Team -DataVerse

PHISHIUNITO (PHISHING) URL DATASET

DATA SCIENCE MID SEM PROJECT SUBMISSION

Colab Link

Dataset Explanation

Dataset Composition:

- Total Instances: 235,795
- Legitimate URLs: 134,850
- Phishing URLs: 100,945
- Features: 54, derived from webpage source code and URL.

- Associated Task: Classification (Phishing vs. Legitimate URLs)
- Feature Types: Real, Categorical, Integer
- No Missing Values in the dataset

Key Features:

- URL Features: URL Length, Domain, TLD, Char Continuation Rate
- Webpage Content Features: Title, HasSocialNet, No Popup

Label: label (1 = Phishing, 0 = Legitimate)

Problem Statement and Importance

- Problem Statement: To develop a model that can accurately classify URLs as phishing or legitimate, helping to detect and prevent phishing attacks.
- Importance:
  - Phishing is a major cybersecurity threat, often used to steal sensitive data like passwords and financial information.
  - A robust URL-based detection model can act as a frontline defense, providing quick identification of malicious sites without needing to visit them.
- Application: Such a model could be implemented in browsers, email filters, and security systems to protect users from phishing threats.

Exploratory Data Analyses

- Total Features : 54
- Missing Values : NA
- Data Type : Object (5) / int64 (41) / float54(10)

01 Percentage of unique/total instances in each feature

	count	unique	Distinct %	is_categorical
FILENAME	235795	235795	100.0	NA
URL	235795	235370	99.819759	NA
URLLength	235795.0	482	0.204415	NA
Domain	235795	220086	93.337857	NA
DomainLength	235795.0	101	0.042834	NA
IsDomainIP	235795.0	2	0.000848	Categorical
TLD	235795	695	0.294748	NA
URLSimilarityIndex	235795.0	36360	15.420174	NA
CharContinuationRate	235795.0	898	0.380839	NA
TLDLegitimateProb	235795.0	465	0.197205	NA
URLCharProb	235795.0	227421	96.44861	NA
TLDLength	235795.0	12	0.005089	NA
NoOfSubDomain	235795.0	10	0.004241	NA
HasObfuscation	235795.0	2	0.000848	Categorical
NoOfObfuscatedChar	235795.0	20	0.008482	NA
ObfuscationRatio	235795.0	146	0.061918	NA
NoOfLettersInURL	235795.0	421	0.178545	NA
LetterRatioInURL	235795.0	709	0.300685	NA

We get 5 features that have more than 82% distinct features indicating that they do not contribute much to the target.

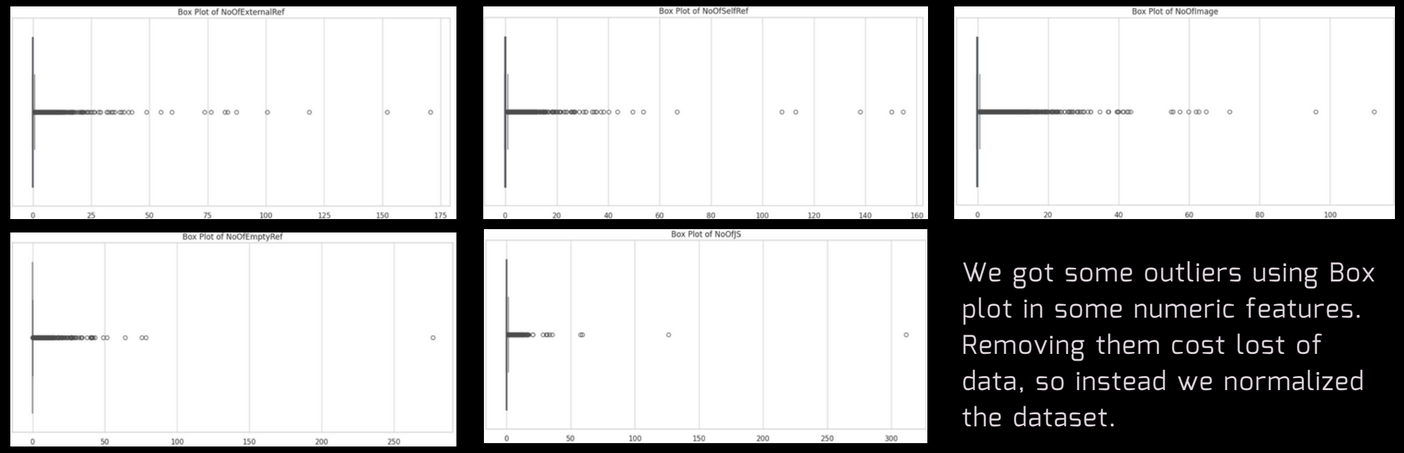
FileName :100%
URL :99.819%
URLCharProb:96.4%
Domain : 93.33 %
Title : 83.91 %

