IBM PROJECT REPORT

WEB PHISHING DETECTION

Submitted by

Pratibha Senthil - 2019105040

Sindhuja G - 2019105054

Tharun V Darshan- 2019105062

Yamini K– 2019105605

TEAM ID :PNT2022TMID35406

INTRODUCTION

1.1 PROJECT OBJECTIVES:

BY THE END OF THE PROJECT:

• We'll be able to understand the problem to classify if it is a regression or a classification kind of problem.

• We will be able to know how to pre-process/clean the data using different data preprocessing techniques.

• Applying different algorithms according to the dataset

• We will be able to know how to find the accuracy of the model.

• We will be able to build web applications using the Flask framework.

1.2 PURPOSE:

There are a number of users who purchase products online and make payments through e-banking. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc., often for malicious reasons. This type of e-banking website is known as a phishing website. Web service is one of the key communications software services for the Internet. Web phishing is one of many security threats to web services on the Internet. Common threats of web phishing :

1. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity.

2. It will lead to information disclosure and property damage.

3. Large organizations may get trapped in different kinds of scams.

4. This Guided Project mainly focuses on applying a machine-learning algorithm to detect Phishing websites.

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

2.LITERATURE SURVEY

2.1 EXISTING PROBLEM

Phishing is the most popular attack vector for criminals and has grown 65% in the last year. Malicious links will lead to a website that often steals login credentials or financial information like credit card numbers. Attachments from phishing emails can contain malware that once opened can leave the door open to the attacker to perform malicious behavior from the user's computer.

Due to their low bar of skill required to launch, phishing is a popular choice for cyber criminals. Many of them use phishing kits, which include all the technical materials needed to launch a phishing campaign.

More advanced phishing methods like spoofing (pretending to send emails from a legitimate source), spear phishing (personalizing emails to target specific people), and whaling (targeting high-level executives) remain popular and are even harder to detect by eye alone.

2.2 References

S.No	Title of the Journal	Year of publication	Inference
1.	Detection of phishing websites using an efficient feature-based machine learning framework	2018	In this, they have classified extracted features into three categories such as URL Obfuscation features, Third-Party-based features, Hyperlink-based features. Moreover, the proposed technique gives 99.55% accuracy. Drawback of this is that as this model uses third party features, classification of websites depends on the speed of third-party services. Also this model purely depends on the quality and quantity of the training set and Broken links feature extraction has a limitation of more execution time for the websites with more number of links.

2.	Finding effective classifier for malicious URL detection	2018	In this they have combined statistical analysis of URL with machine learning technique to get a result that is more accurate for classification of malicious URLs. Also they have compared six machine-learning algorithms to verify the effectiveness of the proposed algorithm which gives 99.7% precision with false positive rate less than 0.4%.
3.	Detecting Phishing Websites Using Rule-Based Classification Algorithm	2018	They have proposed rule based classification techniques for phishing website detection. They have concluded that association classification algorithms are better than any other algorithms because of their simple rule transformation. They achieved 92.67% accuracy by extracting 16 features but this is not up to mark so the proposed algorithm can be enhanced for efficient detection rate.
4.	A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms	2016	In this paper, a proposed model was carried out in two phases. In phase 1,they individually perform classification techniques, and select the best three models based on high accuracy and other performance criteria. While in phase 2, they further combined each individual model with the best three models and made a hybrid model that gives better accuracy than individual models. They achieved 97.75% accuracy on the testing dataset. There is limitation of this model that it requires more time to build hybrid model

5.	A Framework for Auto-Detection of Phishing Websites	2017	For phishing websites, machine-learning data can be created using this framework. In this, they have used reduced features set and using python for building query .They build a large labeled dataset and analyze several machine-learning classifiers against this dataset .Analysis of this gives very good accuracy using machine-learning classifiers. These analyses how long it takes to train the model.

2.3 Problem Statement definition

Phishing attacks are becoming more and more sophisticated, and our algorithms are suffering to keep up with this level of sophistication. They have low detection rate and high false alarm especially when novel phishing approaches are use. The blacklist-based method is unable to keep up with the current phishing attacks as registering new domains has become easier. Moreover, comprehensive blacklist can ensure a perfect up-to-date database. Various other techniques such as page content inspection algorithms have been used to combat the false negatives but as each algorithm uses a different approach, their accuracy varies. Therefore, a combination of the two can increase the accuracy while implementing different error detection methods.

