Leveraging Machine Learning for Credit Card Fraud Detection

Kannan N
Department of Artificial Intelligence
Kongu Engineering College
Perundurai-638052
kannan.ai@kongu.ac.in

Saravana P
Artificial Intelligence and Machine Learning
Kongu Engineering College
Perundurai-638052
saravanap.23aim@kongu.edu

Sanjay N
Artificial Intelligence and Machine Learning
Kongu Engineering College
Perundurai-638052
sanjayn.23aim@kongu.edu

Saruha S M
Artificial Intelligence and Machine Learning
Kongu Engineering College
Perundurai-638052
saruhasm.23aim@kongu.edu

***Abstract*-The rise, in credit card fraud cases has prompted the development of detection systems that rely on machine learning technology. This study delves into the use of different machine learning techniques like Logistic Regression,XGBoost and Decision Trees to detect activities in credit card transactions. The effectiveness of each algorithm is evaluated based on its capability to differentiate between fraudulent transactions by considering factors such, as transaction amount, merchant details and customer behavioral trends. Logistic Regression is often utilized as a starting point model because of its simplicity and ease of understanding. XGBoost stands out as a gradient boosting method known for its effectiveness, in managing imbalanced datasets while maintaining high accuracy levels. Ensemble techniques such as Extra Trees and Random Forest are explored for their ability to mitigate overfitting issues and improve the performance of the models. Moreover included is the Decision Tree model due, to its interpretability and implementation process. The models undergo assessment based on metrics, like accuracy and precision as measures such as Mean Squared Error (MSE) and R^2 Score with a focus on reducing both false positives and false negatives occurrences in the results obtained point to ensemble models like XGBoost and Random Forest surpassing individual models in providing a reliable approach, for promptly detecting credit card fraud in real time.**

Keywords—Credit Card Fraud, Machine Learning, XGBoost, Random Forest, Fraud Detection.

I.INTRODUCTION

This has opened floodgates to credit card fraud, while it challenges consumers and financial institutions. Traditional rule-based systems fail to provide solutions to combat the evolving tactics of fraudsters. These developments underpin the urgency for more advanced and automated solutions. Machine learning has emerged as a mainstay tool in fraud detection, exploiting historical transaction information to find suspicious patterns very accurately. Algorithms deployed in the algorithm include Logistic Regression, XGBoost, Extra Trees, Random Forest, and Decision Tree algorithms that can handle high-dimensional imbalanced datasets whereby genuine transactions take a large majority.

The paper analyzes several machine learning models to represent the performance of these models in fraud detection using accuracy, precision, recall as well as the reduction of false positives and false negatives. This work is meant to identify which algorithm would eventually become the most reliable for building more vigorous fraud detection systems that could be used in real-time. These will eventually lead to more secure credit card transactions for consumers and financial institutions. As the tactics of fraud evolve, the necessity for dynamic and adaptive systems becomes critical, and scalability in machine learning is needed to detect new patterns of fraud. These algorithms are constantly learning from transaction data and, therefore, can adapt to fraudsters' new strategies and remain continuously effective in their task. The most important goals are to minimize financial losses caused by fraud and the inconvenience the users experience because of false alerts. These may eventually be part of a future trend where credit card fraud is recognized in real time and the process goes without a hitch for all users.

The inclusion of ensemble techniques such as

Random Forest and XGBoost suggests that the models can improve their resistance to model overfitting and bias in predictions.
Advancements in computational power mean that models of this kind will be able to process large volumes of transaction data with near-instantaneous fraud detection capabilities. Such machine learning models can also be fine-tuned to adapt to different regions, transaction types, and consumer behaviors, hence giving a personalized approach to fraud detection. These models will continue to improve and reduce the operational costs of financial institutions in the quest to improve customer trust in secure transaction processing. Finally, the integration of machine learning into fraud detection systems is a crucial step in combating the global surge in credit card fraud.

II.LITERATURE SURVEY

1. Abdul Rehman Khalid et al., 2024 (MDPI)

The work essentially targets fraud detection improvement by ensemble machine learning strategies. The paper particularly investigates multiple classifiers and their ensemble application toward achieving better fraud-detection rates and enhancing the robustness against changing fraud patterns. Such techniques as stacking and voting classifiers are discussed and clearly depicted with significant performance gains.

2. K. A. Bakar et al., 2023 (JCSTS)

This paper makes a comparative, detailed study of the various models of machine learning for credit card fraud. The models that have been considered here are the Decision Tree, Random Forest, and Logistic Regression, which also take into account accuracy, recall, and F1 score. The study is in support of using Random Forest because it promises good interpretability and precision.

