Application of Machine Learning Algorithms in Detecting Credit Card Fraud: A Comparative Analysis

Yiting Ren*

Department of Mathematics, University College London, London, United Kingdom

*Corresponding author: zcahyre@ucl.ac.uk

Abstract. Credit card transaction have grown increasingly prevalent in the digital era, and along with them, so have incidents of associated fraud. Hence, identification and prevention of such frauds are critically crucial. Machine learning algorithms are predominantly employed in the realm of credit fraud detection. According to current literature, class imbalance of data, a great disparity in ratio between normal and fraudulent transactions, could severely affect the result in detection. In this paper, a combination of imbalanced classification methods, specifically the Synthetic Minority Random Oversampling Technique (SMOTE) and Under-sampling, is utilized to harmonize the dataset. Some popular machine learning algorithms are applied to detect frauds are compared and analyzed, including Logistic Regression, Decision Tree, Random Forest and XG Boost. The accuracy, precision, recall, F-1 score and Area Under Curve (AUC) of each algorithm are used as metrics of performance evaluation. The research findings indicated that among the four models tested, XG Boost, when coupled with balanced data yielded overall optimal results for classifying fraudulent activities.

Keywords: Credit card fraud; machine learning; class imbalance; XG Boost.

1. Introduction

The evolution of credit card market from its humble origins in the 1920s to its explosive growth fueled by the computerization of society has indeed transformed consumer spending and business operations. Credit cards have become a pivotal component of modern economies, playing a crucial role in household finances, business transactions, and global economic activities [1]. In this digital age, electronic transaction using credit card become increasingly common and convenient. However, this convenience has also brought about their fair share of trouble of frauds.

Credit card fraud constitutes a form of financial rascality, encompassing the illicit acquisition of credit card details of another person with the intention of making unauthorized purchases or withdrawing funds from the account. According to a recent report on card fraud, the total card fraud losses constantly increase from 2019, and it is predicted to reach 13.73 billion dollars by 2014 [2]. In particular, Card-Not-Present fraud will boost from 54% pre-pandemic in 2019 to 74% by 2024, which create new prevention challenges for merchants and consumers. Moreover, digital criminals are continually devising innovative and inventive fraud tactics, which will inevitably contribute to an increase in fraudulent activities. Consequently, it becomes imperative to employ effective and secure techniques for detecting fraud.

To combating frauds, machine learning and predictive analysis utilize extensive sets of example data of the underlying domain. These datasets are used to establish computational model that can classify and categorize future data seen in the domain [3]. This paper will focus on counterfeit fraud using machine learning algorithms, to make a contribution in minimization of damages caused by credit card-based crimes.

1.1. Challenge Description

In identification of the credit card fraud, the most challenges countered in the research process is Class Imbalance. This problem has been addressed and discussed in numerous current academic literatures [4-6]. The proportion of fraudulent or criminal activities is significantly lower compared to the volume of normal and legitimate ones. Fig. 1 illustrates the Highly Imbalanced Dataset [7] used

in this paper from prominent dataset repository Kaggle, transaction fraud is only 0.173% of total transactions.

Class Imbalance will cause a difficulty in detecting the characteristic of fraudulent transaction using machine learning algorithms. Machine learning models learn from the data they are trained on. The models have limited exposure to samples of positive class (fraudulent transactions), the classification algorithms might produce bias results to exact the fraud patterns from the dominant class. Due to scarcity of positive class, models may struggle to distinguish between normal and fraudulent transactions effectively. This could lead to false negatives and an overall decrease in detection accuracy. Thus, this will lead to severe financial losses to individuals and organizations.

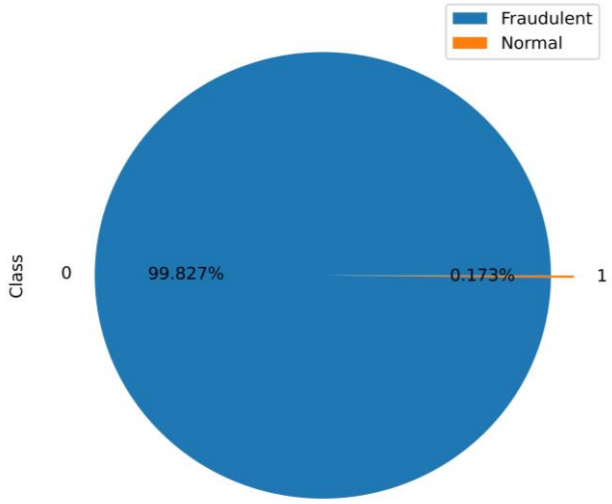

Fig. 1 Highly Imbalanced Dataset

Much related research works purposed solutions to this problem, such as data mining, resampling and feature selection [8], however, each solution still has its limitations in solving Class Imbalance. The goal for this paper is to tackle the imbalanced classification issue to achieve a high precision result.

