Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/115/2025.18483

Efficient Credit Card Fraud Detection Based on Binary Logistic Regression

Junfan Chen[1], Hanyang Qian[2], Weijian Yao[3,a*]

[1]*Faculty of Computational Mathematics and Cybernetics, Shenzhen Msu-Bit University, Shenzhen, China*
[2]*School of Communication and Information Engineering, Shanghai University, Shanghai, China*
[3]*School of Computer Engineering, Jiangsu Ocean University, Lianyungang, China*
a. 2023122279@jou.edu.cn
**corresponding author*

Abstract: With the rapid increase in credit card usage, instances of credit card fraud are also on the rise. The aim of this paper is to design a credit card fraud detection model using binary logistic regression. By using effective detection techniques, the model increases detection accuracy, safeguarding consumer interests and preserving financial market stability. The findings demonstrate that the binary logistic regression model developed for this investigation has a 93.9% accuracy rate in identifying credit card fraud. Important metrics like recall rate and accuracy rate performed exceptionally well, reaching 93.1% and 94.5%, respectively. The model significantly lowers false positives and incorrect assessments in addition to being very good at spotting fraudulent transactions. In addition to offering a reference for resolving other financial fraud detection issues, the paper presents a new method of credit card fraud detection. By improving the model and incorporating additional data characteristics, its performance and applicability can be further enhanced to provide financial institutions with stronger support against future fraud threats.

Keywords: Credit card fraud, Fraud detection, Binary logistic regression.

1. Introduction

A major crime that jeopardizes consumer rights and financial stability is credit card fraud. Credit card fraud action has risen as a result of the recent increase in credit card use. By the end of 2022, there will be 798 million credit cards in total, according to pertinent reports. 86.58 billion yuan, or 0.6% more than the previous year, was the amount owed on credit cards that were past due by more than six months [1]. This demonstrates that even while credit cards are becoming more and more popular, financial institutions are still having a difficult time identifying and stopping fraudulent behavior. Despite advances in technology and increased security measures, credit card fraud remains a persistent and complex problem, which requires continuous improvement of strategies and systems to prevent fraud.

Finding unusual activity that differs from typical transaction patterns is the main difficulty in detecting credit card fraud. There is a serious issue with data imbalance because fraudulent transactions make up a very small percentage of total transactions and are very rare in comparison to regular transactions [2]. The daily transactions on credit cards have turned into a sort of vast stream

DOI: 10.54254/2755-2721/115/2025.18483

of data due to the current size of credit card holdings and usage frequency, making human verification of credit card anti-fraud manifestly impractical [3]. The banking sector frequently uses machine learning-based methods to improve the effectiveness and precision of fraud detection in order to overcome this difficulty. By analyzing patterns in historical transaction data, machine learning models create empirical frameworks that can be used to predict the likelihood of fraudulent activity in future transactions [4]. These models are designed to detect subtle irregularities that may not be easily identifiable by traditional methods, helping financial institutions better mitigate risks associated with fraudulent behavior.

In all kinds of machine learning algorithms, the binary logistic regression model has been widely used because of its superiority in dealing with binary response variable problems. Binary logistic regression models effectively capture the relationship between binary outcome variables (such as the occurrence of fraudulent transactions) and one or more explanatory variables (such as transaction amount, time, etc.). Compared with traditional linear models, it has significant advantages in handling unbalanced data and complex relationships [5]. Xiao and Huang pointed out that binary logistic regression models are able to differentiate between normal and fraudulent transactions by estimating probabilities, thus providing important decision support for financial institutions [6].

The application of the binary logistic regression model in the area of detecting credit card fraud is gradually maturing and has been widely recognized and practiced. Ohlson et al. used the Logistic regression method to build an early warning model for the financial risks of enterprises and analyzed the bankruptcy probability of sample companies [7]. Using the advantages of this model, financial institutions can not only quickly identify potential fraud in large-scale transaction data, but also rationally allocate resources and improve the efficiency of risk management. Specifically, many banks and financial services companies have used binary logistic regression models in combination with other data analysis tools to build more comprehensive anti-fraud systems. In practical applications, binary logistic regression models are usually combined with data preprocessing, feature engineering, and other technologies to improve the predictive performance of the models. Feature selection plays an important role in model training. By extracting key features related to fraud (such as transaction amount, transaction location, etc.), the discriminant ability of the model can be significantly improved. At the same time, with advances in data technology, the training and updating of models have become more efficient, allowing financial institutions to monitor transactions and adjust risk strategies in real time.

