Descriptions of Prediction Features and Target Variable for Detection of Phishing Webpages

Feature #	Feature Identifier	Feature Name	Feature Description
1	DomainIdentityInPage	Domain identity in a webpage	Number of times a domain appears in the webpage structure and contents.
2	DomainIdentityInCopyright	Domain identity in copyright	Domain in a URL is checked if it matches with the copyright information in the contents or not. If the two are mismatching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
3	DomainIdentityInCanonical	Domain in canonical URL	Domain in a URL is compared against a common domain retrieved from canonical URLs. If the two are mismatching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
4	DomainIdentityInAlternate	Domain in alternate URL	Domain in a URL is compared against a common domain retrieved from alternate URLs. If the two are mismatching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
5	ForeignHyperlinks	Foreign domains in links	Domain in a URL is compared against a common domain retrieved from all non-object hyperlinks. If the two are mismatching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
6	VoidHyperlinkRatio	Proportion of void and same webpage links	Ratio of sum of number of void (empty) and number of links pointing to the same webpage divide by the total number of all non-object links.
7	ForeignFormHandler	Foreign form handler	Domain of a form handler of a webpage is compared against a domain in URL and a common domain in non-object links.
8	DomainEncoding	Encoded hostname	Presence of ASCII encoded characters (presented as % followed by two hexadecimal digits) in the hostname is checked. If found, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
9	PathEncodedCharacters	Encoded URL path	Presence of ASCII encoded characters (presented as % followed by two hexadecimal digits) in the URL path is checked. If found, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
10	RedirectionalChar	Use of @ character in a URL	Presence of @ character or its equivalent ASCII representation (%40) in the URL path is checked. If found, the webpage is flagged as a phishing one otherwise it is legitimate webpage.

11	OutPositionedDomain	Domain out of position	The characters *http://*, *https://* and *www* characters and generic or country code Top Level Domain are checked if they have been used more than once in a URL. If not, their positions in the URL will be determined if they are different from the standard ones. If any of the condition is true, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
12	NoDotsFQDN	# dots in hostname	Number of dots in a hostname is counted.
13	NoDotsPath	# dots in the URL path	Number of dots in a URL path is counted.
14	NonStandardPort	Non-standard port number	For a URL that has used a port number, the number is compared against its http protocol. If the number is not 80 for http and 443 for https, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
15	ObfuscationCharFQDN	# obfuscation characters in hostname	Number of '_', '-' and '=' characters in a hostname is counted.
16	ObfuscationCharPath	# obfuscation characters in URL path	Number of '_', '-' and '=' characters in a URL path is counted.
17	NoOfSlash	# forward slashes	Number of '/' in a URL is counted.
18	NoOfCharFQDN	# characters in hostname	Total number of characters in a hostname is counted.
19	NoOfCharPath	# characters in URL path	Total number of characters in a URL path is counted.
20	IPInURL	IP address in a hostname	Presence of an IP address in a hostname is checked. If found, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
21	NumericFQDN	Numeric in a hostname	Number of numeric characters in a hostname is counted.
22	NumericPath	Numeric in a URL path	Number of numeric characters in a URL path is counted.
23	ShortURL	Shortened URLs	Use of shortened URL is checked comparing a hostname of the URL against a list of 242 hostnames of the collected shortening URL services (see appendix III). If found, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
24	SubdomainService	Free domain services	Use of free domain is checked by comparing a domain of a webpage domain against a list of domains of the most abused free domain services we compiled from Anti-Phishing Working Group (APWG)'s reports on global phishing survey between 2008 and 2017. The list is in appendix IV.
25	DomainNameValidity	Domain validity	An expiry date of a webpage's domain registration (from WHOIS database) is compared with the current date to check if it is still valid or not. If it overdue, the webpage is flagged as a phishing one otherwise it is legitimate webpage.

26	FHDomainValidity	Form handler's domain validity	An expiry date of a form handler's domain registration (from WHOIS database) is compared with the current date to check if it is still valid or not. If it overdue, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
27	DomainAge	Domain age	A difference between the current date and the webpage's domain first date of registration (from WHOIS database) is computed.
28	FHDomainAge	Form handler domain's age	A difference between the current date and the form handler's domain first date of registration (from WHOIS database) is computed.
29	SSLCertificateType	Type of SSL certificate	A type of SSL certificate used by the webpage's domain is determined.
30	SSLGeoCountryMatch	Domain, certificate and geolocation country matching	Country names in the ccTLD (for URLs with ccTLDs only), SSL certificate and location of the hosts are compared. If they do not match, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
31	URLSearchEngineRanking	URL ranking in search engines	A URL is searched in the Google and Bing search engines. URLs in the top five results returned by each engine are compared against the searched URL. If none of the results are matching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
32	FQDNSearchEngineRanking	Hostname ranking in search engines	A hostname is searched in the Google and Bing search engines. Hostnames in the top five results returned by each engine are compared against the searched hostname. If none of the results are matching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
33	DomainSearchEngineRanking	Domain ranking in search engines	A domain is searched in the Google and Bing search engines. Domains in the top five results returned by each engine are compared against the searched domain. If none of the results are matching, the webpage is flagged as a phishing one otherwise it is legitimate webpage.
34	FQDNBlacklistIPCounts	Counts of matched hostname's IP address in a phishing blacklist of IP addresses	Number of times an IP address of a hostname appears in a list of IP addresses of blacklisted phishing URLs is counted. A 3-month-old data of a blacklist is used.
35	DomainBlacklistIPCounts	Counts of matched domain's IP address in a phishing blacklist of IP addresses	Number of times an IP address of a domain appears in a list of IP addresses of blacklisted phishing URLs is counted. A 3-month-old data of a blacklist is used.

36	Class	A webpage classified as phishing or legitimate	This is a target labelled variable with two labels: phishing or legitimate