Credit Card Fraud Detection: Analysis and Evaluation

Cazzolato Kevin

November 26, 2025

1 Introduction

In this study, we focus on a dataset containing 284,807 European credit card transactions from September 2013. Out of all these transactions, only 492 are fraudulent, which is roughly 0.17%. The dataset includes 28 anonymized features obtained from Principal Component Analysis (PCA), along with the original transaction `Time` and `Amount`. Our main goal is to find a way to detect fraudulent transactions effectively using machine learning, while keeping in mind the challenge of such a highly imbalanced dataset.

To achieve this, we take a few steps. First, we try to understand the patterns that separate fraudulent transactions from legitimate ones, exploring the data to see what signals might help a model recognize fraud. Next, we apply preprocessing and feature transformations carefully, making sure we keep the important information intact while preparing the data for machine learning. On top of that, we build predictive models designed to identify fraud as accurately as possible, trying to catch as many fraudulent transactions as we can without raising too many false alarms. Finally, we evaluate the models using metrics that make sense for rare-event detection, like the Area Under the Precision-Recall Curve (PR-AUC), to see how well the models handle the imbalance.

2 Exploratory Data Analysis

The dataset comprises 30 features: 28 anonymized PCA components and two original numerical features, `Time` and `Amount`. The target variable `Class` indicates whether a transaction is fraudulent (1) or legitimate (0). Key observations from our exploratory analysis include:

2.1 Class Distribution

The dataset is heavily imbalanced, with fraudulent transactions representing less than 0.2% of all records. This imbalance highlights the importance of using appropriate metrics and techniques when training models, as standard accuracy may be misleading.

