CS – GY 6923 Machine Learning

Professor: Dr. Raman Kannan

HW1: Exploratory Data Analysis

Avinash Authipudi

Section 1 - Overview of the Dataset

The dataset "Phishing_Attacks.csv" was taken from the website "https://data.mendeley.com /datasets/72ptz43s9v/1".

Phishing is a fraudulent activity where an attacker tries to obtain private and sensitive information from the victim. It is usually done via emails, messages or websites. Phishing websites are very similar to legitimate websites.

The dataset consists of a collection of legitimate and phishing website instances. Each instance is represented by a set of features that represent whether the website is legitimate or not. The features for each instance is based on the Uniform Resource Locator (URL) properties.

There are a total of 58,645 rows or instances out of which 27,998 instances are legitimate websites and 30,647 instances are phishing websites. There are 112 features for each instance out of 1 variable is the dependent variable which indicates whether the website is legitimate (0) or a phishing website (1) and the rest are independent variables. The dependent variable is either a '0' or a '1' which means that it is a binary classification problem.

The presented dataset was collected 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:

- attributes based on the whole URL properties presented in Table 1,
- attributes based on the domain properties presented in Table 2,
- attributes based on the URL directory properties presented in Table 3,
- attributes based on the URL file properties presented in Table 4,
- attributes based on the URL parameter properties presented in Table 5 and
- attributes based on the URL resolving data and external metrics presented in Table 6.

Table 1: Dataset attributes based on URL

No	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

Table 2: Dataset attributes based on domain URL

No	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

Table 3: Dataset attributes based on URL directory

No	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

Table 4: Dataset attributes based on URL file name

No	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

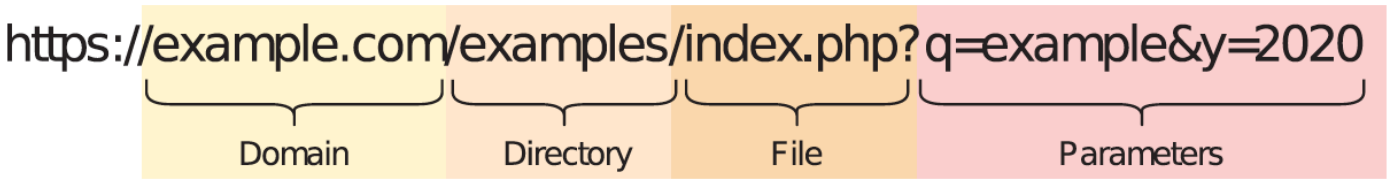


Table 5: Dataset attributes based on URL parameters

No	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 parameter characters	Numeric
19	tld_present_params	TLD present in parameters	Boolean
20	qty_params	Number of parameters	Numeric

Table 6: Dataset attributes based on resolving URLs and external services

No	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 NS	Numeric
8	qty_mx_servers	Number of MX 5 servers	Numeric
9	ttl_hostname	Time-To-Live (TTL)	Numeric
10	tls_ssl_certificate	Has valid TLS 6 /SSL 7 certificate	Boolean
11	qty_redirects	Number of redirects	Numeric
12	url_google_index	Is URL indexed on Google	Boolean
13	domain_google_index	Is domain indexed on Google	Boolean
14	url_shortened	Is URL shortened	Boolean
15	phishing	Is phishing website	Boolean

Section 2 - Loading the Data

The dataset being used is “phishing_data.csv”. It is loaded using the “read.csv” command into the variable data. We can see that the dataset has 58,645 observations and 112 features. The dependant variable is “phishing” which indicates whether the particular observation is a phishing instance or a legitimate instance. The description of each feature is provided in the above tables. Most of the independent features are numerical, hence they are non-categorical.

loadData <- ‘/Users/avinashauthipudi/Downloads/Machine Learning/phishing_data.csv’
data <- read.csv(loadData, header = T, stringsAsFactors = F)

```
> dim(data)
[1] 58645   112
> str(data)
'data.frame':   58645 obs. of  112 variables:
 $ qty_dot_url               : int  2 4 1 2 1 1 1 2 2 1 ...
 $ qty_hyphen_url            : int  0 0 0 0 1 1 0 0 0 0 ...
 $ qty_underline_url         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_slash_url             : int  0 2 1 3 4 4 3 0 2 1 ...
 $ qty_questionmark_url      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_equal_url             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_at_url                : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_and_url               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_exclamation_url       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_space_url             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_tilde_url             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_comma_url             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_plus_url              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_asterisk_url          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_hashtag_url           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_dollar_url            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_percent_url           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_tld_url               : int  1 1 1 1 1 1 1 1 1 1 ...
 $ length_url                : int  14 38 24 38 46 45 32 29 18 11 ...
 $ qty_dot_domain            : int  2 4 1 2 1 1 1 2 2 1 ...
 $ qty_hyphen_domain         : int  0 0 0 0 1 0 0 0 0 0 ...
 $ qty_underline_domain      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_slash_domain          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_questionmark_domain   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_equal_domain          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_at_domain             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ qty_and_domain            : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
> names(data)
  [1] "qty_dot_url"               "qty_hyphen_url"            "qty_underline_url"
  [4] "qty_slash_url"             "qty_questionmark_url"      "qty_equal_url"
  [7] "qty_at_url"                "qty_and_url"               "qty_exclamation_url"
 [10] "qty_space_url"             "qty_tilde_url"             "qty_comma_url"
 [13] "qty_plus_url"              "qty_asterisk_url"          "qty_hashtag_url"
 [16] "qty_dollar_url"            "qty_percent_url"           "qty_tld_url"
 [19] "length_url"                "qty_dot_domain"            "qty_hyphen_domain"
 [22] "qty_underline_domain"      "qty_slash_domain"          "qty_questionmark_domain"
 [25] "qty_equal_domain"          "qty_at_domain"             "qty_and_domain"
 [28] "qty_exclamation_domain"    "qty_space_domain"          "qty_tilde_domain"
 [31] "qty_comma_domain"          "qty_plus_domain"           "qty_asterisk_domain"
 [34] "qty_hashtag_domain"        "qty_dollar_domain"         "qty_percent_domain"
 [37] "qty_vowels_domain"         "domain_length"             "domain_in_ip"
 [40] "server_client_domain"      "qty_dot_directory"         "qty_hyphen_directory"
 [43] "qty_underline_directory"   "qty_slash_directory"       "qty_questionmark_directory"
 [46] "qty_equal_directory"       "qty_at_directory"          "qty_and_directory"
 [49] "qty_exclamation_directory" "qty_space_directory"       "qty_tilde_directory"
 [52] "qty_comma_directory"       "qty_plus_directory"        "qty_asterisk_directory"
 [55] "qty_hashtag_directory"     "qty_dollar_directory"      "qty_percent_directory"
 [58] "directory_length"          "qty_dot_file"              "qty_hyphen_file"
 [61] "qty_underline_file"        "qty_slash_file"            "qty_questionmark_file"
 [64] "qty_equal_file"            "qty_at_file"               "qty_and_file"
 [67] "qty_exclamation_file"      "qty_space_file"            "qty_tilde_file"
 [70] "qty_comma_file"            "qty_plus_file"             "qty_asterisk_file"
 [73] "qty_hashtag_file"          "qty_dollar_file"           "qty_percent_file"
 [76] "file_length"               "qty_dot_params"            "qty_hyphen_params"
 [79] "qty_underline_params"      "qty_slash_params"          "qty_questionmark_params"
 [82] "qty_equal_params"          "qty_at_params"             "qty_and_params"
 [85] "qty_exclamation_params"    "qty_space_params"          "qty_tilde_params"
 [88] "qty_comma_params"          "qty_plus_params"           "qty_asterisk_params"
 [91] "qty_hashtag_params"        "qty_dollar_params"         "qty_percent_params"
 [94] "params_length"             "tld_present_params"        "qty_params"
 [97] "email_in_url"              "time_response"             "domain_spf"
[100] "asn_ip"                    "time_domain_activation"    "time_domain_expiration"
[103] "qty_ip_resolved"           "qty_nameservers"           "qty_mx_servers"
[106] "ttl_hostname"              "tls_ssl_certificate"       "qty_redirects"
[109] "url_google_index"          "domain_google_index"       "url_shortened"
[112] "phishing"
```

Section 3 - Exploratory Data Analysis

Since this is a large dataset, we need to do some preliminary analysis on the dataset before any classification model can be used on the dataset.

