Application of Machine Learning in Credit Card Fraud Detection: A Case Study of F Bank

Yuan-Fa Lin[1], Chou-Wen Wang[1(✉)], and Chin-Wen Wu[2]

[1] National Sun Yat-Sen University, Kaohsiung, Taiwan
chouwenwang@gmail.com
[2] Nanhua University, Dalin, Chiayi County, Taiwan

Abstract. Due to the Covid-19 pandemic, people are increasingly engaging in non-face-to-face credit card transactions in their daily lives. However, this trend has also provided opportunities for malicious actors to obtain customer credit card information through various illicit means, leading to a continuous rise in credit card fraud. Traditional fraud detection methods, relying on extensive rules and manual judgment, struggle to effectively prevent the evolving techniques of fraud and often result in significant false positives, requiring substantial time for transaction verification. In recent years, the development of big data and machine learning algorithms has offered an effective solution to this challenge. This study employs three common machine learning algorithms—Logistic Regression, Random Forest, and Extreme Gradient Boosting—for predicting credit card fraud. Utilizing transaction data from Bank F time period from January 2021 to May 2023, including fields such as transaction ID, credit limit, occupation, transaction date, transaction time, transaction amount, etc., the study addresses the issue of imbalanced data in credit card fraud through sampling methods. Different ratios of normal to fraud samples, coupled with varying sampling frequencies, are employed along with ensemble learning techniques to enhance the accuracy and stability of the predictive model. Subsequently, various commonly used machine learning evaluation metrics are applied to identify the best model. The empirical results indicate that the Extreme Gradient Boosting model performs best in detecting credit card fraud. In scenarios with different sampling ratios of normal to fraud samples, the study identifies key features such as changes in the cardholder's transaction behavior concerning transaction region, frequency, and amount. The results of this study provide the bank with references on how to develop more effective strategies for fraud prevention.

Keywords: Credit Card Fraud Detection · Machine Learning · Logistic Regression · Random Forest · Extreme Gradient Boosting

1 Introduction

The COVID-19 pandemic has significantly increased non-face-to-face credit card transactions, leading to a surge in credit card fraud as fraudsters exploit various illicit means to acquire customer data. Traditional fraud detection methods, heavily reliant on rules

F. F.-H. Nah and K. L. Siau (Eds.): HCII 2024, LNCS 14720, pp. 210–222, 2024.
https://doi.org/10.1007/978-3-031-61315-9_15

and manual judgement, have become less effective against evolving fraud techniques and often produce a high number of false positives, resulting in the need for time-consuming transaction verifications. However, the recent advancements in big data and machine learning algorithms offer a powerful solution to this challenge.

In this context, the current study utilizes three popular machine learning algorithms - Logistic Regression, Random Forest, and Extreme Gradient Boosting - to predict credit card fraud. The study is methodically structured into five stages: establishing research motivation and objectives, formulating research methods via literature review, planning research methodology and data sources, conducting empirical model analysis, and drawing conclusions. The data set comprises 176,473 transactions from Bank F's credit card records between January 2021 and May 2023, with a minority marked as fraudulent. The study addresses the issue of data imbalance through under-sampling and ensemble learning techniques, and employs feature engineering to enhance the accuracy of fraud prediction. To tackle the problem of imbalanced data in credit card fraud, the study employs sampling methods with ratios of 10:1 and 20:1 for normal and fraudulent samples, respectively. Through various iterations of these sampling techniques, multiple independent classifiers were trained using ensemble learning methods, with the results of these models then integrated to generate the final prediction.

This approach is designed to improve the predictive accuracy and stability of the models. The study applies various machine learning evaluation metrics, such as the Confusion Matrix, Accuracy, Precision, Recall, F1-Score, and ROC Curve, to identify the most effective model. Empirical findings reveal that the Extreme Gradient Boosting model excels in detecting credit card fraud, outperforming the other models across all metrics. Key features that emerged as significant in different sampling ratios include changes in cardholder transaction behavior, particularly in terms of transaction region, frequency, and amount. These insights are invaluable for banks in developing more effective fraud prevention strategies.

