Proceedings of the 8th International Conference on Economic Management and Green Development
DOI: 10.54254/2754-1169/122/20242522

Enhancing Bank Credit Card Transaction Fraud Detection with Machine Learning Techniques

Yuhao Zhu[1,a,*]

[1]*University of Maryland Robert H. Smith School of Business, College Park, Maryland*
a. yzhu1212@umd.edu
**corresponding author*

Abstract: Because of the increasing prevalence and sophistication of credit card theft, standard detection measures frequently fail. This study investigates the use of several machine learning algorithms to improve fraud detection in bank credit card transactions. The research aims to develop enhanced fraud detection systems by employing an integrated strategy involving data preprocessing, the application of various classification algorithms, and performance evaluation. The study thoroughly examines algorithms such as Random Forest, Logistic Regression, and Neural Networks, focusing on their predictive capabilities and practical applications. The findings indicate that machine learning provides a robust framework for increasing the accuracy and efficiency of fraud detection systems, consequently assisting financial institutions in protecting customer transactions and strengthening security measures. Moreover, by prioritizing robust model selection and feature engineering, banks can significantly enhance their fraud detection performance. This research could greatly influence how financial institutions handle fraud, potentially reducing losses, improving operational efficiency, and securing transaction environments. The insights gained from this study are also valuable for informing broader financial security strategies and guiding future research in the field.

Keywords: Credit Card Fraud, Machine Learning, Random Forest, Neural Network, XGBoost

1. Introduction

In today's fast-paced digital world, financial transactions, particularly credit card purchases, are increasingly vulnerable to fraudulent behavior. As the global economy continues to digitize, credit card theft is becoming more prevalent, causing significant problems for both financial institutions and their customers. Consider, for example, one of the most notable incidents in history, in which a criminal ring stole more than £17 million by creating more than 32,000 fraudulent credit cards. This case illustrates the sheer scale and complexity of modern financial crime. [1] This study seeks to improve fraud detection in financial systems using machine learning to maintain consumer confidence and system integrity. It will explore how machine learning, through data collection, preprocessing, and classification algorithms, can enhance the detection of fraud in bank credit card transactions and respond to evolving threats. The findings could greatly influence how financial institutions handle fraud, reduce losses, secure transactions, and inform broader financial security strategies.

DOI: 10.54254/2754-1169/122/20242522

2. Case Study and Results

2.1. Machine Learning Algorithms for Classification

This research focuses on the predictive performance of various model-learning algorithms in credit card fraud detection, primarily assessing the likelihood of cardholder default. Yeh and Lien [2] evaluated six data mining techniques: discriminant analysis, logistic regression, nearest neighbor, artificial neural networks, and two decision tree methods. They introduced a "Sorting Smoothing Method" for more accurate default predictions, finding that precise probability estimates are more effective than binary classifications for financial risk management. According to Hastie, Tibshirani, and Friedman [3], understanding the theoretical underpinnings of these algorithms is essential for effectively applying them in practice. The main machine learning models used in this research include XGBoost, Random Forest, Decision Tree, Linear Regression, Logistic Regression, and Neural Network. Model evaluation relies on metrics derived from a confusion matrix, including the F1 score, precision, recall, and accuracy. [4] Precision is key for minimizing false alarms, recall assesses detection of actual fraud cases, and accuracy gauges overall model effectiveness. The F1 score is especially valuable for imbalanced datasets like fraud detection. These metrics are crucial for assessing performance and refining fraud detection algorithms, with each algorithm chosen based on specific needs and data traits to optimize predictions.

2.2. Implementation of the Machine Learning Algorithms

The data used in this work were taken from a dataset that is being publicly shared at Kaggle; the initial dataset was borrowed from the file "creditcard.csv" [5] and represented transactions made by European credit cardholders in a period of two days. Principal Component Analysis (PCA) was used to anonymize sensitive features into 28 components (V1-V28) to protect privacy and simplify data structures for fraud detection. The "Time" and "Amount" features were kept unchanged to preserve crucial transaction data, aiding in trend analysis and fraud anomaly detection. This showcases how feature engineering through PCA improves fraud detection by enhancing data analysis in financial settings. [6].

