Low-Level Document (LLD)

Phishing Domain Detection Project

Table of Contents

1. Introduction

The Low-Level Document (LLD) for the Phishing Domain Detection project provides a detailed overview of the various components and steps involved in developing a predictive model for Phishing Domain Detection

2. Data Collection and Preprocessing

- Identify and collect relevant data sources containing historical sites data, and default labels
- Perform data cleaning by checking missing values, duplicates, and outliers.
- Conduct data integration and transformation as required.
- Split the dataset into training and testing sets for model development and evaluation.

3. Exploratory Data Analysis (EDA)

- Perform descriptive statistics to understand the characteristics of the dataset.
- Visualize the data through plots, histograms, and other graphical representations.
- Identify any patterns, correlations, or anomalies in the data.
- Gain insights into the distribution of features and the target variable (default/non-default).

4. Feature Engineering

- Create additional relevant features from the existing dataset that can potentially improve the predictive power of the model.
- Conduct feature selection techniques to identify the most important features.

- Perform feature scaling or normalization to ensure all features are on a similar scale.

5. Model Development

- Select appropriate machine learning algorithms for credit card default prediction, such as SVC, decision trees, random forests, or gradient boosting.
- Implement the selected algorithms using suitable libraries or frameworks (e.g., scikit-learn).
- Split the training dataset further into training and validation sets for model development and hyperparameter tuning.

6. Model Evaluation and Selection

- Define evaluation metrics for assessing the model's performance, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
- Evaluate the initial models using the validation dataset.
- Compare and analyze the performance of different models to identify the most promising candidate(s) for further development.

7. Model Training and Hyperparameter Tuning

- Train the selected model(s) on the training dataset using the chosen algorithm.
- Perform hyperparameter tuning using techniques like grid search, random search, or Bayesian optimization to optimize the model's performance.
- Validate the tuned model using the validation dataset.

8. Model Validation

- Assess the final model's performance using the testing dataset, using the same evaluation metrics as defined earlier.
- Conduct a thorough analysis of the model's strengths, weaknesses, potential biases, and limitations.
- Validate the model's generalizability and robustness through cross-validation techniques.

9. Model Deployment

- Deploy the trained model into a production environment.
- Integrate the model into the existing processing system for real-time predictions.
- Ensure scalability, efficiency, and reliability of the deployment process.

10. Monitoring and Maintenance

- Establish a monitoring system to track the model's performance in the production environment.
- Implement regular model retraining and updating strategies to account for concept drift and changing patterns in system usage.

- Continuously monitor the model's accuracy, precision, and recall, and investigate any significant deviations or anomalies.
- Maintain documentation and version control of the model and its associated code to facilitate future updates and enhancements.

This Low-Level Document provides a comprehensive outline of the various stages involved in the Phishing Domain Detection project, including data collection and preprocessing, exploratory data analysis, feature engineering, model development, evaluation and selection, training and tuning, validation, deployment, monitoring, and maintenance. Following this document will help ensure a systematic and well-structured approach to the project, resulting in an accurate and reliable prediction model.