Low Level Document

Phishing Domain Detection

Project by – Tejas Muley

Content

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. This paper presents dataset variations that consist of 58,645 websites labelled as legitimate or phishing and allows the others to train their classification models, build phishing detection systems.

1. Introduction

1.1 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.

The main objective of the project is to predict weather the website is real or malicious. So that we can take further actions on malicious/fake websites and avoid phishing.

This project can be delivered in three phases:

1. Building Machine Learning Model depending on the requirements.
2. Integration of UI and database to all the functionalities.
3. Deployment of project on cloud.

1.2 Scope

This software system will be a web application, this system will be designed to predicts weather the website is real or malicious/fake based on the user's input in which there are several categories to fill in like the Length of the URL, Length of the Directory, Length of the file. Activation Domain Time and Expiration Domain Time.

Based on these features model will predict weather the website is real or fake. We make sure that all the given features should be available at that time in order to get the optimum utilization and earn maximum profits by the company.

1.3 Introduction

Phishing is a form of fraud in which the attacker tries to learn sensitive information such as login credentials or account information by sending as a reputable entity or person in email or other communication channels.

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.

Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defence systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents.

1.4 Constraints

We will be using only few features (Selection of features using Feature Selection method).This system only predict weather the website is real or fake.

1.5 Out of Scope

System will not perform well, if user input is not accurate.

2. Problem Statement

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.

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

3. Dataset Information

3.1Dataset overview

Dataset consist of 58654 records and 112 features. This dataset mainly divided into four parts:

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

The presented dataset was collected and prepared for the purpose of building and evaluating various classificatio n methods for the task of detecting phishing websites based on the uniform resource locator (URL) properties, U RL resolving metrics, and external services. The attributes of the prepared dataset can be divided into six groups :

Table 1. Dataset attributes based on URL.

Nr.	Attribute	Format	Description	Values
1	qty_dot_url	Number of "." signs	Numeric	
2	qty_hyphen_url	Number of "-" signs	Numeric	
3	qty_underline_url	Number of "_" signs	Numeric	
4	qty_slash_url	Number of "/" signs	Numeric	
5	qty_questionmark_url	Number of "?" signs	Numeric	
6	qty_equal_url	Number of "=" sings	Numeric	
7	qty_at_url	Number of "@" signs	Numeric	
8	qty_and_url	Number of "&" signs	Numeric	
9	qty_exclamation_url	Number of "!" signs	Numeric	
10	qty_space_url	Number of " " signs	Numeric	
11	qty_tilde_url	Number of "~" signs	Numeric	
12	qty_comma_url	Number of "," signs	Numeric	
13	qty_plus_url	Number of "+" signs	Numeric	
14	qty_asterisk_url	Number of "*" signs	Numeric	
15	qty_hashtag_url	Number of "#" signs	Numeric	
16	qty_dollar_url	Number of "$" signs	Numeric	
17	qty_percent_url	Number of "%" signs	Numeric	
18	qty_tld_url	Top level domain character length	Numeric	
19	length_url	Number of characters	Numeric	
20	email_in_url	Is email present	Boolean	[0, 1]

Table 2. Dataset attributes based on domain URL.

Nr.	Attribute	Format	Description	Values
1	qty_dot_domain	Number of "." signs	Numeric	
2	qty_hyphen_domain	Number of "-" signs	Numeric	
3	qty_underline_domain	Number of "_" signs	Numeric	
4	qty_slash_domain	Number of "/" signs	Numeric	
5	qty_questionmark_domain	Number of "?" signs	Numeric	
6	qty_equal_domain	Number of "=" signs	Numeric	
7	qty_at_domain	Number of "@" signs	Numeric	
8	qty_and_domain	Number of "&" signs	Numeric	
9	qty_exclamation_domain	Number of "!" signs	Numeric	
10	qty_space_domain	Number of " " signs	Numeric	
11	qty_tilde_domain	Number of "signs	Numeric	
12	qty_comma_domain	Number of "," signs	Numeric	
13	qty_plus_domain	Number of "+" signs	Numeric	
14	qty_asterisk_domain	Number of "*" signs	Numeric	
15	qty_hashtag_domain	Number of "#" signs	Numeric	
16	qty_dollar_domain	Number of "$" signs	Numeric	
17	qty_percent_domain	Number of "%" signs	Numeric	
18	qty_vowels_domain	Number of vowels	Numeric	
19	domain_length	Number of domain characters	Numeric	
20	domain_in_ip	URL domain in IP address format	Boolean	[0, 1]
21	server_client_domain	"server" or "client" in domain	Boolean	[0, 1]

Table 3. Dataset attributes based on URL directory.

