Project Proposal

Project Title:

Bank Transaction Fraud Detection

SubmittedBy :
Team Members Name:
1. Sajrudin Aalam
2. Riya Rawat
3. Rekha Kumari Bheel
4. Payal Dhokane

Subject – Major Project
Organization: Global Next Consulting India Private Limited

Abstract

This project aims to build a fraud detection system for bank transactions using machine learning techniques. The workflow includes preprocessing, exploratory data analysis (EDA), and handling data imbalance using CTGAN (Conditional Generative Adversarial Network). The preprocessing phase focuses on cleaning, encoding, scaling, and feature extraction, while CTGAN generates synthetic samples for the minority (fraudulent) class. The next phase will include training and visualizing classification models to detect fraudulent activities. This project ultimately seeks to produce a robust, accurate, and balanced fraud detection system that helps financial institutions identify and prevent fraudulent transactions effectively.

Introduction

Financial fraud detection is one of the most critical challenges faced by banking and digital payment systems. Fraudulent transactions often form a very small portion of total data, creating imbalance and bias in machine learning models. This project focuses on preparing high-quality, balanced data through preprocessing and CTGAN-based synthetic data generation. In later stages, this dataset will be used for machine learning-based fraud detection. The project integrates multiple stages — data preparation, visualization, and predictive modeling — to achieve an efficient, end-to-end data pipeline.

Objectives

- To load, clean, and preprocess the bank transaction dataset for modeling.
- To perform EDA to understand data patterns and correlations.
- To balance the dataset using CTGAN for synthetic data generation.
- To train and evaluate machine learning models for fraud detection.
- To visualize performance metrics (accuracy, precision, recall, ROC curve).
- To prepare a complete, optimized dataset ready for deployment in fraud detection systems.

Problem Statement

Fraudulent transactions in banking data are rare, leading to imbalanced datasets where traditional models tend to misclassify minority (fraud) cases. Additionally, raw transaction data often contain missing values, mixed data types, and redundant information. The project aims to create a clean, balanced, and informative dataset using CTGAN and preprocessing techniques, followed by model training and visualization to accurately predict fraudulent behavior.

Scope of Work

The project will cover:

- **Data Preprocessing:** Cleaning, feature extraction, encoding, and scaling.
- **EDA:** Statistical and graphical analysis to understand relationships and detect anomalies.
- **Synthetic Data Generation:** Using CTGAN to balance fraud and non-fraud data.
- **Model Training :** Building and evaluating machine learning models such ----------for fraud detection.
- **Visualization:** Plotting feature importances, confusion matrices, and fraud pattern charts.

Tools and Technologies Used

.

- Programming Language: Python
- Libraries: pandas, numpy, scikit-learn, matplotlib, seaborn, SDV (CTGAN), PCA
- Environment: Google Colab /Jupyter Notebook
- Techniques:
  - Data Cleaning and Feature Engineering
  - Encoding and Scaling
  - Dimensionality Reduction (PCA)
  - Synthetic Data Generation (CTGAN)
  - Fraud Detection (ML Models – upcoming)
- Dataset: Bank_Transaction.csv

Methodology

Phase 1: Data Preprocessing

- Handle missing values using SimpleImputer.
- Scale numeric columns with StandardScaler.
- Encode categorical data with OneHotEncoder.
- Extract new datetime features (year, month, hour).
- Apply PCA to retain 95% of variance.

Phase 2: EDA

- Analyze data distributions, correlations, and transaction trends.
- Identify imbalance in fraud vs. non-fraud records.

Phase 3: Synthetic Data Generation (CTGAN)

- Train CTGAN on minority fraud data.
- Generate realistic synthetic fraud transactions.
- Combine synthetic and real data for balance.

Phase 4: Model Training

- Evaluate models using Accuracy, Precision, Recall, F1-score.

Phase 5: Visualization

- Plot feature importance, confusion matrix, ROC curve.
- Visualize real vs. synthetic data distributions.

Timeline

Days	Task
Days 1	Dataset study and requirement analysis
Days 2	Data cleaning, preprocessing, and feature extraction
Days 3	Exploratory Data Analysis (EDA)
Days 4	Synthetic data generation using CTGAN
Days 5	Model training and evaluation (fraud detection models)
Days 6	Visualization of results and final report preparation

Expected Outcome

- Cleaned and preprocessed dataset ready for modeling.
- Balanced dataset with realistic fraud samples using CTGAN.
- Fraud detection model capable of distinguishing fraudulent and legitimate transactions.
- Visual insights showcasing model performance and fraud detection trends.

Team Details

Name	Role
Sajrudin Aalam	Data Preprocessing , EDA, CTGAN Implementation
Payal Dhokane	EDA , Synthetic Data Generation
Rekha Kumari Bheel	Preprocessing,Model Development
Riya Rawat	Visualization and Report Preparation

References

- Scikit-learn Documentation – https://scikit-learn.org/stable/
- SDV (CTGAN) Documentation – https://docs.sdv.dev/
- Kaggle: *Bank Transaction Fraud Detection Dataset*
- Python Official Documentation – https://www.python.org/doc/