3.1 Summary of the Dataset

Running the summary command on the dataset gives us the mean and median values of each feature in the dataset.

```
> summary(data)
  qty_dot_url     qty_hyphen_url    qty_underline_url qty_slash_url    qty_questionmark_url
 Min.   : 1.000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000   Min.   :0.0000
 1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.:0.0000
 Median : 2.000   Median : 0.0000   Median : 0.0000   Median : 1.000   Median :0.0000
 Mean   : 2.284   Mean   : 0.4571   Mean   : 0.1713   Mean   : 1.938   Mean   :0.0141
 3rd Qu.: 3.000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 3.000   3rd Qu.:0.0000
 Max.   :24.000   Max.   :35.0000   Max.   :21.0000   Max.   :44.000   Max.   :9.0000
 qty_equal_url       qty_at_url        qty_and_url     qty_exclamation_url qty_space_url
 Min.   : 0.0000   Min.   : 0.00000   Min.   : 0.000   Min.   : 0.000000   Min.   :0.000000
 1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.: 0.000   1st Qu.: 0.000000   1st Qu.:0.000000
 Median : 0.0000   Median : 0.00000   Median : 0.000   Median : 0.000000   Median :0.000000
 Mean   : 0.3112   Mean   : 0.03346   Mean   : 0.213   Mean   : 0.004451   Mean   :0.001535
 3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.: 0.000   3rd Qu.: 0.000000   3rd Qu.:0.000000
 Max.   :23.0000   Max.   :43.00000   Max.   :26.000   Max.   :10.000000   Max.   :9.000000
 qty_tilde_url      qty_comma_url        qty_plus_url       qty_asterisk_url   qty_hashtag_url
 Min.   :0.000000   Min.   : 0.000000   Min.   : 0.000000   Min.   : 0.00000   Min.   :0.00e+00
 1st Qu.:0.000000   1st Qu.: 0.000000   1st Qu.: 0.000000   1st Qu.: 0.00000   1st Qu.:0.00e+00
 Median :0.000000   Median : 0.000000   Median : 0.000000   Median : 0.00000   Median :0.00e+00
 Mean   :0.004877   Mean   : 0.003274   Mean   : 0.004212   Mean   : 0.00685   Mean   :7.67e-04
 3rd Qu.:0.000000   3rd Qu.: 0.000000   3rd Qu.: 0.000000   3rd Qu.: 0.00000   3rd Qu.:0.00e+00
 Max.   :5.000000   Max.   :11.000000   Max.   :19.000000   Max.   :60.00000   Max.   :1.30e+01
 qty_dollar_url      qty_percent_url     qty_tld_url       length_url      qty_dot_domain
 Min.   : 0.000000   Min.   :  0.0000   Min.   : 0.000   Min.   :   4.00   Min.   : 0.0
 1st Qu.: 0.000000   1st Qu.:  0.0000   1st Qu.: 1.000   1st Qu.:  18.00   1st Qu.: 1.0
 Median : 0.000000   Median :  0.0000   Median : 1.000   Median :  29.00   Median : 2.0
 Mean   : 0.002865   Mean   :  0.1625   Mean   : 1.068   Mean   :  44.96   Mean   : 1.8
 3rd Qu.: 0.000000   3rd Qu.:  0.0000   3rd Qu.: 1.000   3rd Qu.:  52.00   3rd Qu.: 2.0
 Max.   :10.000000   Max.   :174.0000   Max.   :12.000   Max.   :4165.00   Max.   :21.0
 qty_hyphen_domain qty_underline_domain qty_slash_domain qty_questionmark_domain qty_equal_domain
 Min.   : 0.0000   Min.   :0.0000000    Min.   :0        Min.   :0               Min.   :0
 1st Qu.: 0.0000   1st Qu.:0.0000000    1st Qu.:0        1st Qu.:0               1st Qu.:0
 Median : 0.0000   Median :0.0000000    Median :0        Median :0               Median :0
 Mean   : 0.1333   Mean   :0.0002899    Mean   :0        Mean   :0               Mean   :0
 3rd Qu.: 0.0000   3rd Qu.:0.0000000    3rd Qu.:0        3rd Qu.:0               3rd Qu.:0
 Max.   :11.0000   Max.   :2.0000000    Max.   :0        Max.   :0               Max.   :0
```

```
qty_at_domain      qty_and_domain qty_exclamation_domain qty_space_domain qty_tilde_domain
Min.   :0.00e+00   Min.   :0      Min.   :0              Min.   :0        Min.   :0
1st Qu.:0.00e+00   1st Qu.:0      1st Qu.:0              1st Qu.:0        1st Qu.:0
Median :0.00e+00   Median :0      Median :0              Median :0        Median :0
Mean   :1.71e-05   Mean   :0      Mean   :0              Mean   :0        Mean   :0
3rd Qu.:0.00e+00   3rd Qu.:0      3rd Qu.:0              3rd Qu.:0        3rd Qu.:0
Max.   :1.00e+00   Max.   :0      Max.   :0              Max.   :0        Max.   :0
qty_comma_domain qty_plus_domain qty_asterisk_domain qty_hashtag_domain qty_dollar_domain
Min.   :0        Min.   :0       Min.   :0           Min.   :0          Min.   :0
1st Qu.:0        1st Qu.:0       1st Qu.:0           1st Qu.:0          1st Qu.:0
Median :0        Median :0       Median :0           Median :0          Median :0
Mean   :0        Mean   :0       Mean   :0           Mean   :0          Mean   :0
3rd Qu.:0        3rd Qu.:0       3rd Qu.:0           3rd Qu.:0          3rd Qu.:0
Max.   :0        Max.   :0       Max.   :0           Max.   :0          Max.   :0
qty_percent_domain qty_vowels_domain domain_length     domain_in_ip      server_client_domain
Min.   :0          Min.   : 0.000    Min.   :  4.00   Min.   :0.000000   Min.   :0.00000
1st Qu.:0          1st Qu.: 4.000    1st Qu.: 14.00   1st Qu.:0.000000   1st Qu.:0.00000
Median :0          Median : 5.000    Median : 17.00   Median :0.000000   Median :0.00000
Mean   :0          Mean   : 5.441    Mean   : 18.07   Mean   :0.003427   Mean   :0.00353
3rd Qu.:0          3rd Qu.: 7.000    3rd Qu.: 21.00   3rd Qu.:0.000000   3rd Qu.:0.00000
Max.   :0          Max.   :61.000    Max.   :231.00   Max.   :1.000000   Max.   :1.00000
qty_dot_directory  qty_hyphen_directory qty_underline_directory qty_slash_directory
Min.   :-1.00000   Min.   :-1.00000     Min.   :-1.0000         Min.   :-1.00
1st Qu.:-1.00000   1st Qu.:-1.00000     1st Qu.:-1.0000         1st Qu.:-1.00
Median : 0.00000   Median : 0.00000     Median : 0.0000         Median : 1.00
Mean   : 0.02234   Mean   :-0.03381     Mean   :-0.2109         Mean   : 1.59
3rd Qu.: 1.00000   3rd Qu.: 0.00000     3rd Qu.: 0.0000         3rd Qu.: 3.00
Max.   :19.00000   Max.   :23.00000     Max.   :17.0000         Max.   :22.00
qty_questionmark_directory qty_equal_directory qty_at_directory  qty_and_directory
Min.   :-1.0000            Min.   :-1.000      Min.   :-1.0000   Min.   :-1.0000
1st Qu.:-1.0000            1st Qu.:-1.000      1st Qu.:-1.0000   1st Qu.:-1.0000
Median : 0.0000            Median : 0.000      Median : 0.0000   Median : 0.0000
Mean   :-0.2985            Mean   :-0.287      Mean   :-0.2934   Mean   :-0.2917
3rd Qu.: 0.0000            3rd Qu.: 0.000      3rd Qu.: 0.0000   3rd Qu.: 0.0000
Max.   : 0.0000            Max.   : 5.000      Max.   :43.0000   Max.   :26.0000
qty_exclamation_directory qty_space_directory qty_tilde_directory qty_comma_directory
Min.   :-1.0000           Min.   :-1.0000     Min.   :-1.0000     Min.   :-1.0000
1st Qu.:-1.0000           1st Qu.:-1.0000     1st Qu.:-1.0000     1st Qu.:-1.0000
Median : 0.0000           Median : 0.0000     Median : 0.0000     Median : 0.0000
Mean   :-0.2961           Mean   :-0.2973     Mean   :-0.2937     Mean   :-0.2975
3rd Qu.: 0.0000           3rd Qu.: 0.0000     3rd Qu.: 0.0000     3rd Qu.: 0.0000
Max.   : 9.0000           Max.   : 9.0000     Max.   : 5.0000     Max.   : 5.0000
qty_plus_directory qty_asterisk_directory qty_hashtag_directory qty_dollar_directory
Min.   :-1.0000    Min.   :-1.0000        Min.   :-1.0000       Min.   :-1.0000
1st Qu.:-1.0000    1st Qu.:-1.0000        1st Qu.:-1.0000       1st Qu.:-1.0000
Median : 0.0000    Median : 0.0000        Median : 0.0000       Median : 0.0000
Mean   :-0.2969    Mean   :-0.2929        Mean   :-0.2985       Mean   :-0.2965
3rd Qu.: 0.0000    3rd Qu.: 0.0000        3rd Qu.: 0.0000       3rd Qu.: 0.0000
Max.   :19.0000    Max.   :60.0000        Max.   : 0.0000       Max.   :10.0000
qty_percent_directory directory_length   qty_dot_file      qty_hyphen_file   qty_underline_file
Min.   : -1.0000      Min.   :  -1.00   Min.   :-1.00000   Min.   :-1.0000   Min.   :-1.0000
1st Qu.: -1.0000      1st Qu.:  -1.00   1st Qu.:-1.00000   1st Qu.:-1.0000   1st Qu.:-1.0000
Median :  0.0000      Median :   7.00   Median : 0.00000   Median : 0.0000   Median : 0.0000
Mean   : -0.2192      Mean   :  16.92   Mean   :-0.04575   Mean   :-0.2111   Mean   :-0.2605
3rd Qu.:  0.0000      3rd Qu.:  27.00   3rd Qu.: 0.00000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
Max.   :174.0000      Max.   :1286.00   Max.   :12.00000   Max.   :21.0000   Max.   :17.0000
qty_slash_file    qty_questionmark_file qty_equal_file    qty_at_file       qty_and_file
Min.   :-1.0000   Min.   :-1.0000       Min.   :-1.000   Min.   :-1.0000   Min.   :-1.0000
1st Qu.:-1.0000   1st Qu.:-1.0000       1st Qu.:-1.000   1st Qu.:-1.0000   1st Qu.:-1.0000
Median : 0.0000   Median : 0.0000       Median : 0.000   Median : 0.0000   Median : 0.0000
Mean   :-0.2985   Mean   :-0.2985       Mean   :-0.296   Mean   :-0.2981   Mean   :-0.2964
3rd Qu.: 0.0000   3rd Qu.: 0.0000       3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
Max.   : 0.0000   Max.   : 0.0000       Max.   : 3.000   Max.   : 2.0000   Max.   : 3.0000
qty_exclamation_file qty_space_file    qty_tilde_file    qty_comma_file    qty_plus_file
Min.   :-1.000       Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000
1st Qu.:-1.000       1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
Median : 0.000       Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
Mean   :-0.297       Mean   :-0.2979   Mean   :-0.2981   Mean   :-0.2977   Mean   :-0.2974
3rd Qu.: 0.000       3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
Max.   : 4.000       Max.   : 9.0000   Max.   : 4.0000   Max.   : 5.0000   Max.   :19.0000
```

