**International Academic Institute for Science and Technology**

International Academic Journal of Science and Engineering
Vol. 8, No. 4, 2021, pp. 1-5.

ISSN 2454-3896

**International Academic Journal of Science and Engineering**

www.iaiest.com

# Handling Imbalanced Data for Credit Card Fraud Detection: A Machine Learning Approach

Deepak Patel [a], Pooja Singh [b]

[a] *Department of Civil Engineering, Sunrise College of Structural Sciences, Shillong, India.*
[b] *Department of Civil Engineering, Sunrise College of Structural Sciences, Shillong, India.*

## Abstract

Using credit cards has become a routine part of the typical person's life. Credit cards are the most versatile and convenient method of payment because they are one of the simplest ways to conduct transactions. The process of identifying credit card theft is both time-consuming and essential. This process should be carried out carefully because it also affects the interests and flexibility of the customer. Due to the rise in credit card usage and the number of scams discovered, this study has recently gained significant attention. Despite the high priority of this procedure, it is crucial to do it efficiently because a big number of customers are engaged. A minor error in the detecting procedure could result in a significant loss. Predicting the categorical labels of unknown things is the process of classification. In addition to helping users distinguish between objects of different classes—which are really one single entity on which operations can be performed—this also helps users organize the data in blocks for simple comprehension. Initially, the fundamental label categorization is done using the supplied data. Since this classification is just temporary, each cluster undergoes additional processing to achieve perfection.

**Keywords:** Credit Card, Fraud, Detection, Machine Learning.



DOI:10.71086/IAJSE/V8I4/IAJSE0823
Received: 05 September 2021/Revised: 07 October 2021/Accepted: 02 November 2021/Published: 31 December 2021

## Introduction

Since the 1970s, electronic payments have been used. Numerous payment mechanisms were developed (Thennakoon et al., 2019). The world has witnessed an explosion in the number of individuals utilizing the internet since its creation and widespread integration. In the late 1990s, electronic payments began to become a popular way to make purchases. During this time, concepts for cashless transactions emerged. These provided the foundation for cashless or electronic trade. The emergence of electronic commerce also occurred during this time. This prompted scholars to work on the procedure, and several scholarly investigations were carried out in this field. Numerous commercial agencies were also interested in this field because this process involved a lot of commercial interest, and as a result, this area of study began to be quite active. Numerous concepts for executing these cashless transactions were put forth, and some of them were even introduced to the market (Warghade et al., 2020). However, many failed because they were unable to reach the intended audience. Although these systems were widely deployed, they did not produce a profit. Many businesses launched new payment options for the business-to-consumer sector at the same time.

Large volumes of credit card transactions are handled by businesses and financial institutions (Varmedja et al., 2019). A plastic card known as a credit card, which is given to many people as a way to make financial transactions, makes these transactions possible. This card grants authorized users the ability to procure goods and services, founded on the user's commitment to reimburse the expenses at a subsequent date. Credit cards have progressively gained prevalence in personal finance circles over recent years, with the number of credit cardholders being an evident reflection of admiration and endorsement rates. According to information taken from the "United States credit card statistics" available on the Statista website, credit card issuance has tripled compared to debit card issuance (Awoyemi et al., 2017). The importance of credit cards in contemporary financial activities is demonstrated by the approximately 1.1 trillion credit cards that were issued between 2012 and 2018. With over 300 million credit cards issued to its customers, Visa was the largest credit card issuer in 2019. To progress the field of e-business, it is essential to guarantee the security of credit card services provided by financial institutions and to develop a trustworthy technique for identifying fraud. Analysing the historical purchase data of cardholders for fraud detection holds significant potential in lowering the occurrence of credit card fraud. Fraud detection systems become especially relevant when perpetrators evade fraud prevention measures and initiate deceitful transactions (Mînăstireanu et al., 2020).

## Proposed Methodology

The suggested approach involves several steps: first, obtaining the initial dataset; then, performing data sampling to ensure data balance; next, training classifier models; and finally, identifying instances of fraud. Fig. 1 shows a graphic representation of this process.

