Credit Card Fraud Detection Prediction: Machine Learning Algorithm

Yi Qu[1(✉)] and Jiani Jin[2]

[1] Department of Finance, Wenzhou-Kean University, Wenzhou, China
qyi@kean.edu
[2] Department of Mathematics, Wenzhou-Kean University, Wenzhou, China
jinji@kean.edu

Abstract. Someone commits payment fraud when they obtain the payment information of another person and use it for unauthorized transactions or purchases. Owing to the ease and convenience of e-commerce, digital purchasing is becoming increasingly popular today and because of the convenience of online shopping, many individuals prefer to shop online. This has resulted in a substantial rise in credit card fraud. Detecting and preventing payment fraud is difficult because the standard rules-based anti-fraud systems deployed by banks cannot manage the high volume of online transactions. This creates unique difficulties for banks and a substantial increase in losses. Therefore, it is crucial to effectively identify and eliminate fraud. In our research, we use machine learning methods to construct models that can detect and analyze fraudulent payments. We primarily employ the Generalized Linear, Decision Tree, Gradient Boosting, and Naive Bayes Models, and determine that the Generalized Linear Model is the most effective.

Keywords: credit card fraud · machine learning · Generalized Linear Model · Decision Tree · Gradient Boosting · Naïve Bayes

1 Introduction

Financial fraud has become a significant problem in modern society. When the payment information does not match the actual beneficiary, the business may object, and the bank may be required to pay a large sum of money, issue a refund, and conduct an investigation to resolve the dispute. Considering the risk of fraud, if a merchant repeatedly discovers that a transaction is not secure, the business ceases working with the bank. It negatively impacts the bank's business and leads to financial losses and socioeconomic instability for the company [1]. Therefore, banks must detect and prevent financial transaction fraud in advance.

One of the most popular financial tools, is the credit card payment card issued by a bank to a user (cardholder). It has the advantage of eliminating the need for metal currency and banknotes while expanding the scope of exchange. Its widespread use also means that business credit card fraud has become one of the main types of payment fraud [2]. In 2020, nearly 400,000 Americans fell victim to credit card fraud; in 2021,

S. H. B. D. M. Zailani et al. (Eds.): ICMSEM 2023, 259, pp. 760–767, 2024.
https://doi.org/10.2991/978-94-6463-256-9_77

the Federal Trade Commission received nearly 390,000 reports of credit card fraud, making it the most prevalent type of fraud in the United States [3].

Our research makes extensive use of machine learning (ML) techniques to analyze and forecast credit card fraud. During a research collaboration on extensive data mining and fraud detection, the ML departments of Worldline and the Free University of Brussels gathered and analyzed a dataset consisting of credit card transactions, which we use in here [4]. Notably, fraud accounts for only 0.172% of all transactions, indicating the imbalanced nature of the dataset.

We divide the data set into two parts, namely the training and test datasets, and the ratio between them is 7:3. The optimal ML strategy is selected by comparing and evaluating the accuracy of four models: the Generalized Linear Model (GLM), the Decision Tree (DT) Model, the Gradient Boosting Model (GBM), and the Naive Bayesian (NB) Model. We use the area under the curve (AUC) as the criterion for evaluating the models' precision.

The GLM has the highest AUC (0.971), followed by the DT Model (0.909), the GBM (0.950), and the NB Model (0.938). Thus, credit card fraud is best modeled using the GLM.

2 Related Work

Numerous studies have used ML models to detect bank payments, especially credit card payments. Owing to advances in information technology, fraud detection must be improved and changed in response to the global expansion of payment.

Nami and Shajari created a technique for identifying fraudulent payment card transactions in 2018, which consists of two steps: using sliding windows and K-Nearest Neighbors (KNN) in the first stage of the detection, and the Distributed Random Forest algorithm and cost-sensitive detection in the second stage [5]. A minimum-risk model was also used. Their investigation utilized a private bank's actual dataset. When there are a large number of input features, the Distributed Random Forest algorithm is the most efficient option because it is characterized as a precise and rapid learning strategy that can effectively operate on massive datasets.

Popat and Chaudhary's 2018 study focused on fraud prevention and detection [6]. The purpose of fraud detection is to determine the difference between fraudulent and legitimate transactions. Classification, clustering, prediction, outlier detection, regression, and visualization are the six data mining techniques they used. The study introduced an Artificial Immune System, Bayesian Belief Network, Logistic Regression (LR) Neural Network, Support Vector Machine, Genetic Algorithm (GA), Decision Tree (DT), Self-organizing Map, hybrid methods, etc., and concluded that ML methods are superior to conventional prediction method, clustering, and outlier detection.