The rapid pace of technological innovation, marked by faster internet speeds, increased computing power, and reduced hardware costs, has greatly accelerated the fields of big data analysis and machine learning. Machine learning models, which learn from training data and improve as they process more data, become increasingly accurate over time. Customized fraud detection models, based on customers' historical transaction data, can be developed through machine learning. However, with the evolution of diverse machine learning models, choosing the most suitable one for credit card fraud detection has become crucial.

In conclusion, the study's findings indicate that the Extreme Gradient Boosting model is superior in detecting credit card fraud, offering key features crucial for fraud detection. These results provide banks with valuable references to enhance their fraud detection systems and reduce fraud risks, thereby helping to minimize the time and cost associated with manual transaction verification.

2 Literature Review

2.1 Application of Machine Learning Models in the Finance

Machine Learning (ML) algorithms learn to predict outcomes from data, automating decision-making with minimal human intervention. They adapt and improve over time by identifying patterns within large datasets. ML has become essential across various fields, such as quantitative finance, computer vision, and natural language processing, thanks to the rise of big data and computational advancements.

ML models are built on training datasets and adjust their parameters through iterative learning to enhance performance based on new data inputs. These algorithms can be trained in several ways, each with its advantages and challenges, broadly categorized into four types:

1. Supervised Learning: Utilizes labeled training data to learn the mapping between inputs and the correct outputs, aiming to predict outcomes for new data. It includes algorithms like Linear Regression, SVM, and Decision Trees, suitable for classification and regression tasks.
2. Unsupervised Learning: Works with unlabeled data to discover hidden structures or patterns, without predefined outcomes. It's used for clustering, dimensionality reduction, and association analysis, employing algorithms like K-Means and PCA.
3. Semi-Supervised Learning: Combines elements of both supervised and unsupervised learning, using a small amount of labeled data alongside a larger set of unlabeled data to improve model performance.
4. Reinforcement Learning: Focuses on training agents to make decisions by interacting with an environment to achieve maximum cumulative rewards. It involves defining agents, states, actions, rewards, and policies to learn the best strategies for given objectives.

In the realm of machine learning applications for credit card fraud detection, several challenges and strategies emerge from recent studies. Weston et al. (2008) utilize real credit card transaction data for peer group analysis, monitoring account holders' behaviors against a peer group with expected similar behaviors. Significant deviations indicate potential fraud, leveraging similarities within peer groups to detect anomalies or suspicious transactions. Prusti and Rath (2019) suggest an innovative approach using ensemble learning, which constructs multiple independent classification algorithms. Each learns and predicts independently before their results are combined into a single prediction, forming a more stable model with superior predictive power for fraud detection.

Baabdullah et al. (2020) apply various machine learning algorithms to handle imbalanced data and detect credit card fraud. Their comparative study reveals that without resampling techniques, focusing on accuracy, sensitivity, and the area under the precision/recall curve (PRC) can enhance fraud detection accuracy and reduce fraudulent incidents.Goyal and Manjhvar (2020) point out the lack of a universal benchmark in evaluating fraud detection systems, leading many studies to use multiple metrics for a comprehensive model assessment. Beyond basic accuracy, precision, recall, F1 scores, confusion matrices, and ROC curves are considered to minimize Type I and Type II errors.Cherif et al. (2023) highlight the issue of imbalanced datasets, where fraudulent

transactions are significantly outnumbered by legitimate ones, posing a challenge for prediction models. Most algorithms assume an equal distribution of classes, but in fraud detection, the rarity of fraud cases can diminish algorithm effectiveness. Addressing this issue involves enhancing algorithm adaptability to handle imbalanced datasets or employing sampling methods to balance data distribution.

Collectively, these studies underscore the multifaceted challenges in credit card fraud detection. Imbalanced datasets significantly impact algorithm performance, addressed either by adapting algorithms to manage imbalances or using sampling to equalize data representation. Ensemble learning and peer group analysis offer effective fraud detection methods by combining multiple classifiers' predictions or comparing transaction behaviors within peer groups. However, the absence of a unified evaluation standard necessitates using various metrics for a thorough performance assessment, aiming to enhance accuracy and reduce the occurrence of fraud.

