DA 204o: **Data Science in Practice**

Course Project Presentation

PhishSense (Phishing Analysis System using Machine Learning Algorithms)

SUDIPTA GHOSH, IISc, sudiptag@iisc.ac.in
DEEPANSH SOOD, IISc, deepanshsood@iisc.ac.in
SHAMBO SAMANTA, IISc, shambos@iisc.ac.in
SOURAJIT BHAR, IISc, sourajitbhar@iisc.ac.in

Data Science Canvas

Project: Data Science

Team: Sudipta, Deepansh, Shambo, Sourajit

Problem Statement

Business Case & Value Added
Which business case should be analyzed and what added value does it generate?

Phishing analysis system that will help to add value in below –

1. Reduction In fraud losses
2. Enhance Data security
3. Improve customer trust
4. User education & awareness

Data Landscape
Which data is required for this and which is already available? Which additional data has to be collected?

list of phishing urls & legitimate urls and other details.

https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset

Model Selection
Which analysis methods can be considered on the basis of the specific data landscape and the business case?

For phishing detection, machine learning models like Random Forest, Decision Tree, SVM, and Naive Bayes are suitable due to their ability to process structured data effectively. Random Forest ensures robustness and minimizes overfitting, while Decision Tree offers interpretability for business insights. SVM excels in handling high-dimensional data, and Naive Bayes is computationally efficient for rapid predictions. These models align with the goal of distinguishing phishing from legitimate URLs accurately.

Model Requirements
Which model requirements must be complied with in order to obtain a valid model?

1. Appropriate algorithm selections
2. Training data
3. Validation techniques
4. Hyperparameter tunning
5. Accuracy, precision, Recall,F1 score,

Software & Libraries
Which software should be used? Is there already a standard solution? Which libraries are used?
Programming Language: Python
Libraries:
Data Manipulation: pandas, NumPy, Matplotlib,Seaborn
Machine Learning: scikit-learn for traditional ML algorithms

Skills
What skills are needed to provide the data and model development?

1.Data cleaning, handling, normalization techniques.
2.feature engineering
3.Understanding of machine learning algorithms, model training, hyperparameter tuning.
4.Performance metrics, cross-validation, model validation techniques.
5.Proficiency in Python, understanding of object-oriented programming
6.Understanding of phishing techniques, cybersecurity principles
7.Team collaboration, effective communication, documentation

Execution & Evaluation

Model Evaluation
Which indicators require quality control and validation and how should they be interpreted? Is real-time monitoring necessary?

Accuracy: Overall correctness of the model.
Precision: Proportion of true positive predictions among all positive predictions.
Recall: Proportion of true positive predictions among all actual positives.
F1 Score: Harmonic mean of precision and recall.

Data Storytelling
What requirements does the target group have for the presentation of the results and how do I effectively communicate this data?

The target group requires clear and simple presentations, using straightforward language and avoiding jargon. The results should be relevant to the business case, providing context and highlighting key findings related to business objectives. Actionable insights are necessary to inform decision-making, with clear recommendations based on data analysis. Detailed performance metrics, including accuracy, are crucial for evaluating the model's effectiveness.

Data Collection & Preparation

Data Selection & Cleansing
Which of the available data is relevant? Do the data have to be cleaned up?
We collected URLs from a public repository containing phishing and legitimate links, and through web scraping, we automatically extracted relevant data such as domains, URL structures, and protocol types (HTTP/HTTPS), without needing additional cleaning.

Data Collection
How and with which methods should additionally required data be collected? What properties has this data to fulfil?
To collect additional data, we could use web scraping to extract real-time, accurate, and relevant information from websites, focusing on attributes like domain age, SSL certificate validity, URL length, suspicious keywords, and redirection behaviors for phishing detection.