3. Mohammed Asrarulhaq Khadir, 2023

The paper introduces the multi-model approach in fraud detection with the use of classifiers, XGBoost, and Extra Trees combined to get better performance. It also touches on aspects dealing with integrating visualization techniques to make fraud detection insights actionable for users at the end.

4. Haritha Nair et al., 2022 (IEEE Access)

This proposed research presents an optimized Light Gradient Boosting Machine as a model to detect credit card fraud. It addresses the efficiency of the model in processing large datasets and its effective ability to manage the imbalance of fraud cases. It therefore shows that optimized gradient boosting techniques have the capacity for lowering the computational overheads with good precision.

5. Seyedeh Khadijeh Hashemi et al., 2022 (IEEE Access)

The study concludes various techniques of machine learning that can be utilized for fraud detection in banking data. Algorithm selection and feature engineering have been given important considerations in formulating strong systems. The analysis concludes that ensemble methods such as Random Forest and Gradient Boosting hold much promise for improved solutions within prediction capabilities.

6. Zhi-Hua Zhou et al., 2021: IEEE Access

A number of machine learning methods evaluate the performance of credit card fraud detection, particularly in handling class imbalance via Synthetic Minority Over-sampling Technique (SMOTE). Boosting AdaBoost can better improve the model's performance by raising the accuracy of fraud detection and reducing false positives.

III.OBJECTIVES

This study aims to evaluate and compare different machine learning methods for the identification of fraudulent credit card transactions using models such as Logistic Regression, XGBoost, Extra Trees, Random Forest, and Decision Tree. Through the analysis of merchant details, amount, and user behavior patterns of features within transactions, the research establishes a difference in genuine and fraudulent transactions. Another objective of the study is to overcome issues with imbalanced datasets, by using techniques such as SMOTE, to improve model performance. The study will also focus on reducing false positives and false negatives to ensure greater accuracy and reliability in real-time fraud detection.

The motivation behind the project is to investigate the computational complexity and scalability of these models in order to assess their readiness for practical deployment in the context of large-volume transactions. Important feature analysis is planned to be done to determine what aspects contribute to fraud detection, SHAP values, for example, to make the models more interpretative. This study further focuses on combining ensemble methods like AdaBoost and Gradient Boosting towards gaining robust and flexible performance at fraud pattern evolution.

These techniques will also include explainable AI, making the fraud detection process transparent and understandable to users. The proposed system will be benchmarked against traditional rule-based systems to quantify improvements in detection accuracy and adaptability. Finally, this research develops a scalable and cost-efficient fraud detection framework that can be adopted effectively by financial institutions to enhance security in transactions.

IV.PROPOSED MODEL

Our credit card fraud detection system has been developed and tested using four known supervised machine learning algorithms: Logistic Regression, Extra Trees Classifier, Decision Tree, Random Forest, and XGBoost. These are selected based on their performances in dealing with classification problems and for compatibility with different data characteristics.

Logistic Regression Logistic regression is a linear model with a logistic function applied in order to predict the probability of a transaction being fraud or not. It is an efficient and quite interpretable model, and it works very well if data is linearly separable. The model doesn't capture a lot of complex relationships as other algorithms, but it's good for a baseline of comparison.

Extra Trees Classifier: This ensemble method builds a huge number of decision trees and combines the output by averaging. It also works great, particularly in big datasets, and is less likely to overfit at every split because it randomly selects features. The high strength of this is in reducing the variance through averaging multiple decision trees, which enhances its predictiveness.

Decision Trees. Here, decision trees split up the data into smaller segments ranked by importance. It will easily extract patterns related to features. They are intuitive and easy to interpret and can handle categorical data as well as numerical data. They have the ability also to capture non-linear interaction that could be represented in the data. Therefore, they are useful in capturing complex relationships between features.

Random Forest is an ensemble algorithm that can aggregate the decisions of more than one decision tree to increase the accuracy of those decisions and to decrease overfitting. The algorithm is robust and especially useful for imbalanced datasets, which is exactly what fraud detection tasks require-the fraudulent transactions are much fewer in number than legal transactions. In other words, by generating the results from hundreds or even thousands of trees, Random Forest can make more reliable predictions.

XGBoost: XGBoost is an advanced version of gradient boosting, incorporating regularization techniques to optimize prediction performance. It is highly efficient and effective in handling imbalanced datasets, making it well-suited for fraud detection. XGBoost's ability to improve accuracy by boosting weak learners makes it one of the top-performing algorithms in classification tasks, particularly in scenarios with high-dimensional and imbalanced data.

This allowed for a structured comparison of the strengths and weaknesses of each algorithm and was able to identify which model is the most effective in fraudulent-transaction detection. All of these algorithms' unique advantages were exploited for maximizing predictive accuracy and ensuring that a practical yet efficient solution would work in real-world fraud detection systems.