3 Research Methodology

3.1 Data Source

The data for this study were derived from the transaction records of credit card customers at F Bank, spanning from January 1, 2021, to May 31, 2023. The dataset encompasses transactions from 743 accounts, totaling 176,473 credit card transactions. Among these, 175,330 were normal transactions, and 1,143 were identified as fraudulent, with the proportions of fraudulent to normal transactions being 0.65% and 99.35% respectively. The transaction data includes various types of credit card activities within the specified period, such as general purchases, authenticated transactions, tax payments, installment plans, and rewards redemption. Out of the sample, 290 customers had experienced credit card fraud, leaving 453 customers who had not been subjected to fraud. In the entire dataset, fraud transactions accounted for a small fraction of the total transactions, indicating that while the majority of credit card activities are legitimate, fraudulent transactions tend to be concealed within normal activities and are designed to evade detection by the issuing bank.

As shown in Table 1, this study encompasses a total of 35 feature fields, including transaction ID, credit limit, occupation, marital status, education level, transaction date, transaction time, transaction amount, among others. These feature fields provide extensive information about credit card transactions. Preliminary data exploration and processing were conducted using these fields, and feature engineering was applied to generate derived features to assist in enhancing the accuracy of fraud transaction prediction. Derived fields, such as the cumulative transaction amount over the last 360 days, cumulative number of transactions over the last 360 days, maximum transaction amount over the last 360 days, and minimum transaction amount over the last 360 days, offer additional insights into cardholders' transaction behavior patterns. These patterns may help predict potential fraudulent transactions, as fraud may exhibit unusual patterns or changes in behavior in these features. Utilizing these features in training machine learning models can improve the models' effectiveness in detecting fraudulent transactions. Subsequently, this study will employ these data to develop credit card fraud detection models and compare the performance of different machine learning models

in fraud detection. By analyzing the data and evaluating the performance of machine learning models, the study aims to identify superior machine learning models for timely approval of legitimate transactions and immediate detection and prevention of fraudulent activities.

The transaction amount ranges for both normal and fraudulent transactions. The majority of normal transactions are concentrated in amounts below 1000 TWD, accounting for 70.53% of the total. However, the interval below 1000 TWD also exhibits a higher risk of fraud, with fraudulent transactions in this range constituting about 27.38%, suggesting a relatively higher proportion of fraud in low-amount transactions.

Data Source: Compiled for this Study. The tendency for fraudulent transactions to opt for lower transaction amounts may be attributed to smaller amounts being less likely to attract the attention of the issuing institution and the cardholder. Another reason might be to minimize the risk of detection, as issuing institutions usually implement control mechanisms for large transactions, such as instant notifications through app push notifications, SMS, or email to the cardholder, or verification calls from personnel. In Taiwan, to enhance the security of online credit card transactions, the Financial Supervisory Commission mandates that issuing banks send transaction confirmation SMS for online transactions exceeding 3000 TWD. Thus, Table 3–3 reveals that transactions below 5000 TWD encompass 86.09% of fraudulent transactions. Nevertheless, the segment above 5000 TWD still covers 13.91% of fraud, likely because fraudsters aim to quickly exhaust the cardholder's credit limit by attempting large-amount transactions to maximize their illicit gains. Therefore, when cardholders conduct unusually large transactions, issuing banks should also pay special attention to whether these transactions are legitimate.

3.2 Data Imbalancing Treatment

In credit card fraud detection, imbalanced data is a key issue because most machine learning algorithms are affected by data imbalance as identified in prior literature reviews. However, this problem can be effectively addressed through sampling methods or ensemble learning. Common sampling methods include data duplication and deletion, which help to form a more balanced dataset. Here are introductions to some common sampling methods:

1. Oversampling: Oversampling increases the number of samples in the minority class by repeatedly sampling to equalize the number of examples across classes. The advantage is maintaining data integrity without losing important information since it balances the dataset by reusing existing data. However, it can be time-consuming and increase computational costs. Moreover, if the proportion of fraud samples is extremely low, it might lead to model overfitting because the model learns predominantly from the repeated minority samples and might struggle to predict accurately in real-world scenarios.
2. Under-sampling: Under-sampling reduces the number of samples in the majority class through random sampling to balance the dataset. For example, in a credit card transaction record, if there are 200,000 normal transactions and only 1,000 fraudulent ones, under-sampling would randomly remove samples from the normal transactions