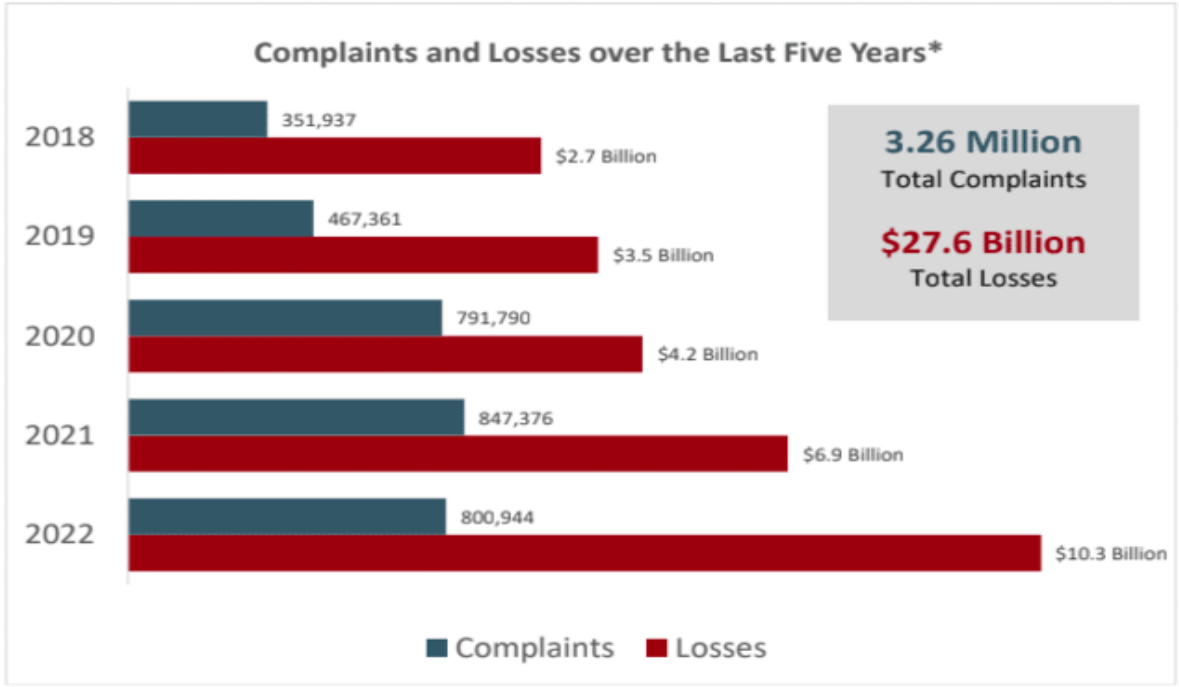
Data Integration
In which system should the data from different sources be migrated?
The data is generated from a custom-designed script that outputs in the desired format, which is then stored in a structured database or file system for easy analysis and model training.

Explorative Data Analysis
Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.
EDA involves identifying outliers, anomalies, and patterns in the dataset, such as URL length and suspicious keywords, to understand data characteristics and determine preprocessing needs.

Adopted from: https://github.com/tomalytics/datasciencecanvas

Background of the problem

Phishing attacks have become one of the most common and dangerous cybersecurity threats in recent years. As internet usage increases and more sensitive information is shared online, cybercriminals exploit vulnerabilities in digital communication to deceive users. These attacks often involve fraudulent emails, websites, or links designed to mimic legitimate entities, tricking individuals into sharing personal or financial information. Phishing poses a significant risk to businesses, government organizations, and individuals, leading to financial losses, identity theft, and compromised security systems. Despite advancements in cybersecurity, phishing remains a persistent problem, largely due to the increasing sophistication of tactics used by attackers.



Why it is important?

- Rising frequency of phishing attacks globally.
- Prevention helps avoid financial and reputational damage.
- Phishing attacks are becoming more sophisticated.
- Protects sensitive personal and corporate information.
- Helps maintain and build user trust in online platforms.

Here are some statistics on phishing

- According to GreatHorn, **57% of organizations** experience phishing attempts on a weekly or daily basis.
- Almost 1.2% of all emails sent are malicious, amounting to approximately 3.4 billion phishing emails each day.
- In 74% of breaches, human factors played a role, encompassing social engineering tactics, mistakes, or misuse.
- IBM identifies phishing as the leading initial attack vector, responsible for **41% of incidents**.
- CSO Online notes that more than **80% of reported security** incidents are due to phishing.
- CSO Online also reports a loss of **$17,700 every minute** due to phishing attacks.

