Sardar Patel Institute of Technology

A Report On

“Fraud Webpage Detection using machine Learning”

Second Year Computer Engineering
Sardar Patel Institute of Technology

April 2021

A report submitted in partial fulfillment of the requirement of Mini Project Course

CERTIFICATE

This is to certify that the work on the project entitled “Fraud Webpage Detection using Machine Learning” has been carried out by the following students, who are bonafide students of Sardar Patel Institute of Technology, Mumbai, in partial fulfillment of the syllabus requirement in the Mini Project.

1. **Sejal Gurkhe (2019130017)**
2. **Ishita More (2019130039)**

Project Guide: _____

(Prof. Kiran Gawande)

Acknowledgement

We are pleased to present our project on the topic “Fraud Webpage Detection using Machine Learning” and take this opportunity to express our profound gratitude to all those people who guided us in this project.

First, we are thankful to the Sardar Patel Institute of Technology, Computer Engineering Department, and Professor Kiran Gawande and external guide professor Reeta Koshy for their guidance and support for our project work.

We would like to thank them for their patience, and faith in our capabilities and for giving us flexibility in working.

Table of Content

Introduction

Phishing is a fraudulent practice in which an attacker tries to obtain sensitive information by impersonating someone else to benefit himself/herself in a malicious way. Today, most of the users are accessing the services online, so it has become very easy for phishers to obtain user's confidential information. The website contents of phishing websites look very similar to that of legitimate websites and hence prompts people to provide their sensitive information.

In a conventional phishing attack, a user's sensitive information like usernames and passwords can be revealed by tricking the user to click on some link to a phishing website. The attacker may duplicate the authorized website and may have some click tactics on this website that can be used to trick a user so as to reveal their sensitive credentials. These credentials can further be made use of by the phisher for digital identity thefts or some financial profits as well.Phishing attacks can be prevented by making users distinguish between phishing and legitimate websites. s. Most of the phishers use images rather than text which are difficult to detect. Various tools and mechanisms have been developed to detect phishing websites and to prevent attacker from obtaining sensitive information. Blacklisting is one the easy way to detect phishing websites but can't be used to find new phishing websites. It is also a time consuming process.

In this paper, Logistic Regression and Multinomial naive Bayes model have been discussed which were implemented with accuracies of 96.369% and 95.74% respectively.

Problem Definition

Phishing websites are duplex web pages created to resemble real websites in-order to cheat people to get their personal information. Because of the plasticity of their tactics with little cost in detecting and identifying them. Web phishing intends to steal confidential data, such as usernames, keys, etc, by imitating a legitimate entity. It will lead to information exposure and property damage.

The purpose of the project is to detect fraud websites that might encounter. The task is to determine if a user can still be trusted or if it should be flagged for potential fraud through that website. Since we have data, we can use various ML algorithms to analyze the data and find possible results.

The aim of the project is to detect phishing URLs as well as narrow them down to best machine learning algorithm by comparing accuracy rate.

Literature Survey

Sr.no	Title	Publication	Short Description	link
1	Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation	2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)	Random Forest classification method, SVM classification algorithm, and Neural Network classification algorithm were enhanced. SVM classification algorithm gave better accuracy.SVM is selected as the final classifier algorithm.	https://ieeexplore.ieee.org/document/9277256
2	Phishing Website Detection Based on Machine Learning Algorithm	2020 International Conference on Computing and Data Science (CDS)	Out of Logistic regression classifier, SVM, Naive Bayes, Decision tree algorithm Logistic regression classifier is selected.	https://ieeexplore.ieee.org/document/9275957
3	Phishing Detection Using Machine Learning Technique	2020 First International Conference of Smart Systems and Emerging Technologies (SMART TECH)	The technique discussed, SVM which has high accuracy. The proposed technique can detect new temporary phishing sites and reduce the damage caused by phishing attacks.	https://ieeexplore.ieee.org/document/9283771
4	Detecting Phishing Website Using Machine Learning	2020 16th IEEE International Colloquium on Signal Processing & Its	This software discussed in this paper is designed to show awareness of the extensive level of its	https://ieeexplore.ieee.org/document/9068728

		Applications (CSPA)	functionality, features that can be displayed in the monitoring. The system fosters many features in comparison to other software.	