In addition, this study used StandardScaler to normalize the “Amount” and “Time” in the dataset, which can also increase the stability of the machine learning model. Normalization is important since it can reduce the influence of size disparities, which can make the performance of the algorithm less skewed. By converting these variables to a standard scale, we can improve the model's capacity to recognize outliers, especially in fraud detections, and make it more generalized. [7]

Here is how the data was cleaned.

Missing Values: this paper checked the dataset for missing values. Fortunately, the initial data showed no missing values, so we can confirm all features.

Scaling: the "Amount" and "Time" were scaled to match the PCA-transformed features. Therefore, StandardScaler was used to normalize these features to reduce the impact of scale differences on the machine learning algorithm.

Outliers: this research also applied outlier detection algorithms to identify the impact of transactions that are outside the normal range, especially when scrutinizing large transactions.

Duplicate Transactions: Duplicate transactions are examined to prevent biased results due to too many data points.

Category Imbalance: given the imbalanced nature of the fraud detection dataset, in which fraudulent transactions exceed valid ones, this project used the Synthetic Minority Oversampling Technique (SMOTE) to improve the prediction performance of minority category identification.

DOI: 10.54254/2754-1169/122/20242522

SMOTE combines fresh instances from the minority class to produce a balanced class distribution, which is required for good model training and assessment.

Saputra et al. [8] employed the SMOTE technique to balance a severely imbalanced Kaggle dataset, where only 0.093% of records were fraudulent transactions. Their use of SMOTE improved the dataset's balance, enhancing the average F1-Score and G-Score, thus demonstrating its effectiveness in preparing datasets for efficient fraud detection models.

The methodology section details using machine learning algorithms like linear regression and neural networks on a full dataset to detect fraudulent transactions. By grouping and testing key features through feature engineering, the study enhances model performance. This approach develops models that not only detect fraud effectively but also provide deep insights, serving as robust fraud prevention tools for banks and identifying new fraud patterns for further research.

2.3. Model evaluation

First, this study applied a full-scale dataset to test a series of machine learning models to identify credit card fraud. Specifically, models such as linear regression, logistic regression, decision trees, random forests, gradient boosting, and neural networks were included.

Linear regression: although commonly used for regression problems, the project tried the linear regression model to see how well it handles classification problems. The mean square error (MSE) and root mean square error (RMSE) of the model were 0.00083 and 0.0288, respectively, with an R-squared value of 0.5166, which shows some fitting ability.

Logistic regression: better suited for classification problems, the results of logistic regression show high precision and recall, especially for category 0 (non-fraudulent), which exhibits near-perfect recognition. However, for category 1 (fraud), although the precision is 0.86, the recall is lower at 0.58, indicating that the model still has room for improvement in recognizing true fraud.

Decision Tree: this model demonstrates a high recognition rate for non-fraudulent transactions and also performs well in the detection of fraudulent transactions with a precision of 0.71 and a recall of 0.79.

Random Forest: as an integrated learning method, Random Forest performs well in this task with almost perfect recognition of non-fraudulent transactions and also high precision and recall for fraudulent transactions with 0.97 and 0.79 respectively.

Neural Networks: finally, the research experimented with a neural network model that showed very high precision and loss reduction during training. On the validation set, the accuracy remained above 99.95%, demonstrating the potential of neural networks to handle complex pattern recognition problems.

After initial testing, the study refined our feature selection by identifying V14 as highly correlated with fraud detection and using it as the primary input. This simplified approach focuses on V14's impact to evaluate machine learning models in a constrained feature scenario, involving performance evaluations using only V14.

