Credit Card Fraud Detection Using Machine Learning Algorithms

Kavita Rani[1], Anjali Garg[2], Saba Zaidi[3*], Kunal Saini[4], Sunidhi[5]

[1]Associate Professor, Panipat Institute of Engineering & Technology, Samalkha, Panipat, Haryana, India
[2,3]Assistant Professor, Panipat Institute of Engineering & Technology, Samalkha, Panipat, Haryana, India
[4,5]Student, Panipat Institute of Engineering & Technology, Samalkha, Panipat, Haryana, India

*Email- sabazaidi220@gmail.com

* Corresponding author
doi: https://doi.org/10.21467/proceedings.7.6.48

Abstract

Targets of credit card fraud are plain and amiable. The growth of online payment methods brought about by e-commerce and a number of other websites has raised the possibility of online fraud. Researchers are using a range of machine learning techniques to identify and examine online transaction schemes as a result of an increase in fraud rates. The main objective of the project is to develop a novel fraud detection technique for streaming transaction data, which assigns cards to groups according to the number of transactions they complete, allowing for the analysis of past consumer transaction data and behavioural trends. The sliding window technique is then used to aggregate the transactions that cards from different groups have performed, allowing for the extraction of each group's unique behavioural style. After that, distinct classifiers are trained for every category. One of the best fraud prediction methods can then be chosen from among the classifiers with the highest rating score.

1 Introduction

When a consumer is given a credit card, they can usually withdraw cash in advance or make purchases up to the credit limit. By enabling customers to roll over their balance to the next payment cycle and repay it later, credit cards give their users the advantage of timeframe. Targets of credit card fraud are calm. It is possible to remove a significant quantity without the owner's knowledge in a short-term, risk-free manner. Since scammers are always working to make every illegal transaction acceptable, detecting fraud is a difficult and time-consuming task [1].The majority of individuals worldwide are victims of various forms of fraud, the bulk of which involve credit cards and have been widely reported by the media in recent years are fully aware of frauds. Because there will be a large imbalance in the credit card dataset, with more transactions than fraudulent ones. Card-not-present fraud rates remain high even with the widespread use of EMV cards by banks, which are smart cards that store data on integrated circuits rather than magnetic stripes, making some on-card payments safer.

2 Literature Survey

Significant class difference, the existence of labelled and unlabelled samples, and growing processing capacity are three major problems with datasets pertaining to card fraud that must be resolved. Fraud is detected using a variety of supervised and semi-supervised machine learning methods. SVM, Decision Trees, Naive Bayes Classification, Logistic Regression, and Least Squares Regression are just a few of the supervised machine learning techniques that are used to detect fraudulent transactions in real-time

Proceedings DOI: 10.21467/proceedings.7.6; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-9-5

datasets [1],[2].Two distinct random forest algorithms are used to learn the evident features of normal and anomalous transactions. These are random trees and random forests based on CART [3][7]. Random forests still have certain problems when dealing with unbalanced data, despite its ability to yield adequate results on smaller data sets[2], [1][4]. Future studies will concentrate on the problems. The random forest technique itself needs to be improved. Credit card fraud that is extremely skewed is analysed using K-Nearest Neighbour, Naïve Bayes, and Staging of Logistic Regression data. In order to handle extremely unbalanced credit card fraud data, metaclassifiers and meta-learning techniques are being researched[5]. Despite its applicability, supervised learning methods may not always be effective in detecting fraud. a deep auto-encoder model and restricted Boltzmann machine (RBM) that can generate normal transactions and distinguish abnormalities from normal patterns [6][3][4]. Additionally, a hybrid strategy that uses the Majority Voting method is employed.

3 Problem Statement

Because they let you make purchases and pay off the balance later, credit cards are a necessary financial tool. For credit card users, the option to settle the bill at a later time is beneficial. Credit cards are therefore a popular target for fraudsters. By pretending that the actual cardholders performed the transaction, these scammers are able to obtain a substantial amount of money without the owners' knowledge. It is difficult to stop or even catch the scammers since they work so carefully and covertly. In 2017, around 179 million records were hacked[7] . with the most common kind being credit card theft. Addressing credit card theft is crucial because it is the most common kind of fraud worldwide. Since there will be more genuine data than fraudulent data, the credit card dataset is extremely unbalanced. Financial institutions are using EMV cards more frequently because they employ integrated circuits to carry data. Even while certain card purchases are safer as a result, scams involving non-card payments nevertheless occur often. According to a 2017 US Payments Forum survey, fraudsters are less interested in conditioning linked to CNP operations as chip card security has improved.

