

# Low Level Design (LLD)

## Phishing Domain Detection

Revision Number: 1

Last date of revision: 21/08/2022

Saurabh Naik

## Document Version Control

---

Date Issued	Version	Description	Author
21th August 2022	1.1	First Draft	Saurabh Naik

## Table of Contents

Document Version Control .....	2
Abstract.....	4
1.Introduction .....	5
1.1 Why this Low-Level Design Document? .....	5
1.2 Scope.....	6
1.3 Constraints .....	6
1.4 Risks .....	6
1.5 Out of Scope.....	6
2.Technical specifications .....	7
2.1 Dataset Name and Source.....	7
2.2 Dataset overview .....	7
2.3 Data Description .....	11
2.4 Logging .....	14
2.5 Database .....	15
3. Deployment.....	16
4.Technology stack .....	17
5. Proposed Solution.....	18
6. Proposed Architecture.....	19
7. Model training/validation workflow.....	20
8.User I/O workflow.....	21
9. Error Handling.....	22
10.Model performances .....	23
11.Output.....	24
12.Key performance indicators (KPI).....	26
13. Conclusion .....	27
14. References.....	28

## Abstract

---

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 of this end-to-end application is to predict whether the domains are real or malicious using various combinations of machine learning algorithms by dividing the dataset into various clusters and applying a suitable algorithm for that particular cluster.

# 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 LLD 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

## 1.2 Scope

The LLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The LLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system. This software system will be a Web application this system will be designed to detect malicious websites.

## 1.3 Constraints

We will only be detecting malicious websites.

## 1.4 Risks

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

## 1.5 Out of Scope

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

## 2. Technical specifications

---

### 2.1 Dataset Name and Source

Data Set Name	Finalized	Source
Dataset_full.csv	Yes	<a href="#">Phishing Websites Dataset - Mendeley Data</a>

### 2.2 Dataset overview

- These data consist of a collection of legitimate as well as phishing website instances. Each website is represented by the 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 Short description of the full variant dataset: Total number of instances: 88,647 Number of legitimate website instances (labelled as 0): 58,000 Number of phishing website instances (labelled as 1): 30,647 Total number of features: 111

dataset\_full - Excel (Product Activation Failed)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	qty_dot_url	qty_hyph	qty_unde	qty_slash	qty_quest	qty_equal	qty_at	qty_and	qty_exclai	qty_space	qty_tilde	qty_comn	qty_plus	qty_asteri	qty_hasht	qty_dollar	qty_perce	qty_tld	length_ur	qty_dot_c	qty_h
2	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	25	2
3	5	0	1	3	0	3	0	2	0	0	0	0	0	0	0	0	0	0	3	223	2
4	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	15	2
5	4	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	81	2
6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	2
7	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	22	1
8	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	27	2
9	2	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	46	2
10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	2
11	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	24	1
12	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	19	2
13	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	58	1
14	2	2	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	45	1
15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	21	2
16	3	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	33	3
17	3	0	1	5	0	3	0	2	0	0	0	0	0	0	0	0	0	0	1	213	2
18	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	2
19	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	30	3
20	4	0	0	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2	57	1
21	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	17	3
22	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	21	4
23	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	20	2

dataset\_full - Excel (Product Activation Failed)

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ
1	qty_dollar	qty_perce	qty_tld	length_ur	qty_dot_c	qty_hyph	qty_unde	qty_slash	qty_quest	qty_equal	qty_at	qty_and	qty_exclai	qty_space	qty_tilde	qty_comn	qty_plus	qty_asteri	qty_hasht	qty_dollar	qty_p
2	0	0	1	25	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	3	223	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	15	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	81	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	1	19	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	1	27	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	1	46	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	1	16	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	1	24	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	1	19	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	1	58	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	1	45	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	1	21	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	1	33	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	1	213	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	1	13	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	1	30	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	2	57	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	1	17	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	1	21	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	1	20	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0



dataset\_full - Excel (Product Activation Failed)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number Styles Cells Editing

Calibri 11 A A

B I U Merge & Center

General Conditional Formatting Cell Styles Insert Delete Format AutoSum Sort & Find & Filter Select