In 2019, Vidanelage et al. imitated the bank payment system, created a synthetic dataset for fraud detection research, performed data preprocessing, and classified fraudulent payments using KNN, Multilayer Perceptron (MLP), Gaussian NB, and Multinomial NB [7]. The results showed that all methods achieved an accuracy of over 90%, with MLP being significantly more accurate than the others.

To increase the precision of fraud detection, Fabrizio et al. studied a hybrid technique in 2019 that combines supervised and unsupervised methods [8]. The supervised method

takes advantage of past fraudulent activities, whereas the unsupervised method focuses on discovering new ones. Fraud category. The study identified outliers using the Z-score, PC-1, PCA-RE-1, IF, and GM-1 outlier scores and obtained global, local, and cluster granularity. The hybrid approach demonstrated superior performance, and the influence of granularity on the method's precision revealed the importance of hierarchical dataset analysis.

Emmanuel et al. proposed an ML-based credit card fraud detection engine in 2022 that uses a GA to select features [9]. After optimizing the features, the following ML classifiers were employed: DT, Random Forest, LR, ANN, and NB. The results indicated that GA-DT obtained an AUC of 1 and a precision of 100%. This is then followed by GA-ANN, which had an AUC of 0.94 and 100% accuracy.

In 2022, Alfaiz and Fati examined a total of 66 MLmodels to investigate credit card fraud, and the best model was the combination of AllKNN and CatBoost (AllKNN-CatBoost) [10]. Furthermore, AllKNN-CatBoost was contrasted to previous research utilizing the same dataset and comparable techniques. The F 1-score (87.40%) demonstrates that AllKNN-CatBoost provides a significant improvement. Alharbi et al. proposed an ML solution in 2022 to detect fraudulent credit card transactions, a novel text2IMG transformation technique that produces compact images. Deep Learning (DL) and ML techniques were utilized to validate the system's robustness and efficacy. Using the deep features of the proposed convoluted neural network, Coarse-KNN achieved 99.87% accuracy [11].

3 Datasets and Methods

3.1 Dataset

Kaggle provided the dataset we used for our research, which includes credit card purchases conducted throughout Europe in September 2013 [4]. This data mostly comprises information from the most recent two days, and out of 284,807 transactions, there have been 492 instances of fraudulent activity. Thus, fraudulent activity accounts for only 0.172% of all transactions.

The data consists of 31 columns, which are the digital input variables of the Principal component analysis (PCA) conversion results. The features V1, V2,… V28 are the principal components derived via PCA, whereas the amount, number, and class are not converted via PCA.

The transaction amount can be used for instance-dependent cost-sensitive learning. The time feature stores the number of seconds that have passed since the beginning of the dataset until each subsequent transaction in the dataset is processed. Class is assigned the value of 1 if fraud is detected, and otherwise, 0 is assigned.

Since fraud accounts for a small proportion of all transactions, the data set is highly imbalanced, and the confusion matrix accuracy is not suitable for imbalanced classification, so we use a different method to make predictions, and we use AUC to measure accuracy.

3.2 Methods

Four methodologies will be utilized for forecasting and analysis. Generalized Linear Model, Decision Tree, Gradient Boosting, and Naive Bayes are these four techniques.

The first GLM, which is an umbrella term for models that allow the response variable y to have a normal distribution and a distinct error distribution. Linear Regression, LR, and the Poisson distribution are all examples of such models. Furthermore, GLM models helps to establish a linear relationship between the response variable and the predictor variables, even if the basic relationship between the response variable and the predictor variables is not linear [12]. In Nonlinear Regression models, the response variable's error distribution does not need to be normal. The errors in the response variable are assumed to follow a binomial distribution in this data set.

DTs, also known as prediction trees, are nonparametric supervised learning models for classification and regression. The tree structure is used to represent decision sequences and outcomes, as illustrated in Fig. 1. The objective is to develop a model that can predict whether a credit card is fraudulent using ground rules inferred from the characteristics of the dataset [13]. Trees can be regarded as approximations that are piecewise constant. However, they are susceptible to overfitting and may not perform as well as other methods when intricate nonlinear relationships are present in the data.