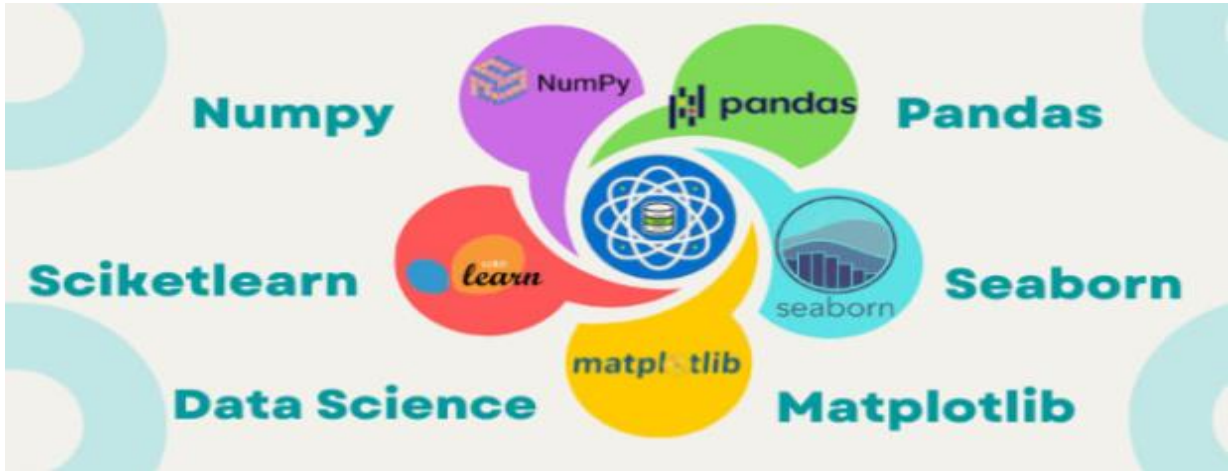
Objectives of the project

The goal is to develop a machine learning model that classifies URLs as phishing or legitimate based on various features such as URL length, domain characteristics, and content elements. This model aims to enhance online security by accurately identifying malicious links, protecting users from potential phishing attacks and fraud.

Requirement for the Project

Using library , application & website

Jupiter notebook , Deepnote



Data Landscape

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.

Quality of Data

The dataset is complete with no missing values or duplicate records.

Key Features

The dataset includes the following key features:

The dataset encompasses key features such as URL characteristics (e.g., URLLength, DomainLength), legitimacy indicators (e.g., TLDLegitimateProb, URLSimilarityIndex), and content/behavioral attributes (e.g., HasTitle, NoOfPopup). It also includes HTTPS metadata (e.g., IsHTTPS, Robots) and additional indicators like social network links and security features.

Relevance to Project Goals

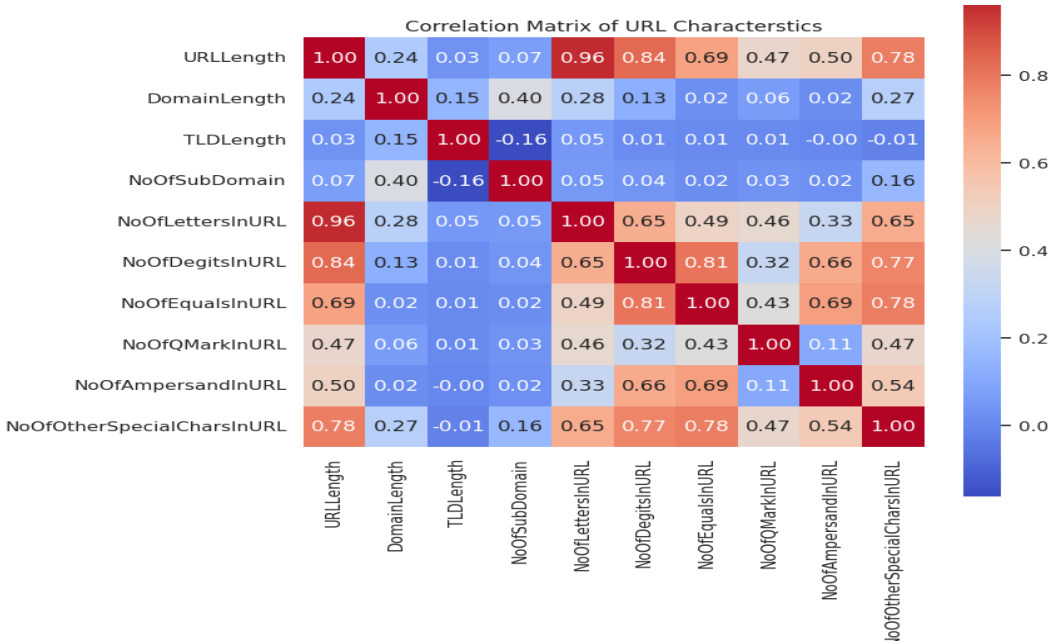
The dataset provides diverse and high-quality features that enable effective phishing detection. It combines characteristics of the URL structure with page content and behavioral indicators, ensuring a well-rounded approach to model training.

Exploratory Data Analysis (EDA)

