International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

A Novel Machine Learning Technique for Fraud Detection on Credit Card Financial Data

Dr. Neeraj Sharma[1],Dr. M M Venkata Chalapathi[2]

[1]Associate Professor, Department of Computer Science and Engineering, School of Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore, Bhopal, M.P., India.
[2]Assistant Professor Senior Grade-1, School of Computer Science and Engineering, VIT-AP University, Amaravati, A.P., India.

Abstract

The credit card exchanges has expanded decisively in the new year's, as has how much fakes and card robberies. Ordinary robbery, skimming fakes, fake cheats and erroneous card application extortion are ways that Visa misrepresentation can be committed. Albeit the fraudster in every one of these cases uses a material card, an actual belonging isn't expected to commit Visa extortion. "Cardholder-not-present" misrepresentation, in which simply the card's information are introduced, is one of the most well-known kinds of extortion (e.g., via telephone). The most direct kind of credit card robbery is the utilization of a taken card. In this present circumstance, the fraudster generally attempts to burn through however much cash as could reasonably be expected in as brief period as doable. A progression of AI models are utilized to battle misrepresentation, with the ideal arrangement being picked after an assessment. The AI models are utilized to identify fakes and recognizes the certifiable and deceitful exchanges.

KEYWORDS: Credit card frauds, Machine learning, Detecting Frauds, stolen card, fraud detection in real-time, cardholder-not-present.

I. INTRODUCTION

Fraud has increased dramatically due to advanced technology and advances in global communication. Fraud can be prevented in two ways: prevention and detection. Prevention acts as a layer of protection to prevent attacks from fraudsters. Detection occurs after prevention has already failed. Therefore, detection allows you to identify and warn of malicious transactions as soon as they are triggered. Recently, cardless credit card transactions have become popular among web payment gateways. According to the October 2021 Nilson Report, online payment systems generated more than $ 40 trillion worldwide in 2022, an increase of 20.8% from 2018. However, fraudulent transactions have increased significantly and are having a major impact on the economy. There are different types of credit card fraud. Card-not-present fraud and card-present fraud are the two types of fraud that can be recognized in a series of transactions. Bankruptcy fraud, theft/tampering, demand fraud, and behavioral fraud are examples of these two categories of fraud. Our investigation focuses on four types of fraud that fall under the category of cardless fraud and we propose a mechanism to detect these frauds in real time. Machine learning is the solution of this generation to replace such approaches and work with huge data sets that are difficult for humans to manage. There are two types of machine learning techniques: supervised learning and unsupervised learning. Fraud detection can be performed in any approach and the dataset determines when which method is applied. Prior classification of anomalies is required for supervised learning. Several verified algorithms have been used in recent years to detect credit card fraud. The data in this study is analyzed in two ways: as categorical data and as numerical data. Initially, the dataset contained

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

categorical data. Data cleaning and other basic preparation procedures can be used to prepare the raw data. First, the categorical data can be converted into numerical data, then the appropriate evaluation procedures can be used. Second, to select the best method, categorical data is used in machine learning approaches. The objective of this work is to find the best algorithms for each of the four categories of fraud by comparing machine learning approaches and using an effective performance measure to detect fraudulent credits. card transactions. The overview of the literature is presented.

II. LITERATURE REVIEW

Several approaches, including supervised, unsupervised, and hybrid algorithms, have been used in past studies to detect fraud. Every day, new types of fraud and patterns emerge. It's critical to have a thorough understanding of the fraud detection technology. In this section, we'll go through some of the machine learning models, algorithms, and fraud detection models that have been employed in previous research.

In [1] discusses data mining techniques, which require time when dealing with large amounts of data. Another issue with the production of credit card transaction data is overlap. Using sampling strategies, an unbalanced data distribution can be overcome.

In [2] discusses skewed data, which is data that has been skewed. When compared to normal transactions, fraud transactions are quite rare. When a regular transaction appears to be fraudulent or when a fraudulent transaction appears to be legal. Also, talk about the challenges of dealing with categorical data. Categorical data will be ignored by many machine learning techniques. Consider the cost of detection and adaptability as a challenge. The cost of fraud prevention and the cost of fraudulent behaviour are both taken into account.

For fraud detection, [3] uses a variety of models. Different algorithms are utilised in each model. If fresh data exhibits substantial changes in fraud trends, detecting credit card fraud for new frauds will be difficult. It's dangerous to replace the model because machine learning algorithms require a long time to train rather than predict.

