**Project Proposal: Credit Card Fraud Detection Using Machine Learning**

**Overview**

Credit card fraud is a significant concern for financial institutions, leading to billions of dollars in losses annually. Fraudulent transactions can be difficult to detect due to their rarity and the complexity of identifying patterns in real-time. Our project aims to build a machine learning model that can effectively detect fraudulent credit card transactions using historical transaction data.

**Problem Statement**

The main objective of this project is to predict whether a given credit card transaction is fraudulent based on the data. We will focus on building a binary classification model that classifies transactions as either fraudulent or non-fraudulent.

**Dataset**

We are using the Credit Card Fraud Data dataset, provided by Neha Roychoudhury on Kaggle. This dataset contains detailed credit card transaction data, including both fraudulent and non-fraudulent transactions. The dataset is specifically designed to help train machine learning models to detect fraud in real-time.

* Dataset Source: [Kaggle Credit Card Fraud Data by Neha Roychoudhury](https://www.kaggle.com/datasets/neharoychoudhury/credit-card-fraud-data)
* Number of records: Approximately 280,000 transactions.
* Number of features: It includes transaction amount, time, and 30 other anonymized features. The target variable (Class) indicates whether the transaction is fraudulent (1) or legitimate (0).

**Objectives**

1. Detect fraudulent transactions using machine learning models.
2. Handle the class imbalance (fraudulent transactions are a small portion of the dataset).
3. Evaluate model performance using appropriate metrics, such as precision, recall, and F1 score.
4. Provide insights into the key features that contribute to detecting fraud.

**Proposed Solution**

We will use machine learning algorithms to classify transactions as either fraudulent or non-fraudulent. Here’s how we’ll approach the problem:

1. Data Preprocessing:
   * Handle any missing data and normalize features like Amount to ensure a consistent scale across features.
   * Address class imbalance using techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or undersampling of non-fraudulent transactions.
   * Engineer features based on patterns like transaction anomalies, time gaps, and geographic information.
2. Model Selection:
   * Implement multiple classification algorithms to determine the most effective model for fraud detection:
     + Logistic Regression: A simple and interpretable model for baseline comparison.
     + Random Forest: A robust model known for handling class imbalance and providing insights into feature importance.
     + XGBoost: A powerful boosting algorithm that often achieves high performance in classification tasks.
3. Evaluation:
   * We will evaluate the models based on metrics such as precision, recall, F1 score, and AUC-ROC to ensure we are balancing the detection of fraud with minimizing false positives.
   * Use confusion matrices to show how well the models distinguish between fraudulent and non-fraudulent transactions.
4. Visualization:
   * Visualize the performance of models using confusion matrices, ROC curves, and feature importance charts.
   * Show patterns in transaction amounts, times, and other key features that are most indicative of fraud using scatter plots and heatmaps.

Technologies to Be Used

* Python Pandas: For data cleaning and preprocessing.
* Scikit-learn: For model building and evaluation.
* XGBoost and Random Forest: For classification and performance improvement.
* SQL or MongoDB: For any database storage and query-related tasks.
* Matplotlib/Plotly: For data visualization and performance analysis.

Team Responsibilities

* Pamala: Project Coordinator & Documentation
* Love: Data Specialist (handles data preprocessing, feature engineering, and class imbalance issues)
* Sacide: Machine Learning Engineer (builds and tunes models, including Logistic Regression, Random Forest, and XGBoost)
* Anthony: Visualization Specialist (creates visualizations for model results and feature importance)

Expected Outcomes

* A well-trained machine learning model capable of accurately detecting fraudulent credit card transactions.
* Visual insights into the key features that contribute most to fraud detection.
* A comprehensive project report detailing our approach, model performance, and potential improvements.