02 Categorical features = 21 Unique instances <=5

'IsDomainIP', 'HasObfuscation', 'NoOfQMarkInURL', 'IsHTTPS', 'HasTitle', 'HasFavicon', 'Robots', 'IsResponsive', 'NoOfURLRedirect', 'NoOfSelfRedirect', 'HasDescription', 'HasExternalFormSubmit', 'HasSocialNet', 'HasSubmitButton','HasHiddenFields', 'HasPasswordField', 'Bank', 'Pay', 'Crypto', 'HasCopyrightInfo', 'label'

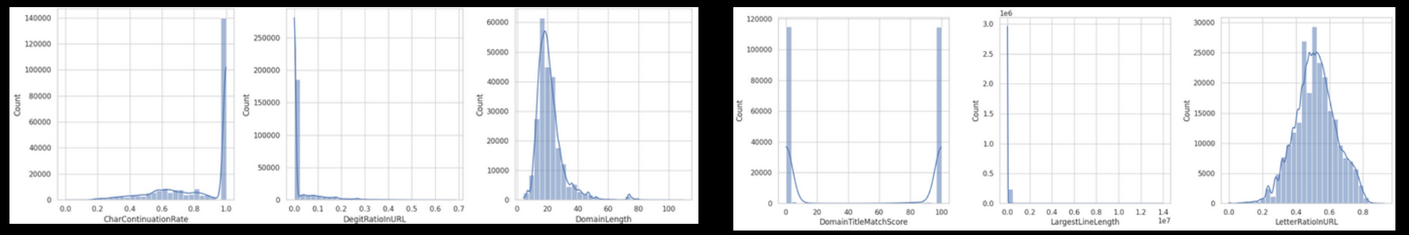
03 Outlier Detection

Box plot on all the numerical features



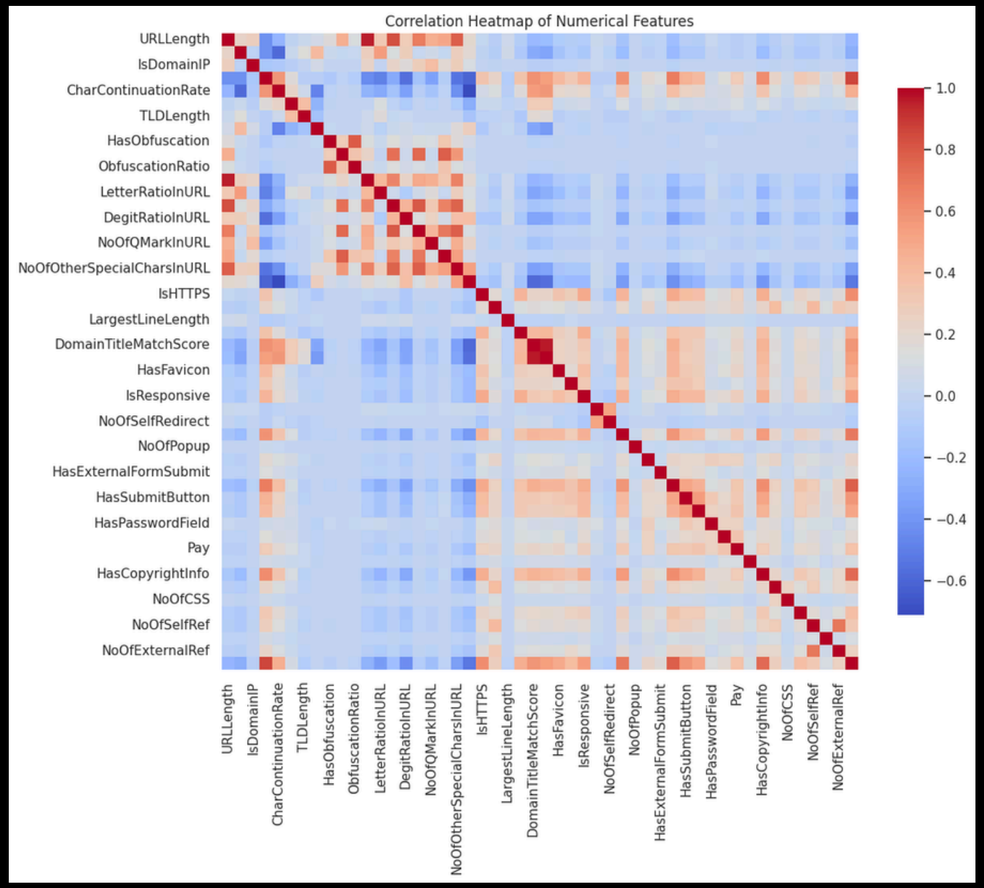
We got some outliers using Box plot in some numeric features. Removing them cost lost of data, so instead we normalized the dataset.

04 Data Distribution using histogram



This helped in understanding data distribution and its skewness . Moreover we get categorical features too after visualizing this .