4 PROPOSED SYSTEM

Card payments are always different when compared to past client payments. Conception drift is the result of this[8]. To put it simply, concept drift is a variable that varies over time and in unexpected ways. The data is highly unbalanced as a result of these conditions. Our investigation's primary goal is to tackle the concept drift issue so that real-world scripts can use it. Machine learning algorithms like Random Forest and decision trees will be used in our suggested system. The algorithm with the best accuracy score will be chosen after their scores have been calculated. In addition to the accuracy score, we will compute the confusion matrix for each algorithm, which we will take into account while selecting one. The optimal algorithm. The fact that the data collection we will be analysing is incredibly unbalanced must also be taken into account.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

5 DATASET DESCRIPTION

The study's Credit Card Fraud Detection dataset can be found on Kaggle. This dataset consists of two days' worth of transactions completed by European cardholders in September 2013.The dataset has 31 numerical features. There were three traits that didn't change. There were 284,807 transactions in the dataset. Of these, 492 were fake, but the rest were authentic. The dataset includes credit card transactions made by European cardholders in September 2013.This dataset contains 492 fraudulent transactions, all of which were finished in less than two days. With 0.172 of all transactions falling into the positive class (frauds), the data set is essentially unstable. Only "time" and "quantum" have not been converted by PCA; v1, v2, v3, and v28 are the main features that PCA has generated.

6 METHODOLOGY

In order to divide the cardholders into three groups or clusters high, medium, and low based on the volume of their transactions, we first use range partitioning and the clustering approach. We use the Sliding-Window technique to arrange the transactions, which entails identifying patterns in the cardholders' behavior by extracting specific window properties. Features include the average amount for the period, the maximum and minimum transaction amounts, and even the amount of time that has elapsed. Algorithm 1: This algorithm uses the sliding window technique to extract cardholder information from aggregated transaction records. input: the customer's ID when holding a card, window size w, and a sequence of transactions t; output: an assortment of transaction data, both genuine and fraudulent, and the cardholder's details;

```
“l: length of T Genuine= [];
Fraud= [];
For i in range 0 to l-w+1: T: [];
/* sliding window features*/
For j in range i+w-1:
/*Add the transaction to window */
T=T+tjid;
End
/* features extraction related to amount */
ai1=MAX_AMT(Ti);
ai2=MIN_AMT(Ti);
ai3=AVG_AMT(Ti);
ai4=AMT(Ti);
For j in range i+w-1:
/* Time elapse */
xi= Time(tj)-Time(tj-1) End Xi= (ai1, ai2,ai3,ai4,ai5,);
Y= LABEL(Ti);
/* classifying a transaction into fraud or not */
if Yi=0 then
Genuine =Genuine U Xi;
Else
Fraud =Fraud U Xi;
End”
```

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

Every time a new transaction is added to the window, the preceding ones are removed, and step two is completed for each group of transactions. These are the algorithms for the sliding window-based aggregation approach. After pre-processing, we train multiple classifiers and extract fraud indicators based on the cardholders' behavior patterns in each category. Even when we utilize classifiers, their performance is poor due to the dataset's imbalance. This method's algorithm modifies the classifier's rating score to ascertain the accuracy of the model. Enter the cardholder's ID along with a history of transactions. Output: The model's rating score following each transaction.

"T: current transaction with w-1 transaction from window.
C: represents the classifier
Label: true value of the incoming/current transaction.
K: total of transactions processed by model.
If the predicted value ≠ label and label==0 then,
For i in range (0, K):
If the predicted value ≠ label then, rsi= rsi-1;
Else
rsi =rsi+1;
End"

7 RESULT ANALYSIS AND DISCUSSION

492 of the 284,807 transactions in the dataset under study were fraudulent, while the rest transactions were valid. With only 0.173 percent of transactions being categorized as fraudulent, the numbers show how biased this dataset is. Class is the only one of the 31 qualities that can have two alternative values: 0 otherwise and 1 in the case of a fraudulent transaction. The PCA dataset selection is evident in accuracy, the most fundamental performance parameter, which is simply the ratio of properly predicted observations to all observed data. If our model is accurate, it can be considered the best. Although accuracy is a useful statistic, only symmetric datasets can use it. The performance achieved by each of the four classifiers under investigation is shown in table1,2,3&4

