**Name: Sakshi Rajesh Gailwar (#16840)**

**Point 1]** **Business Problem Understanding Document**

A concise summary of the business problem, including the scope and importance of detecting phishing websites.

Key insights and takeaways from the literature on phishing, focusing on common characteristics, detection challenges, and potential solutions.

1. **Summary of the Business Problem**

Phishing websites are deceptive sites designed to trick users into sharing sensitive information like usernames, passwords, and credit card details. Detecting these websites is crucial because:

* **Impact**- Phishing attacks lead to identity theft, financial loss, and security breaches.
* **Challenge**- Phishing websites often look legitimate, making detection difficult.

1. **Importance of Detecting Phishing Websites**

* **Trust**: Reduces damage to brand reputation from being associated with phishing activities.
* **Security**: Helps protect users from fraud and malicious activities.
* **Prevention**: Early detection can prevent large-scale attacks.

1. **Key Insights from Literature**

* Phishing websites often contain suspicious patterns in their URLs, IP addresses, and metadata.
* Common characteristics include:
  + **URL length**: Long URLs may be more suspicious.
  + **Presence of special characters**: Unusual characters or multiple subdomains.
  + **IP addresses**: Phishing sites sometimes use direct IP addresses instead of domains.
* Challenges in detection:
  + Phishing websites continuously evolve to mimic legitimate websites.
  + Detection methods often need to be updated to account for new tactics.
* Potential solutions:
  + Machine learning models that classify websites based on known patterns.
  + Domain analysis and WHOIS data to check registration details and history.

**Point 2] Dataset Exploration Report**

So, from given data we can observe that there are 89 columns and 11430 rows.

**Number of features**: 88

['url', 'length\_url', 'length\_hostname', 'ip', 'nb\_dots', 'nb\_hyphens',

'nb\_at', 'nb\_qm', 'nb\_and', 'nb\_or', 'nb\_eq', 'nb\_underscore',

'nb\_tilde', 'nb\_percent', 'nb\_slash', 'nb\_star', 'nb\_colon', 'nb\_comma',

'nb\_semicolumn', 'nb\_dollar', 'nb\_space', 'nb\_www', 'nb\_com',

'nb\_dslash', 'http\_in\_path', 'https\_token', 'ratio\_digits\_url',

'ratio\_digits\_host', 'punycode', 'port', 'tld\_in\_path',

'tld\_in\_subdomain', 'abnormal\_subdomain', 'nb\_subdomains',

'prefix\_suffix', 'random\_domain', 'shortening\_service',

'path\_extension', 'nb\_redirection', 'nb\_external\_redirection',

'length\_words\_raw', 'char\_repeat', 'shortest\_words\_raw',

'shortest\_word\_host', 'shortest\_word\_path', 'longest\_words\_raw',

'longest\_word\_host', 'longest\_word\_path', 'avg\_words\_raw',

'avg\_word\_host', 'avg\_word\_path', 'phish\_hints', 'domain\_in\_brand',

'brand\_in\_subdomain', 'brand\_in\_path', 'suspecious\_tld',

'statistical\_report', 'nb\_hyperlinks', 'ratio\_intHyperlinks',

'ratio\_extHyperlinks', 'ratio\_nullHyperlinks', 'nb\_extCSS',

'ratio\_intRedirection', 'ratio\_extRedirection', 'ratio\_intErrors',

'ratio\_extErrors', 'login\_form', 'external\_favicon', 'links\_in\_tags',

'submit\_email', 'ratio\_intMedia', 'ratio\_extMedia', 'sfh', 'iframe',

'popup\_window', 'safe\_anchor', 'onmouseover', 'right\_clic',

'empty\_title', 'domain\_in\_title', 'domain\_with\_copyright',

'whois\_registered\_domain', 'domain\_registration\_length', 'domain\_age',

'web\_traffic', 'dns\_record', 'google\_index', 'page\_rank']

**Target Variable**: The status column represents whether a website is "legitimate" or "phishing" (binary classification).

**Data Types**:

* Numerical data (e.g., length\_url, nb\_dots, nb\_hyphens, etc.).
* Categorical data (e.g., status—"legitimate" or "phishing").

In general, all the feature variables are numerical type and target variable is categorical.