P1 X ✓ fx qty\_dollar\_url

	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD
1	qty_perce	qty_vowe	domain_l	domain_i	server_cli	qty_dot_c	qty_hyph	qty_under	qty_slash	qty_quest	qty_equal	qty_at_dir	qty_and	c_qty_excl	qty_space	qty_tilde	qty_comn	qty_plus	qty_asteri	qty_hasht	qty_d
2	0	4	17	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	5	16	0	0	3	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	3	14	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
5	0	7	19	0	0	2	0	2	5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	5	19	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	0	4	17	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
8	0	9	27	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	0	9	28	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
10	0	3	16	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	0	4	14	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
12	0	5	19	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	0	6	16	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0
14	0	5	13	0	0	1	1	0	4	0	0	0	0	0	0	0	0	0	0	0	0
15	0	5	21	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
16	0	7	27	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
17	0	6	20	0	0	1	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
18	0	2	13	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
19	0	7	30	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
20	0	3	11	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
21	0	5	17	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
22	0	5	21	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
23	0	5	19	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

dataset\_full

dataset\_full - Excel (Product Activation Failed)

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number Styles Cells Editing

Calibri 11 A A

B I U Merge & Center

General Conditional Formatting Cell Styles Insert Delete Format AutoSum Sort & Find & Filter Select

P1 X ✓ fx qty\_dollar\_url

	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX
1	qty_dollar	qty_perce	directory	qty_dot_f	qty_hyph	qty_under	qty_slash	qty_quest	qty_equal	qty_at_f	qty_and	f_qty_excl	qty_space	qty_tilde	qty_comn	qty_plus	qty_asteri	qty_hasht	qty_dollar	qty_perce	file
2	0	0	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	42	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	62	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
7	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
9	0	0	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
11	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
13	0	0	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
16	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
19	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
20	0	0	20	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
22	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
23	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

dataset\_full

dataset\_full - Excel (Product Activation Failed)

FILEHOMEINSERTPAGE LAYOUTFORMULASDATAREVIEWVIEW



ClipboardFontAlignmentNumberStylesCellsEditing

Calibri11A<sup>+</sup>




dataset\_full - Excel (Product Activation Failed)

FILEHOMEINSERTPAGE LAYOUTFORMULASDATAREVIEWVIEW

Calibri11A<sup>+</sup>

B I U  




Font




Alignment

Wrap Text




General

 %  




Number



Conditional FormattingTableStyles

InsertDeleteFormat

AutoSumFillClear

Sort & Find & Filter & Select

P1: X ✓ fx qty\_dollar\_url

	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI				
1	qty_perce	params	letld_prese	qty_param	email_in	time_res	domain_s	asn_ip	time_dom	time_dom	qty_ip	re_qty	name	qty_mx	s_ttl	hostn	tls_ssl	cei	qty_redir	url	googl	domain	g_url	shorte	phishing
2	-1	-1	-1	-1	-1	0	0.207316	0	60781	-1	-1	1	2	0	892	0	0	0	0	0	0	0	0	1	
3	0	165	0	3	0	0.499566	-1	36024	579	150	1	2	1	9540	1	0	0	0	0	0	0	0	0	1	
4	-1	-1	-1	-1	-1	0.935901	0	4766	-1	-1	1	2	3	589	1	0	0	0	0	0	0	0	0	0	
5	-1	-1	-1	-1	-1	0.410021	0	20454	-1	-1	1	2	0	292	1	0	0	0	0	0	0	0	0	1	
6	-1	-1	-1	-1	-1	0.410761	0	53831	6998	306	1	2	1	3597	0	1	0	0	0	0	0	0	0	0	
7	-1	-1	-1	-1	-1	0.458436	-1	25535	9	355	1	3	3	3591	1	0	0	0	0	0	0	0	0	1	
8	-1	-1	-1	-1	-1	0.710173	0	13446	-1	-1	1	2	2	291	0	0	0	0	0	0	0	0	0	0	
9	-1	-1	-1	-1	-1	0.244512	0	55053	-1	-1	1	2	1	3134	1	0	0	0	0	0	0	0	0	0	
10	-1	-1	-1	-1	-1	0.1712161	0	13335	778	316	1	4	2	3596	1	1	0	0	0	0	0	0	0	0	
11	-1	-1	-1	-1	-1	0.78757	1	20013	4805	307	1	2	1	14397	1	0	0	0	0	0	0	0	0	1	
12	-1	-1	-1	-1	-1	0.799702	0	31815	6241	1063	1	2	1	43197	0	1	0	0	0	0	0	0	0	0	
13	-1	-1	-1	-1	-1	0.381044	1	46606	935	160	1	2	1	13174	1	0	0	0	0	0	0	0	0	1	
14	-1	-1	-1	-1	-1	0.1024603	-1	198610	1065	30	1	6	2	583	0	0	0	0	0	0	0	0	0	1	
15	-1	-1	-1	-1	-1	0.127553	-1	12488	7348	320	1	2	1	14396	1	0	0	0	0	0	0	0	0	0	
16	-1	-1	-1	-1	-1	0.310562	0	26496	-1	-1	-1	2	2	-1	0	0	0	0	0	0	0	0	0	1	
17	0	165	0	3	0	0.372825	1	20013	3144	142	1	2	5	9632	1	1	0	0	0	0	0	0	0	1	
18	-1	-1	-1	-1	-1	0.1188645	0	4765	6889	781	1	2	5	210	1	1	0	0	0	0	0	0	0	0	
19	-1	-1	-1	-1	-1	0.535979	0	13446	-1	-1	1	2	1	299	0	0	0	0	0	0	0	0	0	0	
20	0	26	1	1	1	0.855594	-1	20454	270	94	1	2	1	14390	0	0	0	0	0	0	0	0	0	1	
21	-1	-1	-1	-1	-1	0.327323	0	1680	-1	-1	1	3	1	3440	0	0	0	0	0	0	0	0	0	0	
22	-1	-1	-1	-1	-1	0.804266	0	-1	8120	-1	1	2	1	298	1	0	0	0	0	0	0	0	0	0	
23	-1	-1	-1	-1	-1	0.400331	0	-1	-1	-1	1	2	0	1792	0	0	0	0	0	0	0	0	0	1	

