Phishing Domain Detection

Architecture Design Document

Ramchandra Tukaram Padwal

Revision Number: 1.0

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.

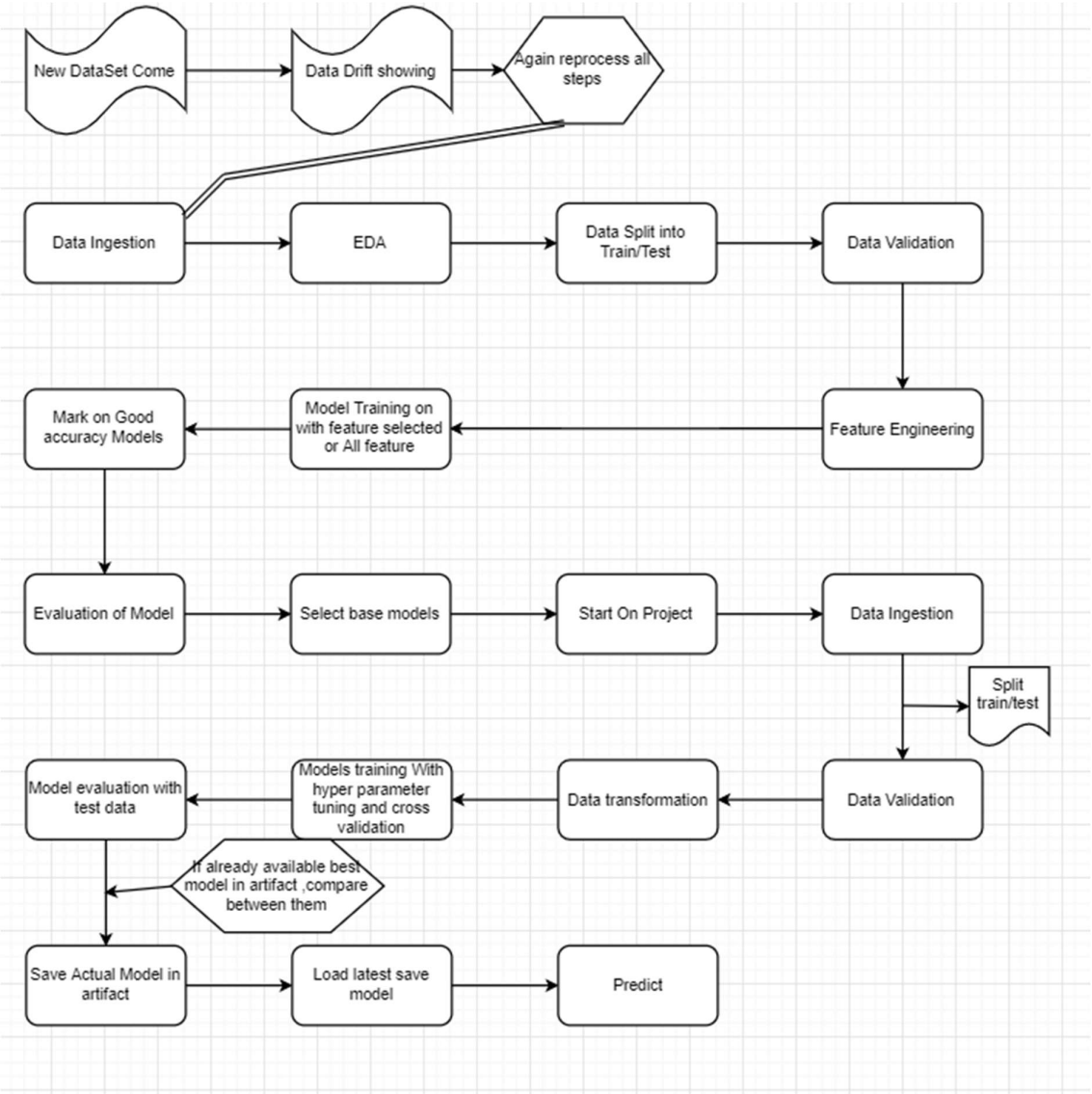
Introduction

Why this Architecture Design Document?

This document describes the architecture of the phishing domain detection system.

It describes:

- A general description of the system
- The logical architecture of software, the layers and top-level components
- The physical architecture of the hardware on which runs the software
- The justification of technical choices made
- The traceability between the architecture and the system requirements.



This project shall be delivered in two phases:

Phase 1: All the functionalities with PyPi packages.

Phase2: Integration of UI to all the functionalities.

Scope

This software system will be a Web application This system will be designed to detect the phishing sites at earliest for better safe browsing.

