1. INTRODUCTION

1.1 Project Overview

The world is now being digitalized and the internet is playing the best part in it.

Though the internet is making things easier and secured there is also other side of internet which is a threat to the user. So there are many possible ways to steal the user's personal information. So the developers made up a research on the various domains and picked the top notch technologies. The machine learning domain has the various opportunities to protect the data from being stealing it. The web phishing is the way in which the personal data is stolen. So Web phishing detection project is the method to reduce the risk of website's security when u surf in.

1.2 Project Purpose

The main purpose of this project is to avoid personal information to be stealed by the fraudulent websites. The Website is been build where the user is to be given a website link, the machine learning model we built will give the percentage of genunity the website you access.

2. LITERATURE SURVEY

According to this paper we people are highly dependent on the internet. For performing online shopping and online activities like banking, mobile recharge and more activities are done only through internet. Here phishing is nothing but a type of website threat which illegally collects the original website information such as login id, password and credit card information. Here we will use an efficient machine learning based web phishing detection technique

2.1 Existing Problem

There are many users who purchase products through online platform and the payment is done through e-banking. There are some fake banking websites in which they collect the more sensitive information like username, password, credit card details etc , for illegal purpose. This type of websites are called phishing website. Here web phishing is one of the security threat to web services on the internet.

2.2 References

[1] Higashi no, M., et al. An Anti-phishing Training System for Security Awareness and Education
Considering Prevention of Information Leakage. in 2019 5th International Conference on Information
Management (ICIM). 2019.
[2] H. Bleau, Global Fraud and Cybercrime Forecast,. 2017.
[3] Michel Lange, V., et al., Planning and production of grammatical and lexical verbs in multi-word

