**Project Proposal: Credit Card Fraud Detection**

**Team 3 Members: Anthony Banks, Love Lorissaint, Pamala Cox, Sacide Belair**

**Overview**

Credit card fraud is a huge issue for banks and credit card companies, leading to billions in losses each year. Fraudulent transactions can easily slip by since they don’t happen often, and it's tricky to spot them right away. Our goal is to create a machine learning model that can help detect these fraudulent transactions based on past data.

**What We’re Trying to Do**

The main point of this project is to figure out if a transaction is fraudulent or not using the information we have about it. We're focusing on using machine learning to classify each transaction as either fraud or not fraud.

**The Dataset**

We’re using the [**Credit Card Fraud Detection Dataset 2023**](https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023) from Kaggle, which has a lot of details on transactions, both fraud and legit. The data is set up to help train models to spot fraud in real-time.

* **Dataset Link**: Kaggle Credit Card Fraud Detection Dataset 2023
* **Number of Records**: Around 1 million
* **Number of Features**: It includes details like transaction time, amount, and other anonymized data points. It has a target column where 1 means fraud and 0 means legit.

**Our Goals**

1. Spot fraudulent transactions using machine learning models.
2. Handle the fact that there are way fewer fraud transactions compared to legit ones (class imbalance).
3. Measure how well our model performs using metrics like precision, recall, and F1 score.
4. Find out which features are most helpful in detecting fraud.

**Our Approach**

We’re planning to use a few machine learning models to get the best results. Here’s the breakdown:

1. **Data Preprocessing**:
   * Clean up any missing data and normalize certain features.
   * Deal with the class imbalance by using techniques like **SMOTE** or reducing the number of non-fraudulent transactions.
   * Create new features based on patterns like transaction anomalies, time gaps, and transaction location.
2. **Model Selection**:
   * Try a few different models to see which works best:
     + **Logistic Regression**: Simple and easy to understand.
     + **Random Forest**: Great for imbalanced data and shows feature importance.
     + **XGBoost**: A more advanced model that usually gives top-notch results.
3. **Visualization**:
   * Use confusion matrices to see how well the models are separating fraud from legit transactions.
   * Show which features are most important in spotting fraud.
   * Create scatter plots or heatmaps to visualize things like transaction time, amount, and fraud likelihood.

**Tools We’ll Be Using**

* **Pandas**: To clean and prep the data.
* **Scikit-learn**: For building and testing the models.
* **XGBoost/Random Forest**: For improving our model performance.
* **SQL/MongoDB**: If we need a database for any storage-related tasks.

**Team Roles**

* **Data Specialist (Love)**: Prepares and cleans the data, handles any missing values, and balances out the classes.
* **Machine Learning Engineer (Sacide)**: Builds and fine-tunes the models like Logistic Regression, Random Forest, and XGBoost.
* **Visualization Specialist (Anthony)**: Creates the charts, graphs, and visual explanations of the model results.
* **Documentation Specialist (Pamala)**: Writes up the final report and gets the presentation ready.

**What We Expect to Deliver**

* A machine learning model that can accurately detect fraudulent transactions.
* Visual insights that show which features are most useful in detecting fraud.
* A detailed project report explaining everything we did