IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Machine Learning Approach To Credit Card Fraud Detection Using Logistic Regression And Random Forest

Sara Tesema, Diriba Giichile ,Eden Selemon, Dagim Dagmawi, Kemal Mohammed

Student, Student, Student, Student
Department of Computer Science
Mattu University, Oromia, Ethiopia

Abstract: Scientific progress in electronic commerce and communication networks led to credit cards becoming the most popular payment instrument supporting in-store and electronic purchasing transactions. The number of credit card payment fraud incidents has substantially escalated because of current transaction trends. The yearly financial losses from fraud damages credit card issuers to such an extent that they need an advanced system with a complete fraud detection framework. Electronic payments have become a leading factor that drives the increase in fraudulent transactions. The urgent need exists for new methods which can detect credit card transaction fraud before it happens. Our fundamental research develops a transaction fraud system which integrates machine learning elements with original feature modification strategies. Our method received validation through experiments conducted with two real-world public credit card transaction datasets where one of them contained actual fraud cases. Our system achieved the highest success rate compared to other competing methods when processing both datasets. Our proposed method proves highly effective based on the achieved results. The tool from our approach strengthens credit card issuer capabilities to recognize fraudulent transactions thus safeguarding their customers while simultaneously minimizing loss and regulatory expense. A protected transaction environment for credit cards can emerge from implementing the procedures we have recommended.

I. INTRODUCTION

Bank or financial institutions issue credit cards to customers for borrowing funds that reach specific levels. Cardholders can make transactions up to the defined spending limits without requiring physical cash. An increasing number of people experience credit card fraud in modern society. Digital economy operations suffer billions of birrs in annual losses because of credit card frauds. As electronic payments Venues in e-commerce keep experiencing growth which allows fraudsters to take advantage of system weaknesses to execute their fraudulent operations. unauthorized transactions. The current systems which detect fraud find it difficult to adjust to new fraud models while coping with unbalanced data as well as instant transaction flows. The system requires solutions for both dealing with imbalanced datasets and operating efficiently on real-time transactions.

This project aims to develop a robust fraud detection system using advanced machine learning algorithms, specifically Logistic Regression and Random Forest. By leveraging these models, the system reaches high accuracy in detecting suspicious activities while continuously evolving to new types of fraud attempts. operate in real-time. The main purpose remains the reduction of financial losses together with transaction security improvements. Financial institutions will establish stronger relationships with their clientele through trust as a result of this project.

II. BACKGROUND

These days credit cards function as one of the most popular tools for both online and offline purchases. and offline transactions. Their ease of use has driven customers to accept them more rapidly. in the expanding ecommerce and electronic payment sectors[1]. However, this widespread use The massive spread of credit cards brought about a dramatic increase of fraudulent dealings that cost businesses and individuals major financial losses. Businesses and individuals. Unlawful payments executed through credit cards form the basis of credit card fraud. Thieves acquire payment credentials either through stolen account details or hacking incidents or from deceptive phishing attacks.[2]

Current fraud detection practices use both rule-based systems alongside statistical models for their operations. Historically profitable but today these fraud detection methods show major operational weaknesses. Foreigners use advanced techniques to avoid detection in fraud patterns which adapt and transform continuously. The ability of fraud patterns to change frequently causes basic rule-based systems to face errors in their detection processes. The available fraud detection datasets contain too few fraudulent transactions compared to regular ones because they represent only a tiny fraction of the total transaction data. The unbalanced nature of fraud patterns creates difficulties for standard detection systems because their accuracy suffers while they produce numerous wrong alarm alerts.[3]

Machine learning has emerged as a powerful tool to address these challenges. Unlike traditional methods, machine learning models can analyze large datasets, uncover hidden patterns, and learn from past data to predict future fraud. These models are adaptable, scalable, and capable of reducing false positives and false negatives significantly. By integrating advanced algorithms like Logistic Regression and Random Forest, machine learning systems can provide a robust framework for detecting fraudulent transactions.

