Credit Card Fraud Detection Based on MiniKM-SVMSMOTE-XGBoost Model

Yanzhao Gu
Faculty of Applied Sciences, Macao Polytechnic University
China
p2311998@mpu.edu.mo

Junhao Wei
Faculty of Applied Sciences, Macao Polytechnic University
China
p2312195@mpu.edu.mo

Ngai Cheong
Faculty of Applied Sciences, Macao Polytechnic University
China
ncheong@mpu.edu.mo

Abstract

In recent years, the problem of credit card fraud has become more acute with the digitisation of credit cards. For the high data volume, high dimensionality and extreme imbalance of credit card transaction data. This paper explores the application in the field of credit card fraud detection based on MiniBatchKMeans-SVMSMOTE-XGBoost model. Through combining clustering, oversampling and classification algorithms, an improved fraud detection method is proposed. The experimental results show that the model performs well in handling unbalanced data with high accuracy and generalisation ability.

CCS Concepts

• **Computing methodologies** → Machine learning; • **Security and privacy** → Intru-sion/anomaly detection and malware mitigation; • **Theory of computation** → Design and analysis of algorithms.

Keywords

MiniKM, SVMSMOTE, Credit Card Fraud Detection, XGBoost, Imbalance data

ACM Reference Format:
Yanzhao Gu, Junhao Wei, and Ngai Cheong. 2024. Credit Card Fraud Detection Based on MiniKM-SVMSMOTE-XGBoost Model. In *2024 8th International Conference on Big Data and Internet of Things (BDIOT) (BDIOT 2024), September 14–16, 2024, Macau, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3697355.3697397

1 Introduction

In recent years, China's credit card industry has flourished in the wave of digital transformation. Diversified e-credit cards and all-digital services for instant application and instant use have closely catered to the needs of young consumer groups, driving continued rapid growth in credit card issuance. With the widespread use of digital technology, innovative payment methods such as digital credit cards, payments by mobile devices, cardless transactions and QR-code payments have blossomed, which greatly facilitates users but also opens up new ways for criminal activities. In 2019-2021, our credit card fraud caseload consistently leads the pack, totaling 3,375 cases accounting for 50.4% of all caseloads. Data from the "Chinese Blue Book on the Development of the Bank Card Industry (2023)" shows that China's credit card industry had outstanding credit card balances of RMB 8.69 trillion at the end of 2022, up 0.9% from the previous year. The total amount of credit cards overdue for half a year was $86.58 billion, up 0.6% from the previous year. According to Fidelity National Information Services (FIS) information, fraud cases on credit cards and debit cards in the United States jumped dramatically during the COVID-19 period. In April, the size of attempted fraudulent transactions in USD increased by 35% compared to the last year.

The new type of fraud exhibits the characteristics of intangibility of transactions, high concealment and difficulty in tracing compared to traditional fraud. For this reason, financial institutions and payment platforms use big data to build intelligent fraud detection systems. Deep analyses of large amounts of transaction data to identify anomalous transaction patterns accurately and improve the precision and response speed of fraud detection. The core of credit card fraud detection is to mine information from the cardholder's credit history data. Determining the presence of fraud through pattern recognition. It is essentially a binary classification problem. However, the problem of extreme data imbalance is encountered when building the detection model, which can lead to the trained model not being able to extract the features effectively. Therefore, effective data processing is also one of the important issues for research.

In this paper, we propose to improve SVMSMOTE using MiniBatchKMeans, which improves the imbalanced data processing method to effectively deal with the problems of high dimensionality, high data volume and extreme unbalance of credit card transaction data. In this paper, XGBoost classification algorithm is optimised using feature engineering to improve the accuracy and generality of fraud detection.

