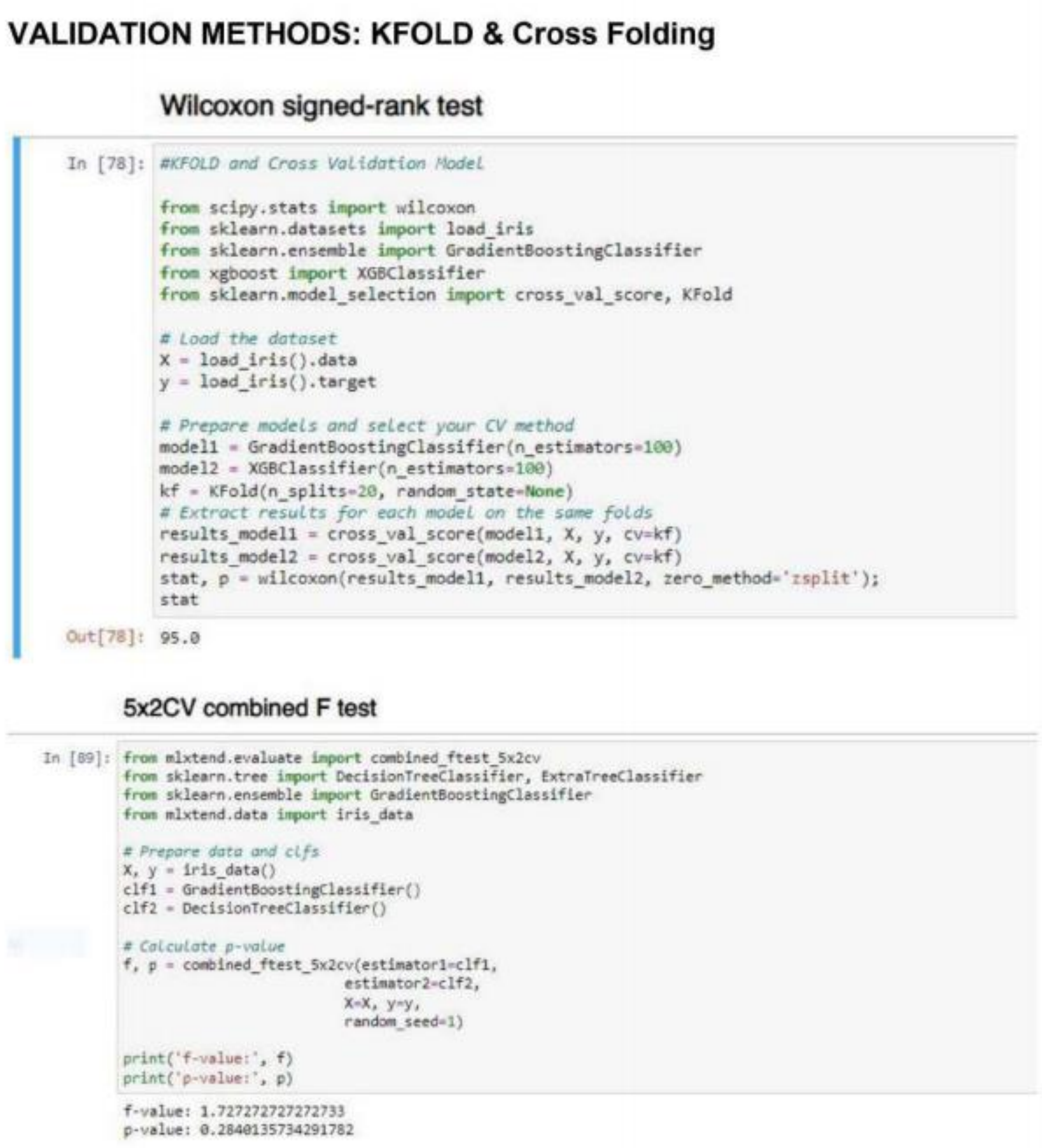


```
In [58]: #HYPERPARAMETER TUNING
         grid.fit(X_train, y_train)
```

```
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: 1.727272727272733
         p-value: 0.2840135734291782
```

10. ADVANTAGES & DISADVANTAGES

Advantages

1. Measure the degrees of corporate and employee vulnerability
2. Eliminate the cyber threat risk level

3. Increase user alertness to phishing risks
4. Instill a cyber security culture and create cyber security heroes
5. Change behavior to eliminate the automatic trust response
6. Deploy targeted anti-phishing solutions
7. Protect valuable corporate and personal data
8. Meet industry compliance obligations
9. Assess the impacts of cyber security awareness training
10. Segment phishing simulation

Disadvantages

1. 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.