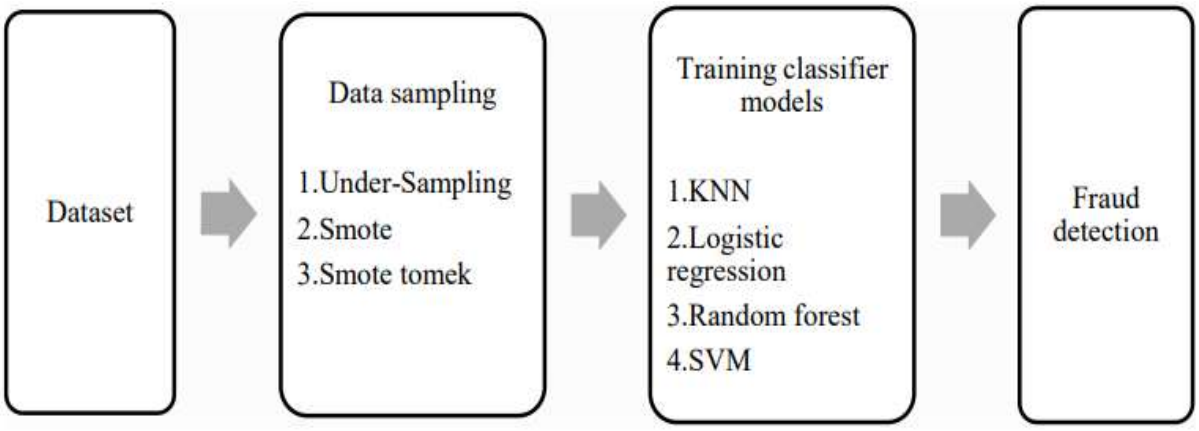



**Fig. 1: Overview of the Proposed Approach**

This study utilizes a publicly accessible dataset that forms part of a challenge on Kaggle. 284,807 credit card transactions that occurred over two days in 2013 are included in the sample. There were 492 instances of fraudulent conduct among these transactions. Since there are so few fraudulent occurrences in the dataset, they can be classified as the minority class, whereas the majority class consists of transactions that are not fraudulent (Makki et al., 2019). The dataset has a notable class imbalance as a result of this circumstance. Principal Component Analysis (PCA) was used to alter the characteristics of the data samples in order to address privacy concerns. This method produced 28 major components, designated V1 through V28, which capture the key characteristics of the data. Notably, only the

'Time' and 'Amount' attributes of the transactions remained unaltered by PCA. The labeling for the target assigns class 0 to valid transactions and class 1 to fraudulent ones.

An algorithm for machine learning is represented by this ensemble bagging technique. The original training dataset is split up into several bootstrap samples for this classification method. These examples are each made up of arbitrary subsets with m rows and n columns that were extracted from the original dataset. The gathered bootstrap samples are then used as input by several decision trees. The ultimate prediction is derived from the evaluations of most of these trees together. The basic setup of the random forest classifier is shown in Fig. 2 (Dornadula & Geetha, 2019).