Potential groupings of the most relevant numeric columns

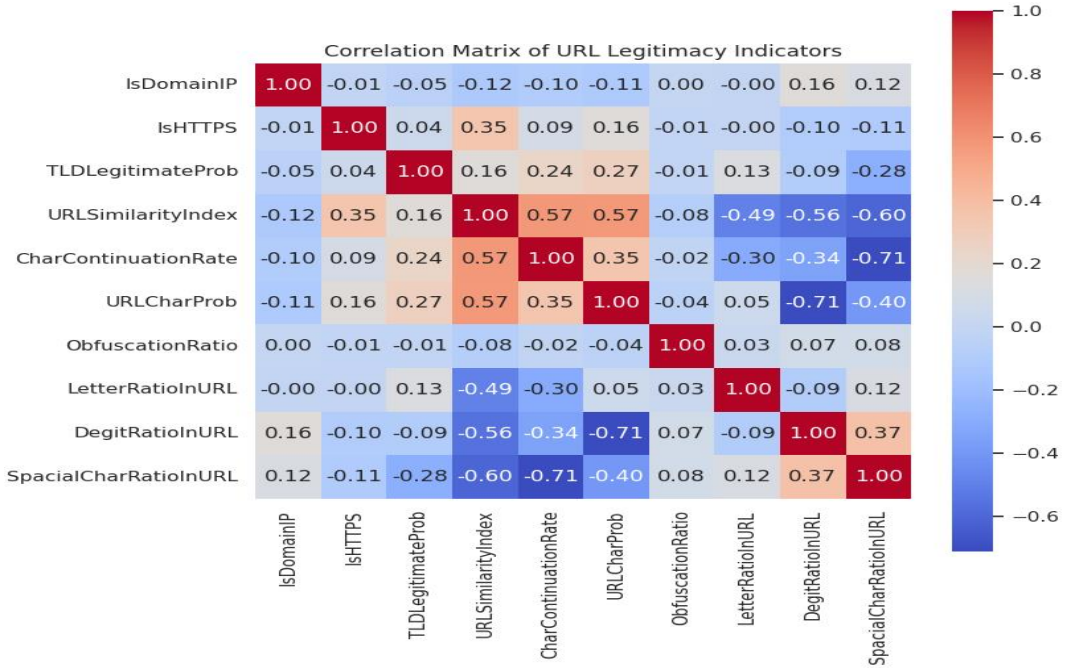
URL characteristics :

Analyzing phishing patterns via relationships.



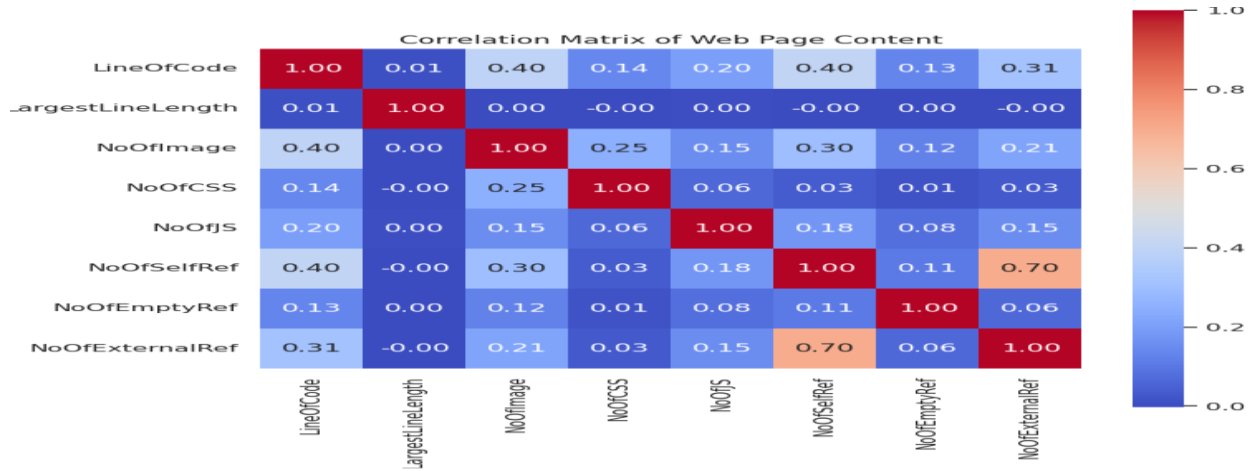
URL legitimacy indicators:

