WEB PHISHING DETECTION

SUBMITTED BY

Kowsalya.S

Petchiammal.s

Pavithra.m

Thenmozhi.m

Monisha.k

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

UDAYA SCHOOL OF ENGINEERING

VELLAMODI

TEAM ID:PNT2022TMID52047

2019-2023

CONTENTS

1.INTRODUCTION

1.1 PROJECT OVERVIEW

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. Phishing may be a style of broad extortion that happens once a pernicious web site act sort of a real one memory that the last word objective to accumulate unstable info, as an example, passwords, account focal points, or MasterCard numbers. all the same, the means that there square measure some of contrary to phishing programming

and techniques for recognizing potential phishing tries in messages and characteristic phishing substance on locales, phishes think about new and crossbreed procedures to bypass the open programming and frameworks. Phishing may be a fraud framework that uses a mixture of social designing what is additional, advancement to sensitive and personal data, as an example, passwords associate degree open-end credit unpretentious elements by presumptuous the highlights of a reliable individual or business in electronic correspondence. Phishing makes use of parody messages that square measure created to seem substantial and instructed to start out from true blue sources like money connected institutions, online business goals, etc, to draw in customers to go to phony destinations through joins gave within the phishing websites.

1.2 PURPOSE

The basic structure of the website is made with the help of HTML. CSS is used to add effects to the website and make it more attractive and user-friendly. It must be noted that the website is created for all users, hence it must be easy to operate with and no user should face any difficulty while making its use. Every naÃ¯ve person must be able to use this website and avail maximum benefits from it.

The website shows information regarding the services provided by us. It also contains information regarding ill- practices occurring in todays technological world. The website is created with an opinion such that people are not only able to distinguish between legitimate and fraudulent website, but also become aware of the mal-practices occurring in current world. They can stay away from the people trying to exploit ones personal information, like email address, password, debit card numbers, credit card details, CVV, bank account numbers, and the list goes on.The dataset consists of different features that are to be taken into consideration while determining a website URL as legitimate or phishing.

2.literature survey

2.1 EXISTING PROBLEM

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

2.2 REFERENCES

H. Huang et al., (2009) proposed the frameworks that distinguish the phishing utilizing page section similitude that breaks down universal resource locator tokens to create forecast preciseness phishing pages normally keep its CSS vogue like their objective pages.