Constraints

We will only be shown legitimate websites and phishing domain detection.

Risks

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

Out of Scope

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

Technical specifications

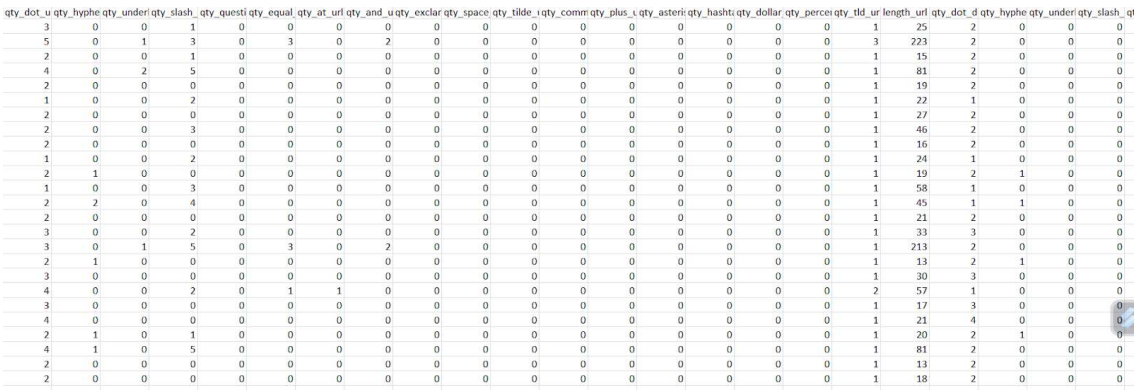
2.1 Dataset

Detection	Finalized	Source
Balance (Small)	No	https://www.sciencedirect.com/science/article/pii/S2352340920313202
Unbalanced (Full)	Yes	https://www.sciencedirect.com/science/article/pii/S2352340920313202

2.1.1 Phishing detection dataset overview

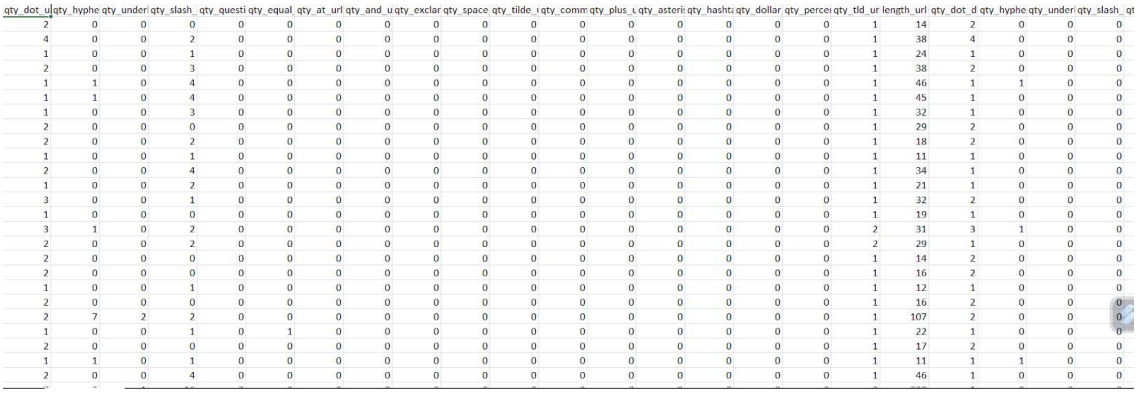
Two dataset variations that consist of 58,645 and 88,647 websites labelled as legitimate or phishing and allow the researchers to train their classification models, build phishing detection systems, and mining association rules.

- Full dataset table



qty_dot_u	qty_hyphe	qty_under	qty_slash_	qty_questi	qty_equal	qty_at_url	qty_and_u	qty_exclar	qty_space	qty_tilde_	qty_comm	qty_plus_	qty_asteri	qty_hashti	qty_dollar	qty_percei	qty_tld_ur	length_url	qty_dot_d	qty_hyphe	qty_under	qty_slash_	qt
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	2	0	0	0	
5	0	1	3	0	3	0	2	0	0	0	0	0	0	0	0	0	3	223	2	0	0	0	
2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	2	0	0	0	
4	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	81	2	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	2	0	0	0	
1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	22	1	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	27	2	0	0	0	
2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	46	2	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	2	0	0	0	
1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	24	1	0	0	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	2	1	0	0	
1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	58	1	0	0	0	
2	2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	45	1	1	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	21	2	0	0	0	
3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	33	3	0	0	0	
3	0	1	5	0	3	0	2	0	0	0	0	0	0	0	0	0	1	213	2	0	0	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	2	1	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	30	3	0	0	0	
4	0	0	2	0	1	1	0	0	0	0	0	0	0	0	0	0	2	57	1	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	17	3	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	21	4	0	0	0	
2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	20	2	1	0	0	
4	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	81	2	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	2	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	18	2	0	0	0	