GBM is notable for its predictive speed and accuracy, particularly for large and complex datasets. It is predicated on the assumption that the next-best model, when combined with earlier models, minimizes the aggregate forecast error [14]. For instance, if a minor change in a case's prediction results in a significant reduction in error, the case's next desired outcome is a high value. Then, fewer errors are present in the new predictions which are closer to the objective. It can perform both regression and classification tasks and is particularly useful when the data has complex nonlinear relationships. However, it can be computationally intensive and may necessitate meticulous hyperparameter tuning for optimal performance.

Under the premise of conditional independence between each pair of features, NB methods involve a type of supervised learning algorithm that applies Bayes' theorem to the values of class variables [15]. As a member of the generative learning algorithm family, it can be used for classification tasks, such as text classification, and modeling the input distribution of a given class or categories. Contrary to discriminative classifiers, like LR, NB does not identify which features are most crucial for distinguishing classes.

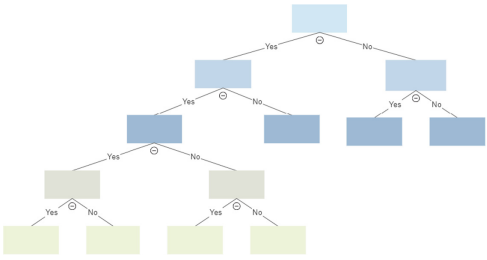

Fig. 1. A structural example of a DT model

It is quick and requires a small amount of training data, however, it implies that the predictors are independent, which is not always the case in practical applications.

4 Results and Conclusion

In this work, the dataset is split into a training set and a test set with a ratio of 70:30. Various criteria have been used to compare algorithms and determine the most effective algorithm for detecting fraudulent, but the most common measures used to assess ML algorithms are accuracy and AUC [16]. We use AUC to evaluate the performance of the four models.

The ROC curve (receiver operating characteristic curve) is the probability curve, and the AUC is the degree of separability, sometimes known as the measure of separability. It indicates how well the model can differentiate between groups. AUC refers to the Area Under the ROC Curve in the most general sense, and it ranges from 0 to 1, with higher values indicating more accurate performance. Moreover, we have decided to use the AUC as our comparative benchmark since the ROC AUC is more significant in cases of imbalanced.

The total number of samples in the dataset is 85,443, and testing accounts for 30% of the dataset; therefore, the training set contains 62,560 samples. The outcomes using GLM is shown in Fig. 2.

The AUC of 0.97 for GLM indicates that it performed exceptionally well. Then, we constructed the DT Model and obtained the ROC results shown in Fig. 3. Although the result is inferior to that of GLM, the AUC of the DT Model still surpassed 0.90. The GBM attained an AUC of 0.95, as shown in Fig. 4, outperforming the DT Model. Finally, the NB Model is depicted in Fig. 5, reaching an AUC of 0.938. To easily compare the four models, we listed the results in Table 1.

In general, when the relationship between the response and predictor variables is nonlinear and the data exhibit nonlinear behavior, GLMs may perform better than DTs, GBM, and NB Models. This may be the main reason why GLM is more suitable for this dataset. GLM is a flexible framework that can accommodate a wide variety of response distributions and link functions, making it especially useful in this context. DTs are more suitable for categorical and continuous data, whereas NB is a probabilistic algorithm that works well with small datasets and is more applicable to text classification tasks.