**Feature Descriptions:**

1. **url**: The full URL/link of the website.
   * **Relevance**: The URL is the first thing to check for phishing. Phishing websites often disguise themselves with suspicious URLs.
2. **length\_url**: Length of the URL string.
   * **Relevance**: Longer URLs are sometimes used in phishing attempts to hide malicious content.
   * The **length of the hostname** in a URL refers to the number of characters in the domain name portion of the URL, which comes before the first /.
   * Legitimate websites typically have **shorter, simpler hostnames** that are easy to remember and type.
   * Phishing websites often use **longer, confusing hostnames** in an attempt to impersonate legitimate sites. They may Use **extra subdomains, Include random or irrelevant words, Try to imitate a real hostname by adding extra words or letters,**
   * **Suspiciously long hostnames** (more than 20-30 characters) might be a red flag for phishing.
3. **length\_hostname**: Length of the hostname (the domain part of the URL).
   * **Relevance**: Abnormal lengths may indicate irregular domains, which are often used in phishing sites.
4. **ip**: Whether the website is hosted on an IP address instead of a domain name.
   * **Relevance**: Phishing sites are more likely to use IP addresses instead of registered domain names.
5. **nb\_dots**: Number of dots in the URL.
   * **Relevance**: Multiple dots may be used to confuse users into thinking a website is legitimate when it is not.
6. **nb\_hyphens**: Number of hyphens in the URL.
   * **Relevance**: Phishing websites often use multiple hyphens to resemble legitimate websites (e.g., "google-com" vs. "google.com").
7. **nb\_at**: Number of "@" symbols in the URL.
   * **Relevance**: Presence of "@" in the URL can indicate suspicious or unauthorized redirection behaviour.
8. **nb\_qm**: Number of question marks (?) in the URL.
   * **Relevance**: A high number of question marks in a URL can be a red flag for phishing sites, as they may be used in query strings for malicious purposes.
   *  **0 question marks**: The URL is **not using query parameters**, which is normal for some URLs (like home pages, blogs, etc.), so this is often legitimate.
   *  **1 question mark**: **Legitimate use**, commonly used for search parameters or page filters.
   *  **2 or more question marks**: **Suspicious** or potentially **phishing**, as this may indicate manipulation, hidden redirects, or attempts to obfuscate the true URL.
9. **nb\_and**: Number of "&" symbols in the URL.
   * **Relevance**: Multiple "&" symbols could suggest that the URL is attempting to pass complex parameters, often seen in phishing sites.
   *  **1 or 2 & symbols**: Likely **legitimate** for separating standard query parameters.
   *  **3 or more & symbols**: May be **suspicious** and should be investigated further, as it could indicate an attempt to obfuscate the true URL or inject multiple deceptive parameters.
   *  **Too many & symbols**: Likely indicates **phishing**, as legitimate websites rarely need to include multiple, complex query parameters in their URLs
10. **nb\_or**: Number of "OR" strings in the URL.
    * **Relevance**: Used to confuse search engines or URL filters.
11. **nb\_eq**: Number of "=" symbols in the URL.
    * **Relevance**: An unusual number of equal signs could indicate that the URL is trying to pass parameters in a suspicious way.
    *  **1 to 3 = symbols**: **Normal** for most legitimate websites.
    *  **4 to 6 = symbols**: **Suspicious**, but might still be legitimate for certain complex websites or search engines.
    *  **7 or more = symbols**: **Outlier**, **highly suspicious**, and likely indicative of a **phishing** attempt, especially if combined with other suspicious URL features.
12. **nb\_underscore**: Number of underscores (\_) in the URL.
    * **Relevance**: Some phishing websites use underscores to make the URL appear legitimate, as legitimate sites typically avoid underscores in domain names.

* The **maximum** number of underscores used in phishing URLs is generally **5 or more**, but **URLs with 3-4 underscores** are also considered **suspicious**. The excessive use of underscores in URLs should raise concerns, as legitimate websites rarely use these many underscores. In phishing detection, any URL with **5 or more underscores** would likely be flagged as **outlier** or **suspicious**.

