**LOW LEVEL DESIGN (LLD)**

**PHISHING DOMAIN DETECTION**

**(PDD)**

Revision Number: 1

Last date of revision: 12/03/2023

Shaik mahammad rafi

# Document Version Control

**Contents**

Document Version Control 2

Abstract 5

1 Introduction 6

1.1 Why this Low-Level Design Document? 6

1.2 Scope 7

1.3 Constraints 7

1.4 Risks 6

1.5 Out of Scope 7

2. Technical specifications 11

2.1 Dataset 11

2.2 Dataset overview 11

2.3 Input schema 11

2.4 Logging 11

2.5 Database 11

3. Deployment 11

4. Technology stack 12

5. Proposed Solution 13

6 Model training/validation workflow 14

9.Error Handling 18

10.Test Cases 19

11..Conclusion 20

**Abstract**

Phishing is a type of fraud in which an attacker impersonates a reputable company or

person in order to get sensitive information such as login credentials or account

information via email or other communication channels. Phishing is popular among

attackers because it is easier to persuade someone to click a malicious link that appears

to be authentic than it is to break through a computer's protection measures.

# 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.

Phishing involves attempts by Internet fraudsters to access and obtain personal and sensitive information, such as usernames, passwords, and financial information, by utilizing social engineering techniques. Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user’s computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

implement the following use cases.

* + - Whether the website is legitimate or not.

Computer security enthusiasts can find these datasets interesting for building firewalls, intelligent ad blockers, and malware detection systems.

## 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 unusual activity , and fire disasters.

## 1.3 Constraints

We will find the website is legitimate or not.

## 1.4 Risks

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

## 1.5 Out of Scope

Delineate specific activities, capabilities, and items that are out of scope for the project.

# 2.Technical specifications

## 2.1 Dataset

|  |  |  |
| --- | --- | --- |
| cases | Finalized | source |
| Detecting website is legitimate or not | yes | Phishing domain detection website link.  https://data.mendeley.com/datasets/72ptz43s9v/1 |

## 2.2 Dataset overview

These data consist of a collection of legitimate as well as phishing website instances. Each website is represented by the set of features which denote, whether website is legitimate or not. Data can serve as an input for machine learning process.

Total number of instances: 88,647 Number of legitimate website instances (labeled as 0): 58,000 Number of phishing website instances (labeled as 1): 30,647 Total number of features

1. URL-Based Features
2. Domain-Based Features
3. Page-Based Features
4. Content-Based Features

## URL-Based Features

URL is the first thing to analyse a website to decide whether it is a phishing or not. As we mentioned before, URLs of phishing domains have some distinctive points. Features which are related to these points are obtained when the URL is processed. Some of URL-Based Features are given below.

* Digit count in the URL
* Total length of URL
* Checking whether the URL is Typosquatted or not. (google.com → goggle.com)
* Checking whether it includes a legitimate brand name or not (apple-icloud-login.com)
* Number of subdomains in URL
* Is Top Level Domain (TLD) one of the commonly used one?

## Domain-Based Features

The purpose of Phishing Domain Detection is detecting phishing domain names. Therefore, passive queries related to the domain name, which we want to classify as phishing or not, provide useful information to us. Some useful Domain-Based Features are given below.

* Its domain name or its IP address in blacklists of well-known reputation services?
* How many days passed since the domain was registered?
* Is the registrant name hidden?

## Page-Based Features

Page-Based Features are using information about pages which are calculated reputation ranking services. Some of these features give information about how much reliable a web site is. Some of Page-Based Features are given below.

* Global Pagerank
* Country Pagerank
* Position at the Alexa Top 1 Million Site

Some Page-Based Features give us information about user activity on target site. Some of these features are given below. Obtaining these types of features is not easy. There are some paid services for obtaining these types of features.

* Estimated Number of Visits for the domain on a daily, weekly, or monthly basis
* Average Pageviews per visit
* Average Visit Duration
* Web traffic share per country
* Count of reference from Social Networks to the given domain
* Category of the domain
* Similar websites etc.

## Content-Based Features

Obtaining these types of features requires active scan to target domain. Page contents are processed for us to detect whether target domain is used for phishing or not. Some processed information about pages are given below.

* Page Titles
* Meta Tags
* Hidden Text
* Text in the Body
* Images etc.

By analysing these information, we can gather information such as;

* Is it required to login to website
* Website category
* Information about audience profile etc.

All of features explained above are useful for phishing domain detection. In some cases, it may not be useful to use some of these, so there are some limitations for using these features.

## 

## 

## 

## 

## 2.4 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.
* Developers can choose logging methods. You can choose database logging/ File logging as well.
* 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.

## 2.5 Database

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

1. The User chooses the activity dataset.

2. The User gives required information.

3. The system stores each and every data given by the user or received on request to the database. Database you can choose your own choice whether MongoDB/ MySQL.

**3. Deployment**

1. AWS 2. 3.

# 4.Technology stack

|  |  |
| --- | --- |
| **Database** | MongoDB/MySql |
| **Deployment** | AWS |
| **Visualization** | Matplotlib,Seaborn ,Plotly |
| **Dashboard** | Tableau/Power BI |
| **version control** | GitHub |

# 5. Proposed Solution

Phishing involves attempts by Internet fraudsters to access and obtain personal and sensitive information, such as usernames, passwords, and financial information, by utilizing social engineering techniques. Typically a victim receives a message that appears to have been sent by a known contact or organization. The message contains malicious software targeting the user’s computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.

# 6.Model training/validation workflow

### Model Training and Evaluation

New data

Train dataset

Dataset

New data prediction

Model building

EDA and PREPROCESSING

Test dataset

Training

solitting Evaluation

## 7