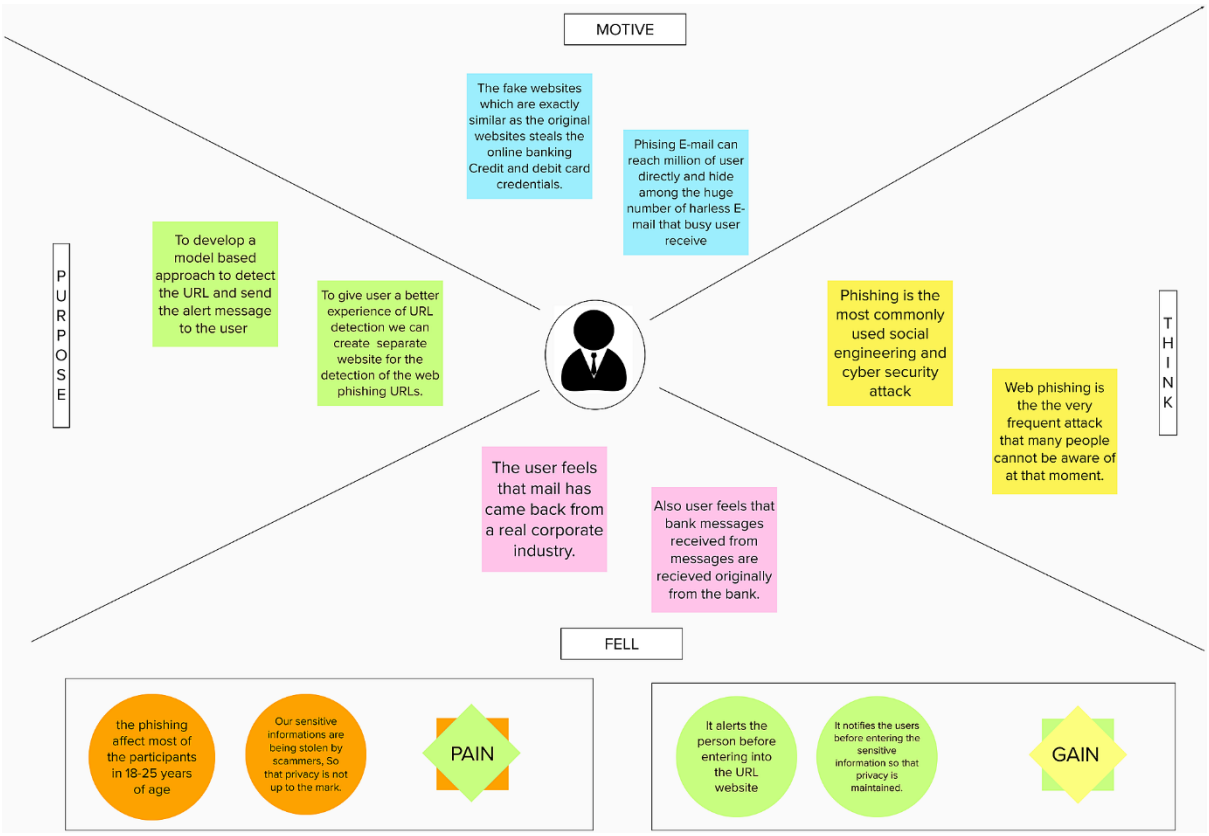
2.3 Problem Statement Definition

To overcome the problem of phishing website whenever we are clicking on one website it must show an alert box like it is a secure website or it is not a secure website. Then another way is that we can scan the website in order to prevent our system or mobile from the phishing attack. Even though technologies are there we as the user have to be aware of the websites whether it is secure or not. We should not click any unwanted websites.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

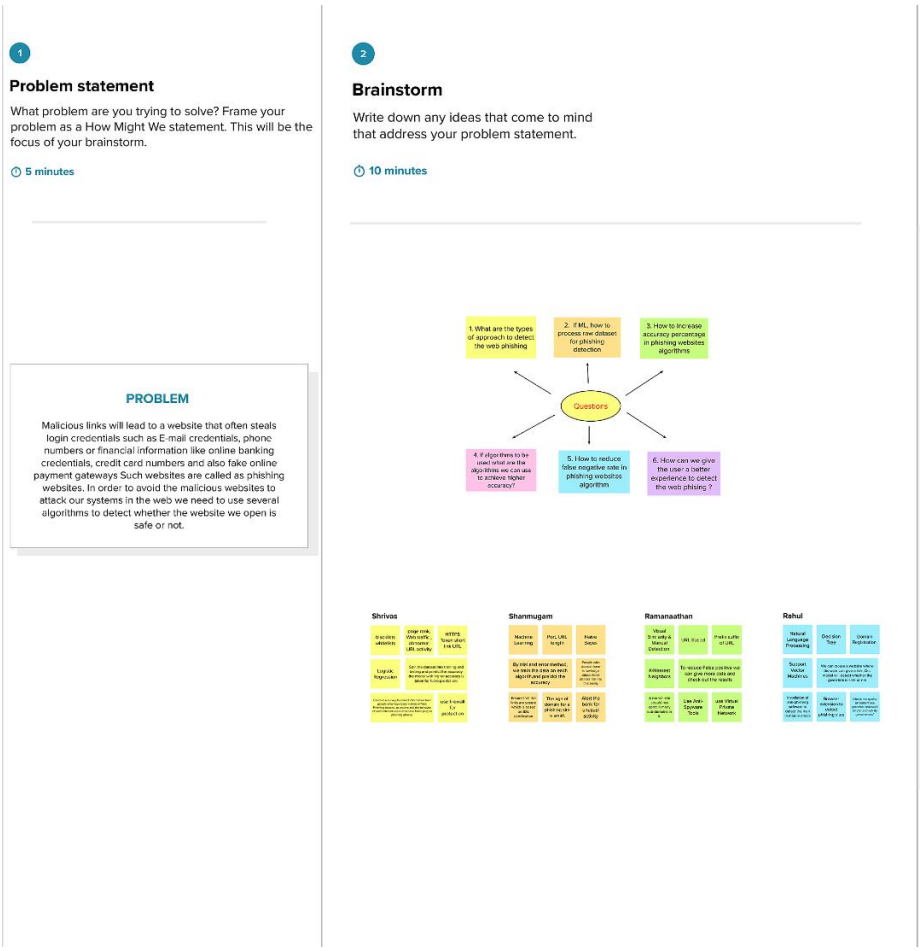
An empathy map is a collaborative tool teams can use to gain a deeper insight into their customers. Much like a user persona, an empathy map can represent a group of users, such as a customer segment. The empathy map was originally created by Dave Gray and has gained much popularity within the agile community.