Insights on URL legitimacy factors



Key findings

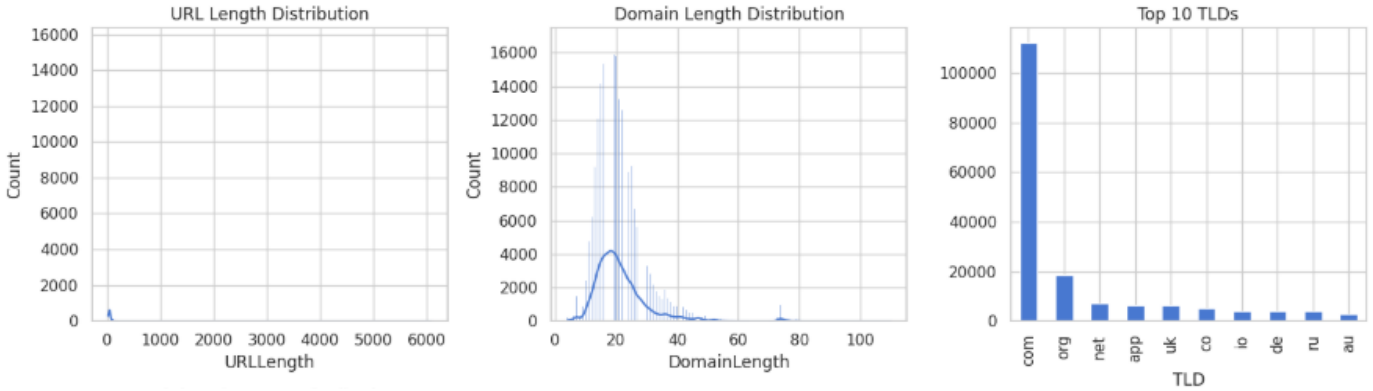
Web Content features

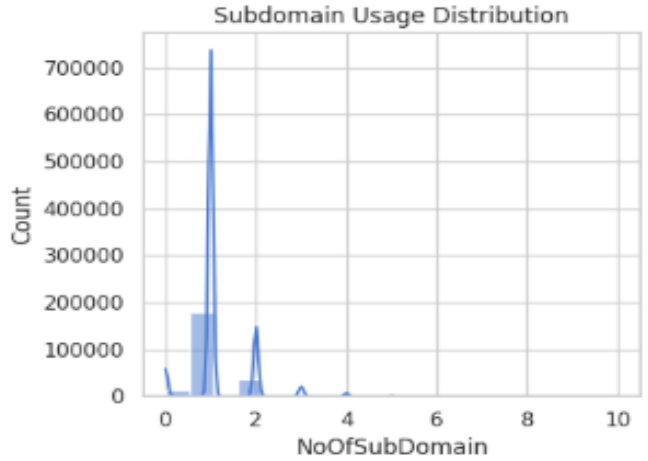


Top Domain influence

```
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
```