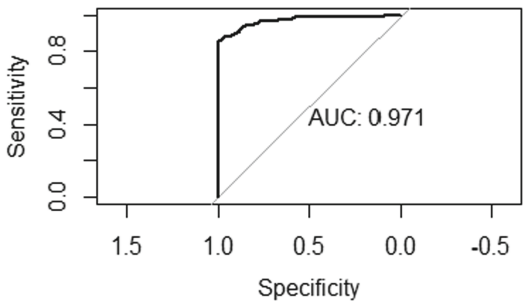

Fig. 2. GLM ROC results

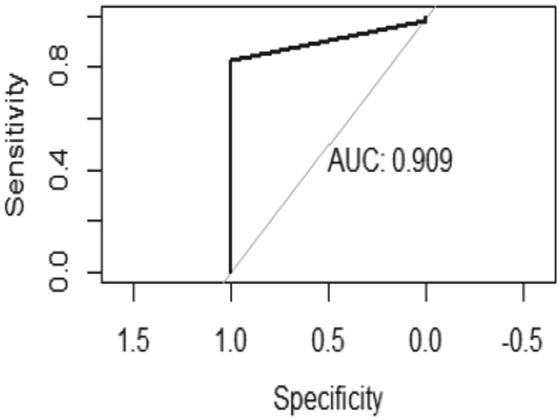

Fig. 3. DT Model ROC results

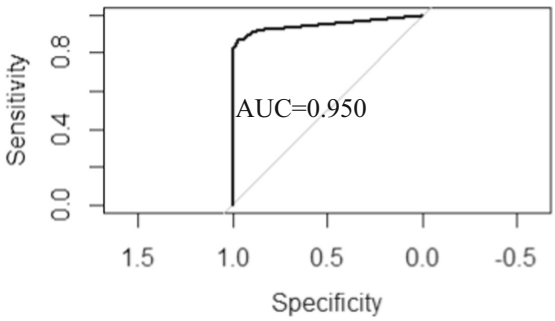

Fig. 4. GBM ROC results

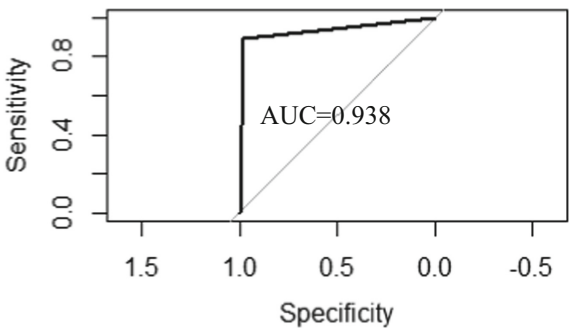

Fig. 5. NB Model ROC results

Table 1. Comparison table

Model	AUC
Generalized Linear Model	0.971
Decision Tree Model	0.909
Gradient Boosting Model	0.950
Naïve Bayes Model	0.938

5 Conclusions

This article compares the GLM, the DT Model, GBM, and the NB Model. AUC has been used to evaluate the precision of these four models, and the GLM is the most appropriate for learning and predicting cases of fraud using a large and imbalanced dataset.

Credit card fraud is a significant problem because it can result in substantial financial losses for enterprises and individuals, in addition to the damage it causes to a business's or individual's reputation. Therefore, detecting and preventing fraud not only be a continuous effort but a priority. Future research should concentrate on enhancing the precision of each classifier and selecting a more appropriate model to supplement the current body of knowledge on credit card fraud.

References

1. Ghosh, S., & Reilly, D. L. (1994, January). Credit card fraud detection with a neural-network. In System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on (Vol. 3, pp. 621–630). IEEE.
2. Sakharova, I. (2012, June). Payment card fraud: Challenges and solutions. In 2012 IEEE international conference on intelligence and security informatics (pp. 227–234). IEEE.
3. Federal Trade Commission, & Federal Trade Commission. (2013). Consumer sentinel network data book for January-December 2012. Washington, DC.
4. Kaggle.com. (2023). Credit Card Fraud Detection. [online] Available at: https://www.kaggle.com/mlg-ulb/creditcardfraud [Accessed 18 Apr. 2023].
5. Nami, S. & Shajari, M. (2018, June). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors, Expert Systems with Applications,110, 381–392.
6. Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. Engineering Applications of Artificial Intelligence, 76, 130–157.
7. Vidanelage, H. M. M. H., Tasnavijitvong, T., Suwimonsatein, P., & Meesad, P. (2019, October). Study on machine learning techniques with conventional tools for payment fraud detection. In 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1–5). IEEE.
8. Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, 317–331.
9. Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data, 9(1), 1–17.

10. Alfaiz, N. S., & Fati, S. M. (2022). Enhanced credit card fraud detection model using machine learning. Electronics, 11(4), 662.
11. Alharbi, A., Alshammari, M., Okon, O. D., Alabrah, A., Rauf, H. T., Alyami, H., & Meraj, T. (2022). A novel text2IMG mechanism of credit card fraud detection: a deep learning approach. Electronics, 11(5), 756.
12. Hastie, T. J., & Pregibon, D. (2017). Generalized linear models. In Statistical models in S (pp. 195–247). Routledge.
13. Kotsiantis, S. B. (2013). Decision trees: a recent overview. Artificial Intelligence Review, 39, 261–283.
14. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937–1967.
15. Zheng, F., & Webb, G. I. (2005). A comparative study of semi-naive Bayes methods in classification learning. AUSDM05, 141–155.
16. Jiménez-Valverde, A. (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology and Biogeography, 21(4), 498–507.