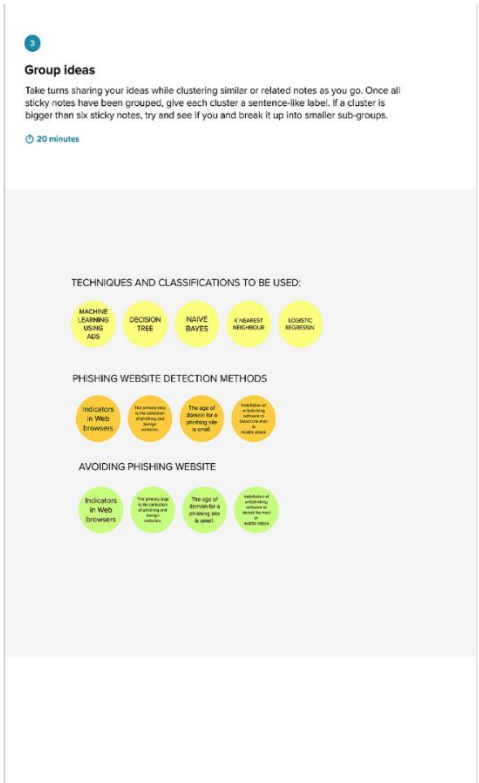


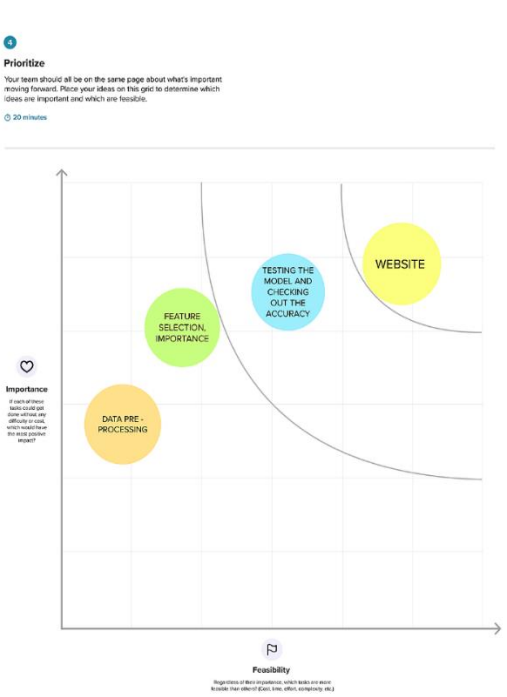
3.3 Proposed Solution

To overcome the problem of phishing website whenever we are clicking on one website it must show an alert box like it is a secure website or it is not a secure website.
Then another way is that we can scan the website in order to prevent our system or mobile from the phishing attack. Even though technologies are there we as the user have to be aware of the websites whether it is secure or not. We should not click any unwanted websites.

3.2 Ideation & Brainstorming







3.4 Problem Solution fit

Project Title: Web Phishing Detection **Project Design Phase - I** **Team ID: PNT2022TMID15077**

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

An internet user who is willing to shop products online.

An enterprise user surfing through the internet for some information

6. CUSTOMER CONSTRAINS CC

Customers have very little awareness on phishing websites. They don't know what to do after losing data.

They don't know what to do after losing data

5. AVAILABLE SOLUTIONS AS

The already available solutions are blocking such phishing sites and by triggering a message to the customer about dangerous nature of the website. But the blocking of phishing sites are not more effective as the attackers use a different/new site to steal potential data thus a AI/ML model can be used to prevent customers from these kinds of sites from stealing data

Explore AS, differentiate

Focus on J&P tap into BE understand RC

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

The phishing websites must be detected in a earlier stage . The user can be blocked from entering such sites for the prevention of such issues.

9. PROBLEM ROOT CAUSE RC

The hackers use new ways to cheat the naïve users. Very limited research is performed on this part of the internet.

7. BEHAVIOUR BE

The option to check the legitimacy of the Websites is provided.
Users get an idea what to do and more importantly what not to do.