Scope of Work & Objectives

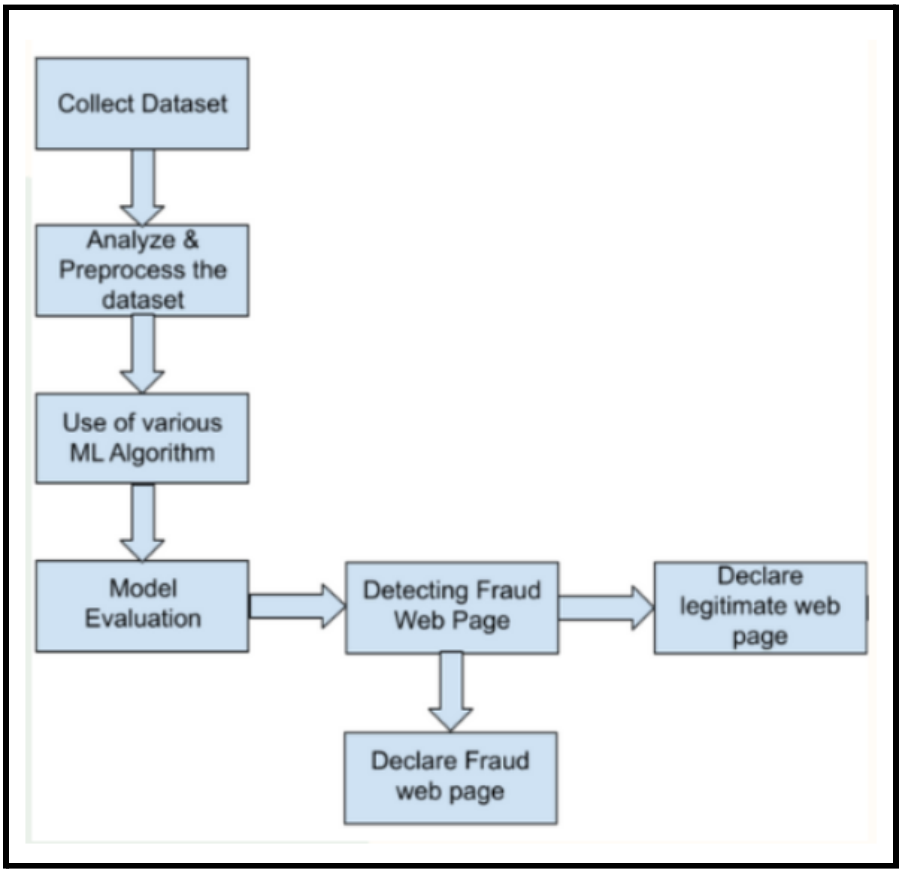
The objective of this project is

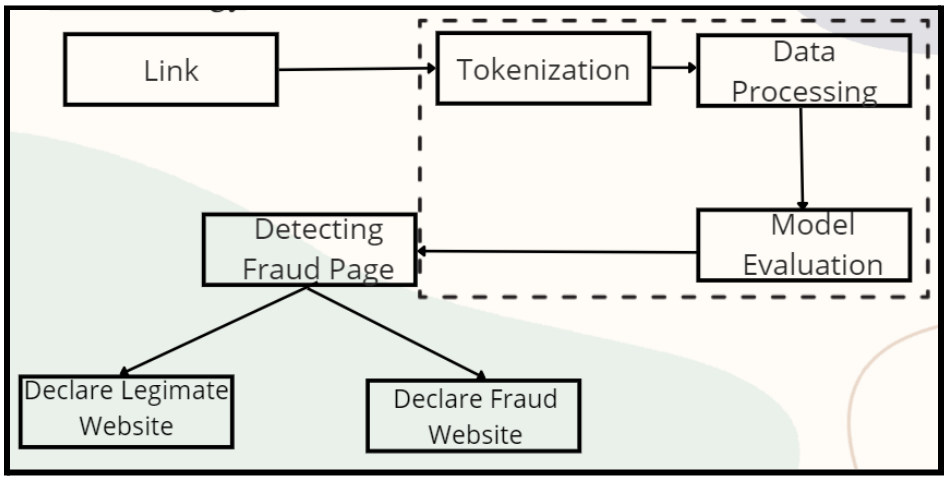
- To train machine learning models on the dataset created to predict fraud web pages.
- To use the dataset of fraud websites data and from that data use it to detect Fraud Web pages.
- The performance level of each algorithm is measured and compared.
- The objective of phishing website URLs is to purloin the personal information like user names, passwords and etc.

Machines are much fitter than humans at dealing with large datasets. They are apt to catch and recognize thousands of patterns on a user's purchasing journey instead of the few captured by creating rules.

