FINAL Week 6 Final Project Report

Fraud Detection with CTGAN

Name : Abhienaya Sri

Batch: AIML

1. Introduction

Fraudulent transactions pose a significant challenge in financial systems, leading to billions of dollars in losses every year. Detecting such fraud is challenging because fraud cases are **rare** compared to legitimate transactions, resulting in **imbalanced datasets**.

Traditional machine learning models tend to underperform on such datasets because they are biased toward the majority class (non-fraud). This project addresses the issue using **CTGAN (Conditional Tabular GAN)** to generate synthetic fraudulent transaction data, thereby balancing the dataset and improving detection performance.

---

2. Objectives

The main goals of this project are:

1. To **balance the dataset** using CTGAN-generated synthetic fraud data.
2. To improve **model recall** for rare fraudulent transactions without sacrificing precision.
3. To develop an **end-to-end pipeline** from data preprocessing to model evaluation.

---

3. Dataset Description

- **Source:** Kaggle Credit Card Fraud Detection Dataset
- **Size:** 284,807 transactions
- **Features:**
    - Numerical features (V1 to V28) generated from PCA transformations for privacy
    - Amount – Transaction amount
    - Class – Target label (0 for non-fraud, 1 for fraud)

Class Distribution:

Class	Count	Percentage
Legitimate (0)	284,315	99.83%
Fraudulent (1)	492	0.17%

---

4. Challenges in Fraud Detection

- **Data Imbalance –** Very few fraud cases make model training difficult.
- **Overfitting Risk –** Models may memorize minority samples.
- **Generalization –** Models may fail on unseen fraudulent patterns.

---

5. Proposed Solution

We used **CTGAN (Conditional Tabular GAN)** to generate synthetic fraud transaction records. CTGAN is designed for **tabular data with mixed data types** and can model complex feature distributions conditioned on the target variable.

Advantages of CTGAN in Fraud Detection:

- Generates **realistic minority class samples** without simply duplicating data (unlike SMOTE).
- Preserves statistical relationships between features.
- Handles skewed and imbalanced datasets effectively.

---

6. Methodology

The project followed this pipeline:

Step 1 – Data Preprocessing

- Removed null/missing values (none in this dataset).
- Scaled features (StandardScaler for Amount).
- Split dataset into **fraudulent** and **non-fraudulent** subsets.

Step 2 – CTGAN Training

- Trained CTGAN on the **fraudulent subset** (Class = 1) only.
- Set conditional column = Class to ensure correct label generation.
- Generated **10,000 synthetic fraudulent transactions**.

Step 3 – Dataset Augmentation

- Combined original dataset with synthetic fraud transactions.
- Result: **Balanced dataset** with ~50% fraud and 50% legitimate transactions.

Step 4 – Model Training

Trained three models on the **augmented dataset**:

- Logistic Regression

- Random Forest
- XGBoost

Step 5 – Model Evaluation

- Metrics: Precision, Recall, F1-score, ROC-AUC
- Compared results before and after augmentation.

---

7. Results

Model	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Logistic Regression	0.85	0.72	0.78	0.95
Random Forest	0.91	0.88	0.89	0.98
XGBoost	0.93	0.90	0.91	0.99

Key Findings:

- **Recall** increased significantly after augmentation.
- XGBoost performed the best overall.
- CTGAN-generated data improved rare-event detection without large precision loss.

---

8. Visualizations

- **Class distribution before and after augmentation** (bar charts).
- **ROC curves** for each model.
- **Feature importance** for Random Forest & XGBoost.

---

9. Conclusion

CTGAN proved to be an effective method for **balancing imbalanced fraud datasets**. By generating high-quality synthetic fraud samples, the model performance improved, especially in detecting rare fraud cases.

Impact:

- Higher recall ensures more fraudulent transactions are flagged.
- Balanced datasets improve fairness and robustness in model training.

---

10. Future Work

- Compare CTGAN with **TVAE** and other generative methods.

- Explore **ensemble models** combining multiple classifiers.
- Deploy fraud detection model as a **real-time API**.
- Test across different industries (e-commerce, banking).

---

11. References

- Credit Card Fraud Detection Dataset – Kaggle
- Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). **Modeling Tabular data using Conditional GAN.**
- CTGAN GitHub Repository