dataset\_full

## 2.3 Data Description

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

- 1) attributes based on the whole URL properties presented in table 1.

Nr.	Attribute	Format	Description
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 "=" signs	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

- 2) attributes based on the domain properties presented in Table 2

Nr	Attribute	Format	Description
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
21	server_client_domain	"server" or "client" in domain	Boolean

### 3) attributes based on the URL directory properties presented in Table 3

Nr	Attribute	Format	Description
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

### 4) attributes based on the URL file properties presented in Table 4

Nr	Attribute	Format	Description
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

5) attributes based on the URL parameter properties presented in Table 5

Nr	Attribute	Format	Description
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 present in parameters	Boolean
20	qty_params	Number of parameters	Numeric

6) Attributes based on the URL resolving data and external metrics presented in Table 6.

Nr	Attribute	Format	Description
1	time_response	Domain lookup time response	Numeric
2	domain_spf	Domain has SPF	Boolean
3	asn_ip	ASN	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
7	qty_nameservers	Number of resolved NS4	Numeric
8	qty_mx_servers	Number of MX 5servers	Numeric
9	ttl_hostname	Time-To-Live (TTL)	Numeric
10	tls_ssl_certificate	Has valid TLS 6/SSL 7certificate	Numeric
11	qty_redirects	Number of redirects	Boolean
12	url_google_index	Is URL indexed on Google	Numeric
13	domain_google_index	Is domain indexed on Google	Boolean
14	url_shortened	Is URL shortened	Boolean
15	phishing	Is phishing website	Boolean

The first group is based on the values of the attributes on the whole URL string, while the values of the following four groups are based on the particular sub-strings, as presented in [Figure 1](#). The last group attributes are based on the URL resolve metrics as well as on the external services such as Google search index.

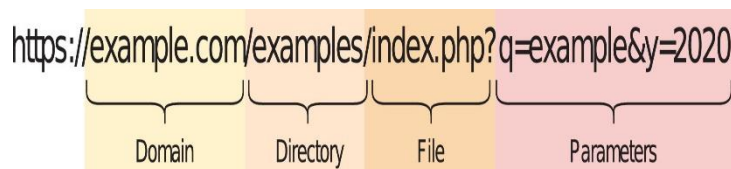


Fig. 1. Separation of the whole URL string into sub-strings.

## 2.4 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.

## 2.5 Database

System needs to store every request into the database and we need to store it in such a way that it is easy to retrain the model as well.

1. The User chooses the activity dataset.
2. The User gives required information.
3. The system stores each and every data given by the user or received on request to the database. Database chosen in this case is Cassandra.