Nr.	Attribute	Format	Description	Values
1	qty_dot_directory	Number of "." signs	Numeric	
2	qty_hyphen_directory	Number of "-" signs	Numeric	
3	qty_underline_directory	Number of "_" signs	Numeric	
4	qty_slash_directory	Number of "/" signs	Numeric	
5	qty_questionmark_directory	Number of "?" signs	Numeric	
6	qty_equal_directory	Number of "=" signs	Numeric	
7	qty_at_directory	Number of "@" signs	Numeric	
8	qty_and_directory	Number of "&" signs	Numeric	
9	qty_exclamation_directory	Number of "!" signs	Numeric	
10	qty_space_directory	Number of " " signs	Numeric	
11	qty_tilde_directory	Number of "signs	Numeric	
12	qty_comma_directory	Number of "," signs	Numeric	
13	qty_plus_directory	Number of "+" signs	Numeric	
14	qty_asterisk_directory	Number of "*" signs	Numeric	
15	qty_hashtag_directory	Number of "#" signs	Numeric	
16	qty_dollar_directory	Number of "$" signs	Numeric	
17	qty_percent_directory	Number of "%" signs	Numeric	
18	directory_length	Number of directory characters	Numeric	

Table 4. Dataset attributes based on URL file name.

Nr.	Attribute	Format	Description	Values
1	qty_dot_file	Number of "." signs	Numeric	
2	qty_hyphen_file	Number of "-" signs	Numeric	
3	qty_underline_file	Number of "_" signs	Numeric	
4	qty_slash_file	Number of "/" signs	Numeric	
5	qty_questionmark_file	Number of "?" signs	Numeric	
6	qty_equal_file	Number of "=" signs	Numeric	
7	qty_at_file	Number of "@" signs	Numeric	
8	qty_and_file	Number of "&" signs	Numeric	
9	qty_exclamation_file	Number of "!" signs	Numeric	
10	qty_space_file	Number of " " signs	Numeric	
11	qty_tilde_file	Number of "signs	Numeric	
12	qty_comma_file	Number of "," signs	Numeric	
13	qty_plus_file	Number of "+" signs	Numeric	
14	qty_asterisk_file	Number of "*" signs	Numeric	
15	qty_hashtag_file	Number of "#" signs	Numeric	
16	qty_dollar_file	Number of "$" signs	Numeric	
17	qty_percent_file	Number of "%" signs	Numeric	
18	file_length	Number of file name characters	Numeric	

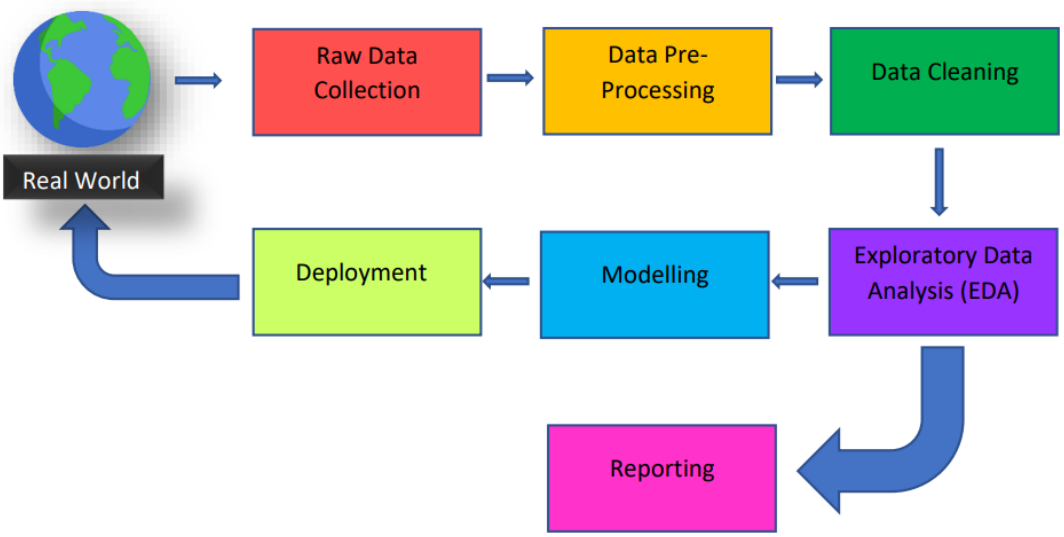
Table 5. Dataset attributes based on URL parameters.