Table 1. Feature Description

NO	Variables	Explanation	Data Type
1	IDNO	Transaction ID	Category
2	CRLIMIT	Credit Limit	Number
3	POSITION	Occupation	Category
4	MARRIAGE	Marital Status	Category
5	EDUCATION	Education Level	Category
6	TX_DATE	Transaction Date	Number
7	TX_TIME	Transaction Time	Number
8	TX_AMT	Transaction Amount	Number
9	POS_NO	Terminal Machine Number	Category
10	REV_FLAG	Cancellation Mark	Category
11	ADJ_FLAG	Adjustment Mark	Category
12	MCHT_NO	Store Code	Category
13	MCC	Store Category	Category
14	ACQ_BIN	Acquiring Bank Code	Category
15	POS_ENTRY	Terminal Entry Method	Category
16	STIP	Proxy Authorization Mark	Category
17	MANUAL	Authorization Method	Category
18	CUS_CLASS	Credit Rating	Category
19	CUS_AVAIL	Available Balance	Number
20	CUS_LIMIT	Cardholder Limit	Number
21	RESP_ACT	Authorization Result	Category
22	APPR_CODE	Authorization Number	Category
23	POS_COND	Terminal Acquirer Status	Category
24	MERCH_NAME	Store Name	Category
25	COUNTRY	Transaction Country	Category
26	EDC_FUNC	EDC Transaction Code	Category
27	INSTALL_FLAG	Installment Mark	Category
28	BONUS_FLAG	Reward Redemption Mark	Category
29	FALLBACK	Chip to Magnetic Stripe Transaction Mark	Category
30	AC_FLAG	Mobile Payment Code	Category
31	DISPFLG	Fraud Mark	Category
32	FLAG_3D	3D Transaction Mark	Category

(*continued*)

Table 1. (*continued*)

NO	Variables	Explanation	Data Type
33	AGE	Age	Number
34	SEX_H	Cardholder Gender	Number
35	DATEOPEN1	Credit Card Issuance Date	Number

to bring their numbers down to around 1,000. This method can decrease model training time, which is advantageous for processing large datasets. However, its major drawback is the potential loss of representative data samples.

3. Synthetic Minority Oversampling Technique (SMOTE): Based on minority class data, SMOTE doesn't just copy existing minority class samples but generates synthetic new samples near selected data points to increase data diversity. The algorithm generates similar samples along the path connecting minority class fraud data with its neighbors, thus increasing the volume of data and achieving class balance. The advantage is maintaining data integrity and reducing overfitting risks, but it may also lead to the generation of specific patterns, which could cause overfitting to these patterns and increase computational costs for large samples.

3.3 Brief Review of Machine Learning Models

This study consists of 176,473 data entries, spanning from January 1, 2021, to May 31, 2023. Given the sequential nature of transaction data, this study segmented the data temporally, designating data from January 1, 2021, to December 31, 2022, as the training set, which totals 147,081 entries. Data from January 1, 2023, to May 31, 2023, was used as the test set, comprising 29,392 entries. Within the training data, there were 600 instances of fraudulent transactions and 146,481 normal transactions. In the test data, there were 543 instances of fraudulent transactions and 28,849 normal transactions.

To address the issue of data imbalance, the study employed sampling methods with ratios of 10:1 and 20:1 for normal and fraudulent samples. Through several iterations of different samplings and the application of ensemble learning techniques, multiple independent classifiers were trained. The results of these models were then consolidated to produce the final predictions. We employ three machine learning models for training: Logistic Regression, Random Forest, and XGBoost. Below is a concise concise description of the three models mentioned.

1. Logistic Regression:Logistic regression is a statistical model commonly used for binary classification problems, ideal for scenarios with two possible outcomes, such as default/no default, male/female, etc. It models the probability of an event occurring by transforming the linear combination of input variables (features) into a probability value between 0 and 1 using the logistic (or sigmoid) function. Its strength lies in its simplicity and interpretability, being a go-to model for many classification issues. However, it assumes a linear relationship between features and is sensitive to feature engineering quality.

