Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.39306

Construction and Evaluation of Credit Card Fraud Detection Model Based on Logistic Regression — A Case Study of Kaggle Dataset

Huiwen Liu

Beijing Haidian Kaiwen Academy, Beijing, China
judyliuvip@163.com

Abstract. The widespread use of online payment has led to frequent credit card fraud incidents, causing significant economic losses to financial institutions and users. Efficient fraud detection models are the focus of research in financial risk control. Based on the publicly available credit card transaction dataset from Kaggle, this paper constructs a logistic regression fraud detection model. The study involves data exploration of the dataset, handling of imbalanced data, feature selection, as well as model training and tuning. This study uses stratified sampling, StandardScaler normalization, and cross-validation to enhance data processing and model reliability. Model evaluation is conducted using confusion matrices, precision, recall rate, and AUC values. Experimental results show that the model achieves high recall but low precision under SMOTE oversampling. This model achieved an AUC-ROC of 0.9708 for the fraud category, with a recall rate of 0.92, but the accuracy rate of this model is only 0.06, which indicating a significant trade-off between detecting fraud and controlling false positives.

Keywords: Credit Card Fraud Detection, Logistic Regression, Class Imbalance, SMOTE, ROC-AUC

1. Introduction

The rapid development of internet finance has led to the continuous expansion of the scale of online credit card transactions. At the same time, the methods of credit card fraud have also been constantly updated and improved. Various fraudulent transactions not only cause property losses to users but also increase the risk control pressure on financial institutions. As a lightweight and highly interpretable linear model, logistic regression still warrants further research on its actual performance and optimization potential in the credit card fraud detection scenario. Many existing studies have explored logistic regression for credit-card fraud detection tasks [1, 2]. In response to the practical problems of class imbalance in the credit card dataset and the failure of the model to detect fraudulent samples, this paper will address the following issues: First, analyze the data distribution characteristics of the Kaggle credit card dataset [3] and explore the impact of class imbalance on the classification task; Second, complete data cleaning, balance processing, and feature processing, build a logistic regression fraud detection model and perform parameter tuning;

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.39306

Third, use multiple evaluation indicators to conduct a comprehensive assessment of the model performance, and analyze the advantages and shortcomings of logistic regression in the fraud identification task. This paper uses a quantitative research method, and uses the publicly available credit card transaction dataset from Kaggle as the experimental basis. This paper constructs and evaluates a credit card fraud detection model based on logistic regression, revealing the trade-off relationships among the performance indicators of the model under the condition of class imbalance. It provides empirical references for the selection and optimization of fraud detection models in financial risk control.

2. Dataset and data preprocessing

2.1. Introduction of the Kaggle credit card dataset

This experiment uses the public credit card fraud detection dataset available on the Kaggle platform [1]. This dataset records the credit card transaction records of European cardholders within two days in September 2013. The original dataset consists of 284,807 transaction samples, among which there are 492 fraudulent transactions, accounting for only 0.172% of the total. It exhibits an extreme class imbalance characteristic.

The dataset contains a total of 31 variables. Due to user privacy protection, most of the original business features have been processed through PCA (Principal Component Analysis) transformation, resulting in 28 numerical features from V1 to V28, which no longer have direct business meanings. The fields that have not undergone PCA transformation include time and amount: time represents the time interval between this transaction and the first transaction in the dataset, measured in seconds; amount represents the transaction amount. The target label is class, where 0 represents a normal transaction and 1 represents a fraudulent transaction.

2.2. Data exploratory analysis

After confirming the basic structure of the dataset, this paper will conduct exploratory data analysis (EDA), which focuses on examining the distribution of the target variable, the types of variables, and the basic differences between different transaction categories.

As shown in Figure 1 below, the number of normal transactions is much higher than the fraudulent transactions. If the original data is directly used to train the classifier, even if the model will predict most of the samples as normal transactions, it may still achieve a relatively high overall accuracy. However, this result does not suggest that the model has a good detection ability for fraud. Therefore, this paper does not only rely on accuracy, but focuses on evaluating indicators such as precision, recall, F1, and ROC-AUC.

In addition, the numerical scales of amount and time differ from the PCA features. To avoid unnecessary influence of large-scale variables during model optimization, the time and amount will be standardized in the preprocessing stage.