Focus on J&P tap into BE understand RC

3. TRIGGERS TR

A trigger message can be popped warning the user about the site. Phishing sites can be blocked by the ISP and can show a "site is blocked" or "phishing site detected" message

4. EMOTIONS: BEFORE / AFTER EM

How do customers feel when they face a problem or a job and afterwards? The customers feel lost and insecure to use the internet after facing such issues. Unwanted panicking of the customers is felt after encountering loss of potential data to such sites.

10. YOUR SOLUTION SL

An option for the users to check the legitimacy of the websites is provided. This increases the awareness among users and prevents misuse of data, data theft etc.,

8.CHANNELS of BEHAVIOR CH

8.1 ONLINE

Customers tend to lose their data to phishing sites.

Nothing teaches like experience. When employees click on a link or an attachment in a simulated phishing email, it's important to communicate to them that they have potentially put both themselves and the organization at risk

8.2 OFFLINE

Customers try to learn about the ways they get cheated from various resources viz., books, other people etc.,

Simulated phishing campaigns reinforce employee training, and to understand risk and improve workforce resiliency as these can take many forms, such as mass phishing, spear phishing

4.REQUIREMENT ANALYSIS

4.1 Functional requirement

FR-1 - User Input

User inputs an URL in required field tocheck its validation.

FR-1 - Website Comparison

Model compares the websites using Blacklist and Whitelist approach.

FR-3 - Feature extraction

After comparing, if none found oncomparison then it extracts feature using heuristic and visual similarity approach

FR-4 - Prediction

Model predicts the URL using MachineLearning algorithms such as LogisticRegression, KNN, SWA.

FR-5 - Classifier

Model sends all output to classifier and produces final result.

FR-6 - Announcement

Model then displays whether website is alegal site or a phishing site.

FR-7 - Events

This model needs the capability of retrieving and displaying accurate resultfor a website.

4.2 Non-functional requirements

NFR-1 - Usability

Usability is commonly considered to be the enemy of security. In general, being secure means taking extra steps to avoid falling for different attacks.This is especially true of phishing where the bestways to prevent most phishing attacks arecommonly known, but cybersecurity guidance is rarely followed.

NFR-2 - Security

Implementation of updated securityalgorithms and techniques.

NFR-3 - Reliability

The reliability factor evaluates if a suspected site is legitimate or not.

NFR-4 - Performance

A phishing website has two key characteristics: it closely resembles a real website and has at least one field for users to enter their credentials. A suspicious attachment is frequently used as a phishing attempt warning sign.