Problem statements:

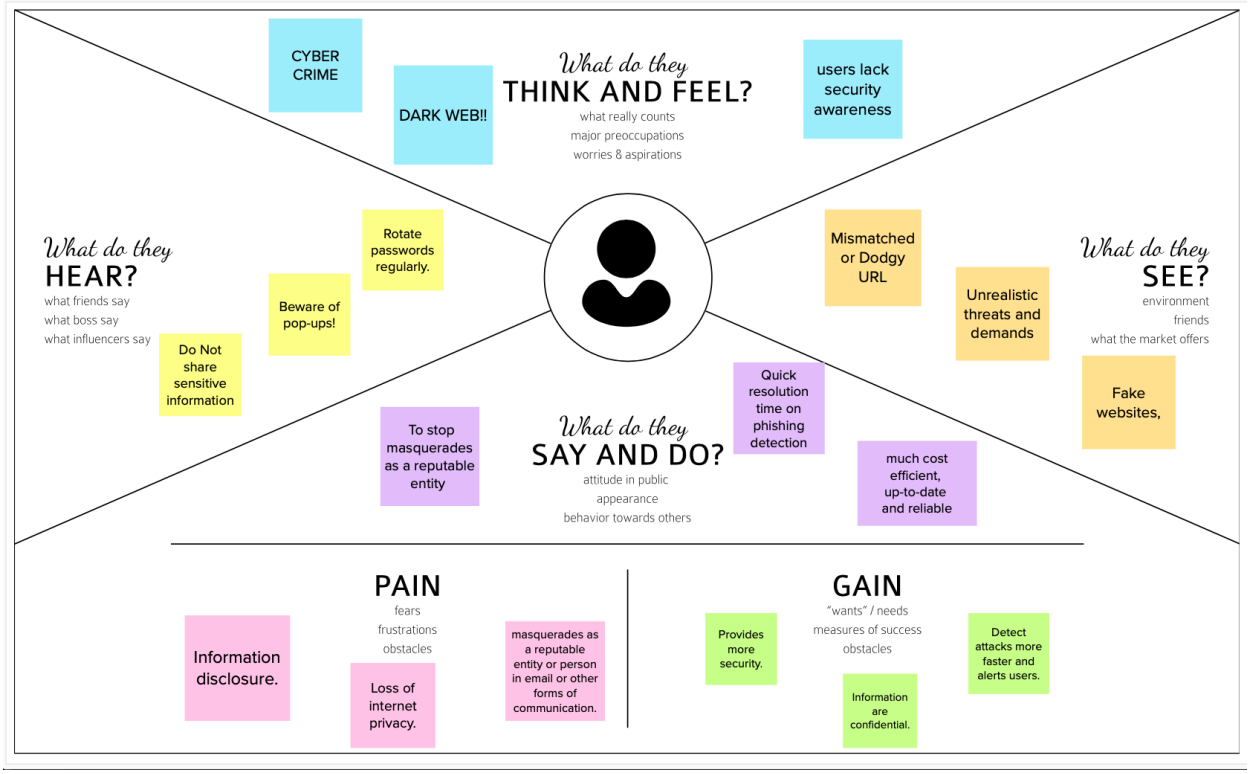
1. Phishing is the technique of extracting user credentials and sensitive data from users by masquerading as a genuine website.
2. Phishing attacks can lead to huge financial losses for customers of banking and financial services.

