Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/47/20230366

Credit Card Fraud Prediction Based on Machine Learning Algorithms

Xinman Wang[1,a,*]

[1]*Information Management for Business, University College of London, Gower Street, London, United Kingdom*
a. xinman.wang.21@ucl.ac.uk
**corresponding author*

Abstract: The escalating use of the Internet has led to a surge in online shopping and e-commerce, resulting in a corresponding increase in credit card fraud incidents. Therefore, this research focuses on employing machine learning techniques, which offer enhanced precision and efficiency compared to manual detection, to identify fraudulent activities. To establish the association between credit card transaction attributes and the presence of fraudsters, this study initially gathers data from Kaggle, subsequently normalizing the collected data. Furthermore, the data exhibits severe imbalance, leading to overfitting concerns. To ascertain feature correlations, a correlation heatmap is constructed. Moreover, this investigation selects three models for analysis. Finally, the performance of each model is evaluated using a confusion matrix and derived metrics. The findings reveal that both the decision tree and random forest models exhibit optimal performance, achieving 100% across all indicators. The most influential factors in determining credit card fraud involve the ratio to median purchase price and the geographical proximity of the transaction location to the cardholder's residence.

Keywords: machine learning, credit card fraud prediction, business analysis

1. Introduction

A credit card is a payment card that people can use to pay in the form of a loan when they need money or use it to buy something they need, and they must be repaid with interest [1]. The advent and exponential expansion of E-Commerce have engendered a substantial upswing in the utilization of credit cards as a means of conducting online transactions. This surge in credit card usage has concomitantly precipitated a notable escalation in the occurrence of credit card fraud. As credit cards have become the preferred payment method for both online and regular purchases, the incidence of associated fraud has also risen [2]. Credit card fraud encompasses two main categories: authorized and unauthorized transactions. Authorized fraud occurs when a legitimate cardholder processes a payment to a criminal-controlled account, whereas unauthorized fraud takes place when the account owner fails to approve the payment, and a third party conducts the fraudulent act. In the United Kingdom, the financial losses due to unauthorized fraud reached a staggering £844.8 million in 2018. However, the proactive efforts of banks and card companies successfully prevented £1.66 billion in unauthorized fraud during the same period, effectively stopping around £2 out of every £3 of attempted fraudulent activities [3]. Within the United States, retailers encounter an average monthly incidence of approximately 1,740 fraudulent activities, where more than half of these endeavors result

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/47/20230366

in successful perpetration, thereby signifying a significant shift in the year 2021 wherein the prevalence of fruitful fraud attempts supersedes that of thwarted ones [4]. Identity theft affects approximately 7-10% of U.S. adults annually, with an alarming 80% of credit cards in circulation having been compromised [5]. To combat credit card fraud, two methods are employed: traditional financial models and artificial intelligence. Traditional models exhibit limited accuracy and relatively sluggish predictive capabilities in identifying fraudulent activities. In contrast, artificial intelligence methods aim to address the deficiencies inherent in these models and provide more effective fraud detection and prevention mechanisms.

In this research, predicting credit card fraud used supervised learning. To be more specific about supervised learning, the labeled dataset is given to the model for training, the intended result is known [6]. When it comes to the model prediction phase, the trained model will predict new data, different from the training set data [6]. In the early stages of machine learning for credit card fraud prediction, neural networks and Bayesian networks were commonly utilized. Neural networks consist of interconnected artificial neurons and the feedforward network, comprising input, hidden, and output layers, it processes data and transfers it to the following layer [7-9]. For the neural network, this model is first trained on non-credit card fraud data, and then predicts data with fraud, and separates suspicious and non-suspicious data through the model to predict credit card fraud [2]. Moreover, for Bayesian networks, which also called belief networks, they excel at modelling scenarios when incoming data is ambiguous or partially unavailable, but some information is already known, and the belief is then utilized for identification of patterns and data categorization [2].

In this article, three different machine learning models are used in predicting credit card fraud. Given the nature of the dataset, which exhibits an imbalance between the number of fraud cases and non-fraud cases, certain data processing techniques were employed to address this challenge. One crucial step in data processing was the normalization of the data. This process ensures that all the features are on a uniform scale, enabling fair comparisons and avoiding any characteristic from being the sole driver of the model's learning process. Besides, oversampling was also adopted. Among the three models, it was observed that the decision tree and random forest model achieved the highest performance, demonstrating its superior accuracy in predicting credit card fraud.

2. Methodology

2.1. Data Collection

The data set used is from Kaggle [10], which is about credit card fraud. The data provided with seven details about customer, including the distance from home etc. [10]. The total amount of data is 1,000,000. The purpose of the data set is to judge whether there is credit card fraud, where 0 means no fraud and 1 means there is fraud.