As credit card usage increases, credit card fraud is on the rise, and traditional detection methods are no longer able to cope with the increasing complexity and scale of fraud. Therefore, this paper aims to propose an effective fraud detection model through the analysis of existing machine learning technologies to help financial institutions better identify and prevent credit card fraud. Credit card fraud detection is of great practical significance in the field of financial risk control, and utilizing the binary logistic regression model for efficient data processing and analysis is an effective way to improve detection capability. Through these advanced technical means, the security risk brought by illegal transactions can be effectively reduced and the overall security of the financial system can be improved.

This paper mainly compares and analyzes the specific performance of the binary logistic regression model in credit card fraud detection, based on a dataset from Kaggle. Accuracy, recall, F1-score, and confusion matrix are used to evaluate the model's effectiveness and horizontally compare the performance of other models on this dataset.

DOI: 10.54254/2755-2721/115/2025.18483

2. Methods

2.1. Data Sources

The dataset for the project is sourced from Kaggle and the original dataset is provided by the author, Dhanush Narayanan R. This paper uses the complete dataset which contains 8 features and 1000000 samples. Table 1 shows the description of all the variables:

Table 1: Different types of variables

Variable	Description	Type
distance_from_home	the distance from the home where the transaction happened.	Numeric
distance_from_last_transaction	the distance from last transaction that happened.	Numeric
ratio_to_median_purchase_price	Ratio of purchased price transaction to median purchase price.	Numeric
repeat_retailer	Is the transaction happened from same retailer.	Boolean
used_chip	Is the transaction through chip (credit card).	Boolean
used_pin_number	Is the transaction happened by using PIN number.	Boolean
online_order	Is the transaction an online order.	Boolean
fraud	Is the transaction fraudulent.	Boolean

2.2. Method

This paper employs the Binary Logistic Regression model to predict the binary classification of the target variable (e.g., TARGET outcome, 0 or 1). Logistic regression is a commonly used statistical technique for classification challenges, particularly when the dependent variable is binary.
The basic structure of the logistic regression model is:

$$\text{Logit}(\text{p}) = \ln\left(\frac{p}{1-p}\right) \quad (1)$$

Where p is the probability of the event occurring (i.e., Y=1), and $\ln(\frac{\text{p}}{1-\text{p}})$ represents the log odds. This model can be expressed as a linear combination of the independent variables:

$$\text{Logit}(\text{p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (2)$$

2.3. Evaluation Metrics

To assess the model's performance, this paper utilizes accuracy, recall, and precision metrics. Accuracy represents the proportion of correctly classified samples out of the total samples and is suitable for evenly distributed classes; however, it may not be effective when a class imbalance exists. Recall measures the proportion of true positives (TP) out of all actual positives (TP + FN), highlighting the model's ability to identify positive cases. This is especially crucial in contexts where missing positive instances (false negatives, FN) is a concern, such as in medical diagnostics for disease detection.

Precision refers to the proportion of true positives (TP) out of all predicted positives (TP + FP). It emphasizes the accuracy of predicting positive cases and is suitable for scenarios where false positives (FP), or false alarms, are a concern, such as in spam filtering.

The F1 score is recall and accuracy's harmonic average, taking both into account. It is suitable for situations where a precision and recall balance is needed, such as in fraud detection [8].

DOI: 10.54254/2755-2721/115/2025.18483

A confusion matrix is a table that contains the prediction results of a classification model, typically used for binary classification problems [9,10].

3. Result

First, a binary logistic regression model is built based on the dataset obtained from training. Then, the importance of features is calculated by constructing the logistic regression model. Then, the logistic regression model is tested with the trained dataset and the test set, and the evaluation results of classification are obtained. The key point is that the logistic regression model utilizes a data shuffling function, so the results of the operation may be different each time. Even so, with this trained model, new data can be passed directly into the model for classification calculations as a way to streamline future analysis processes.

