Proceedings of ICFTBA 2026 Symposium: Strategic Management & Business Analytics: Data-Driven Performance Models
DOI: 10.54254/2754-1169/2026.LD36536

Credit Card Fraud Detection Using Machine Learning and Risk-Based Alert Strategies

Ziyue Meng

School of Science, Monash University Malaysia, Bandar Sunway, Malaysia
zmen0034@student.monash.edu

Abstract. Credit card fraud has become a central issue for the financial security of banks and cardholders. However, fraudulent transactions account for only a very small proportion of all transaction cases. If it is missed, it may directly cause huge property damage. As a result, this study focuses on how to identify fraudulent transactions as much as possible, while controlling false positives and human audit costs. Specifically, logistic regression, gradient boosting, XGBoost, and random forest are compared, while class weighting, SMOTE, and Random Undersampling are evaluated for handling class imbalance. It is more critical to translate the model results into a risk warning system and a loss simulation system. The results show that random forest with class weighting has the best overall performance, with 90.6% F1-Score, 94.1% precision, and 87.3% recall. Meanwhile, under the cost assumptions, this model reduces simulation costs by 85.4% relative to the no-model baseline. These findings suggest that in data environments with severe class imbalances, a machine learning model can be designed as an effective triage tool. Its actual value is not only to predict fraud, but also to establish a low-risk release, medium-risk verification, and high-risk manual review of the decision-making process for institutions and truly reduce losses.

Keywords: Credit card fraud detection, Machine learning, Imbalanced classification, Random Forest, Risk-based alert system

1. Introduction

Credit card fraud has been a major concern in financial services, highlighting the need for effective fraud detection methods [1]. Credit card fraud detection not only reduces money loss but also maintains customer trust. However, because transactions account for a very low proportion of all transactions, the model is susceptible to the class imbalance problem [1]. In a real scene where the fraud rate is only 0.1%, a system that blindly guesses that all transactions are normal can also achieve a falsely high accuracy rate of 99.9%, but all frauds are missed. Therefore, evaluation should not only focus on overall correctness but also on the model's ability to identify fraudulent transactions.

To solve the recognition difficulties caused by the extreme imbalance of categories, Alarfaj et al. compared the performance of a variety of machine learning and deep learning algorithms in the credit card fraud detection task, indicating that there may be significant differences in the

DOI: 10.54254/2754-1169/2026.LD36536

recognition ability of different models in the same kind of fraud detection problem. Therefore, the model selection itself is an important link in the research of credit card fraud detection, which also provides the basis for this paper to compare the models of Logistic Regression, Random Forest, XGBoost, and Gradient Boosting [2, 3]. Saito and Rehmsmeier emphasized that Precision-Recall analysis is useful for imbalanced binary classification because it focuses on the performance of the positive class [4]. Moreover, so far, many studies have focused on comparing model metrics, and few have paid attention to translating risk probabilities into audit-level and cost decisions.

Therefore, this paper will further analyze how to transform the model into a system that can help the organization optimize decision-making. To build the system, this study compared a variety of models and imbalance treatment methods. After obtaining the fraud probability of the optimal model, the transaction is divided into different risk levels. In addition, a cost simulation system will be established to test whether the proposed early warning system can really reduce the potential fraud-related costs.

2. Methodology

2.1. Dataset and data quality

This study uses a public credit card transaction dataset from Kaggle to examine how machine learning models can detect fraudulent transactions and support risk-based alert decisions [5]. The dataset contains 284,807 transactions over 2 days. For modelling, the study uses a stratified sample of 160,000 transactions. This reduces the running time and preserves the class distribution. Additionally, since the original features have been handled with PCA, the main independent variables are anonymized and represented as V1 to V28, while the main dependent variable is class, which indicates whether the transaction is fraudulent or legitimate. The dataset contains no missing values. Although 1081 repeated rows were found, they are not removed because the dataset does not include transaction IDs, so it is unclear whether these are duplicate rows or repeated real transactions.