1.2. Contribution

This paper will outline and assess various commonly used machine learning methods, examining their ability to precisely classify fraudulent transactions using an authentic real-world dataset as a basis. The list of contribution is as follow:

This paper pre-processed the highly imbalanced dataset from Kaggle using a combination of data sampling approaches suggested in the literature review.

This paper developed and compared four machine learning methods that are commonly utilized in fraud identification and categorization in associated studies.

This paper presented a detailed examination of the outcomes for each logarithm, along with an exploration of the experiment's limitations.

The remaining part of the paper is structed into 6 parts. In Section 2, an overview of the published studies with regard to detecting credit card fraud will be provided. Section 3 describes the experimental procedures and provides the basic concepts of imbalanced classification approaches, machine learning algorithms used and the metrics of performance evaluation. In Section 4, the results obtained across different comparison metrics are presented and subject to discussion. Section 5 suggested limitations in this work and possible improvements for them. Section 6 serves as the paper's conclusion, providing a brief overview of future research direction.

2. Related Work

Data researchers have been attempted to resolve fraud detection with machine learning. Numerous fraud detection techniques have been suggested by researchers, providing degree of effectiveness in

combating credit card fraud. In recent studies, strategies frequently used to tackle this problem includes Decision Tree, Genetic Algorithms, Support Vector Machine, Bayesian Networks, Artificial Neural Networks and Gradient Boosting techniques. A review of research findings on the identification and prevention of credit card fraud will be outlined in this section.

Mittal and Tyagi [3] discussed various popular algorithms in the categories of supervised, ensemble and unsupervised learning, using different metrics. They found that unsupervised algorithms is more effective in dataset skewness management, resulting in superior performance across all metrics compared to other logarithms.

Itoo et al. [4] selected Logistic Regression, Naïve Bayes, and K-Nearest Neighbour to conduct the comparative analysis. In order to obtain a better result, re-sampling techniques are used to pre-process the imbalanced dataset. Logistic Regression showed an optimal outcome for all data proportions. In contrary of the research of Mittal and Tyagi, they concluded that supervised techniques are more suitable for fraud detection in each case as compared to unsupervised techniques.

Gupta et al. [5] examined the performance of machine learning models trained on a public dataset, including Logistic Regression, Decision Tree, XG Boost and Artificial Neural Networks. They conducted data processing through feature selection, which resulted in the removal of independent variables, leaving only pertinent data for model training. Before using any method for data balancing, XG Boost produced a better result, which has precision of 91.3% and recall of 80.5%. The substantial efforts they have invested in improving classifier performance are expected to introduce favourable outcomes when implementing data balancing techniques, including random over-sampling, random under-sampling, and the Synthetic Minority Random Oversampling Technique (SMOTE), in combination with the most optimal model, i.e., XG Boost. In summary, random over-sampling with XG Boost has provided the greatest outcomes from all aspects of evaluation.

Makki et al. [6] assessed and compared eight machine learning models in the context of identification and prevention of credit card fraudulent activity, along with their respective limitations. Additionally, they evaluated various imbalanced classification methods and assessed their efficacy when dealing with extremely imbalanced datasets. Logistic Regression, Artificial Neural Networks, C5.0 decision tree algorithm and Support Vector Machine stand out as the top four methods when it comes to performance across Accuracy, Sensitivity and Area Under Precision-Recall Curve. Their findings indicated that the commonly employed approaches for resolution could lead to undesirable outcomes in solving imbalanced problems. They also highlighted the misleading nature of relying solely on a single performance measure for imbalanced learning.

3. Methodology

In this paper, the experimental procedures are briefly shown in Fig. 2.

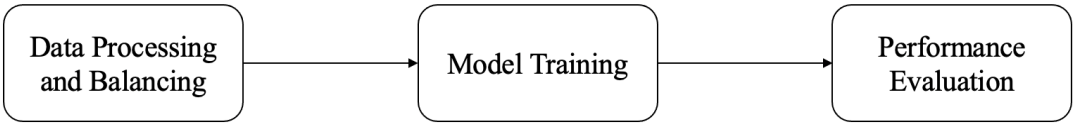

Fig. 2 Experimental Procedures

3.1. Data Exploration

The dataset used is from prominent dataset repository Kaggle [7]. This collection of data include authentic credit card transactions conducted by European cardholders during the month of September 2013.