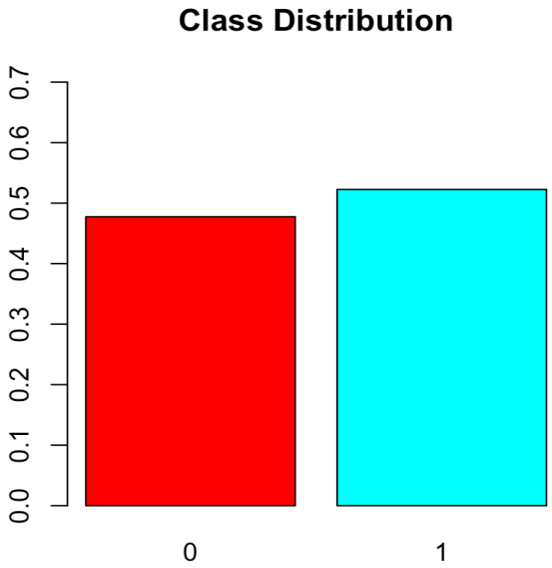
```
qty_asterisk_file qty_hashtag_file  qty_dollar_file   qty_percent_file    file_length
Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000   Min.   : -1.0000   Min.   :  -1.000
1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.: -1.0000   1st Qu.:  -1.000
Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :  0.0000   Median :   0.000
Mean   :-0.2968   Mean   :-0.2985   Mean   :-0.2985   Mean   : -0.2423   Mean   :   4.659
3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:  0.0000   3rd Qu.:   9.000
Max.   :60.0000   Max.   : 0.0000   Max.   : 0.0000   Max.   :174.0000   Max.   :1232.000
qty_dot_params    qty_hyphen_params qty_underline_params qty_slash_params
Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000      Min.   :-1.0000
1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000      1st Qu.:-1.0000
Median :-1.0000   Median :-1.0000   Median :-1.0000      Median :-1.0000
Mean   :-0.7145   Mean   :-0.8165   Mean   :-0.7918      Mean   :-0.8309
3rd Qu.:-1.0000   3rd Qu.:-1.0000   3rd Qu.:-1.0000      3rd Qu.:-1.0000
Max.   :23.0000   Max.   :35.0000   Max.   :21.0000      Max.   :43.0000
qty_questionmark_params qty_equal_params  qty_at_params     qty_and_params
Min.   :-1.0000         Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000
1st Qu.:-1.0000         1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
Median :-1.0000         Median :-1.0000   Median :-1.0000   Median :-1.0000
Mean   :-0.8607         Mean   :-0.5877   Mean   :-0.8462   Mean   :-0.6808
3rd Qu.:-1.0000         3rd Qu.:-1.0000   3rd Qu.:-1.0000   3rd Qu.:-1.0000
Max.   : 9.0000         Max.   :23.0000   Max.   :10.0000   Max.   :22.0000
qty_exclamation_params qty_space_params  qty_tilde_params  qty_comma_params  qty_plus_params
Min.   :-1.0000        Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.0000   Min.   :-1.000
1st Qu.:-1.0000        1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.000
Median :-1.0000        Median :-1.0000   Median :-1.0000   Median :-1.0000   Median :-1.000
Mean   :-0.8727        Mean   :-0.8733   Mean   :-0.8734   Mean   :-0.8712   Mean   :-0.871
3rd Qu.:-1.0000        3rd Qu.:-1.0000   3rd Qu.:-1.0000   3rd Qu.:-1.0000   3rd Qu.:-1.000
Max.   :10.0000        Max.   : 4.0000   Max.   : 1.0000   Max.   :11.0000   Max.   : 6.000
qty_asterisk_params qty_hashtag_params qty_dollar_params qty_percent_params params_length
Min.   :-1.0000     Min.   :-1.0000    Min.   :-1.0000   Min.   :-1.0000    Min.   :  -1.000
1st Qu.:-1.0000     1st Qu.:-1.0000    1st Qu.:-1.0000   1st Qu.:-1.0000    1st Qu.:  -1.000
Median :-1.0000     Median :-1.0000    Median :-1.0000   Median :-1.0000    Median :  -1.000
Mean   :-0.8733     Mean   :-0.8734    Mean   :-0.8729   Mean   :-0.7916    Mean   :   8.482
3rd Qu.:-1.0000     3rd Qu.:-1.0000    3rd Qu.:-1.0000   3rd Qu.:-1.0000    3rd Qu.:  -1.000
Max.   : 4.0000     Max.   : 0.0000    Max.   : 4.0000   Max.   :65.0000    Max.   :4094.000
tld_present_params   qty_params       email_in_url     time_response       domain_spf
Min.   :-1.0000    Min.   :-1.0000   Min.   :0.00000   Min.   :-1.0000   Min.   :-1.00000
1st Qu.:-1.0000    1st Qu.:-1.0000   1st Qu.:0.00000   1st Qu.: 0.2147   1st Qu.: 0.00000
Median :-1.0000    Median :-1.0000   Median :0.00000   Median : 0.4263   Median : 0.00000
Mean   :-0.8364    Mean   :-0.6368   Mean   :0.02771   Mean   : 0.7764   Mean   :-0.03015
3rd Qu.:-1.0000    3rd Qu.:-1.0000   3rd Qu.:0.00000   3rd Qu.: 0.8540   3rd Qu.: 0.00000
Max.   : 1.0000    Max.   :23.0000   Max.   :1.00000   Max.   :38.4024   Max.   : 1.00000
    asn_ip       time_domain_activation time_domain_expiration qty_ip_resolved  qty_nameservers
Min.   :    -1   Min.   :   -1          Min.   :   -1.0        Min.   :-1.000   Min.   : 0.00
1st Qu.: 13335   1st Qu.:   -1          1st Qu.:   -1.0        1st Qu.: 1.000   1st Qu.: 2.00
Median : 20013   Median : 1488          Median :  125.0        Median : 1.000   Median : 2.00
Mean   : 33141   Mean   : 2532          Mean   :  293.6        Mean   : 1.111   Mean   : 2.83
3rd Qu.: 36351   3rd Qu.: 4754          3rd Qu.:  319.0        3rd Qu.: 1.000   3rd Qu.: 4.00
Max.   :395754   Max.   :17775          Max.   :22574.0        Max.   :24.000   Max.   :16.00
qty_mx_servers   ttl_hostname   tls_ssl_certificate qty_redirects     url_google_index
Min.   : 0.00   Min.   :   -1   Min.   :0.0000      Min.   :-1.0000   Min.   :-1.000000
1st Qu.: 1.00   1st Qu.:  288   1st Qu.:0.0000      1st Qu.: 0.0000   1st Qu.: 0.000000
Median : 1.00   Median : 1191   Median :1.0000      Median : 0.0000   Median : 0.000000
Mean   : 1.63   Mean   : 5057   Mean   :0.5014      Mean   : 0.3034   Mean   : 0.001279
3rd Qu.: 2.00   3rd Qu.: 9866   3rd Qu.:1.0000      3rd Qu.: 1.0000   3rd Qu.: 0.000000
Max.   :20.00   Max.   :86400   Max.   :1.0000      Max.   :17.0000   Max.   : 1.000000
domain_google_index url_shortened        phishing
Min.   :-1.000000   Min.   :0.000000   Min.   :0.0000
1st Qu.: 0.000000   1st Qu.:0.000000   1st Qu.:0.0000
Median : 0.000000   Median :0.000000   Median :1.0000
Mean   : 0.002234   Mean   :0.008287   Mean   :0.5226
3rd Qu.: 0.000000   3rd Qu.:0.000000   3rd Qu.:1.0000
Max.   : 1.000000   Max.   :1.000000   Max.   :1.0000
```