2.2. Data Processing

In the process of data processing, there are two problems with the data. The first is that the values between different features are not on the same scale. To solve this problem, the data normalization method, standardization, is adopted. To be more specific, different feature scales are scaled to a similar scale range.

Second, the data is unbalanced. The amount of data that was not fraudulent was approximately nine times the amount of data that was fraudulent. If the data is not balanced, it will lead to inaccurate model presdiction results. Therefore, the method of oversampling data is adopted. To be more specific, the number of ‘0’ is 912597, accounting for 91.3%, and the number of '1' is 87403, accounting for 8.7%, after oversampling the data, the number of '1' is equal to the number of '0'.

DOI: 10.54254/2754-1169/47/20230366

2.3. Data Analysis

There is a data imbalance problem in the original data, the number of ‘0’ is 912597, accounting for 91.3%, and the number of '1' is 87403, accounting for 8.7%. After oversampling the data, the number of '1' is equal to the number of '0'.

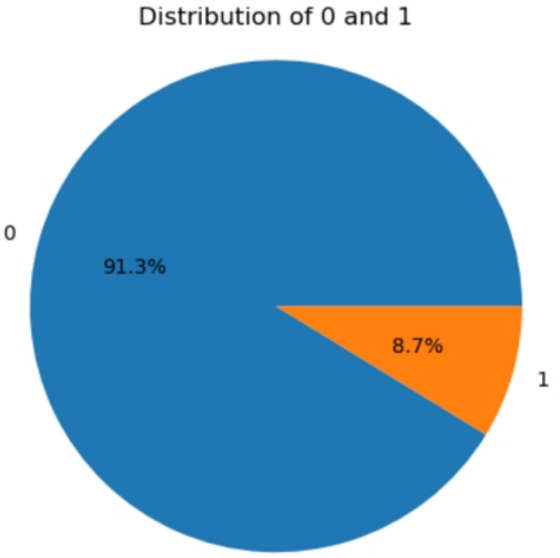

Figure 1: The pie chart of the distribution of 0 and 1 in original data (Photo/Picture credit: Original).

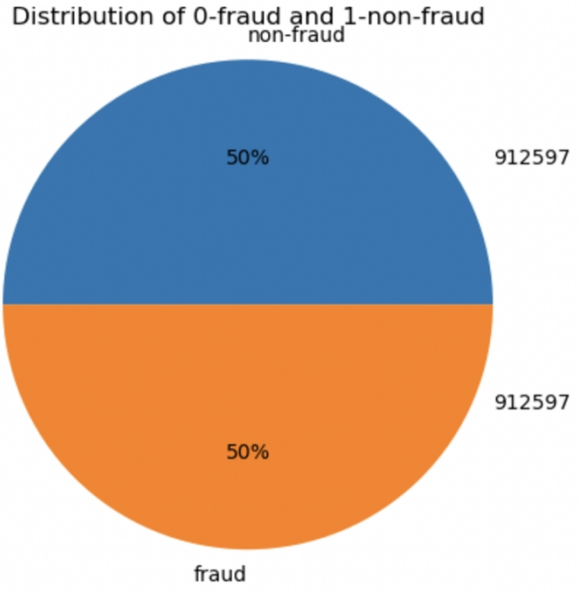

Figure 2: The pie chart of the distribution of 0 and 1 after oversampling the data (Photo/Picture credit: Original).

To understand the degree of correlation between features, a correlation heatmap is created. From the figure, it can be found that the two features with the strongest positive correlation are 'distance from home' and 'repeat retailer', which is 0.16, and the most negative correlation is 'distance from home' and 'used chip', which is -0.11. However, the absolute value of these two correlations is lower

DOI: 10.54254/2754-1169/47/20230366

than 0.5 and closer to 0, reflecting a relative lower correlation. Thus, feature selection and dimensionality reduction operations are considered unnecessary.