Table 2: Model parameters

Parameter name	Parameter value
Training time	12.445s
Data slicing	0.7
Data shuffling	yes
Cross-validation	none
Regularization	none
Set constant term	true
Error convergence condition	0.001
Maximum number of iterations	1000

Table 2 shows the configuration of the model's parameter and the training time. The model training time was 12.445 seconds. The data was split with 70% used for training and 30% for testing. The data was shuffled randomly before the split. The error convergence condition was set to 0.001.

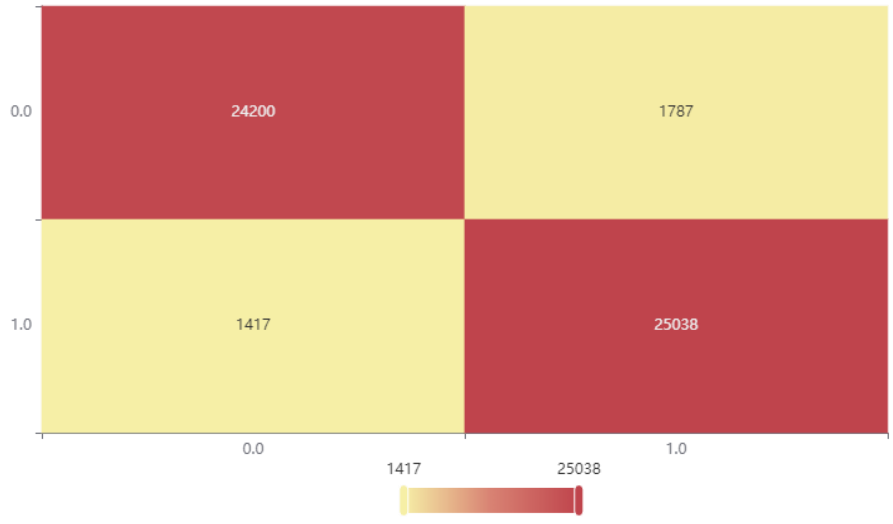

Figure 1: Confusion Matrix Heat Map of Test Set

Figure 1 displays a heat map of the test set's confusion matrix. Here, 0 denotes the positive class, and 1 denotes the negative class. The model correctly predicted 24,200 samples with a true label of 0. There were 1,417 samples of 0 that the model incorrectly predicted as 1. Additionally, 1,787

DOI: 10.54254/2755-2721/115/2025.18483

samples of 1 were wrongly predicted as 0. Lastly, the model accurately predicted 25,038 samples with a true label of 1.

Table 3: Model evaluation results

	Accuracy	Recall	Precision	F1
Training set	0.941	0.933	0.948	0.941
Test set	0.939	0.931	0.945	0.939

Table 3 above presents the prediction evaluation metrics for the cross-validation, training, and test sets, utilizing quantitative measures to assess the logistic regression model's predictive performance. The accuracy scores for the training and test sets are 0.941 and 0.939, respectively, suggesting the model maintains good consistency between training and testing, with no significant overfitting evident.

The recall is recorded at 0.933 for the training set and 0.931 for the test set, demonstrating the model's strong capability to identify positive instances. Precision values are 0.948 for the training set and 0.945 for the test set, indicating the model's high reliability in predicting positive cases. The minimal differences in accuracy, recall, and precision between the training and test sets reflect that the model performs effectively across both datasets. The similar values of these metrics suggest the model avoids overfitting and possesses strong generalization ability.

Table 4: Model evaluation results

	Accuracy	Recall	Precision	F1
Logistic Regression	0.939	0.931	0.945	0.939
Random Forest	1.000	1.000	1.000	1.000
SVM	0.935	0.961	0.953	0.957

Comparing the three models horizontally shows that the random forest model performed the best on this dataset, achieving a perfect classification result of 100% (Table 4). However, it is necessary to monitor for potential overfitting risks. The SVM also showed strong performance, with an accuracy of 93%, making it suitable for more complex data patterns. Lastly, the logistic regression model performed slightly better than the SVM, with an accuracy of 94%. Its strong interpretability makes it well-suited for linear classification problems.