### 3. Deployment

---

Heroku





## 4. Technology stack

---

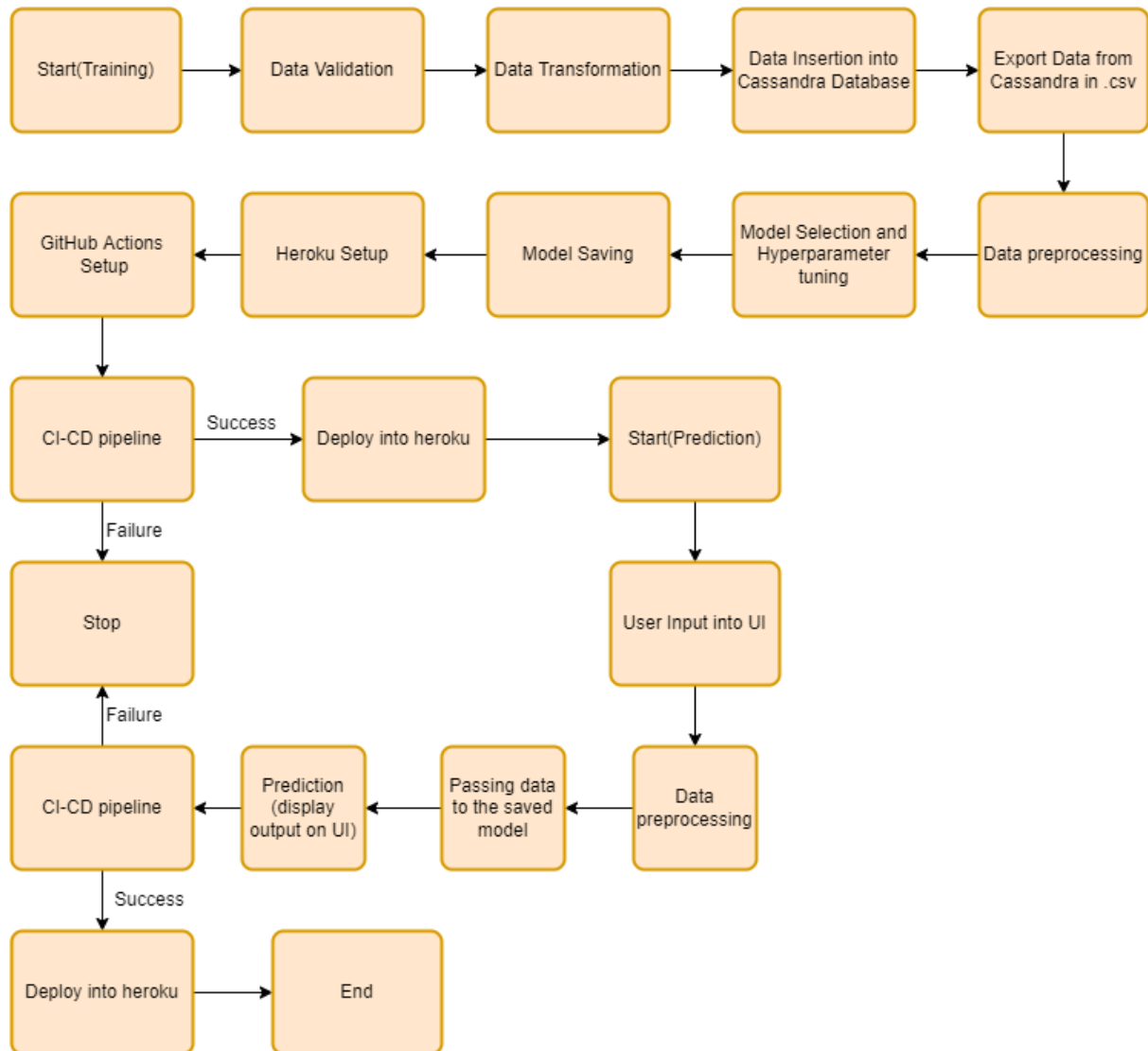
<b>Front End</b>	HTML/CSS/JS
<b>Backend</b>	Python Flask
<b>Database</b>	Cassandra
<b>Deployment</b>	heroku
<b>Visualization</b>	Sea born
<b>Data version control</b>	DVC
<b>Source version control</b>	GitHub
<b>CI-CD pipeline</b>	GitHub actions
<b>Model Selection and Hyper parameter tuner</b>	Optuna