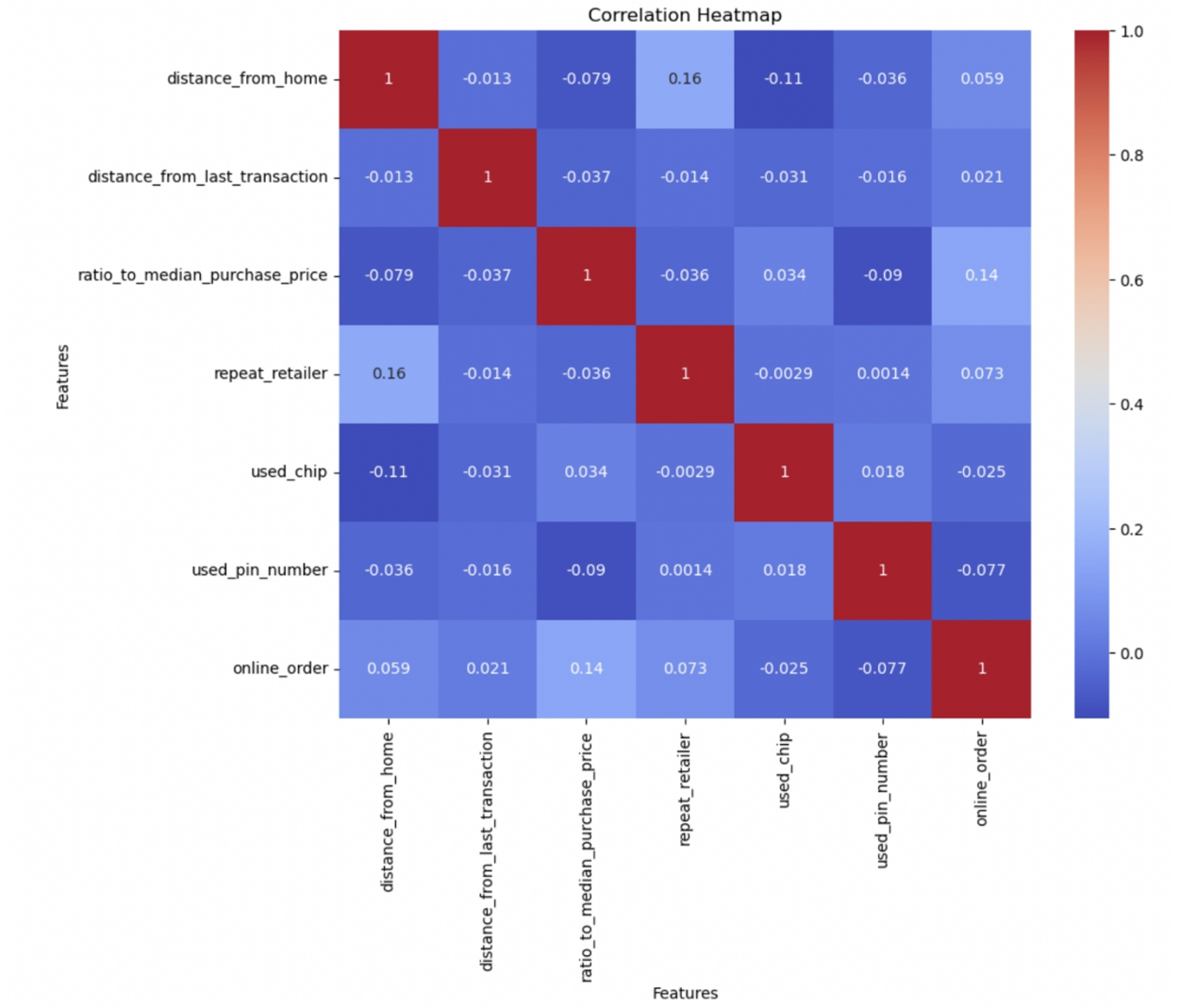

Figure 3: The correlation heatmap of the data (Photo/Picture credit: Original).

2.4. Machine Learning Models

Three models are used in this research. In order to evaluate the performance of the models, this research measures the performance of each model through the confusion matrix, and some derived indicators calculated.

2.4.1. Logistic Regression

The link between the possibility of a targeted variable and a linear combination of independent factors is constructed using a logistic regression model [11]. Since the target variable is binary, a logistic regression model was preferred. In view of the strong interpretability of the logistic regression model, this paper calculates the coefficient of each feature.

DOI: 10.54254/2754-1169/47/20230366

2.4.2. Decision Tree

In the realm of classification and prediction problems, the decision tree model serves as a prominent machine learning technique. It assumes the form of a hierarchical structure resembling a tree, wherein each internal node represents a specific attribute test, every branch corresponds to the outcome of the test, and each leaf node signifies a particular classification or prediction [12].

In this study, the maximum depth of the hyperparameters of the decision tree model is set to be equal to seven.

2.4.3. Random Forest

In many research contexts, the machine learning technique of categorization of random forests is employed to create prediction models. In order to relieve the load of data collection and boost efficiency, the primary objective of modelling for prediction is typically to minimise the number of variables required to achieve a forecast [13].