Random Forest:
Classification Report for Random Forest:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56864
1	0.97	0.79	0.87	98
accuracy			1.00	56962
macro avg	0.99	0.89	0.93	56962

DOI: 10.54254/2754-1169/122/20242522

weighted avg 1.00 1.00 1.00 56962

Confusion Matrix for Random Forest:
[[56862 2]
[21 77]]

Neural Network:

	precision	recall	f1-score	support
Non-Fraud	1.00	1.00	1.00	56864
Fraud	0.63	0.87	0.73	98
accuracy			1.00	56962
macro avg	0.81	0.93	0.86	56962
weighted avg	1.00	1.00	1.00	56962

Confusion Matrix for Decision Tree:

	Predicted Negative	Predicted Positive
Actual Negative	56814	50
Actual Positive	13	85

These initial full data model tests provide the basis for subsequent feature engineering and further model optimization. The next steps include more in-depth data analysis and feature selection to further improve the model's ability to predict fraudulent behavior.

After initial testing, this paper refined the feature selection by focusing on V14 due to its strong correlation with fraud detection. Using V14 as the primary model input, the study evaluated various machine learning models in a simplified feature space to identify the most effective model under feature-constrained conditions. This involved detailed performance evaluations using only the V14 feature.

Logistic regression: while logistic regression performs perfectly when dealing with the non-fraud category, its recall in the fraud category is only 0.35, indicating its low sensitivity in fraud detection.

Linear regression: linear regression is not suitable as a classification tool due to its extremely low R-squared value (0.0828), which suggests that it explains little of the variation in categories.

Decision trees: decision trees perform relatively well, but have a recall of 0.52 for fraud detection, which may be due to the over-reliance of decision trees on a single feature.

Random Forest: although the overall performance of Random Forest is usually strong, its recall for fraudulent transactions is also only 0.52 in this test, showing its limitations in the case of a single feature.

Neural network: the neural network performs the best among all the tested models, especially with a recall of 0.58 in the fraud category, which is higher than the other models, although there is still room for improvement.

In summary, the **neural network** model performs optimally when using a single feature V14.

Feature Importance
0 V14 1.0

	precision	recall	f1-score	support
0	1.00	0.96	0.98	56864

DOI: 10.54254/2754-1169/122/20242522

1	0.03	0.87	0.06	98
accuracy			0.96	56962
macro avg	0.52	0.91	0.52	56962
weighted avg	1.00	0.96	0.98	56962

Confusion Matrix for Decision Tree:

	Predicted Negative	Predicted Positive
Actual Negative	54423	2441
Actual Positive	13	85

Despite challenges in identifying complex fraud patterns, neural networks showed superior classification abilities in fraud detection, outperforming other models. This result indicates that complex models like neural networks can effectively discern nuances between categories using single features. To improve accuracy, the research emphasized a comprehensive feature set, choosing V14, V4, V12, and V8 based on their predictive value for credit card fraud, and then retrained and evaluated each model on these features to identify the most effective one.

Below is the performance and analysis of each model's performance with these four features:

Logistic regression: precision and recall perform well in this configuration, especially for the fraud category, with a precision of 0.82, a recall of 0.50, and an F1 score of 0.62. This shows the effectiveness of logistic regression for more complex combinations of features.

Linear regression: although not typically used for classification tasks, linear regression improved its R-squared value to 0.1627, indicating that the inclusion of more relevant features can slightly improve its explanatory power.

Decision Tree: performance improved, with precision and recall of 0.69 and 0.73 respectively, showing that decision trees are better at capturing the complexity of fraudulent behavior when dealing with multi-featured data.

Random Forest: of all the models, Random Forest performs the best with a precision and recall of 0.95 and 0.74, respectively, and an F1 score of 0.83. This further validates the efficiency of Random Forest when dealing with class-imbalanced and feature-rich datasets.

Neural Network: In tests involving multiple features and complex relationships, the neural network excelled, particularly in identifying non-fraudulent transactions, and also demonstrated high efficiency and accuracy in detecting fraud, achieving a precision of 0.88, a recall of 0.68, and an F1 score of 0.77.

Taken together, the Random Forest model performs optimally after accounting for the second significance difference.