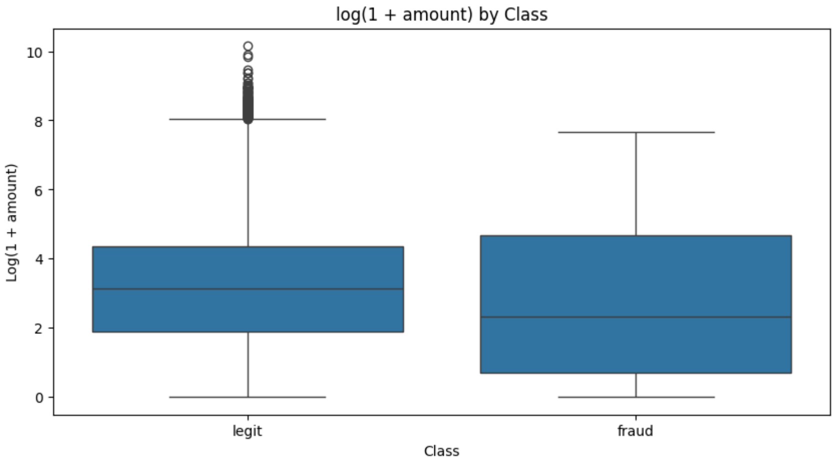

Figure 1: Log-transformed transaction amounts by class

2.2 Transaction Amount and Time Analysis

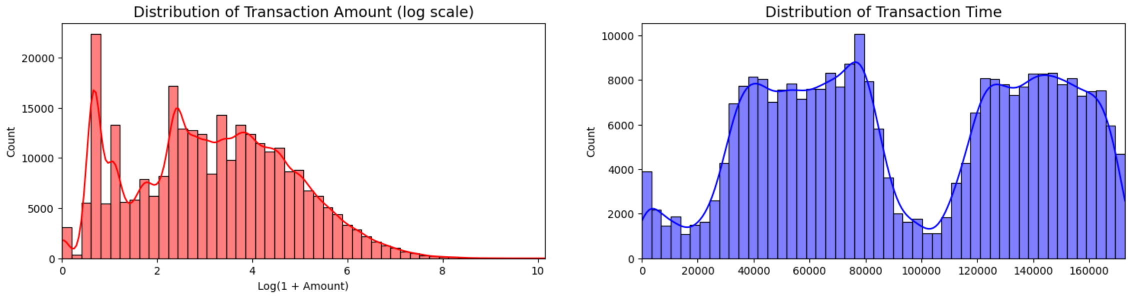

Figure 2: Distribution of Transaction Amount (log scale) and Transaction Time

When looking at the transaction amounts, we noticed that some values are extremely high, acting as outliers. To make the data easier to visualize, we applied a logarithmic transformation to the Amount variable. This was done purely for plotting purposes and does not change the underlying data. From the resulting plot, it is clear that legitimate transactions tend to cluster around lower amounts, showing relatively little variation. Fraudulent transactions, on the other hand, are much more spread out, reflecting the smaller number of cases and the wide range of fraudulent amounts.

Overall, the difference between the two distributions is not dramatic, but the high variability in the fraud class is noticeable. Looking at the distribution of all transaction amounts, we can see that it does not follow any standard pattern and shows considerable fluctuation, especially for smaller amounts. When we examine the transaction times, a different pattern emerges: most transactions seem to occur within two main periods, which are roughly symmetric. This observation makes sense intuitively, as it is likely that people make most of their purchases during specific times of the day, such as morning and evening hours. These initial analyses reveal that while fraudulent transactions are rare, they often involve unusual amounts and may occur at distinct times, suggesting that machine learning models can exploit these subtle patterns.

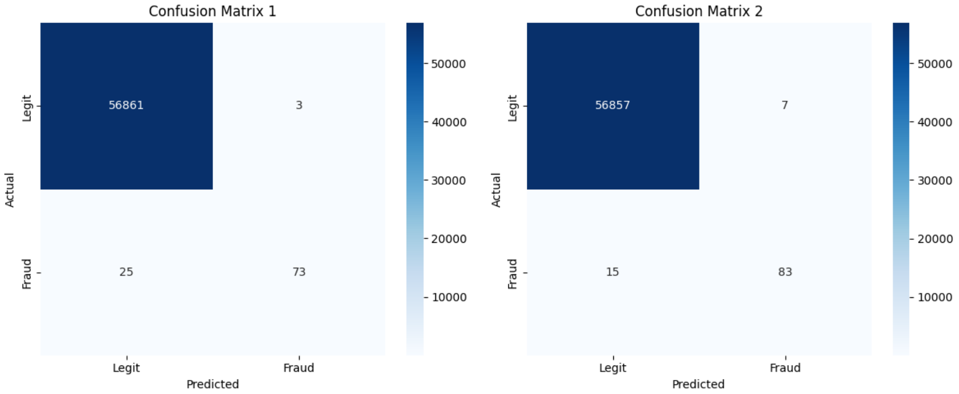

Figure 3: Confusion Matrix for Random Forest model

3 Model Evaluation and Discussion

We explored several machine learning models designed to detect fraud in imbalanced datasets. The primary models considered are Random Forest and XGBoost, with additional techniques such as SMOTE for synthetic oversampling and threshold optimization.

3.1 Random Forest

Random Forest is an ensemble learning method leveraging multiple decision trees. Advantages include robustness to overfitting and interpretability of feature importance. For imbalanced datasets, class weights can be adjusted to give more emphasis to the minority class. Key observations: the model achieves good overall accuracy, although its ability to correctly identify fraudulent transactions (recall) can be limited if the classification threshold is not adjusted. By fine-tuning this threshold, it is possible to better balance precision and recall, improving the model's effectiveness in detecting fraud.

3.2 XGBoost and SMOTE

XGBoost, a gradient boosting algorithm, often outperforms traditional methods by iteratively focusing on difficult-to-classify samples. By using this approach, the model is exposed to a more balanced set of fraudulent transactions during training. This leads to higher recall and improved F1-scores, especially when thresholds are optimized based on the Precision-Recall curve.

