High-Level Document (HLD)

Phishing Domain Detection Project

Table of Contents

1. Introduction

The Phishing Domain Detection project aims to develop a predictive model that can accurately identify legitimate sites on the basis of site features which will help in identifying phishing domain which will help in addressing phishing problem.

2. Project Overview

The project will involve the following key activities:

- Collecting and prepossessing sites data, and default labels.
- Developing a machine learning model that can predict the likelihood of sites default.
- Evaluating and validating the model's performance using appropriate metrics.
- Deploying the model into a production environment and establishing a monitoring system.

3. Objectives

The main objectives of the Credit Card Default Prediction project are:

- Develop a robust machine learning model that accurately predicts sites behaviour.
- Improve the risk assessment process for financial institutions by providing timely and reliable predictions.
- Enhance the decision-making process by identifying high-risk users and taking proactive measures to prevent defaults.

4. Scope

The scope of this project includes:

- Collecting websites data, customer information, and default labels.
- Prepossessing the data to handle missing values, outliers, and feature engineering.
- Exploratory data analysis to gain insights into the dataset.
- Developing and training a machine learning model using appropriate algorithms.
- Evaluating the model's performance using relevant metrics such as accuracy, precision, recall, and F1-score.
- Deploying the model into a production environment and setting up a monitoring system to track its performance over time.

5. Architecture

The Phishing Domain Detection project will follow a typical architecture consisting of the following components:

- Data Collection: Retrieve historical websites behavioural data, domain information, and default labels from appropriate sources.
- Data Preprocessing: Handle missing values, outliers, and perform feature engineering to prepare the data for model development.
- Model Development: Train and validate a machine learning model using suitable algorithms and techniques.
- Evaluation and Validation: Assess the model's performance using metrics such as accuracy, precision, recall, and F1-score.
- Deployment: Deploy the trained model into a production environment to make real-time predictions.
- Monitoring: Establish a monitoring system to track the model's performance and ensure its effectiveness over time.

6. Data Collection and Preprocessing

- Identify and retrieve data, and default labels from reliable sources.
- Preprocess the data by handling missing values, outliers, and performing feature engineering.
- Split the data into training and testing sets for model development and evaluation.

7. Model Development

- Select appropriate machine learning algorithms based on the project requirements.
- Develop and train the model using the training dataset.
- Optimize the model parameters and hyperparameters to improve its performance.
- Validate the model using the testing dataset.

8. Evaluation and Validation

- Evaluate the model's performance using metrics such as accuracy, precision, recall, and F1-score.
- Analyze the model's strengths, weaknesses, and potential biases.
- Validate the model's effectiveness and generalizability using cross-validation techniques.

9. Deployment and Monitoring

- Deploy the trained model into a production environment, ensuring scalability and real-time prediction capabilities.
- Set up a monitoring system to track the model's performance and detect any performance degradation or concept drift.
- Implement appropriate measures for model retraining and updating as new data becomes available.

10. Conclusion

The Phishing Domain Detection project aims to develop a reliable and accurate machine learning model for predicting legitimate sites. By leveraging historical data and advanced analytics techniques, institutions can make informed decisions and proactively manage information risk. The project will encompass various stages, including data collection, preprocessing,model development, evaluation, deployment, and monitoring.
Successful implementation of this project will lead to improved risk assessment, reduced information losses, and better decision-making for users.