Phishing Domain Detection

Low Level Design (LLD)

Ramchandra Tukaram Padwal

Revision Number: 1

Last date of revision: 28/05/2023

Document Version Control

Date Issued	Version	Description	Author
28/05/2023	1.1	First Draft	Ramchandra

Contents

Abstract

Phishing stands for a fraudulent process, where an attacker tries to obtain sensitive information from the victim. Usually, these kinds of attacks are done via emails, text messages, or websites. Phishing websites, which are nowadays in a considerable rise, have the same look as legitimate sites. However, their backend is designed to collect sensitive information that is inputted by the victim. Discovering and detecting phishing websites has recently also gained the machine learning community's attention, which has built the models and performed classifications of phishing websites.

1.Introduction

Why this Low-Level Design Document?

The purpose of this Low-Level Design (LLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The main objective of the project is to detect phishing domain detection which will affect any type to user.

- Security.
- Reliability
- Maintainability
- Portability
- Reusability
- Application compatibility
- Resource utilization

1.2 Scope

The LLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The LLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system. This software system will be a Web application This system will be designed to detect malicious vs real website.

1.3 Constraints

We will only be selecting detecting malicious phishing domains vs real.

1.4 Risks

Document specific risks that have been identified or that should be considered.

1.5 Out of Scope

Other than phishing domain detection security threats are out of scope.

2.Technical specifications

2.1 Dataset

Cases	Finalized	Source
Imbalance dataset	yes	ScienceDirect research paper
Balance dataset	Yes	ScienceDirect research paper

2.2 Dataset overview

The presented dataset was collected and prepared for the purpose of building and evaluating various classification methods for the task of detecting phishing websites based on the uniform resource locator (URL) properties, URL resolving metrics, and external services. The attributes of the prepared dataset can be divided into six groups:

- Whole URL
- Domain URL
- URL Directory
- URL File Name
- URL Parameters
- Resolving URL and external service
-

2.3 Logging

We should be able to log every activity done by the incidents.

- The System identifies at what step logging required
- The System should be able to log each and every system flow.
- System should not be hung even after using so many loggings. Logging just because we can easily debug issues, so logging is mandatory to do.

3.Technology stack

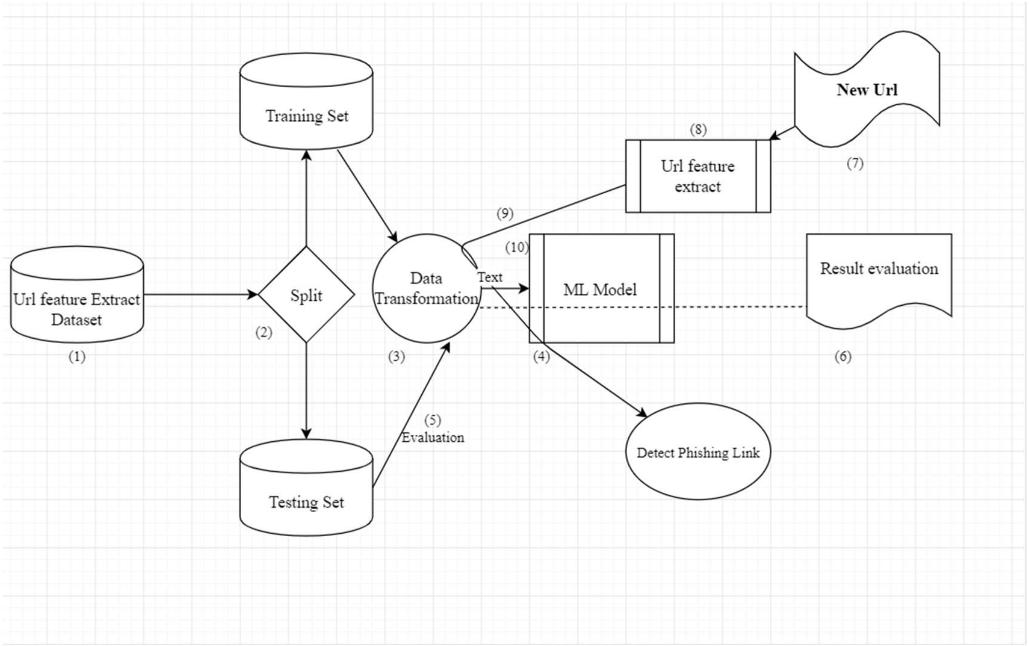
Front End	HTML/CSS
Backend	Python Flask
Visualization	Matplotlib, Seaborn
Dashboard	Pandas Profiling
Version control	GitHub

4. Proposed Solution

The solution proposed here is a website interface where we will input the URL of the website into a field to check if the URL is fake and malicious. Checking and detection goes with the help of machine learning algorithms.