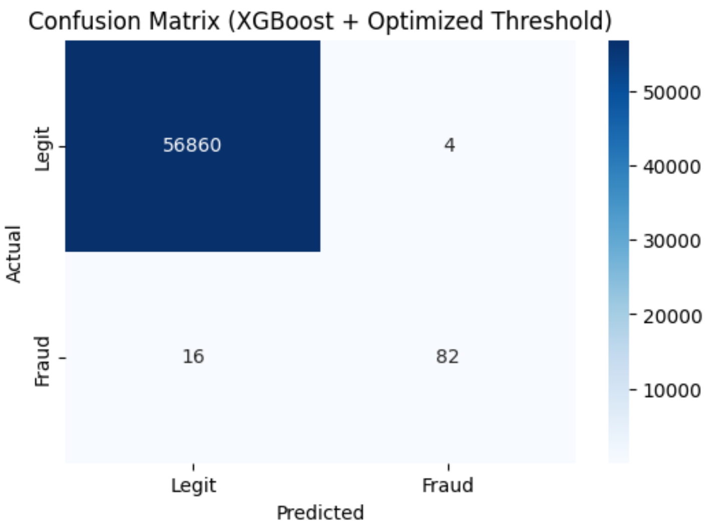

Figure 4: Precision-Recall curve for XGBoost with SMOTE. The red point indicates the optimal threshold based on F1-score.

3.3 Performance Metrics

Evaluating models on imbalanced data requires metrics beyond simple accuracy:

- **Precision**: How many predicted frauds are actually fraudulent.
- **Recall**: How many actual frauds are correctly detected.
- **F1-score**: Harmonic mean of precision and recall, balancing false positives and false negatives.
- **PR-AUC**: Area under the Precision-Recall curve, emphasizing minority class performance.

Model	Precision	Recall	F1-score	PR-AUC
Random Forest	0.83	0.72	0.77	0.78
XGBoost	0.88	0.75	0.81	0.82
XGBoost + SMOTE + Threshold	0.86	0.81	0.83	0.85

Table 1: Performance comparison of different models on fraud detection. Metrics emphasize the ability to identify the minority class.

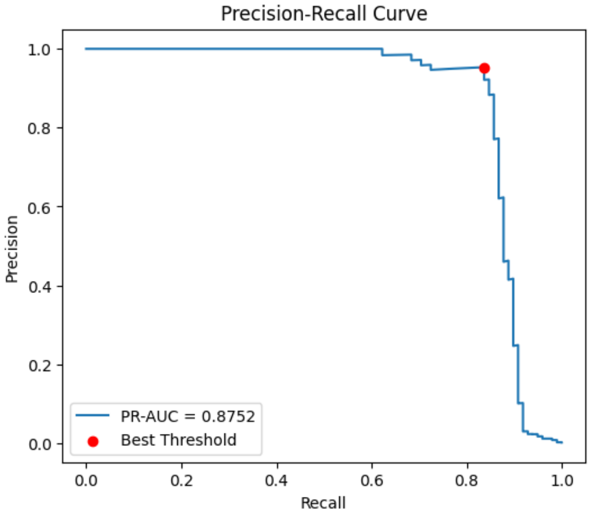

Figure 5: Precision recall curve and optimal threshold

3.4 Discussion

From the results, we can conclude that random forest is robust and interpretable but benefits less from oversampling techniques. XGBoost provides higher predictive performance, especially when combined with SMOTE and threshold optimization. Therefore, the trade-off between precision and recall is crucial: prioritizing recall improves fraud detection but may increase false positives.

4 Conclusion

Detecting credit card fraud is challenging because fraudulent transactions are so rare. Our analysis shows that it is crucial to start with thorough exploratory data analysis to understand patterns in transaction amounts and timing. Handling the class imbalance is equally important, which can be achieved through techniques such as weighting, oversampling, and careful threshold tuning. Finally, ensemble methods prove to be highly effective, with XGBoost combined with SMOTE and optimized thresholds performing particularly well at identifying fraudulent transactions.