Phishing Domain Detection System

Internship Project

Low Level Design(LLD) Report

Mann Lahoti

Content:

Phishing Domain Detection System

This document details the functionalities of each component within the Phishing Domain Detection System, referencing the provided high-level design (HLD)

1. Data Acquisition Module

- **Function:** Downloads the phishing domain dataset from the provided source (https://data.mendeley.com/datasets/72ptz43s9v/1).
- **Functionality Breakdown:**
  - Establishes a connection to the dataset source.
  - Downloads the dataset in its designated format (e.g., CSV).
  - Stores the downloaded data in a designated location.

2. Data Preprocessing Module

- **Function:** Cleans, transforms, and prepares the data for further processing.
- **Functionality Breakdown:**
  - Handles missing values (e.g., imputation, deletion).
  - Identifies and corrects inconsistencies in data formats.
  - Converts categorical data into numerical representations for machine learning models.
  - Normalizes numerical features to a common scale.

3. Feature Engineering Module

- **Function:** Extracts relevant features from various aspects of the domains (URL, domain, page, content).
- **Functionality Breakdown:**
  - **URL-Based Features:**
    - Calculates URL length.
    - Identifies presence of special characters (e.g., hyphens, underscores).
    - Extracts subdomain information.
  - **Domain-Based Features:**
    - Queries WHOIS database to retrieve domain age and registration information.
    - Checks for presence on blacklists of known phishing domains.
  - **Page-Based Features (if applicable):**
    - Simulates browser interaction to access the domain.
    - Checks for presence of SSL certificates.
    - Analyses visual elements for suspicious characteristics.

    - **Content-Based Features (if applicable):**
        - Extracts text content from the domain webpage.
        - Identifies keywords commonly associated with phishing attempts.
        - Uses sentiment analysis to detect urgency or negativity in the content.

4. Model Building Module

- **Function:** Trains various machine learning models to classify domains as real or malicious.
- **Functionality Breakdown:**
    - Splits the preprocessed data into training and testing sets.
    - Implements multiple machine learning algorithms (e.g., Random Forest, Support Vector Machine).
    - Trains each model on the training data set.
    - Tunes hyperparameters of each model to optimise performance.

5. Model Evaluation Module

- **Function:** Evaluates the performance of the trained machine learning models.
- **Functionality Breakdown:**
    - Uses the testing data set to evaluate model predictions.
    - Calculates metrics like accuracy, precision, recall, and F1-score for each model.
    - Compares the performance of different models and selects the one with the best overall performance.

6. Model Selection Module

- **Function:** Selects the best performing model for deployment based on evaluation results.
- **Functionality Breakdown:**
    - Analyses evaluation metrics from the Model Evaluation Module.
    - Consider factors like accuracy, precision, and computational efficiency.
    - Selects the model that best balances these factors for real-world application.

7. API/UI Module

- **Function:** Provides an interface for users to interact with the system and test the model with new domains.
- **Functionality Breakdown:**
  - **API:**
    - Defines endpoints for users to submit domain URLs for testing.
    - Receives domain URLs from users.
    - Forwards URLs to the selected model for prediction.
    - Returns the model's prediction (real or malicious) to the user.
  - **UI (Optional):**
    - Provides a user-friendly interface for entering domain URLs.
    - Triggers the API module to submit the URL for testing.
    - Displays the model's prediction in an easy-to-understand format.

8. Reference to Public Github Repo

The provided GitHub repository (https://github.com/MannPlayXZ47/Phishing-Domain-Detection) seems to focus on a different approach using association classification for phishing detection. While this LLD is designed