DOI: 10.54254/2753-8818/2026.39306

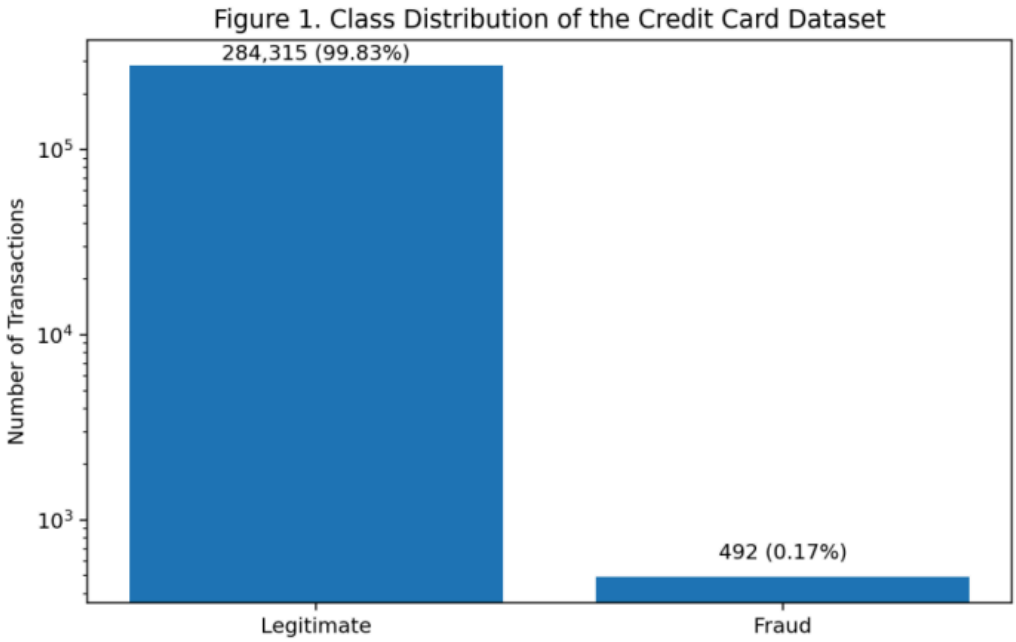

Figure 1. Class distribution of the credit card dataset

2.3. Missing value and outlier handling

First, an integrity check is conducted on the dataset, including missing values, duplicate records, and abnormal ranges of numerical variables. The description of the dataset provided publicly indicates that all 31 fields are of numerical type and no large-scale missing value issues have been reported. Therefore, in this study, the mean, median, or mode filling method is not employed to avoid artificially altering the original transaction distribution.

For outliers, this study distinguishes them from missing values. Extreme transaction amounts in credit card fraud detection are not necessarily data errors because real transactions may have high consumption; similarly, large absolute values in V1--V28 after PCA transformation cannot simply be regarded as invalid records.

Before model training, time and amount are standardized using StandardScaler to make different variables have similar scales.

2.4. Handling of imbalanced dataset classes

Class imbalance is the most significant data processing issue in this study. There were only 492 fraudulent transactions in the original data, while there were 284,315 normal transactions. The proportion of fraudulent samples was approximately 0.172%. If the model were trained directly according to the original ratio, the classifier might tend towards the majority class, resulting in high accuracy but low recall.

To alleviate this problem, this paper uses SMOTE (Synthetic Minority Over-sampling Technique) to oversample the minority class in the training set. The basic idea of SMOTE is to generate new synthetic samples by leveraging the proximity relationships among minority class samples, thereby expanding the representation proportion of the minority class in the training data [4].

According to the research design of this paper, SMOTE can only be applied to the training set and cannot be used to process all the data before dividing the dataset. Otherwise, the information of

DOI: 10.54254/2753-8818/2026.39306

the synthetic samples might indirectly enter the test set, causing data leakage. The research of Dal Pozzolo et al. also indicates that the sampling method in imbalanced learning can affect probability estimation [5].

Therefore, the processing flow of this paper is: first, divide the training set and test set; second, apply SMOTE only to the training set; and finally, use the balanced training set to train the logistic regression model and conduct the final evaluation on the test data that has not been processed by SMOTE.

2.5. Dataset splitting and feature processing

To ensure the comparability of experimental results, stratified sampling is adopted in this paper to split the training and test sets, so that the proportions of legitimate and fraudulent transactions in both subsets remain roughly consistent. Given that there are only 492 fraudulent samples, stratified splitting can reduce the fluctuation of class proportions caused by random splitting. The training set is used for model training, while the test set is reserved solely for final performance evaluation.