Since the hyperparameters were all default values during the first model fitting, and the predicted result accuracy rate was very high, reaching 1.0, so this article did not adjust the hyperparameter values. In addition, this paper also calculates the value of feature importance for the random forest model.

3. Results and Discussion

3.1. The Performance of Various Models

To evaluate the performance of each model, this research uses the confusion matrix and its derived indexes. To be more specific, it shows the counts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions. True or False represent the actual labels, while Positive or Negative represent the predicted labels.

It should be noted that, TP means the number of instances that are actually positive (belong to the positive class) and are correctly predicted as positive by the model. But in the confusion matrix of the model, True and Positive represent the actual and predicted quantities of ‘0’, and ‘0’ does not mean that there is credit card fraud. The labels in the confusion matrix cannot represent whether the actual data is positive or negative.

3.1.1. Logistic Regression

The performance evaluation of logistic regression is presented in Table 1, where diverse indicators are utilized as measures of its effectiveness.

Table 1: The performance of the logistic regression.

Accuracy score	0.941
F1 score	0.942
Recall score	0.962
Precision score	0.923

According to the confusion matrix shown in Figure 4, it can be observed that there are 18, 978 credit card fraud and 175, 266 are non-fraud. A total of 2123 of them were predicted incorrectly.

DOI: 10.54254/2754-1169/47/20230366

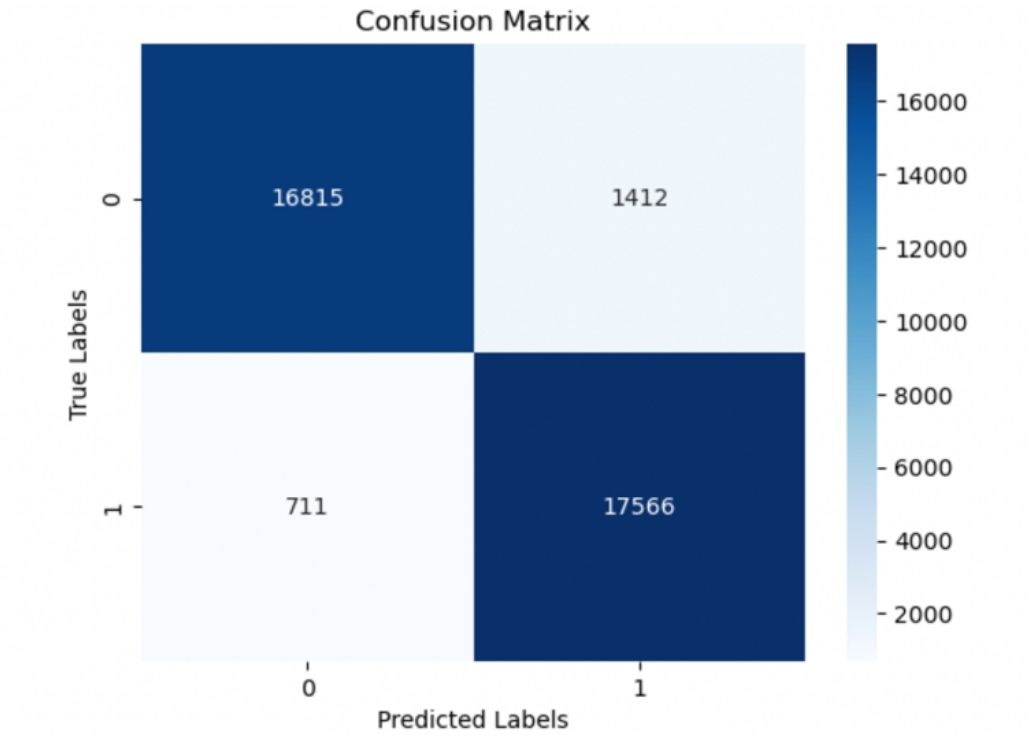

Figure 4: The confusion matrix based on the logistic regression model (Photo/Picture credit: Original).

3.1.2. Decision Tree

At the beginning, the hyperparameter max_depth of the decision tree model is set to 3, and it can be observed that False Negative (FN) is 412 and False Positive (FP) is 530. However, when the maximum depth is gradually increased until all leaf nodes of the model only contain samples of the same type (max_depth=7), the performance of the model is gradually improved to 0 for both FN and FP.