## 5. Proposed Solution

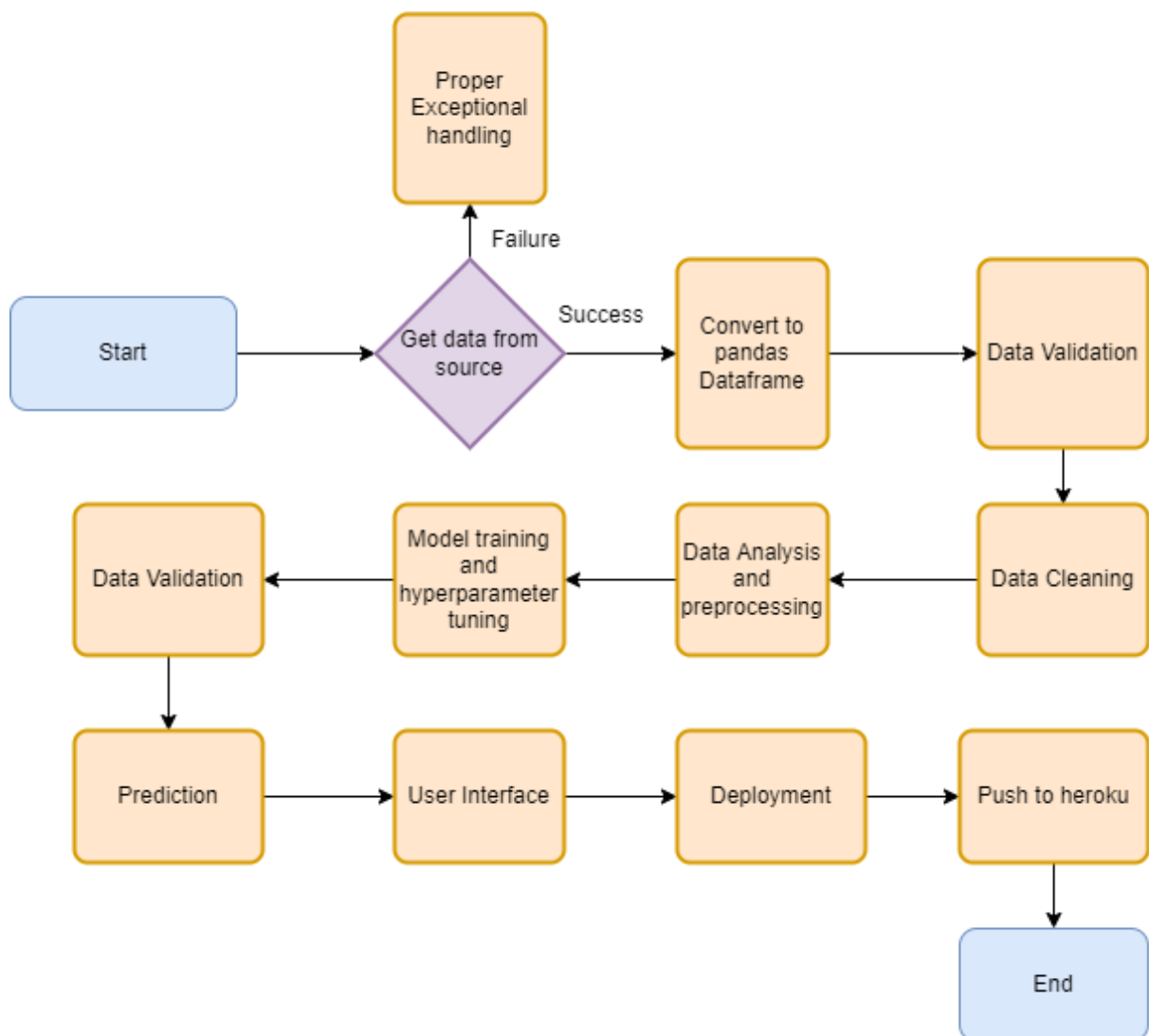
---

The solution proposed here is a Phishing domain detection based machine learning system to tackle phishing. In these approach we perform some classical machine learning task like Data Exploration, Data Cleaning, Feature Exploration and Feature Selection. Then building multiple models using different parameters and testing the same and selecting the best model giving good performance metrics. Then creating 2 flask APIs for training and prediction respectively and binding them with a frontend created using HTML, CSS, and Bootstrap. Finally hosting this solution on heroku by dockerizing complete module.