In terms of feature processing, class, as the target variable, is excluded from the feature matrix. Time, amount, and V1–V28 constitute the input variables of the model. Since the scales of time and amount differ from those of the PCA-transformed variables, StandardScaler is applied to standardize them. Meanwhile, V1–V28 are not reprocessed with PCA, as these variables are already anonymized PCA features.

3. Fraud detection model construction

3.1. Model construction

This paper selects logistic regression as the model for credit card fraud detection. For binary classification problems, logistic regression converts the linear combination into a probability value between 0 and 1 through the Sigmoid function. Its basic form is:

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\ldots+\beta_kx_k)}} \quad (1)$$

When the predicted probability exceeds the set threshold, the model classifies the transaction as fraud; otherwise, it is classified as normal. Compared with complex ensemble learning models, logistic regression has a simple structure, lower training cost, and can explain the relationship between the variables and the fraud probability through regression coefficients. Therefore, it is suitable as a benchmark model for credit card fraud detection.

Model construction mainly consists of four steps. The first step is to perform standardization and SMOTE on the training data; the second step is to establish the Logistic Regression classifier; the third step is to select model hyperparameters through cross-validation or a validation set; the fourth step is to output the prediction labels and probabilities on the test set while maintaining the original class proportions. The model does not ultimately use Accuracy as the sole optimization objective, but focuses on observing the Recall, Precision, and F1 of the minority class.

3.2. Model parameter tuning

The main hyperparameters of logistic regression include the regularization strength C, the type of regularization, and the class weights. In this paper, C is taken as the main adjustment parameter, and

DOI: 10.54254/2753-8818/2026.39306

the stability of the model under different parameter settings is evaluated through cross-validation. A larger C indicates a weaker regularization constraint, and the model is more inclined to fit the training data; a smaller C results in stronger regularization, which helps to prevent overfitting.

Since the class imbalance has already been handled by SMOTE in this paper, the parameter tuning stage no longer simply adds overly strong class weights to avoid over-amplifying the weights of the minority class. In this experiment, the regularization hyperparameter is set to C=1, which represents moderate regularization strength for the logistic regression model.

It should be noted that there is no fixed optimal combination between SMOTE and the parameters of logistic regression. Existing studies have shown that different imbalance handling methods may improve the Recall of the minority class, but at the same time increase false alarms [5].

3.3. Model training process

After completing the parameter selection, the final logistic regression model is trained using the standardized and SMOTE-processed training data. During the training process, the model estimates the regression coefficients by maximizing the likelihood function and controls the model complexity through regularization.

4. Model experimental results and evaluation analysis

4.1. Analysis of confusion matrix results

The confusion matrix divides the prediction results into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Among them, TP represents the successful identification of fraudulent transactions, TN represents the correct identification of normal transactions, FP represents the normal transactions wrongly labeled as fraudulent, and FN represents the undetected fraudulent transactions.

