Classification for Imbalanced Credit Card Transaction Data

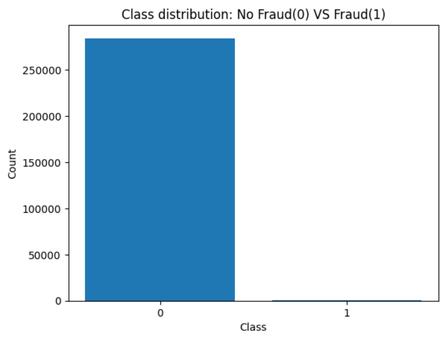
Wen Sun, Dan Li (Faculty Sponsor)
Department of Computer Science & Electrical Engineering

1. Introduction

- It is important for the credit card companies to identify fraudulent transactions to avoid financial loss for the customers and the companies.
- The challenge of fraud detection lies in the imbalanced feature of transaction data which makes traditional classification algorithms infeasible.
- This research investigates the methodologies that are commonly employed to deal with imbalanced datasets. Specifically, over-sampling, under-sampling, and Synthetic Minority Over-sample Technique (SMOTE) are studied.

2. Dataset

- This dataset [1] contains two days of credit card transactions in September 2013 by European cardholders.
- The total number of data records is 284, 807.
- Imbalanced dataset; the positive transactions (frauds) count for 0.17% of all records.

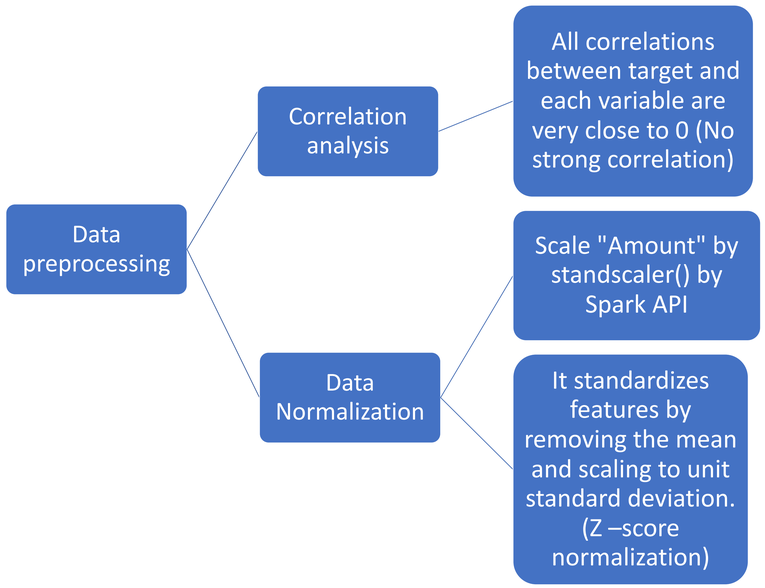


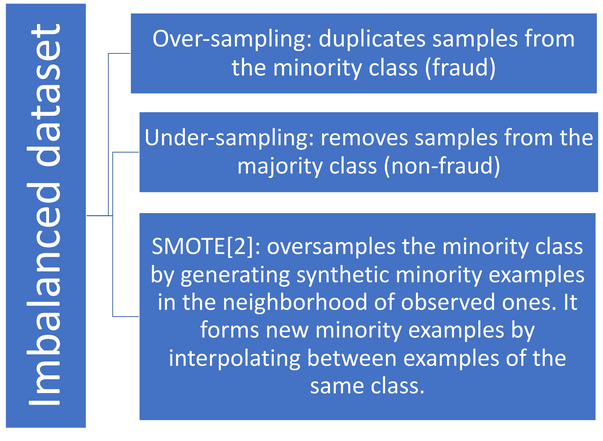
Dataset attributes:

- Time
- Amount
- V1, V2, …, V28: PCA features (confidential information)
- Class (1: fraud; 0: non-fraud)

3. Methods

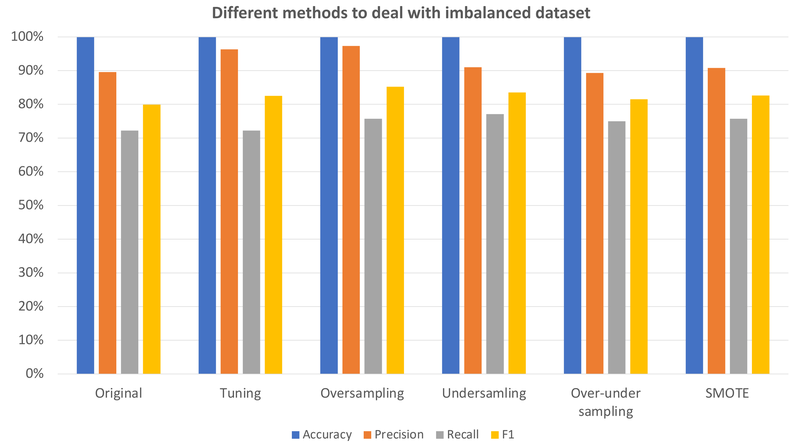
- To solve data imbalanced problems, over-sampling, under-sampling, and Synthetic Minority Over-sample Technique (SMOTE) are studied.





4. Results

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN}$$
$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$
$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$



5. Conclusions

- Over-sampling performs better than other methods, with the accuracy, precision, and f1 score being the highest.
- The recall scores of three methods have been improved a bit, meaning it helps to detect the “fraud” transactions.
- Due to the large imbalanced ratio of this experiment, it is hard to find the ”perfect” over-sampling/under-sampling ratio.

6. Future direction

- Do an outlier removal on our oversampling dataset and see if our test performance will improve.
- Use different models to train the dataset to see if it will improve the performance, e.g., neutral network…[3]

7. References Cited

[1] Andrea, Machine Learning Group-ULB. 2018. Credit Card Fraud Detection, Version 1. Retrieved January 20,2024 from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

[2] Jason Brownlee. Smote for imbalanced classification with python, 2021.

[3] John O. Awoyemi, Adebayo O. Adetunmbi, and Samuel A. Oluwadare. Credit card fraud detection using machine learning techniques: A com- parative analysis. In 2017 International Conference on Computing Net- working and Informatics (ICCNI), pages 1–9, 2017.