**Fraud Detection in Credit Card Transactions with Oversampling**

**Submitted by: Aswin V T – 23BM6JP14**

# 1. Introduction

Credit card fraud detection is crucial in financial security, as fraudulent transactions can result in significant economic losses for both banks and customers. This project aims to build a machine learning-based fraud detection system that accurately distinguishes between genuine and fraudulent transactions. Since fraud cases are rare, this project incorporates oversampling techniques to handle the class imbalance problem.

# 2. Dataset and Preprocessing

The dataset comprises credit card transactions made by European cardholders in September 2013, covering a period of two days. It contains a total of 284,807 transactions, of which only 492 were identified as fraudulent. This results in a highly imbalanced dataset, with fraud cases constituting just 0.172% of all transactions.

All input variables are numerical and have been transformed using Principal Component Analysis (PCA), except for the 'Time' and 'Amount' features. The 'Time' feature records the number of seconds that have elapsed between each transaction and the first transaction in the dataset, while 'Amount' represents the value of the transaction. The target variable, 'Class,' is binary: a value of 1 indicates a fraudulent transaction, while 0 indicates a legitimate one.

This study utilizes a dataset of anonymized credit card transactions, labelled to differentiate between fraudulent and non-fraudulent cases.

The preprocessing steps included:

* Handling missing values
* Feature scaling and transformation
* Exploratory Data Analysis (EDA) to understand class distribution and transaction amount trends
* Application of oversampling techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Edited Nearest Neighbors (ENN) to balance the dataset.

## Class Imbalance

The dataset is highly imbalanced, with most transactions being non-fraudulent. Specifically:

Non-fraudulent transactions: 284,315 (99.83%)

Fraudulent transactions: 492 (0.17%)

This extreme imbalance presents challenges in training machine learning models, as traditional algorithms tend to favor the majority class. Effective strategies such as oversampling are essential to improve model performance in detecting fraudulent transactions.

# 3. Methodology

The following machine learning models were employed:

* Random Forest Classifier
* XGBoost Classifier
* LightGBM Classifier
* CatBoost Classifier
* Artificial Neural Networks (ANN)

The models were trained on both the original imbalanced dataset and the oversampled dataset to assess the impact of data balancing on fraud detection performance.

# 4. Model Evaluation

The models were evaluated using standard performance metrics:

* **Accuracy:** Measures the overall correctness of the model.
* **Precision:** Determines how many predicted fraud cases were actual frauds.
* **Recall** (Sensitivity): Measures the ability of the model to detect fraud cases.
* **F1-score:** Balances precision and recall to provide a single performance measure.

# 5. Results and Findings

Key Observations from the Comparison

SMOTEENN Improved Recall for All Models

* Random Forest recall increased from 80.88% → 96.32%
* XGBoost recall increased from 81.62% → 96.32%
* CatBoost recall increased from 82.35% → 96.32%
* ANN recall saw the largest increase from 74.26% → 97.79%
* LightGBM recall remained the same at 97.06%

Precision Dropped for Most Models (Expected due to Oversampling)

* XGBoost precision dropped from 94.87% → 74.01%
* CatBoost precision dropped from 95.72% → 66.83%
* LightGBM had the largest drop from 99.99% → 51.56%
* ANN saw a significant increase in precision from 0.11% → 36.34%

F1-Score Became More Balanced

* Random Forest F1-score improved from 86.27% → 88.51%
* ANN improved significantly from 0.24% → 52.98%
* LightGBM dropped from 99.92% → 67.34% (due to precision drop)

Accuracy Remained High for All Models

* Minimal change in overall accuracy for most models (~99.9%)
* ANN saw the biggest jump in accuracy from 1.12% → 99.72%

The results indicate that oversampling significantly improves model performance, particularly in recall and F1-score, which are critical for fraud detection. Below are the findings based on model performance:

* SMOTEENN improved recall across models by an average of 15.2%, enhancing fraud detection performance.
* Random Forest recall increased from 80.88% to 96.32%, while ANN recall improved significantly from 74.26% to 97.79%.
* Precision decreased in some models (e.g., XGBoost dropped from 94.87% to 74.01%) due to oversampling, which is expected when optimizing recall.
* ANN saw the most significant impact, improving from an F1-score of 0.24% to 52.98% and accuracy from 1.12% to 99.72%.
* Overall, Random Forest and XGBoost maintained strong balance, making them optimal models for fraud detection.

The results indicate that oversampling significantly improves model performance, particularly in recall and F1-score, which are critical for fraud detection.

Below are the results for each model:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Precision (Train)** | **Precision (Test)** | **Recall (Train)** | **Recall (Test)** | **F1-score (Train)** | **F1-score (Test)** | **Accuracy (Train)** | **Accuracy (Test)** |
| Random Forest Classifier | 1.000000 | 0.818750 | 1.000000 | 0.963235 | 1.000000 | 0.885135 | 100.00% | 99.96% |
| XGBoost Classifier | 1.000000 | 0.740113 | 0.999991 | 0.963235 | 0.999996 | 0.837061 | 100.00% | 99.94% |
| LightGBM Classifier | 0.999996 | 0.515625 | 0.999366 | 0.970588 | 0.999681 | 0.673469 | 99.97% | 99.85% |
| CatBoost Classifier | 1.000000 | 0.668367 | 0.999877 | 0.963235 | 0.999938 | 0.789157 | 99.99% | 99.92% |
| Artificial Neural Networks (ANN) | 0.999965 | 0.363388 | 0.997987 | 0.977941 | 0.996416 | 0.529880 | 99.64% | 99.72% |

* For best overall performance, Random Forest and XGBoost stand out due to their balance between precision, recall, and F1-score.
* If recall is the primary focus, LightGBM and ANN can be considered, but at the cost of more false positives.
* For a well-rounded choice, CatBoost provides a strong balance and is a reliable option for fraud detection.

# 6. Conclusion

This project demonstrates that oversampling techniques effectively mitigate class imbalance in fraud detection datasets. The study highlights the importance of recall-oriented models for fraud detection, as missing fraudulent cases can be more costly than occasional false positives. Future improvements may include ensemble learning techniques and real-time fraud detection deployment.

# 7. Future Scope

* Implementation of deep learning architectures for improved fraud detection
* Real-time fraud detection system deployment
* Integration of additional features such as user behaviour analysis and transaction context to enhance prediction accuracy