2.These effects work together to cause loss of company value, sometimes with irreparable repercussions

3.Phishing emails are frustratingly hard to detect, for humans and digital tools alike. Your best defense is solid training and testing.

11. CONCLUSION

Phishing has becoming a serious network security problem, causing financial loss of billions of dollars to bothconsumers and e-commerce companies. Phishing attacks can be detected through a combination of customerreportage, bounce monitoring, image use monitoring, honey pots and other techniques. Email authenticationtechnologies such as Sender-ID and cryptographic signing, when widely deployed, have the potential to preventphishing emails from reaching users. Personally identifiable information should be included in all email communications. Systems allowing the user to enter or select customized text and imagery are particularly promising.Anti-phishing toolbars are promising tools for identifying phishing sites and heightening security when a potential phishing site is detected. By IPDCM it includes the detection of phishing websitesthrough ensemble classifiers and categorizing the phishing websites according to the various streams as online payments, Banking etc.

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.

13. APPENDIX

Source Code

HTML code:

```
<!DOCTYPE html>
<html lang="en">
<head>
    <center> <h1> Web Phishing Detection  </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="Lokesh P">

    <!-- BootStrap -->
    <link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/css/bootstrap.min.css"
        integrity="sha384-9aIt2nRpC12Uk9gS9baDl411NQApFmC26EwAOH8WgZl5MYYxFfc+NcPb1dKGj7Sk" crossorigin="anonymous">

    <link href="static/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+IbbVYUe
w+OrCXaRkfj"
        crossorigin="anonymous"></script>
    <script
src="https://cdn.jsdelivr.net/npm/popper.js@1.16.0/dist/umd/popper.min
.js"
        integrity="sha384-
Q6E9RHvbIyZFJoft+2mJbHaEWldlvI9IOYy5n3zV9zzTtmI3UksdQRV
voxMfooAo"
        crossorigin="anonymous"></script>
    <script
src="https://stackpath.bootstrapcdn.com/bootstrap/4.5.0/js/bootstrap.min
.js"
        integrity="sha384-
OgVRvuATP1z7JjHLkuOU7Xw704+h835Lr+6QL9UvYjZE3Ipu6Tp75
j7Bh/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> Damon </p> </center>
  </footer>
</html>
```