2. Random Forest:Random forest is an ensemble learning method that builds upon decision tree algorithms, introducing randomness in the construction and prediction processes to enhance performance and generalization. It employs bootstrap sampling for training individual trees and randomly selects features at each node for splitting. The model's predictions are aggregated, usually by majority vote for classification and average or median for regression. Random Forest is powerful, particularly for medium to large datasets, and combats overfitting through its inherent randomness.
3. Extreme Gradient Boosting (XGBoost): XGBoost combines gradient boosting with regularization techniques to improve predictive performance and reduce overfitting, applicable to both regression and classification. Known for high predictive performance, it addresses various problem types and incorporates L1 and L2 regularization. XGBoost can handle missing data, supports parallel computing, offers cross-validation for optimal hyperparameter selection, and allows feature importance evaluation. It typically outperforms other algorithms, especially on large datasets, and provides flexibility with custom loss functions. However, it requires careful hyperparameter tuning and has higher memory demands, which may be challenging in resource-limited environments. XGBoost is primarily used for structured data and is less suited for unstructured data like images and text.

These models are essential tools in machine learning, each with its unique advantages and limitations. They are widely used across various industries for predictive analytics and decision-making processes.

4 Research Methodology

4.1 Model Comparison

In the context of credit card fraud detection, banks aim to identify as many potential fraudulent transactions as possible without generating excessive false positives that inconvenience customers. Therefore, this study utilizes the F1 score as the primary criterion for evaluating model performance, as the F1 score balances precision and recall.

Table 2 shows the predictive results of three machine learning models, including an ensemble learning method, under a sample ratio of normal to fraudulent transactions of 10:1. Overall, the Extreme Gradient Boosting (XGBoost) model demonstrates the best composite performance with the highest F1 and AUC scores, making it the optimal predictive model among the three. The Random Forest model also performs well but slightly less so compared to XGBoost. Logistic Regression, on the other hand, performs poorly in this highly imbalanced scenario, indicating room for model optimization.

In detail, Logistic Regression shows improved performance with an increased number of samplings, yet it scores lower in precision, recall, and F1 across all samplings, with fluctuating AUC values. Random Forest exhibits stability across all samplings, with high precision, recall, F1 scores, and AUC, indicating excellent performance despite the imbalance. XGBoost maintains high accuracy, precision, recall, and F1 scores across all samplings, with consistently high AUC values, suggesting superior comprehensive performance in fraud detection. Individual model results are analyzed as follows:

1. Logistic Regression:

- Accuracy gradually increases with the number of samples, ranging from 83.97% to 93.00%.
- Precision increases with the number of samples, indicating a decrease in the chance of making errors in predicting fraud transactions.
- Recall improves with the number of samples, indicating better capture of fraud transactions.

Table 2. Comparison of Model Results at Different Sampling Numbers

# of Samples	1	3	5	7	9
Logistic Regression					
Accuracy	83.97%	91.63%	93.23%	92.49%	93.00%
Precision	7.09%	20.68%	23.61%	21.26%	21.87%
Recall	61.69%	73.85%	76.76%	78.87%	79.15%
F1 Score	12.72%	31.30%	35.40%	32.71%	33.63%
AUC	73.05%	86.16%	88.47%	90.42%	86.26%
Random Forest					
Accuracy	98.58%	98.50%	98.46%	98.51%	98.49%
Precision	61.07%	59.08%	57.95%	59.10%	58.64%
Recall	69.61%	68.63%	69.61%	69.61%	69.70%
F1 Score	65.06%	63.49%	63.18%	63.90%	63.62%
AUC	97.78%	97.57%	97.62%	97.70%	97.67%
XGBoost					
Accuracy	98.32%	98.42%	98.32%	98.41%	98.44%
Precision	54.04%	55.92%	53.82%	55.73%	56.34%
Recall	76.43%	79.01%	81.69%	79.06%	79.23%
F1 Score	63.31%	65.48%	64.87%	65.36%	65.83%
AUC	98.42%	98.73%	98.50%	98.66%	98.60%

Data Source: Compiled for this study.

- F1 score increases gradually with the number of samples, but it is lower compared to the other two models.

2. Random Forest:

- Performance is stable across different sampling frequencies, with high accuracy (>98%) consistently.
- Precision remains around 60% across different sampling frequencies, indicating stable performance in predicting fraud.
- Recall shows relatively stable performance between 69–70%.
- F1 score maintains a high level across different sampling frequencies, indicating good balance between precision and recall.