05 Heatmap



	URLLength	DomainLength	IsDomainIP	URLSimilarityIndex	CharContinuationRate	TLDLegitimateProb	TLDLength	NoOfSubDomain
URLLength	1.000000	0.243091	0.277272	-0.421104	-0.241104	-0.016656	0.033339	0.071058
DomainLength	0.243091	1.000000	-0.041398	-0.420694	-0.577425	0.027209	0.152669	0.399157
IsDomainIP	0.277272	-0.041398	1.000000	-0.121439	-0.101307	-0.053220	-0.011125	0.084619
RLSimilarityIndex	-0.421104	-0.420694	-0.121439	1.000000	0.572938	0.162042	-0.053915	-0.058481
arContinuationRate	-0.241104	-0.577425	-0.101307	0.572938	1.000000	0.237221	0.051438	-0.482314
LDLegitimateProb	-0.016656	0.027209	-0.053220	0.162042	0.237221	1.000000	0.386794	-0.250048
TLDLength	0.033339	0.152669	-0.011125	-0.053915	0.051438	0.386794	1.000000	-0.162431
NoOfSubDomain	0.071058	0.399157	0.084619	-0.058481	-0.482314	-0.250048	-0.162431	1.000000

Strong Positive Correlations:

- **URLLength** with **LargestLineLength** and **NoOfQMarkInURL** suggesting that longer URLs may have more special characters and a loner line length.
- **NoOfSelfRedirect** with **NoOfPopup**, indicating that sites with self-redirects may also use pop-ups, possibly a characteristic of phishing sites.

Strong Negative Correlations:

- **IsDomainIP** with **IsHTTPS** and **HasPasswordField**, indicating that URLs using IP addresses instead of domain names are less likely to use HTTPS or display favicons.
- **HasObfuscation** with **IsHTTPS** and **HasPasswordField** which could suggest that obfuscated URLs are less likely to employ HTTPS and include password fields, perhaps indicative of phishing tendencies.

Other notable Correlations:

- **DomainTitleMatchScore** and **IsResponsive** have a positive correlation, which may imply that legitimate URLs with a matching title are often responsive, contrasting with some phishing sites that may ignore responsiveness.
- **Pay** and **HasPasswordField** are positively correlated, as payment fields are often found alongside password fields.

Data Preprocessing

1. Encoding categorical features

There are binary and nominal categorical features for example IsDomainIP, TLD. Since binary categorical features are already encoded, we shall apply Label Encoding only on TLD column

2. Feature Scaling

Firstly we find all the numerical continuous features excluding categorical datas. To those features we apply Standard Scaler function which standardizes those continuous data.

3. Handling Imbalanced Data

As the target feature, ie, 'label' might be imbalanced , we use Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes.

Hypothesis Testing on URL Characteristics

01 URL Length Hypothesis

- Null Hypothesis (H_0): There is no significant difference in the average URL length between phishing URLs and legitimate URLs.
- Alternative Hypothesis (H_1): The average URL length differs significantly between phishing and legitimate URLs.

02 HTTPS Usage Hypothesis

- Null Hypothesis (H_0): The proportion of phishing URLs that use HTTPS is equal to the proportion of legitimate URLs that use HTTPS.
- Alternative Hypothesis (H_1): The proportion of phishing URLs using HTTPS differs from that of legitimate URLs

03 Number of Subdomains Hypothesis

- Null Hypothesis (H_0): There is no significant difference in the average number of subdomains between phishing URLs and legitimate URLs.
- Alternative Hypothesis (H_1): The average number of subdomains is significantly different between phishing and legitimate URLs.

Experiments Conducted to Validate Hypothesis Tests

01

URL Length Hypothesis

- Test Used: Independent t-test
- Result: Significant difference (t-statistic: -101.05, p-value: 0.0)
- Conclusion: Significant difference in URL length between phishing and legitimate URLs.

02

HTTPS Usage Hypothesis

- Test Used: Chi-square test
- Result: Significant difference (χ^2-statistic: 87,486.79, p-value: 0.0)
- Conclusion: Significant difference in HTTPS usage between phishing and legitimate URLs.

03

Number of Subdomains Hypothesis

- Test Used: Independent t-test
- Result: Significant difference (t-statistic: -2.66, p-value: 0.0079)

Conclusion: Significant difference in the average number of subdomains between phishing and legitimate URLs.

Summary of Hypothesis Testing Insights

- Significant Features Identified:
  - URL Length: Phishing URLs have different lengths compared to legitimate URLs.
  - HTTPS Usage: HTTPS is used differently between phishing and legitimate URLs.
  - Number of Subdomains: Phishing URLs tend to have fewer subdomains compared to legitimate URLs.

Overall Conclusion: These features provide strong discriminatory power for distinguishing phishing URLs from legitimate ones, supporting the use of these features in a classification model.

Thank You