In [4] makes use of the Risk Based Multi - model concept. This model can produce good results for data with issues, and the Nave Bayes technique is used to reduce implicit noise in transactions.

Past researchers discovered many issues with fraud detection after analysing various detection models. They identified a lack of real-life data as a major difficulty in [10] and [5]. Because of data sensitivity and privacy concerns, real-world data is scarce. Imbalance data or skewed distribution of data was studied in papers [5] and [7]. The reason for this is that there are many fewer frauds in transaction datasets than there are non-frauds.

They stated that choosing detection methods and feature selection as a hurdle in detecting frauds in [5][6], because most machine learning techniques take much longer to train than to forecast. The feature selection is another important factor in detecting financial fraud. Its goal is to eliminate the characteristics that best describe fraud detection and its qualities.

Except for accuracy, the investigation of extremely imbalanced data in paper [11] demonstrates that KNN performs exceptionally well in terms of sensitivity, specificity, and MCC. The paper [12] reviewed and evaluated commonly utilised supervised learning algorithms. They've also demonstrated that, depending on the challenge, all algorithms alter.

In [8], the K closest Neighbor approach produces good results in terms of performance criteria such as specificity and sensitivity, but the accuracy of the findings is poor.

In [9] focuses primarily on supervised procedures. They conducted a comparative analyses of all algorithms and discovered that algorithms respond differently in different problem situations.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

III. PROPOSED METHOD

A. DATA DESCRIPTION

The fraudulent transaction log file and the all transaction log file were used to create the dataset. The fraud transaction log contains all instances of online credit card fraud, while the all transaction log contains all transactions recorded by the respective bank over a period of time. Some sensitive attributes, such as card number, were encrypted as part of a private disclosure agreement between the bank and the study authors. Due to the odd number of legal transactions and fraudulent events, the shape of the data was very skewed when analyzing the combined dataset. The fraud file contained 200 records, while there were 917781 records in the transaction logfile

The attributes of our two data sources are as follows:

TABLE 1: GENUINE TRANSACTIONS LOG ATTRIBUTES

ATTRIBUTE	**EXPLANATION**
Card No	Card number of the credit cardholder
Date	Transaction date
Time	Transaction Time
Transaction Amount	Amount of the Transaction
Merchant Name	Name of the merchant related to the transaction
Merchant City	Merchant city in which he/she registered
Merchant Country	Merchant country in which he/she registered
Response Code	Transaction relevant ISO response code
Merchant Category Code	Category code of the merchant
Accepted/Not Accepted	Transaction Status

TABLE 2: FRAUDULENT TRANSACTIONS LOG ATTRIBUTES

ATTRIBUTE	**EXPLANATION**
Card No	Card number of the credit cardholder
Date	Transaction date
Time	Transaction Time
Identifier number	Unique Identifier number which was given to frauds
Fraud Nature	Card present/card not present
MCC	Merchant Category
Amount of Fraud	Amount of the transaction
Reversal	Bank related field

B. DATA PREPARATION

First, the raw data was split into four data sets based on the fraud pattern. The information obtained by the bank was used for this purpose. The four datasets are:

1. High Risk Commerce Code (MCC) Transactions.
2. Transactions over $100.
3. Transactions with a potentially dangerous ISO response code.
4. Transactions with unknown web addresses.

These 4 datasets were used in two different ways.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

1. By converting raw data into digital form. (Form A)
2. Categorize the raw data and do not transform it in any way. (Category B)

Datasets 1, 2, and 3 have been assigned to type A, while dataset 4 has been assigned to type B. Data is cleaned, converted, integrated, and minified during data preparation. To prepare the data numerically, the first three datasets underwent all the above steps. With the exception of data transformation, all processes were used to categorize the data.

The basic steps of type A are described below.

Data Cleaning- When cleaning data, it is essential to fill in missing values. There are a number of solutions to this problem, including ignoring the entire tuple, but most will likely skew the data. Since the source file, which contained real transactions, did not contain any missing data, filling it was no longer a problem because the source file, which contained legitimate transactions, did not contain any records with missing data. Unnecessary tuples have been removed from the files as they do not contribute to the production of useful data and do not skew the data. Additionally, adjustments have been made, such as removing unnecessary columns and splitting the datetime column into two.

Data integration – Since the dummy and genuine logs were in two separate files, the two data sources were combined before any other changes were made. The mapping method is illustrated in Figure 1.