User need:

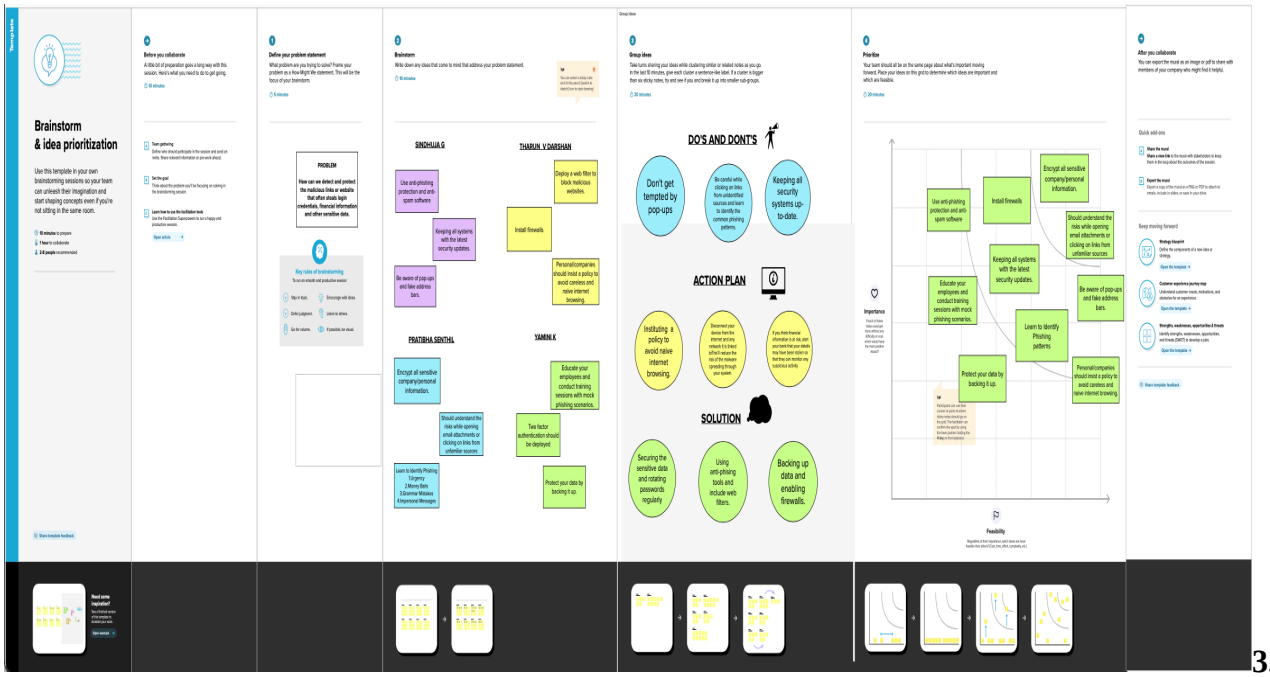
1.To stop masquerades as a repudiation entity.
2.To have internet safety .

3. IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTROMING



3.3 3.3

PROPOSED SOLUTION

S.No	Parameter	Description
1.	Problem Statement (Problem to be solved)	Our main aim of this paper is classification of a phishing website with the aid of various machine learning techniques to achieve maximum accuracy and concise model.
2.	Idea / Solution description	This strategy has a strong generalisation capacity to find unknown malicious URLs compared to the blacklist approach. Anti-spyware and firewall settings should be used to prevent phishing attacks and users should update the software automatically.
3.	Novelty / Uniqueness	Python serves as a powerful tool to execute the application with Low false positives, High accuracy. It can easily differentiate the fake and safe URLs. If it's fake means, a warning message will be intimate to the users
4.	Social Impact / Customer Satisfaction	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. As an impact of this model, people can able to find out fraudulent websites or fake ones.
5.	Business Model (Revenue Model)	Our model can be used by all people to secure their data from malicious websites. It's an open source tool.
6.	Scalability of the Solution	The methods are evaluated in terms of learning rate, accuracy, and precision. It presents the learning rate of the methods during the training phase. The performance of three detectors during the training phase are similar. It is evident that the learning ability of methods are same.

3.4 PROPOSED SOLUTION FIT

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 Input**	User types in URL to be checked.
FR-2	**Feature Extraction**	The model must extract appropriate features from the URL given by the user,so that we can apply the ML algorithm.

