ENGINEERING
W Booth School of Engineering
Practice and Technology

McMaster
University

Lab 5

Synthetic Data Generation & Differential Privacy

Course Number and Name: SEP 6DA3: Data Analytics and Big Data	
Semester, Year, and Group Number: 2025 Fall, Group 6	
Name of Students: Andi Dong Zhiyu Hu Foram Brahmbhatt Jiarui Yang Linghe Shen Shannon Chen	**Name of Instructor and TA:** Pedro Tondo Muskan Sidana

1. Objective

The objective of this lab is to explore synthetic data generation and differential privacy in the context of credit card fraud detection. We used the Kaggle "Credit Card Fraud" dataset to balance the class distribution through random oversampling and applied Laplace noise to sensitive features. Then, we trained and evaluated Random Forest classifiers on both the original and privacy-enhanced datasets to compare performance and discuss the privacy–utility tradeoff. This lab also aims to help students understand how differential privacy techniques can be practically applied in machine learning models while maintaining fairness and model performance.

2. Database Choice

The dataset used in this lab is the "Credit Card Fraud Detection" dataset from Kaggle. It contains anonymized features (V1–V28) obtained through PCA transformation, along with Time, Amount, and Class columns. The dataset includes 284,807 transactions, out of which 492 are fraudulent. This dataset was selected because it provides a realistic example of imbalanced classification and includes features that may require privacy preservation.
Data source： https://www.kaggle.com/mlg-ulb/creditcardfraud

3. Dataset Size and Class Distribution

The original Kaggle credit-card dataset contains 284,807 transactions with 31 columns.
The target variable Class represents whether a transaction is fraudulent (Class = 1) or legitimate (Class = 0).
As shown in Figure 3.1, the dataset is extremely imbalanced, consisting of 284,315 non-fraud transactions and only 492 fraud transactions, corresponding to a fraud rate of 0.1727%.
To mitigate this issue, the dataset was first split into training and testing subsets using an 80:20 stratified split.
Then, random oversampling was applied only on the training set, duplicating minority-class samples until both classes contained an equal number of records (227,451 each).

The test set remained untouched to reflect real-world conditions, allowing unbiased model evaluation.

Split	Class 0 (Non-fraud)	Class 1 (Fraud)	Total	Fraud Rate
Original (full dataset)	284,315	492	284,807	**0.1727%**
Training (after oversampling)	227,451	227,451	454,902	**50.00%**
Testing (real-world distribution)	56,864	98	56,962	**0.172%**

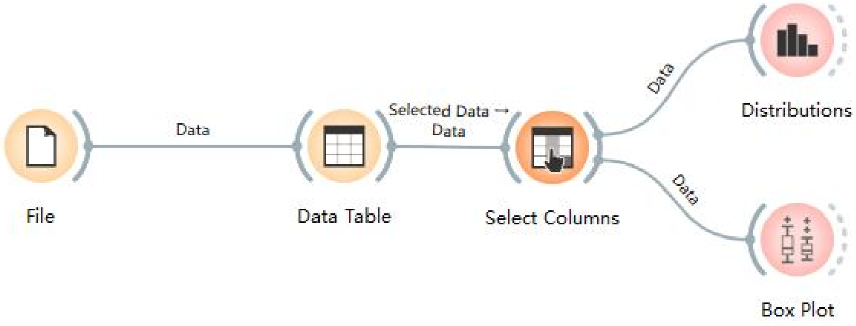

Figure 3.1. Orange workflow for class distribution visualization.