By visual inspection of the dataset and running the below series of commands, we can conclude that the dataset is a binary classification dataset.

```
> tab = table(data$phishing)
> print(ifelse(length(tab)==2, "Binary Classification", "MultiClass
+ Classification"))
[1] "Binary Classification"
```

3.2 Class Imbalance Handling

By plotting the mean of number of legitimate instances vs the number of phishing instances, we can check if there exists a class imbalance issue.

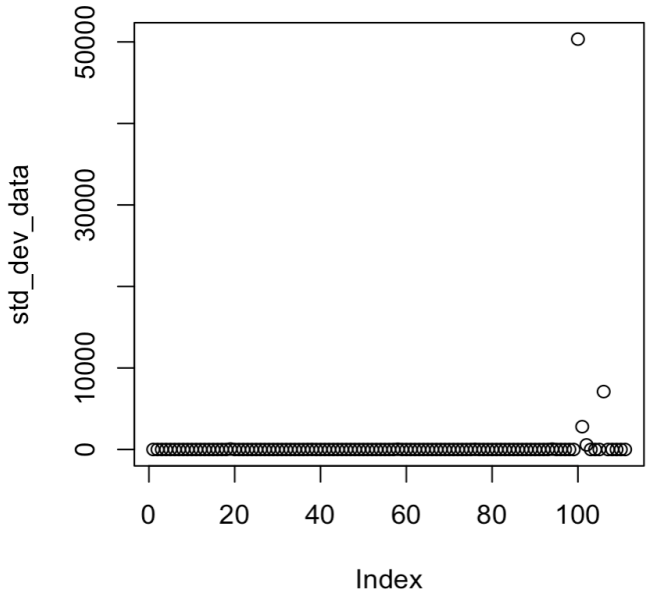


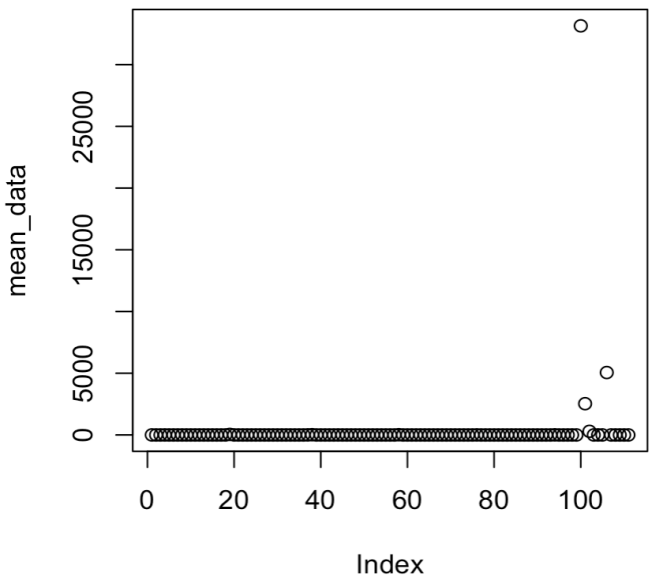
Since the ratio of the number of legitimate instances to the number of phishing instances is 27,998 : 30,647 , we conclude that the dataset is almost balanced. Hence, there is no class imbalance issue in the dataset.