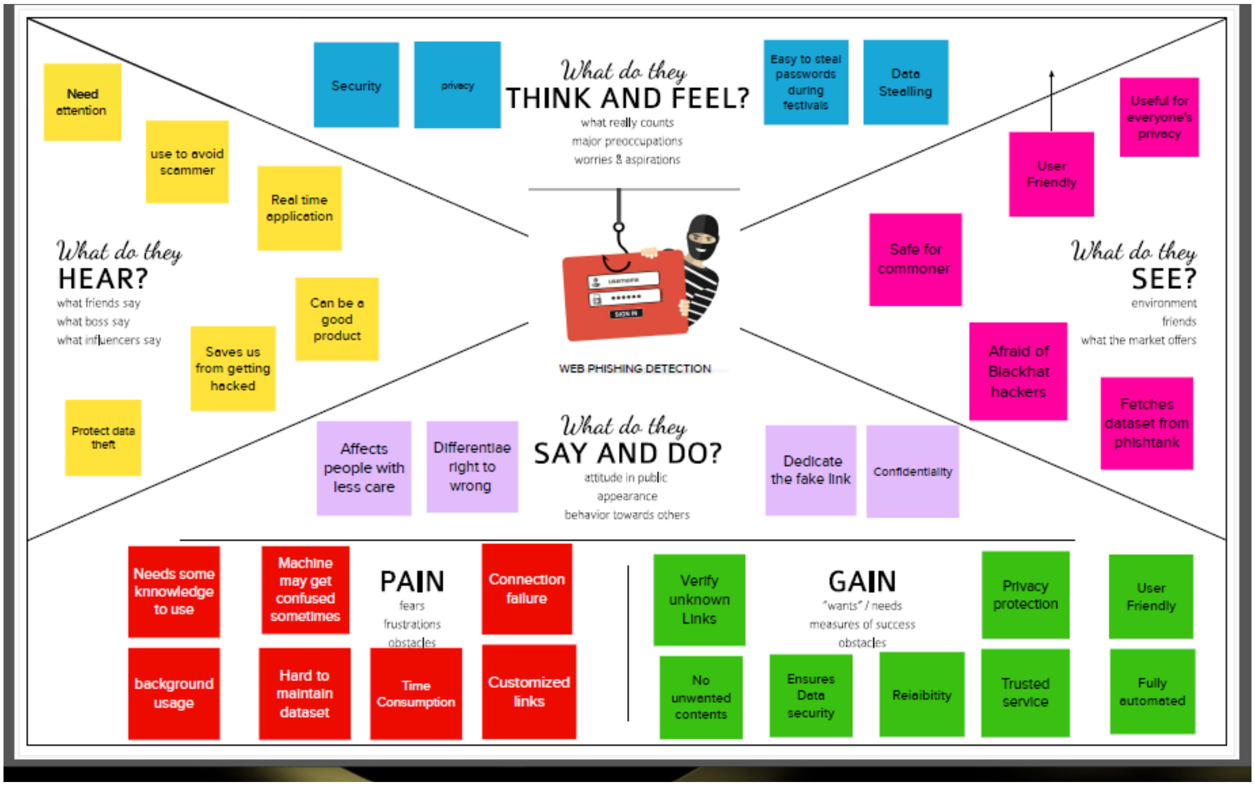
S. Marchal et al., (2017) proposed this technique to differentiate Phishing website depends on the examination of authentic site server log knowledge. An application Off-the- Hook application or identification of phishing website. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, and nice language-freedom, speed of selection, flexibility to dynamic phish and flexibility to advancement in phishing ways.

2.3 problem statement definition

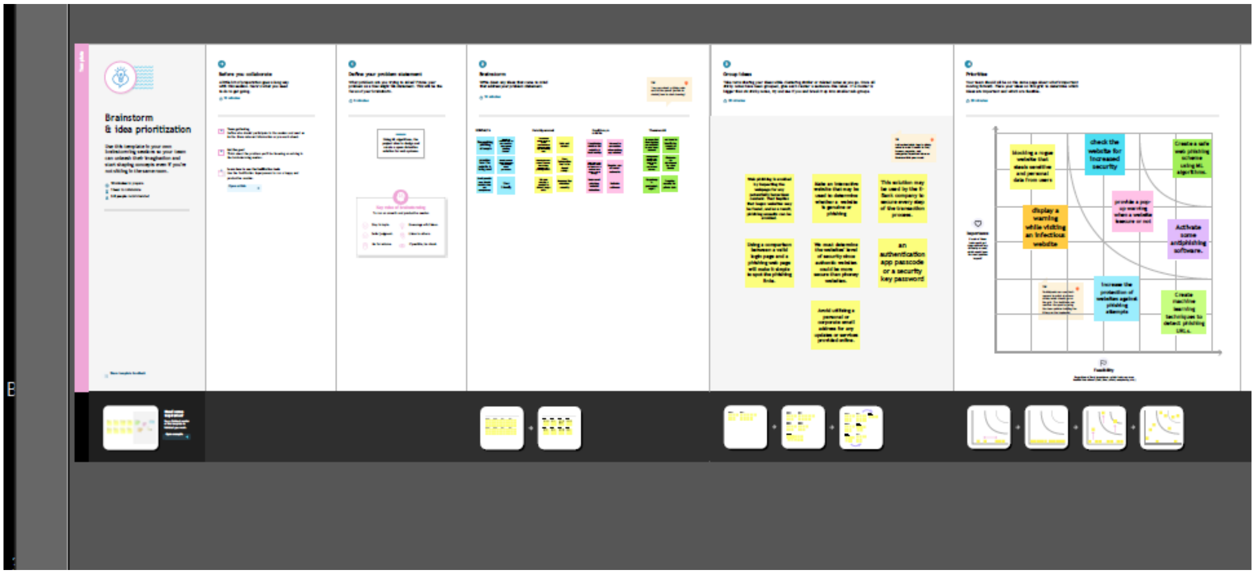
Many consumers utilize e-banking to make purchases and payments for goods they find online. There are several e-banking websites that often request sensitive information from users—such as usernames, passwords, and credit card information—for harmful purposes. Phishing websites are this kind of online banking websites. One of the most important software services for communications on the Internet is the web service. One of the many security risks to online services on the Internet is web phishing. By seeming to be a trustworthy organization, web phishing seeks to obtain sensitive information including usernames, passwords, and credit card numbers. It will result in data leakage and property damage. Large businesses may fall victim to various schemes