DOI: 10.54254/2754-1169/47/20230366

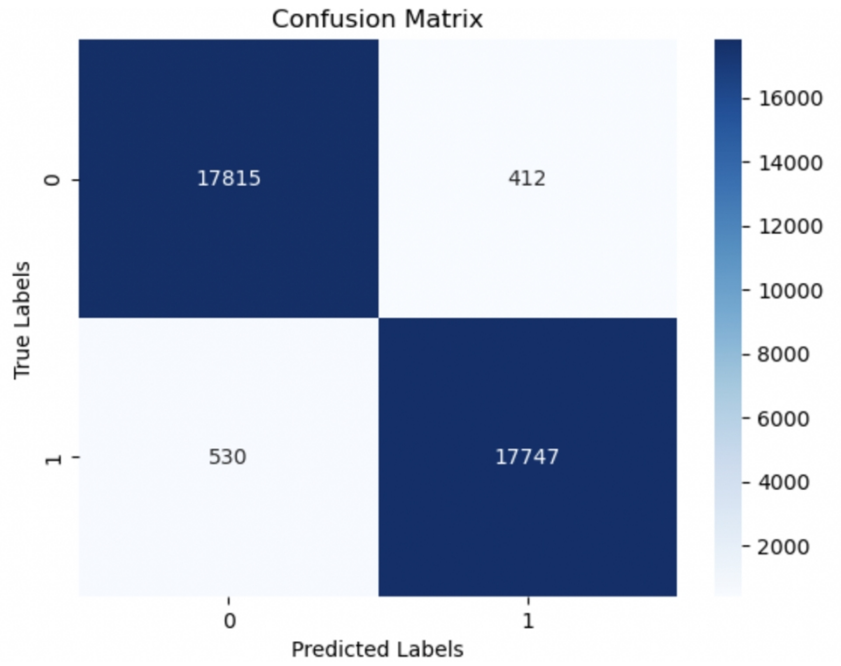

Figure 5: The confusion matrix based on the decision tree model (Photo/Picture credit: Original).

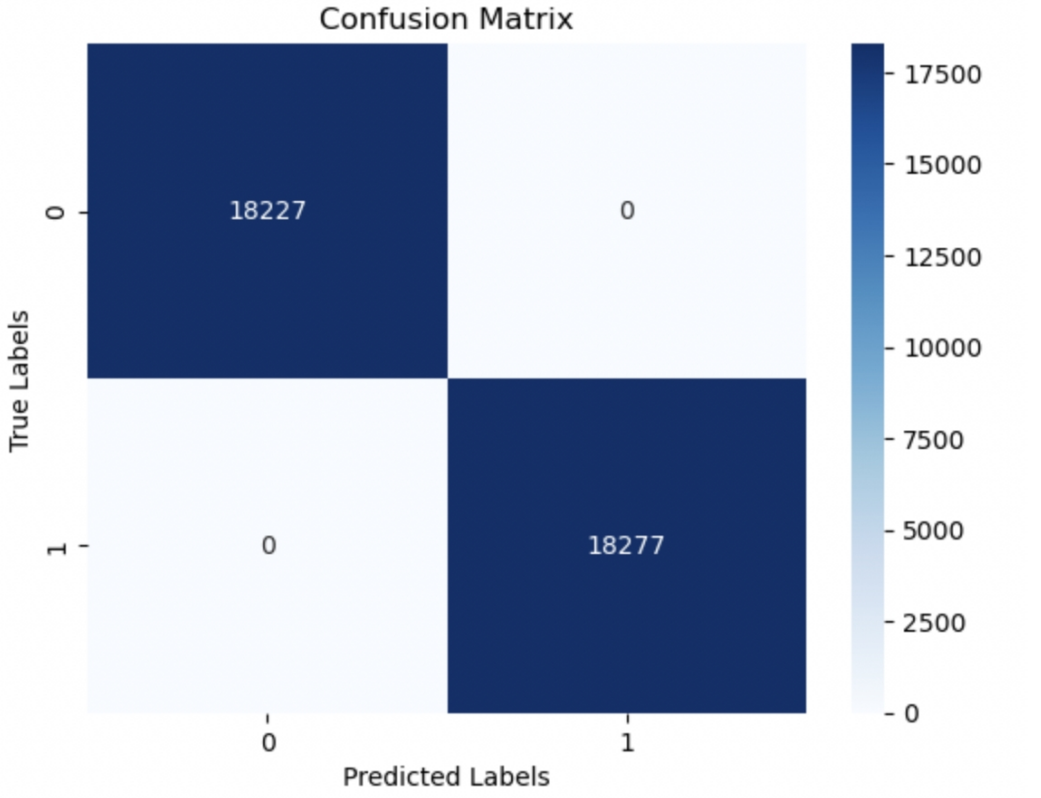

Figure 6: The confusion matrix based on the decision tree model after modification (Photo/Picture credit: Original).

DOI: 10.54254/2754-1169/47/20230366

Following an intricate process of hyperparameter tuning, the decision tree model exhibited exemplary performance, demonstrating exceptional accuracy in its predictions shown in Figure 6 and Table 2.

