Sure! Below is a **short explanation of each column** in your dataset and how to obtain the data for a given URL:

◇ Basic URL Features

Column Name	Explanation	How to Get Data?
URL	The full website link	Direct input
LengthOfURL	Total characters in the URL	`len(url)`
Domain	Main domain name	Extract using `tldextract.extract(url).domain`
DomainLengthOfURL	Length of domain name only	`len(domain)`
TLD	Top-Level Domain (e.g., .com, .edu)	Extract using `tldextract.extract(url).suffix`
TLDLength	Length of the TLD	`len(tld)`

◇ Complexity & Special Characters

Column Name	Explanation	How to Get Data?
URLComplexity	Number of unique characters in the URL	`len(set(url))`
CharacterComplexity	Ratio of unique characters to total characters	`len(set(url)) / len(url)`
dot_count	Number of `.` in the URL	`url.count('.')`
hyphen_count	Number of - in the URL	`url.count('-')`
slash_count	Number of / in the URL	`url.count('/')`
digit_count	Number of digits (0-9) in URL	`sum(c.isdigit() for c in url)`
URLLetterRatio	Ratio of letters to total characters	`sum(c.isalpha() for c in url) / len(url)`
URLDigitRatio	Ratio of digits to total characters	`digit_count / len(url)`

EqualCharCntInURL	Count of = characters	`url.count('=')`
QuesMarkCntInURL	Count of ? in URL	`url.count('?')`
AmpCharCntInURL	Count of & in URL	`url.count('&')`
OtherSpclCharCntInURL	Count of other special characters (%, _, #, etc.)	`sum(url.count(c) for c in ['%', '_', '#', '!', '$'])`
URLOtherSpclCharRatio	Ratio of special characters in URL	`OtherSpclCharCntInURL / len(url)`

◇ Path & Query Analysis

Column Name	Explanation	How to Get Data?
NumberOfHashtags	Count of # in URL	`url.count('#')`
NumberOfSubdomains	Number of subdomains	`len(tldextract.extract(url).subdomain.split('.'))`
HavingPath	Whether the URL has a path (1 = Yes, 0 = No)	`1 if urlparse(url).path else 0`
PathLength	Length of the URL path after domain	`len(urlparse(url).path)`
HavingQuery	Whether the URL has a query string	`1 if urlparse(url).query else 0`
HavingFragment	Whether the URL has a fragment (`#...`)	`1 if urlparse(url).fragment else 0`
HavingAnchor	Whether the URL has an anchor link	`1 if '#' in url else 0`

◇ Security Features

Column Name	Explanation	How to Get Data?
HasSSL	Whether the URL uses HTTPS (1 = Yes, 0 = No)	`1 if url.startswith('https') else 0`

IsUnreachable	Whether the website is unreachable	Check response using `requests.get(url, timeout=5)`
IsDomainIP	Whether the domain is an IP address	`1 if domain.replace('.', '').isdigit() else 0`

◇ Website Content & HTML Analysis

Column Name	Explanation	How to Get Data?
LineOfCode	Number of lines in the page source	`len(response.text.split('\n'))`
LongestLineLength	Length of the longest line in source code	`max(len(line) for line in response.text.split('\n'))`
HasTitle	Whether the page has a `<title>` tag	`1 if '<title>' in response.text else 0`
HasFavicon	Whether the page has a favicon	Check `<link rel="icon">` in HTML
HasRobotsBlocked	Whether robots.txt is blocking access	Check `/robots.txt` response
IsResponsive	Whether the page is mobile-friendly	Check viewport meta tag in `<head>`
HasDescription	Whether meta description exists	Check `<meta name="description">`
HasPopup	Whether the page has popups	Search for JavaScript popup functions (`window.alert`, etc.)
HasIFrame	Whether the page contains `<iframe>` elements	`1 if '<iframe' in response.text else 0`

◇ Redirection & Social Engineering

Column Name	Explanation	How to Get Data?
IsURLRedirects	Whether the URL redirects to another page	Check `response.history` in `requests.get(url)`
IsSelfRedirects	Whether the page redirects to itself	Compare final URL with initial URL

IsFormSubmitExternal	Whether a form submits data to an external domain	Extract `action` attribute in `<form>`

◇ Banking, Payment, and Phishing Indicators

Column Name	Explanation	How to Get Data?
HasSocialMediaPage	Whether the site has social media links	Check for `facebook.com`, `twitter.com` in HTML
HasSubmitButton	Whether the page has a submit button	Search `<input type="submit">`
HasHiddenFields	Whether forms contain hidden input fields	Search `<input type="hidden">`
HasPasswordFields	Whether the page has a password field	Search `<input type="password">`
HasBankingKey	Whether banking-related words exist	Check for `bank`, `account`, `secure`, etc.
HasPaymentKey	Whether payment-related words exist	Check for `pay`, `checkout`, `credit card`, etc.
HasCryptoKey	Whether crypto-related words exist	Check for `bitcoin`, `crypto`, `wallet`, etc.
HasCopyrightInfoKey	Whether copyright information exists	Check for ©, `copyright`, etc.

◇ External Resources & Linking Behavior

Column Name	Explanation	How to Get Data?
CntImages	Number of images on the page	Count `<img>` tags
CntFilesCSS	Number of external CSS files	Count `<link rel="stylesheet">` tags
CntFilesJS	Number of JavaScript files	Count `<script>` tags with `src`
CntSelfHRef	Number of self-referencing links	Count `<a href>` with same domain
CntEmptyRef	Number of empty links (`href="#"`)	Count `<a href="#">`

CntExternalRef	Number of links pointing to external domains	Compare `href` domains to site domain
CntPopup	Number of JavaScript popups	Search for `window.alert`, `window.prompt`
CntIFrame	Number of iframe elements	Count `<iframe>` tags

◇ Advanced Security Metrics

Column Name	Explanation	How to Get Data?
UniqueFeatureCnt	Number of unique security features	Count nonzero security-related values
WAPLegitimate	Whether the site is a well-known web application	Compare against a list of trusted apps
WAPPhishing	Whether the site is a known phishing app	Compare against phishing database
ShannonEntropy	Measures randomness in URL (higher = more phishing)	Compute entropy formula
FractalDimension	Measures complexity of page structure	Compute fractal dimension of DOM tree
KolmogorovComplexity	Measures compressibility of HTML source	Compute using Lempel-Ziv complexity
HexPatternCnt	Number of hexadecimal patterns in URL	Count occurrences of `[0-9A-Fa-f]{4,}`
Base64PatternCnt	Number of Base64 encoded patterns	Count occurrences of `[A-Za-z0-9+/]{10,}`
LikelinessIndex	Probability of being phishing based on features	Model-derived score

◇ Label (Target)

Column Name	Explanation
Label	The actual classification (Legitimate/Phishing)

🚀 Next Steps

- Use `BeautifulSoup`, `requests`, `tldextract`, and `re` to extract most of these features automatically.
- Train your **LightGBM** and **CatBoost** models on this dataset.

This will **improve accuracy and reduce false positives** for login-based legitimate websites! 🚀