Business Problem Understanding

1. Problem Summary

Phishing attacks pose a serious threat in the digital age, where attackers create fake websites that mimic legitimate ones to steal sensitive user data such as login credentials, financial information, and personal details. These deceptive websites are often indistinguishable from genuine websites to the average user, which increases the risk of data breaches and financial fraud.

2. Scope and Importance

With the increasing number of internet users and online transactions, the ability to automatically detect phishing websites has become essential. Traditional blacklists are no longer effective due to the dynamic and short-lived nature of phishing sites. Machine learning models offer a scalable and adaptive solution that can classify websites as legitimate or phishing based on various features extracted from the URLs and content.

The goal is to develop a system that can detect phishing websites with high accuracy by leveraging patterns in URL structures, domain information, web content, and metadata.

3. Literature Review: Key Insights

- **Common characteristics of phishing sites**:
    - Use of IP addresses instead of domain names
    - Long or obfuscated URLs with special characters
    - Suspicious use of HTTPS or absence of SSL certificates
    - Use of popups, iframes, and embedded forms
- **Detection challenges**:
    - Fast-evolving nature of phishing websites
    - Mimicry of legitimate websites
    - Class imbalance in datasets
- **Proposed solutions**:
    - Feature-based machine learning models (e.g., Random Forest, SVM, XGBoost)
    - Real-time URL analysis
    - Ensemble learning and feature selection techniques

Dataset Exploration Report

1. Dataset Overview

- **Total Samples**: 11,430
- **Total Features**: 89
- **Target Variable**: status (binary classification: phishing or legitimate)
- **Class Distribution**:
    - phishing: 5,715 (50%)
    - legitimate: 5,715 (50%)

The dataset is balanced, which is beneficial for training machine learning models without requiring class weighting or oversampling.

2. Feature Description

The dataset includes various features derived from the structure and content of URLs and web pages. Some examples include:

- ip: Indicates whether the URL uses an IP address
- https_token: Indicates suspicious use of HTTPS
- nb_dots, nb_hyphens, nb_slash: Count of specific characters in the URL
- web_traffic: Website traffic ranking data
- domain_age, domain_registration_length: WHOIS-based features
- iframe, popup_window, submit_email: Behavioral features common in phishing pages

3. Data Types

- **Text**: URL string (url)
- **Numerical**: URL length, character counts, traffic ranking, domain age
- **Categorical/Binary**: Flags indicating presence of phishing indicators

4. Missing Values

There are no missing values in the dataset. All features are complete.

5. Preliminary Observations

- The dataset is clean and ready for analysis.
- A variety of features are available, enabling detailed behavioral and structural analysis