Table 2: The performance of the decision tree.

Accuracy	1.0
F1 score	1.0
Recall	1.0
Precision	1.0

3.1.3. Random Forest

The performance of the random forest model is similar to that of the decision tree model shown in Table 3 and Figure 7, and all indicators have reached the highest.

Table 3: The performance of the random forest.

Accuracy	1.0
F1 score	1.0
Recall	1.0
Precision	1.0

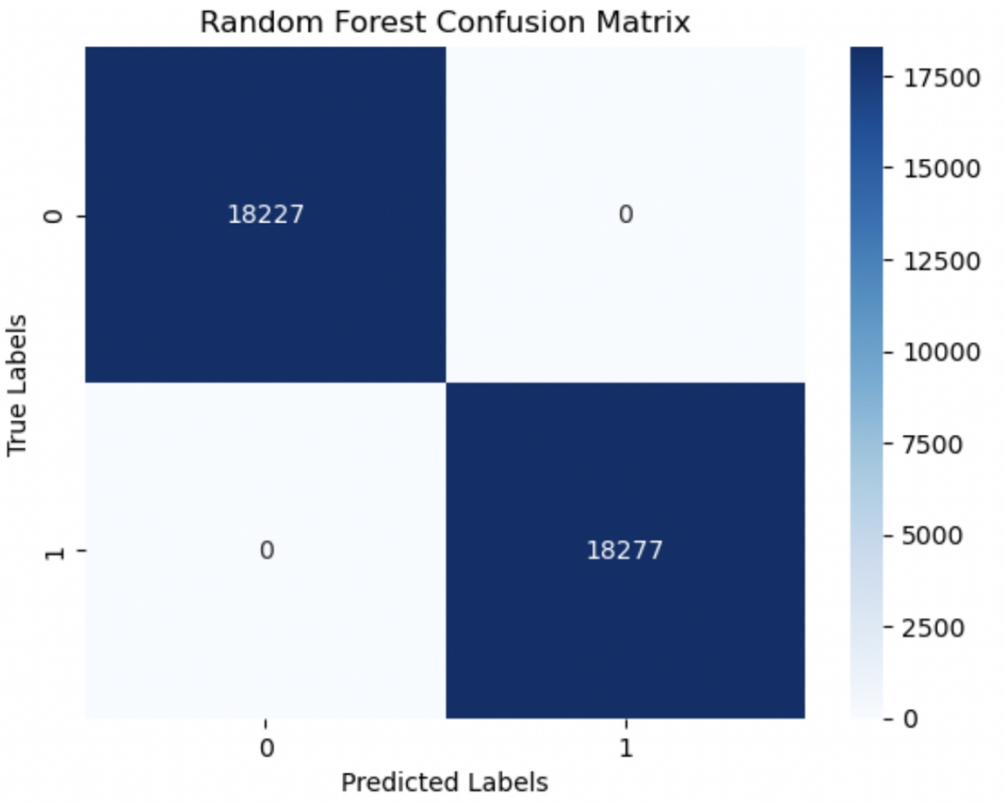

Figure 7: The confusion matrix based on the random forest model (Photo/Picture credit: Original).

DOI: 10.54254/2754-1169/47/20230366

3.1.4. Discussion

Table 4 is the summary of the four indicators of the three models.

Table 4: The summary of the four indicators of the three models.

	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.941	1.0	1.0
F1 score	0.942	1.0	1.0
Recall	0.962	1.0	1.0
Precision	0.923	1.0	1.0

According to the comparison of these four indicators, it is apparent that both the Decision Tree and Random Forest models exhibit superior performance compared to the Logistic Regression model in predicting credit card fraud. Consequently, it is not advisable to rely on the logistic regression model for this specific task. The reason is that this study believes that judging credit card fraud is an important decision, and in the original data set, the number of frauds is much smaller than the number of non-frauds, so the accuracy of the model is required to be higher. In this study, an f1 value of 0.94 may not be considered high enough, so a more precise model is recommended.

3.2. The Feature Importance of Models

The main purpose of calculating the feature coefficients is to understand and analyze the influence degree of the features on the target variable. The feature coefficients can help to determine which features have a strong correlation with the prediction of the target variable, which can be used for feature selection, feature engineering, and model interpretation. Table 5 provides coefficients of features of logistic regression.

Table 5: The feature importance and corresponding coefficient.

