1. Business Problem Summary

Phishing attacks have become a widespread threat in the digital era, targeting users through deceptive websites that mimic legitimate platforms to steal sensitive information such as usernames, passwords, and credit card numbers. The main challenge lies in accurately identifying and blocking phishing websites before users fall victim. The goal of this project is to develop an automated system for detecting phishing websites based on key features extracted from URLs and website characteristics.

Scope and Importance:

- Protect users from online fraud and identity theft.
- Enhance cybersecurity infrastructure.
- Reduce economic loss caused by phishing.
- Improve trust in digital platforms by ensuring user safety.

2. Literature Insights

Common Characteristics of Phishing Websites:

- Suspicious URLs with IP addresses or misleading domain names.
- Use of "https" in a deceptive manner.
- High number of redirects.
- Presence of abnormal or mismatched URL components.
- Lack of proper SSL certificates.

Challenges in Detection:

- Rapid evolution of phishing tactics.
- High similarity to legitimate websites.
- Imbalanced datasets due to fewer phishing samples.
- Overfitting in models due to noisy or redundant features.

Potential Solutions:

- Use of machine learning and deep learning models.
- Feature engineering and selection to improve accuracy.

- Real-time detection using URL-based heuristics.
- Ensemble models to improve generalizability.

Dataset Exploration Report

1. Dataset Overview

- **Total Records:** 11055
- **Number of Features:** 31 (30 predictor variables and 1 target variable)
- **Data Types:** Mix of numerical (binary and integer) features.
- **Target Variable:** Result (1 for phishing, -1 for legitimate)

2. Feature Descriptions and Relevance

- **Having_IP_Address:** Indicates use of an IP address in URL (common in phishing).
- **URL_Length:** Longer URLs often indicate phishing.
- **Shortining_Service:** Use of URL shorteners may hide malicious intent.
- **Having_At_Symbol:** "@" in URLs is suspicious.
- **Double_slash_redirecting:** Presence of "//" beyond protocol is anomalous.
- **Prefix_Suffix:** Hyphenated domains can indicate fake sites.
- **Web_Traffic, Page_Rank, Links_pointing_to_page, Statistical_report:** These indicate popularity and legitimacy.
- **SSLfinal_State:** Indicates SSL certificate validity.

3. Exploratory Data Analysis (EDA)

- **Target Distribution:**
    - Phishing: ~55%
    - Legitimate: ~45%
    - Slight class imbalance observed.
- **Histograms:**
    - Most features are binary or categorical with values -1, 0, 1.

    - Phishing sites show a strong correlation with certain values in Having_IP_Address, Prefix_Suffix, and SSLfinal_State.
- **Correlation Heatmap:**
    - Features such as SSLfinal_State, URL_Length, and Web_Traffic show higher correlation with the target.
    - Some features are highly correlated with each other (e.g., Request_URL and URL_of_Anchor).

4. Data Quality Issues

- **Missing Values:** None detected in the dataset.
- **Outliers:** Binary data shows minimal outliers; some features like Web_Traffic require scaling.
- **Redundant Features:** Some correlated features may be dropped during modeling to reduce overfitting.

Handling Strategies:

- Standardization of numeric features like Web_Traffic.
- Dimensionality reduction techniques (e.g., PCA) if required.
- Feature selection to improve model performance.