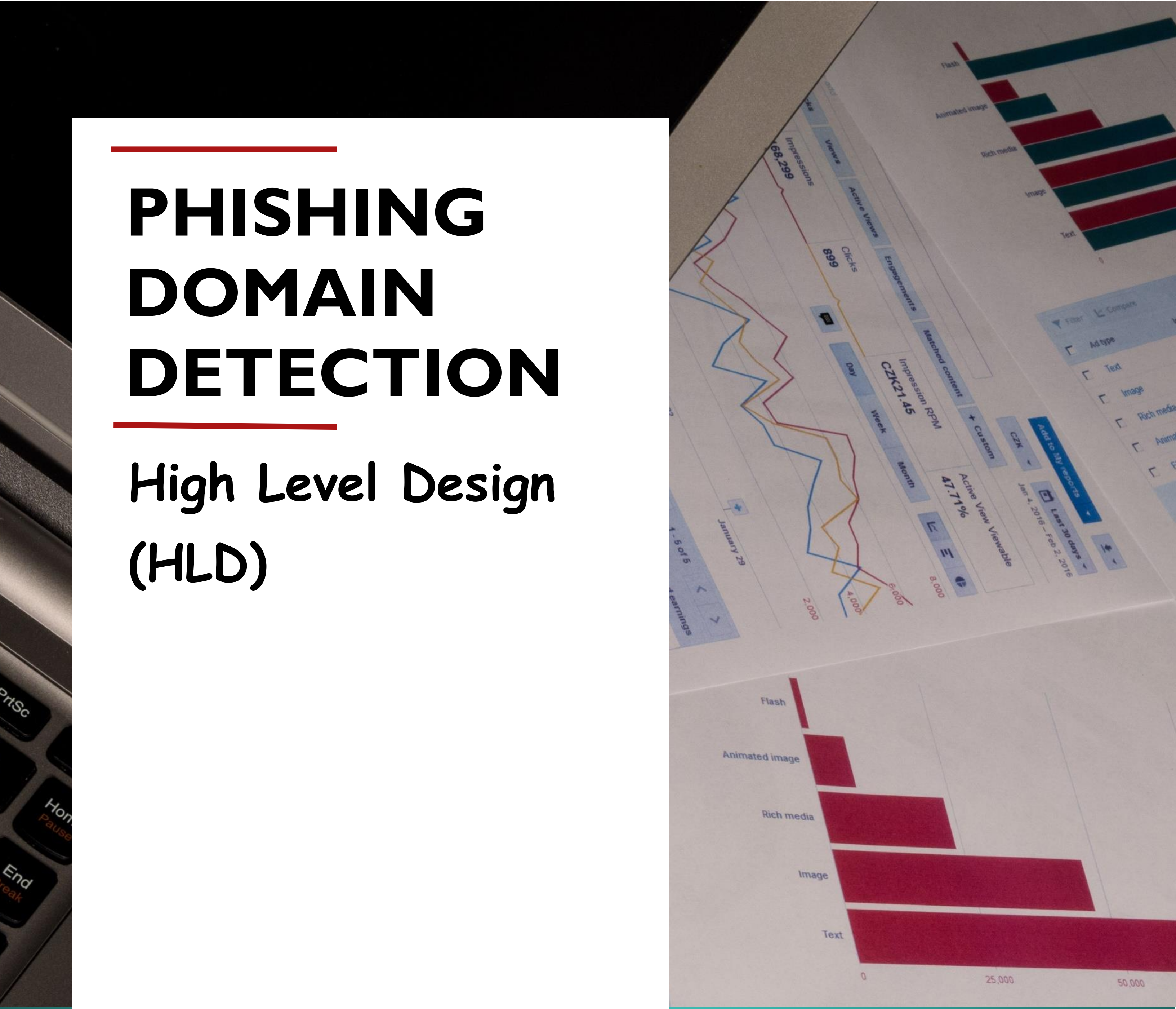


PHISHING DOMAIN DETECTION

High Level Design (HLD)

NOVEMBER 3

iNeuron.ai | PW Skills
Authored by: Suyash Waykar

Abstract

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

Approach -

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

Dataset -

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.
This data consists of a collection of legitimate as well as phishing website instances. Each website is represented by a set of features which denote whether website is legitimate or not. Data can serve as an input for machine learning process. In this repository the two variants of the Phishing Dataset are presented.

Full variant - dataset_full.csv
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: 111