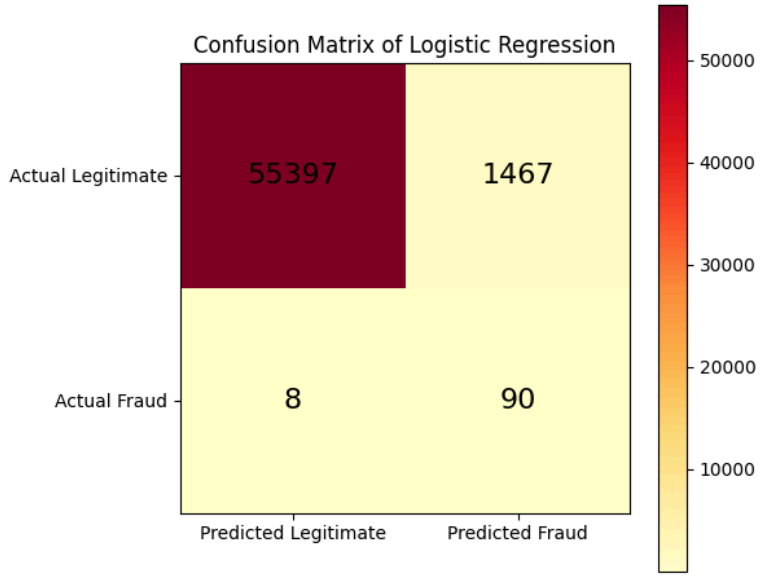

Figure 2. Confusion matrix of logistic regression

DOI: 10.54254/2753-8818/2026.39306

After training the logistic regression model with SMOTE oversampling only on the training set, the model was evaluated on the held-out test set. The confusion matrix (Figure 2) shows TN=55397, FP=1467, FN=8, TP=90. The model successfully detects 90 out of 98 real fraud transactions, with only 8 fraud cases missed. However, the model produces 1467 false positives: many normal transactions are incorrectly flagged as fraud. For financial scenarios, low false negative rate is critical to avoid economic loss from fraud; but excessive false positives will increase manual review workload for bank staff.After training the logistic regression model with SMOTE oversampling only on the training set, the model was evaluated on the held-out test set. The confusion matrix (Figure 2) shows TN=55397, FP=1467, FN=8, TP=90. The model successfully detects 90 out of 98 real fraud transactions, with only 8 fraud cases missed. However, the model produces 1467 false positives: many normal transactions are incorrectly flagged as fraud. For financial scenarios, low false negative rate is critical to avoid economic loss from fraud; but excessive false positives will increase manual review workload for bank staff.

4.2. Analysis of Precision, Recall and F1-score

Precision, Recall, and F1 evaluate the model from different perspectives. Precision represents the proportion of all transactions determined by the model as fraudulent that are actually fraudulent; Recall represents the proportion of all truly fraudulent transactions that are identified by the model; F1 is the harmonic average of Precision and Recall.

The classification report shows that for the fraud class (class=1), recall = 0.92, precision = 0.06, and F1-Score = 0.11. The high recall satisfies the business priority of discovering fraud as much as possible, but the precision is very low. This phenomenon is common when using SMOTE on extremely imbalanced datasets: the model becomes sensitive to fraud and tends to classify more samples as positive, which introduces many false alarms. The F1-score is low because of the huge gap between precision and recall. Overall accuracy reaches 0.97, but accuracy is misleading for this highly imbalanced dataset and cannot be used as the main evaluation metric.

4.3. Analysis of AUC-ROC results

The ROC curve has the False Positive Rate on the x-axis and the True Positive Rate on the y-axis, and is used to describe the model's ability to distinguish positive and negative samples as the classification threshold changes. The closer the ROC-AUC is to 1, the more typically it indicates that the model has a strong ability to distinguish under different thresholds.

As shown in Figure 3, the AUC-ROC of this logistic regression model is 0.9708, which demonstrates strong overall discriminative ability to separate legitimate and fraudulent transactions. However, ROC-AUC can be over-optimistic under extreme class imbalance. In further research, PR-AUC (Precision-Recall AUC) can be adopted as a supplementary metric to better reflect model performance on minority fraud samples.

It should be particularly pointed out that in a situation where the fraud rate is only about 0.172%, ROC-AUC may be more "optimistic" than Precision, Recall, and F1.

DOI: 10.54254/2753-8818/2026.39306

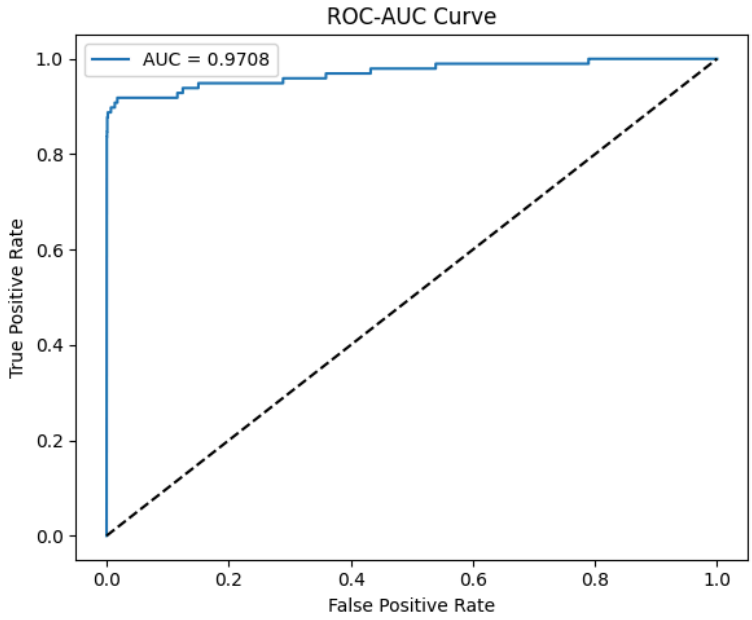

Figure 3. ROCAUC curve of logistic regression

4.4. Reflection

Although logistic regression can achieve a high ROC-AUC on this dataset and identify a considerable proportion of fraudulent transactions, the model still has significant limitations.