2 Related work

2.1 Credit card fraud detection model

Researchers have been working on detection models in this area. Asha [1] and her team explored the application of multivariate machine learning techniques in the field of credit card fraud identification. An Artificial Intelligence Neural Network (AIN) model is proposed through deep comparison. It achieves almost 100 per cent accuracy and significantly exceeds the traditional unsupervised learning models. Lebichot [2] et al. focus on dynamic learning environments. They constructed a novel credit card fraud detection

BDIOT 2024, September 14–16, 2024, Macau, China

ACM ISBN 979-8-4007-1752-9/24/09
https://doi.org/10.1145/3697355.3697397

system. The core strategy stems from deep mining and optimisation of incremental learning evaluation strategies. Feng Zhao [3] et al. proposed a credit card fraud detection (VAE-GWO-LightGBM) method combining variational autocoder (VAE), grey wolf algorithm (GWO) and lightweight gradient boosting machine (LightGBM), which greatly improves the recognition rate of credit card fraud detection. HONGHAO ZHU [4] and others proposed an adaptive dandelion algorithm that takes much less time than other algorithms for credit card detection. Zhang [5] et al, used SVM algorithm to greatly improve the detection performance. Prasetiyo [6] et al used Random Forest algorithm to train the synthetic dataset, and the recognition effect is better than the traditional machine learning algorithm. Li Meng-tao [7] and others applied the random forest approach to the credit card fraud detection problem. ZHANG [8] et al. proposed an XGBoost and LR fusion model, which possesses higher accuracy than traditional algorithms. Researchers of these models have focused on the research of methods and algorithms but neglected the processing of imbalanced data. Credit card transaction data is characterised by extreme imbalance. The preliminary feature engineering and imbalanced data preprocessing are important work points.

2.2 Imbalance data processing

To solve the data imbalance problem, researchers have proposed a variety of methods and algorithms. Zhang Yihao [9] and others used weighted random forest for imbalanced samples, which effectively improved the algorithm accuracy. Ying Liu [10] et al. used boosting algorithm to generate base classifier clusters combining SVM and random forest (RF). The problem of imbalanced data distribution bias is effectively solved by integrating the multivariate prediction of base classifier clusters using deep belief network (DBN). Chunhua [11] and others constructed a credit card fraud detection model based on KNN-SMOTE-LSTM using LSTM combined with Smote algorithm and KNN classification algorithm. It overcoming the blindness and limitations of the Smote algorithm when generating new samples. Su-Mei Nguyen [12] et al. combined feature selection of FS method, imbalanced data processing of IsolationForest-KMeans++-KNN(IFKK) method and heterogeneous model integration of Stacking method. The overfitting problem due to feature redundancy and sample imbalance is solved, which can detect credit card fraudulent transactions more accurately. WANG [13] et al. integrated DBSCAN and SMOTE to improve sample diversity. It is proved by experiment that DB-MCSMOTE adjusts the performance of classification and recognition. Xu L [14] et al. used SMOTE to effectively improve the classification performance of CatBoost. The model convergence speed and classification accuracy were improved. Nguyen [15] et al. proposed an improved method called SMOTE-CD. It generates synthetic data by calculating linear combinations of selected existing data points and the results show that all indicators are improved. YANG [16] et al. proposed an oversampling algorithm AGNES-SMOTE based on hierarchical clustering and improved SMOTE. Sample noise and sample marginalisation during sample synthesis were dealt with. WU HaiYan [17] et al. proposed an adaptive kernel synthetic minority oversampling technique support vector machine(SMOTE-SVM) classifycation algorithm. The algorithm can be modified to improve the robustness and accuracy of the classification algorithm through experiments on multiple datasets. SONG Yinghua [18] and others used Tomek Link to improve the sampling method and constructed the SmoteTomek- GBDT model. Researchers have begun to move from single methods to improved methods. Without exception, the processing of extremely imbalanced credit card data is an important task for improving model accuracy and increasing generalisation. In order to improve the recognition rate of credit card fraud detection, this paper constructs a MiniKM-SVMSMOTE sampling method. The advantages of KMeans clustering are used to improve the model classification and recognition performance.