This project builds on these advancements to design a machine learning-based fraud detection system that is accurate, adaptable, and capable of real-time decision-making, ensuring secure transactions and reducing financial risks for stakeholders.[4]

III. STATEMENT OF THE PROBLEM

The expansion of online payments and e-commerce while boosting credit card trades has simultaneously made credit card fraud prevalent across the market. Widespread fraudulent credit card activities harm people and enterprises together with financial organizations while annually causing billions of dollar losses. Paid criminals take advantage of payment system deficiencies through multiple methods including card information theft and phishing techniques and data breach exploitation. The detection of credit card fraud remains difficult to achieve because of multiple important consideration points. The smaller number of fraudulent transactions compared to total transactions results in a very unbalanced dataset. This imbalance makes it difficult for detection systems to accurately identify fraud without generating a high number of false positives, which can inconvenience customers and erode trust.

Additionally, The methods which fraudsters use to deceive systems continuously develop sophisticated patterns to replicate authentic transactions and discover weaknesses in system framework. Traditional rule-based detection systems, which are static and inflexible, often fail to adapt to these new fraud trends, resulting in missed detections and financial losses . Numerous differences in genuine transactions between users because of seasonal variations and geographic factors and individual behavioral conduct proved challenging for detecting genuine from fraudulent activities when relying on static models based on arbitrary thresholds or rules. The existing limitations show that society needs sophisticated fraud detection methods. A solution needs to succeed in managing unbalanced data while adapting to fraud pattern changes together with real-time processing to stop financial losses. The proposed solution uses Logistic Regression and Random Forest algorithms to develop a machine learning-based system because it targets the current data challenges in fraud detection. The models effectively locate underlying data patterns along with adjusting their responses to fresh fraud methods which leads to an efficient large-scale solution against credit card fraud practices.[2]

IV. OBJECTIVE OF PROJECT

The primary goal of this project is to develop a reliable and efficient system for detecting credit card fraud through advanced machine learning techniques. It aims to enhance detection accuracy by effectively distinguishing between legitimate and fraudulent transactions while minimizing false positives and negatives. A key focus is on real-time response, enabling the system to analyze transactions as they occur

and take immediate action against potential fraud. Additionally, the project addresses the challenge of imbalanced datasets, employing strategies to ensure accurate learning from both legitimate and fraudulent transactions. Ultimately, a working prototype will be created to demonstrate the feasibility and effectiveness of using Logistic Regression and Random Forest models for fraud detection, there by improving transaction security and fostering trust between consumers and financial institutions.[1]

Abbreviations and Acronyms

CCFD = Credit Card Fraud Detection

ML = Machine Learning

1.1 Scope of Project

This project is dedicated to creating an advanced fraud detection system tailored specifically for credit card transactions, leveraging the power of machine learning. At its core, the system focuses on accurately identifying fraudulent activities within vast datasets using sophisticated models like Logistic Regression and Random Forest. By enabling real-time processing, it ensures that suspicious transactions are flagged immediately, helping to minimize financial losses and protect users. One of the major challenges in fraud detection is the imbalance in data, where fraudulent transactions make up only a tiny fraction of the total. To address this, the system incorporates specialized techniques to improve detection rates and reduce false alarms, ensuring a smoother experience for customers. Additionally, the system is designed to be adaptable and scalable, capable of evolving alongside ever-changing fraud tactics and handling the growing volume of transactions in today's digital economy. It's also built with integration in mind, making it easy to incorporate into existing payment infrastructures, thereby strengthening overall fraud prevention efforts. By tackling these challenges head-on, the project aims to not only enhance security for users but also build greater trust in digital payment systems, making online transactions safer and more reliable for everyone.[8]