By reason of confidentiality issue, original features V1, V2, …, V28 have been transformed with Principal Component Analysis (PCA). These are three features have not undergone PCA transformation:

Feature "Time" records duration in seconds that have passed since the initial transaction, for each consecutive transaction.

Feature “Amount” shows the magnitude of transaction volume.

The response variable "Class" Feature returns a value of 1 in cases of fraud transaction and 0 in all other circumstances.

3.2. Data Processing and Balancing

As Section 1.1 mentioned, the most critical problem of the dataset is class imbalance. As models will deal with a substantial dataset, under-sampling is used to reduce the total amount of data. Then, SMOTE will used on the reduced dataset to tackle class imbalance problem.

Under-sampling involves reducing the quantity of instances or samples belonging to the majority target category [9]. This approach aims to systematically filter out repeated or redundant information from dataset, consequently reducing the inclusion of noise sample. However, a trade-off will emerge: this strategy could result in the removal of data from the positive class, which leads to the loss of informative patterns during the process of reducing the size of dataset.

SMOTE stands as a well-established over-sampling method utilized to generate synthetic data, primarily employed in addressing class imbalance. This data balancing method applies a K-Nearest Neighbour algorithm to craft additional instances of the minority class, thereby enhancing class balance within the dataset [10]. It improves model performance without causing overfitting issue. However, SMOTE fails to account for scenario where neighboring instances might belong to distinct classes, potentially leading to increased class overlap and the introduction of extra noise [11].

A combination of under-sampling and SMOTE could help to reduce the size of dataset and prevent the loss of information due to this reduction.

3.3. Machine Learning Algorithms

In this paper, 4 machine learning classifiers for classification task are used, specifically Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and XG Boost. The implementation is carried out using the Python programming language.

Logistic Regression (LR) is a supervised machine learning model that often utilized in classification of data into distinct categories. The frequently utilized LR models are typically applied to assess the association between binary predictors and categorical variables [12]. In this case, GridSearch is used to optimise the parameter “C” in the model, i.e., a parameter is responsible the balance between under-fitting and overfitting.

A Decision Tree (DT) is a hierarchical model that resembles a tree-like structure comprising decision and their potential consequences. The root node represents at the beginning of a decision tree, where the population starts to divide according to various features. Each decision node represents a choice or condition based on a feature, and each leaf node represent a predicted or prediction.

A Random Forest (RF) is a classifier based on ensemble methods that amalgamates numerous decision trees, each constructed using distinct subsets of the training data [3]. It helps address the problem of overfitting. Individual trees are less likely to memorize noise or outliers in the data since each tree is trained on a separate sample of data and has a restricted depth owing to the random feature selection procedure at each node split.

XG Boost is also an ensemble supervised learning method that leverages a decision-tree-based algorithms. In contrary of RF, boosting involves building trees with fewer splits. It also addresses the presence of missing values within the dataset and incorporates an integrated cross-validation mechanism that runs at every stage [5].

3.4. Metrics of Performance Evaluation

Accuracy is typically the most fundamental measure of the performance in classification problems. However, comparison solely in accuracy is not adequate to conclude the effectiveness of models in class imbalanced case. Therefore, other performance measures are introduced.

Confusion matrix represent counts from predicted and actual values. Table 1 presents the definition of confusion matrix used for evaluation.

Table 1. Confusion Matrix for Evaluating Classification

Actual	Predicted	
	Normal (0)	Fraudulent (1)
Normal (0)	True Negative (TN)	False Positive (FP)
Fraudulent (1)	False Negative (FN)	True Positive (TP)

Accuracy of a model through a confusion matrix is computed using the given formula below:

$$Accuracy = (TN + TP)/(TN + FP + FN + TP) \quad (1)$$

Precision is a measure of True Positive result among all positive predictions, which is defined as:

$$Precision = TP/(TP + FP) \quad (2)$$

Recall is the metric that assesses the ability of a models to correctly recognize True Positives. The mathematical expression of recall is:

$$Recall = TP/(TP + FN) \quad (3)$$

F-1 Score is an error metric that calculates the harmonic mean of precision and recall. It represents the balanced ability between these two measures, which is defined as:

$$\text{F-1 Score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Area Under Curve (AUC) of Receiver Operator Characteristic (ROC) Curve measures the ability of model to distinguish between classes. As the AUC value rises, the ability of the classifier to differentiate between positive and negative classifications becomes more effective [13].