3.3 Analyzing Basic Metrics

```
> library(psych)
> describe(data)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
qty_dot_url	1	58645	2.28	1.47	2	2.07	0.00	1	24	23	4.71	35.70	0.01
qty_hyphen_url	2	58645	0.46	1.34	0	0.16	0.00	0	35	35	6.81	77.69	0.01
qty_underline_url	3	58645	0.17	0.80	0	0.00	0.00	0	21	21	10.40	165.85	0.00
qty_slash_url	4	58645	1.94	2.04	1	1.63	1.48	0	44	44	1.61	6.83	0.01
qty_questionmark_url	5	58645	0.01	0.14	0	0.00	0.00	0	9	9	18.93	797.02	0.00
qty_equal_url	6	58645	0.31	1.16	0	0.03	0.00	0	23	23	5.93	45.43	0.00
qty_at_url	7	58645	0.03	0.34	0	0.00	0.00	0	43	43	74.60	8622.97	0.00
qty_and_url	8	58645	0.21	1.13	0	0.00	0.00	0	26	26	7.75	73.50	0.00
qty_exclamation_url	9	58645	0.00	0.11	0	0.00	0.00	0	10	10	52.93	3917.58	0.00
qty_space_url	10	58645	0.00	0.09	0	0.00	0.00	0	9	9	79.92	7118.85	0.00
qty_tilde_url	11	58645	0.00	0.10	0	0.00	0.00	0	5	5	29.98	1175.98	0.00
qty_comma_url	12	58645	0.00	0.09	0	0.00	0.00	0	11	11	53.92	4458.40	0.00
qty_plus_url	13	58645	0.00	0.14	0	0.00	0.00	0	19	19	71.46	7644.80	0.00
qty_asterisk_url	14	58645	0.01	0.37	0	0.00	0.00	0	60	60	97.85	12983.96	0.00
qty_hashtag_url	15	58645	0.00	0.08	0	0.00	0.00	0	13	13	131.86	19514.76	0.00
qty_dollar_url	16	58645	0.00	0.12	0	0.00	0.00	0	10	10	57.83	3882.48	0.00
qty_percent_url	17	58645	0.16	2.12	0	0.00	0.00	0	174	174	35.01	1996.57	0.01
qty_tld_url	18	58645	1.07	0.31	1	1.00	0.00	0	12	12	5.18	61.68	0.00
length_url	19	58645	44.96	54.71	29	34.41	20.76	4	4165	4161	12.28	615.46	0.23
qty_dot_domain	20	58645	1.80	0.79	2	1.70	0.00	0	21	21	2.97	35.30	0.00
qty_hyphen_domain	21	58645	0.13	0.47	0	0.00	0.00	0	11	11	5.90	59.95	0.00
qty_underline_domain	22	58645	0.00	0.02	0	0.00	0.00	0	2	2	76.82	6535.83	0.00
qty_slash_domain	23	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_questionmark_domain	24	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_equal_domain	25	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_at_domain	26	58645	0.00	0.00	0	0.00	0.00	0	1	1	242.15	58638.00	0.00