First, there is an extreme imbalance in the data categories. Fraud samples account for only about 0.172%, so the model evaluation is easily influenced by the category ratio. SMOTE can improve the category distribution during the training stage, but the generated samples are not real transactions and may change the local structure of the original data and affect probability calibration.

The model achieves high recall but very low precision. A large number of false positives will raise manual auditing costs in real financial systems. Although SMOTE helps reduce fraud omission, it does not fully solve the trade-off between false positives and false negatives. Besides, the threshold of the logistic regression is default 0.5; threshold tuning can be used in future work to adjust the balance between recall and precision.

Second, logistic regression essentially establishes a linear decision boundary. If there is a significant non-linear relationship or high-order feature interaction between normal transactions and fraudulent transactions, only relying on a linear model may not fully capture the complex fraud patterns.

Third, V1-V28 are all anonymized PCA variables, which give the model a certain degree of mathematical interpretability, but lack direct business explanations. The model cannot clearly answer "which merchant behavior" or "which transaction scenario" leads to an increase in risk, which limits the model's ability to transform into actual risk control rules.

Fourth, the dataset only covers the European cardholder transactions for two consecutive days in September 2013. Therefore, the experimental results cannot directly represent the real fraud patterns in other years, regions, or financial institutions.

In summary, the main advantages of logistic regression are simple structure, low computational cost, and relatively strong interpretability; the main problems lie in category imbalance, limited linear expression ability, insufficient business meaning of anonymous features, and limited

DOI: 10.54254/2753-8818/2026.39306

representativeness of the public dataset. Open-source community projects also provide reference experimental results on this public dataset [6]. Ensemble-based methods are alternative solutions for fraud detection tasks [7].

5. Conclusion

In conclusion, This paper is based on the publicly available credit card transaction dataset from Kaggle, and constructs a logistic regression model for credit card fraud detection. Due to the extreme class imbalance issue where fraud samples only account for 0.172% of the total records, this study only uses SMOTE oversampling on the training set to avoid data leakage. The research discards the overall accuracy rate that is prone to misleading and uses multiple indicators such as confusion matrix, precision rate, recall rate, F1 score, and ROC-AUC to conduct a comprehensive evaluation of the model.

The experimental results show that after SMOTE processing, the recall rate of this logistic regression model for fraud samples reaches 0.92, which can identify the vast majority of fraudulent transactions and reduce potential economic losses. However, the model's precision rate is only 0.06, resulting in a large number of false alarms, which will bring heavy manual review work to financial institutions. The ROC-AUC value of this model is 0.9708, and the overall discrimination performance is good. However, in the condition of severe class imbalance, the evaluation results of ROC-AUC are often overly optimistic. As a lightweight linear benchmark model, logistic regression has low computational cost and can achieve good mathematical interpretability through regression coefficients.

This study also has some limitations: the features de-sensitized by PCA lack practical business meaning, and the geographical and time coverage of the dataset is also very limited. Future research can adjust the classification threshold, introduce PR-AUC as a supplementary evaluation indicator, and adopt ensemble learning algorithms to balance recall rate and precision rate, so that the model can be better applied to real risk control scenarios.

References

[1] Alenzi, H. Z., & Aljehane, N. O. (2020). Fraud detection in credit cards using logistic regression. *International Journal of Advanced Computer Science and Applications, 11*(12), 484–490.

[2] Priya, M. J., & Ahammad, H. (2025). Balancing interpretability and performance in credit card fraud detection using machine learning techniques with SMOTE and logistic regression. In *2025 IEEE International Conference on Future Machine Learning and Data Science* (pp. 743–748). IEEE.

[3] Kaggle. (2018). Credit card fraud detection dataset [Data set]. *Kaggle*.

[4] 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.

[5] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). IEEE.

[6] Amakaamazu. (2026). Fintech fraud detection: Logistic regression and random forest on the Kaggle credit card dataset. Public project results.

[7] MDPI. (2024). Enhancing credit card fraud detection: An ensemble machine learning approach. *Big Data and Cognitive Computing, 8*(1), 6. https://www.mdpi.com/2504-2289/8/1/6