css code:

```
*,
*::after,
*::before {
  margin: 0;
  padding: 0;
  box-sizing: inherit;
  font-size: 62,5%;
}
.image {
  width: 500px;
  height: 500px;
}
.image-contain {
  object-fit: contain;
  object-position: center;
}

.image-cover {
  object-fit: cover;
  object-position: center;
}
body {
  padding: 10% 5%;
  background: #0f2027;
  background: linear-gradient(to right,#2c5364, #203a43, ##55FFFF);
  justify-content: center;
  align-items: center;
  height: 100vh;
  color: #fff;
}

.form__label {
  font-family: 'Roboto', sans-serif;
  font-size: 1.2rem;
  margin-left: 2rem;
  margin-top: 0.7rem;
  display: block;
  transition: all 0.3s;
  transform: translateY(0rem);
}

.form__input {
  top: -24px;
  font-family: 'Roboto', sans-serif;
  color: #333;
  font-size: 1.2rem;
  padding: 1.5rem 2rem;
  border-radius: 0.2rem;
  background-color: rgb(255, 255, 255);
  border: none;
  width: 75%;
  display: block;
  border-bottom: 0.3rem solid transparent;
  transition: all 0.3s;
}
```

```
.form__input:placeholder-shown + .form__label {
  opacity: 0;
  visibility: hidden;
  -webkit-transform: translateY(+4rem);
  transform: translateY(+4rem);
}


.button {
  appearance: button;
  background-color: transparent;
  background-image: linear-gradient(to bottom, #fff, #f8eedb);
  border: 0 solid #e5e7eb;
  border-radius: .5rem;
  box-sizing: border-box;
  color: #482307;
  column-gap: 1rem;
  cursor: pointer;
  display: flex;
  font-family: ui-sans-serif,system-ui,-apple-system,system-ui,"Segoe UI",Roboto,"Helvetica Neue",Arial,"Noto Sans",sans-serif,"Apple Color Emoji","Segoe UI Emoji","Segoe UI Symbol","Noto Color Emoji";
  font-size: 100%;
  font-weight: 700;
  line-height: 24px;
  margin: 0;
  outline: 2px solid transparent;
  padding: 1rem 1.5rem;
  text-align: center;
  text-transform: none;
  transition: all .1s cubic-bezier(.4, 0, .2, 1);
  user-select: none;
  -webkit-user-select: none;
  touch-action: manipulation;
  box-shadow: -6px 8px 10px rgba(81,41,10,0.1),0px 2px 2px rgba(81,41,10,0.2);
}

.button:active {
  background-color: #f3f4f6;
  box-shadow: -1px 2px 5px rgba(81,41,10,0.15),0px 1px 1px rgba(81,41,10,0.15);
  transform: translateY(0.125rem);
}
```

```
.button:focus {
  box-shadow: rgba(72, 35, 7, .46) 0 0 0 4px, -6px 8px 10px
rgba(81,41,10,0.1), 0px 2px 2px rgba(81,41,10,0.2);
}



.main-body{
  display: flex;
  flex-direction: row;
  width: 75%;
  justify-content:space-around;
}

.button1{
  appearance: button;
  background-color: transparent;
  background-image: linear-gradient(to bottom, rgb(160, 245, 174),
#37ee65);
  border: 0 solid #e5e7eb;
  border-radius: .5rem;
  box-sizing: border-box;
  color: #482307;
  column-gap: 1rem;
  cursor: pointer;
  display: flex;
  font-family: ui-sans-serif,system-ui,-apple-system,system-ui,"Segoe UI",Roboto,"Helvetica Neue",Arial,"Noto Sans",sans-serif,"Apple Color Emoji","Segoe UI Emoji","Segoe UI Symbol","Noto Color Emoji";
  font-size: 100%;
  font-weight: 700;
  line-height: 24px;
  margin: 0;
  outline: 2px solid transparent;
  padding: 1rem 1.5rem;
  text-align: center;
  text-transform: none;
  transition: all .1s cubic-bezier(.4, 0, .2, 1);
  user-select: none;
  -webkit-user-select: none;
  touch-action: manipulation;
  box-shadow: -6px 8px 10px rgba(81,41,10,0.1),0px 2px 2px rgba(81,41,10,0.2);
  display: none;
}

.button2{
```

```css
  appearance: button;
  background-color: transparent;
  background-image: linear-gradient(to bottom, rgb(252, 162, 162), #ee3737);
  border: 0 solid #e5e7eb;
  border-radius: .5rem;
  box-sizing: border-box;
  color: #482307;
  column-gap: 1rem;
  cursor: pointer;
  display: flex;
  font-family: ui-sans-serif,system-ui,-apple-system,system-ui,"Segoe UI",Roboto,"Helvetica Neue",Arial,"Noto Sans",sans-serif,"Apple Color Emoji","Segoe UI Emoji","Segoe UI Symbol","Noto Color Emoji";
  font-size: 100%;
  font-weight: 700;
  line-height: 24px;
  margin: 0;
  outline: 2px solid transparent;
  padding: 1rem 1.5rem;
  text-align: center;
  text-transform: none;
  transition: all .1s cubic-bezier(.4, 0, .2, 1);
  user-select: none;
  -webkit-user-select: none;
  touch-action: manipulation;
  box-shadow: -6px 8px 10px rgba(81,41,10,0.1),0px 2px 2px
rgba(81,41,10,0.2);
  display: none;
}

.right {
  right: 0px;
  width: 300px;
}

@media (max-width: 576px) {
  .form {
    width: 100%;
  }
 }
.abc{
  width: 50%;
}
```