1.2 Significance of Project

The significance of this project lies in its potential to tackle the critical and ever-growing issue of credit card fraud. As digital payments and e-commerce continue to expand, safeguarding transactions has become a top priority for financial institutions, businesses, and consumers alike. This project introduces an advanced fraud detection system that leverages machine learning algorithms to provide a secure, reliable, and scalable solution for identifying fraudulent transactions. By doing so, it aims to achieve several key objectives: reducing financial losses by catching fraud early, enhancing customer trust and experience by minimizing false alarms and ensuring smoother transactions, and addressing the constantly evolving tactics of fraudsters with adaptable and intelligent systems. Additionally, the system is designed to be scalable and versatile, capable of handling high transaction volumes and integrating seamlessly into various payment infrastructures. Ultimately, this project supports financial institutions in their efforts to combat fraud, fostering a safer and more trustworthy digital payment ecosystem for everyone.

1.3 Limitations of project

While the proposed credit card fraud detection system offers significant advancements, it also has certain limitations that need to be addressed to ensure optimal performance. These limitations arise from the inherent complexities of fraud detection and the constraints of current methodologies. First, the system's accuracy is highly dependent on the quality and completeness of the dataset. Incomplete, inconsistent, or noisy data can significantly hinder its performance, leading to less reliable results. Second, the issue of imbalanced datasets poses a major challenge, as fraudulent transactions typically represent only a tiny fraction of all transactions. This imbalance makes it difficult to achieve high detection accuracy without generating excessive false positives, which can undermine trust and usability. Third, the project currently focuses on Logistic Regression and Random Forest models, which, while effective, have their own limitations. Logistic Regression may struggle to capture non-linear patterns in the data, and Random Forest can become computationally expensive when dealing with very large datasets. To overcome these limitations, integrating more advanced models like XGBoost (Extreme Gradient Boosting), K-Nearest Neighbors (KNN), or deep learning techniques could improve accuracy and adaptability. However, these approaches often require greater computational resources, expertise, and time to implement effectively.

Addressing these limitations will be crucial for refining the system and ensuring it meets the demands of modern fraud detection.

V. RESEARCH METHODOLOGY

The project methodology contains a structured approach which leads to the creation of an efficient fraud detection system that can expand to handle large volumes. By leveraging machine learning techniques, the project ensures accurate identification of fraudulent credit card transactions while addressing key challenges like data imbalance and real-time processing requirements. This procedure includes multiple important sequential steps to achieve its objective.[14]

2.1 Data Preparation and Preprocessing

The project starts with obtaining prepared historical credit card transaction records which have both fraudulent and legitimate transactions. The stage begins with data cleaning procedures which eliminate missing data together with duplicates and unneeded information to maintain the dataset quality. The process of handling missing data uses techniques that involve either mean substitution or the removal of entire rows containing missing information. All features must receive normalization and scaling procedures due to algorithms needing equal magnitude values to operate effectively. A series of procedures clean up data while making it consistent and satisfying all analysis requirements.[15]

2.2 Feature Engineering

The project starts by choosing appropriate data from historical credit card transactions which incorporate authentic and fraudulent transactions. After selection of the data set for historical credit card transactions the data cleaning process begins to eliminate missing and duplicate and irrelevant information to keep the data quality high. Mean imputation and removing incomplete rows represent two techniques used specifically for dealing with missing values. All features require feature scaling alongside normalization so algorithms which depend on even magnitude scales can process the data. The established procedures transform data into a cleaned state which becomes ready for data analysis.[13]

2.3 Data Splitting

The available dataset receives a training subset and testing subset distribution. Training occurs using data in the training set but model performance testing requires information from the testing set. Such separation between data ensures the models achieve effective generalization for real-life applications.