3.IDEATION & PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTORMING



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.

Before you collaborate

Define your problem statement

Key rules of brainstorming

Stay in topic

Encourage wild ideas

Defer judgment

Listen to others

Go for volume

If possible, be visual

Brainstorm

Group ideas

Web phishing is avoided by inspecting the webpage for any potentially hazardous content. This implies that bogus websites may be found, and as a result, phishing assaults can be avoided.

Make an interactive website that may be used to determine whether a website is genuine or phishing

This solution may be used by the E-Bank company to secure every step of the transaction process.

Using a comparison between a valid login page and a phishing web page will make it simple to spot the phishing links.

We must determine the websites' level of security since authentic websites could be more secure than phoney websites.

an authentication app passcode or a security key password

Avoid utilising a personal or corporate email address for any updates or services provided online.

Prioritize

blocking a rogue website that steals sensitive and personal data from users

check the website for increased security

Create a safe web phishing scheme using ML algorithms.

provide a pop-up warning when a website insecure or not

display a warning while visiting an infectious website

Activate some antiphishing software.

Increase the protection of websites against phishing attempts

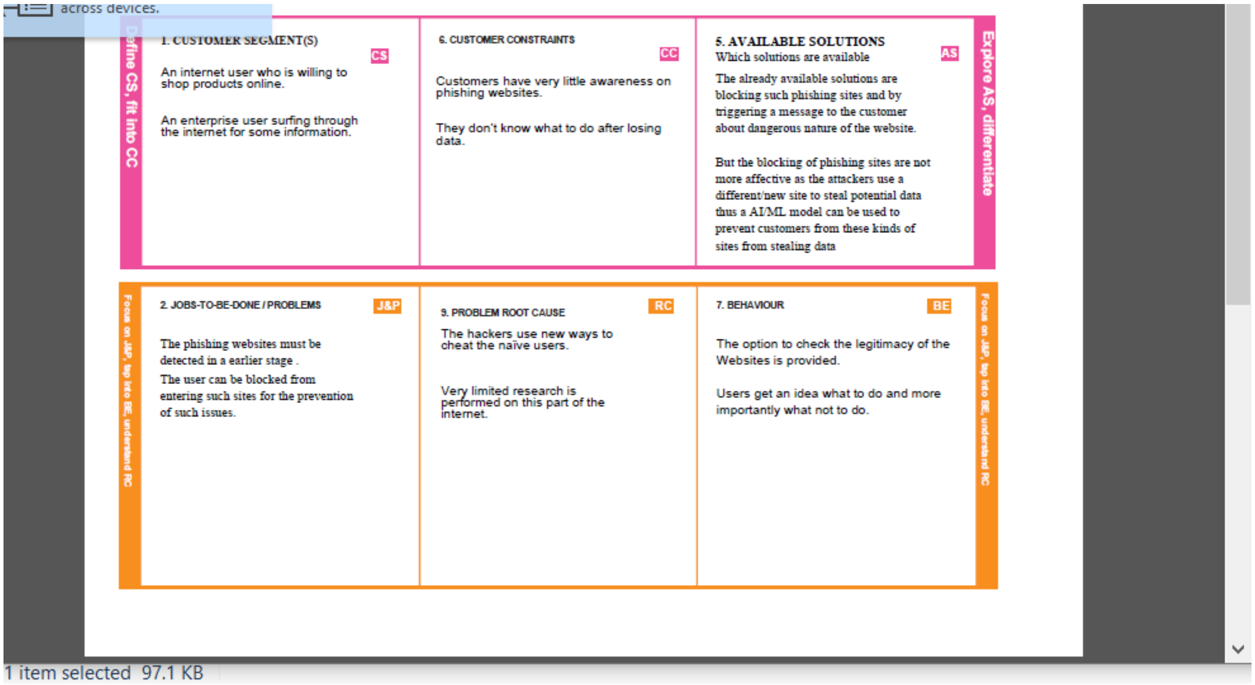
Create machine learning techniques to detect phishing URLs.