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56864
1	0.95	0.74	0.83	98
accuracy			1.00	56962
macro avg	0.97	0.87	0.92	56962
weighted avg	1.00	1.00	1.00	56962

Confusion Matrix for Decision Tree:

	Predicted Negative	Predicted Positive

DOI: 10.54254/2754-1169/122/20242522

Actual Negative	56860	4
Actual Positive	25	73

The high precision and better recall of Random Forest indicate that it provides robust performance in dealing with fraud detection tasks with complex relationships and more features. This finding supports the use of the random forest model for credit card fraud detection in practical applications, especially in scenarios where feature selection is carefully analyzed.

2.4. Insights of the Analysis and implication for bank sector

Below are the combined evaluations of each model under these three data configurations (full-volume data, the first GAP, and the second GAP):

Logistic regression

Full data: moderate performance, recall and precision are OK, but there is room for improvement in the identification of fraudulent categories.

First gap: lower recall and higher precision for fraud categories.

Second gap: more balanced precision and recall, especially improvement in fraud category identification.

Linear regression

Full data: poorer performance, very low R-squared values, not applicable to classification.

First gap: still low R-squared values and limited classification ability.

Second gap: improved but still shows limited classification ability.

Decision Tree

Full data: good performance overall, especially in recognizing non-fraudulent categories.

First gap: average performance, with improved recall for fraud categories.

Second gap: significant improvement, especially in complexity and fraud recognition.

Random Forest

Full data: best performance among all models.

First gap: maintains a high level of performance, especially in recall.

Second gap: continues to demonstrate its superiority in handling datasets with complex features and class imbalance.

Neural Networks

Full data: performs well on complex datasets, especially in terms of recall and precision for fraud detection.

First gap: performs well on recall and precision for fraud categories.

Second gap: maintains high performance, especially in multi-feature environments.

Random forests and neural networks excel in all data configurations. Random forest stands out for its efficiency and stability in handling imbalanced and feature-rich datasets. Meanwhile, neural networks are favored for their flexibility and strong performance in complex, multivariate patterns.

When choosing a model, it should be considered based on the actual business requirements, expected model interpretability, and the complexity of the operation. For example, if the cost of false positives is high, Random Forest may be preferred; if the goal is to maximize the identification of potential frauds, neural networks may be a more appropriate choice. Krishna and Praveenchandar [9] analyzed data mining techniques like logistic regression and decision trees for credit card default prediction, noting the importance of selecting models based on data specifics and prediction needs. Their work offers practical insights for applying machine learning in fraud detection and guiding future research.

DOI: 10.54254/2754-1169/122/20242522

2.5. XGBoost

This study explored the efficiency of XGBoost [10] in credit card fraud detection, known for its efficiency and robust performance in handling large datasets and classification challenges. The models used the full dataset and tested two subsets of features identified by their importance—the first and second "big gaps." These tests provided insights into how XGBoost performs with different feature configurations. Below is a comprehensive evaluation of each model under these three data configurations (full data, first gap, second gap):

2.5.1. Full Data Model Analysis

	Feature	Importance
13	V14	0.563830
3	V4	0.072744
11	V12	0.046294
7	V8	0.031549
18	V19	0.019538
25	V26	0.017921
29	NormalizedAmount	0.015550
6	V7	0.015352
14	V15	0.014150
26	V27	0.013740
16	V17	0.013533
22	V23	0.013439
9	V10	0.012772
19	V20	0.012238
0	V1	0.011676
8	V9	0.011266
17	V18	0.011222
28	NormalizedTime	0.009967
1	V2	0.009962
20	V21	0.009852
2	V3	0.009749
15	V16	0.009495
24	V25	0.009347
21	V22	0.009022
4	V5	0.008496
10	V11	0.006673
23	V24	0.006402
5	V6	0.005652
27	V28	0.004605
12	V13	0.003964

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56864
1	0.99	0.81	0.89	98
accuracy			1.00	56962
macro avg	0.99	0.90	0.94	56962

DOI: 10.54254/2754-1169/122/20242522

weighted avg	1.00	1.00	1.00	56962

Confusion Matrix for Decision Tree:

	Predicted Negative	Predicted Positive
Actual Negative	56863	1
Actual Positive	19	79

Precision (Precision): reaches a perfect 1.00 for the non-fraud category and an unusually high 0.99 for the fraud category.