2.2. Data split and model evaluation

Before analysis, a data split process was conducted. In this study, the ratio of 60:20:20 was used to divide the dataset. 60% of them are used as the training set for model training; 20% as a validation set for model tuning and threshold selection; The remaining 20% is used as the test set for the final model performance evaluation. As the proportion of fraudulent transactions is very low, stratified sampling should be used in the division to ensure that the proportion of fraud in the training set, verification set, and test set is close to the original data distribution.

The confusion matrix is one of the most common methods for evaluating machine learning classifiers. This study uses a confusion matrix to obtain the values of Precision, Recall, F1-score, and PR-AUC. Based on the above data sets and unified evaluation criteria, this paper selects a variety of machine learning models for comparison and optimizes them with the category imbalance strategy.

2.3. Models and imbalanced strategies

To compare the performance of different methods in the part of credit card detection, this study employs different Machine Learning models. To begin with, Logistic Regression is used as a

DOI: 10.54254/2754-1169/2026.LD36536

baseline model because it is simple and widely used in binary classification problems. Second, the class weighting approach used in this paper addresses class imbalances and allows models to be trained to pay more attention to fraudulent transactions as a minority. Moreover, resampling methods are used, including SMOTE and Random Undersampling, which are used to change the proportion of categories in the training data, with both methods configured to balance samples at a 1:1 ratio [6]. In terms of model selection, this study further compared tree models such as Random Forest, Gradient Boosting, and XGBoost to examine whether non-linear models can have better performance in identifying fraudulent transactions. These models represent the baseline model, the category imbalance treatment, and the complex classification model. Therefore, it is feasible to conduct a comprehensive comparison of the performance across various modeling strategies.

2.4. Risk tier design and cost simulation

This study set three different risk levels based on the model's probability. These levels are used to support different fraud control actions. Thus, two thresholds were generated to divide the probability into three parts corresponding to three risk levels. The high-risk threshold was set to the highest F1-score in the validation set, and the low-risk threshold was also chosen to distinguish low-risk and medium-risk transactions. Furthermore, to make the model performance more interpretable, a cost simulation was also built to estimate the potential financial impact of the model. In this simulation, a missed fraud case was assigned a hypothetical cost of 500 dollars. This represents possible reimbursement, chargeback, and investigation expenses. A correctly detected fraud alert was assigned a simulated cost of 10 dollars for staff review or identity verification. A false alert included the same 10-dollar assumed review cost and an additional 5 dollars for customer delay or support. Therefore, a false alert had a total hypothetical cost of 15 dollars. A legitimate transaction that was correctly approved had no additional cost.

3. Results

3.1. Data characteristics

The exploratory data analysis confirms two important data characteristics.

Table 1. Distribution of transaction amount by class

Metric	Value
Rows	284,807
Columns	31
Missing Values	0
Exact Duplicated Row	1,081
Fraud Cases	492
Fraud Rate	0.173%

First, as seen in Table 1, the fraud cases are extremely rare, with only a 0.173% fraud rate in the original dataset. This confirms that the model should focus on the minority class.

DOI: 10.54254/2754-1169/2026.LD36536

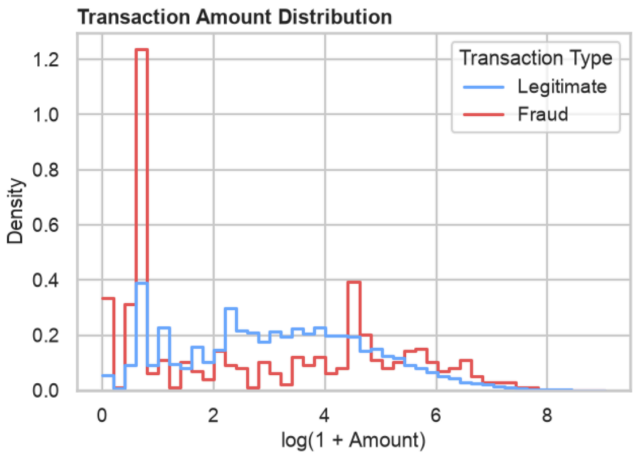

Figure 1. Distribution of log-transformed transaction amounts by transaction class (photo/picture credit: original)