3 Method

3.1 MiniKM

MiniBatchKMeans(MiniKM) is a variant of the KMeans algorithm that uses a small subset of data to greatly reduce computation time. It is very suitable for running on large sample datasets. MiniKM is able to greatly reduce the computation time while maintaining clustering accuracy [19]. The steps of MiniKM algorithm are as follows:

1. In the first step, extracted part of the dataset. A model of K clustered points is constructed using KMeans algorithm.
2. Continue to extract portions of the training dataset and add them to the model, then assigning them to the closest cluster centroids.
3. Recalculation of centre point values.
4. Cycle step 2 and step 3 until the centre point is stable or the number of iterations is reached.

The function of KMeans clustering is shown in equation(1):

$$E = \sum_{i=1}^{k} \sum\nolimits_{x \in C_i} ||x - \mu_i||^2 \tag{1}$$

Where x denotes the data sample, C_i denotes the ith clustering cluster, μ_i is the mean vector of C_i denotes the cluster centre of the ith clustering cluster. The smaller the E value, the more similar the samples within the cluster are. In order to find the most suitable k value quickly and accurately, the elbow rule is used in this paper with the following steps:

1. Iterate the calculation n times for a data set of n points. After each clustering is completed, calculate the sum of the squares from each point to the centre of the clusters.
2. It is progressively smaller because each point is the centre of the cluster itself in which it is located.
3. In this process, there is an ”elbow” point, where the rate of decline suddenly slows down. It is considered to be a good k value.

3.2 SVMSMOTE

Support Vector Machine Synthetic Minority Oversampling Technique(SVMSMOTE) is a variant of SMOTE. SMOTE improves the balance of the dataset by interpolating between minority class samples to synthesise new samples, in this way increasing the number of minority class samples. New samples generated by the SMOTE algorithm may be distributed at the edges of the original data distribution, resulting in less distinguishable demarcation lines between

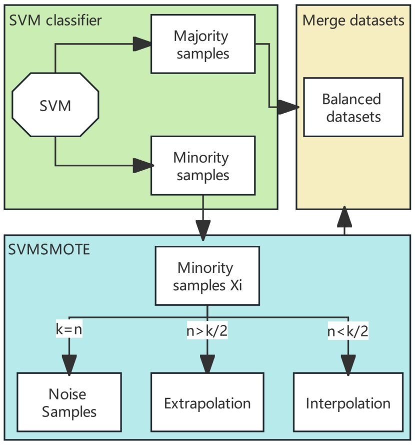


Figure 1: The data processing steps for SVMSOMTE

samples. Meanwhile, the samples generated by the SMOTE algorithm are based on the original few samples, which contain some noisy data and are easy to cause the distribution marginalisation problem [20]. SVMSMOTE improves on this by using a support vector machine (SVM) to determine which minority class samples are more important to be oversampled. It provides a more refined and effective oversampling strategy than the traditional SMOTE method, especially in improving the performance of the classification boundary region.

Specifically, SVMSMOTE starts by training the original dataset using an SVM classifier. The SVM will find a hyperplane that distinguishes between majority and minority class samples as correctly as possible. In this process, some minority class samples are determined as support vectors, which are usually located near the decision boundary. Then, SVMSMOTE will preferentially select these support vector samples close to the decision boundary for oversampling. Such a strategy can more effectively enhance the model's ability to recognise a small number of classes, especially near the decision boundary, which is usually the most difficult region to correctly classify in classification tasks. By oversampling samples close to the decision boundary, SVMSMOTE helps to generate more representative and diverse minority class samples which in turn improves the generalisation ability of the model. The data processing steps for SVMSOMTE are shown in Fig.1. First, the SVM model is run to determine the classification boundaries. In SVMSMOTE, special attention was paid to those samples that belonged to minority classes. Then for each minority class x_i, find its k neighbouring minority samples in the data space. After determining the k neighbouring minority samples for each minority class x_i, new data points are next generated by inserting. Record the number n of majority classes of neighboring samples k. If k = n then it is a noise sample and relabeled. If n>k/2 the value is extrapolated; if n<k/2 the value is interpolated. Repeat the above interpolation process so that multiple new data points are generated.

