PhishSense (Phishing Analysis System) Report

Team Members:

SUDIPTA GHOSH

DEEPANSH SOOD

SHAMBO SAMANTA

SOURAJIT BHAR

Abstract

Phishing attacks are a growing cybersecurity threat, using deceptive websites to steal sensitive information from users. This project leverages machine learning to develop an effective phishing URL detection system, analyzing URL structures and webpage features to classify URLs as phishing or legitimate. A dataset of 235,795 URLs with 50 detailed features was used for training and testing four machine learning models: Decision Tree, Random Forest, SVM, and Naive Bayes. Exploratory data analysis (EDA) was conducted to uncover feature patterns and relationships, guiding model selection. Random Forest emerged as the best-performing model, achieving 100% accuracy with robust performance across all metrics. The system demonstrates scalability and reliability for real-world applications, with plans for deployment and integration into cybersecurity frameworks.

Introduction

Phishing attacks exploit users by creating websites that imitate trusted entities to harvest sensitive information, such as passwords and financial details. These attacks pose a severe threat to individuals and organizations, with significant financial and privacy implications.

The goal of this project is to develop a machine learning-based phishing detection system. By analyzing characteristics of URLs and webpage features, the system identifies phishing attempts with high accuracy.

The project emphasizes feature analysis, model training, and evaluation to build a reliable, scalable system. The outcome demonstrates how machine learning can strengthen cybersecurity defences against phishing attacks.

Materials and Methods

1. Dataset Overview: The dataset used for this project contains a comprehensive collection of URLs and their corresponding webpage characteristics to analyze and detect phishing attempts. It is publicly available and taken from PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository. The dataset contains 235,795 rows, with 134,850 legitimate URLs and 100,945 phishing URLs, spanning 55 columns. It includes a mix of numerical, categorical, and textual fields to support comprehensive analysis.

2. Key Features: The dataset includes the following key features:

(1) URL Characteristics: URL, URLLength, Domain, DomainLength, NoOfSubDomain, etc.

(2) Legitimacy Indicators: IsDomainIP, TLDLegitimateProb, URLSimilarityIndex, etc.

(3) Content and Behavioral Features: HasTitle, HasDescription, NoOfPopup, NoOfRedirects, etc.

(4) HTTPS and Metadata: IsHTTPS, Robots, HasCopyrightInfo, etc.

(5) Other Indicators: Social network links (HasSocialNet), security features (HasPasswordField), and obfuscation metrics (HasObfuscation).

3. Quality of Data: The dataset is complete with no missing values or duplicate records.

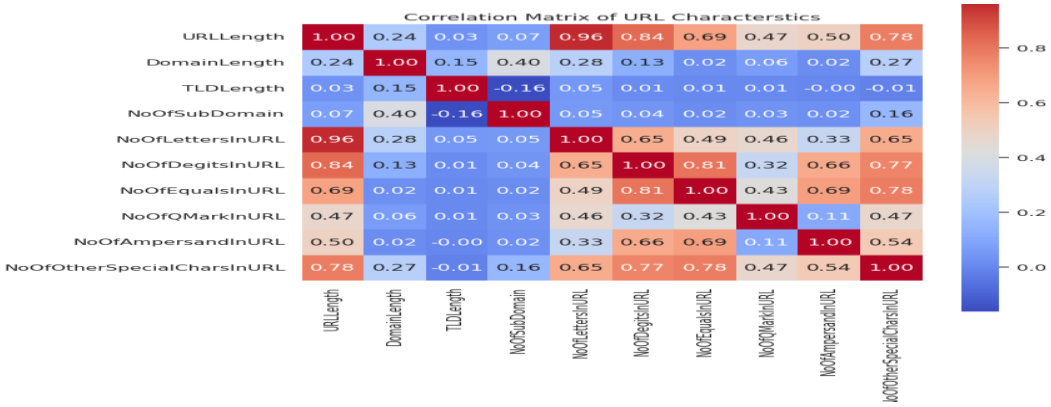
4. Relevance to Project Goals: The dataset provides diverse and high-quality features that enable effective phishing detection. It combines characteristics of the URL structure with page content and behavioral indicators, ensuring a well-rounded approach to model training.