Recall (Recall): also a perfect 1.00 for non-fraudulent transactions.

F1 Score: 1.00 for the non-fraud category and 0.89 for the fraud category, showing a good balance.

Confusion matrix: very few false positives (only 1), while 79 cases of fraud were correctly identified.

2.5.2. First gap (feature V14)

Feature Importance

0	V14	1.0

	precision	recall	f1-score	support
0	1.00	0.96	0.98	56864
1	0.03	0.87	0.06	98
accuracy			0.96	56962
macro avg	0.52	0.91	0.52	56962
weighted avg	1.00	0.96	0.98	56962

Confusion Matrix for Decision Tree:

	Predicted Negative	Predicted Positive
Actual Negative	54423	2441
Actual Positive	13	85

Precision (Precision): reached 1.00 for the non-fraud category, but very low at 0.03 for the fraud category.

Recall (Recall): 0.96 for the non-fraud category and 0.87 for the fraud category.

F1 Score: 0.98 for the non-fraud category, but very low at 0.06 for the fraud category.

Confusion matrix: while most non-fraudulent transactions were correctly categorized, misclassification of fraudulent transactions was extremely high (2441 misclassified as non-fraudulent).

2.5.3. Second gap (features V14, V4, V12, V8)

Feature Importance

0	V14	0.752498
1	V4	0.113067
2	V12	0.080442
3	V8	0.053992

	precision	recall	f1-score	support

DOI: 10.54254/2754-1169/122/20242522

0	1.00	1.00	1.00	56864
1	0.79	0.80	0.79	98
accuracy			1.00	56962
macro avg	0.89	0.90	0.90	56962
weighted avg	1.00	1.00	1.00	56962

Confusion Matrix for Decision Tree:

	Predicted Negative	Predicted Positive
Actual Negative	56843	21
Actual Positive	20	78

Precision (Precision): remains perfect for non-fraud categories and improves dramatically to 0.79 for fraud categories.

Recall (Recall): remains perfect for non-fraud categories and improves to 0.80 for fraud categories.

F1 Score: 1.00 for non-fraud categories and 0.79 for fraud categories.

Confusion Matrix: misclassification was reduced significantly, with only 21 non-fraudulent transactions misclassified as fraudulent.

The XGBoost model demonstrates its unique performance in each of the three different data configurations for credit card fraud detection (full data, first big gap, second big gap). The following is a comprehensive analysis of XGBoost for these three scenarios:

Full Data Configuration

XGBoost achieved nearly perfect precision and recall in the full data configuration, showcasing its strength with all available features. The model's high precision and recall suggest effective identification and distinction between fraudulent and non-fraudulent transactions, with minimal false positives and omissions. This robust performance makes it highly suitable for real-world applications where accuracy is paramount.

First big gap configuration (Feature V14 only)

The first big-gap configuration, using only feature V14, shows good recall (0.87) but very low precision (0.03), leading to a high false alarm rate. This indicates that while V14 effectively identifies fraudulent transactions, it lacks sufficient discriminatory power when used alone, causing many normal transactions to be incorrectly flagged as fraudulent. This could result in significant misclassification costs in practical settings.

Second big gap configuration (features V14, V4, V12, V8)

The second big-gap configuration uses features V14, V4, V12, and V8, enhancing precision to 0.79 and recall to 0.80. This multi-feature approach reduces false alarms and omissions, proving effective for accurate fraud prediction in comprehensive evaluations.

Comprehensive Analysis

Comparing the three configurations, XGBoost performs best with full-volume data, maximizing the use of multiple features. The single-feature setup of the first big gap, while capturing some fraud, has low accuracy, limiting its utility. The multi-feature second big gap improves fraud detection and reduces false alarms, making it ideal for balanced feature selection and model design.

In summary, the selection and combination of features have a decisive impact on the model performance when XGBoost handles the task of credit card fraud detection. The full amount of data provides the best results, but in the case of limited feature availability or processing power, a reasonable combination of features can also achieve satisfactory results.

