Phishing URL Classification Rules

Reference: https://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf

Rule Name	Feature	Logic	Why It Matters	Classification
IP Address in URL	has_ip_in_url	URL contains raw IP instead of domain	Legitimate websites rarely use raw IPs; phishing sites often do	PHISHING
'@' Symbol in URL	has_at_symbol	URL contains '@'	'@' can redirect to another domain, tricking users	PHISHING
No DNS Record	dns_record_found	Domain does not resolve	Phishing sites may use non-existent or new domains	PHISHING
'https' Token in Domain	https_token_in_domain	Domain name contains 'https'	Used to trick users into thinking site is secure	PHISHING
Domain Age	domain_age_months	Age of domain since registration	Newly created domains are often malicious	PHISHING if 0 months
Domain Expiration	months_to_expire	Months left until domain expires	Short-lived domains often used for temporary phishing campaigns	PHISHING if ≤3 months

Rule Name	Feature	Logic	Why It Matters	Classification
URL Reachability	url_alive	URL responds to requests	Dead or unreachable URLs may indicate abandoned or malicious sites	SUSPICIOUS if unreachable
Free Hosting Domain	is_free_hosting	Domain belongs to known free hosting provider	Free hosting is easier for phishing setup	SUSPICIOUS
Suspicious Keywords in Path	suspicious_path_keyword	URL path contains keywords like login, verify, account	Mimics login or verification pages	SUSPICIOUS
Dash in Domain	has_dash_in_domain	Domain name contains '-'	Used to imitate legitimate domains	SUSPICIOUS
Suspicious File Extensions	suspicious_extension	Path ends with .php, .asp, .aspx, .cgi, .exe	Phishing websites often use dynamic scripts or executables	SUSPICIOUS if combined with other weak signals
Suspicious Query Tokens	suspicious_query_token	Query string contains id=, rand=, login, session	Typical in URLs capturing sensitive info	SUSPICIOUS if combined with other weak signals
Path Depth	path_depth	Number of '/' segments in URL	Very deep URLs can obscure real target	SUSPICIOUS if combined with 2+ weak signals

Rule Name	Feature	Logic	Why It Matters	Classification
Multiple Weak Suspicious Signals	weak_signals	Combination of extension, query token, path depth>=3	Single weak signals might be harmless; multiple together indicate phishing	SUSPICIOUS