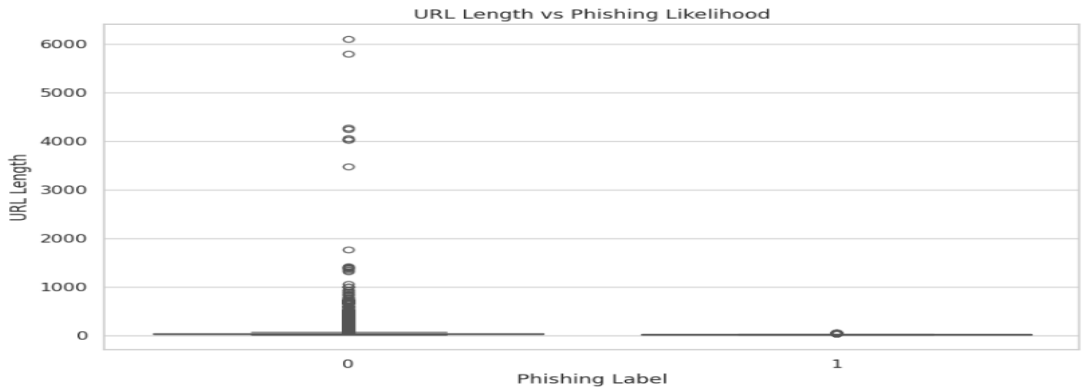
URL Structure Analysis



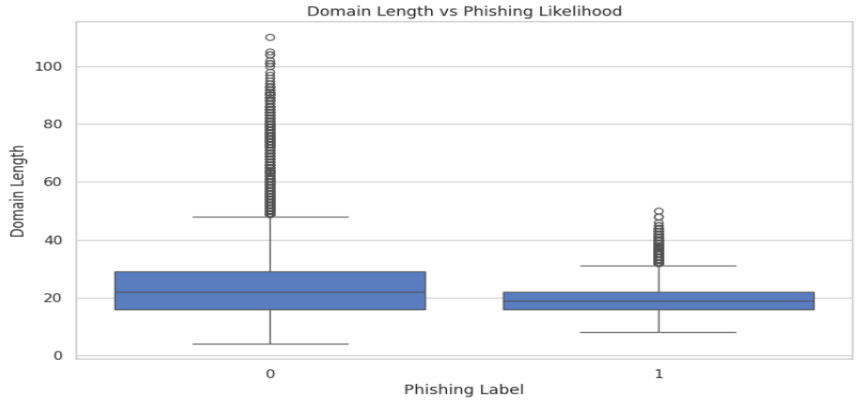


Hypothesis

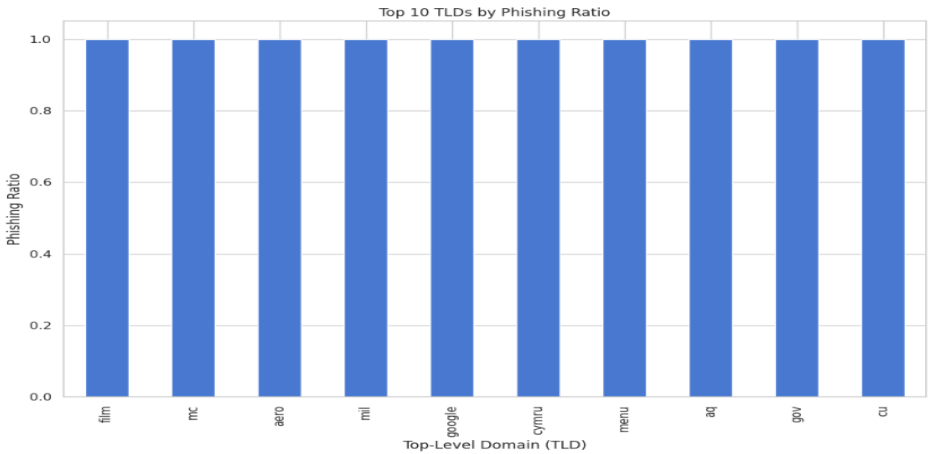
URL Length and Phishing Likelihood



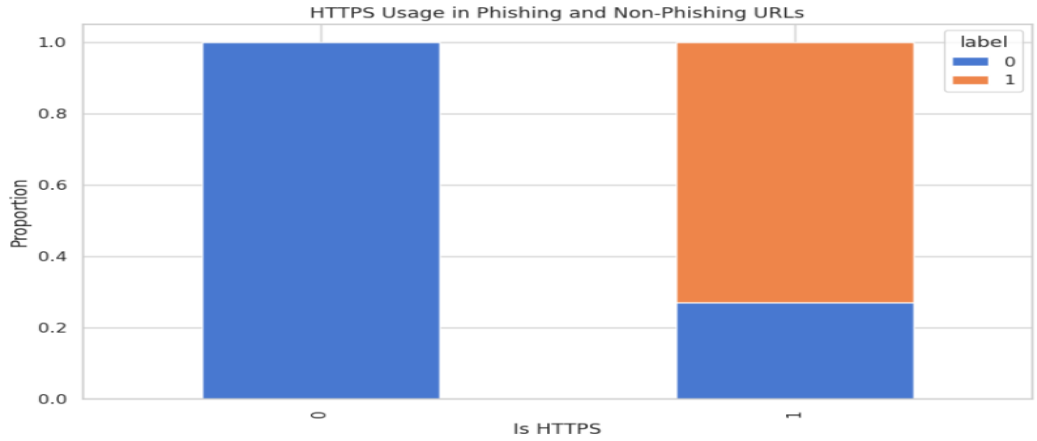
Domain Characteristics and Legitimacy



Top-Level Domain (TLD) Influence

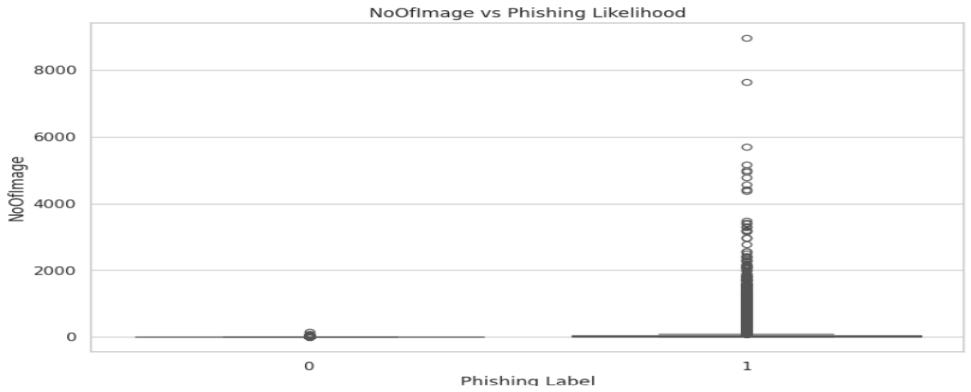


HTTPS Usage and Phishing Likelihood

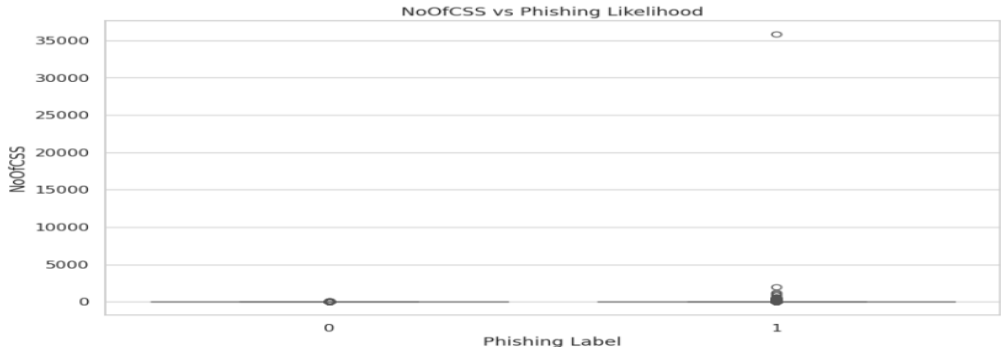


Hypothesis

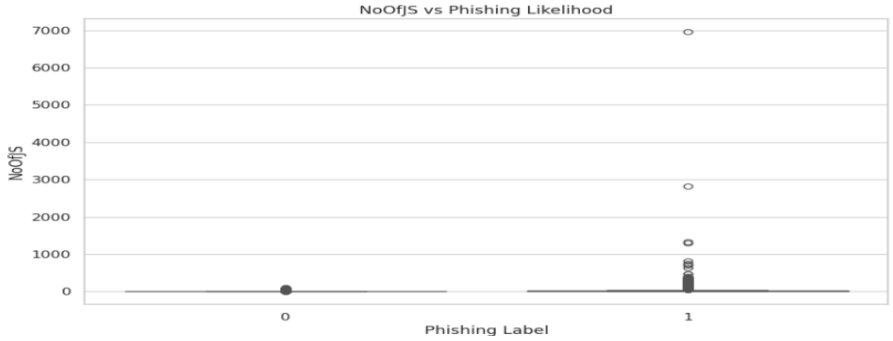
Content Features and Phishing Likelihood

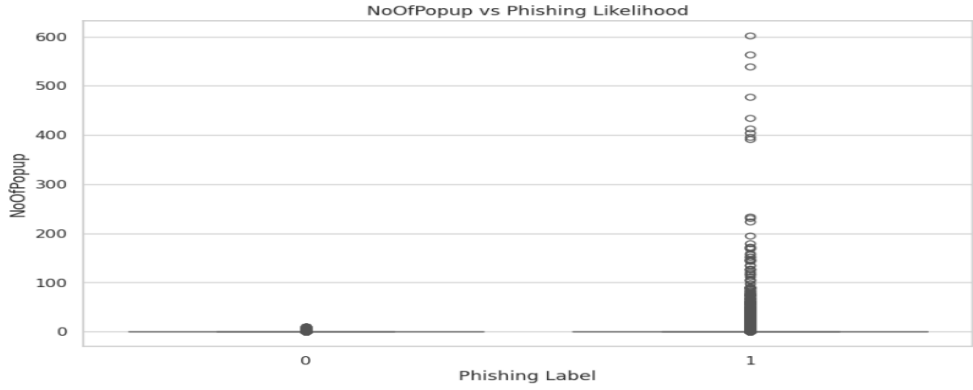


Social Network Links - phishing and non-phishing



JS and Popups and phishing likelihood





Model Training

Feature Encoding

Domain, TLD

Drop unnecessary features and target

Like file name , url , title, label

Data Split into Train and Test