3. XGBoost:

- Accuracy remains around 98% across different sampling frequencies, showing very stable performance overall.
- Precision remains above 50% across different sampling frequencies, indicating stable performance in predicting positive cases.
- Recall remains relatively stable around 79% across different sampling frequencies.
- F1 score maintains a high level across different sampling frequencies, indicating good balance between precision and recall.

From Table 2, it is evident that both Random Forest and XGBoost models outperform Logistic Regression in overall performance across different sampling frequencies. They exhibit good performance on highly imbalanced datasets and maintain stable performance. Overall, based on the evaluation metrics, the XGBoost model is the best fraud detection model, providing efficient predictive performance in practical applications.

In this study, we further explores the XGBoost model's top ten important features selected in the 10:1 sampling ratio scenario and provides explanations for their significance:

1. COUNTRY_dchange_180 (Change in transaction country in the last 180 days): Indicates whether there has been a change in the transaction country for the account in the past 180 days. Checking for recent changes in transaction countries may help detect abnormal transaction behavior.
2. TX_AMT (Transaction amount): Indicates the amount of each transaction. Large or very small transactions may carry significant risk, hence the model focuses on this feature.
3. COUNTRY1 (Transaction region is the USA, Canada): Indicates whether the transaction region is in the USA or Canada, suggesting transactions in these regions might be more complex and susceptible to fraud.
4. cum_past360_cnt (Cumulative transaction count in the last 360 days): Indicates the cumulative number of transactions for the account in the past 360 days. This feature helps identify accounts with unusual transaction frequencies.
5. MCC7273 (Merchant category is dating website): Transactions categorized as dating websites may pose risks, as these sites are common targets for fraudulent activities.
6. cum_past30_min_cnt (Cumulative transaction count in the last 30 min): Indicates the cumulative number of transactions for the account in the last 30 min. A sudden increase in transaction frequency may indicate abnormal account behavior.
7. POS_ENTRYphysical (Is it a physical transaction): Physical transactions may pose different fraud risks compared to online or virtual transactions.
8. RESP_ACTD (Is it a declined transaction): This feature indicates whether the transaction is marked as declined. Declined transactions may be triggered by abnormal behavior, making them important indicators for fraud detection.
9. COUNTRY2 (Transaction region is the UK): Similar to COUNTRY1, this feature indicates whether the transaction region is in the UK.
10. MERCH_NAME_dchange_360 (Change in transaction merchant in the last 360 days): Indicates whether there has been a change in the transaction merchant for the account in the past 360 days. Changes may indicate unusual transaction behavior for the account.

Considering these features, the XGBoost model primarily focuses on transaction country, amount, region, frequency, and merchant category. Changes in these transaction behaviors can be considered important information for fraud detection. For issuing banks, any changes in cardholders' transaction behaviors involving these features should warrant increased scrutiny for potential fraud or alternative methods of transaction verification.

4.2 Robustness Check

Apart from establishing models under a 10:1 ratio of fraud samples, Table 3 shows the model prediction results under different sampling frequencies with a fraud sample ratio of 20:1. Overall, when the normal to fraud sample ratio increases to 20:1, all three models still maintain high overall accuracy. Random Forest and XGBoost models remain superior, demonstrating stable and excellent performance in this highly imbalanced dataset. Logistic Regression performs relatively worse in this scenario but still shows some improvement with increasing sampling frequency.

These analyses demonstrate that even with a higher fraud sample ratio of 20:1, Random Forest and XGBoost models maintain superior performance, while Logistic Regression shows relatively poorer performance but still improves with increasing sampling frequency. Additionally, Table 3 presents the results under a 20:1 normal to fraudulent transaction ratio, using ensemble learning methods across different sampling frequencies. The Random Forest and XGBoost models remain preferable, maintaining stability and superior performance in this highly imbalanced dataset. Logistic Regression, while showing improvement with increased sampling, still lags behind the other two models.

In summary, based on the evaluation metrics, XGBoost is determined as the best model for fraud detection, providing efficient predictive performance for practical applications. This suggests that banks should closely monitor transactions with changes in the identified key features to promptly detect and prevent fraudulent activities.