NFR-5 - Availability

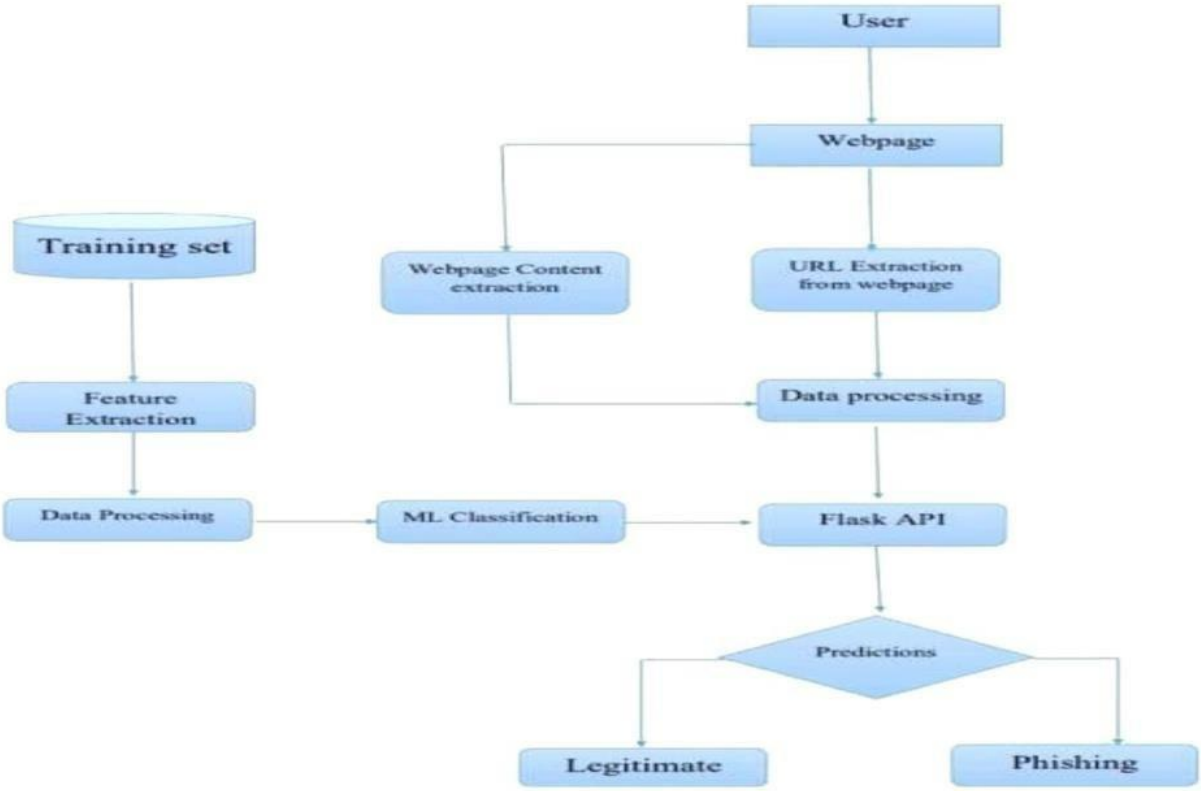
A phishing website has two key characteristics: it closely resembles a real website and has at least one field for users to enter their credentials. A suspicious attachment is frequently used as a phishing attempt warning sign.

NFR-6 - Scalability

5.PROJECT DESIGN

5.1 Data Flow Diagrams

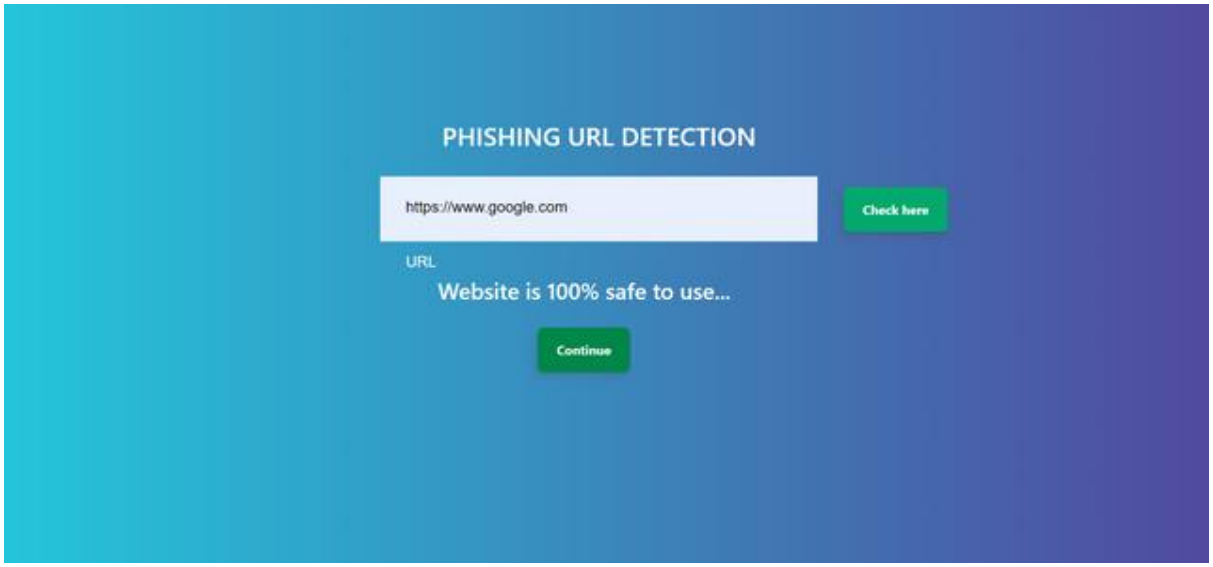
A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount ofthe system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.