V. DATASET AND PREPROCESSING

Kaggle hosts the Credit Card Fraud Detection dataset, with 284,807 total transactions, but only 492 have been classified as fraudulent, meaning only 0.172% of all transaction instances, which is heavily skewed. Such class imbalance is common in the detection of fraud, where a vast majority of the transactions are legitimate, while only a small fraction of transactions is fraudulent.

Each transaction in the dataset includes 30 features: 28 anonymized using Principal Component Analysis (PCA) to protect user privacy, and two additional features, Time and Amount, which are critical for understanding transaction behavior. Preprocessing the dataset involves several steps to enhance model performance, with a focus on balancing the class distribution and preparing the data for effective learning.

To balance this, we use SMOTE, which means Synthetic Minority Over-sampling Technique, to over-sample the fraudulent transactions and undersample the legitimate ones, thereby creating a relatively more balanced dataset for training models. Algorithms like XGBoost can handle imbalanced data in particular with weighted learning: All of the algorithms in the minority class (fraudulent transactions) are paid with more importance in model training.

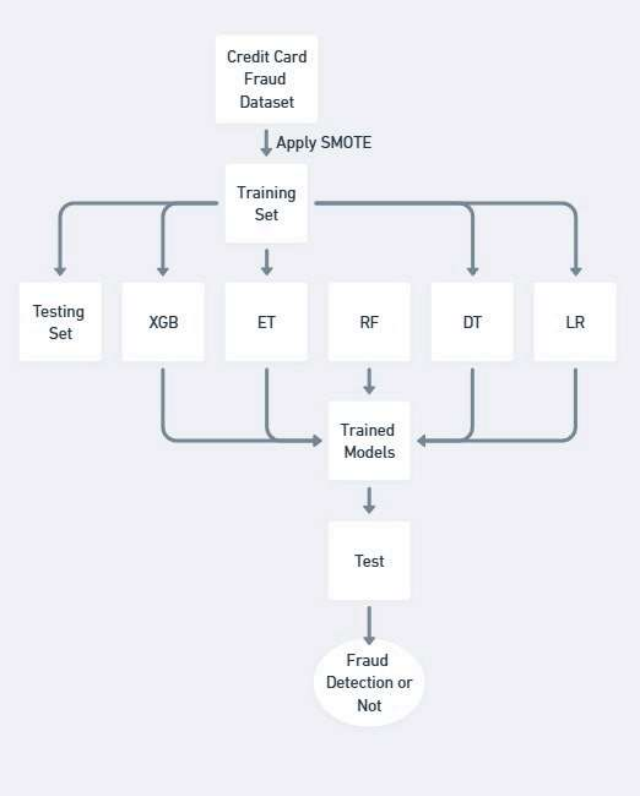
Feature scaling is applied to the Amount feature in order to ensure it is standardized, meaning the values taken may range widely, and hence several

algorithms, such as Logistic Regression, will not perform well. Time is another important feature but is either transformed or excluded based on relevance so that only the most informative features are used in the model.

Feature selection techniques are used to identify redundant or irrelevant features and remove them. Such techniques improve the efficiency of a model and prevent overfitting. Feature importance analysis is another technique that enables the prioritization of variables that are most impactful on the variables, in this case ensuring that fraud detection models concentrate on the most essential information.

The dataset has been divided into training with 80% and testing with 20% for ensuring fair evaluation. The training-testing split can ensure that the performance of the model will be measured based on unseen data. This allows for proper evaluation of its ability to generalize. To further reduce overfitting and ensure that the model performs well in cross-subsets of the data, k-fold cross-validation is implemented. This would divide the training set into k smaller subsets and train on different combinations of those subsets. Such an approach would provide a more reliable estimate of model performance.

Class weights are applied in models like Logistic Regression and Random Forest to give more importance to fraudulent transactions while training. By this way, misclassification would be penalized much more for fraudulent transactions than it would for legitimate ones. In algorithms such as XGBoost, cost-sensitive learning techniques further minimize false negatives, as the fraud detection system requires that not detecting a fraudulent transaction can be pivotal in loss.



VI. FUTURE SCOPE

1. It will be highly profitable for financial houses as the transaction can be easily identified for the fraud in real-time hence saving real-time financial losses.

2. In the future, we may use a combination of more machine learning models to predict future trends and help the system build more good decisions.

3. We can raise features related to user behavior, location, and device details which will eventually make the process of detecting frauds accurate and reliable.

4. The system is scalable to process vast amounts of transaction data, thereby capable of being designed for high frequency applications and global financial networks.

5. We can use techniques based on explainable AI to ensure the transparency of fraud detection decisions so as to gain user trust, and also comply with certain regulatory requirements.

6. Adaptive learning mechanism can be used to allow the system dynamically to keep updating and detecting new patterns that fraud frauds take as they evolve.