5. Exploratory Data Analysis (EDA)

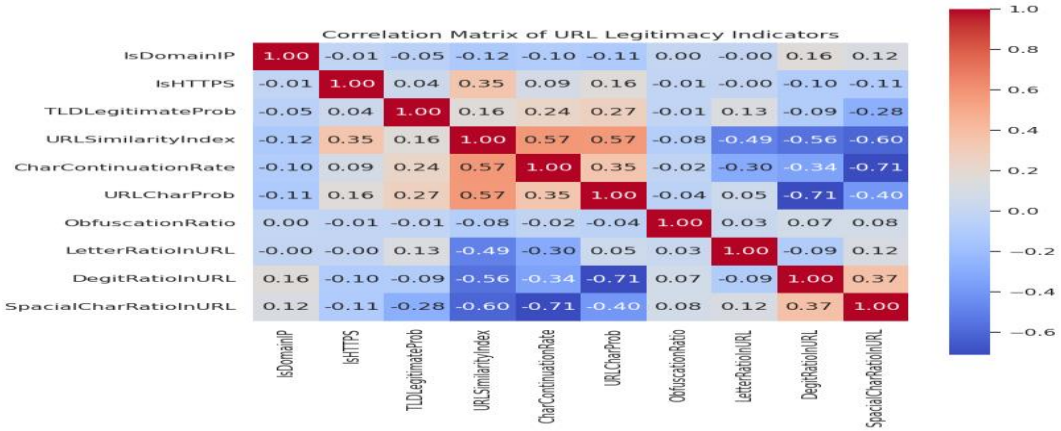
Exploratory Data Analysis (EDA) was conducted to understand the dataset's structure, identify patterns, and inform feature selection for model training. The analysis focused on identifying key characteristics of phishing and legitimate URLs, leveraging statistical summaries and visualizations.

1.Potential groupings of the most relevant numeric columns

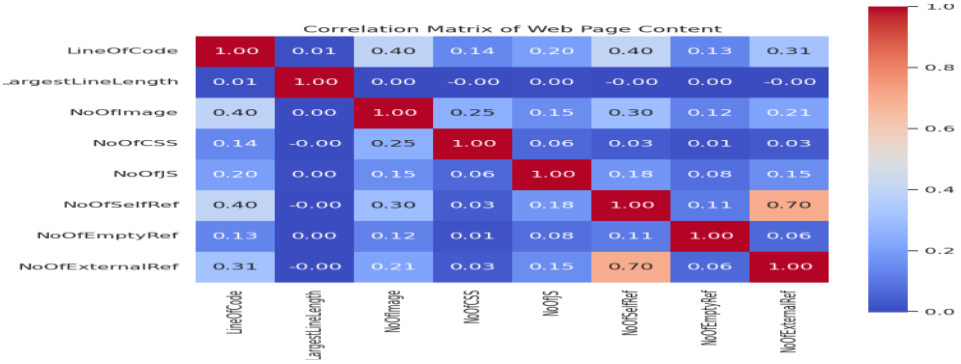
(1) URL Characteristics: This table displays the correlation coefficients between structural properties like URLLength, DomainLength, TLDLength, and NoOfSubDomain, helping to identify relationships among URL features for analyzing phishing patterns.



(2) URL Legitimacy Indicators: It helps to identify the strength and direction of relationships among indicators assessing URL legitimacy, providing insights into how various characteristics may influence the likelihood of a URL being legitimate or malicious.