FR-3	**Prediction**	Model must use an ML algorithm such as KNN, Logistic regression etc to make predictions.
FR-4	**Verify the link provided by the user**	User inputs the link to be verified
FR-5	**Display the result**	If the site link is a phishing site, user must be aware and read the precautions displayed If the site link is legit, exit the application
FR-6	**Sharing and clearing the Queries**	If any doubts, send query Read FAQs

4.2 NON-FUNCTIONAL REQUIREMENTS:

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

NFR.No	NON-FUNCTIONAL REQUIREMENT	DESCRIPTION
NFR-1	**USABILITY**	Engage the user about the process to ensure that the functionality can meet design and usability requirements. It relates to overall satisfaction of the user.
NFR-2	**SECURITY**	Users need to be protected from malicious attacks when using the site.
NFR-3	**RELIABILITY**	It focuses on preventing failures during the lifetime of the product or system, from commissioning to decommissioning.
NFR-4	**PERFORMANCE**	It is the ability of the application to always run acceptably. In time-critical scenarios, even the smallest delay in processing data can be unacceptable.
NFR-5	**AVAILABILITY**	Fault tolerance , Ensuring that the application can meet its availability targets to be resilient

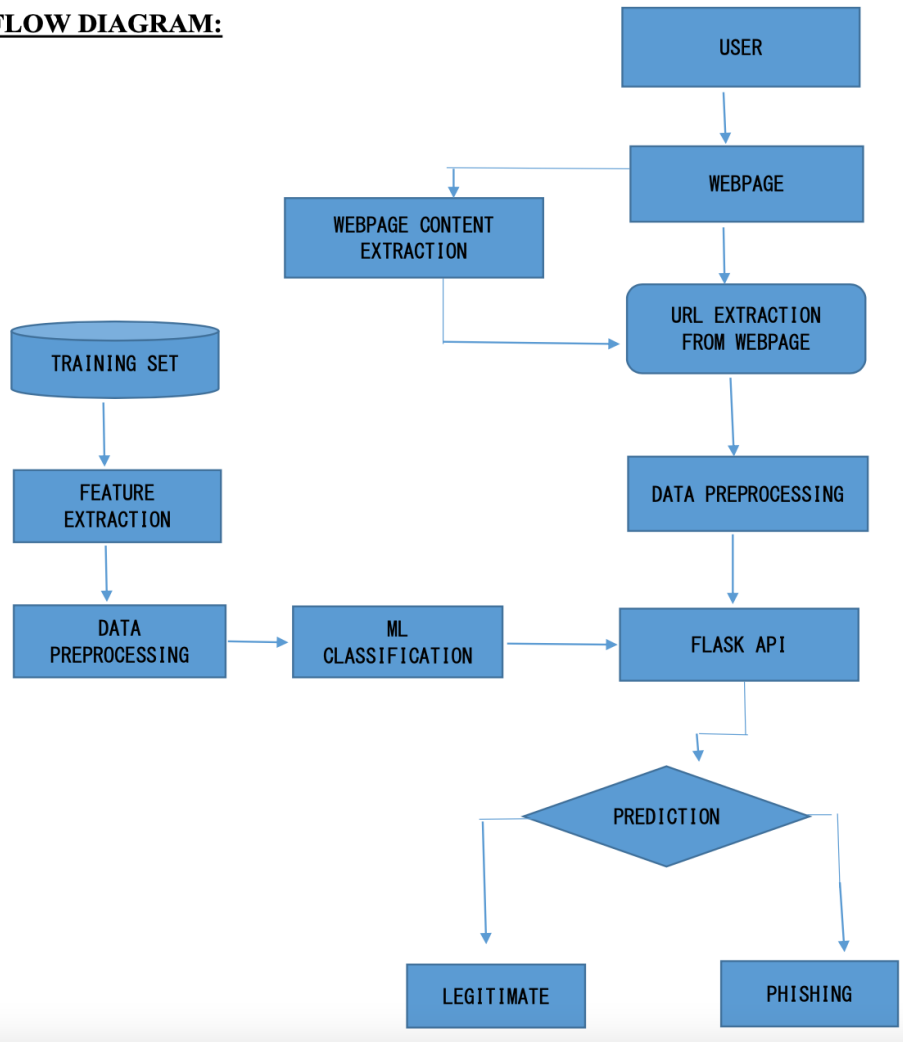
NFR-6	**SCALABILITY**	It is the ability for the application to scale to meet increasing demands; for example, at peak times or as the system becomes more widely adopted.