3.3 PROPOSED SOLUTION

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	Web phishing tends to steal a lots of information from the user during online transaction like username, password, important documents that has been attached to that websites. There are Multiple Types of Attacks happens here every day, but there is no auto detection Process through Machine Learning is achieved
2.	Idea / Solution description	Through ML and data mining techniques like classification algorithm user can able to attain a warning signal to notify these phishing websites which helps the user to safeguard their identities and their login credentials etc. python is the language that helps to enable these techniques for the online users

3.	Novelty / Uniqueness	This project not only able to identify the malicious websites it also has the ability to automatically block these kind of websites completely in the future when it has been identified and also blocks some various mails /ads from these malicious websites
4.	Social Impact / Customer Satisfaction	This web phishing detection project attains the customer satisfaction by discarding various kinds of malicious websites to protect their privacy. This project is not only capable of using by an single individual ,a large social community and a organisation can use this web phishing detection to protect their privacy. This project helps to block various malicious websites simultaneously.
5.	Business Model (Revenue Model)	This developed model can be used as an enterprise applications by organisations which handles sensitive information and also can be sold to government agencies to prevent the loss of potential important data.
6.	Scalability of the Solution	This project's performance rate will be high and it also provide many capabilities to the user without reducing its efficieny to detect the malicious websites. thus scalability of this project will be high .

3.4 PROBLEM SOLUTION FIT


across devices.
Define CS, fit into CC
Explore AS, differentiate
Focus on J&P, tap into BE, understand RC
Focus on J&P, tap into BE, understand RC
1 item selected 97.1 KB

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 CONSTRAINTS CC Customers have very little awareness on phishing websites. They don't know what to do after losing data.	5. AVAILABLE SOLUTIONS AS Which solutions are available 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 affective 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
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.

4. REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS

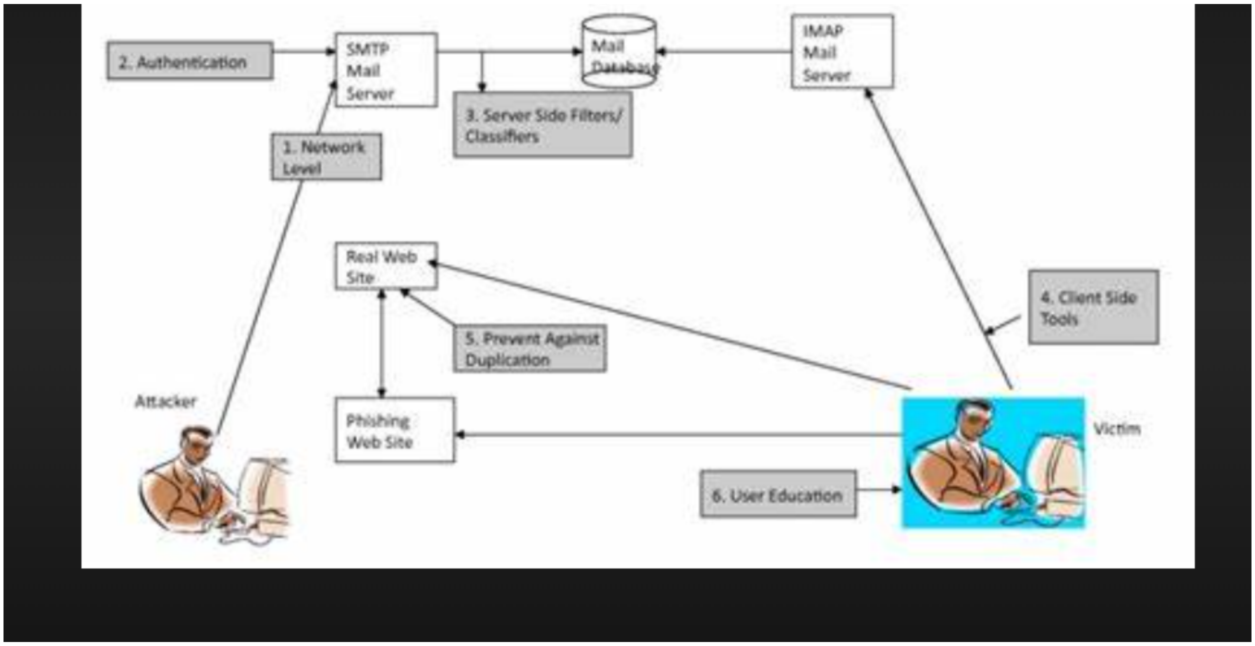
FR NO.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Input	User inputs an URL in required field to check its validation.
FR-2	Website Comparison	Model compares the websites using Blacklist and Whitelist approach.
FR-3	Feature extraction	After comparing, if none found on comparison then it extracts feature using heuristic and visual

		similarity approach.
FR-4	Prediction	Model predicts the URL using Machine Learning algorithms such as Logistic Regression, KNN
FR-5	Classifier	Model sends all output to classifier and produces final result.
FR-6	Announcement	Model then displays whether website is a legal site or a phishing site.
FR-7	Events	This model needs the capability of retrieving and displaying accurate result for website

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.