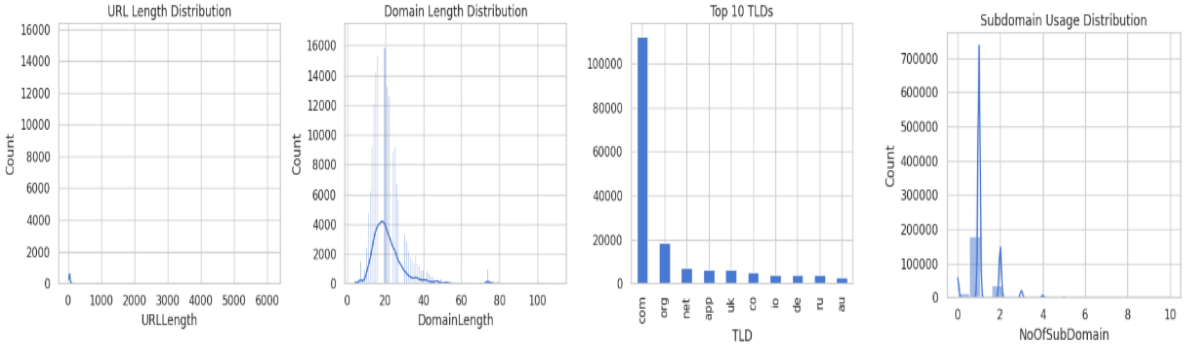


(3) Web Content Features: This matrix helps identify relationships among webpage content and structure characteristics, offering insights into usability, performance, and security risks, and guiding design decisions while highlighting potential security vulnerabilities.



2. Key Analysis and Findings

(1) URL Structure Analysis: URL Structure Analysis examines elements like URL length, domain length, TLD distribution, subdomain usage, and IP addresses in domains to identify patterns in phishing URLs. Phishing sites often manipulate these features, using longer URLs with additional subdomains or obfuscation to evade detection.



(2) Top-Level Domains (TLDs): the distribution of TLDs in URLs, highlighting common ones like .com, .org, and .uk, which may suggest certain TLDs are more associated with phishing

```
TLD
com        112554
org         18793
net          7097
app          6508
uk           6395
           ...
47              1
sharp           1
faith           1
243             1
ne              1
Name: count, Length: 695, dtype: int64
```

3. Hypotheses:

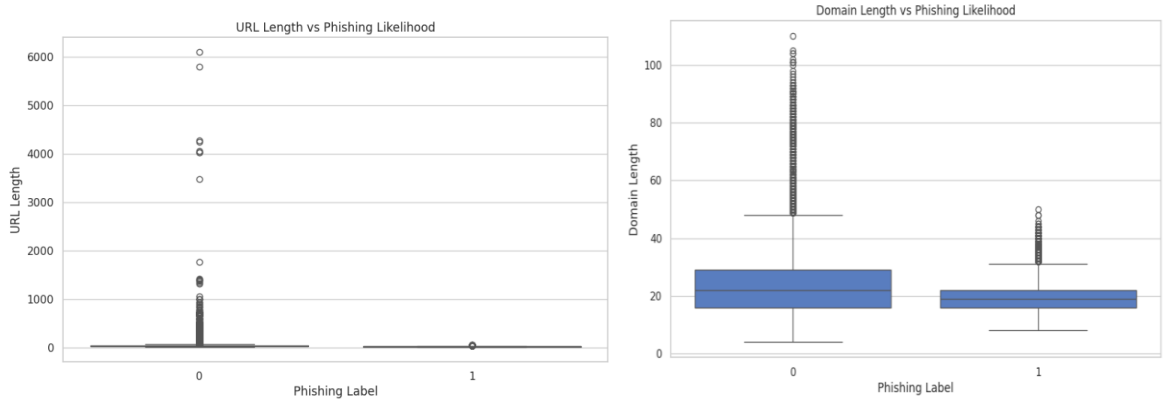
The following hypotheses were explored during the EDA phase:

(1) URL Length and Phishing Likelihood Hypothesis: Longer URLs may increase the likelihood of phishing attempts, as they can obscure the true destination and include misleading information.

Observation: Phishing URLs are typically longer, hiding malicious domains within a sea of characters, whereas shorter URLs are more common among legitimate sites, making URL length a valuable indicator for phishing detection.

(2) Domain Characteristics and Legitimacy Hypothesis: Legitimate domains are shorter and more recognizable.