Based on the relatively good predictive performance of this model, it can be considered that there is a significant connection between the selected features and the occurrence of fraud, and machine learning algorithms such as logistic regression can effectively predict the occurrence of credit card fraud [11].

In addition, it is worth noting that the features selected in this article are different from traditional financial indicators (such as personal information, user characteristics, transaction details, financial status, etc.). The features selected in this article focus more on behavioral data, consumption habits, and transaction patterns. Such feature selection can more accurately reflect the user's actual usage scenarios, thereby more accurately determining whether there is a risk of fraud in a transaction.

Currently, models for detecting credit card fraud that rely on conventional financial data are quite advanced. However, models based on behavioral data have not been widely used, partly because such data involves more privacy issues. Through data masking, anonymization, and other technical means, the problem of privacy protection can be effectively solved. Predictive models based on behavioral data combined with machine learning have broad development prospects in the area of detecting credit card fraud and are worthy of further in-depth research and application.

DOI: 10.54254/2755-2721/115/2025.18483

4. Conclusion

This paper employs a binary logistic regression model for detecting credit card fraud action. Among the three models assessed, the Random Forest achieved perfect scores Close to 1.00 on all evaluation metrics, including accuracy, recall, precision, and F1 score. This remarkable performance indicates that the Random Forest model effectively identifies fraudulent transactions within the current credit card fraud detection datasets, exhibiting a low rate of false positives. However, it is crucial to acknowledge the potential for overfitting, which may arise from data set peculiarities or other influencing factors.

In contrast, both Logistic Regression and Support Vector Machines (SVMs) performed less favorably than Random Forest on all metrics. Despite recording higher F1 scores of 0.939 and 0.957, respectively, these models did not achieve perfection.

Consequently, based on this quantitative comparative analysis, the Random Forest model emerges as the most suitable choice for the current dataset because of its superior performance across all key metrics.

While traditional financial data-based models of detecting credit card fraud action are relatively mature, there is still a limited adoption of models that leverage behavioral data, primarily due to heightened privacy concerns associated with such information. These privacy challenges could potentially be mitigated through technical measures, including data obfuscation and anonymization. Predictive models that utilize behavioral data and machine learning techniques offer significant promise in the realm of detecting credit card fraud action and merit further in-depth research and application. Future studies should focus on identifying more effective data protection methods and enhancing the robustness and adaptability of these models for real-world scenarios.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] *China Banking Magazine. (2023). China's bank card industry development blue book (2023) released: The total number of bank cards issued in China maintains year-on-year growth, and the risk control situation of the bank card industry remains serious. China Banking Magazine, 2023(10), 102-104.*

[2] *Makki, S., Assaghir, Z., & Taher, Y., et al. (2019). Experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, 93010-93022.*

[3] *Jiang, H. X., Jiang, J. Y., & Liang, X. (2023). Review of credit card transaction fraud detection based on machine learning. Journal of Computer Engineering and Applications, 59(21), 1-25.*

[4] *Yang, F., Zou, Y., Zhu, M. Z., et al. (2024). Credit card fraud detection model based on graph attention transformer neural network. Journal of Computer Applications, 44(08), 2634-2642.*

[5] *Xiao, S., & Huang, J. W. (2024). Correction Liu estimator in binary logistic regression models. Journal of Liaoning Institute of Technology (Natural Science Edition), 44(01), 64-70.*

[6] *Huang, S. S. (2018). Study on performance optimization methods for shuffle process in Spark big data platform (Master's thesis). Beijing University of Technology.*

[7] *Li, C. (2018). Construction of a corporate financial risk early warning model based on logistic regression method. Statistics and Decision, 34(06), 185-188.*

[8] *Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.*

[9] *Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint, arXiv:2010.16061.*

[10] *SPSSPRO. (2021). Scientific platform serving for statistics professional (Version 1.0.11) [Online application software]. Retrieved from https://www.spsspro.com.*

[11] *Sun, Y., & Lin, W. (2018). Application of gradient descent method in machine learning. Journal of Suzhou University of Science and Technology: Natural Science Edition.*