2.4 Machine Learning Models

This section elaborates the proper statistical/econometric/financial models which are being used to forward the study from data towards inferences. The detail of methodology is given as follows. Two machine learning algorithms are employed for fraud detection: Logistic Regression and Random Forest. Logistic Regression is a probabilistic model that predicts the likelihood of a transaction being fraudulent based on historical patterns. Random Forest, an ensemble learning method, combines multiple decision trees for robust predictions, effectively handling non-linear data patterns. The models are trained on the training dataset and tested for accuracy, precision, recall, and F1-score on the testing dataset.[16][17]

2.5 Performance Evaluation

The system uses various metrics to evaluate model performance

Table 1:Performance Evaluation

Metric	Description
Accuracy	Measures overall correct predictions.
Precision	Ensures flagged frauds are genuinely fraudulent.
Recall	Captures how well the model identifies all fraudulent transactions.
F1-Score	Balances precision and recall for a comprehensive evaluation

Confusion matrices and visualizations provide additional insights into the models' strengths and weaknesses.

SYSTEM DIAGRAMS

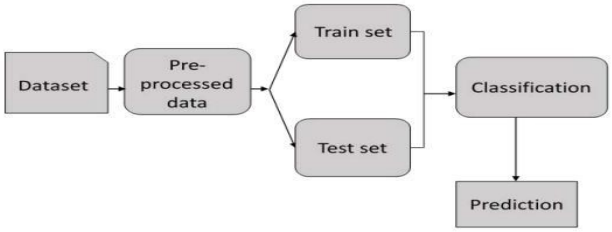

Figure 1 architecture diagram

VI. SYSTEM IMPLEMENTATION METHOD

***Modules*:-**The system implementation method for this project is structured around several key modules that work together to create an effective credit card fraud detection system. These modules include Data Loading and Selection, Preprocessing of the Data, Splitting, Classification, Prediction, and Result Generation.

4.1 Overview of the Modules

A, Choosing and loading data

The process of selecting the right data type, source, and gathering techniques is known as data selection. Prior to the actual data collecting procedure, data selection establishes the pertinent data for analysis. The phrase "data loading" describes the process of retrieving, combining, organizing, and preparing data in preparation for loading it into a cloud data warehouse or other storage system. This project makes use of a credit card dataset to detect fraudulent activity, including time, amount, class, v1, v2, and other pertinent data[9], [19]

B, Preprocessing of data

Data preprocessing is a crucial step in data analysis, as raw data often contains missing or unnecessary components that need to be addressed. Data cleaning involves several strategies to handle missing values, such as manually entering the most likely value, using the attribute mean, or employing a combination of methods. In cases where a tuple contains many missing values within a large dataset, ignoring the tuple can be an effective approach. Another critical aspect of preprocessing is dividing the dataset into training and testing sets, a process known as data splitting. This is typically done for cross-validation, where the available data is partitioned into two parts: one for constructing a predictive model and the other for evaluating the model's performance. Separating the data into training and testing sets is an essential step in assessing the effectiveness of data mining algorithms. Typically, the majority of the data is allocated for training, while a smaller portion is reserved for testing, ensuring a robust evaluation of the model's accuracy and generalizability.

C, Classification

Machine learning, a branch of artificial intelligence, is a data analysis technique that automates the creation of analytical models. Its tenet is that machines are capable of identifying patterns in data, drawing conclusions from those patterns, and making decisions on their own with little assistance from humans. [10]

1.Random forests :- also known as random choice forests, are a type of ensemble learning technique that can be applied to tasks relating to regression, classification, and other related fields. During the training phase of the technique, many decision trees are constructed. In the process of classification, it generates a class based on the individual trees that represents the mode of classes in classification or the mean/average prediction in regression. Modeling the probability of a discrete result given an input variable is known as logistic regression [4].

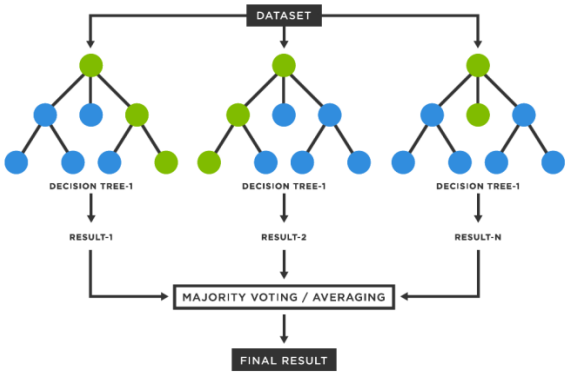

Figure 2: Random forest model

2. Logistic regression models :- which are frequently used to analyze binary outcomes (i.e., true or false, yes or no), reflect the relationship between the input factors and the outcome's log-odds. Predictive analytics algorithms utilize methods like "boosting" and "bagging" to reduce prediction mistakes.[20]

