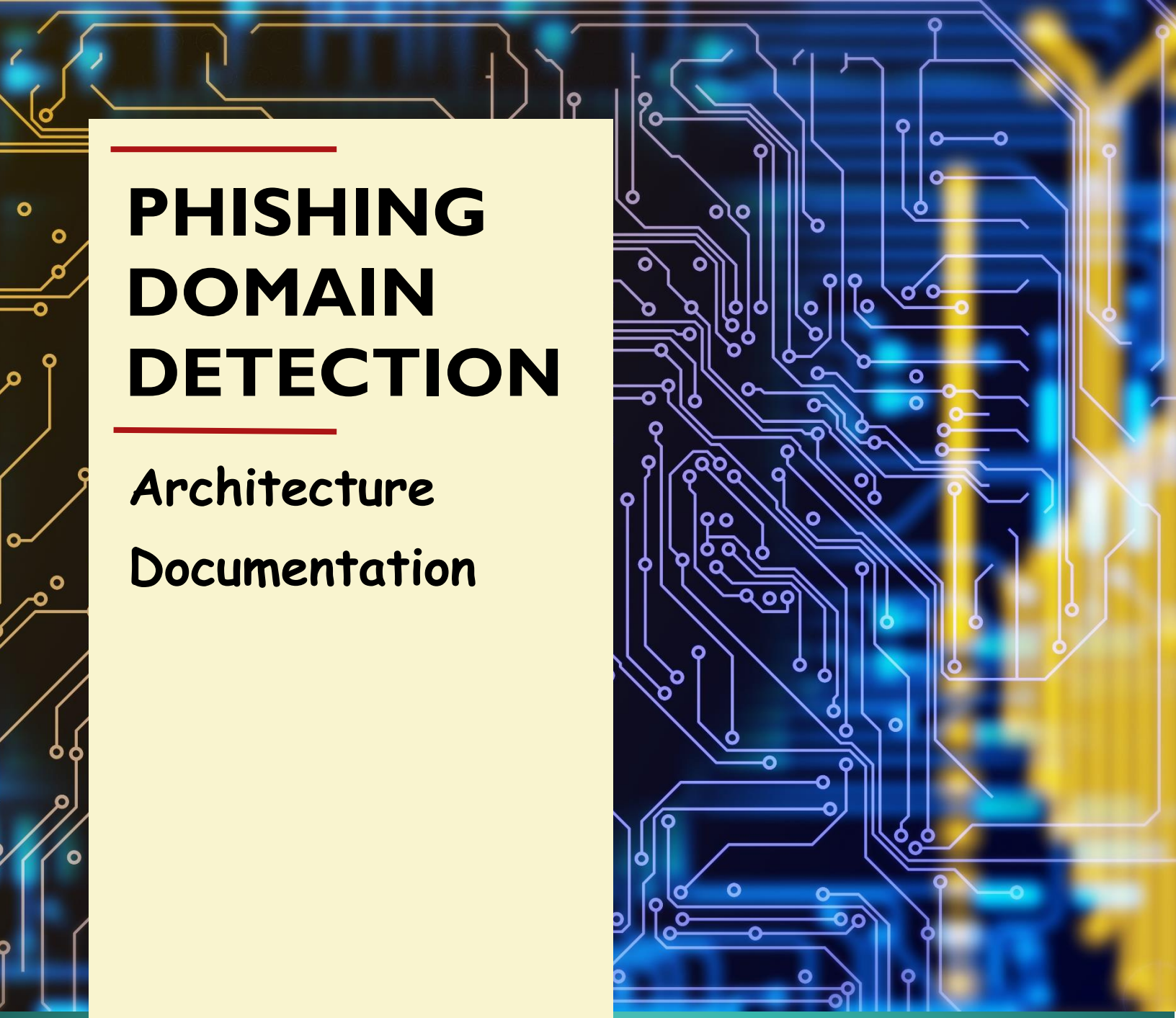


PHISHING DOMAIN DETECTION

Architecture Documentation

NOVEMBER 5

iNeuron.ai | PW Skills

Authored by: Suyash Waykar

Abstract

What is Architecture Design?