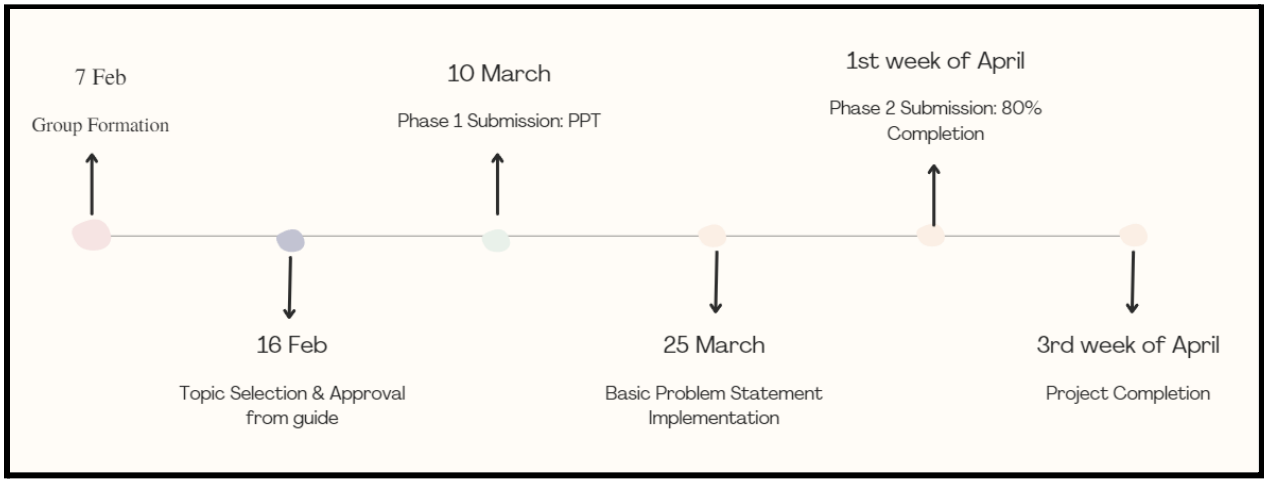
We can predict fraud in a large volume of actions by assigning cognitive computing technologies to crude data. That is why we use ML algorithms.

Block Diagram





Project Plan / Timeline



Implementation Details

The Project focuses on detecting fraud websites that might encounter. The task is to determine if a user can still be trusted or if it should be flagged for potential fraud through that website. Since we have data, we can use various ML algorithms to analyze the data and find possible results. Another way is where users will add the blacklisted word. If that word is encountered on the site, then the website is phishing. In short, the system will predict fraud websites.

Data Collection and Processing

The Dataset for the implementation of this system was taken from the Kaggle. The Dataset includes 2 attributes. The attributes include: websites and label. Label is Good or Bad, which is basically showing not phishing and phishing website.

Sr. no	Attributes	Description	Values
1	Website URL	Links	String
2	Label	-	Good: Legitimate Bad: Phishing

While creating the ML model, we used the target column as the class and another column as the features for the model, to predict whether the given website is fraudulent or not. The dataset is split up into training and testing in the ratio of 3:1. While training a classifier, it is needed to sort the data into two parts: training and testing sets. The training set will have targets, while the testing set won't contain these values.

The Algorithms like Logistic Regression, Multinomial naive Bayes are used for analysis purpose. From which, the Logistic Regression algorithm gives the best result. Features like link is the inputs that are given to the ML algorithm, which will be used to calculate an output value.

```
Training Accuracy : 0.9783572688946115
Testing Accuracy : 0.9638334898825517

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.91      0.96      0.93     36442
        Good       0.99      0.96      0.98    100895

    accuracy                           0.96    137337
   macro avg       0.95      0.96      0.95    137337
weighted avg       0.97      0.96      0.96    137337
```

The coding portion was carried out to prepare the data, visualize it, pre-process it, build the model and then evaluate it. The experiments and all the model building are done based on python libraries.