1. **nb\_tilde**: Number of tilde (~) symbols in the URL.
   * **Relevance**: Phishing websites may use the tilde symbol to mimic official websites (e.g., example.com/~username).
2. **nb\_percent**: Number of percent signs (%) in the URL.
   * **Relevance**: Excessive percent signs in a URL can indicate encoding issues or an attempt to deceive users.
   *  **1 to 3 percent signs**: This is **normal** for legitimate websites that need to encode special characters in their URLs.
   *  **4 or more percent signs**: **Suspicious**. This could indicate URL manipulation, obfuscation, or an attempt to hide the actual content or redirection path in phishing URLs.
3. **nb\_slash**: Number of slashes (/) in the URL.
   * **Relevance**: Phishing sites may contain multiple slashes in their URLs to confuse users.
   *  Phishing websites tend to use tactics like **domain name obfuscation**, **redirects**, or hiding malicious URLs with **encoded characters**, but it’s very rare for a legitimate or even a typical phishing URL to have **this many slashes**.
   *  If you encounter URLs with **20-33 slashes**, it would be an **outlier** in your dataset and should be **flagged** for further inspection.
4. **nb\_star**: Number of asterisks (\*) in the URL.
   * **Relevance**: Special characters like asterisks are often used in phishing sites to evade detection or confuse URL filters.
5. **nb\_colon**: Number of colons (:) in the URL.
   * **Relevance**: Colons are part of valid URLs but can be used unusually in phishing sites.
6. **nb\_comma**: Number of commas (,) in the URL.
   * **Relevance**: Commas are unusual in URLs and could be a sign of manipulation or suspicious behaviour.
7. **nb\_semicolumn**: Number of semicolons (;) in the URL.
   * **Relevance**: Similar to commas, semicolons are rare in URLs and may be used for malicious purposes.
8. **nb\_dollar**: Number of dollar signs ($) in the URL.
   * **Relevance**: Dollar signs may indicate suspicious URLs used for redirecting to fraudulent pages.
9. **nb\_space**: Number of spaces in the URL.
   * **Relevance**: Spaces are not typically valid in URLs and can be used to create misleading URLs.
10. **nb\_www**: Number of occurrences of "www" in the URL.
    * **Relevance**: Phishing sites may repeat "www" to appear like a legitimate site (e.g., www.example.www.com).
11. **nb\_com**: Number of occurrences of "com" in the URL.
    * **Relevance**: Multiple uses of "com" can trick users into thinking a site is legitimate (e.g., examplecom.com).
12. **nb\_dslash**: Number of double slashes (//) in the URL.
    * **Relevance**: Double slashes in URLs could be an indication of a phishing attack.
13. **http\_in\_path**: Whether "http" appears in the path of the URL.
    * **Relevance**: Presence of "http" in the path could indicate a misleading URL.
14. **https\_token**: Whether the URL contains "https" (secure protocol).
    * **Relevance**: Phishing websites may still use "https" to appear legitimate, though this alone doesn’t guarantee safety.
15. **ratio\_digits\_url**: Ratio of digits in the URL.
    * **Relevance**: Phishing sites often use numerical strings in the URL to look more like legitimate sites.
16. **ratio\_digits\_host**: Ratio of digits in the hostname.
    * **Relevance**: A higher ratio of digits in the hostname can suggest suspicious behaviour, often seen in phishing sites.
17. **punycode**: Whether the URL uses Punycode (used for non-Latin characters in domain names).
    * **Relevance**: Phishing sites often use Punycode to disguise a domain name by using visually similar characters from other alphabets.
18. **port**: Port number in the URL.
    * **Relevance**: Phishing sites might use unusual ports to avoid detection.
19. **tld\_in\_path**: Whether the top-level domain (TLD) appears in the path of the URL.

* **Relevance**: Phishing sites might include the TLD (e.g., .com) in the URL path to deceive users into thinking the website is legitimate.

1. **tld\_in\_subdomain**: Whether the TLD appears in the subdomain of the URL.

* **Relevance**: Phishing websites may try to include TLDs in the subdomain to make a site look like a trusted one (e.g., bank.com.mysite.com).

1. **abnormal\_subdomain**: Indicates if the subdomain of the URL appears unusual or suspicious.

* **Relevance**: Phishing sites may use strange or excessive subdomains to confuse users (e.g., login.bank.account.example.com).

1. **nb\_subdomains**: Number of subdomains in the URL.

* **Relevance**: A high number of subdomains might indicate a phishing attempt, as malicious sites often use multiple subdomains to appear authentic.

1. **prefix\_suffix**: Indicates if the domain uses prefixes or suffixes that seem unusual (e.g., "www" or "login" in odd places).

* **Relevance**: Phishing sites often use strange prefixes or suffixes to try to look like legitimate sites (e.g., login.example.com).

1. **random\_domain**: Whether the domain appears random or nonsensical.

* **Relevance**: Phishing sites often use random domain names to make detection harder and appear authentic.

1. **shortening\_service**: Whether the URL uses a URL shortening service (e.g., bit.ly).

* **Relevance**: Phishing websites often use URL shorteners to hide the true destination of the link.

1. **path\_extension**: Extension of the path in the URL (e.g., .php, .html, .asp).