VII. RESULT

Techniques	Accuracy
RandomForestClassifier	99.9868%
Logistic Regression	93.9035%
XGBoost(XGB)	99.9745%
Extra Trees Classifier	99.9832%
Decision Tree Classifier	99.8030

The performance of each of these models when evaluated in totality shows that Naïve Bayes performs superbly well since it balances well the given dataset and maintains high values of precision for fraudulent transactions while the SVM demonstrates a strong performance, particularly in the balancing both precision and recall with good values. However, both Random Forest and KNN have difficulties, especially in handling rare fraudulent transactions, leading to a higher number of false positives and false negatives. The Logistic Regression and the Extra Trees Classifier are much more balanced between accuracy and interpretability, making them appropriate for situations where the model transparency is important. Future work would be in hyperparameter tuning and making use of more sophisticated techniques such as ensemble methods or deep learning to improve underperforming models. Additionally, the real-time applicability of each model to the scenario is important because, in a production environment, computational efficiency and detection accuracy form a critical trade-off.

VIII.CONCLUSION

In conclusion, we studied the performances of various machine learning algorithms—Logistic Regression, XGBoost, Extra Trees, Random Forest, and Decision Tree—on credit card fraud detection with a highly imbalanced real-world dataset. At its core, the research aimed to determine how these models classify transactions into fraudulent or legitimate categories. The experiment results showed that ensemble methods such as XGBoost and Random Forest are better in terms of handling class imbalance than the traditional models, which include logistic regression and decision tree models, when accuracy, precision, and recall are being evaluated.

The ensemble models performed well in identifying complex patterns in the data and successfully avoiding overfitting by using multiple decision trees. Also, class weighting and data balancing through oversampling and undersampling proved to be efficient techniques to further improve fraud detection while keeping both false positives and negatives on the low side. The results show the promising potential of ensemble learning methods for improving the reliability and performance of a fraud detection system.

It establishes the foundation for further innovations, including the inclusion of additional features such as behavioral biometrics or history patterns from transactions that could improve the system further. Continuing to update the model in real-time is required to stay up with new fraud techniques and keep the system highly effective over time. In addition, hybrid models, by leveraging the strengths from different algorithms, may produce even better results.

Deploying such systems in real time would require their optimization to process low-latency processing to facilitate fraud detection at the right time. The aim will be to reduce false negatives because undetected fraudulent transactions may bring dire consequences for the consumer and financial institutions alike. Overall, this work demonstrates the capability of machine learning, especially ensemble learning methods such as XGBoost and Random Forest to operate within real-time credit card fraud detection processes. With continuous advancement in machine learning and data preprocessing, the future fraud detection systems are going to be more accurate, adaptable, and efficient to combat ever-evolving fraud strategies.

VIII. REFERENCE

[1] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

[3] D. Geissler, M. Schmidt, and R. Mikut, "Extra Trees: An ensemble learning algorithm for feature importance estimation and its application to correlated input datasets," *Pattern Recognition Lett.*, vol. 149, pp. 82-90, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167865521003447

[4] C. Cortes and V. Vapnik, "Support-vector

networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995. [Online]. Available: https://link.springer.com/article/10.1007/BF00994018

[5] D. Dua and C. Graff, "Credit Card Fraud Detection Dataset," *UCI Machine Learning Repository*, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/credit+card+fraud+detection

[6] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Lett.*, vol. 27, no. 8, pp. 861-874, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865505000673

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, 2002. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302

[8] G. Brown and C. Mues, "An experimental comparison of classification algorithms for credit scoring," *J. Oper. Res. Soc.*, vol. 63, no. 12, pp. 1674-1685, 2012. [Online]. Available: https://www.jstor.org/stable/41508832

[9] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179-188, 1936. [Online]. Available: https://www.cambridge.org/core/journals/annals-of-eugenics/article/abs/use-of-multiple-measurements-in-taxonomic-problems/06FF2F1F64E7B1AA098C07C8D7E9EBD1

[10] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189-1232, 2001. [Online]. Available: https://projecteuclid.org/euclid.aos/1013203451

[11] R. R. Rojas, Neural Networks: A Systematic Introduction, Berlin, Germany: Springer, 1996. [Online]. Available: https://link.springer.com/book/10.1007/978-3-642-60920-8

[12] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol. 19, no. 6, pp. 716-723, Dec. 1974. [Online]. Available: https://ieeexplore.ieee.org/document/1100705

[13] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119-139, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0022000097915127

[14] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993. [Online]. Available: https://www.sciencedirect.com/book/9781558602380/c45-programs-for-machine-learning

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199-222, 2004. [Online]. Available: https://link.springer.com/article/10.1023/B:STCO.0000035301.49549.88