1. Actual Model: Random Forest classifier.

5.Model training/validation workflow



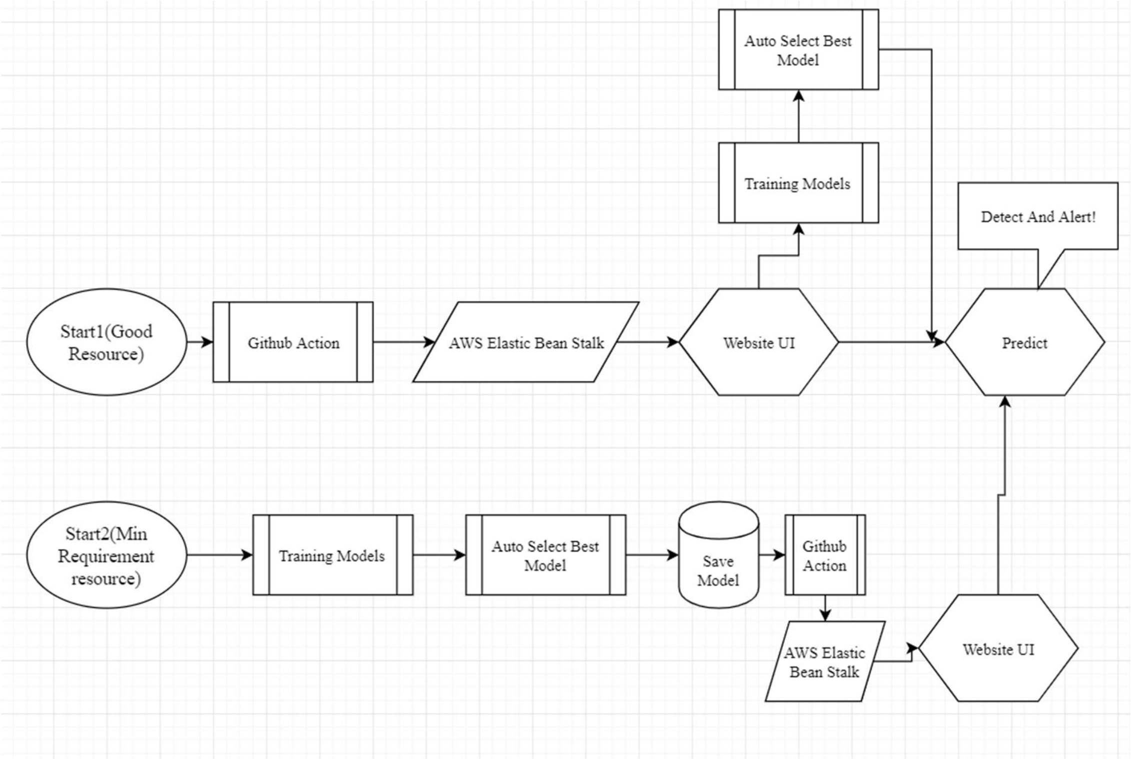
6. Requirements

6.1 Hardware Requirements

a. For full functionality it has to have a good resource server with minimum 2GB ram, 2 vcpu and above 10GB storage.

b. Minimum server requirement with resources 1GB ram and 8GB storage and 1 vcpu for detection working

7.User I/O workflow



8. Error Handling

Should errors be encountered, an explanation will be displayed as to what went wrong?

An error will be defined as anything that falls outside the normal and intended usage.

9.Test cases

Use case	Module	Accuracy
Detection	XG boost	65%
Detection	Random forest classifier	73%

10.Key performance indicators (KPI)

- Key indicators displaying a summary of the phishing domain detection.
- Show alert message legitimate vs phishing domain website

11. Conclusion

The phishing domain detection will detect a malicious website based on machine learning models. After this user will be safe from any loss or safe internet browsing.