Project Proposal: Credit Card Fraud Detection using Machine Learning

---

Problem Statement: This project is aimed at developing a robust credit card fraud detection system using machine learning techniques. With the increasing prevalence of online transactions, credit card fraud has become a serious concern for both consumers and financial institutions. Detecting fraudulent activities accurately and promptly is crucial to prevent financial losses and maintain people's trust in the banking system.

Value Proposition: The proposed solution holds immense value for financial institutions, cardholders, and online merchants. By effectively identifying fraudulent transactions, it enables timely intervention on the fraudulent transactions thereby reducing financial losses for banks and preventing inconvenience for customers. Moreover, it helps in enhancing the overall reliability and security of online transactions. As a result, the customer's trust in the financial institution is elevated.

Data Source: The primary data source for this project is a Credit Card Fraud Detection Dataset collected from credit card transactions made by European cardholders in 2023, with sensitive information removed to ensure privacy and compliance with ethical guidelines.The dataset consists of about 31 columns and over 550000 records.

The data set is available at:
https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023/data.

Key Features:

- **id**: Unique identifier for each transaction
- **V1-V28**: features representing various anonymized transaction attributes (e.g., time, location, etc.) derived from PCA transformation aimed at protecting user privacy.
- **Amount**: The transaction amount
- **Class**: Binary label indicating whether the transaction is fraudulent (1) or not (0)

Machine Learning Techniques to be implemented:

- **Supervised Learning:**
  - **Decision Trees:** Decision trees partition the feature space based on attribute values, making them well-suited for classification tasks.
  - **Random Forest:** Random Forest, an ensemble learning technique, will be utilized to improve the robustness and accuracy of the classification model. By combining multiple decision trees trained on different subsets of the data, Random Forest reduces overfitting and enhances generalization.

    - **Regression:** Regression techniques, such as Logistic Regression, will be explored to model the probability of fraudulent transactions based on the input features.
- **Unsupervised Learning:**
    - **K-Means:** K-Means clustering will be employed for unsupervised grouping of transactions into clusters based on their feature similarities. K-Means can help in detecting anomalies and suspicious patterns in the data.
    - **Deep Learning:** Deep Learning techniques will be used for their ability to automatically learn complex patterns and representations from raw transaction data. Deep Learning models offer the advantage of capturing intricate relationships in high-dimensional data, potentially improving the detection of subtle fraudulent activities.
- **Model Evaluation:**
    - Cross Validation: Cross-validation will be used to assess the generalization performance of the models by splitting the dataset into multiple train-test subsets. This technique helps in estimating the model's performance on unseen data and reduces the risk of overfitting.
    - Precision, Recall, F1 Score: Precision, Recall, and F1 Score metrics will be computed to evaluate the performance of the classification models. Precision measures the accuracy of positive predictions, Recall quantifies the ability of the model to identify all positive occurrences, and F1 Score provides a balance between precision and recall, making it a suitable metric for imbalanced datasets like credit card fraud detection.

By leveraging a combination of supervised and unsupervised learning techniques, along with model evaluation using cross-validation and performance metrics, we aim to develop a robust credit card fraud detection system capable of accurately identifying fraudulent transactions while minimizing false positives.

Anticipated Challenges:

- Imbalance data: The primary challenge we anticipate is the imbalance between the number of fraudulent and non-fraudulent transactions in the dataset. Addressing this class imbalance is crucial to prevent the model from being biased towards the majority class and to ensure accurate detection of fraudulent activities.
- Processing huge data: Due to processing limits, the data processing would be challenging and models could run longer compared to the other datasets used.

Conclusion: In summary, this project aims to develop an effective credit card fraud detection system leveraging machine learning techniques. By addressing the problem of fraudulent transactions, the proposed solution offers substantial value to financial institutions, cardholders, and online merchants, contributing to enhanced security and trust in online transactions.