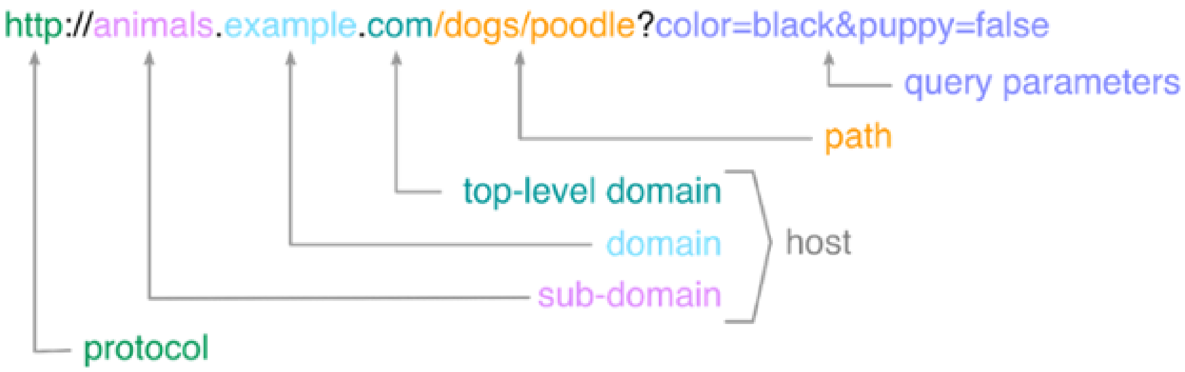
Machine Learning Project

Explaining the different features of the data set and what they represent.



THRESHOLD = 500
(By Default - Will be a hyperparameter within the algorithms)
(Where we saying deleted, it will DEPEND on the THRESHOLD)

Feature	Description
url	The 'x' value (in subject)
length_url	Length of the URL
length_hostname	Length of the Host Name of the URL
ip	The IP is in the hostname
nb_hyphens	How many - signs there are in the URL
nb_dots	How many . signs there are in the URL
nb_at	How many @ signs there are in the URL
nb_qm	How many ? signs there are in the URL

nb_and	How many & signs there are in the URL
nb_or	How many || signs there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_eq	How many = signs there are in the URL
nb_underscore	How many _ signs there are in the URL
nb_tilde	How many ~ signs there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_percent	How many % signs there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_slash	How many / signs there are in the
nb_star	How many * signs there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_colon	How many : signs there are in the URL
nb_comma	How many , signs there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_semicolumn	How many ; signs there are in the URL

	DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_dollar	How many $ signs there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_space	How many spaces there are in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_www	How many 'www' there are in the URL
nb_com	How many 'com' there are in the URL
nb_dslash	Number of '//' in the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD Not sure if we should delete.
http_in_path	How many 'http(s)' there are in the URL (doesn't include starting http/s protocol) DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_redirection	When accessing a particular URL, does it redirect to another URL within the same domain ? If so then how many times
nb_external_redirection	When accessing a particular URL, does it redirect to another URL outside the domain ? If so then how many times

	DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
length_words_raw	How many raw words there are in the URL excluding the “https://www” part
char_repeat	Number of repeated characters before the top level domain
shortest_words_raw	The shortest word in the URL
shortest_word_host	The shortest word in the host
shortest_word_path	The shortest word in the path
longest_words_raw	The longest word in the URL
longest_word_host	The longest word in the host
longest_word_path	The longest word in the path
avg_words_raw	The average word length
avg_word_host	The average word length in the host
avg_word_path	The average length of words in the path
phish_hints	It looks like it gives a hint if it’s phishing where = 0 means not phishing and != 0 means it is phishing
domain_in_brand	1 if the domain is the brands name and 0 otherwise.
brand_in_subdomain	If the brand name is contained in the subdomain of the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
brand_in_path	If the brand name is contained in the path of the URL DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD

suspecious_tld	In the name - a suspicious top-level domain. = 0 (usually legitimate) = 1 (mostly phishing) DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
statistical_report	If the IP address matches one of the main phishing domains across the web. DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_hyperlinks	I think it's the number of hyperlinks on the website.
ratio_intHyperlinks	Ratio of internal hyperlink tags in the website.
ratio_extHyperlinks	Ratio of external hyperlink tags in the website.
ratio_nullHyperlinks	Deleted as there is only 1 unique value for the column DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_extCSS	Number of external CSS files. *External file sheet def - With an external style sheet, you can change the look of an entire website by changing just one file!*
ratio_intRedirection	Ratio of internal redirections into the same website. DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
ratio_extRedirection	Ratio of external redirections into the same website.
abnormal_subdomain	If the URL does NOT contain a subdomain in the host or has a shortened version of a subdomain DELETING AS THERE ARE 2 UNIQUE

	VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
random_domain	If the URLs use words formed from random characters
ratio_intErrors	Deleted as there is only 1 unique value for the column DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
ratio_extErrors	The ratio between connection errors of external hyperlinks.
login_form	If login form contains external action links or empty actions like "", "#", "#nothing", "#doesnotexist", "#null", "#void", "#whatever", " #content "," javascript :: void (0) "," javascript :: void (0); "," javascript ::; "," javascript ".
external_favicon	If the website is using an external (not stored locally on server) link to a favicon
links_in_tags	Ratio of internal links (links to the same domain).
submit_email	If the form actions contain "mailto:" or "mail ()". DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
ratio_intMedia	Ratios of internal media file links (same domain).
ratio_extMedia	Ratios of external media file links (same domain).
sfh	Does the form have an empty string or "about: blank" DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
iframe	Is there an invisible frame on the site.

	DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
popup_window	Does the site have pop-up windows with text fields. DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
https_token	Is it https protocol? If so, it will be 0. Else it is the http protocol, 1
ratio_digits_url	The length of the URL divided by the number of digits in URL
punycode	Is the URL in ASCII or is it encoded in punycode as an internationalized domain names (IDN) DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
port	Does the URL contain a port address DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
tld_in_path	Is there a Top Level Domain in the path of the address (refer to the document image)
tld_in_subdomain	Is there a Top Level Domain in the subdomain of the address DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
nb_subdomains	Number of subdomains in the URL
prefix_suffix	If there is either a prefix or a suffix somewhere in the URL, 1 for True, 0 for False
shortening_service	If the website uses a shortening service to shorten their URL in some

	form

safe_anchor	An anchor is an element defined by the tag <a>* The number of unsafe anchors. Tags with one of the following links {'#', 'javascript', 'mailto'} are considered dangerous.
onmouseover	Does the site contain the “OnMouseOver“ attribute in javascript. DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
right_clic	If you right click on a element with the “OnMouseOver“ attribute. DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
empty_title	Whether the URL contains a title tag or not
domain_in_title	Domain name in title tag
domain_with_copyright	Whether or not the domain name has a copyright tag within it.
whois_registered_domain	We not sure what this is so we shall delete it, it has a low correlation with the answer column. DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
domain_registration_length	The period that the domain has been registered for.
domain_age	The age of the domain over a

	certain period of time.
web_traffic	This represents the amount of web traffic the site gets over a certain period of time
dns_record	Has a DNS record that provides important information about the domain or hostname DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD
google_index	Check if site is indexed by google
page_rank	A Google search ranking to determine how important a page is
path_extension	Are there any malicious path extensions such as, "exe "," js" DELETING AS THERE ARE 2 UNIQUE VALUES WITH ONE OF THEM BEING LESS THAN THE THRESHOLD