Phishing Website Data Analysis: Identifying Key Patterns and Trends in Cybersecurity Threats

By Aditi Gupta, Devashree Palav, Uday Rawat, Anusha Sarkar

Objective

To explore and analyze pattern in phishing websites and build a Threat Score model using EDA, hypothesis testing, and multiple linear Regression.

Dataset

PhiUSIIL Phishing URL Dataset from UCI ML Repository.
- 235,795 URLs | 56 features
- Label: 1 = Legitimate, 0 = Phishing
- No missing values

Methodology

DATA PREPROCESSING
- Dropped irrelevant columns, encoded TLD, removed 120 duplicates, scaled features.

EDA:
- Explored stats and visual patterns. Phishing URLs showed suspicious traits like length, special characters, and lack of social/media tags.

Correlation & Feature Selection:
- Selected top 40 features based on correlation with label. Removed weak predictors to reduce noise.

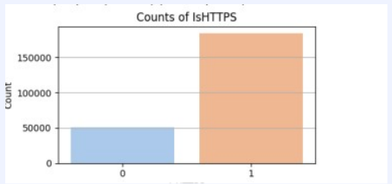
Train-Test Split:
- Stratified 80:20 split to maintain label proportions.

MLR Modeling:
- Built regression model using 14 top features to predict URLSimilarityIndex.

Findings

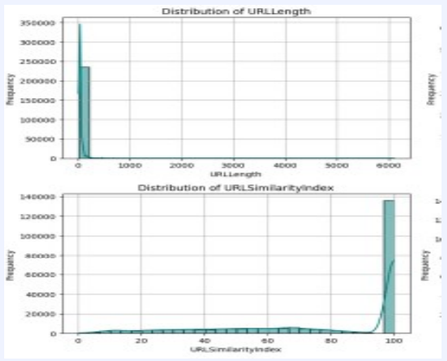
- URLSimilarityIndex is the most indicative feature (corr = 0.86).
- Phishing URLs are longer, unstructured, and use more special characters.
- They often lack favicons, metadata, and social media links.
- Mild class imbalance: 57% phishing, 43% legitimate

Analysis of Counts



The majority of URLs in the dataset do not use HTTPS, indicating a potential security risk. Legitimate sites typically use HTTPS, while phishing sites often do not.

Features Distribution



The distribution of various feature (eg.URLlength,URLSimilarity Index)reveal that many phishing website tend to have specific characteristics such as shorter length or unusual character distribution which can be used for detection

Modelling-Multiple Linear Regeression

Objective:
- Predict a numeric phishing threat score using top correlated features.

Target Variable:
- URLSimilarityIndex — represents similarity to safe URL patterns.

Approach:
- Applied Multiple Linear Regression on 14 most predictive features.

Performance:
- Achieved $R^2 = 0.725$, indicating strong model fit.

Significance:
- All predictors were statistically significant ($p < 0.001$), confirming reliable contribution to threat score prediction.

OLS Regression Table

Feature	Coefficient	Impact
DomainTitleMatchScore	+15.63	**Very strong positive** — higher match implies more legitimate site.
URLCharProb	+6.60	Strong positive — URLs with common character patterns seem safer.
HasSocialNet	+6.13	Presence of social media links increases legitimacy score.
CharContinuationRate	+5.23	Longer consistent strings in the URL are positively associated.
HasCopyrightInfo	+3.64	Boosts perceived trust.
HasDescription	+3.30	A meta description increases legitimacy.
HasFavicon	+1.26	Trusted websites often use favicons.
HasTitle	+1.22	Websites with a proper title appear more legitimate.
IsResponsive	+0.79	Positive influence — mobile compatibility matters.
HasSubmitButton	+0.30	Mild positive impact.
SpacialCharRatioInURL	−3.12	**Negative impact** — excessive symbols in URLs are suspicious.
URLTitleMatchScore	−12.87	Strong negative — low match between URL and page title is a phishing indicator.
IsHTTPS	−0.19	Surprisingly, slight negative. This could indicate that HTTPS is no longer a reliable differentiator.
HasHiddenFields	+0.71	Mild positive, possibly context-dependent.

Conclusion

Conclusion & Takeaways
- Phishing sites use obfuscated URLs and lack metadata or social links.
- Legitimate sites show structured content and title-URL consistency.
- Key indicators: domain-title similarity, special character patterns.
- Threat score from MLR helps in automated risk detection.

Future Scope

The MLR model, which predicts a **phishing "threat score"**, can be integrated into an actual system that could:

1.Monitor websites or links in real time
2.Apply your model to assign a threat score
3.Alert users if a link seems suspicious or risky

Reference

- PhiUSIIL Dataset – UCI Repository
- Research inspired by phishing pattern detection and ML modeling