Nr.	Attribute	Format	Description	Values
1	qty_dot_params	Number of "." signs	Numeric	
2	qty_hyphen_params	Number of "-" signs	Numeric	
3	qty_underline_params	Number of "_" signs	Numeric	
4	qty_slash_params	Number of "/" signs	Numeric	
5	qty_questionmark_params	Number of "?" signs	Numeric	
6	qty_equal_params	Number of "=" signs	Numeric	
7	qty_at_params	Number of "@" signs	Numeric	
8	qty_and_params	Number of "&" signs	Numeric	
9	qty_exclamation_params	Number of "!" signs	Numeric	
10	qty_space_params	Number of " " signs	Numeric	
11	qty_tilde_params	Number of "signs	Numeric	
12	qty_comma_params	Number of "," signs	Numeric	
13	qty_plus_params	Number of "+" signs	Numeric	
14	qty_asterisk_params	Number of "*" signs	Numeric	
15	qty_hashtag_params	Number of "#" signs	Numeric	
16	qty_dollar_params	Number of "$" signs	Numeric	
17	qty_percent_params	Number of "%" signs	Numeric	
18	params_length	Number of parameters characters	Numeric	
19	tld_present_params	TLD[1] present in parameters	Boolean	[0, 1]
20	qty_params	Number of parameters	Numeric	

Table 6. Dataset attributes based on resolving URL and external services.

Nr.	Attribute	Format	Description	Values
1	time_response	Domain lookup time response	Numeric	
2	domain_spf	Domain has SPF [2]	Boolean	[0, 1]
3	asn_ip	ASN [3]	Numeric	
4	time_domain_activation	Domain activation time (in days)	Numeric	
5	time_domain_expiration	Domain expiration time (in days)	Numeric	
6	qty_ip_resolved	Number of resolved IPs	Numeric	
8	qty_nameservers	Number of resolved NS[4]	Numeric	
9	qty_mx_servers	Number of MX [5] servers	Numeric	
10	ttl_hostname	Time-To-Live (TTL)	Numeric	
11	tls_ssl_certificate	Has valid TLS [6]/SSL [7] certificate	Boolean	[0, 1]
12	qty_redirects	Number of redirects	Numeric	
13	url_google_index	Is URL indexed on Google	Boolean	[0, 1]
14	domain_google_index	Is domain indexed on Google	Boolean	[0, 1]
15	url_shortened	Is URL shortened	Boolean	
16	**phishing**	**Is phishing website**	**Boolean**	**[0, 1]**

4. Architecture



4.1 Architecture Description

1. Raw Data Collection

The Dataset was taken from iNeuron's Provided Project Description Document.

Dataset Link: - https://data.mendeley.com/datasets/72ptz43s9v/1

2. Data Pre-Processing

Before building any model, it is crucial to perform data pre-processing to feed the correct data to the model to le arn and predict. Model performance depends on the quality of data needed to the model to train. This Process in cludesa) Handling Null/Missing Values b) Handling Skewed Data c) Outliers Detection and Removal.

3. Data Cleaning and Exploratory Data Analysis

Data Cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or inco mplete data within a dataset. a) Remove duplicate or irrelevant observations b) Filter unwanted Outliers c) Rena ming required attributes.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover p atterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphi cal representations.

4. Model Building

Model Building is the process of developing a probabilistic model that best describes the relationship between th e dependent and independent variables.

5. Reporting

Reporting is a most important and underrated skill of a data analytics field. Because being a Data Analyst you sh ould be good in easy and self-explanatory report because your model will be used by many stakeholders who are not from technical background. a) High Level Design Document b) Low Level Design Document (LLD) c) Arc hitecture d) Wireframe e) Detailed Project Report f) Power Point Presentation

6. Modelling

Data Modelling is the process of analysing the data objects and their relationship to the other objects It is used to analyse the data requirements that are required for the business processes. The data models are created for the da ta to be stored in a database. The Data Model's main focus is on what data is needed and how we have to organiz e data rather than what operations we have to perform.

7. Deployment using Flask on Azure Cloud.

5. 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.