5.PROJECT DESIGN

5.1 Data flow diagram

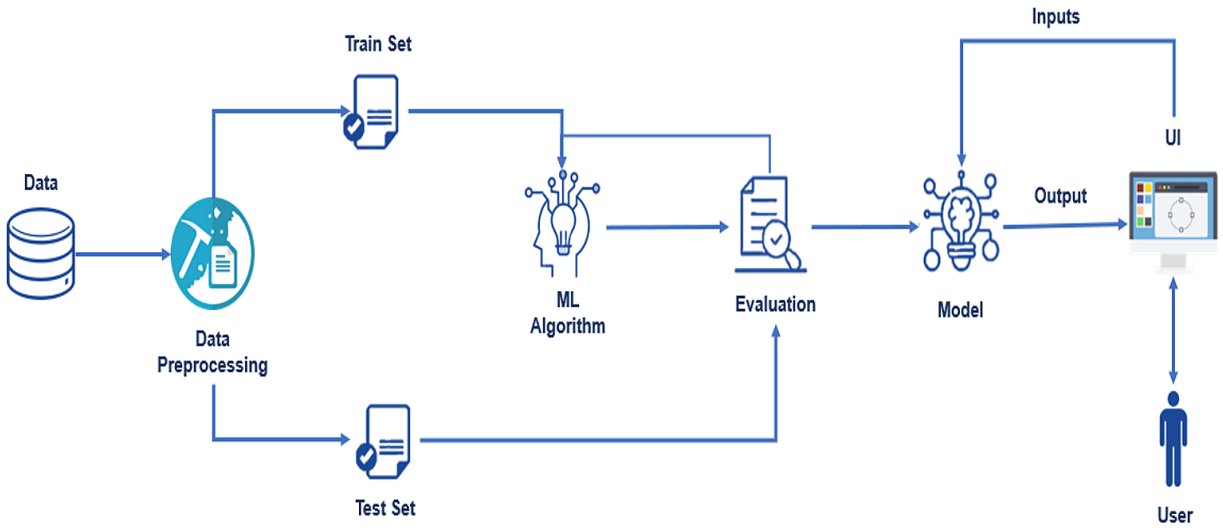
A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

DATA FLOW DIAGRAM:



5.2 SOLUTION AND TECHNICAL ARCHITECTURE

TECHINCAL 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 an user, I must register for the application by entering my User Id, password, confirming my password and other personal details.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user,I should create strong passwords.	I can access my Account more safely.	High	Sprint-1
		USN-3	As an user, ,I can register in websites which does not navigate me to any other websites.	I can store my details incorrect websites.	Low	Sprint-2

	Login	USN-4	As an user, I must log into the application by entering User id & password.	I can login.	High	Sprint-1
Customer (Web user)	Registration	USN-5	As a user,I can register my personal details only in officialwebsites..	I can register and access.	High	Sprint-1
	Login & Dashboard	USN-6	As an user, I must easily navigate through dashboard and I must be able to use the dashboard to get details about app and make use of available resources.	I can login to the application using the same User id and password to access the resource.	High	Sprint-1
Customer Care Executive	Login	USN -7	As a CCE I must login to application using User id & Password and I must be able to	I can complain.	High	Sprint-1
			interact with user and check if my account is phished			
	Dasshboard	USN-8	As a CCE I should not take others information.	I can be punished for it.	High	Sprint-1
Administrator	Login and Dashboard	USN-9	As an administrator, I must login and access dashboard and manage and direct activities and databases carefully&safely.	This reduces from Phishing.	High	Sprint-1

6.PROJECT PLANNING AND SCHEDULING

6.1 SPRINT PLANNING AND ESTIMATION

Sprint	Functional requirement (Epic)	User story/ input	Userstory/task	Story points	Priority	Team member
Sprint -1	User inputs	USN – 1	User can drop the URLin the given space for verification and waitfor validation	10	High	Tharun VDarshan
Sprint -1	Websitecomparison	USN – 2	The detection modelchecks for the phishing activity.	10	High	Sindhuja G
Sprint -2	Featureextraction	USN - 3	After completion, ifnone is found then the model extracts the necessary features from the URL for furtherprocess.	10	High	PratibhaSenthil