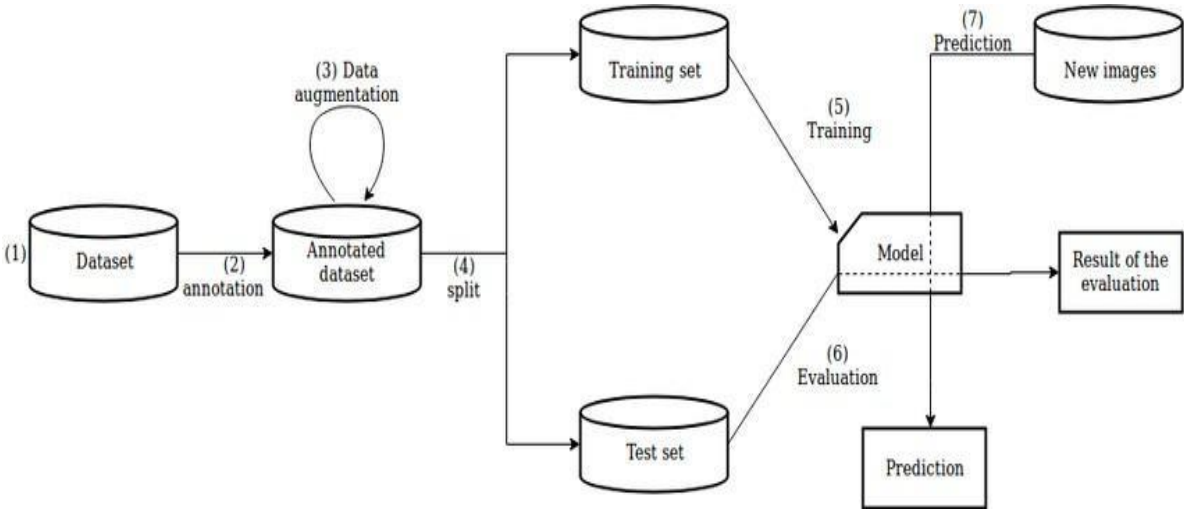
Small variant - dataset_small.csv
Total number of instances: 58,645
Number of legitimate website instances (labeled as 0): 27,998.
Number of phishing website instances (labeled as 1): 30,647.
Total number of features: 11

Tools Used -

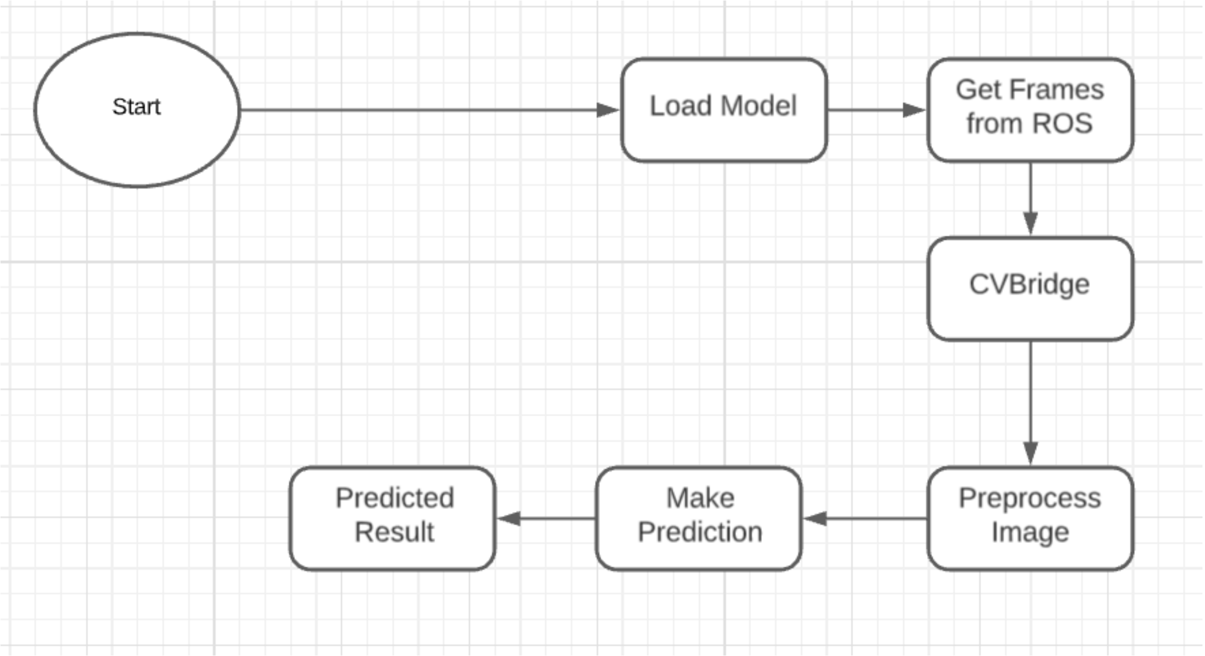
- Pandas
- NumPy
- Flask
- Cassandra
- Matplotlib
- Plotly
- Scikit-Learn
- HTML
- GitHub
- PyCharm
- Python Programming

Design Flow -

Model training and Evaluation Diagram -



Development process -



Design Details -

The system logs 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 is required.
2. The System can log every system flow.
3. Developers can choose logging methods. You can choose database logging/ File logging as well.
4. System doesn't hang even after using so many loggings. Logging just because we can easily debug issues, so logging is mandatory to do.

Errors are being 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 -

- Phishing Domain detection is used for detection of malicious websites to prevent cyber-crimes. Also, model retraining is very important to improve performance.
- The code written and the components used should have the ability to be reused with no problems.
- 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.
- When any task is performed, it will likely use all the processing power available until that function is finished.
- 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.
- Key indicators display 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 cyber-crimes and imbalance in the society in early stages and can take necessary action to stop them immediately.

THANKS!