SOLUTION AND TECHNICAL ARCHITECTURE

SOLUTION ARCHITECTURE

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

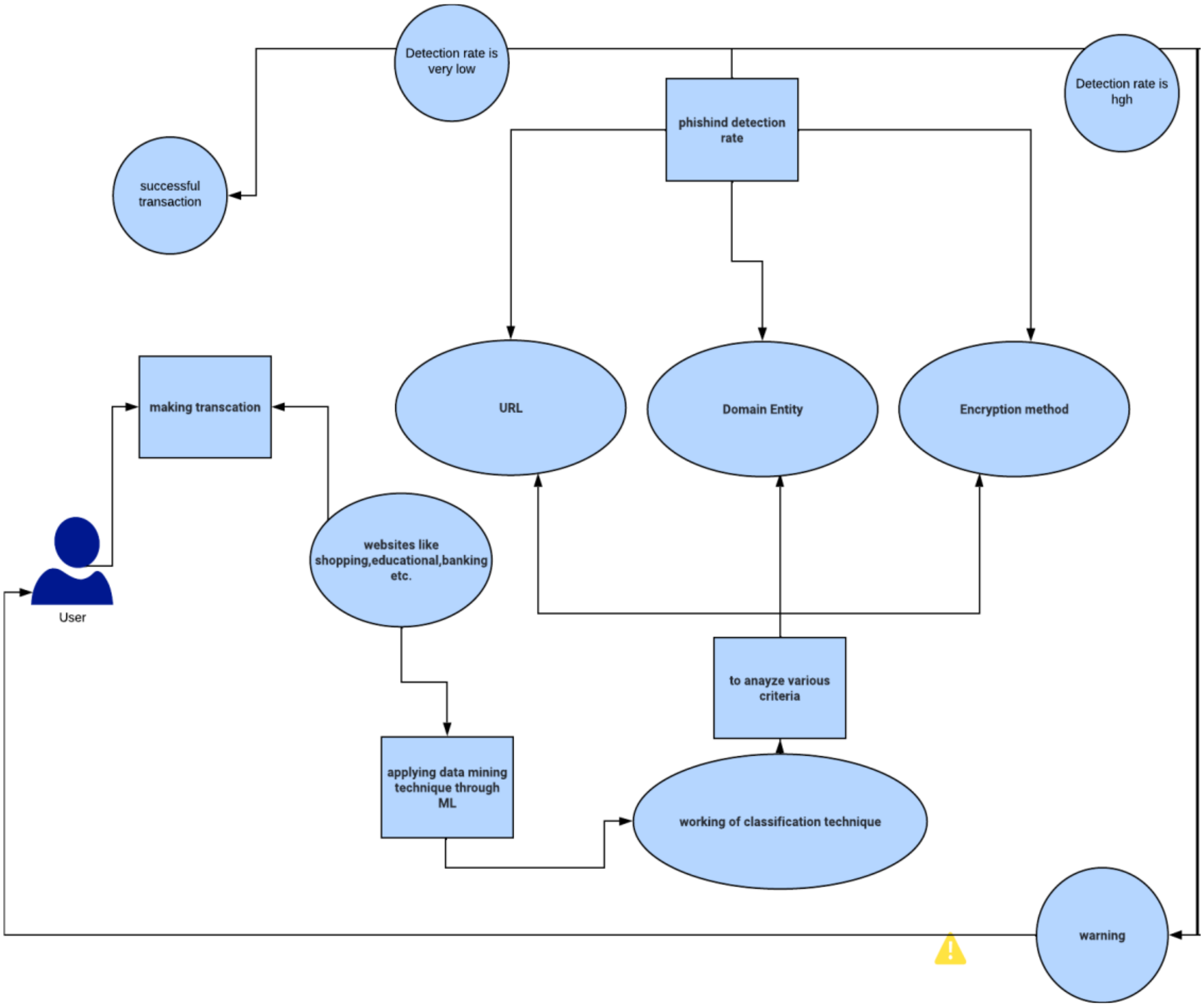
Find the best tech solution to solve existing business problems.

Describe the structure, characteristics, behavior, and other aspects of the software to project stakeholders.

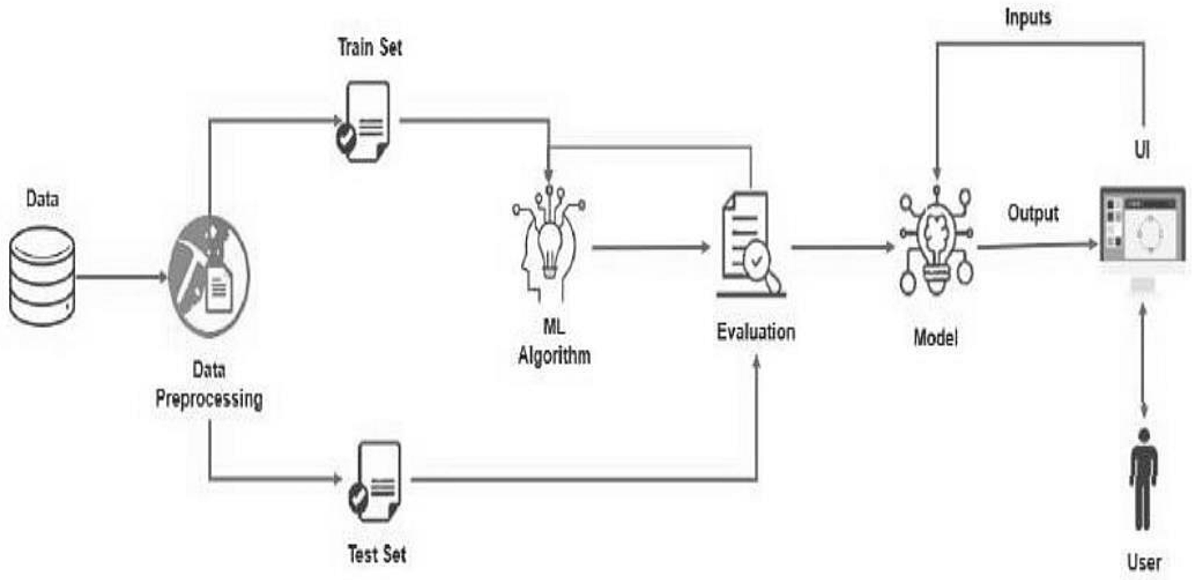
Define features, development phases, and solution requirements.