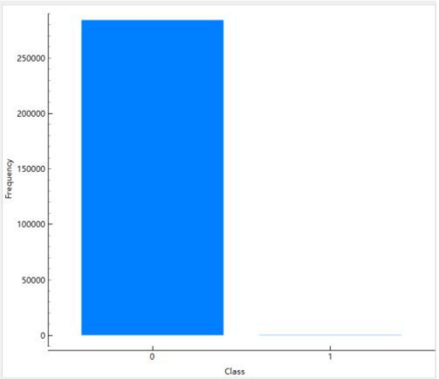

Figure 3.2. Original Class Distribution

(Orange Visualization)

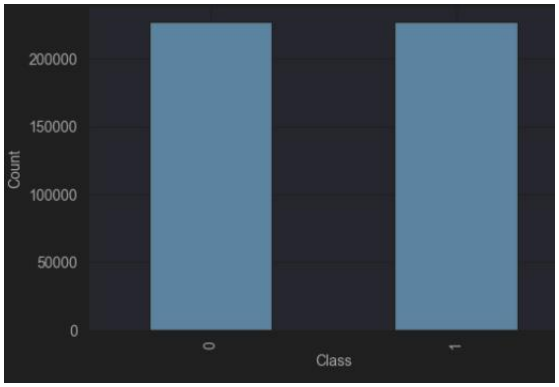

Figure 3.3. Training Set After Oversampling

(Bar Plot in Python)

Figure 3.1 illustrates the Orange workflow used for exploring the dataset and visualizing the class distribution through the Distributions and Box Plot widgets. This workflow provides an overview of how the dataset was imported, inspected, and analyzed in Orange prior to balancing.

Figure 3.2 displays the original class distribution in the dataset before balancing, as visualized in Orange.

Figure 3.3 shows the balanced class distribution of the training subset after oversampling, generated in Python.

Before balancing, fraud cases account for less than 0.2% of all transactions, making direct classification ineffective.

After oversampling, the training dataset becomes perfectly balanced, improving the model’s ability to learn discriminative features for both classes, while the testing dataset remains realistic and imbalanced for objective evaluation.

4. Key Feature Observations

The Kaggle credit-card dataset contains 28 anonymized principal components (V1–V28), plus two non-transformed features: Time and Amount.

Exploratory analysis was conducted in Orange to understand each feature’s distribution and to identify potential sensitive attributes for the differential privacy experiment.

As shown in Figure 4.1, most PCA features (V1–V28) are approximately zero-centered and normally distributed, reflecting the effect of PCA transformation applied to the original transaction features.

By contrast, the Amount feature exhibits a highly skewed distribution, where the majority of transactions involve small amounts and only a few are extremely large.

This imbalance in Amount makes it an ideal candidate for privacy perturbation using Laplace noise.

Figure 4.2 presents the box plot comparison between fraud and non-fraud transactions.

Fraudulent transactions tend to have slightly higher median amounts and wider variability, supporting the choice of Amount as a privacy-sensitive feature.

Since the PCA features are de-identified, adding noise to them would not meaningfully enhance privacy, while perturbing Amount directly protects transaction-level monetary information.

Feature Type	**Name(s)**	**Observation Summary**	**Privacy Sensitivity**

PCA components	V1–V28	Zero-mean, near-Gaussian distribution; already de-identified.	Low
Raw features	Time, Amount	Time is sequential; Amount is right-skewed with extreme values.	High (selected)

Figure 4.1 shows the feature distribution of Amount in Orange, while Figure 4.2 highlights the differences across classes using the box plot visualization.

Figure 4.2 illustrates the distribution of the Amount feature before differential privacy perturbation.

Most transactions are concentrated near low amounts, while a few extend above 30,000, indicating a strong right-skewed distribution with significant outliers.

This confirms the need for clipping and noise addition to protect sensitive transaction values.