- Small dataset table



qty_dot_u	qty_hyphe	qty_under	qty_slash_	qty_questi	qty_equal	qty_at_url	qty_and_u	qty_exclar	qty_space	qty_tilde_	qty_comm	qty_plus_	qty_asteri	qty_hashti	qty_dollar	qty_percei	qty_tld_ur	length_url	qty_dot_d	qty_hyphe	qty_under	qty_slash_	qt
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	2	0	0	0	
4	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	38	4	0	0	0	
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	24	1	0	0	0	
2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	38	2	0	0	0	
1	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	46	1	1	0	0	
1	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	45	1	0	0	0	
1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	32	1	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	29	2	0	0	0	
2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	18	2	0	0	0	
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	1	0	0	0	
2	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	34	1	0	0	0	
1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	21	1	0	0	0	
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	32	2	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	1	0	0	0	
3	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	31	3	1	0	0	
2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	2	29	1	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	2	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	2	0	0	0	
1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	12	1	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	2	0	0	0	
2	7	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	107	2	0	0	0	
1	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	22	1	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	17	2	0	0	0	
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	1	1	0	0	
2	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	1	46	1	0	0	0	

2.2 Phishing domain detection

- The system displays the message legitimate site vs phishing.
- The system presents the set of inputs required from the user.
- The user gives required information.
- The system should be able to predict the detection between malicious vs real.

2.3 Logging

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

- 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.