## 6. Proposed Architecture

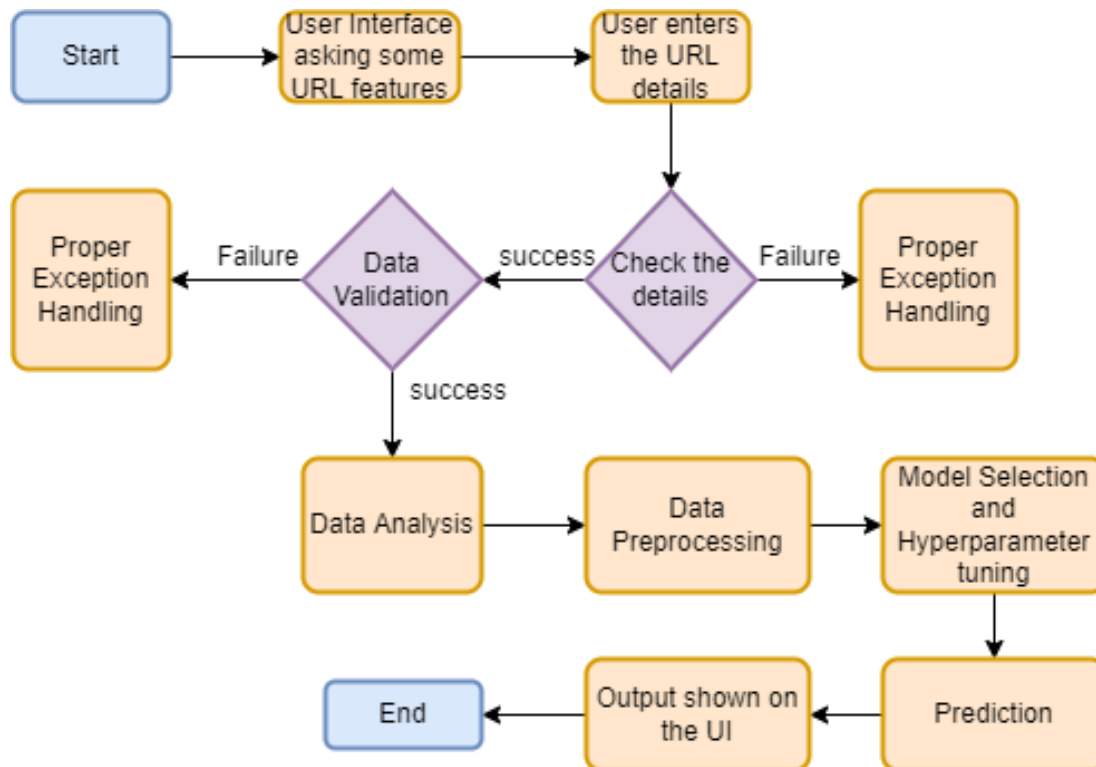


## 7. Model training/validation workflow



## 8. User I/O workflow

---



## 9. 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.

A python function is created to display the error message in detail.

The format of message will be current date -- current time -- error message.

## 10. Model performances

---

Use case (Binary Classification)	ML Model	Accuracy
1.	Logistic Regression	92.5%
2.	Random Forest	95%
3.	Support Vector Classifier	93.8%
4.	Naïve Bayes Classifier	92.1%
5.	Decision Tree Classifier	94.4%
6.	XGBoost Classifier	95.1%

## 11. Output

The image displays two screenshots of a web application titled "Phishing domain detection" at the URL `phishing-domain-detector123.herokuapp.com`.

**Top Screenshot (Input Form):**

The form contains the following input fields:

- Enter no of slash in entire URL: 5
- Enter no characters in entire URL: 6
- Enter no dots in URL domain section: 5
- Enter no dots in URL directory section: 6
- Enter no hyphen in URL directory section: 5
- Enter no of filename characters: 5
- Enter no of underlines in directory section of URLs: 4
- Enter ASN ip no of the URL: 5
- Enter time domain activation (in days): 4
- Enter time domain expiration (in days): 5
- Enter time to live: 4

A yellow "Predict" button is located at the bottom left of the form.