4. Results

Fig. 3 below presents the confusion matrices for LR, DT, RF and XG Boost. The ideal result of a model is supposed to have low FN and FP. RF and XG Boost both have limited number of FN and FP, compared to LR and DT.

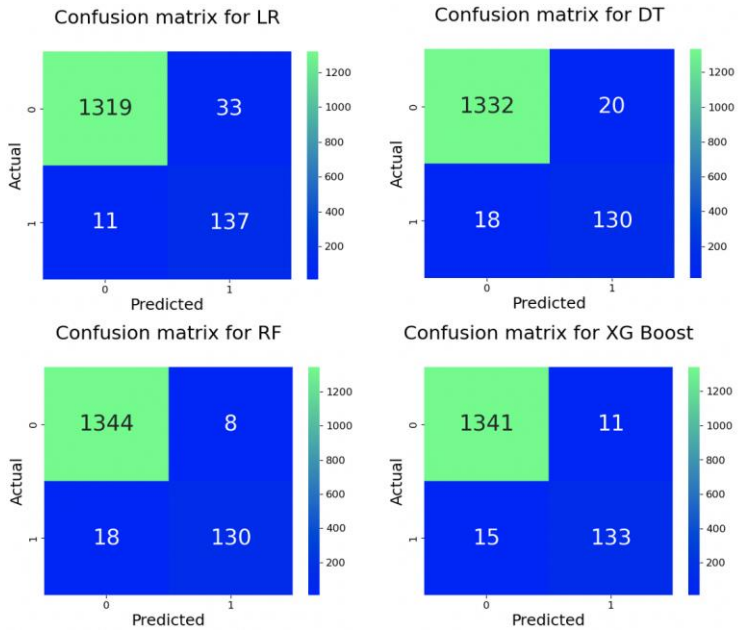

Fig. 3 Confusion Matrices for 4 Selected Algorithms

Fig. 4 below displays the ROC curves plotted for predicting credit card fraud detection. The rationale behind creating separate plots for different models is to avoid the issue of curve overlap. Similarity among the curves could lead to visual confusion, making it challenging to identify them distinctly.

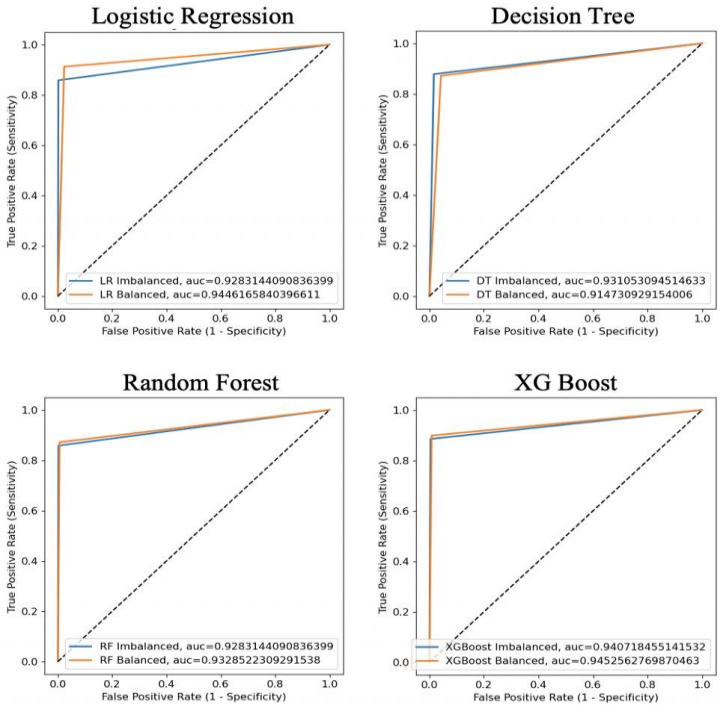

Fig. 4 ROC Curve for 4 Selected Algorithms

The ROC curve in blue illustrates the performance before data balancing, while the orange curve represents the performance after data balancing. The AUC shows slight improvements in most models after data balancing, except for DT.

