‘PW Skills’ Project for Internship

Project Title: Phishing Domain Detection

Technology: Machine Learning Technology

Domain: Cyber Security

Project Difficulty Level: Intermediate

Executive Summary:

The Phishing Domain Detection project aims to develop a predictive model that accurately identifies sites who are likely to be fraud or unsecure on the basis of domain. By leveraging historical websites data and applying machine learning techniques, the project aims to assist users in identifying high-risk websites, enabling them to take proactive measures to minimize security losses.

Introduction:

Using online sites poses significant security risks for users.
By accurately predicting website legitimatacy, companies can implement effective risk management strategies, such as detecting sites legitimacy, offering security and serve as one of the measures against cyber crimes.

Objectives:

- Develop a predictive model that accurately predicts legitimate sites.
- Identify key factors contributing to riskful sites.
- Enable users to prevent themselves from revealing personal information
- which could be used against them and serve as a measure in cyber security.

Methodology:

a) ***Data Collection:*** Gather a comprehensive dataset containing historical data, websites demographics, and other relevant variables.

b) Data Preprocessing: Cleanse and preprocess the data to remove duplicates, handle missing values, and perform feature engineering, including scaling, normalization, and one-hot encoding.

c) Feature Selection: Identify the most significant features using techniques such as correlation analysis, feature importance ranking, and domain knowledge.

d) Model Development: Apply machine learning algorithms, such as support vector classifier, decision trees, random forests, gradient boosting, or neural networks, to develop a predictive model for legitimate sites prediction. Tune hyperparameters to optimize model performance.

e) Model Evaluation: Evaluate the model's performance using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Utilize techniques like cross-validation and train-test splits to assess the model's generalization ability.

f) Model Interpretation: Interpret the model's results to identify the key factors influencing sites. Analyze feature importances, coefficients, or SHAP values to gain insights into the underlying patterns and factors contributing to defaults.

Project Deliverables:

- Detailed analysis and documentation of the dataset used.
- Preprocessed and cleaned dataset ready for modeling.
- Trained default prediction model with optimized hyperparameters.
- Model evaluation metrics and performance analysis.
- Insights into key features influencing Phishing via websites.
- Recommendations for risk management strategies based on the model's findings.

Timeline:

- Data collection and preprocessing: 1 Day
- Feature selection and engineering: 1 Day
- Model development and tuning: 2 Days
- Model evaluation and interpretation: 4 Days
- Documentation and report preparation: 1 week

Risks and Challenges:

- Availability and quality of historical websites data.
- Dealing with imbalanced datasets where defaults are relatively rare.
- Overfitting or underfitting of the predictive model.
- Interpretability of complex machine learning models.

Conclusion:

The Phishing Domain Detection Project helps in identifying legitimate sites which is an helpful step in contributing as one of the measures for cyber security, as phishing is one of the common and very dangerous phenomenon as it can manipulate personal information and can even lead to worst.