**Bottom Screenshot (Output):**

After clicking the "Predict" button, the application displays a message: "This is malicious website". Below this message, the input fields are updated with ranges:

- Enter no of slash in entire URL: Enter between 0 to 8
- Enter no characters in entire URL: Enter between 5 to 100
- Enter no dots in URL domain section: Enter between 0 to 19
- Enter no dots in URL directory section: Enter between 0 to 19
- Enter no hyphen in URL directory section: Enter between 1 to 23
- Enter no of filename characters: Enter between 0 to 18
- Enter no of underlines in directory section of URLs: Enter between 0 to 18
- Enter ASN ip no of the URL: Enter between 0 to 7000
- Enter time domain activation (in days): Enter between 0 to 14000
- Enter time domain expiration (in days): Enter between 0 to 850
- Enter time to live: Enter between 0 to 28000

A yellow "Predict" button is also present at the bottom left of the output screen.



The image displays two screenshots of a web application titled "Phishing domain detection" hosted on `phishing-domain-detector123.herokuapp.com`.

**First Screenshot (Input Form):**

The form contains the following input fields:

- Enter no of slash in entire URL: 5
- Enter no characters in entire URL: 6
- Enter no dots in URL domain section: 5
- Enter no dots in URL directory section: 6
- Enter no hyphen in URL directory section: 5
- Enter no of filename characters: 5
- Enter no of underlines in directory section of URLs: 4
- Enter ASN ip no of the URL: 5
- Enter time domain activation (in days): 4
- Enter time domain expiration (in days): 5
- Enter time to live: 4

A yellow "Predict" button is located at the bottom left of the form.

**Second Screenshot (Output):**

After clicking the "Predict" button, the application displays the message: "This is legitimate website". Below this message, the form fields are updated with ranges:

- Enter no of slash in entire URL: Enter between 0 to 8
- Enter no characters in entire URL: Enter between 5 to 100
- Enter no dots in URL domain section: Enter between 0 to 19
- Enter no dots in URL directory section: Enter between 0 to 19
- Enter no hyphen in URL directory section: Enter between 1 to 23
- Enter no of filename characters: Enter between 0 to 18
- Enter no of underlines in directory section of URLs: Enter between 0 to 18
- Enter ASN ip no of the URL: Enter between 0 to 7000
- Enter time domain activation (in days): Enter between 0 to 14000
- Enter time domain expiration (in days): Enter between 0 to 850
- Enter time to live: Enter between 0 to 28000

A yellow "Predict" button is also present at the bottom left of the second form.

## 12.Key performance indicators (KPI)

---

- Key indicators displaying a summary of the anomaly detection in the URLs.
- Total number of slash in URL
- Total number of characters in URL
- Total number of dots in URL domain section
- Total number of dots in URL directory section
- Total number of hyphen in URL directory section
- Total number of filename characters
- Total number of underlines in directory section of URL
- ASN IP number of URL
- Enter no of days of time domain activation
- Enter no of days of time domain expiration
- Enter time to live

## 13. Conclusion

---

The Phishing Domain Detection System will detect whether the domains is real for fake and avoid user to become a victim of phishing based on various data used to train our algorithm, so we can identify the fake domain(URL) so prevent the loss of crucial data by avoiding phishing.

## 14. References

---

- 1) [Datasets for phishing websites detection - ScienceDirect](#)
- 2) [Phishing Websites Dataset - Mendeley Data](#)
- 3) <https://getbootstrap.com/docs/4.3/getting-started/introduction/>
- 4) <https://www.youtube.com/watch?v=1BSwYlJUXK0&list=PLZoTAE LR MXVOK1pRcOCaG5xtXxgMalpIe>
- 5) [https://www.youtube.com/watch?v=ioN1jcWxbv8&list=PLZoTAE LR MXVPQyArDHyQVjQxjj\\_YmEuO9](https://www.youtube.com/watch?v=ioN1jcWxbv8&list=PLZoTAE LR MXVPQyArDHyQVjQxjj_YmEuO9)
- 6) <https://www.youtube.com/watch?v=uMIU2JaiOd8&list=PLZoTAE LR MXVPgJwJ8VyRoqmfNs2CJwhVH>
- 7) <https://www.youtube.com/user/krishnaik06/playlists>