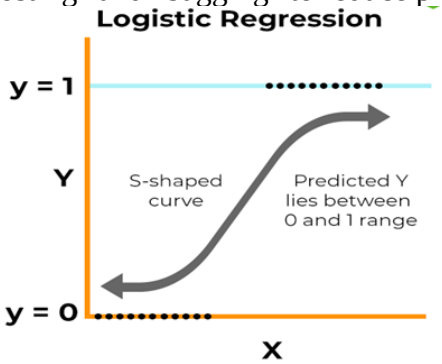

Figure 3: Logistic regression models

D, Prediction

Three key qualities define the strength and reliability of a model: accuracy, robustness, and scalability. Accuracy is the cornerstone of any good classifier, reflecting its ability to correctly predict class labels or estimate attribute values for new, unseen data. Think of it as the model's report card—how often it gets things right. Robustness takes things a step further, measuring how well the model can maintain its performance even when faced with noisy or imperfect data. In real-world scenarios, data is rarely clean or perfect, so a robust model is one that can handle these challenges without faltering. Finally, scalability is about the model's ability to efficiently process and learn from large datasets. As data grows in size and complexity, a scalable model ensures that it can handle the increased workload without sacrificing performance. Together, these qualities—accuracy, robustness, and scalability—make a model not only effective but also practical for real-world applications. [21]

E, Result generation

The goal of this project is to propose a method for detecting fraudulent credit card transactions by leveraging supervised machine learning models. Specifically, we focus on two powerful algorithms: Random Forest and Logistic Regression. These models were implemented in a Python environment using the Scikit-learn (Sklearn) library, ensuring a robust and efficient workflow. we trained and evaluated these models to achieve precise and reliable results.

To assess the performance of each model, we utilized key evaluation metrics such as accuracy, precision, recall, and F1-score. After thorough testing and comparison, Random Fores emerged as the top-performing algorithm, demonstrating the highest accuracy in identifying fraudulent transactions. Logistic Regression also performed well, offering a strong alternative with its simplicity and interpretability. Together, these models provide a balanced and effective solution for credit card fraud detection, combining the strengths of ensemble learning and linear classification to address the challenges of imbalanced datasets and noisy transaction data.

VII. GUI IMPLEMENTATION METHOD

Our credit card fraud detection application builds its graphical user interface through Tkinter which serves as the standard Python library for making user interfaces. Our application begins its process by establishing the primary window which receives its title describing the application goal. The system features an easy-to-use interface which lets users easily operate the software through various simple buttons. The application provides buttons that let users perform actions including dataset loading and model training and comparison graph generation. The application contains a text area which presents output information about model training results alongside evaluation statistics including accuracy reports and classification data. The system design includes a well-ordered interface which allows users to move through the application in an easy and efficient way leading to unencumbered engagement with the fraud detection system. The GUI implements by the system enables users to access complex machine learning operations more easily while utilizing system features.

VIII. CONCLUSION

In conclusion, this project presents a robust machine learning approach to credit card fraud detection using Logistic Regression and Random Forest algorithms. By addressing critical challenges such as data imbalance and the dynamic nature of fraud patterns, the proposed system demonstrates superior performance in accurately identifying fraudulent transactions. The findings indicate that Random Forest is particularly effective, providing high accuracy and reliability in real-time detection scenarios. This work not only contributes to enhancing the security of financial transactions but also fosters greater trust between consumers and financial institutions. Additionally, the insights gained from this study can serve as a foundation for future research and development in fraud detection systems, ultimately leading to more secure digital payment environments. The implementation of these advanced techniques promises to significantly reduce financial losses associated with fraud, ensuring a safer experience for users in the evolving landscape of electronic commerce.