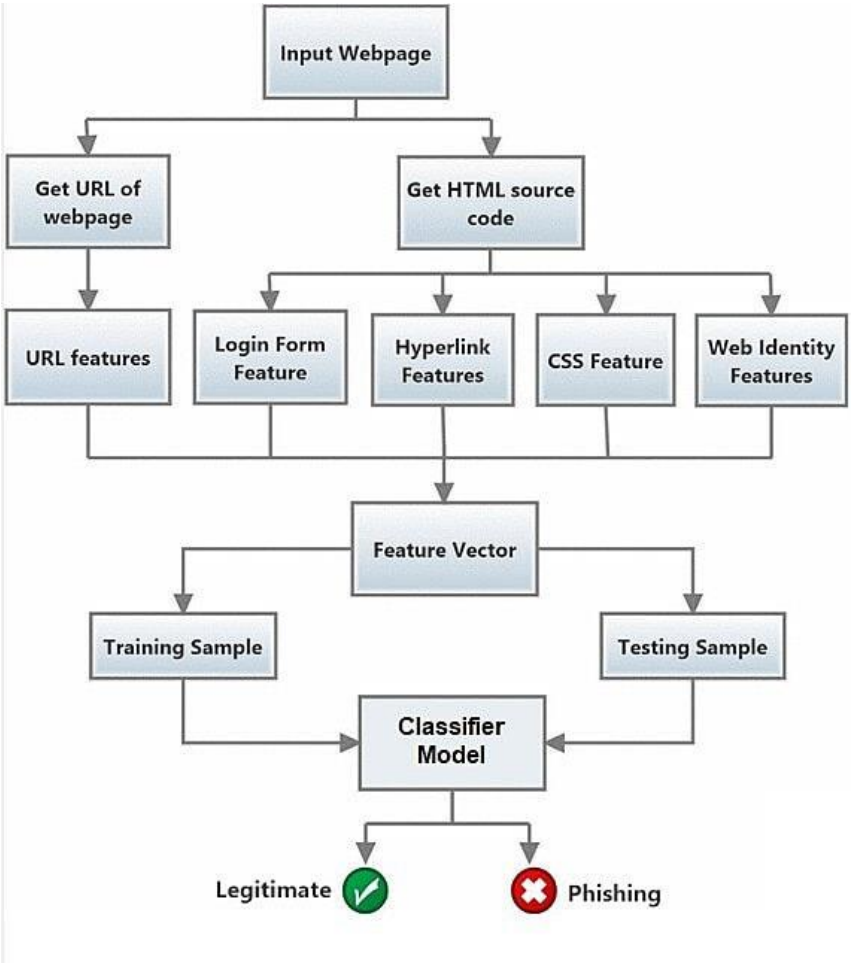
5.2 Solution & Technical Architecture

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

1. We can use various machine learning algorithms to detect webphishing.
2. At first, we need to preprocess the dataset provided and also detect the important features using feature importance and alsoby using feature selection techniques which can be done by inbuilt libraries.
3. We are able to split the dataset into training and testing.
4. We train the dataset on several algorithms such as
   logistic regression, naive bayes, decision tree, support vector machineetc.,
5. We need to test the models with the test dataset and calculate theaccuracy of those models.
6. So, With higher accuracy model we are able to use the model asthe default
   model and predict the URL to be safe or not.
7. UI has been created by web developers, where the user can givea URL in the website in the text box provided in the .
8. The model will be stored in the cloud. The website will get therequired parameters
   from the URL and pass it to the model prediction using API. This will predict the URL is legitimate orphishing.
9. Those results will be sent to the user on the website.



Solution Architecture Diagram:



5.3 User Stories

Use the below template to list all the user stories for the product.