DOI: 10.54254/2754-1169/122/20242522

3. Challenges and Limitations

The credit card fraud detection dataset has a significant class imbalance, with far fewer fraudulent transactions than non-fraudulent ones. This imbalance can bias models toward predicting the majority class, potentially neglecting an accurate classification of fraud. Techniques like adjusting the scale_pos_weight in XGBoost are used to address this, but class imbalance still significantly impacts model generalization and effectiveness in practical applications.

Although the first and second thresholds in feature engineering are based on data insights, they may still fail to capture all the useful or critical factors for fraud detection. In particular, when conducting a model with the first threshold, the use of only a single feature, V14, leads to a high false positive rate. This suggests that when selecting a limited number of features for model training, the lack of critical information may cause the ineffectiveness of the model.

The differences in performance amongst models under different settings indicate that model selection and parameter setup must be fine-tuned for distinct application scenarios and data characteristics. For example, while XGBoost performed admirably under the full data scenario, performance plummeted dramatically when the features were limited. This necessitates more research into the model's application and adjustment method in order for it to demonstrate high generalization performance in a variety of settings.

The performance of the model includes precision, recall, and F1 score. However, these may be insufficient to accurately demonstrate full real-world model performance, most especially when issues of cost sensitivity (like the economic impact of false positives and omissions) come into play. Accordingly, in the future, the assessment criteria should be more dimensional, for instance, cost-benefit analysis, in order to cover the practical utility of the model.

4. Conclusion

The study's use of XGBoost for credit card fraud detection offers significant benefits for the banking sector by enhancing fraud prevention strategies, operational efficiency, and customer trust. Prioritizing model selection and feature engineering improves fraud detection accuracy, while XGBoost's scalability supports real-time data processing, essential for proactive fraud management. This leads to fewer service disruptions and higher customer satisfaction. This study employs diverse machine learning models, notably XGBoost, in credit card fraud detection, analyzing different feature configurations. While yielding promising results, further research expansion and depth are warranted. According to a recent bibliometric analysis of 11,870 conference papers, the IEEE International Conference on Data Mining has the highest publication rates in the field, underscoring the expanding scope of research topics and methodologies within the data mining community, which directly impacts advancements in fraud detection technologies. Future research could leverage deep learning methods like SVM, CNN, and RNN to improve fraud detection, develop real-time systems, and expand cross-domain approaches, enhancing security and adaptability in financial services.

References

[1] *uSwitch. (2019, December 29). Credit card fraud: the biggest card frauds in history. Retrieved from https://www.uswitch.com/credit-cards/guides/credit-card-fraud-the-biggest-card-frauds-in-history/*

[2] *Yeh, I., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems With Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020*

[3] *Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning. In Springer series in statistics. https://doi.org/10.1007/978-0-387-84858-7*

[4] *Lucas, Y., & Jurgovsky, J. (2010). Credit Card Fraud Detection Using Machine Learning: A survey. https://ar5iv.labs.arxiv.org/html/2010.06479#S1.SS2.SSS1*

[5] Credit card fraud Detection. (2018). [Dataset]. Kaggle. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

[6] Jolliffe IT, Cadima J. 2016Principal component analysis: a review andrecent developments. Phil.Trans.R.Soc.A374:20150202. http://dx.doi.org/10.1098/rsta.2015.0202

[7] Kaplan, S. (n.d.). Choosing between data standardization vs normalization: Key considerations [Ensure optimal model. EML. https://enjoymachinelearning.com/blog/data-standardization-vs-normalization-in-data-science/

[8] Saputra, A., & Suharjito, S. (2019). Fraud Detection using Machine Learning in e-Commerce. International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications, 10(9). https://doi.org/10.14569/ijacsa.2019.0100943

[9] Krishna, M., & Praveenchandar, J. (2022). Comparative Analysis of Credit Card Fraud Detection using Logistic regression with Random Forest towards an Increase in Accuracy of Prediction. 2022 International Conference on Edge Computing and Applications (ICECAA). https://doi.org/10.1109/icecaa55415.2022.9936488

[10] Chen, T., & Guestrin, C. (2016). XGBOOST: a scalable tree boosting System. https://arxiv.org/pdf/1603.02754