Second, Figure 1 shows the density distribution of legitimate transactions and fraudulent transactions in terms of transaction amounts. In this graph, the red line represents fraudulent transactions, and the blue line represents legitimate transactions. It is worth noting that in Figure 1, the transaction amount after log processing is used. The purpose of this processing is to make the distribution of transaction amounts easier to observe and model [7]. To analyze this plot, it is obvious to see that the amount distribution of the two types of transactions has a centralized range, but the whole amount segment is distributed. Thus, it is not possible to distinguish between fraud and normal transactions by the transaction amount.

3.2. Model comparison

Table 2. Test-set performance comparison of six representative models using Precision, Recall, F1-score, and PR-AUC

	Experiment	Precision	Recall	F1	PR-AUC
0	Random Forest Class Weight	94.1%	87.3%	90.6%	88.4%
1	XGBoost weight	92.2%	85.5%	88.7%	88.2%
2	SMOTE + XGBoost	90.2%	83.6%	86.8%	85.2%
3	LogReg Class Weight	88.2%	81.8%	84.9%	80.9%
4	UnderSample + Random Forest	83.9%	85.5%	84.7%	84.2%
5	Gradient Boosting	47.1%	29.1%	36.0%	21.5%

Given that credit card fraud is an extremely unbalanced binary classification problem and a single model cannot deal with fraud samples well, this study compared a variety of machine learning algorithms and imbalance processing methods (Table 2) [8]. In this paper, there are six representative models/model combinations, including Random Forest with class weight, XGBoost weighted, SMOTE + XGBoost, Logistic Regression with class weight, UnderSample + Random Forest, and Gradient Boosting.

DOI: 10.54254/2754-1169/2026.LD36536

The model mainly uses Precision, Recall, F1 score, and PR-AUC for comparison, where Precision measures the accuracy of alerts, Recall shows fraud recognition coverage, and F1 score reflects the balance between the two.

3.3. Selected random forest performance

Figure 2 shows the performance of various models/model combinations in the test set.

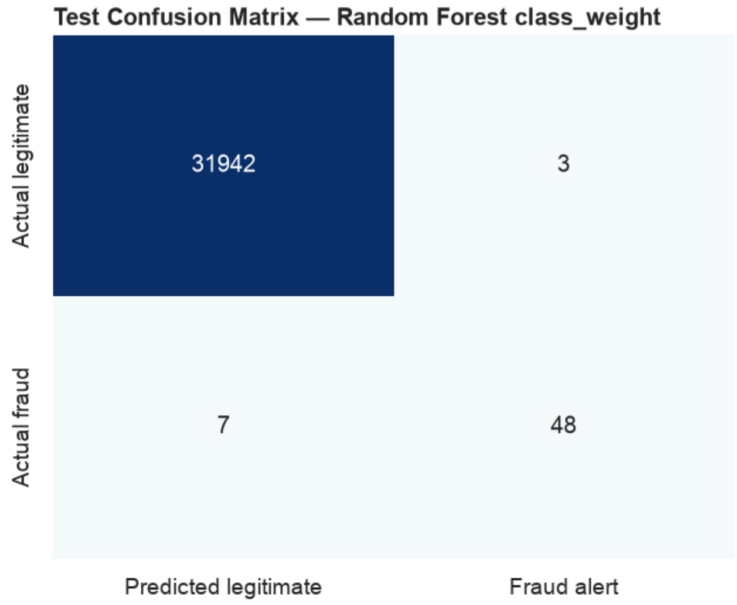

Figure 2. Test-set confusion matrix of the selected Random Forest model (photo/picture credit: original)

Based on the model comparison in Section 3.2, the study has chosen Random Forest as the subsequent analysis model for the following analysis. Fig 2 presents the results of the Random Forest on the test set. The model correctly detected 48 fraudulent transactions, missed 7 fraud cases, falsely flagged 3 legitimate transactions, and correctly approved 31,942 legitimate transactions.