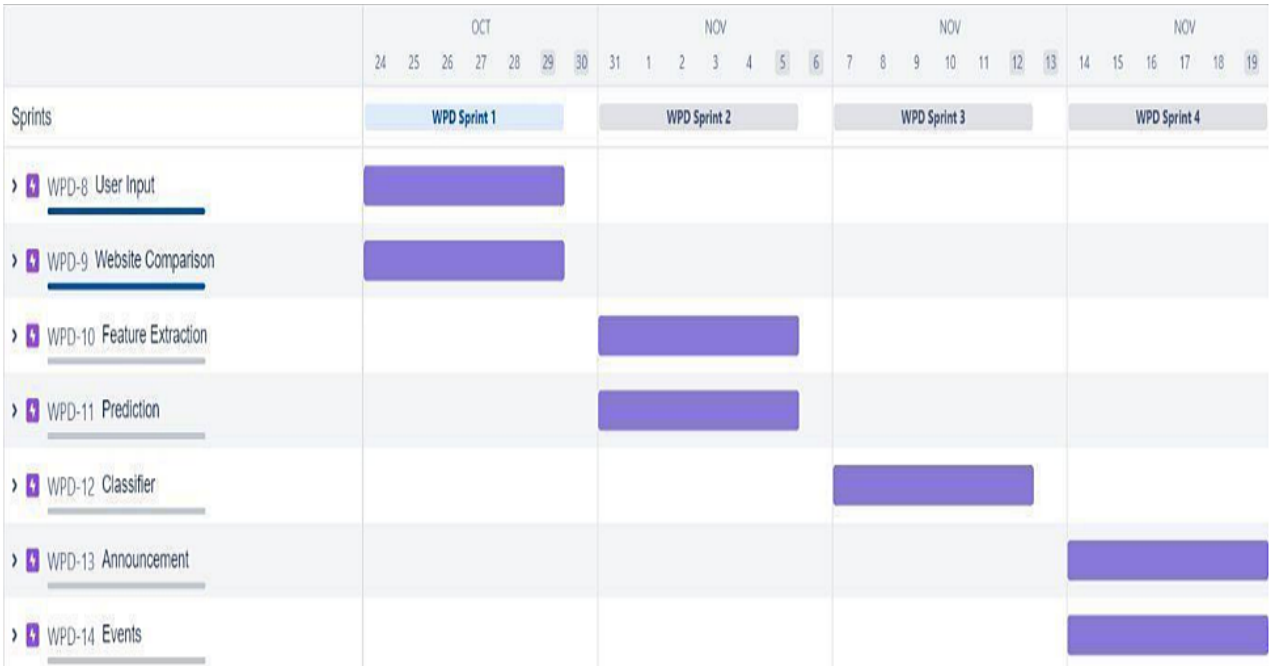
Sprint -2	Prediction	USN - 4	Model predicts the URL using Machinele arning algorithms such as logistic Regression, KNN.	10	High	Yamini K
Sprint -3	Classification	USN - 5	Classificati on isdone based on the prediction made toprovide a result.	20	High	Tharun VDarshan &Sindhuja G
Sprint -4	Announceme nt	USN - 6	Result is displayedwhet her the website is a phishing website ornot.	10	High	PratibhaSenthil
Sprint -4	Events	USN – 7	Should check themodel for itscapabilities and efficiency.	10	High	Yamini K

6.2 SPRINT DELIVERY SCHEDULE

Sprint	Total story points	Duration	Sprint started date	Sprint ended date(Planned)	Story Points Completed (as on Planned End	Sprint Release Date (Actual)

					Date)	
Sprint-1	20	6days	29October2022	4November2022	20	4November2022
Sprint-2	20	6days	4November2022	10November 2022	20	10November 2022
Sprint-3	20	6days	11 November 2022	18November 2022	20	18November 2022
Sprint-4	20	6days	18November 2022	24November 2022	20	24November 2022

6.3 REPORTS FROM JIRA



7.CODING AND SOLUTIONING

7.1 FEATURE CODE 1

Feature sets are divided into four main categories:

Address Bar-Based Features – these are features extracted from the URL itself like,

- URL length : Function returns 1 if URL length is less than 54, 0 if length is between 54 and 75 , else returns -1 which indicates phishing.
- Whether it contains an IP address: Functions returns -1 if IP address is detected otherwise 1 is returned.
- Uses an URL shortening service like Tiny URL or Bitly:- A number of shortening services are stored in a separate variable, and re.search is used to see if any of there are in the URL given and if so, -1 is returned indicating it is a phishing URL.
- Employs redirection:- Checks for //, if it is in a position greater than seven in the URL, -1 is returned indicating it is phishing.