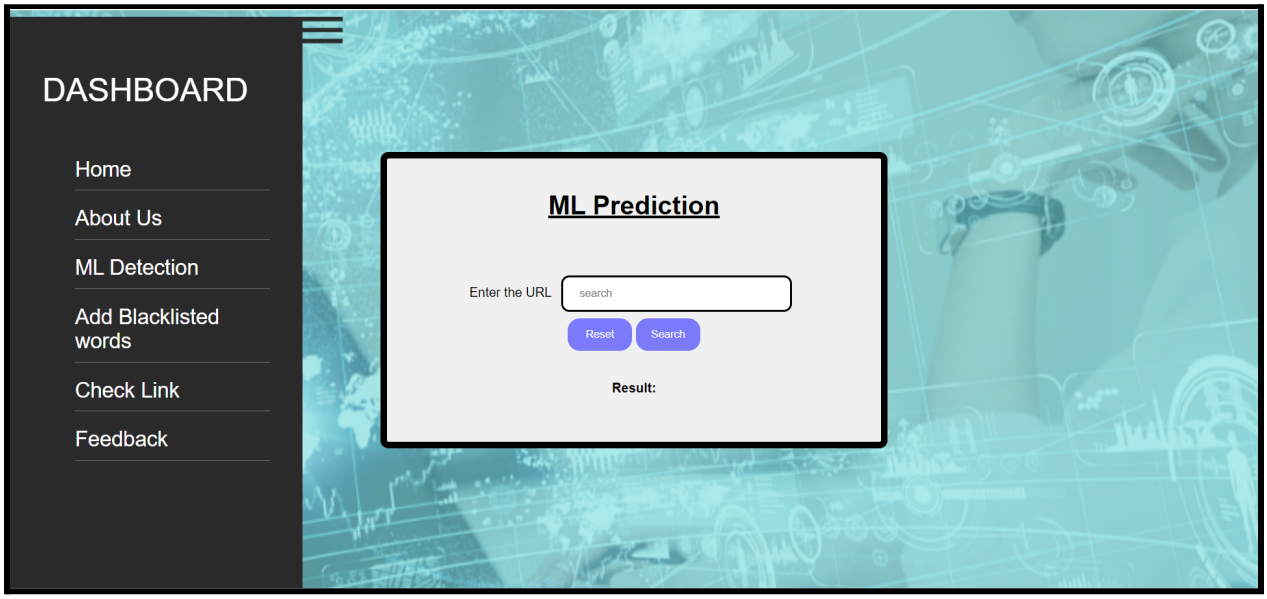
Model Deployment:

The web application is built using the Flask web application framework. We created the web app code (API) to load the model and get user input from the HTML template and made the prediction to return the result.