Observation: Confirmed, with legitimate domains having fewer subdomains; shorter and simpler domain names are more likely to be perceived as trustworthy, while longer, complex ones may raise suspicion and signal potential phishing attempts.

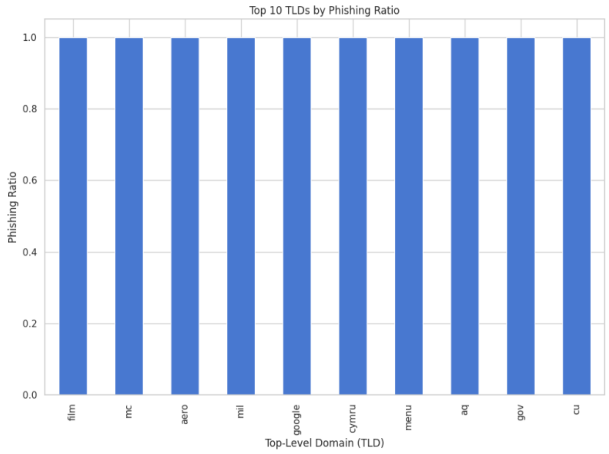


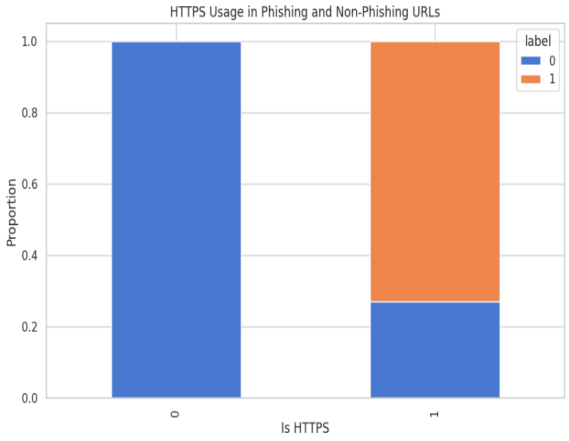
(3) Top-Level Domain (TLD) Influence, HTTPS Usage Hypothesis: Certain TLDs are disproportionately associated with phishing.

Observation: With TLDs like .film and .tk being more frequently linked to phishing URLs, while more established TLDs like .com and .net show lower phishing ratios.

(4) Obfuscation Techniques Hypothesis: Phishing URLs employ obfuscation more often than legitimate URLs.

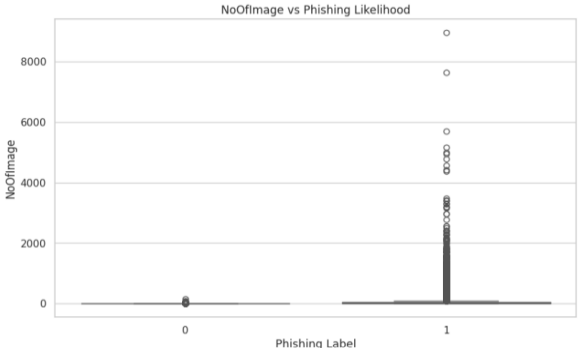
Observation: With significantly higher obfuscation metrics for phishing, as phishing URLs often use tactics like character substitution and misleading subdomains to disguise their malicious intent.