Splits the dataset into training and testing sets: Training Set (80%): Used for model training. Testing Set (20%)

Model Development

Since it is a **classification problem**, we selected four classifier algorithms to build the model and identify the best one.

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.

Model Development

Hyperparameter Tuning : GridSearchCV was used to optimize key parameters (e.g., n_estimators, max_depth, min_samples_split) for each model.

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

Performance Metrics

Accuracy: Measures the percentage of correct classifications.

Precision: Indicates the proportion of true positives among predicted positives.

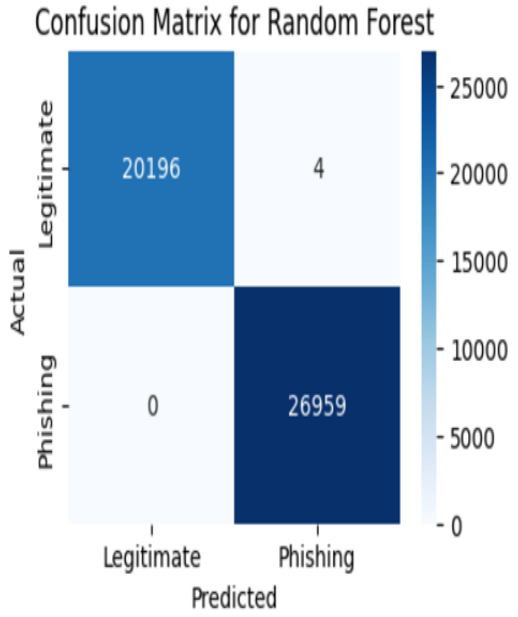
Recall: Represents the proportion of true positives among actual positives.