* **Relevance**: Phishing sites may have unusual file extensions or unexpected extensions that can make them seem suspicious.

1. **nb\_redirection**: Number of redirection sequences in the URL.

* **Relevance**: Phishing sites may redirect users through multiple links to confuse them or hide the true destination.

1. **nb\_external\_redirection**: Number of external redirections in the URL.

* **Relevance**: Multiple external redirections can be a sign of phishing, as the user may be redirected to a fraudulent website.

1. **length\_words\_raw**: Total length of words in the raw URL.

* **Relevance**: A high length of words could indicate an unusually long and complicated URL structure, potentially pointing to a phishing attempt.

1. **char\_repeat**: Number of repeated characters in the URL.

* **Relevance**: Repeated characters (like aaaaa) could be used to hide the true nature of the URL or trick filters.

1. **shortest\_words\_raw**: Length of the shortest word in the raw URL.

* **Relevance**: Phishing URLs may contain unusually short words to hide malicious intent.

1. **shortest\_word\_host**: Length of the shortest word in the hostname.

* **Relevance**: Shorter words in the hostname might indicate an attempt to hide the true nature of the domain.

1. **shortest\_word\_path**: Length of the shortest word in the path.

* **Relevance**: Similar to the hostname, shorter words in the path may signal unusual, potentially dangerous URLs.

1. **longest\_words\_raw**: Length of the longest word in the raw URL.

* **Relevance**: Longer words in URLs can sometimes be used to make the site appear legitimate or mislead security checks.

1. **longest\_word\_host**: Length of the longest word in the hostname.

* **Relevance**: Phishing sites may have unusually long words in the hostname to confuse users and evade detection.

1. **longest\_word\_path**: Length of the longest word in the path.

* **Relevance**: Phishing sites often use long words in the path to make the URL appear less suspicious.

1. **avg\_words\_raw**: Average word length in the raw URL.

* **Relevance**: A high average word length could indicate a complex and potentially deceptive URL.

1. **avg\_word\_host**: Average word length in the hostname.

* **Relevance**: Longer average words in the hostname may be a sign of phishing.

1. **avg\_word\_path**: Average word length in the path.

* **Relevance**: Phishing sites might use a larger number of complex words in the path to avoid detection.

1. **phish\_hints**: The number of hints (e.g., keywords, phrases) suggesting that the URL might be a phishing attempt.

* **Relevance**: A high number of phishing hints suggests that the website is likely malicious.

1. **domain\_in\_brand**: Whether the domain contains the brand name (e.g., "paypal.com").

* **Relevance**: Phishing sites often try to include brand names in their domains to seem more credible.

1. **brand\_in\_subdomain**: Whether the brand name appears in the subdomain of the URL.

* **Relevance**: Phishing sites frequently use the brand name in subdomains to look like legitimate sites (e.g., paypal.example.com).

1. **brand\_in\_path**: Whether the brand name appears in the path of the URL.

* **Relevance**: This could be another tactic used by phishing sites to deceive users into trusting them.

1. **suspecious\_tld**: Whether the top-level domain (TLD) is suspicious.

* **Relevance**: Some TLDs are often associated with phishing sites or less-trustworthy domains.

1. **statistical\_report**: A flag or value representing whether statistical reports are available for the domain.

* **Relevance**: Legitimate sites often have statistical reports and other indicators of trustworthiness.

1. **nb\_hyperlinks**: Number of hyperlinks on the page.

* **Relevance**: An unusually high number of links could be a sign of a phishing site trying to overwhelm the user.

1. **ratio\_intHyperlinks**: Ratio of internal hyperlinks to total hyperlinks.

* **Relevance**: A high ratio of internal hyperlinks could suggest a more legitimate site, while phishing sites may rely more on external links.

1. **ratio\_extHyperlinks**: Ratio of external hyperlinks to total hyperlinks.

* **Relevance**: Phishing sites might include a high number of external links to redirect users to malicious websites.

1. **ratio\_nullHyperlinks**: Ratio of null (broken or undefined) hyperlinks to total hyperlinks.

* **Relevance**: A high ratio of broken links may indicate that the site is incomplete or fake.

1. **nb\_extCSS**: Number of external CSS files linked in the webpage.

* **Relevance**: External CSS files are often linked to legitimate sites for styling, but phishing sites may use them differently.

1. **ratio\_intRedirection**: Ratio of internal redirections to total redirections.

* **Relevance**: Phishing sites might use redirection techniques to send users to malicious pages.