3.4. Risk tiers and cost simulations

After obtaining the fraud probability output of Random Forest, this study further divides transactions into different risk levels, rather than simply processing the results into 0/1 classifications. The reason for doing so is that in actual financial risk control scenarios, transactions with different levels of risk are usually not treated in the same way. Therefore, this study divides transactions into three categories: low-risk, medium-risk, and high-risk, corresponding to automatic release, secondary verification, and manual review, respectively. Among them, the high-risk threshold adopts the threshold corresponding to the optimal F1 score in the validation set, while the medium-risk boundary is set as a business rule to distinguish transactions that require additional validation but have not reached the level of manual review.

DOI: 10.54254/2754-1169/2026.LD36536

Table 3. Risk-tier results based on the Random Forest fraud risk score

	Risk Tier	Risk-Score Range	Transactions	Fraud Cases	Fraud Rate	Action
0	Low Risk	< 0.1	31375	4	0.013%	Approve automatically
1	Medium Risk	0.1 to <0.462	574	3	0.523%	Set-up verification
2	High Risk	>=0.462	51	48	94.1%	Manual review

Table 3 summarizes the distribution of transactions and fraud cases across the three risk tiers. The results showed that the low-risk layer contained 31,375 transactions, of which only 4 were fraudulent transactions, indicating a low fraud rate and therefore suitable as automatic release transactions. The medium-risk layer contains 574 transactions, including 3 fraudulent transactions, indicating that the risk of this layer is higher than that of the low-risk layer, but it is still not suitable for all to enter manual review, so secondary verification is more suitable. The high-risk layer only includes 51 transactions, but 48 of them are genuine fraudulent transactions, with a fraud rate of 94.1%. This indicates that the model can concentrate most fraudulent transactions into a small set of high-risk transactions.

To further illustrate the potential application value of this risk stratification strategy, this study established a simple simulation cost model. It should be noted that this cost analysis is based on assumed parameters and aims to demonstrate the cost changes that the model may bring under different processing strategies, rather than representing the actual returns of real banks. Based on the cost assumptions listed in section 2.5, all fraudulent transactions are considered underreported, resulting in a simulation cost of $ 27,500. After adopting the risk warning strategy, the simulation cost decreased to $4025. This corresponds to a simulated cost reduction of approximately 85.4%, indicating that the model not only has classification significance but can also be transformed into an executable risk management process.

Therefore, the ultimate result of this study is not simply to find the best-performing classification model, but to construct a risk stratification and cost perception decision-making framework based on the probability output of the model.

4. Discussion

4.1. Interpretation of model performance

Random Forest with class weight is suitable for this task because it cannot only capture complex nonlinear fraud patterns, but also make the model pay more attention to a few types of fraudulent transactions through class weight [9]. At the same time, it does not need to generate composite samples, so it may be more stable than SMOTE in PCA anonymized feature space.

4.2. Practical implications of risk tiering

The value of the model is not only to judge whether a transaction is fraudulent, but also to help financial institutions decide how the transaction should be handled. For low-risk transactions, the system can automatically release, thus reducing the disturbance to normal users and unnecessary audit costs. For medium-risk transactions, the system can require secondary verification, such as an

DOI: 10.54254/2754-1169/2026.LD36536

SMS verification code, APP confirmation, or additional authentication. This secondary verification is gentler than the direct manual audit, which can not only increase security, but also bring too much pressure to the operation Department. For high-risk transactions, the system can give priority to manual review, because these transactions are most likely to be true fraud. This allows the limited manual review resources to focus on the transactions that are most worth checking, rather than being evenly distributed to all transactions. Therefore, risk stratification is solving a resource allocation problem: which transactions can pass automatically, which need additional verification, and which must be manually involved.