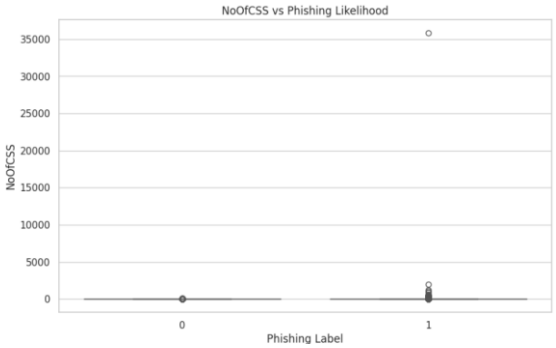


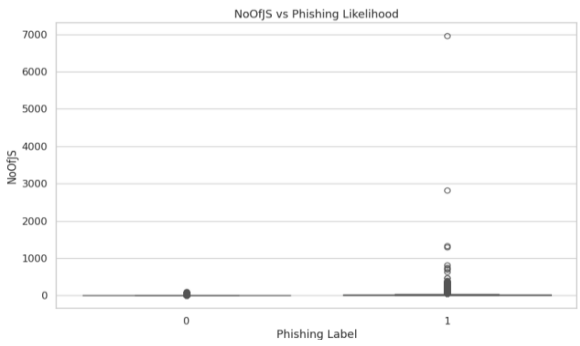


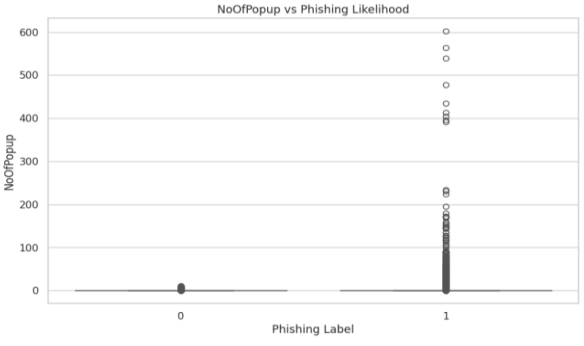
(5) Content Features, popup, social network link related Hypothesis: Phishing websites typically have fewer images, CSS files, and JavaScript, use deceptive or missing social network links, and employ minimal popups and redirects to obscure their true intent.

Observation: Phishing sites were found to have minimal content, misleading or absent social media links, and limited popups and redirects, while legitimate websites feature a balanced content mix, authentic social media links, and controlled popups/redirects for a more user-friendly experience.









6. Model Training

To build a reliable phishing URL detection system, we experimented with four machine learning algorithms: Decision Tree, Random Forest, Support Vector Machine (SVM), and Naive Bayes. This section outlines the steps taken for model development, evaluation, and selection.

1. Feature Encoding:

Categorical features like TLD and IsDomainIP were encoded using Label Encoding to make them suitable for the machine learning models.

2. Data Splitting: The dataset was split into 80% training data and 20% testing data to ensure robust model evaluation. The target variable label was set to 0 for phishing URLs and 1 for legitimate URLs.

3. Model Development

We trained the following machine learning models using the pre-processed dataset:

Models Chosen: Random Forest, Decision Tree, SVM, and Naive Bayes for their classification capabilities.

Random Forest: Chosen for its ability to combine multiple decision trees, improving accuracy through ensemble learning. It is robust against overfitting and performs well with large datasets and high-dimensional features.

Decision Tree: Chosen for its simplicity and interpretability, allowing for clear visualization of decision paths. It works well with both categorical and numerical data, making it suitable for initial analysis and feature importance evaluation.

SVM: Chosen for its effectiveness in high-dimensional spaces and robustness to overfitting.

Naive Bayes: Selected for its simplicity, efficiency, and strong performance with categorical data.

4. Hyperparameter Tuning and Cross-Validation: GridSearchCV was used to optimize key parameters (e.g., n_estimators, max_depth, min_samples_split) for each model. Cross-validation was applied to evaluate model performance across different subsets of the training data, ensuring robustness and preventing overfitting. The best hyperparameters identified for each model are as follows:

Random Forest Classifier: n_estimators=200,max_depth=20, min_samples_split=5 , min_samples_leaf=1

Decision Tree Classifier: criterion='gini', max_depth=5, min_samples_leaf=1, min_samples_split=2

Support Vector Classifier (SVC): C=0.1, kernel='linear', gamma='scale'

Naive Bayes: var_smoothing=1e-09

5. Performance Metrics

The models were evaluated using key metrics: **Accuracy** (correct classifications), **Precision** (true positives among predicted positives), **Recall** (true positives among actual positives), F1-Score (balance between precision and recall), **Confusion Matrix** (true/false positives and negatives), and the AUC-ROC Curve (model's ability to discriminate between classes). These metrics provide a comprehensive assessment of model performance.