Abnormal Features

- Loading images loaded in the body from a different URL :- Favicon module is used and if domain of image and URL match it returns 1 else -1 is returned indicating phishing.
- Minimal use of meta tags: If more than one link is associated with page, it returns -1.
- The use of a Server Form Handler (SFH):- If SFH refers to a domain then it returns -1.
- Submitting information to email:-Using mail() or mailto() returns -1.
- An abnormal URL :-If host name is not in URL, then it is classified as an abnormal URL.

7.2 Feature Code 2

HTML and JavaScript-Based Features – these can include things like:

- Disabling the ability to right-clicks:- If right click is disabled , function returns -1 else it returns 1.
- On mouseover:-If status bar changes , function returns -1.
- Using pop-up windows:-If popup window contains textfield , then it is considered phishing and returns -1.
- iFrame redirection:-If iframe is used it returns -1, else it returns 1.

Domain-Based Features – these can include:

- Unusually young domains:-If age of domain is greater then 6 months it returns 1 and it considered legitimate.
- Suspicious DNS record:-If there is no DNS record, then it is considered a phishing URL.
- Low volume of website traffic:-If website rank it lesser than 100, it returns 1.
- PageRank:- Most phishing websites do not have a page rank, the function checks for prank and if not, available it is considered a phishing URL.
- Whether the site has been indexed by Google:- Most phishing websites are not indexed on Google, if webpage is indexed , function returns 1 and is considered legitimate.

8.TESTING

8.1 Testcases Report

				Date	19-Nov-22								
				Team ID	PNT2022TMID35404								
				Project Name	Project - 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	Comments	TC for Automation(Y/N)	BUG ID	Executed By
1	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://www.google.com	Should Display the Webpage	Working as expected	Pass		N		Pratibha Senthil
2	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://www.google.com	Should Wait for Response and then gets Acknowledge	Working as expected	Pass		N		Yamini K
3	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://www.google.com	User should observe whether the website is legitimate or not.	Working as expected	Pass		N		Sindhuja G
4	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://www.google.com		Working as expected	Pass		N		Tharun V Darshan
5	Functional	Home Page	Testing the website with multiple URLs		1.EnteRURL(https://phishingshield.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	https://127.0.0.1	User can identify whether the websites is secure or not	Working as expected	Pass		N		Pratibha Senthil

8.2 User Acceptance Testing

UAT Execution & Report Submission

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	4	2	1	0	7
URL	3	0	2	3	8
External	2	4	0	1	7
Model	8	2	4	2	16
Skipped	0	5	1	0	6
Prediction	4	0	1	1	6
Won't Fix	6	2	2	1	11
Totals	27	15	11	8	61

Test Case Analysis

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

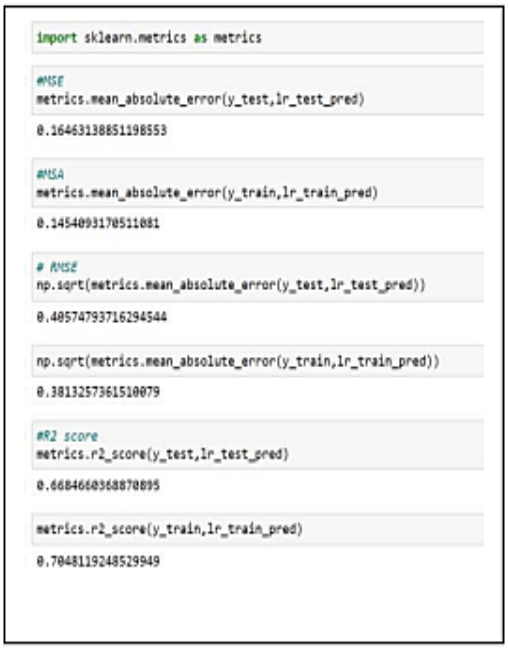
Section	Total Cases	Not Tested	Fail	Pass
By Design	7	0	0	7
URL	51	0	0	51
External	2	0	0	2
Model	3	0	0	3
Skipped	9	0	0	9
Prediction	4	0	0	4
Won't Fix	2	0	0	2