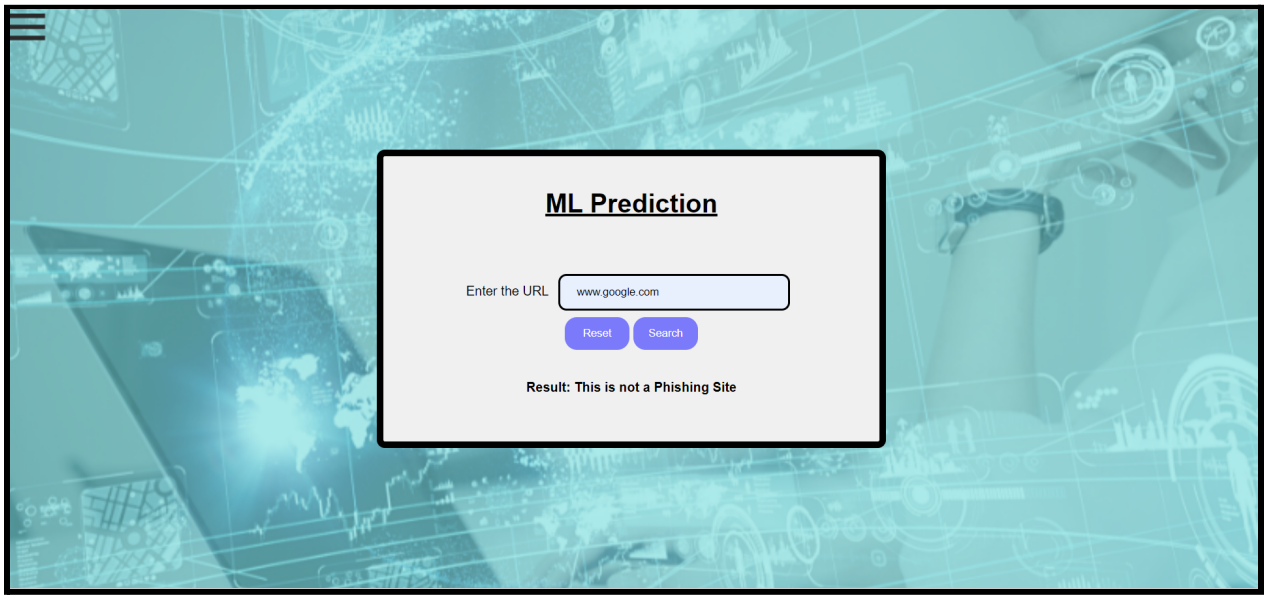
System Implementation

1. Home Page

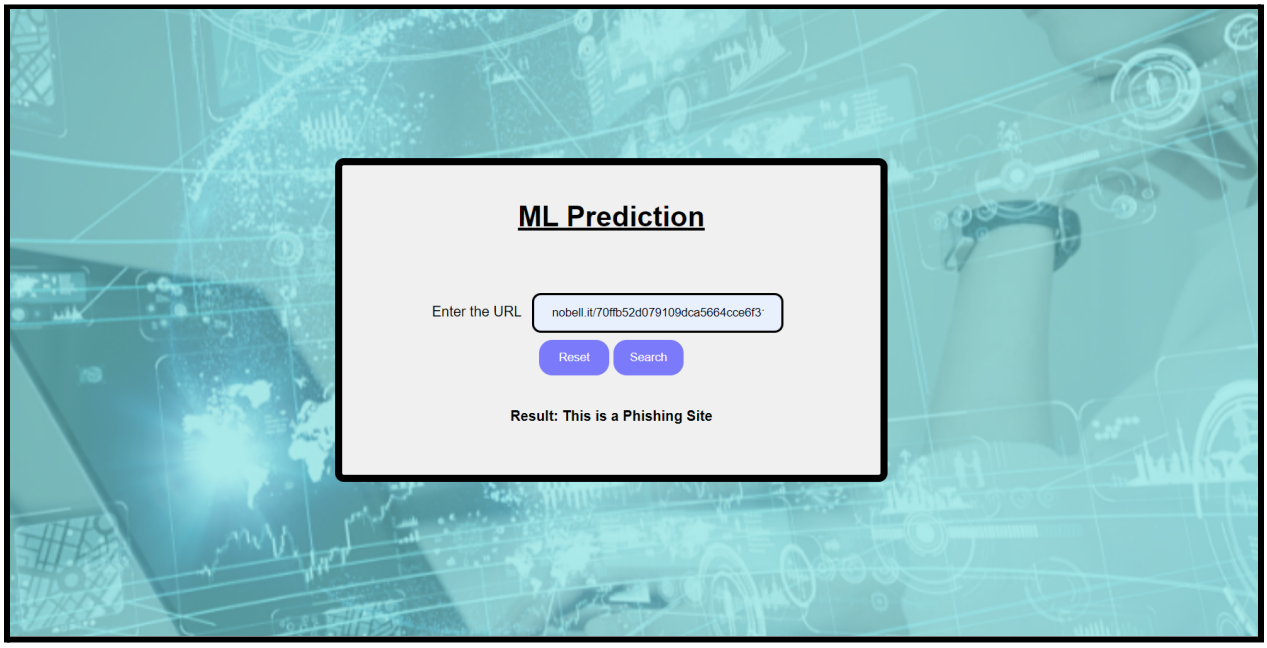
2. Prediction Page



3. Prediction where the input link is not a Phishing site



3. Prediction where the input link is a Phishing site



Future Work

For future enrichment, we aim to form the phishing detection system as an expandable web service that will integrate with online learning so that new phishing attack patterns can easily be learned and enhance the accuracy of our models with better feature extraction.

Conclusion

URL phishing analysis is very useful in determining whether a certain URL is a legitimate URL or not and whether it should be visited or not. This helps the users a lot in knowing which of the websites should be avoided. Thus, it prevents them from revealing their sensitive information to unknown or illegitimate sources.Since, Logistic Regression algorithm gave better accuracy as compared to that of Multinomial Naive Bayes algorithm, Logistic Regression is chosen as final classifier algorithm for classification of websites as phishing or legitimate.

References

[1] A. S. F. R. Smita Sindhu, Sunil Parameshwar Patil, “Phishing detection using random forest, svm and neural network with backpropagation,” International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020.

[2] W. Bai, “Phishing website detection based on machine learning algo-rithm,” International Conference on Computing and Data Science (CDS), 2020.

[3] M. W. N. T. N. Junaid Rashid, Toqeer Mahmood, “Phishing detection using machine learning technique,” First International Conference of Smart Systems and Emerging Technologies (SMART TECH) , 2020.

[4] Z. S. X. L. Feng Vue, Jianmin Pang, “Detection and prevention of phish- ing websites using machine learning approach,” 16th IEEE InternationalColloquium on Signal Processing Its Applications (CSPA), 2020.

[5] U. P. A. using Random Forest, “Url phishing analysis using random forest,” International Journal of Pure and Applied Mathematics , 2018.