High Level Design (HLD)

Phishing Domain Detection

Revision Number: 1.0

Last date of revision: 03/11/2022

# Document Version Control

|  |  |  |  |
| --- | --- | --- | --- |
| **Date Issued** | **Version** | **Description** | **Author** |
| 03/11/2022 | 1 | Initial HLD – V1.0 | Triveni M |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**Contents**

[Document Version Control 2](#_heading=h.gjdgxs)

[Abstract 4](#_heading=h.30j0zll)

1. [Introduction 5](#_heading=h.1fob9te)
   1. [Why this High-Level Design Document? 5](#_heading=h.3znysh7)
   2. [Scope 5](#_heading=h.2et92p0)
   3. [Definitions 5](#_heading=h.tyjcwt)
2. [General Description 6](#_heading=h.3dy6vkm)
   1. [Problem statement 6](#_heading=h.4d34og8)
   2. [Proposed Solution 6](#_heading=h.2s8eyo1)
   3. [Data Requirements 7](#_heading=h.26in1rg)
   4. [Tools used 8](#_heading=h.lnxbz9)
   5. [Constraints 9](#_heading=h.44sinio)
   6. [Assumptions 9](#_heading=h.2jxsxqh)
3. [Design Details 10](#_heading=h.z337ya)
   1. [Process Flow 10](#_heading=h.3j2qqm3)
      1. [Model Training and Evaluation 10](#_heading=h.1y810tw)
      2. [Deployment Process 11](#_heading=h.4i7ojhp)
   2. [Event log 11](#_heading=h.2xcytpi)
   3. [Error Handling 11](#_heading=h.1ci93xb)
   4. [Performance 12](#_heading=h.3whwml4)
   5. [Reusability 12](#_heading=h.2bn6wsx)
   6. [Application Compatibility 12](#_heading=h.qsh70q)
   7. [Resource Utilization 12](#_heading=h.3as4poj)
   8. [Deployment 12](#_heading=h.1pxezwc)
4. [Dashboards 13](#_heading=h.49x2ik5)
   1. [KPIs (Key Performance Indicators) 13](#_heading=h.2p2csry)
5. [Conclusion 14](#_heading=h.147n2zr)

# 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.

# Introduction

## Why this High-Level Design Document?

The purpose of this High-Level Design (HLD) 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 HLD will:

* + - Present all of the design aspects and define them in detail
    - Describe the user interface being implemented
    - Describe the hardware and software interfaces
    - Describe the performance requirements
    - Include design features and the architecture of the project
    - List and describe the non-functional attributes like:
      * Security
      * Reliability
      * Maintainability
      * Portability
      * Reusability
      * Application compatibility
      * Resource utilization
      * Serviceability

## Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

## Definitions

|  |  |
| --- | --- |
| *Term* | *Description* |
| *Database* | Collection of all the information monitored by this system |
| *IDE* | Integrated Development Environment |
| *AWS* | Amazon Web Services |

# General Description

#### Problem statement

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

person 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.

The mail goal is to predict whether the domains are real or malicious.

#### PROPOSED SOLUTION

Refer: <https://www.sciencedirect.com/science/article/pii/S2352340920313202>

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Try out different machine learning algorithms that’s best fit for the above case.

For Feature Engineering show:-

1. URL-Based Features

2. Domain-Based Features

3. Page-Based Features

4. Content-Based Features

Baseline Model: Logistic Regression since this is a classification problem.

Actual model: XG Boast

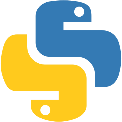
## Data Requirements

Data requirements completely depend on our problem statement.

* + - We need the data to be in the csv file format
    - We require all the columns to be non-negative values
    - Phishing columns to be having 0 or 1 which is categorical and the remaining columns are expected to be either integers or float

## Tools used

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn , are used to build the whole model.





* + - PyCharm is used as IDE.
    - For visualization of the plots, Matplotlib, Seaborn and Plotly are used.
    - AWS is used for deployment of the model.
    - Cassandra is used to retrieve, insert, delete, and update the database.
    - Front end development is done using HTML/CSS
    - Python Flask is used for backend development.
    - GitHub is used as version control system.