3.3 MiniKM-SVMSMOTE

The main purpose of fusing MiniKM and SVM-SMOTE algorithms is to combine fast clustering methods with efficient oversampling techniques. First, the data were initially clustered using the MiniBatch K-Means algorithm. The clustering generated by MiniBatch K-Means and allows the dataset to be divided into multiple clusters of minority samples. Each cluster contains a portion of the minority class data. Such segmentation helps in further processing, especially when the amount of data is very large. Then, the SVMSMOTE algorithm is applied individually to each clustered generated subset to increase the balance of the data. After processing all the subsets, they are recombined into one dataset. The method aims to balance processing efficiency and preprocessing effectiveness, and is particularly suitable for those application scenarios where the sample size is large and the category is seriously imbalanced.

3.4 XGBoost

The name of XGBoost is Extreme Gradient Boosting, It improved on the original GBDT, making the model much more effective. XGBoost is composed of multiple Classification And Regression Tree (CART) as shown in equation (2). It provides good generalisation capabilities through powerful regularisation and systematic optimisation methods and shows superior performance on a wide range of datasets. Its core strengths are speed and performance, and it typically delivers better results than traditional gradient boosting methods. Thus, XGBoost can show strong performance in credit card fraud detection.

$$\hat{y}_i^{(t)} = \sum\nolimits_{k=1}^{t} f_k\left(x_i\right) = \hat{y}_i^{(t-1)} + f_t\left(x_i\right) \tag{2}$$

In this equation, $\widehat{y_i}^{(t)}$ is the prediction of sample i after the t-th iteration. $\hat{y}_i^{(t-1)}$ is the prediction of the t-1st tree and $f_t(x_i)$ is the model for the t-th tree. The loss function of XGBoost can be represented by the predicted value and the true value. As shown in equation 3:

$$L = \sum\nolimits_{i=1}^{n} l\left(y_i, \hat{y}\right) \tag{3}$$

In this equation, n is the number of samples, $\hat{y}_i$ is the predicted value and y_i is the true value. The a model prediction accuracy is determined by the bias and the variance, so the loss function represents the error of the model. If you want to reduce the error then you need to add a regular term in the obj-function to prevent overfitting, so the obj-function is composed of a loss function L and a regular term Ω. The objective function is defined in equation 4:

$$T = \sum\nolimits_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum\nolimits_{i=1}^{t} \Omega\left(f_i\right) \tag{4}$$

3.5 MiniKM-SVMSMOTE-XGBoost model construction

In order to improve the efficiency and accuracy of credit card fraud detection. In this paper, we choose to improve the traditional SMOTE method by clustering the data first through MiniKM, and then using SVM to identify the boundary of the minority class dataset, aiming to process the minority class data in the dataset. Finally, it is handed over to the XGBoost algorithm for classification. MiniKM-SVMSMOTE-XGBoost model structure diagram is shown in Figure.2 below.

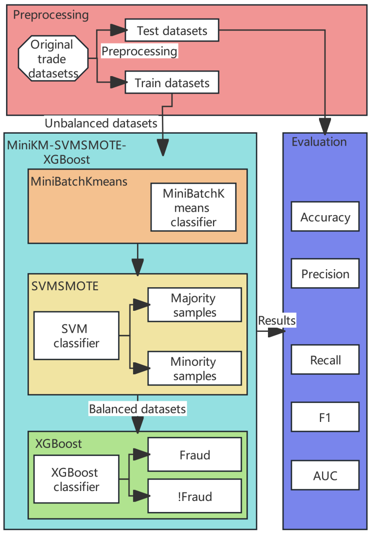


Figure 2: MiniKM-SVMSMOTE-XGBoost Model Structure Diagram

For credit card data with many samples and large dimensions. In this paper, MiniKM is used to cluster the data then SVMSMOTE is applied for minority class of samples to oversample the imbalanced data. The excellent performance of SVMSMOTE can help the model to get accurate and reliable balanced data, and then give the balanced data to XBGoost algorithm for training. The prediction results are compared with the test results, and the training effect is evaluated by the model evaluation metrics.