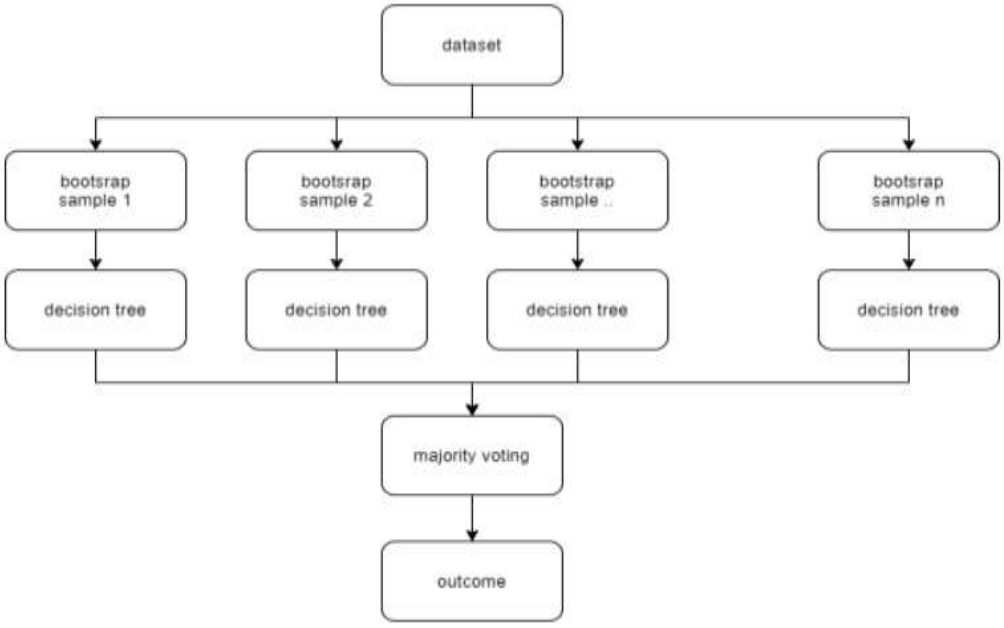
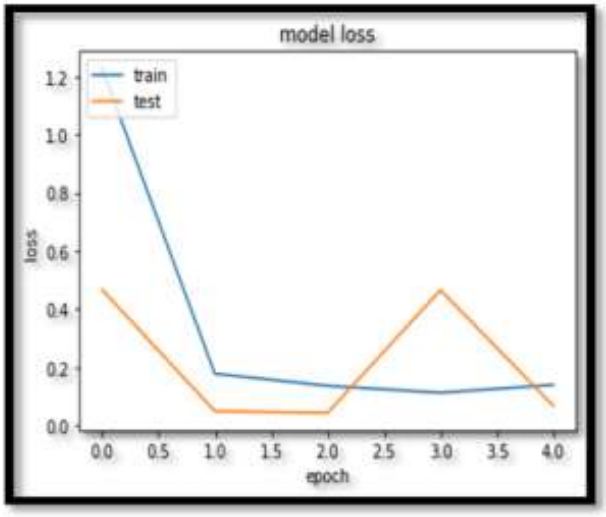



**Fig. 2: Decision Tree of Random Forest Classifier**

## Results

Fig. 3 shows how many data samples were obtained for each class once data sampling procedures were put into practice. When under-sampling is employed, the number of samples is significantly decreased, reaching just 492, which is insufficient for the proper training of machine learning algorithms (Dal Pozzolo et al., 2014). However, there are 227,845 instances of the SMOTE and SMOTE-Tomek approaches in each class. Many classification models are then trained using these preprocessed data sets: Support vector machines (SVM), random forests, logistic regression, and K-nearest neighbors (KNN).



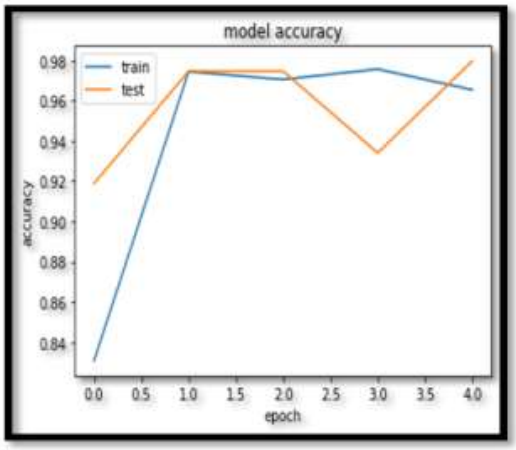



**Fig. 3: Accuracy and Loss Plot**

When the classifiers are compared, it becomes clear that the random forest, which makes use of different data sampling techniques, performs better than the others. Given that it applies to under-sampled data and has the highest F1-score of 0.94, this advantage is evident. On the other hand, the logistic regression classifier produces the least impressive outcome when combined with the SMOTE and SMOTE-Tomek sampling techniques. ROC and PR - Hybrid Approach shown in Fig. 4 and Fig. 5.