1. **ratio\_extRedirection**: Ratio of external redirections to total redirections.

* **Relevance**: A higher ratio of external redirections could be indicative of phishing.

1. **ratio\_intErrors**: Ratio of internal errors to total errors.

* **Relevance**: A high number of internal errors can suggest that the site is poorly maintained, which is common for phishing sites.

1. **ratio\_extErrors**: Ratio of external errors to total errors.

* **Relevance**: External errors may indicate broken or fraudulent links, which are common in phishing attacks.

1. **login\_form**: Whether the website contains a login form.

* **Relevance**: Phishing sites often contain fake login forms to steal user credentials.

1. **external\_favicon**: Whether the website uses an external favicon (the small icon displayed in browser tabs).

* **Relevance**: Legitimate sites typically host their favicon files locally, while phishing sites may use external sources.

1. **links\_in\_tags**: Whether there are links in HTML tags (such as <a> tags).

* **Relevance**: Phishing sites often try to hide links inside tags or other elements to disguise the malicious content.

1. **submit\_email**: Whether the website contains an email submission form.

* **Relevance**: Phishing sites often ask for email addresses to collect personal information.

1. **ratio\_intMedia**: Ratio of internal media links (e.g., images, videos) to total media links.

* **Relevance**: A higher number of internal media links may indicate a more legitimate site.

1. **ratio\_extMedia**: Ratio of external media links to total media links.

* **Relevance**: A higher number of external media links could suggest that the website is not genuine.

1. **sfh**: Whether the website has a "single frame handler" (iframe) structure.

* **Relevance**: Phishing websites often use iframes to hide malicious content.

1. **iframe**: Whether the website uses an iframe (inline frame) to embed another website.

* **Relevance**: Phishing sites often use iframes to mask the true nature of the page.

1. **popup\_window**: Whether the website uses pop-up windows.

* **Relevance**: Pop-ups are common on phishing sites as a means to mislead or trick users.

1. **safe\_anchor**: Whether the website has safe anchor tags (e.g., links that use https://).

* **Relevance**: Unsafe anchors or links without proper HTTPS encryption are often seen in phishing sites.

1. **onmouseover**: Whether the website uses the onmouseover event (e.g., triggering actions when the user hovers over elements).

* **Relevance**: Phishing sites may use JavaScript events like onmouseover to manipulate user behavior.

1. **right\_clic**: Whether the website blocks the right-click function (e.g., to prevent users from inspecting the page).

* **Relevance**: Some phishing sites disable right-click to prevent users from copying URLs or inspecting elements.

1. **empty\_title**: Whether the webpage has an empty title tag.

* **Relevance**: Legitimate sites typically have meaningful title tags, while phishing sites may leave this blank.

1. **domain\_in\_title**: Whether the domain appears in the title of the webpage.

* **Relevance**: The presence of the domain name in the title can indicate a legitimate site, but phishing sites may also use this tactic.

1. **domain\_with\_copyright**: Whether the domain includes copyright symbols or text.

* **Relevance**: Phishing sites may mimic legitimate websites by including copyright symbols to appear more credible.

1. **whois\_registered\_domain**: Whether the domain is registered through WHOIS.

* **Relevance**: Phishing sites may have unregistered or recently registered domains, which are often suspicious.

1. **domain\_registration\_length**: Length of time the domain has been registered.

* **Relevance**: Legitimate sites typically have longer registration periods; phishing sites are often newly registered.

1. **domain\_age**: Age of the domain.

* **Relevance**: Older domains are often more trustworthy, while phishing sites tend to use newer domains.

1. **web\_traffic**: Amount of web traffic the site receives.

* **Relevance**: Legitimate sites typically have higher web traffic, while phishing sites have little or no traffic.

1. **dns\_record**: Whether the website has a valid DNS record.

* **Relevance**: A valid DNS record is usually a sign of a legitimate website, while phishing sites may lack DNS records.

1. **google\_index**: Whether the site is indexed by Google.

* **Relevance**: Legitimate sites are usually indexed by search engines like Google, while phishing sites are less likely to be indexed.

1. **page\_rank**: The site's ranking in search engines.

* **Relevance**: High page rank is generally associated with trustworthy, legitimate websites, while phishing sites often have low or no rank.
* **Link given below where you’ll find EDA, model fitting and conclusion**

[Projects/Code-B\_intern\_Phishing\_task.ipynb at main · SakshiGailwar/Projects](https://github.com/SakshiGailwar/Projects/blob/main/Code-B_intern_Phishing_task.ipynb)