9.RESULTS

Model Performance Testing:

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

S.No.	Parameter	Values	Screenshot

1.	Metrics	**Regression Model:**
		MAE – 0.145
		MSE – 0.1646
		RMSE – 0.3818
		R2 score – 0.704
		
		Classification Model: Confusion Matrix –
		array([[961,53],
		[20, 1177]],
		dtype=int64)
		
		Accuracy Score – 96.69%
		
		Classification Report –
		precision recall f1-score support
		
		-1 0.93 0.89 0.91 1014
		1 0.91 0.94 0.93 1197
		
		accuracy 0.92 2211
		macro avg 0.92 0.92 0.92 2211
			
			`pd.crosstab(y_test, lr_test_pred)`
			
			col_0 -1 1
			row_0
			-1 902 112
			1 70 1127
			
			`print(classification_report(y_test, lr_test_pred))`
			
			precision recall f1-score support
			
			-1 0.93 0.89 0.91 1014
			1 0.91 0.94 0.93 1197
			
			accuracy 0.92 2211
			macro avg 0.92 0.92 0.92 2211
			weighted avg 0.92 0.92 0.92 2211

		weighted avg 0.92 0.92 0.92 2 211 Training accuracy: 0.92 Testing accuracy:0.91	print('Training accuracy: ',accuracy_score(y_train,lr_train_pred)) print('Testing accuracy: ',accuracy_score(y_test,lr_test_pred)) Training accuracy: 0.9272953414744459 Testing accuracy: 0.9176843057440073
2.	Tune the Model	Hyperparameter Tuning - GridSearchCV Validation Method	

10.ADVANTAGES AND DISADVANTAGES

ADVANTAGES OF WEB PHISHING DETECTION:

- The algorithm provides a clear idea about the effective level of each classifier on phishing email detection
- It has a very high level of accuracy
- It can be evolved with time according to the classification of features
- Majority of the work is online
- It provides a secure connection between agent and user
- It can Mitigate zero-hour attacks.

- Requiring low resources on host machine
- We can construct our own ML classification models

DISADVANTAGES OF WEB PHISHING DETECTION:

- It is a time-consuming process
- The process needs feed continuously
- High computational cost involved in certain cases
- It involves a huge number of rules
- Can result in excessive queries with heavily loaded servers.

11.CONCLUSION

The most important way to protect the user from phishing attack is the education awareness. Internet users must be aware of all security tips which are given by experts. Every user should be trained not to blindly follow the links to websites where they must enter their sensitive information. It is essential to check the URL before entering the website. In future, system can upgrade to automatic detection of the web page and the compatibility of the application with the web browser. Additional work also can be done by adding some other characteristics to distinguishing the fake web pages from the legitimate web pages. These can be upgraded in the Darkphish web page.

There are many features that can be improved in the work, for various other issues. The heuristics can be further developed to detect phishing attacks in the presence of embedded objects like flash. Identity extraction is an important operation and it was improved with the Optical Character Recognition (OCR) system to extract the text and images. More effective inferring rules for identifying a given suspicious web page, and strategies for discovering if it is a phishing target, should be designed in order to further improve the overall performance of this system. Moreover, it is an open challenge to develop a robust malware detection method, retaining accuracy for future phishing emails. In addition, the dynamic and static features complement each other, and therefore both are considered important in achieving high accuracy.

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

PROJECT DEMO VIDEO:

https://drive.google.com/file/d/140mBqI2Qq_hOgckDS-g8cYsr27AaODKq/view?usp=share_link

GITHUB REPOSITORY:

https://github.com/IBM-EPBL/IBM-Project-7176-1658849123