qty_and_domain	27	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_exclamation_domain	28	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_space_domain	29	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_tilde_domain	30	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_comma_domain	31	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_plus_domain	32	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_asterisk_domain	33	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_hashtag_domain	34	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_dollar_domain	35	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_percent_domain	36	58645	0.00	0.00	0	0.00	0.00	0	0	0	NaN	NaN	0.00
qty_vowels_domain	37	58645	5.44	2.71	5	5.27	2.97	0	61	61	1.73	13.95	0.01





When we plot the standard deviation and mean of the features in the dataset, we can see that most of the values are between 0 and 1. There are also a few outliers which will be removed in the process. Further, during the training and testing phases, we can decide if the data has to be normalized or not.

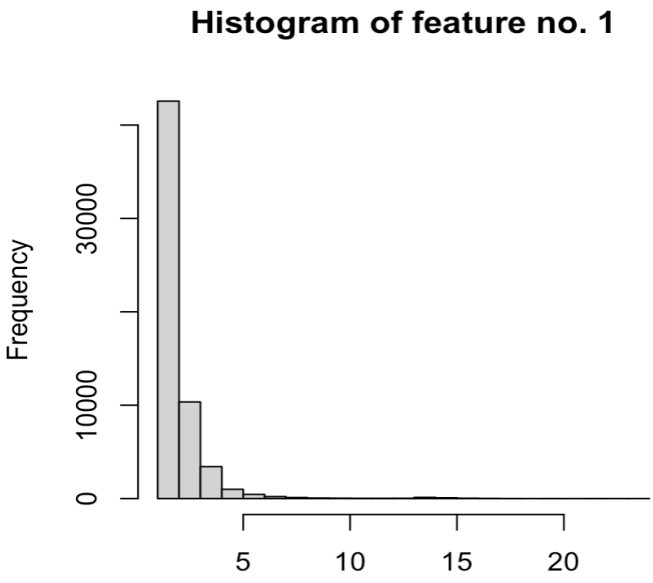
3.4 Missing Values in the dataset

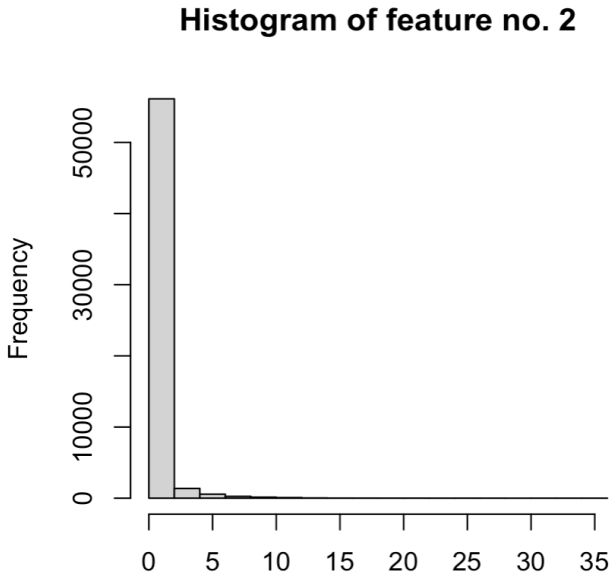
We run “is.na()” command to check if any null values are present in the dataset. As seen from the output of the command below, there are no null values in the dataset.

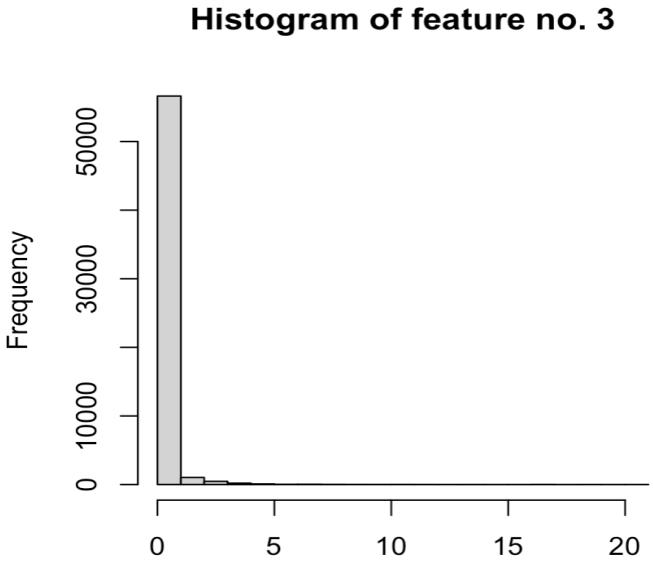
```
> sum(is.na(data))
[1] 0
```

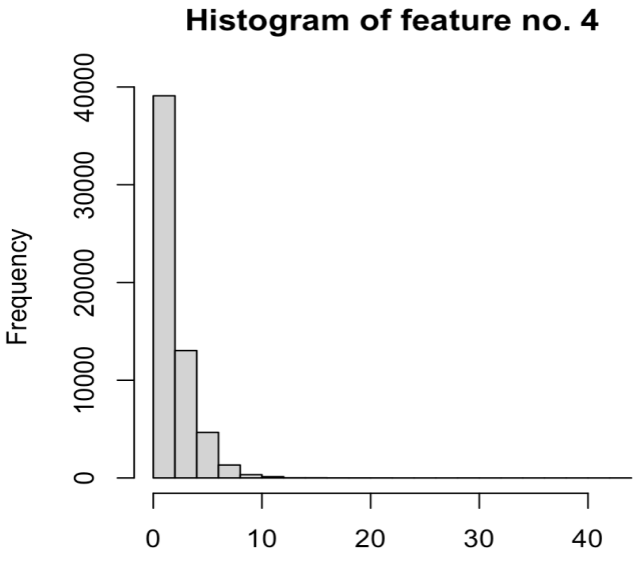
3.5 Outlier Treatment

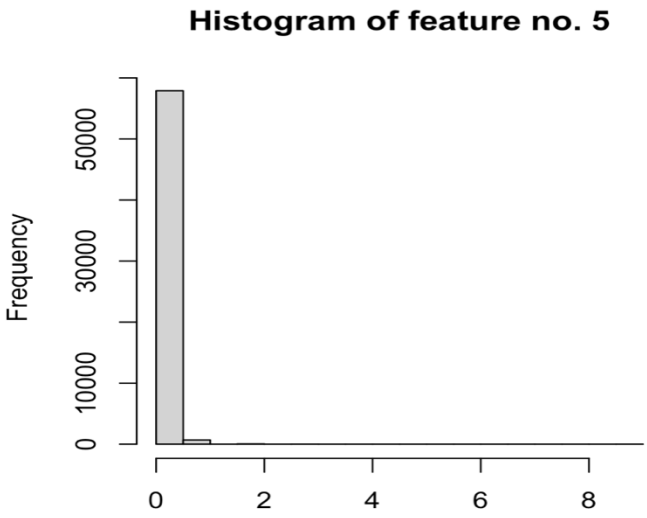
Histograms of the first 8 features are plotted below. Most of the values of these features are less than 5. Since most of the features in the dataset are numerical, they do not have any specific ranges. However, there are outliers which were shown in the standard deviation and mean graph of all features.

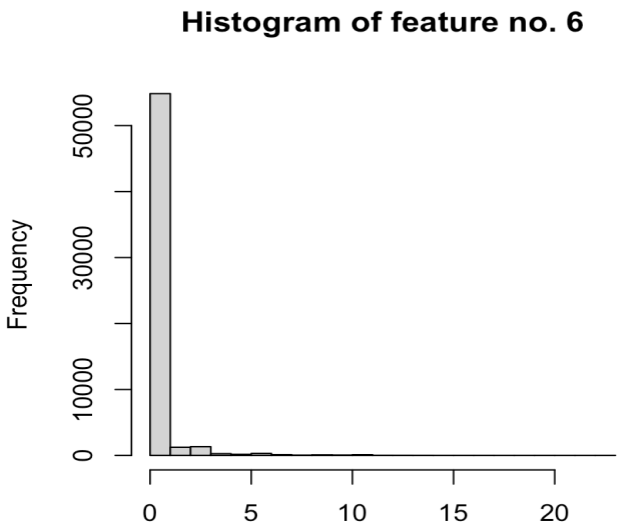


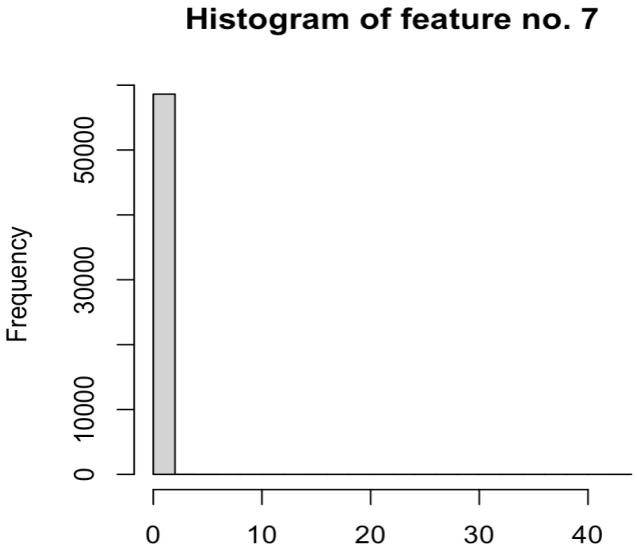


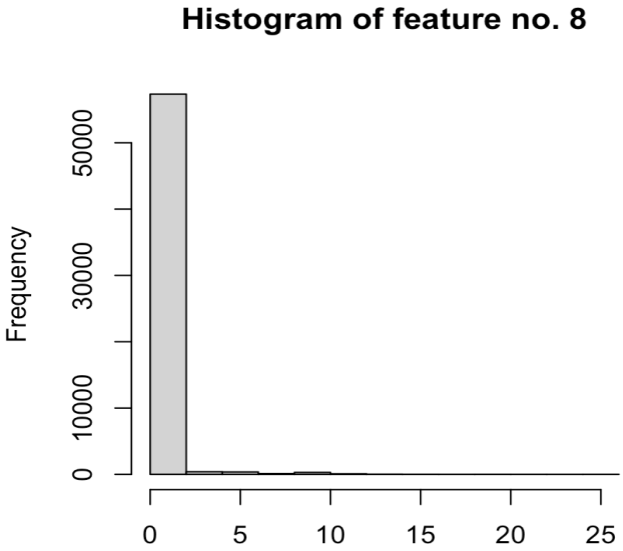













