# Phishing Website Detection Feature Extraction

This document outlines the rationale behind selecting specific features for analyzing URLs in the context of identifying phishing websites. By focusing on these features, we aim to distinguish between legitimate and phishing URLs effectively. The chosen features are derived from patterns observed in URL structures, domain information, and webpage attributes that are indicative of phishing attempts. This selection process enhances the accuracy and efficiency of our phishing detection model.

## Address Bar Based Features

| | |
|---|---|
| Domain of URL | Here, we are just extracting the domain present in the URL. This feature doesn't have much significance in the training. May even be dropped while training the model. |
| IP Address in URL | Checks for the presence of IP address in the URL. URLs may have IP address instead of domain name. If an IP address is used as an alternative of the domain name in the URL, we can be sure that someone is trying to steal personal information with this URL.If the domain part of URL has IP address, the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |
| "@" Symbol in URL | Checks for the presence of '@' symbol in the URL. Using “@” symbol in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.If the URL has '@' symbol, the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |
| Length of URL | Computes the length of the URL. Phishers can use long URL to hide the doubtful part in the address bar. In this project, if the length of the URL is greater than or equal 54 characters then the URL classified as phishing otherwise legitimate.If the length of URL >= 54 , the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |
| Depth of URL | Computes the depth of the URL. This feature calculates the number of sub pages in the given url based on the '/'.The value of feature is a numerical based on the URL. |
| Redirection "//" in URL | The existence of “//” within the URL path means that the user will be redirected to another website. We find that if the URL starts with “HTTP”, that means the “//” should appear in the sixth position. the "//" is anywhere in tURL apart from after the protocal, thee value assigned to this feature is 1 (phishing) or else 0 (legitimate). |

| | |
|---|---|
| "http/https" in Domain name | Checks for the presence of "http/https" in the domain part of the URL. The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users.If the URL has "http/https" in the domain part, the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |
| Using URL Shortening Services “TinyURL” | URL shortening is a method on the “World Wide Web” in which a URL may be made considerably smaller in length and still lead to the required webpage. This is accomplished by means of an “HTTP Redirect” on a domain name that is short, which links to the webpage that has a long URL.If the URL is using Shortening Services, the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |
| Prefix or Suffix "-" in Domain | Checking the presence of '-' in the domain part of URL. The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.If the URL has '-' symbol in the domain part of the URL, the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |

## HTML and JavaScript based Features:

| | |
|---|---|
| IFrame Redirection | IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the “frameBorder” attribute which causes the browser to render a visual delineation.If the iframe is empty or repsonse is not found then, the value assigned to this feature is 1 (phishing) or else 0 (legitimate). |
| Status Bar Customization | Phishers may use JavaScript to show a fake URL in the status bar to users. To extract this feature, we must dig-out the webpage source code, particularly the “onMouseOver” event, and check if it makes any changes on the status barIf the response is empty or onmouseover is found then, the value assigned to this feature is 1 (phishing) or else 0 (legitimate) |
| Disabling Right Click | Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. for this feature, we will search for event “event.button==2” in the webpage source code and check if the right click is disabled. |
| Website Forwarding | The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected. In our dataset, we find that legitimate websites have been redirected one time max. On the other hand, phishing websites containing this feature have been redirected at least 4 times. |