5 Conclusion

This study utilized three commonly used machine learning algorithms, namely Logistic Regression, Random Forest, and XGBoost, for credit card fraud prediction. It conducted preliminary descriptive statistics on the credit card transaction data from F Bank to identify the distribution of features such as sample gender, education level, and transaction time. Moreover, it addressed the issue of imbalanced data in credit card fraud by using undersampling, employing different ratios of normal to fraud samples, and varying sampling frequencies along with ensemble learning methods to enhance model prediction accuracy and stability.

All three machine learning algorithms achieved good accuracy, with XGBoost performing the best among them. It exhibited superior performance in terms of accuracy, precision, recall, and F1 score, while providing key features helpful for fraud detection. Across different sampling ratios of normal to fraud samples, features such as change in transaction country in the last 180 days, transaction amount, transaction region in the USA or Canada, cumulative transaction count in the last 360 days, merchant category as dating websites, cumulative transaction count in the last 30 min, whether it is a physical transaction, and whether it is a declined transaction were selected as important variables. These findings provide reference for fraud detection for issuing banks to improve their fraud detection systems and reduce fraud risks. However, facing the continuously evolving fraud techniques, banks should also continuously monitor changes in cardholder transaction behavior and adjust their control strategies as fraud techniques evolve. This flexibility and continuous adjustment of strategies are crucial for coping with the ever-changing fraud techniques.

This study addressed the issue of imbalanced credit card fraud data while balancing computational efficiency and fully utilizing fraud samples by utilizing undersampling. It established three machine learning models using undersampling with normal to fraud sample ratios of 10:1 and 20:1. However, in practice, the actual ratio of fraud to normal samples may vary. Therefore, future research could explore different sampling ratios and establish different machine learning models to compare results. Additionally, this study only used credit card data from a single bank. If data from multiple banks with different credit cards could be combined, it might be possible to identify more key features and enhance model performance further.

References

Baabdullah, T., Alzahrani, A., Rawat, D.B.: On the comparative study of prediction accuracy for credit card fraud detection with imbalanced classifications. In: 2020 Spring Simulation Conference (SpringSim), pp. 1–12 (2020). https://doi.org/10.22360/SpringSim.2020.CSE.004

Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M., Imine, A.: Credit card fraud detection in the era of disruptive technologies: a systematic review. J. King Saud Univ. – Comput. Inf. Sci. **35**(1), 145–174 (2023). https://doi.org/10.1016/j.jksuci.2022.11.008

Duman, E., Ozcelik, M.H.: Detecting credit card fraud by genetic algorithm and scatter search. Expert Syst. Appl. **38**(10), 13057–13063 (2011). https://doi.org/10.1016/j.eswa.2011.04.110

Gadi, M.F.A., do Lago, A.P., Wang, X.: A comparison of classification methods applied on credit card fraud detection. Technical Report (2016)

Goyal, R., Manjhvar, A.K.: Review on credit card fraud detection using data mining classification techniques & machine learning algorithms. SSRN Scholarly Paper 3677692 (2020). https://papers.ssrn.com/abstract=3677692

Kovalenko, O.: Credit Card fraud detection using machine learning | SPD technology. software product development company (2023. https://spd.tech/machine-learning/credit-card-fraud-detection/

Patidar, R., Sharma, L.: Credit card fraud detection using neural network. Int. J. Soft Comput. Eng. (IJSCE) **1**(32–38) (2011). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0419c275f05841d87ab9a4c9767a4f997b61a50e

Prusti, D., Rath, S.K.: Fraudulent transaction detection in credit card by applying ensemble machine learning techniques. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6 (2019). https://doi.org/10.1109/ICCCNT45670.2019.8944867

Receiver operating characteristic. In: Wikipedia (2023). https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=1184185440

Şahin, Y.G., Duman, E.: Detecting credit card fraud by decision trees and support vector machines (2011). https://openaccess.dogus.edu.tr/xmlui/handle/11376/2366

Weston, D.J., Hand, D.J., Adams, N.M., Whitrow, C., Juszczak, P.: Plastic card fraud detection using peer group analysis. Adv. Data Anal. Classif. **2**(1), 45–62 (2008). https://doi.org/10.1007/s11634-008-0021-8