4 Experimentation

4.1 Data selection

The dataset is from the IEEE-CIS Fraud Detection dataset on Kaggle. The dataset was provided by Vesta and the dataset has characteristics of authenticity, comprehensiveness and reliability. The detailed transaction and identification information is shown in Table 1, Table 2

The dataset used in this paper contains 590,541 transaction data and 144,234 identification data. Among them, there are 20663 fraud data, which is about 3.5% of the all data, and the data is extremely imbalanced. To protect user privacy, the transaction time and some attribute field names are masked.

4.2 Data preprocessing

First, filling in the missing values in the original data. Second, normalise the transaction time data D columns. The DT Columns are "time deltas" from some point in the past. We will transform the DT Columns into their point in the past. This will stop the DT columns from increasing with time. As shown in Equation 5, Equation 6 below:

$$\text{TransactionDay} = \text{TransactionDT}/(24*60*60) \tag{5}$$

$$\text{Dx}_{\text{nor}} = |\text{TransactionDay} - \text{Dx}| \tag{6}$$

Standardizing Dxx facilitates the model's identification of uniform credit card activities. When two transactions exhibit identical normalized values for D1, D7, and D15, it strongly suggests that they originate from the same card. This is due to the fact that the normalized Dxx represents a specific past event timestamp. Since the original dataset contains both values and non-values, the next step is to label encoding the non-values. Before clustering with MiniKM, the optimal value point of k was found by the elbow law as shown in Figure 3. This process identifies the different subsets where minority classes exist and then the clustered data is given to SVMSOMTE for processing. After the above processing, credit card transaction data balancing is achieved, which improves the generalisation of credit card fraud detection.

4.3 Evaluation

Due to the imbalanced feature of credit card transaction data, this paper adopts a comprehensive series of evaluation metrics to assess the processing performance of algorithms on imbalanced datasets. For classification results, the text uses accuracy, precision, recall,

Table 1: Transaction Table

Name	Description
TransactionID	transaction id
TransactionDT	timing gap (not actual timestamp)
TransactionAmt	transaction in USD
ProductCD	product id
card	card information (card type/bank/country etc.)
ADDR	
dist	
P_ emaildomain and R_ emaildomain	user email domain
Cxx	recording (how many addresses/email/number are linking with cards, etc.)
Dxx	timedelta (days between previous transaction)
Mx	match (names, number and address on card, etc.)
Vxxx	Vesta provides rich function: arrangements, records and other properties.

Table 2: Identity Table

Name	Description
TransactionID	transaction id
id_xx	network information and digital signature
DeviceType	PC/Phone, etc.
DeviceInfo	Device details (Name/OS, etc.)

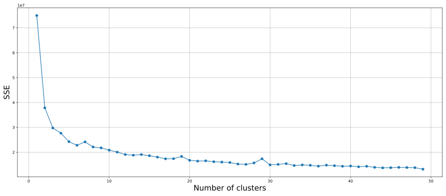


Figure 3: Elbow Law

F1 and AUC as evaluation metrics[21].

$$Accuracy = \frac{\mathrm{TP+TN}}{TP+TN+FP+FN} \tag{7}$$

$$Precision = \frac{\mathrm{TP}}{TP+FP} \tag{8}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{TP+FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

$$TPR = \frac{\mathrm{TP}}{TP+FN} \tag{11}$$

$$FPR = \frac{\mathrm{FP}}{FP+TN} \tag{12}$$

In this paper, the balanced dataset is divided into training set and test set in 8:2 ratio. The XGBoost model is compared with the base models such as Decision Tree, Random Forest, GBDT and Logistic Regression and the results of the experiments are shown in Table 3 below:

As can be seen from Table 3, the prediction results of the traditional tree models DT and Random Forest are relatively similar, slightly falling behind in comparison to the performance of the current more advanced tree models GBDT and tradition model Logistic. As a traditional classification model Logistic model has better performance results with high F1 value, which is a comprehensive measure of credit card fraud classification. This represents that the Logistic model did obtain a better generalisation, which may be related to the clustering of MiniKM, where the generalisation of the model was further improved. After in-depth analysis, it is found that as an improved model of GBDT tree model XGBoost has a high AUC value. Comprehensively analysing its values XGBoost has better overall classification performance on imbalanced data and its generalisation ability is well demonstrated. Therefore, in this paper, XGBoost model is used as a credit card fraud detection model.