The results of performances is summarized in Table 2 below. Despite varying precision, recalls and F-1 scores, all algorithms have accuracy and AUC values greater than 0.900. Therefore, accuracy does not hold the utmost significance as a performance evaluation metric. By comparing other metrics, the results of XG Boost in all metrics is 0.10 – 0.20 higher than that of RF, except for precision. Hence, the XG Boost is the most suitable model for credit card fraud detection among these for algorithms.

Table 2. Results of Performance Evaluation

	Accuracy	Precision	Recall	F-1 Score	AUC
LR	0.971	0.813	0.912	0.860	0.945
DT	0.949	0.694	0.872	0.772	0.915
RF	0.982	0.942	0.872	0.905	0.933
XG Boost	0.983	0.924	0.899	0.911	0.945

5. Limitations and Improvements

This experiment solely constructs and compares 4 machine learning algorithms, employing regression-based and tree-based approaches. Consequently, it lacks the compelling strength to determine the optimal model. More supervised machine learning algorithms, such as Support Vector Machine and Artificial Neural Networks, could also include in evaluation. In addition, it is worthwhile to contemplate the incorporation and assessment of semi-supervised, unsupervised, and reinforcement machine learning techniques [3].

The idea of using an imbalanced classification approach to improve the result in classification of credit card fraudulent activity using 4 different algorithms. However, there are some limitations of this approach. Firstly, the use of random under-sampling could cause an information loss due to random process. To improve this, random under-sampling could be replaced by Tomek link, which is also a under-sampling method using Condensed Nearest Neighbours [14]. Secondly, optimizer GridSearch is only applied on LR. However, the result could be improved by using GridSearch or cross-validation to optimize parameters of RF, DT and XG Boost.

6. Conclusion

As the damage of credit card fraud is irrevocable, it is essential to focus on combating the fraud. The most critical problem of this fraud detection is class imbalance. In this paper, a combination of random under-sampling and SMOTE is used as an attempted solution. Moreover, the application and fitness of four machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest and XG Boost, in detection of credit card fraud has been compared and evaluated. Among these four models. XG Boost stands out as a more suitable choice of model.

In the future, it is important to invest effort in developing strategies to address the limitations outlined. Additionally, efficient approaches to solve issue of class imbalance are necessary for better classification and detection outcomes.

References

[1] Nicole Long, Ashely Donohoe. The Importance of Credit Cards, 2019. Retrieved from: https://budgeting.thenest.com/importance-credit-cards-29514.html.

[2] Toplin, J. Spotlight: US Card Payment Fraud Losses Forecast 2022, 2022. Retrieved from: https://www.insiderintelligence.com/content/spotlight-us-card-payment-fraud-losses-forecast-2022.

[3] Mittal, S., & Tyagi, S. Performance evaluation of machine learning algorithms for credit card fraud detection. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering, 2019: 320-324.

[4] Itoo, F., Meenakshi, & Singh, S. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. International Journal of Information Technology, 2021, 13: 1503-1511.

[5] Gupta, P., Varshney, A., Khan, M. R., Ahmed, R., Shuaib, M., & Alam, S. Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. Procedia Computer Science, 2023, 218: 2575-2584.

[6] Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 2019, 7: 93010-93022

[7] Kaggle. Retrieved from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud.

[8] Wasikowski, M., & Chen, X. W. Combating the small sample class imbalance problem using feature selection. IEEE Transactions on knowledge and data engineering, 2019, 22(10): 1388-1400.

[9] Mohammed, R., Rawashdeh, J., & Abdullah, M. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems, 2020: 243-248.

[10] Sridhar, S., & Sanagavarapu, S. Handling data imbalance in predictive maintenance for machines using SMOTE-based oversampling. In 2021 13th International Conference on Computational Intelligence and Communication Networks, 2021: 44-49.

[11] Paul, S. Diving Deep with Imbalanced Data, 2018. Retrieved from https://www.datacamp.com/tutorial/diving-deep-imbalanced-data.

[12] Prasad, P. Y., Chowdarv, A. S., Bavitha, C., Mounisha, E., & Reethika, C. A Comparison Study of Fraud Detection in Usage of Credit Cards using Machine Learning. In 2023 7th International Conference on Trends in Electronics and Informatics, 2023: 1204-1209.

[13] Bhandari, A. Guide to AUC ROC Curve in Machine Learning: What is Specificity? Retrieved from https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/.

[14] Zeng, M., Zou, B., Wei, F., Liu, X., & Wang, L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In 2016 IEEE International Conference of Online Analysis and Computing Science, 2016, 225-228.