Architecture Design (AD) aims to give the internal design of the actual program code for the 'Phishing Domain Detection'. AD describes the class diagrams with the methods and relation between classes & program specifications. It describes the modules so that the programmer can directly code the program from the document.

Scope -

AD is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code and ultimately performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work and the complete workflow.

This software system will be a Web application. This system will be designed to detect phishing websites and prevent cyber-attacks and improve cyber security.

Constraints -

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.

Out of Scope -

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

Project Details

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 -

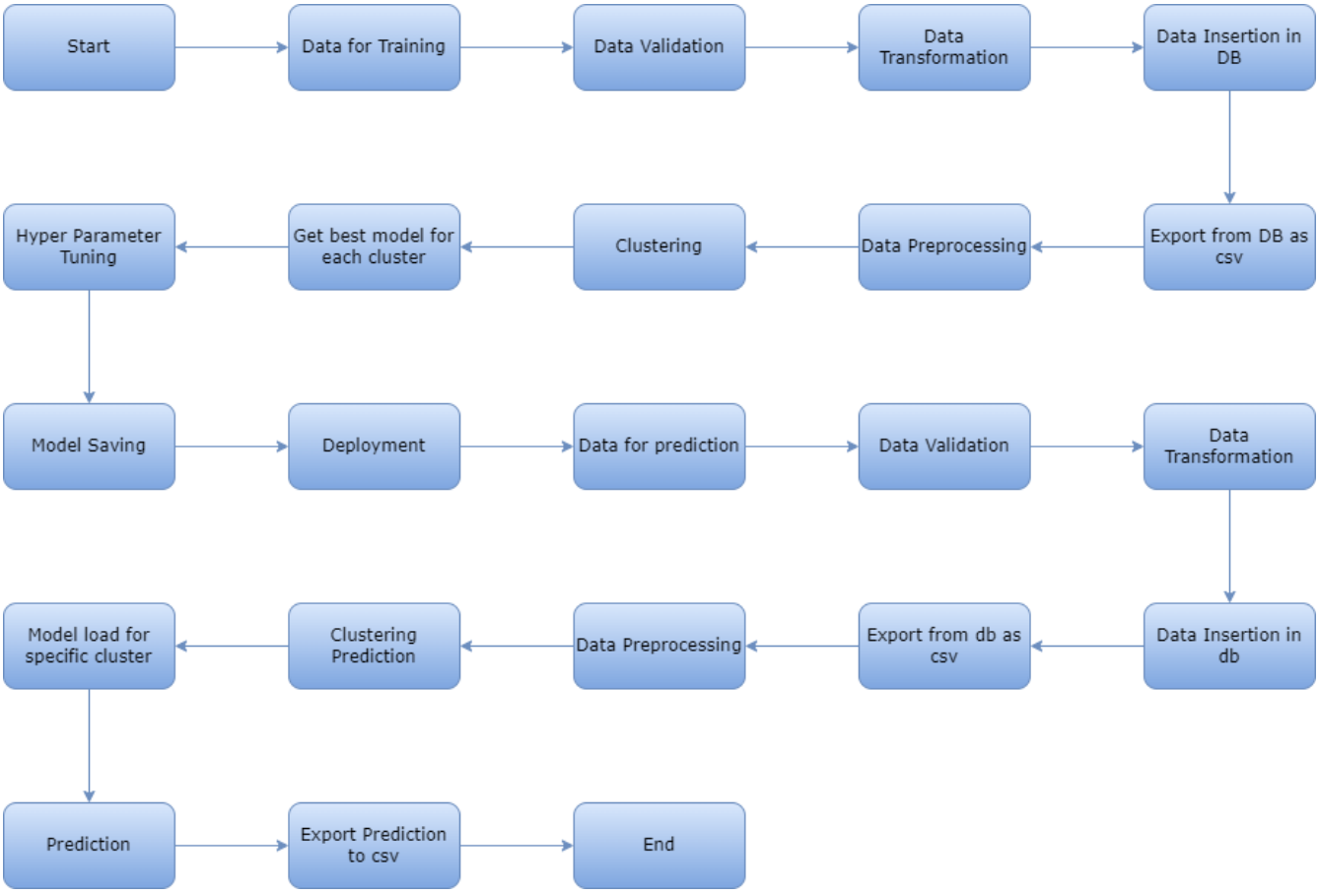
Phishing websites, which are nowadays in a considerable rise, have the same look as legitimate sites. However, their backend is designed to collect sensitive information that is inputted by the victim.
This data consists of a collection of legitimate as well as phishing website instances. Each website is represented by a set of features which denote whether the 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.