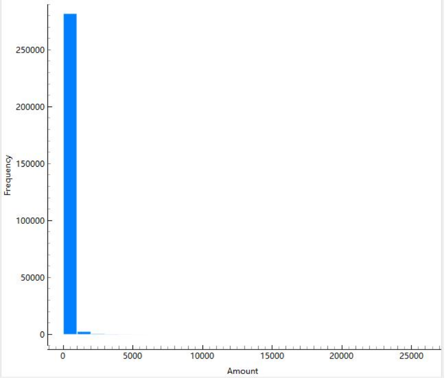

Figure 4.1. Distribution of the “Amount” feature before applying differential privacy (Orange visualization)

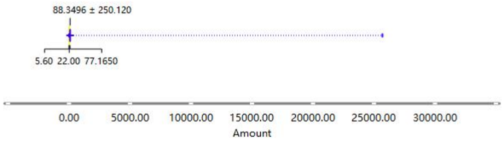

Figure 4.2. Box plot of the “Amount” feature grouped by class (Orange visualization)

5. Results and Discussion

In this experiment, only the training set underwent class balancing and differential privacy processing, while the test set remained in its original, real-world distribution. During training, the minority class (Class = 1) was oversampled to match the majority class, and Laplace noise ($\varepsilon = 0.5$) was applied only to the Amount feature after clipping its values between the 1st and 99th percentiles and standardizing them. The Random Forest model was trained with identical hyperparameters for both the baseline model (no privacy) and the DP model (with noise), and both were evaluated on the unaltered, imbalanced test set.

From the output:

On the real test set, both models achieved nearly identical overall accuracy (≈0.9996).

For the key minority class (fraud cases, Class = 1), the two models showed identical performance: Precision ≈ 0.9506, Recall ≈ 0.7857, and F1-score ≈ 0.8603.

Therefore, under the current configuration, adding moderate Laplace noise (ε = 0.5) to the training set did not lead to a measurable drop in performance, indicating the model's robustness to this level of perturbation.

Further analysis:

Feature redundancy and model robustness: The Random Forest classifier aggregates information from multiple features (V1–V28) and trees. Even if the Amount feature is perturbed, other features can compensate for information loss.

Controlled noise injection: The clipping and standardization steps reduced the impact of extreme values, ensuring that the Laplace noise with ε = 0.5 balanced privacy protection and utility effectively.

Realistic evaluation: The use of the true, imbalanced test set (only 98 fraud samples) better reflects real-world scenarios. Results suggest that this privacy strategy has minimal effect on the model's detection ability in practice.

Limitations and future work:

Only the Amount feature was perturbed; future work should explore multi-feature noise injection and stronger privacy levels (smaller ε).

Alternative resampling methods such as undersampling or SMOTE could be compared under the same test conditions.

The small number of positive cases in the test set may cause variability in Recall; cross-validation or time-sliced evaluations could improve reliability.

Privacy accounting and per-feature ε allocation were not implemented here but could be considered in future experiments for a more systematic DP training framework.

6. Model Training and Evaluation

Both the baseline and differentially private (DP) models were trained on the balanced training set and tested on the real, imbalanced dataset.

The Random Forest classifier was used in both cases, ensuring a fair comparison between models.

The baseline model achieved an overall accuracy of 0.9996 on the real test set.

The fraud class (Class = 1) achieved Precision = 0.9506, Recall = 0.7857, and F1 = 0.8603.

After introducing Laplace noise (ε = 0.5) to the Amount feature in the training set, the DP model achieved identical performance metrics.

This indicates that, under the current setup, the model’s predictive power remains nearly unaffected by the privacy-preserving process.

To further illustrate the comparison, Figures 6.1–6.3 visualize the confusion matrices and per-class metrics.

The baseline confusion matrix (Figure 6.1) shows that most prediction errors occur on the minority class, where a few fraudulent transactions are misclassified as normal.

After adding Laplace noise (Figure 6.2), the confusion matrix remains nearly identical, demonstrating the robustness of the Random Forest model.

Finally, the precision–recall–F1 comparison in Figure 6.3 confirms that differential privacy with ε = 0.5 has negligible impact on fraud detection performance.