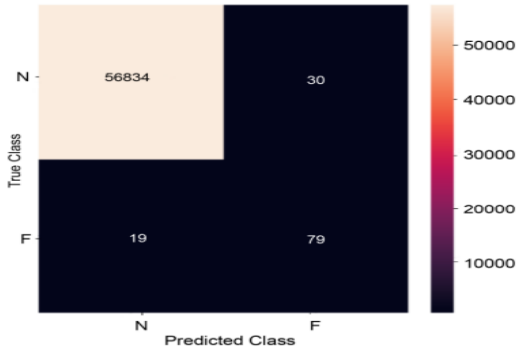

Table 1: Confusion matrix for random forest

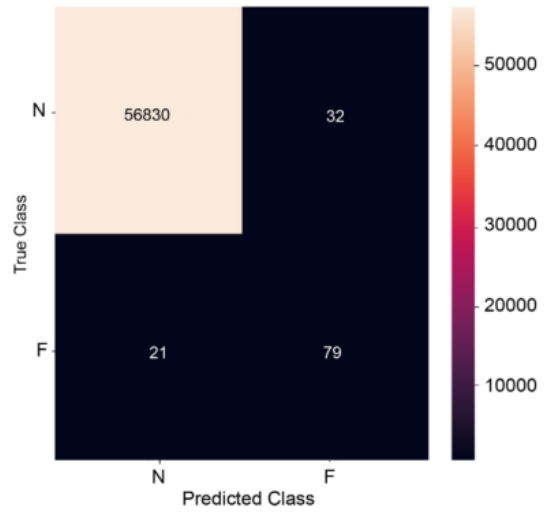

Table 2: Confusion matrix for decision tree

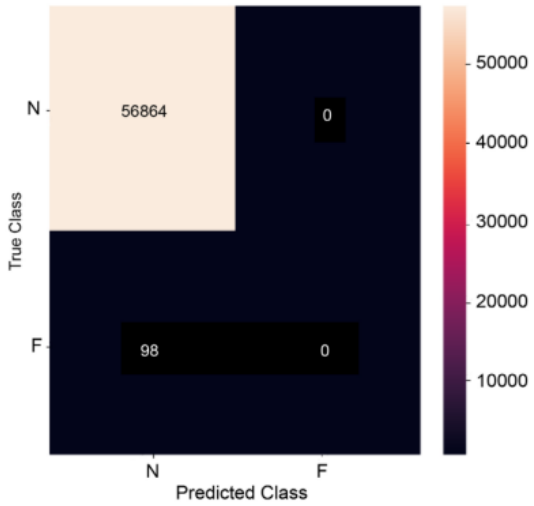

Table 3: Confusion matrix for SVM

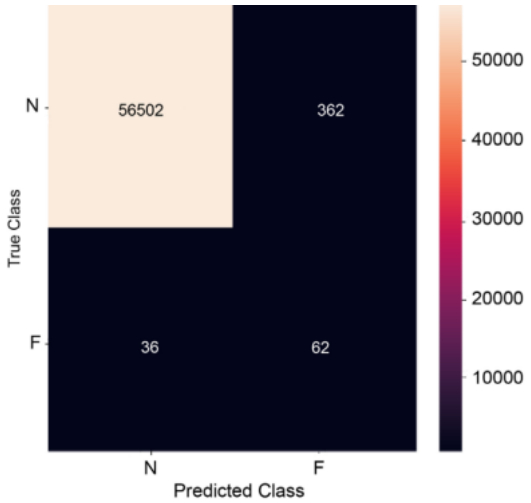

Table 4: Confusion matrix for Naïve Bayes

8 CONCLUSION

A common issue that costs banks, credit card issuers, and customers money is credit card fraud. By using the time and amount variables in the Kaggle dataset, this project aims to create a model that can more effectively distinguish between fraudulent and non-fraudulent transactions, helping banks and customers recover their losses. Using supervised machine learning methods like Naïve Bayes, random forest, decision tree, and support learning, we start by building the model vector machine. In this study, we proposed a solution to address this problem statement. Regarding accuracy, precision, recall, and f1-score, the Random Forest classifier outperformed the other four machine learning classifiers based on the data. As a result, it is the most effective way to spot fraudulent transactions. To find a more effective solution, we will focus on different datasets and machine learning techniques in the future.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922

REFERENCES

[1] X. Niu, L. Wang, and X. Yang, "A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised," 2019, [Online]. Available: http://arxiv.org/abs/1904.10604

[2] O. S. Adebayo, T. A. Favour-Bethy, O. Otasowie, and O. A. Okunola, "Comparative Review of Credit Card Fraud Detection using Machine Learning and Concept Drift Techniques," *Int. J. Comput. Sci. Mob. Comput.*, vol. 12, no. 7, pp. 24–48, 2023, doi: 10.47760/ijcsmc.2023.v12i07.004.

[3] and H. I. A. AlsharifHasan Mohamad Aburbeian, "Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced Data," vol. 17, p. 302, 1385.

[4] K. Randhawa, C. K. Loo, M. Seera, C. P. Lim, and A. K. Nandi, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," *IEEE Access*, vol. 6, pp. 14277–14284, 2018, doi: 10.1109/ACCESS.2018.2806420.

[5] N. R. Dzakiyullah, A. Pramuntadi, and A. K. Fauziyyah, "Semi-Supervised Classification on Credit Card Fraud Detection using AutoEncoders," *J. Appl. Data Sci.*, vol. 2, no. 1, pp. 1–7, 2021, doi: 10.47738/jads.v2i1.16.

[6] L. Y. Apapan Pumsirirat, "Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine," vol. 9, no. 1, pp. 18–25, 2019, [Online]. Available: http://arxiv.org/abs/1912.02629

[7] S. Goel, D. Cards, C. Cards, and P. Cards, "Credit Cards Business- Regulatory nuances from issuance to co- branding Credit Card Issuance and RBI Approval," 2020.

[8] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3637–3647, 2018, doi: 10.1109/JIOT.2018.2816007.

Proceedings DOI: 10.21467/proceedings.7.6
ISBN: 978-81-989164-9-5

Series: AIJR Proceedings
ISSN: 2582-3922