2.5 Deployment

1. AWS
2. Heroku

Technology stack

Front End	HTML/CSS
Backend	Python Flask
Deployment	AWS

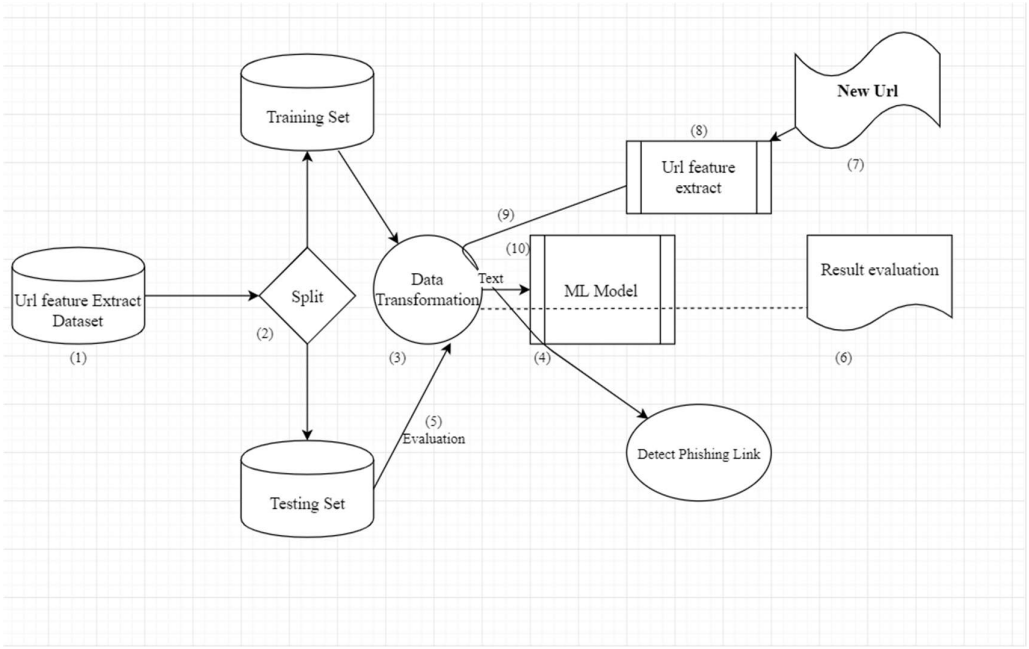
Proposed Solution

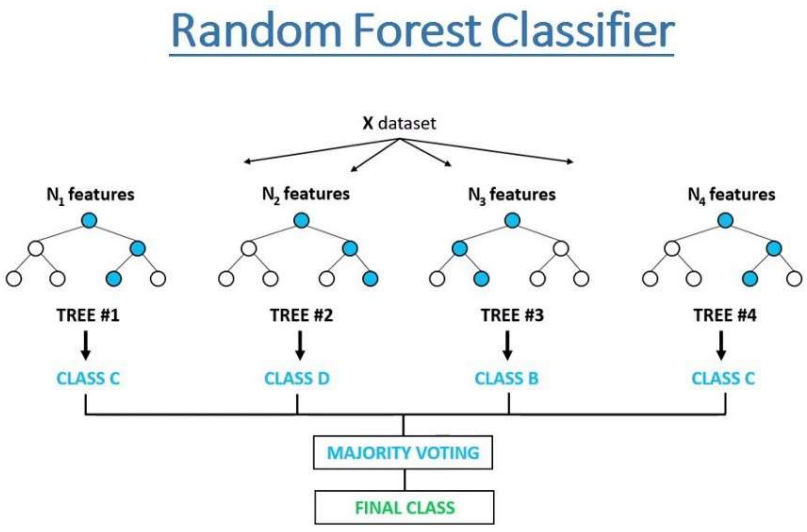
refer: https://www.sciencedirect.com/science/article/pii/S2352340920313202

Based on the actual research paper, 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.

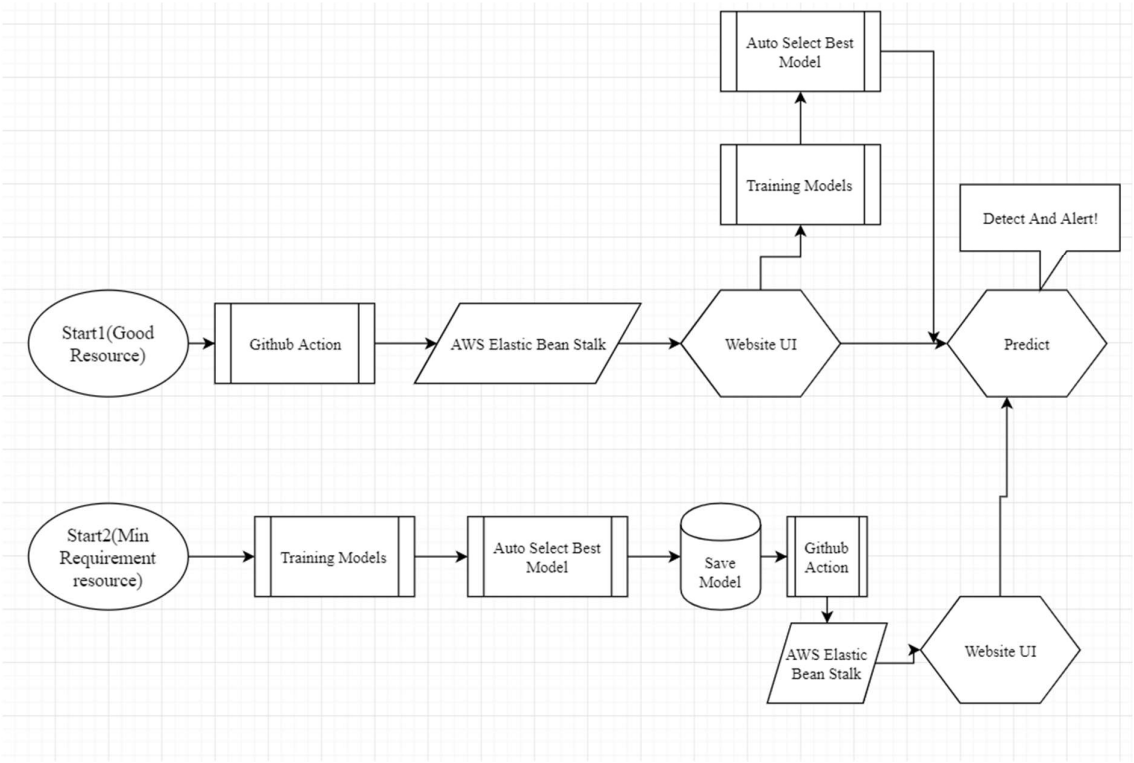
2. Baseline Models: Decision tree classifier and XG Boost classifier, since this is a classification problem.
3. Actual model: Random Forest Classifier.

Model training/validation workflow





User I/O workflow



Test cases

Test case	Steps to perform test case	Module	Pass/Fail
1.	a. Kaggle dataset of URLs with mixup of phishing and real site b. Feature extraction of URL c. Extracted feature into csv d. Predict	Predict app	Pass

Key performance indicators (KPI)

- Already save model also present for minimum requirement server.
- View logs
- View Saved Model and artifacts