```
> for(i in 1:78) {
+     data_mean=mean(ncpdata.new[,i])
+     data_sd=sd(ncpdata.new[,i])
+     low_cutoff=data_mean-3*data_sd
+     upper_cutoff=data_mean+3*data_sd
+     outliers_idx=which(ncpdata.new[,i]<low_cutoff | ncpdata.new[,i]>upper_cutoff)
+     outliers_row=c(outliers_row,outliers_idx)
+ }
> outliers_row=unique(outliers_row)
> print(paste("Number of Outilers =",length(outliers_row)))
[1] "Number of Outilers = 18671"
```

The above code is used to sum up the number of outliers that fall outside three standard deviations from the mean, in each feature column. The total number of outliers are 18,671. These will be handled depending on the fit of the models and accuracy.

3.6 Constant and Correlated Features

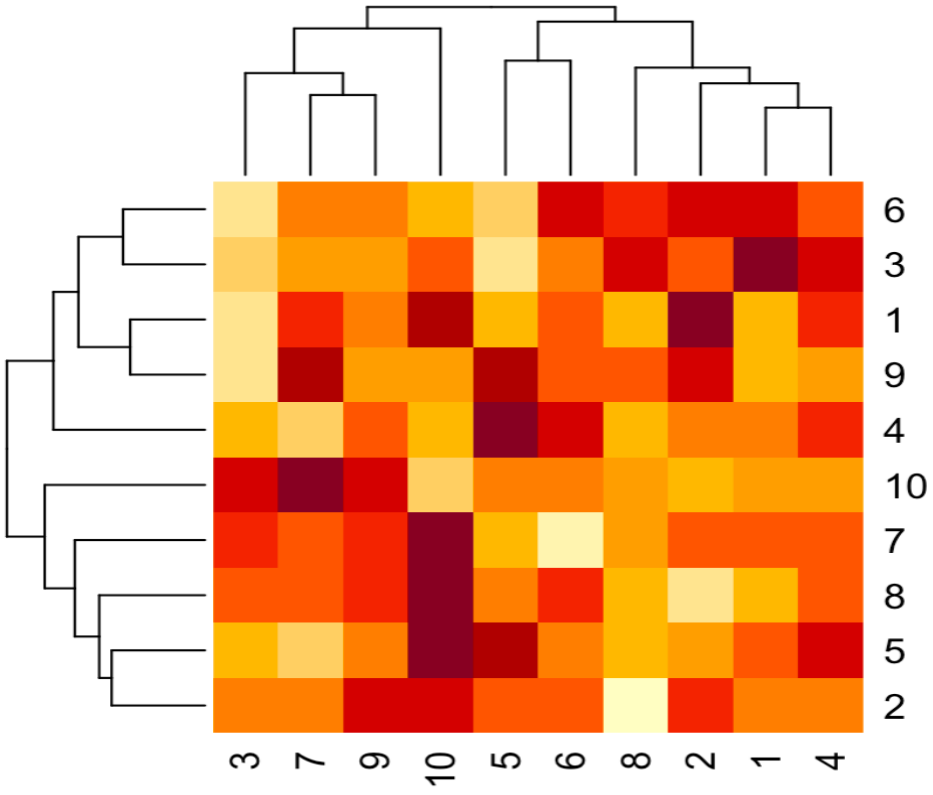
We inspect the dataset for constant features. We eliminate the constant features as shown below. We also inspect the dataset to find correlated features. We then remove the correlated features and store the result in a different data frame to compare the performance metrics. A correlation coefficient of 0 indicates that the predictors are not correlated, and a value of 1 indicates that they are correlated. Accordingly, the criteria to identify two predictors to be correlated is to have a correlation coefficient larger than 0.5 between those predictors. Scatterplot is drawn between the first 6 features and a heatmap is drawn for the first 10 features as shown below. By visual inspection, we can get to know that feature 1 is related to 3 and 4. More detailed analysis is done using the code given below.

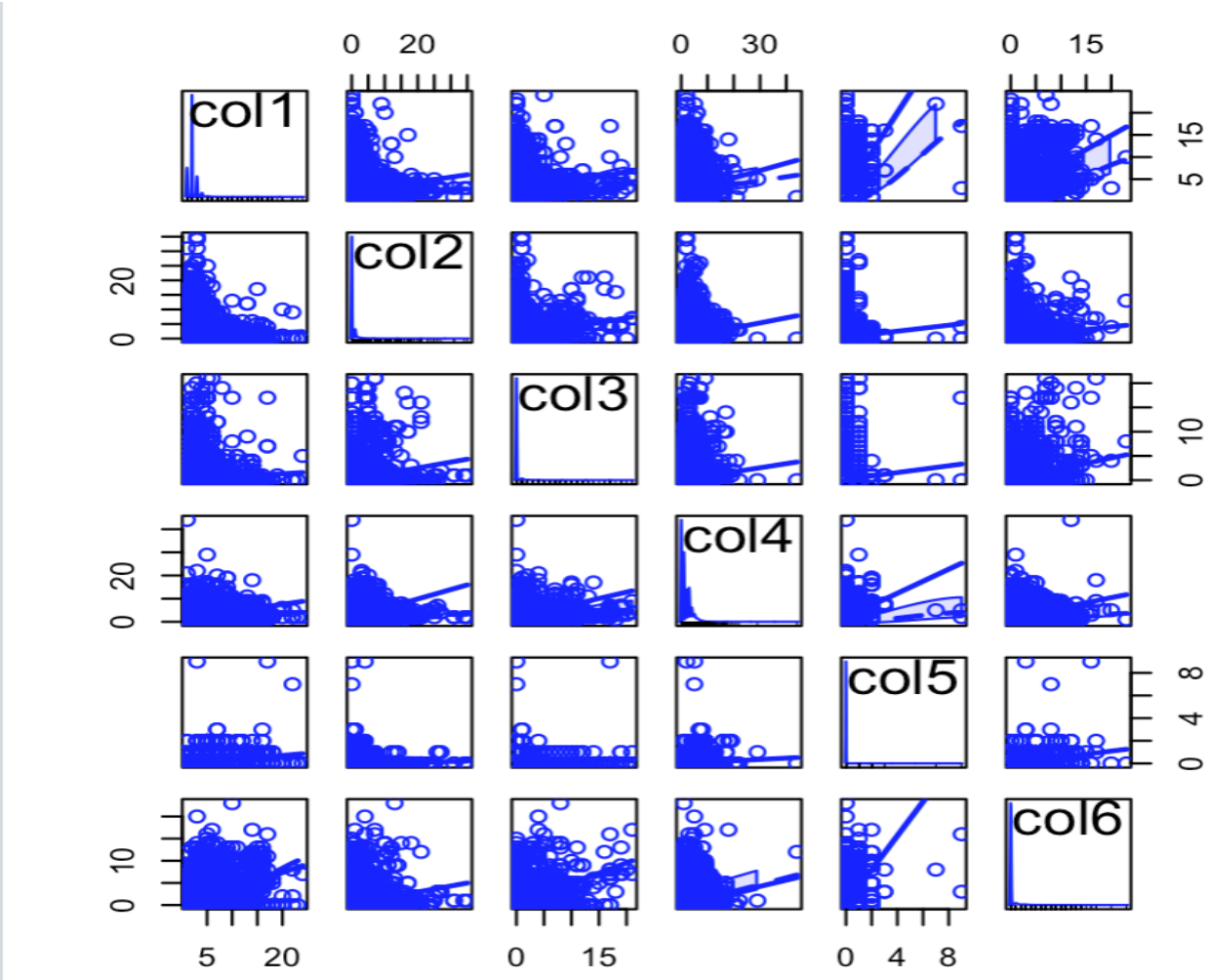
```
> const_pred=unlist(lapply(1:111,FUN=function(x) {
+ TBL=table(data[[x]])
+ ifelse(length(names(TBL))<2,-1*x,x)}))
> print(ifelse(any(const_pred<0),"Constant Predictors Exist","No Constant
+ Predictors"))
[1] "Constant Predictors Exist"
```