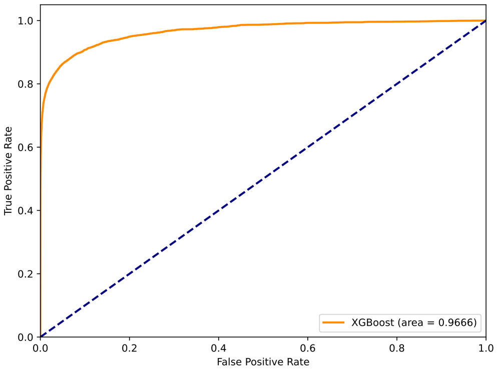


Figure 4: XGBoost model effectiveness without imbalanced data processing

In order to further evaluate the accuracy, reliability, and validity of MiniKM-SVMSOMTE-XGBoost model. This paper also compares the results of KMSMOTE-XGBoost, SVMSMOTE-XGBoost, and XGBoost model training under imbalanced data. the experimental results are shown in Table 4.

Through comparison, it can be found that after imbalanced data processing, the performance of various models has been greatly improved and all the indicators have reached more than 90% except SMOTE-LightGBM. The XGBoost model without imbalance processing is poorly trained, as shown in Figure 4, with Recall and

Table 3: Comparison of various fraud detection models

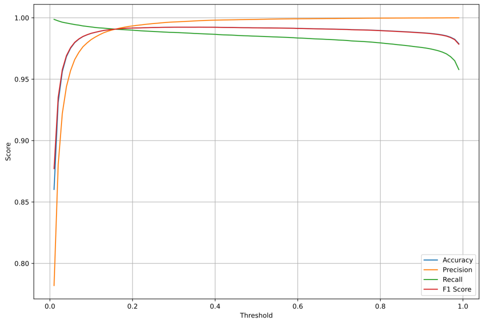
Model	Accuracy	Precision	Recall	F1	AUC
Decision Tree(DT)	0.9171	0.5952	0.9264	0.6315	0.9234
Random Forest(RF)	0.9128	0.5914	0.9258	0.6351	0.9232
GBDT	0.9613	0.6645	0.9189	0.7273	0.9397
Logistic	0.9775	0.7258	0.9483	0.7971	0.9481
XGBoost	0.9740	0.9120	0.9548	0.7345	0.9654

Table 4: Evaluation of various sampling methods

Model	Accuracy	Precision	Recall	F1	AUC
XGBoost(Imbalanced data)	0.9837	0.9489	0.5782	0.7186	0.9666
SMOTE- LightGBM	0.9885	0.9874	0.8524	0.8515	0.9082
KMSMOTE-XGBoost	0.9618	0.9787	0.9247	0.9189	0.9477
SVMSMOTE-XGBoost	0.9713	0.9745	0.9259	0.9289	0.9582
MiniKM-SVMSOMTE-XGBoost	0.9821	0.9889	0.9753	0.9821	0.9988

F1 values much lower than the model after imbalance processing. The processing of imbalance data can be seen as important.

In order to further verify the validity of the MiniKM-SVMSOMTE-XGBoost model, the KMSMOTE-XGBoost model, the SVMSMOTE-XGBoost model and SMOTE-LightGBM model were also added for comparison. The traditional SMOTE is weaker than the optimised processing of imbalance data. Its Recall and F1 values did not reach 90%. The model performance of KMSMOTE-XGBoost and SVMSMOTE-XGBoost converge and both are able to show strong generalisation. However, comparing the MiniKM-SVMSOMTE-XGBoost model with other models, it performs better in 5 evaluation metrics, namely Accuracy, Precision, Recall, F1 and AUC. It is verified that the MiniKM-SVMSOMTE-XGBoost model in this paper is more effective in credit card fraud detection and provides better capability, as shown in Figure 5, which shows that its effectiveness and accuracy have been greatly improved.