Provide specifications according to which the solution is defined, managed, and delivered.

Solution Architecture Diagram


Detection rate is very low
Detection rate is hgh
successful transaction
phishind detection rate
making transcation
URL
Domain Entity
Encryption method
websites like shopping,educational,banking etc.
User
to anayze various criteria
applying data mining technique through ML
working of classification technique
warning

TECHNICAL ARCHITECTURE


Inputs
Train Set
UI
Data
Output
ML
Algorithm
Evaluation
Model
Data
Preprocessing
Test Set
User

6.PROJECT PLANNING AND SCHEDULING

6.1 SPRINT PLANNING AND ESTIMATION

Product Backlog, Sprint Schedule, and Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points
Sprint-1	Homepage	USN-1	As a user, I can explore the resources of the homepage for the functioning	10
Sprint-1		USN-2	As a user, I can learn about the various sides of the web phishing and be aware of the scams	5
Sprint-2	Final page	USN-3	As a user, I can explore the resources of the final page for the functioning	15
Sprint-3	Prediction	USN-4	As a user, I can predict the URL easily for detecting whether the website is legitimate or not	10
Dashboard				
Sprint-4	Chat	USN-5	As a user, I can share the experience or contact the admin for the support	10
Sprint-1	Homepage	USN-6	As a admin, we can design interface and maintain the functioning of the website	5
Sprint-2	Final page	USN-7	As a admin, we can design the complexity of the website for making it user-friendly	5

Sprint-3	Prediction	USN-8	As a admin, we can use various ML classifier model for the accurate result for the detection of URL	10
Dashboard Sprint-4		USN-9	As a admin, we can response to the user message for improvement of the website	10

7.RESULTS

7.1 PERFORMANCE METRICS

Scikit-learn tool has been used to import Machine learning algorithms. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers.Performance of classifiers has been evaluated by calculating classifier's accuracy score.improve the accuracy of our models with better feature extraction

8.CONCLUsION

Thus to summarize, we have seen how phishing is a huge threat to the security and safety of the web and how phishing detection is an important problem domain. We have reviewed some of the traditional approaches to phishing detection; namely blacklist and heuristic evaluation methods, and their drawbacks. We have tested two machine learning algorithms on the Phishing Websites Dataset and reviewed their results. We then selected the best algorithm based on its performance and built a Chrome extension for detecting phishing web pages. The extension allows easy deployment of our phishing detection model to end users. We have detected phishing websites using Random Forest algorithm with and accuracy of 97.31%. For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned

9.FUTURE SCOPE

Although the use of URL lexical features alone has been shown to result in high accuracy (97%), phishers have learned how to make predicting a URL destination difficult by carefully manipulating the URL to evade detection. Therefore, combining these features with others, such as host, is the most effective approach .For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.

1o. APPENDIX

Python:

Python is an interpreted, high-level, general purpose programming language created by Guido Van Rossum and first released in 1991, Python's design philosophy emphasizes code Readability with its notable use of significant White space. Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically type and garbage collected. It supports multiple programming paradigms, including procedural, object oriented ,and functional programming.

Keras :

Keras is a powerful and easy-to-use free open source Python library for developing and evaluating **deep learning** model **.**It wraps the efficient numerical computation libraries **Theano** and **TensorFlow** and allows you to define and train neural network models in just a few lines of code. It uses libraries such as Python, C#,C++ or standalone machine learning toolkits. Theano and TensorFlow are very powerful libraries but difficult to understand neural network.Keras is based on minimal structure that provides a clean and easy way to create deep learning models based on TensorFlow or Theano. Keras is designed to quickly define deep learning models. Well, Keras
is an optimal choice for deep learning applications.