Feature importance Analysis:

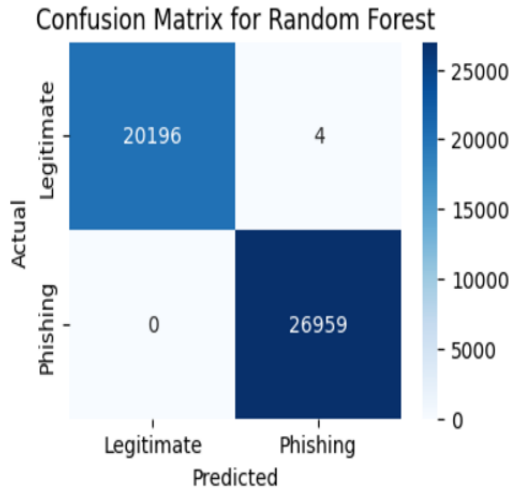
Top 10 Features by Importance:

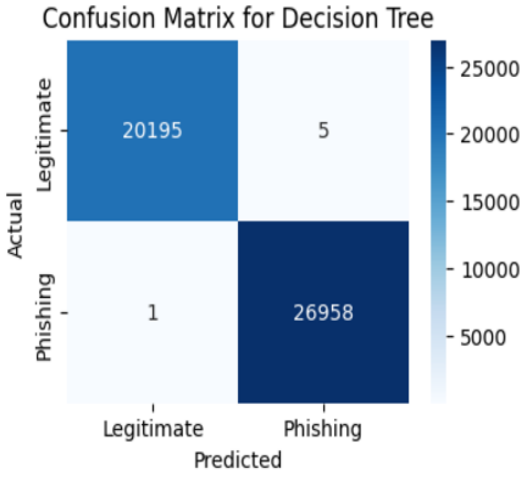
	Feature	Importance
5	URLSimilarityIndex	0.213593
24	LineOfCode	0.115887
51	NoOfExternalRef	0.111005
46	NoOfImage	0.092498
49	NoOfSelfRef	0.076345
48	NoOfJS	0.062096
45	HasCopyrightInfo	0.053735
38	HasSocialNet	0.049124
1	Domain	0.048771
47	NoOfCSS	0.032345

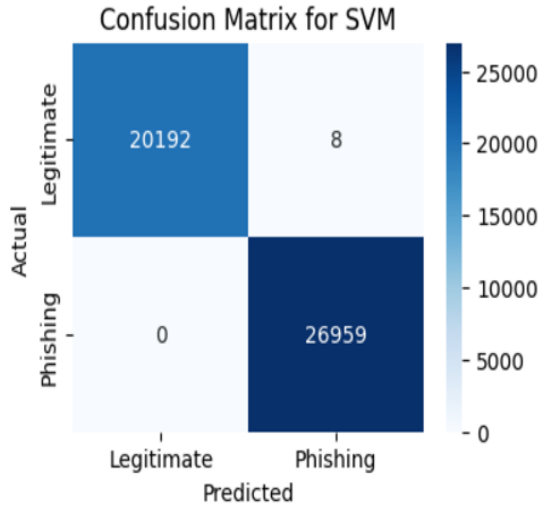
Performance & Evaluation

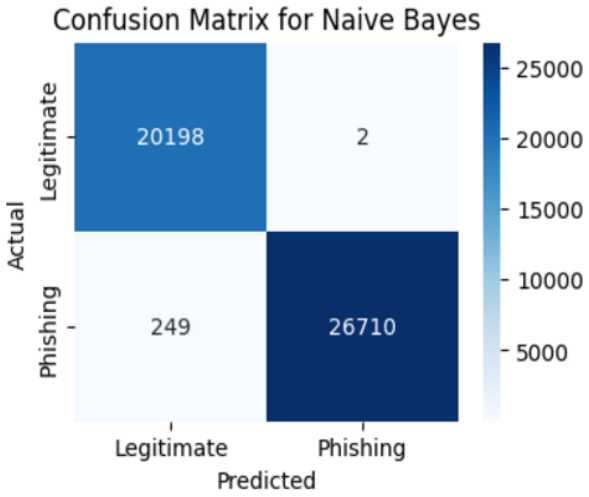
The performance of each model is summarized below:

	Model	Accuracy	Precision	Recall	F1-Score
0	Random Forest	0.999915	0.999852	1.000000	0.999926
1	Decision Tree	0.999873	0.999815	0.999963	0.999889
2	SVM	0.999830	0.999703	1.000000	0.999852
3	Naive Bayes	0.994678	0.999925	0.990764	0.995323









Key Observations

(1) Random Forest outperforms all other models with the highest accuracy (99.99%), precision (99.99%), recall (100%), and F1-score (99.99%), demonstrating its superior ability to classify phishing URLs accurately.

(2) Decision Tree and SVM perform well, with minor differences in metrics, but slightly lag behind Random Forest in recall.

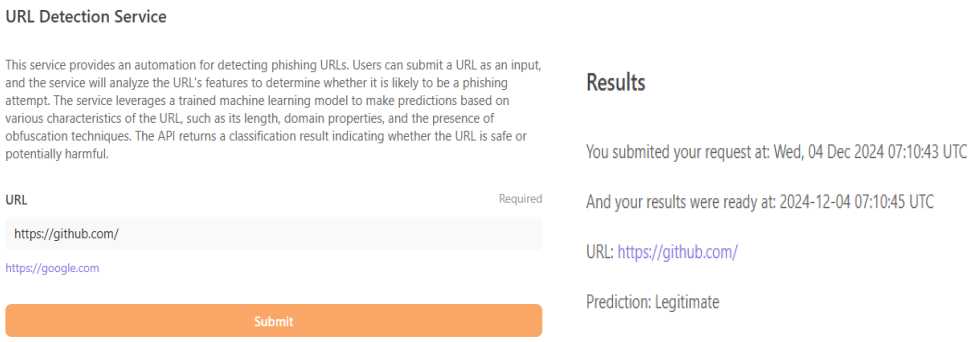
(3) Naive Bayes shows a lower overall performance, particularly in recall, indicating it struggles more with identifying phishing URLs accurately. The Confusion Matrices reveal minimal misclassification across all models, with Random Forest achieving zero false negatives, highlighting its robustness in phishing detection.

Best Model Selection

The models evaluated - Random Forest, Decision Tree, SVM, and Naive Bayes - demonstrated high performance, but **Random Forest** was selected as the best model. It achieved the highest accuracy, precision, recall, and F1-score, and its ensemble nature helps it perform better against overfitting, making it a robust choice for phishing URL detection. Random Forest is also more reliable when handling unseen data, ensuring stable performance in real-world applications.

Deployment:

We have developed a system capable of predicting whether a URL is phishing or legitimate. While it demonstrates promising accuracy, future updates and enhancements will be essential to improve its robustness, accuracy, and adaptability to evolving phishing tactics.



Future Work:

(1) Real-Time Phishing Detection: Develop and deploy real-time systems, such as browser extensions or exchange platforms, to detect and block phishing URLs as they appear, providing immediate protection for users.

(2) Model Retraining and Updates: Continuously update and retrain the model with new phishing data to adapt to evolving tactics and improve detection accuracy.

(3) Advanced Feature Integration: Incorporate additional features such as user interaction data and behavioral analysis to improve model robustness and detect more sophisticated phishing attacks.

Conclusion

This project successfully implemented a machine learning-based approach to detect phishing URLs, leveraging various URL features such as length, domain structure, TLD, and obfuscation techniques. Among the tested models, Random Forest demonstrated superior performance in terms of accuracy, precision, recall, and F1-score, making it the most reliable for phishing detection. Future work will involve integrating the system into cybersecurity frameworks and evaluating its performance in real-time scenarios.

References

GitHub Repository URL: CySentinels/DA-204o: Project repository for DA 204o Data Science in Practice (Aug semester 2024) @ IISc BLR

Demo URL - Input URL

Dataset URL: PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository

https://deepnote.com/

Project Jupyter | Home

Phishing.Database/phishing-links-ACTIVE-today.txt at master · mitchellkrogza/Phishing.Database