Feature	Coefficient
distance_from_home	1.0650302725950727
distance_from_last_transaction	1.0650302725950727
ratio_to_median_purchase_price	1.9757359893784288
repeat_retailer	0.1899534548720653
used_chip	- 0.31917332361128486
used_pin_number	- 0.8139428273342978
online_order	1.1654815594413264

Among them, the absolute value of the correlation coefficient of 'ratio_to_median_purchase_price' is the highest, about 1.98. Moreover, the other two models, decision tree and random forest, are calculated the feature importance shown in Figure 8 and Figure 9. According to figure and figure, it can be easily seen that the feature 'ratio_to_median_purchase_price' is far more important than other features, followed by ‘distance_from_home’.

DOI: 10.54254/2754-1169/47/20230366

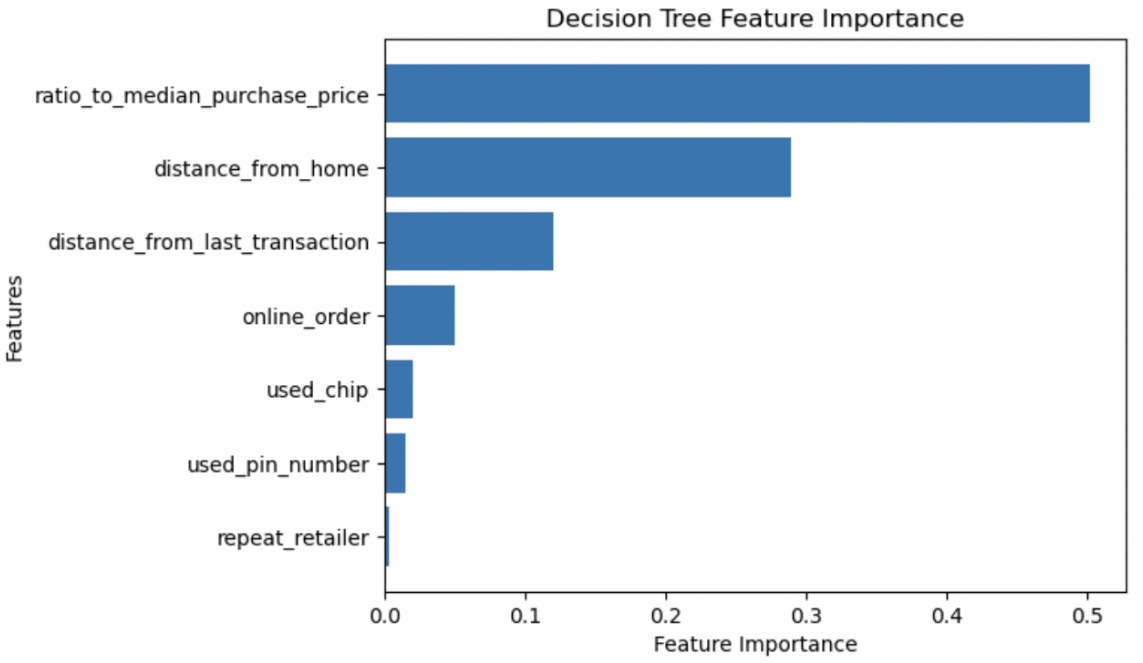

Figure 8: The feature importance of the decision tree (Photo/Picture credit: Original).