5 conclusion

This paper provides an insight into credit card fraud detection methods based on MiniKM-SVMSMOTE-XGBoost model. To investigate the processing of imbalance data, this paper combines MiniKM clustering, SVMSMOTE oversampling and XGBoost classification algorithms to improve the efficiency and accuracy of fraud detection. The experimental results show that the model performs well in processing credit card transaction data. It can effectively identify abnormal transaction patterns and improve the accuracy and response speed of fraud detection. The model provides a useful reference for financial institutions and payment platforms to build intelligent fraud detection systems. By combining advanced data processing technology and machine learning algorithms, the challenges of credit card fraud can be better addressed. Cardholder rights and financial security are protected.

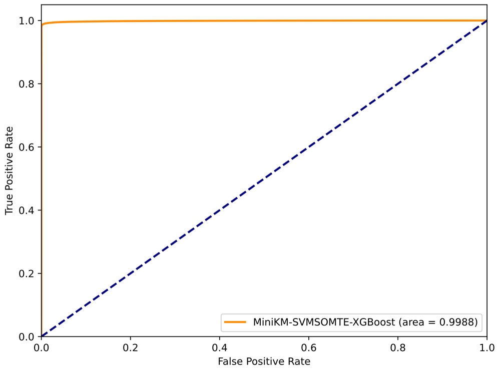


Figure 5: MiniKM-SVMSOMTE-XGBoost model results

The next step can be exploring the combination of multiple sampling methods and feature engineering improvements to further improve the generalisation of the model. In addition, the model in this paper can be applied to other binary classification problems with imbalanced data, for example, disease diagnosis, spam detection, machine breakdown prediction, and so on. Overall, the model is widely used and has high research significance and implementation value.

Acknowledgments

The supports provided by Macao Polytechnic University (MPU Grant no: RP/FCA-06/2022) and Macao Science and Technology Development Fund (FDCT Grant no: 0044/2023/ITP2) enabled us to conduct data collection, analysis, and interpretation, as well as cover expenses related to research materials and participant recruitment. MPU and FDCT investment in our work have significantly contributed to the quality and impact of our research findings.

References

[1] ASHA R B, SURESH KUMAR K R. 2021. Credit card fraud detection using artificial neural network. Global Transitions Proceedings, 2021,2(1):35-41.
[2] LEBICHOT B, PALDINO G M, SIBLINI W, *et al.* 2021. Incremental learning strategies for credit cards fraud detection. International Journal of Data Science and Analytics,2021,12:165-174.
[3] Zhao Feng,Li Niu Niu. 2023. Credit card fraud detection method based on VAE-GWA-LightGBM. Journal of Northeast Normal University (Natural Science Edition), Changchun, Jilin, 2023,55(4):77-84. https://doi.org/10.16163/j.cnki.dslkxb202209210002.
[4] HONGHAO ZHU, MENGCHU ZHOU, YU XIE, *et al.* 2024. A Self-Adapting and Efficient Dandelion Algorithm and Its Application to Feature Selection for Credit

Card Fraud Detection. IEEE/CAA JOURNAL OF AUTOMATICA SINICA, VOL.11, NO.2, FEBRUARY 2024, 377-390. https://ieeexplore.ieee.org/document/10415907/.

[5] ZHANG D F, BHANDARI B, BLACK D. 2020. Credit card fraud detection using weighted support vector machine. Applied Mathematics, 2020,11(12):1275-1291.

[6] PRASETIYO B, ALAMSYAH, MUSLIM M A, *et al.* 2020. Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique. Journal of Physics: Conference Series, 2020,1918(4):1-5.

[7] Li Meng-tao, LV Zhao-hui. 2020. Detection of credit card fraud based on data mining. Journal of Communication University of China Science and Technology, Beijing, 2020,27(6):69-73.