IBM Cloud Deployment:

```
from flask import Flask, request, render_template
import numpy as np
import pandas as pd
from sklearn import metrics
import warnings
import pickle


import requests

# NOTE: you must manually set API_KEY below using information retrieved from
your IBM Cloud account.
API_KEY = "cWGD5yTjEpEGtqPpvHPDBElN5eXFS7eh2JRDyUWhySMW"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token',
data={"apikey":
 API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

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

# NOTE: manually define and pass the array(s) of values to be scored in the next
line
payload_scoring = {"input_data": [{"field":
[["UsingIP","LongURL","ShortURL","Symbol@","Redirecting//","PrefixSuffix-
","SubDomains","HTTPS","DomainRegLen","Favicon","NonStdPort","HTTPSDo
mainURL","RequestURL","AnchorURL","LinksInScriptTags","ServerFormHandl
er","InfoEmail","AbnormalURL","WebsiteForwarding","StatusBarCust","Disable
RightClick","UsingPopupWindow","IframeRedirection","AgeofDomain","DNSRe
cording","WebsiteTraffic","PageRank","GoogleIndex","LinksPointingToPage","St
atsReport"
]], "values": [[1,1,1,1,1,-1,-1,-1,-1,1,1,1,1,-1,-1,1,1,1,0,1,1,1,1,-1,-1,-1,-1,1,0,1]]}]}

response_scoring = requests.post('https://us-
south.ml.cloud.ibm.com/ml/v4/deployments/084b5c52-f617-40ef-a0e8-
3e6cf79ae447/predictions?version=2022-11-06', json=payload_scoring,
 headers={'Authorization': 'Bearer ' + mltoken})
print("Scoring response")
predictions=response_scoring.json()
#print(predictions)
pred=print(predictions['predictions'][0]['values'][0][0])
if(pred != 1):
#if secure print this statement
    print("The Website is secure.. Continue")
else:
#if  not secure print this statement
    print("The Website is not Legitimate... BEWARE!!")
```

Integrating Flask with IBM cloud:

```
#importing required libraries

from flask import Flask, request, render_template
import numpy as np
import pandas as pd
from sklearn import metrics
```

```
import warnings
import pickle
import requests
warnings.filterwarnings('ignore')
from feature import FeatureExtraction

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

# NOTE: you must manually set API_KEY below using information retrieved
from your IBM Cloud account.
API_KEY = "_GTlDRru34jJmAn-oPJwyytYz0reQa0sR-UcO8Ux0bRx"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token',
data={"apikey":
 API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]

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


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)
        payload_scoring = {"input_data": [{"field":
[["UsingIP","LongURL","ShortURL","Symbol@","Redirecting//","PrefixSuffix-
","SubDomains","HTTPS","DomainRegLen","Favicon","NonStdPort","HTTPSDo
mainURL","RequestURL","AnchorURL","LinksInScriptTags","ServerFormHandl
er","InfoEmail","AbnormalURL","WebsiteForwarding","StatusBarCust","Disable
RightClick","UsingPopupWindow","IframeRedirection","AgeofDomain","DNSRe
cording","WebsiteTraffic","PageRank","GoogleIndex","LinksPointingToPage","St
atsReport"
]], "values": [[1,1,1,1,1,-1,-1,-1,-1,1,1,1,1,-1,-1,1,1,1,0,1,1,1,1,-1,-1,-1,-1,1,0,1]]}]}
        response_scoring = requests.post('https://us-
south.ml.cloud.ibm.com/ml/v4/deployments/084b5c52-f617-40ef-a0e8-
3e6cf79ae447/predictions?version=2022-11-06', json=payload_scoring,
        headers={'Authorization': 'Bearer ' + mltoken})
        print("Scoring response")
        predictions=response_scoring.json()
#print(predictions)
        pred=print(predictions['predictions'][0]['values'][0][0])
        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=2020)
```