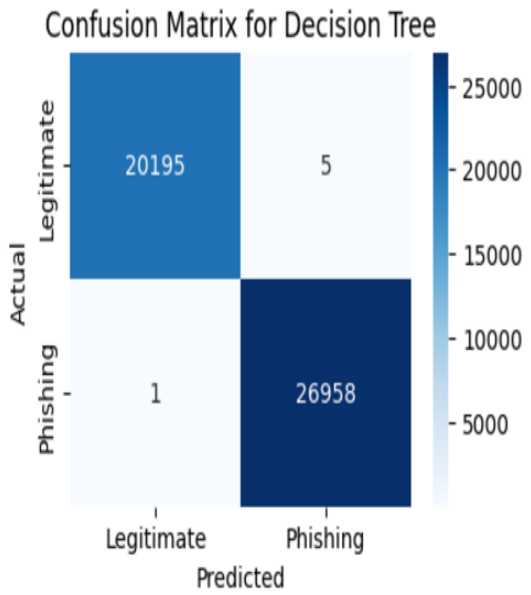
F1-Score: Balances precision and recall.

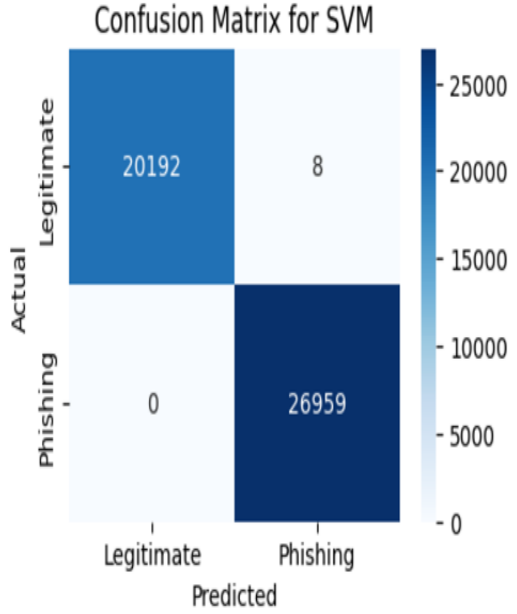
Confusion Matrix: Displays true/false positives and negatives.

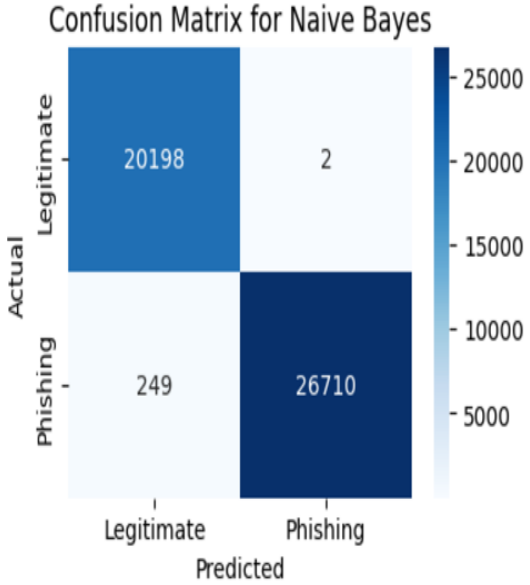
Performance and evaluation

	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









Result Observations

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.

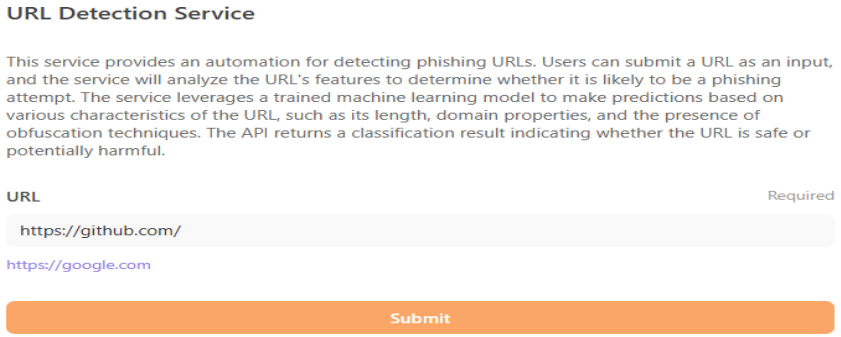
Best model : Random Forest

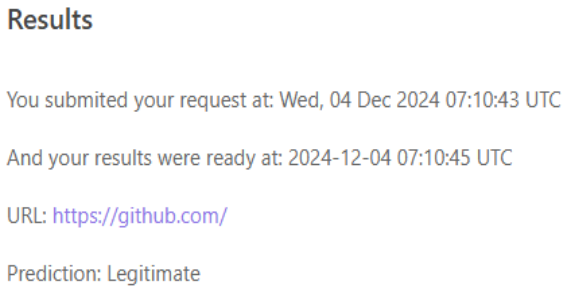
- Random Forest achieved the highest accuracy, precision, recall, and F1-score.
- Its ensemble nature helps prevent overfitting.
- Random Forest handles unseen data effectively, ensuring stable performance.
- Chosen for its robustness and reliability in phishing URL detection.

Deployment and Future work

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

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

Demo URL - Input URL

https://deepnote.com/

Project Jupyter | Home

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

Thank You