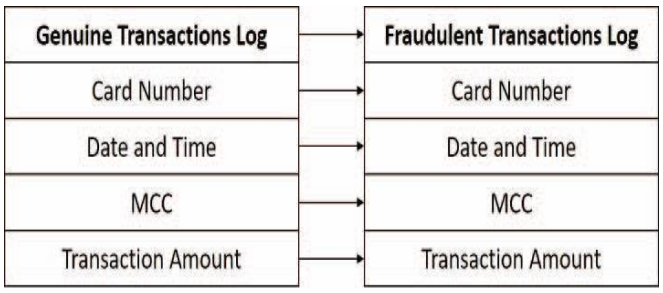

Fig 1: Mapping Data

Data Transformation - All categorical data has been aggregated into an easy-to-understand numerical format. The transaction data set includes a variety of data types and ranges. Therefore, data transformation involves data normalization. Data normalization reduces the numeric range of attribute data to a manageable size.

Data reduction: In this case, dimension reduction was used as a strategy. We must avoid the possibility of learning from erroneous data models, and the selected features must eliminate irrelevant aspects and attributes from the domain of fraud [10]. PCA stands for Principal Component Analysis and is a well-known transformation method. From the point of view of numerical analysis, this strategy addresses the problem of job selection. By determining the correct number of key components, PCA was able to successfully perform feature selection.

Data cleaning and integration was devoted to type B, as well as type A. The data was then sent to the next stage of the procedure C.

C. RESAMPLING TECHNIQUES

The distribution of examples between classes in the two data sources was strongly skewed. The number of examples of fraudulent transactions was significantly lower than the number of genuine transactions. To solve this problem, we use under sampling and oversampling techniques to reduce the number of cases and increase the number of minority cases. Synthetic Minority Oversampling (SMOTE) techniques were used for up sampling, and Packed Nearest Neighbor (CNN) and Random Down sampling (RUS) were used for down sampling. In the SMOTE method there are

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

examples of minority classes. RUS is a non-heuristic method that uses a way to remove random samples from majority classes to balance the class distribution.
We also use a 10-fold cross-validation method. The cross-validated data were then resampled using methods mentioned above.

D. MODELING AND TESTING

The analysis in our study focuses on four different types of fraud. We use the process depicted in Figure 2 to analyze each model. Various methods were used for data analysis. Using the literature, we prioritize four machine learning methods in our research. Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Logistic Regression are the four algorithms. We use our chosen controlled learning classifications to classify our resampled data. The accuracy and performance of each model were considered when selecting the machine learning models capable of absorbing each cheat. Filtering the optimal models against an appropriate performance matrix gave the best results.

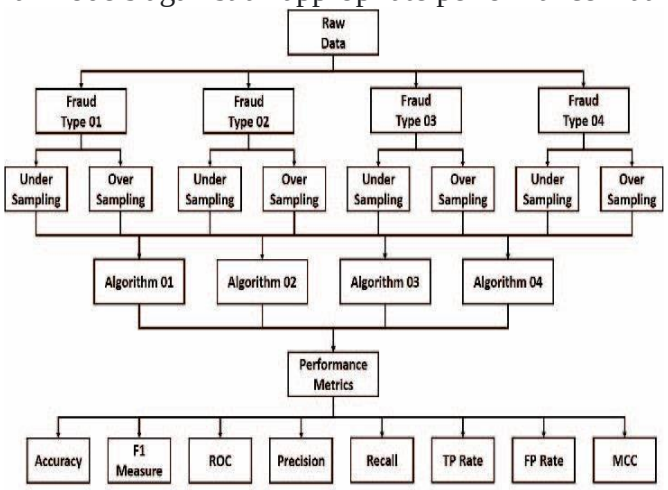

Fig 2: Model Selection

TABLE 3: PERFORMANCE METRICS

Mectrics	**Formula**
Accuracy	(TN+TP)/(TP+FP+FN+TN)
Precision	TP/(TP+FP)
Recall	TP/(TP+FN)
True Positive Rate	TP/(TP+FN)
False Positive Rate	FP/(FP+TN)
F1 Score	2((Precision*Recall)/(Precision+ Recall))
ROC	True positive rate against false positive rate
MCC	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

When the ML classifiers are applied to preprocessed and resampled data, the accuracy rates from the four categories of fraud are shown in the graphs below.