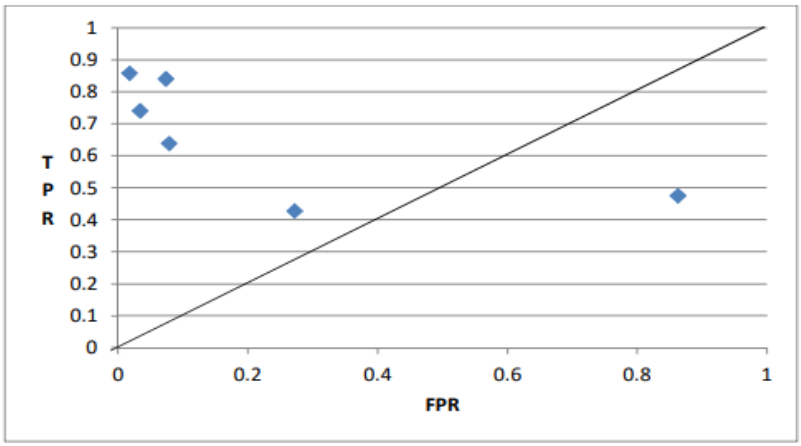



**Fig. 4: ROC**

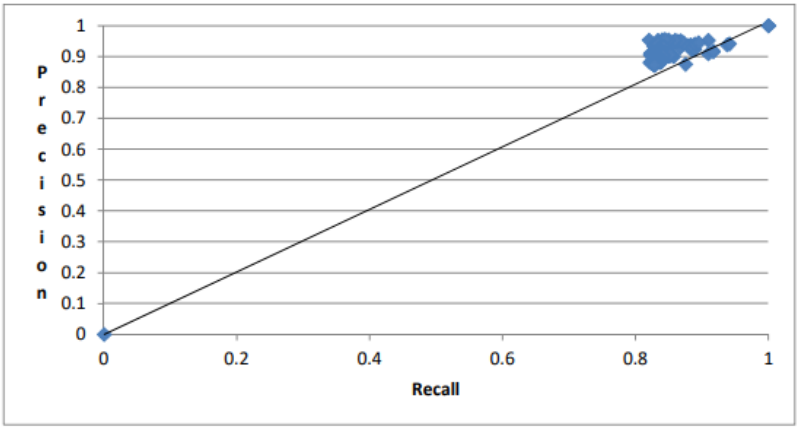



**Fig. 5: PR - Hybrid Approach**

Similar to this, employing these techniques, the SVM classifier performs below averagely. Notably, while both the SVM and logistic regression classifiers exhibit notable sensitivity, their precision is still notably constrained.

## Conclusion

A new problem has emerged as a result of e-commerce. It handles security when carrying out these transactions. Credit card fraud is one of the primary issues that influences the debate over online transactions. The banks have always placed the utmost importance on these frauds. These are the most important issues, and they will only become worse. This issue is likewise expected to be among the top concerns for the clientele. The rise in online transactions in recent years has led to a significant surge in these kinds of frauds. The use of payment cards has led to an increase in related frauds, making it a significant issue. These kinds of scams have evolved into organized crime due to the rise in card usage and associated crimes. This article presents a range of machine learning techniques for identifying credit card fraud. In order to overcome the dataset's unequal class distribution, the study looks into the use of under- and over-sampling techniques. Several classifiers were trained using the altered datasets, and the performance of each classifier was evaluated using measures including precision, recall, and the F1-score.

## References

Thennakoon, A., Bhagyani, C., Premadasa, S., Mihiranga, S., & Kuruwitaarachchi, N. (2019, January). Real-time credit card fraud detection using machine learning. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 488-493). IEEE. https://doi.org/10.1109/CONFLUENCE.2019.8776942

Warghade, S., Desai, S., & Patil, V. (2020). Credit card fraud detection from imbalanced dataset using machine learning algorithm. *International Journal of Computer Trends and Technology*, *68*(3), 22-28.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019, March). Credit card fraud detection-machine learning methods. In *2019 18th International Symposium Infoteh-Jahorina (Infoteh)* (pp. 1-5). IEEE. https://doi.org/10.1109/INFOTEH.2019.8717766

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (ICCNI)* (pp. 1-9). IEEE. https://doi.org/10.1109/ICCNI.2017.8123782

Mînăstireanu, E. A., & Meșniță, G. (2020). Methods of handling unbalanced datasets in credit card fraud detection. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *11*(1), 131-143. https://doi.org/10.18662/brain/11.1/19

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. *IEEE Access*, *7*, 93010-93022. https://doi.org/10.1109/ACCESS.2019.2927266

Dornadula, V. N., & Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. *Procedia computer science*, *165*, 631-641. https://doi.org/10.1016/j.procs.2020.01.057

Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, *41*(10), 4915-4928. https://doi.org/10.1016/j.eswa.2014.02.026