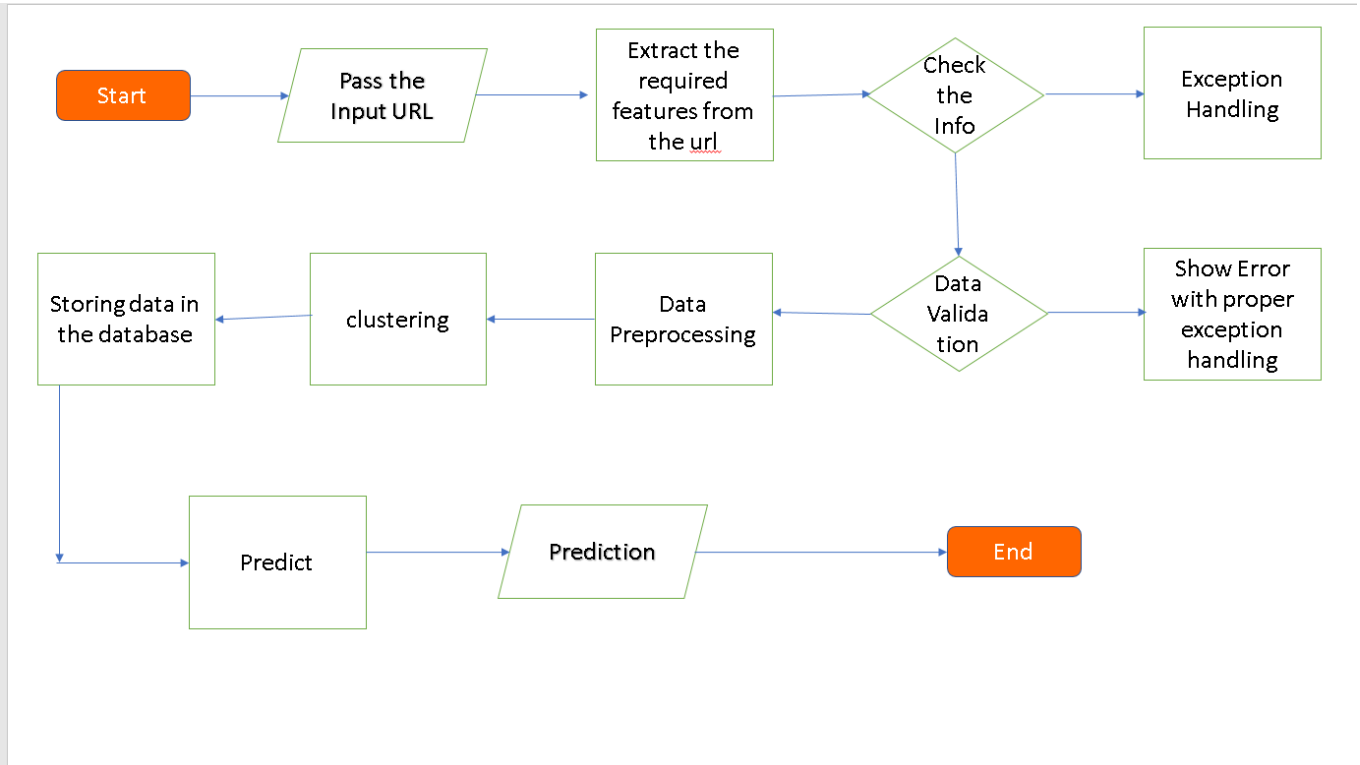
Number of phishing website instances (labeled as 1): 30,647.
Total number of features: 111

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

Architecture



User I/O Workflow



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!