URL Analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category | Rule | Description | File Type | Reason | Score |
| Domain Reputation |  |  |  |  |  |
| Focus: The credibility and history of the domain. | Blacklisted Domain | Check if the domain is listed on known blacklists. | Individual File | Requires API calls to blacklist services (e.g., Google Safe Browsing). | 3 |
|  | Domain Age | Check if the domain is newly registered (e.g., less than 30 days). | Individual File | Requires WHOIS lookup, which can be slow. | 2 |
|  | WHOIS Lookup | Check if the domain registration details are hidden or suspicious. | Individual File | Requires external WHOIS database queries. | 2 |
|  | Domain Popularity | Check if the domain is obscure or rarely seen. | Individual File | Requires domain popularity data, which may involve external APIs. | 1 |
| URL Structure |  |  |  |  |  |
| Focus: The format and components of the URL. | Mismatched URLs | Check if the displayed link text differs from the actual URL. | Common File | Simple string comparison. | 3 |
|  | Suspicious TLDs | Check if the TLD is unusual or highrisk (e.g., `.xyz`, `.top`). | Common File | Quick look up in a predefined list of suspicious TLDs. | 2 |
|  | IP Address in URL | Check if the URL contains an IP address instead of a domain name. | Common File | Simple regexes check for IP addresses. | 2 |
|  | Subdomain Abuse | Check if the URL uses excessive or misleading subdomains. | Individual File |  | 2 |
|  | URL Length | Check if the URL is unusually long or contains random characters. | Common File | Basic length check. | 1 |
| URL Behavior |  |  |  |  |  |
| Focus: How the URL behaves when accessed. | Redirects | Check if the URL redirects to another domain. | Individual File | Requires following the URL to check for redirects, which can be slow. | 2 |
|  | Shortened URLs | Check if the URL is shortened (e.g., `bit.ly`, `tinyurl.com`). | Individual File | Requires expanding the URL, which involves external API calls. | 2 |
|  | HTTPS Usage | Check if the URL uses HTTPS (secure protocol). | Individual File | Requires checking the URL protocol, which is lightweight but grouped with heavyweight rules for consistency. | 1 |
|  | SSL Certificate | Check if the SSL certificate is valid and issued by a trusted authority. | Individual File | Requires validating the SSL certificate, which can involve external checks. | 2 |
| Content Analysis |  |  |  |  |  |
| Focus: The content and context of the URL. | Phishing Keywords | Check if the URL contains phishing related keywords (e.g., `login`, `verify`). | Common File | Simple keyword matching. | 1 |
|  | Brand Impersonation | Check if the URL mimics a well-known brand (e.g., `paypa1.com`). | Individual File | Requires comparing the URL to a list of known brands, which can be computationally expensive. | 3 |
|  | Unusual Paths | Check if the URL path is unusual or contains random strings. | Common File | Basic path analysis using regex. | 1 |
| Contextual Analysis |  |  |  |  |  |
| Focus: The relationship between the URL and the email context. | Email Context | Check if the URL aligns with the email's content. | Individual File | Requires analysing the email content, which can be complex. | 2 |
|  | Sender Reputation | Check if the sender's domain matches the URL domain. | Individual File | Requires checking the sender's domain, which may involve external lookups. | 2 |
|  | Frequency of URLs | Check if the URL appears frequently in known phishing campaigns. | Individual File | Requires checking a database of known phishing URLs, which can be slow. | 1 |

Encoding Strategy for URL Features

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Type | Recommended Encoding | Reason |
| TLDs | Nominal | One-Hot Encoding | Limited unique values (e.g., .com, .org), no ordinal meaning. |
| Protocols | Binary | Label Encoding | Only two options (http, https), 0/1 mapping is sufficient. |
| Keywords | Nominal | One-Hot Encoding | Works well for specific phishing-related terms in URLs. |
| Domains | High Cardinality | Target Encoding | Retains predictive power for a large variety of domains. |

Common Types of Visualizations and Their Uses

|  |  |  |
| --- | --- | --- |
| Visualization | Purpose | Example in Your Project |
| Histogram | Show the distribution of a single feature. | Visualize the distribution of URL lengths for phishing and safe URLs. |
| Box Plot | Show the distribution and detect outliers. | Detect outliers in the number of subdomains. |
| Scatter Plot | Show the relationship between two features. | Visualize the relationship between URL length and domain age. |
| Bar Chart | Compare values across categories. | Compare the average URL length of phishing and safe URLs. |
| Heatmap | Show correlations between features. | Visualize the correlation matrix of all features in your dataset. |