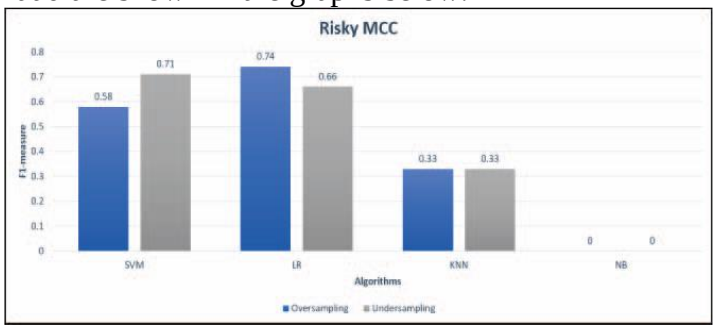

Fig 3: Results of Risky MCC

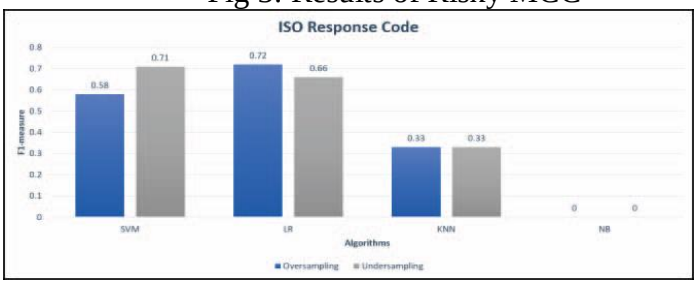

Fig 4: Results of ISO Response Code

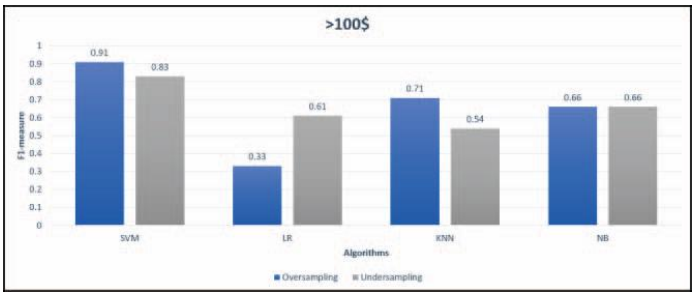

Fig 5: Results >100$

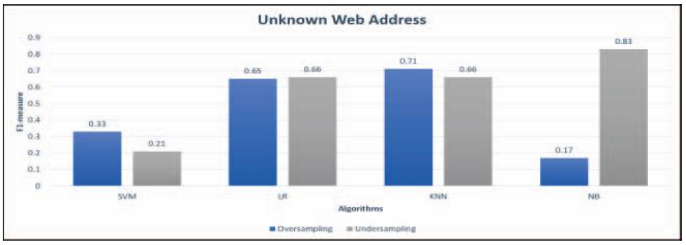

Fig 6: Results of Unknown Web Addresses

E. FRAUD DETECTION IN REAL-TIME

In the past, fraud detection was achieved by analyzing a large number of transactions already made and applying machine learning models to them. Because it takes weeks or months for discoveries to appear, it has been difficult to detect reported fraud and there have been many incidents where scammers were able to make many more fraudulent purchases before being caught. . Running fraud detection models at the time of making an online purchase is known as real-time fraud detection. This allows our system to detect fraud in real time. It sends a notification to the bank, highlighting

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.6 July – 2022
DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

the fraud pattern and its accuracy, so fraud detection teams can easily take the next step without wasting time or money.

F.FRAUD DETECTION SYSTEM

One of the main achievements of the project is the ability to detect credit card theft in real time. The API MODULE, FRAUD DETECTION MODELS and DATA STORE are the three main components of the real-time fraud detection system. All components are simultaneously active in the fraud detection process. Using three verified learning classifications, fraudulent transactions are categorized into four types (risky MCC, ISO response code, unknown web address, and transactions over $100). The API module is responsible for sending transactions in real time between the fraud detection model, the GUI and the data warehouse. Machine learning models live transactions, expected results and other critical data were stored in a data warehouse. The user can interact with the fraud detection system through graphical user interfaces (GUIs), which display real-time transactions, fraud alerts, and historical fraud data in a graphical representation. When the fraud detection model detects a fraudulent transaction, a message is sent to the API module. The API module then sends a notification to the end user and the end user's comments are stored. The complete fraud detection system flow is shown in Figure 7