4.3. Limitations and future work

Although the models in this study perform well, there are still many limitations [1]. First, the dataset used is processed by PCA, so the features will not be seen, so it is unknown what behavior they correspond to in the real business [10]. Secondly, the data is only about 48 hours, so it cannot study the long-term changes, nor can the research judge the impact of holidays, business cycles, and user behavior changes on fraud [11]. In addition, in the process of cost assumption, this study does not use the real data of the bank, so it can only explain the mechanism, not the real profit.

Future research can use data sets containing original business variables to further explain the relationship between different characteristics and real trading behavior. For example, if the data includes user age, consumption location, transaction category, equipment information, or credit score, the model results can provide more specific business explanations. Future research can also use a longer time range of transaction data, rather than just 48 hours of transaction records. In this way, it can observe whether fraud behavior will change with changes in holidays, business cycles, or users' consumption habits. In terms of cost analysis, future research can introduce the real operating cost data of banks or payment platforms. For instance, it can use real refund costs, manual audit costs, customer complaint costs, and service costs caused by false positives. In this way, the cost simulation results can be closer to the real business value, rather than just explaining the model mechanism. Finally, future research can also regularly update the model to deal with changes in fraud strategies and user behavior. This can improve the stability and practical usability of the model in long-term application.

5. Conclusion

In conclusion, this study focuses on the problem of class imbalance in credit card fraud detection and discusses how the machine learning model can identify a few fraudulent transactions and support risk early warning decisions. The research results show that under a seriously unbalanced data environment, accuracy is easy to mask the model's recognition ability for fraudulent transactions. Therefore, this study mainly uses precision, recall, F1 score, and PR-AUC for evaluation. By comparing various models and imbalance treatment methods, random forest with class weighting achieved the best comprehensive performance in this study. The model can identify most fraudulent transactions and control the number of false positives, so it has good application value. More importantly, this study converts the fraud probability output from the model into three processing levels of low risk, medium risk, and high risk, corresponding to automatic release, secondary verification, and manual review, respectively. The cost simulation results show that under given assumptions, the risk early warning process can reduce the processing cost of potential fraud.

Therefore, the main contribution of this study is to transform the classification results of machine learning into a decision-making framework closer to the actual financial risk control scenario.

DOI: 10.54254/2754-1169/2026.LD36536

However, this study is still limited by the anonymization characteristics of PCA, the shorter data time range, and the assumed cost parameters. Future research can use a longer time range, more business variables, and real operating cost data to improve the interpretability and practical application value of the model.

References

[1] Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. *arXiv preprint arXiv:2010.06479*. https://arxiv.org/abs/2010.06479

[2] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms. *IEEE Access, 10*, 39700-39715.

[3] Gupta, P., Varshney, A., Khan, M. R., Ahmed, R., Shuaib, M., & Alam, S. (2023). Unbalanced credit card fraud detection data: A machine learning-oriented comparative study of balancing techniques. *Procedia Computer Science, 218*, 2575-2584.

[4] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), Article e0118432.

[5] Machine Learning Group - ULB. (2018). Credit card fraud detection. *Kaggle*. Retrieved July 13, 2026, from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357.

[7] Lakshmi, N. V., Danish, F., & Alrasheedi, M. (2025). Enhanced estimation of finite population mean via power and log-transformed ratio estimators using an auxiliary variable in solar radiation data. *Journal of Radiation Research and Applied Sciences, 18*, Article 101379.

[8] Ileberi, E., Sun, Y., & Wang, Z. (2021). Performance evaluation of machine learning methods for credit card fraud detection using SMOTE and AdaBoost. *IEEE Access, 9*, 165286-165294.

[9] Kulatilleke, G. K. (2022). Credit card fraud detection - Classifier selection strategy. arXiv preprint arXiv:2208.11900. https://arxiv.org/abs/2208.11900

[10] Salekshahrezaee, Z., Leevy, J. L., & Khoshgoftaar, T. M. (2023). The effect of feature extraction and data sampling on credit card fraud detection. *Journal of Big Data, 10*(1).

[11] Hafez, I. Y., Hafez, A. Y., Saleh, A. M. S., Abd El-Mageed, A. A., & Abohany, A. A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data, 12*(1).