Steps for creating a keras model:

1)First we must define a network model.
2)Compile it, which transforms the simple sequence of layers into a complex group of matrix operations.
3)Train or fit the network.
To import: from keras.models import Sequential

From keras.layers import Dense, Activation, Dropout

TensorFlow:

TensorFlow is a Python library for fast numerical computing created and released by Google. It is a foundation library that can be used to create Deep Learning models directly or by using wrapper librarie sthat simplify the process built on top of **TensorFlow.** TensorFlow tutorial is designed for both beginner and professionals. Our tutorial provides all the basic and advanced concept of machine learning and deep learning concept such as deep neural network, image processing and sentiment analysis. TensorFlow is one of the famous deep learning frameworks, developed by **Google** Team. It is a free and
open source software library and designed in **Python** programming language, this tutorial is designedin such a way that we can easily implements deep learning project on TensorFlow in an easy andefficient way. Unlike other numerical libraries intended for use in Deep Learning like **Theano**,**TensorFlow** was designed for use both in research and development and in production systems. It canrun on single CPU systems, GPUs as well as mobile devices and largescale distributed systems ofhundreds of machines.

Numpy:

NumPy is a Python library used for working with arrays. It also has functions for working in domain oflinear algebra, Fourier transform, and matrices. Numpy which stands for Numerical Python, is a libraryconsisting of multidimensional array objects and a collection of routines for processing those arrays.Using NumPy, mathematical and logical operations on arrays can be performed. This tutorial explains
the basics of NumPy such as its architecture and environment. It also discusses the various array functions, types of indexing, etc. It is an opensource project and you can use it freely.
NumPy stands for Numerical Python. NumPyaims to provide an array object that is up to 50x faster than traditional Python lists. The array object inNumPy is called **ndarray**, it provides a lot of supporting functions that make working with **ndarray** very easy. Arrays are very frequently used in data science, where speed and resources arevery important.

Pillow:

Pillow is a free and ope nsource library for the Python programming language that allows you to easily create &s manipulate digital images. Pillow is built on top of PIL (Python Image Library). PIL is one of the important modules for image processing in Python. However, the PIL module is not supported since 2011 and does n't support python 3.
Pillow module gives more functionalities, runs on all major operating system and support for python

3. It supports wide variety of images such as “jpeg”, “png”, “bmp”, “gif”, “ppm”, “tiff”. You can do almost anything on digital images using pillow module. Apart from basic image processing functionality, including point operations, filtering images using built-in convolution kernels, and color space conversions.

Tkinkter:

Tkinter is the standard **GUI library** for Python. Python when combined with Tkinter provides a fast and easy way to create **GUI applications**. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.We need to import all the modules that we are going to need for training our model. The Keras library already contains some datasets and MNIST is one of them. So we can easily import the dataset through Keras. The mnist.load_data() method returns the training data, its labels along with the testing data and its labels.

Jupyter Notebook:

Jupyter Lab is a web-based interactive development environment for Jupyter notebooks, code, and data. JupyterLab is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.JupyterLab is extensible and modular: writeplugins that add new components and integrate with existing ones.

Machine Learning:

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Deep Learning:

Deep learning is an artificial intelligence (AI) function that imitates the workings of the human brain in processing data and creating patterns for use in decision making. Deep learning is a subset of machine learning in artificial intelligence that has networks capable of learning unsupervised from data that is unstructured or unlabeled. Also known as deep neural learning or deep neural network.

Neural Networks:

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.

GitHub link

https://github.com/IBM-EPBL/IBM-Project-36263-1660293762

PROJECT DEMO LINK

https://s33.aconvert.com/convert/p3r68-cdx67/vdsil-fyj57.mp4