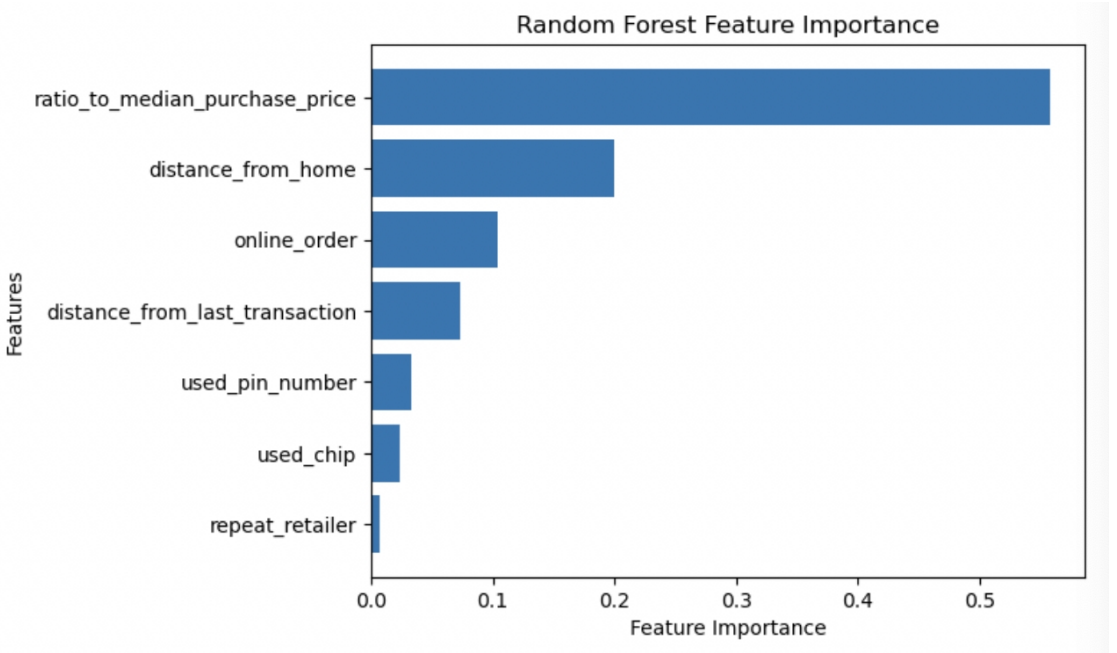

Figure 9: The feature importance of the random forest (Photo/Picture credit: Original).

The results demonstrated that the ratio to median purchase price has a significant impact on the target variable, and the distance from home is also an important indicator for determining whether there is credit card fraud. According to feature importance, this study may provide useful insights for predicting the domain of credit card fraud.

DOI: 10.54254/2754-1169/47/20230366

4. Conclusion

In this study, machine learning methods are used to predict whether there is credit card fraud. Three main models used in this study. In order to improve the performance of the model, some major data processing methods are used, including data normalization and data oversampling. With the aim to measure the performance of the model, the method used is the confusion matrix and its derived four indicators. According to the experimental results, decision tree model and random forest model are recommended to predict credit card fraud. Among the many determinants of credit card fraud, the two most important features are the ratio to median purchase price and the distance from home where the transaction happened. In the future, research can consider integrating more data sources and adding more features, which can further improve the predictive ability of the model. Besides, future research can further explore different feature selection methods and model optimizing techniques to improve the accuracy of the model.

References

[1] *Saltz, J. S. (1996). Another Year of Credit Card Late Charge Cases: The Search for a Definition of "Interest" Continues. The Business Lawyer, 51(3), 925–931. http://www.jstor.org/stable/40687670*

[2] *Benson, E. R. S., & Annie, P. A. (2011, March 1). Analysis on credit card fraud detection methods. https://doi.org/10.1109/ICCCET.2011.5762457*

[3] *UK Finance. (2019). FRAUD THE FACTS 2019 The definitive overview of payment industry fraud. Retrieved June 28, 2023, from UK Finance website: https://www.ukfinance.org.uk/policy-and-guidance/reports-publications/fraud-facts-2019#:~:text=Fraud%20poses%20a%20major%20threat*

[4] *Steele, J. (2021, June 11). Credit card fraud and ID theft statistics. Retrieved from CreditCards.com website: https://www.creditcards.com/statistics/credit-card-security-id-theft-fraud-statistics-1276/*

[5] *Schulte, T. (2021, July 15). 50+ Identity Theft & Credit Card Fraud Statistics (2021). Retrieved from Define Financial website: https://www.definefinancial.com/blog/identity-theft-credit-card-fraud-statistics/*

[6] *Rebala, G., Ravi, A., Churiwala, S. (2019). Machine Learning Definition and Basics. In: An Introduction to Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-15729-6_1*

[7] *Han, S., Pool, J., Tran, J., & Dally, W. (2015) Learning both weights and connections for efficient neural network. Advances in neural information processing systems, 28.*

[8] *Yu, Q., Chang, C. S., Yan, J. L., et al. (2019)Semantic segmentation of intracranial hemorrhages in head CT scans, 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2019: 112-115.*

[9] *Lo, S. C. B., Chan, H. P., Lin, J. S., et al. (1995) Artificial convolution neural network for medical image pattern recognition. Neural networks, 8(7-8): 1201-1214.*

[10] *Kaggle (2023) Credit Card Fraud https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud/code?datasetId=2156255&sortBy=voteCount*

[11] *Srimaneekarn, N., Hayter, A., Liu, W., & Tantipoj, C. (2022). Binary Response Analysis Using Logistic Regression in Dentistry. International Journal of Dentistry, 2022, 1–7. https://doi.org/10.1155/2022/5358602*

[12] *You, J., Li, G., & Wang, H. (2021). Credit Grade Prediction Based on Decision Tree Model. 2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE). https://doi.org/10.1109/iske54062.2021.9755326*

[13] *Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. Expert Systems with Applications, 134, 93–101. https://doi.org/10.1016/j.eswa.2019.05.028*