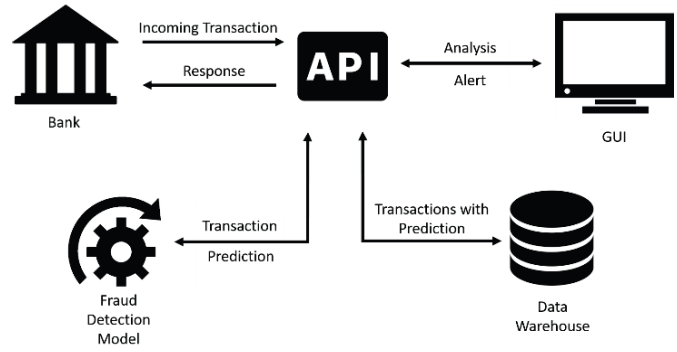

Fig 7: Architecture Diagram

IV. CONCLUSIONS

Investigators have been interested in detecting credit card fraud for years and it will remain an exciting topic of research in the future. This is mainly due to the fact that cheating patterns are constantly changing. In this study, we propose a unique credit card fraud detection system that uses best-in-class matching algorithms to detect four different patterns of fraudulent transactions and addresses relevant concerns noted in previous credit card fraud detection studies. . The end user is notified via the GUI as soon as a fraudulent transaction occurs by addressing real-time credit card fraud detection using predictive analytics and an API module.

Once a suspicious transaction is detected, this part of our system can allow the fraud investigation team to decide whether or not to proceed to the next step. As stated in the methodology, optimal algorithms have been found to deal with four basic types of fraud through literature, experiments, and parameter tuning. We also discuss examples that effectively deal with skewed data distributions. As a result, we can conclude that applying resampling approaches has a significant impact on generating relatively higher classifier performance. The machine learning models with the highest accuracy rates for all four fraud models (risky MCC, unknown web address, ISO response code, transaction over $100) are LR, NB, LR, and SVM. Additionally, the models showed accuracy rates of 74%. , 83%, 72% and 91%, respectively. We plan to work on improving the prediction levels to get better predictions, as the current machine learning models have an average degree of accuracy. Future releases will also focus on location-based fraud.

DOI:10.46647/ijetms.2022.v06i04.0060 ISSN: 2581-4621

REFERENCES

[1] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and G. N. Surname, "Random forest for credit card fraud detection", IEEE 15th International Conference on Networking, Sensing and Control (ICNSC),2018.
[2] Satvik Vats, Surya Kant Dubey, Naveen Kumar Pandey, "A Tool for Effective Detection of Fraud in Credit Card System", published in
International Journal of Communication Network Security ISSN: 2231 – 1882, Volume-2, Issue-1, 2022.
[3] M. Hamdi Ozcelik, Ekrem Duman, Mine Isik, Tugba Cevik, "Improving a credit card fraud detection system using genetic algorithm", published by International conference on Networking and information technology, 2022.
[4] Andreas L. Prodromidis and Salvatore J. Stolfo; "Agent-Based Distributed Learning Applied to Fraud Detection"; Department of Computer Science- Columbia University; 2021.
[5] David Robertson, "Investments & Acquisition-September 2016 Top Card Issuers in Asia-Pacific Card Losses Reach $21.84 Billion, "Nilson Rep., no. 1096, 1090.
[6] J.West and M. Bhattacharya, "An investigation on Experimental Issues in Financial Fraud Mining" ,Procedia Comput.Sci., vol. 80, pp. 1734-1744, 2019.
[7] Z. Zojaji, R. E. Atani, and A. H. Monadjemi, "Survey of Credit Card Fraud Detection Techniques : Data and Technique Oriented Perspective" , pp. 26, 2016.
[8] S.Akila and U. S. Reddy, " Risk based Bagged Ensemble (RBE) for Credit Card Fraud Detection, "Int. Icici, pp. 670-674,
2017.
[9] D.P. Methods, "Data Preprocessing techniques for DataMining,"Science (80-.)., p. 6, 2011.
[10] M.Rafalo, "Real-time fraud detection in credit card transactions," *Data Science Warsaw*. 2017.
[11] J.O.Awoyemi, A.O.Adetunmbbi, and S.A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis,"*2017 Int. Conf. Comput. Netw. Informatics*, pp. 1□9, 2017.
[12] R. Choudhary and H. K. Gianey"Comprehnsive Reviiew On Supervised Machine Learning Algorithms," 2017 *Int. Conf. Mach. Learn. Data Sci.*, pp. 37-43, 2017.
[13] Samaneh Sorournejad, Zahra Zojaji , Reza Ebrahimi Atani , Amir Hassan Monadjemi, "*A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective* ", IEEE 2016