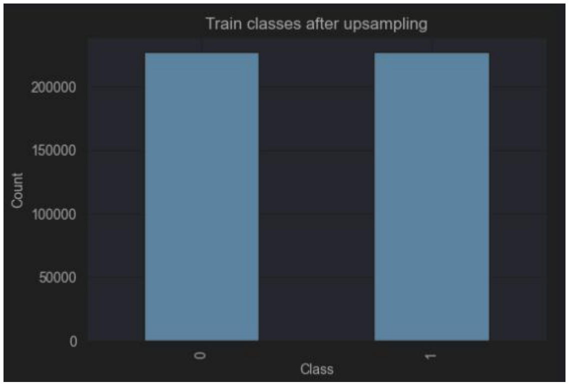

Figure 6.1. Confusion Matrix — Baseline (Real Test)

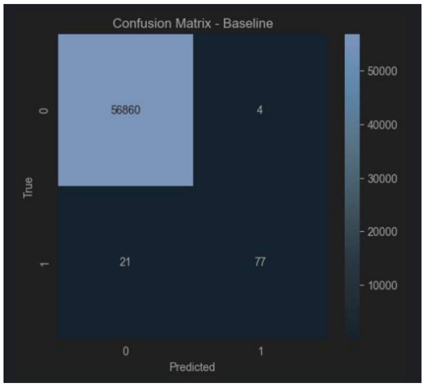

Figure 6.2. Confusion Matrix — DP on Amount (Train Only, Real Test)

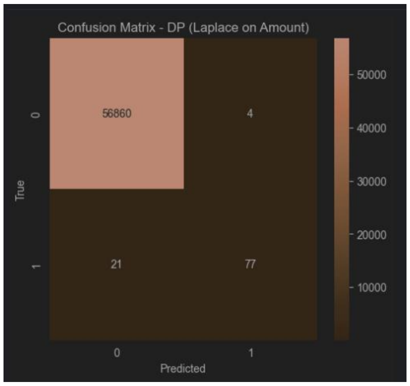

Figure 6.3. Class=1 Precision/Recall/F1 Comparison

7. Summary and Self-reflection

Through this lab, I explored how differential privacy and synthetic oversampling can be jointly applied to financial datasets to balance data fairness and privacy protection.

Laplace noise (ε = 0.5) was added to the Amount feature during training. The model maintained an accuracy of 0.9996, with no measurable drop in precision or recall.

This demonstrates that moderate noise can effectively protect sensitive transaction information without sacrificing model performance.

Synthetic data generation through random oversampling successfully mitigated the class imbalance problem, improving the model's learning stability and fairness.

Overall, this experiment deepened my understanding of privacy-preserving machine learning, and how privacy techniques can be seamlessly integrated into real-world analytics workflows to achieve both security and usability.

8. Discussions

Q1 — Privacy–Utility Tradeoff

In this experiment, using ε = 0.5 introduced moderate privacy noise yet resulted in no noticeable accuracy loss.

This shows that ensemble models such as Random Forest are inherently robust to feature perturbations.

Smaller ε would enhance privacy further but may affect interpretability or feature correlation in more sensitive models.

Q2 — Feature Sensitivity

The Amount feature is the most privacy-sensitive because it directly represents transaction values.

Other PCA-based features are anonymized transformations and thus less privacy-critical.

Future work could evaluate sensitivity by measuring how noise injection in each feature affects accuracy or feature importance rankings.

Q3 — Synthetic Data Ethics & Limitations

While oversampling solves class imbalance, excessive replication can introduce bias or unrealistic minority examples.

To ensure fidelity, techniques like SMOTE or GAN-based generation can be applied with cross-validation to preserve realistic data patterns.

Q4 — Real-World Applications

A practical pipeline for financial institutions could combine:

Synthetic data to balance and anonymize samples;

Differential privacy to protect sensitive attributes before sharing;

Model auditing to verify that no private information leaks through model outputs.

Such an approach aligns with data protection regulations like GDPR and HIPAA, enabling secure, compliant data analysis and collaboration across organizations.