IX. FUTURE WORK

In future, it is possible to provide extensions or modifications to the proposed optimization and classification algorithms using intelligent agents to achieve further increased performance. we want to carry out further research from two aspects.

1. Research on the computational requirements of real-time fraud detection systems forms the basis of the first research direction.
2. The second is to explore the application of more advanced machine learning methods and possible combinations of Decision Tree , XGBoost ,deep learning methods and traditional data mining methods in fraud detection.

IXX. ACKNOWLEDGMENT

First, we would like to extend our sincere gratitude to our mentor/supervisor**, Mr .DirribaG (MSc)** for his constant encouragement and insight. We thank him for suggesting the initial idea of applying machine learning to credit card fraud detection, which eventually formed the core of this project. We greatly appreciate his constant motivation to understand the machine learning concepts that we weren't familiar with. This work would not have been accomplished without his support.

The faculty members of the Computer Science department played an extensive role in preparing us to pursue this research. Without the knowledge gained from the various core courses they taught us, we would not have acquired the technical expertise necessary to undertake such a project. Therefore, on this note, we sincerely thank them for the preparing us for journey that has successfully come to its completion.

Above all, we thank the Almighty God for enabling us to accomplish this proposal.

REFERENCES

[1] N. Sharma and V. Ranjan, "Credit Card Fraud Detection: A Hybrid of PSO and K-Means Clustering Unsupervised Approach," in *Proceedings of the 13th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 445–450. doi: 10.1109/Confluence56041.2023.10048876.

[2] S. Misra, V. O. Matthews, A. Adewumi, O. S. Covenant University (Ota, IEEE Nigeria Section, and Institute of Electrical and Electronics Engineers, *Proceedings of the IEEE International Conference on Computing, Networking and Informatics (ICCNI 2017) : 29-31 October, 2017, Covenant University, Canaanland, Ota, Ogun State, Nigeria*.

[3] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. S. Hacid, and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019, doi: 10.1109/ACCESS.2019.2927266.

[4] P. Anusha, S. Bharath, N. Rajendran, S. D. Devi, and S. Saravanakumar, "Experimental Evaluation of Smart Credit Card Fraud Detection System using Intelligent Learning Scheme," in *Proceedings of the 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems, ICSES 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICSES60034.2023.10465367.

[5] M. Devika, R. Ravi Kishan, L. Sai Manohar, and N. Vijaya, "Credit Card Fraud Detection Using Logistic Regression," in *2nd IEEE International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering, ICATIECE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICATIECE56365.2022.10046976.

[6] S. Maurya, K. Sharma, A. P. Singh, N. P. Tiwari, A. Sharma, and H. Pant, "Credit Card Financial Fraudster Discovery with Machine Learning Classifiers," in *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 2206–2211. doi: 10.1109/IC3I59117.2023.10398009.

[7] B. Bayram, B. Koroglu, and M. Gonen, "Improving Fraud Detection and Concept Drift Adaptation in Credit Card Transactions Using Incremental Gradient Boosting Trees," in *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 545–550. doi: 10.1109/ICMLA51294.2020.00091.

[8] F. Ahmed and R. Shamsuddin, "A comparative study of credit card fraud detection using the combination of machine learning techniques with data imbalance solution," in *Proceedings - 2021 2nd International Conference on Computing and Data Science, CDS 2021*, Institute of Electrical and Electronics Engineers Inc., Jan. 2021, pp. 112–118. doi: 10.1109/CDS52072.2021.00026.

[9] K. Modi, "Review On Fraud Detection Methods in Credit Card Transactions Reshma Dayma."

[10] Univerzitet u Istočnom Sarajevu. Faculty of Electrical Engineering, IEEE Industry Applications Society, Institute of Electrical and Electronics Engineers. Bosnia and Herzegovina Section, Institute