User Type	Functional Requirement(Epic)	User StoryNumber	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobileuser)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirmingpassword.	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 applicationthrough Facebook	I can register & access thedashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through G-mail		Medium	Sprint-1
	Login	USN-5	As a user,I can log into the application by entering email& password		High	Sprint-1
Customer (Web user)	User input	USN-1	As a user i can input the particular URL in the required field and waiting for validation.	I can go access the website without any problem	High	Sprint-1
Customer CareExecutive	Feature extraction	USN-1	After i compare in case if none found on comparison then we can extract feature using heuristic and visual similarity approach.	As a User i can have comparison betweenwebsites for security.	High	Sprint-1
Administrator	Prediction	USN-1	Here the Model will predict the URL websites using Machine Learning algorithms such as LogisticRegression, KNN	In this i can have correct prediction on theparticular algorithms	High	Sprint-1

	Classifier	USN-2	Here i will send all the model output to classifier inorder to produce final result.	I this i will find thecorrect classifier forproducing the result	Medium	Sprint-2

7.CODING & SOLUTIONING

7.1 Feature 1

The most fascinating features that we can insist that is, We are using feature extraction python file where the website link given by the user in webpage is fed into python file using url parser and the url entered by the user will be parsed as string, The URL string should contain various parameters known as several features in machine learning model which is been trained by the dataset. The model is well trained by the known and required features to detect whether the link provided is genuine or not.

FEATURE EXTRACTION .PY

7.2 Feature 2

The another feature we insist is that we use python flask framework for using a pipeline between python and web page rendering. The user will be providing the webpage link in the website to check whether the site is genuine or not. The input is fed into the feature extraction file which is a python file and the model will be stored in the IBM Cloud and when the user opt to check for the genunity of the site.

The app.py is a python file which collects url from the webpage and also fed that url string into feature extraction.py which will return the features associated with the URL. So, the model will be retrieved from the IBM cloud and also the features extracted from the extraction.py will be given to the model for predicting. The predicted result will be sent from the Cloud via API. The result will be sent to the website and percentage of the genunity is sent to the site. So that user may be aware of the site to be genuine they opted for.

FLASK APP . PY
FLASK IBM_APP . PY

8.TESTING & RESULTS

When consdiering the machine learning algorithm to be used, We should look into the various algorithms that are to be considered while building the model and we should choose the algorithm which is very efficient and well performing with the dataset which we are using. We have a dataset of which 70% of the data is used for trainng and 30% of the data is used for tesing. While testing the 30% of the data with the models we built we can caluclate the accuracy of each model and other metrics. Then we choose the model with high accuracy and the model will be stored in IBM Cloud. The best model is retrieved from the cloud and fed to the python file and further proceeded by webpage through python flask.

The various model trainied with particular accuracies and other metrics will be shown below.

	ML Model	Accuracy	f1_score	Recall	Precision
1	CatBoost Classifier	0.972	0.976	0.994	0.987
2	Gradient Boosting Classifier	0.971	0.975	0.992	0.985
3	Random Forest	0.967	0.972	0.994	0.986
4	Decision Tree	0.961	0.965	0.992	0.991
5	Support Vector Machine	0.957	0.963	0.982	0.966
6	K-Nearest Neighbors	0.944	0.950	0.962	0.996
7	Logistic Regression	0.924	0.933	0.947	0.927
8	Naive Bayes Classifier	0.583	0.420	0.291	0.996

10 .ADVANTAGES &DISADVANTAGES

10.1 Advantages

- This system can be used by many E-commerce or other websites in order to have good customer relationship.
- User can make online payment securely.
- Data Mining algorithm used in this system provides better performance as compared to other traditional classifications algorithms.
- A mailbox-level anti-phishing solution offers an additional layer of protection by analysing account information and understanding users communication habits.
- This delivers an enhanced level of phishing protection to detect attacks faster,alert users and remediate threats as quickly as possible.

10.2.1Disadvantages

- If Internet connection fails,this system won't work.
- All websites related data will be stored in one place.
- 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.

11.CONCLUSION

11.1 Conclusion

This paper aims to enhance detection method to detect phishing website using machine learning technology. Also , classifiers generated by machine learning algorithms identify legitimate phishing websites.The proposed technique can detect new temporary phishing sites and reduce the damage caused by phishing attacks. The performance of the proposed technique based on machine learning is more effective that previous phishing detection technologies. In the future, it will be useful to investigate the impact of feature selection using various algorithms.

12.FUTURE SCOPE

12.1 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
GitHub Link
Demo Video Link