## Assumptions

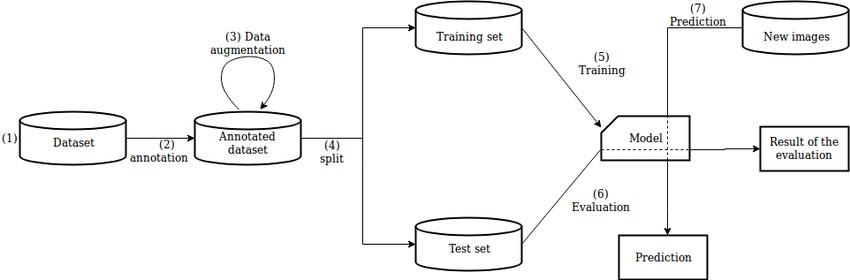
For the phishing websites, only the ones from the Phish Tank registry were included, which are verified from multiple users. For the legitimate websites, we included the websites from publicly available, community labelled and organized lists [1], and from the Alexa top ranking websites.

# Design Details

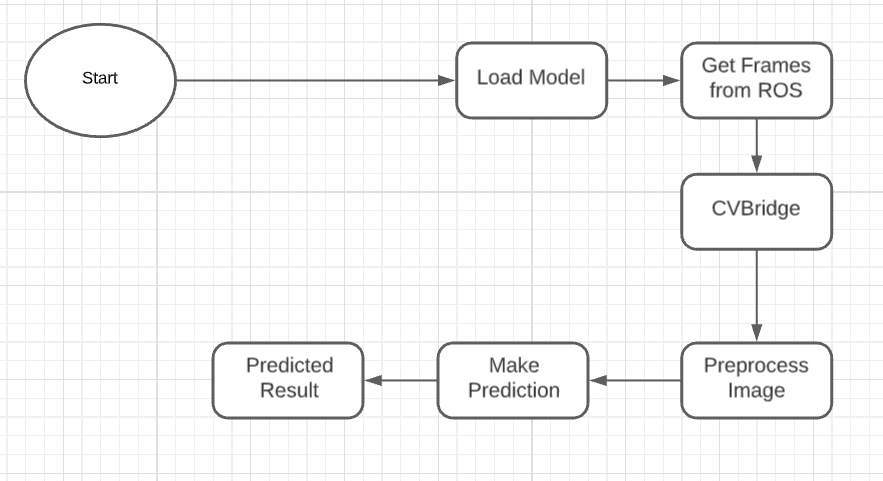
## Process Flow

For identifying the phishing domains, we will use machine learning classification model. Below is the process flow diagram is as shown below.

### Model Training and Evaluation



### Deployment Process



## Event log

The system should log every event so that the user will know what process is running internally.

**Initial Step-By-Step Description:**

1. The System identifies at what step logging required
2. The System should be able to log each and every system flow.
3. Developer can choose logging method. You can choose database logging/ File logging as well.
4. System should not hang even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

## 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.

# Performance

The phishing Domain detection is used for detection of malicious websites to prevent the cyber crimes. Also, model retraining is very important to improve the performance.

## Reusability

The code written and the components used should have the ability to be reused with no problems.

## Application Compatibility

The different components for this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

## Resource Utilization

When any task is performed, it will likely use all the processing power available until that function is finished.

* 1. **Deployment**



# Dashboards

Dashboards will be implemented to display and indicate certain KPIs and relevant indicators for the unveiled problems that if not addressed in time could cause catastrophes of unimaginable impact.



As and when, the system starts to capture the historical/periodic data for a user, the dashboards will be included to display charts over time with progress on various indicators or factors.

## KPIs (Key Performance Indicators)

Key indicators displaying a summary of phishing domain by showing 0 or 1 where 1 means the given url is malicious and 0 means the given url is safe to access

# Conclusion

The Phishing Domain Detection will detect malicious domains based on various parameters for the given url data used to train our algorithm, so we can prevent the cyber crimes and imbalance in the society in early stages and can take necessary action to stop them immediately

# References

1. https://www.sciencedirect.com/science/article/pii/S2352340920313202