[8] ZHANG Haiyang, CHEN Yuming, ZENG Nianfeng, *et al.* 2024. A fusion model for credit card fraud detection based on XGBoost and LR. Journal of Chongqing University of Technology(Natural Science), Xiamen, 2024,38(3):195–200. https://doi.org/10.3969/j.issn.1674-8425(z).2024.03.021

[9] Zhang Yihao, Sheng Danhong, Li Lifang, *et al.* 2021. Application and research on credit card fraud detection based on weighted random forest. Computer Programming Skills and Maintenance, Beijing, 2021(4):111-112,117. https://doi.org/10.3969/j.issn.1006-4052.2021.04.040.

[10] Y. Liu,K. Yang. 2021. Credit Fraud Detection for Extremely Imbalanced Data Based on Ensembled Deep Learning. Computing Research and Development, Changchun, Jilin, 2021,58(3):539-547. https://doi.org/10.7544/issn1000-1239.2021.20200324.

[11] Pui-Chun Hua,Guan-Yu Chen and Fu-Guang Bao. 2021. KNN-Smote-LSTM Based Consumer Financial Risk Detection Model:A Case Credit Card Fraud Detection. System Science and Mathematics, Hangzhou, Zhejiang, 2021,41(2):481-498.

[12] S. M. Ruan, X. S. Sun, and C. X. Gan. 2023. Research on Credit Card Fraud Detection Based on Three−stage Ensemble Learning. Operations Research and Management, Anhui, 2023,32(12):118-123. https://doi.org/10.12005/orms.2023.0395.

[13] WANG Liang, YE Jimin. 2020. Hybrid algorithm of DBSCAN and improved SMOTE for oversampling. Computer Engineering and Applications, Xi'an, 2020, 56(18):111-118. https://doi.org/10.3778/j.issn.1002-8331.1906-0441

[14] Xu L, Jing X N, Yangg Y, *et al.*2023. National surface water quality classification evaluation based on SMOTE-GA·CatBoost method. China Environmental Science, Hefei, 2023, 43(7):3848-3856. https://doi.org/10.3969/j.issn.1000-6923.2023.07.059

[15] Nguyen T, Mengersen K, Sous D, Liquet B. 2023. SMOTE-CD: SMOTE for compositional data. PLOS ONE, Korea, 18(6): e0287705. https://doi.org/10.1371/journal.pone.0287705

[16] QIN Qin, YANG Yue, CHEN Mingsong and WANG Xin. 2022. Improved SMOTE for Oversampling. Journal of Guilin University of Electronic Technology, Guangxi, 2022,42(1):53-59. https://doi.org/10.3969/j.issn.1673-808X.2022.01.007

[17] WU HaiYan, CHEN XiaoLei and FAN GuoXuan. 2021. An adaptive kernel SMOTE-SVM algorithm for imbalanced data classification. Journ-al of Beijing University of Chemical Technology(Natural Science Edition), Beijing, 2023,50(2):97-104. https://doi.org/10.13543/j.bhxbzr.2023.02.012

[18] SONG Yinghua, JIANG Chen, LI Moxiao and QI Shi. 2023. Rockburst prediction model based on improved Smote-GBDT algorithm. China Safety Science Journal, Wuhan, 2023,33(09):25-32. https://doi.org/10.16265/j.cnki.issn1003-3033.2023.09.0850

[19] Lin Kangwei, Xiao Hong, Jiang Wenchao, *et al.* 2022. Optimal Control of Denitrification Processes in Coal-Fired Power Plants Based on Deterministic Policy Gradients with Deep Reinforcement Learning. Computerised measurement and control, Guangzhou, Guangdong, 2022,30(10):132-139. https://doi.org/10.16526/j.cnki.11-4762/tp.2022.10.021

[20] Wang Guotao, Lv Bingze, Sun Zhigang, Liang Xiaowen and Yan Huizhen. 2020. An unbalanced data processing method based on improved SMOTE algorithm. CN202010832796.2, Heilongjiang, 2020-11-20.

[21] Zhang Tianyi, Ding Lixin. 2021. A NEW RESAMPLING METHOD BASED ON SMOTE FOR IMBALANCED DATA SET. Computer applications and software, Wuhan, Hubei, 2021,38(9):273-279. https://doi.org/10.3969/j.issn.1000-386x.2021.09.043