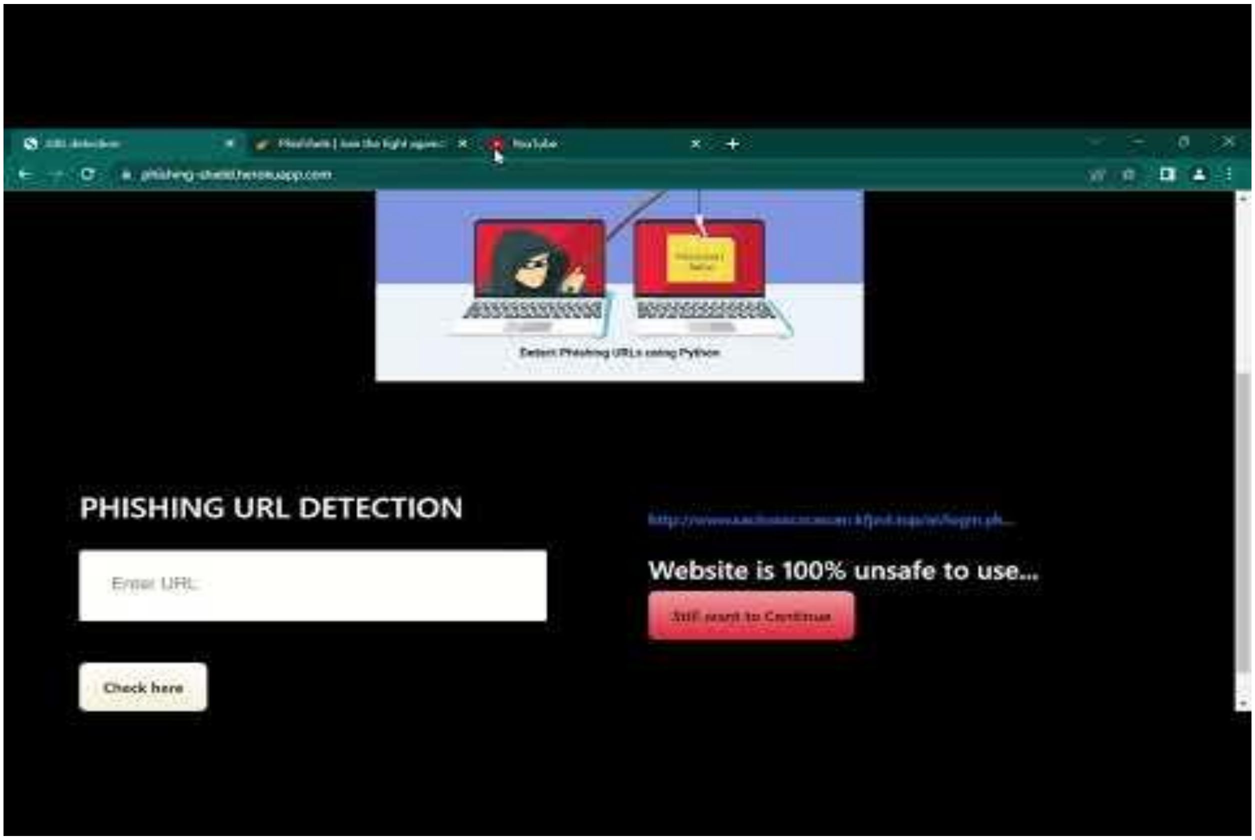
GitHub & Project Demo Link

Github link:

https://github.com/IBM-EPBL/IBM-Project-36046-1660292166

Project Demo Link:

https://youtu.be/adNQuRtZiM8


URL detection
PhishTank | Join the fight agains
YouTube
phishing-shield.herokuapp.com
Detect Phishing URLs using Python
Enter URL
Check here
Still want to Continue

PHISHING URL DETECTION

[illegible]

Website is 100% unsafe to use...