```
> ncpdata=data[,const_pred>0]
> dim(ncpdata)
[1] 58645    99
```

```
> const_pred=unlist(lapply(1:98,FUN=function(x) {
+     TBL=table(ncpdata[[x]])
+     ifelse(length(names(TBL))<2,-1*x,x)}))
> print(ifelse(any(const_pred<0),"Constant Predictors Exist","No Constant Predictors"))
[1] "No Constant Predictors"
```

```
> cordata=cor(ncpdata[,1:98])
> print(ifelse(any(abs(cordata[cordata!=1])>0.5),"Correlated Predictors Exist","
+ No Correlated Predictors"))
[1] "Correlated Predictors Exist"
```


6
3
1
9
4
10
7
8
5
2
3 7 9 10 5 6 8 2 1 4


col1
col2
col3
col4
col5
col6
0 20
0 30
0 15
0 20
0 10
0 20
0 4 8
0 10
5 20
0 15
0 4 8

The below code is used to eliminate correlated features in the dataset. We then divide the dataset into train and test data to determine the feature importance and VIF.

```
> tmp <- cor(ncpdata)
> tmp[upper.tri(tmp)] <- 0
> diag(tmp) <- 0
> ncpdata.new <-
+ ncpdata[, !apply(tmp, 2, function(x) any(abs(x) > 0.99, na.rm = TRUE))]
> dim(ncpdata.new)
[1] 58645    79
> set.seed(43)
> tridx<-sample(1:nrow(ncpdata.new),0.7*nrow(ncpdata.new),replace=F)
> traindata<-ncpdata.new[tridx,]
> testdata<-ncpdata.new[-tridx,]
> table(traindata$phishing)

    0     1
19718 21333
```

3.7 VIF of features

```
> detailswhole = glm(phishing~.,data=traindata,family="binomial")
```

```
> summary(detailswhole)

Call:
glm(formula = phishing ~ ., family = "binomial", data = traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0531  -0.2738   0.0000   0.3218   5.0215

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.929e+00  3.164e-01   6.096 1.09e-09 ***
qty_dot_url             -9.023e-01  6.609e-01  -1.365 0.172137
qty_hyphen_url          -8.061e-02  2.794e-01  -0.289 0.772913
qty_underline_url       -4.600e-02  3.636e-01  -0.126 0.899340
qty_slash_url            5.951e-01  4.090e-01   1.455 0.145691
qty_questionmark_url    -1.402e+00  1.921e+00  -0.730 0.465338
qty_equal_url            4.172e-01  2.452e-01   1.702 0.088818 .
qty_at_url               6.268e+01  9.769e+02   0.064 0.948838
qty_and_url             -5.592e-01  3.096e-01  -1.806 0.070922 .
qty_exclamation_url      6.200e-01  5.465e-01   1.134 0.256609
qty_space_url            1.437e+01  7.202e+02   0.020 0.984077
qty_tilde_url            1.851e+01  4.381e+03   0.004 0.996629
qty_comma_url           -3.857e+00  4.331e-01  -8.906  < 2e-16 ***
qty_plus_url            -1.342e+00  1.047e+00  -1.283 0.199628
qty_asterisk_url         1.229e+01  3.756e+02   0.033 0.973897
qty_hashtag_url          1.284e+01  1.060e+03   0.012 0.990339
qty_dollar_url           1.245e-01  4.418e+02   0.000 0.999775
qty_percent_url          2.887e-01  1.232e+00   0.234 0.814711
qty_tld_url              5.473e-01  1.518e-01   3.605 0.000312 ***
length_url               2.296e-02  1.981e-02   1.159 0.246522
qty_dot_domain          -3.613e-01  6.612e-01  -0.546 0.584748
qty_hyphen_domain        7.109e-01  2.835e-01   2.508 0.012150 *
qty_underline_domain     1.715e+01  2.677e+03   0.006 0.994888
```

```
> vif(detailswhole)
           qty_dot_url          qty_hyphen_url      qty_underline_url           qty_slash_url
          1.067035e+03           3.638519e+02           9.708330e+01           7.928506e+02
  qty_questionmark_url          qty_equal_url             qty_at_url            qty_and_url
          5.265569e+00           4.859351e+01           6.905056e+05           4.831475e+01
   qty_exclamation_url          qty_space_url          qty_tilde_url          qty_comma_url
          1.262617e+00           1.000004e+00           2.098472e+08           1.243927e+00
          qty_plus_url       qty_asterisk_url       qty_hashtag_url         qty_dollar_url
          4.875978e+01           4.508320e+00           1.000000e+00           7.706461e+00
       qty_percent_url            qty_tld_url             length_url         qty_dot_domain
          1.131743e+03           1.552938e+00           4.084662e+02           6.940181e+02
     qty_hyphen_domain   qty_underline_domain          qty_at_domain      qty_vowels_domain
          4.096528e+01           1.000000e+00           1.007286e+00           4.657516e+00
         domain_length           domain_in_ip   server_client_domain      qty_dot_directory
          6.933900e+01           1.137730e+00           1.017637e+00           4.875506e+02
  qty_hyphen_directory qty_underline_directory    qty_slash_directory    qty_equal_directory
          3.707978e+02           1.486416e+02           1.079262e+03           1.585255e+02
      qty_at_directory       qty_and_directory    qty_tilde_directory qty_asterisk_directory
          3.107706e+08           3.241935e+02           7.313833e+09           7.078384e+07
  qty_dollar_directory   qty_percent_directory       directory_length           qty_dot_file
          8.894865e+07           1.144699e+03           2.874417e+02           2.436374e+01
       qty_hyphen_file      qty_underline_file          qty_plus_file       qty_asterisk_file
          5.525177e+00           1.005084e+01           4.994247e+02           8.700574e+07
       qty_dollar_file        qty_percent_file            file_length          qty_dot_params
          7.437324e+09           1.501385e+01           8.248794e+00           1.526203e+02
     qty_hyphen_params    qty_underline_params       qty_slash_params qty_questionmark_params
          2.399240e+01           4.354516e+01           4.338350e+01           6.615586e+02
      qty_equal_params          qty_at_params        qty_and_params  qty_exclamation_params
          2.182651e+02           1.094103e+08           1.348886e+02           5.245511e+07
      qty_comma_params         qty_plus_params      qty_dollar_params      qty_percent_params
          1.981422e+07           1.944629e+02           3.713962e+07           9.235740e+02
         params_length      tld_present_params             qty_params            email_in_url
          1.006946e+02           4.642762e+01           7.521719e+01           8.085958e+04
         time_response             domain_spf                 asn_ip  time_domain_activation
          1.044526e+00           1.026812e+00           1.034804e+00           1.540717e+00
 time_domain_expiration        qty_ip_resolved       qty_nameservers          qty_mx_servers
          1.250294e+00           1.178742e+00           1.112202e+00           1.214904e+00
          ttl_hostname     tls_ssl_certificate          qty_redirects        url_google_index
          1.091567e+00           1.221029e+00           1.173736e+00           1.250909e+00
   domain_google_index          url_shortened
          1.255052e+00           1.022170e+00
```

VIF values less than 5 indicate that the features are not correlated and can be used in the classification models. Some of the features are time_responce, domain_spf, asn_ip, time_domain_activation, time_domain_expiration, qty_ip_resolved, qty_nameservers, qty_mx_servers, ttl_hostname, tls_ssl_certificate and so on. This is also pointed out by the variable importance as shown in the next steps.

3.8 Variable Importance of features

We use decision trees and random forest to determine the variable importance of the features in the dataset as shown below. The top features are listed below. These features will be used in the classification models for training and prediction.

```
> tree.data2 <- rpart(phishing~., data=data.new)
> tree.data2$variable.importance
       directory_length           qty_slash_url     qty_slash_directory              length_url
             6709.497313             5985.149053             5982.084656             5086.582060
     qty_dot_directory    qty_hyphen_directory time_domain_activation time_domain_expiration
             5043.450177             5043.357697             1807.078800              603.176395
 qty_asterisk_directory qty_underline_directory           domain_length          qty_dot_domain
              572.657786              572.657786              213.466304              173.342423
         time_response       qty_vowels_domain          qty_mx_servers             qty_dot_url
              150.650719              138.579912              138.146398               80.893131
         qty_redirects            ttl_hostname         qty_ip_resolved         qty_nameservers
               74.717897               47.963252               46.774618               27.122563
            domain_spf      qty_percent_params                  asn_ip        qty_percent_file
               22.839650               22.543429               11.086192                9.263804
         qty_plus_file           qty_plus_url
                0.184960                0.184960
```

```
> ncpdata.new$phishing=as.factor(ncpdata.new$phishing)
> rfmodel = randomForest(phishing~., ncpdata.new)
> rfmodel_features = varImp(rfmodel)
> rfmodel_features = rfmodel_features[order(rfmodel_features, decreasing = TRUE),,drop = FALSE]
Warning message:
In xtfrm.data.frame(x) : cannot xtfrm data frames
> top15features = row.names(rfmodel_features[1:15,,drop = FALSE])
> top15features
 [1] "directory_length"       "time_domain_activation" "length_url"             "qty_slash_directory"
 [5] "qty_slash_url"          "file_length"            "qty_asterisk_file"      "ttl_hostname"
 [9] "asn_ip"                 "time_domain_expiration" "qty_dot_file"           "qty_plus_file"
[13] "qty_tilde_directory"    "qty_dot_directory"      "time_response"
```

3.9 Dimensionality Reduction with PCA

We perform this step if there is an overfit or underfit of the classification model. Hence, this is implemented if required at a later stage.

Our dataset has been split into training and testing data and is ready to work with classification methods.