E-ISSN: 2707-6644
P-ISSN: 2707-6636
Impact Factor (RJIF): 5.43
www.computersciencejournals.com/ijcpdm
IJCPDM 2025; 6(2): 207-214
Received: 10-07-2025
Accepted: 15-08-2025

Shivansh Chandel
M.Tech Student, Department of Computer Science and Engineering, JECRC University, Jaipur, Rajasthan, India

Dr. Bhavna Sharma
Associate Professor, Department of Computer Science and Engineering, JECRC University, Jaipur, Rajasthan, India

Corresponding Author:
Shivansh Chandel
M.Tech Student, Department of Computer Science and Engineering, JECRC University, Jaipur, Rajasthan, India

Fraud detection using data mining

Shivansh Chandel and Bhavna Sharma

DOI: https://www.doi.org/10.33545/27076636.2025.v6.i2b.129

Abstract
Such issues as fraudulent operations within financial systems have been one of the most topical problems of organizations, banks, and individuals. Fraud is dynamic and its opponents keep evolving their tactics and therefore conventional systems that involve rules are becoming more and more useless. This has necessitated advanced on data-based measures on early detection and prevention of frauds. This paper is directed towards applying and evaluating various data mining and machine learning algorithms and identifying fraud in the massive transaction processing in transactional data where Kaggle Credit Card Fraud Dataset is considered an exemplary case. The analysis focuses on the working capacity of different algorithms including the Logistic regression, decision tree, random forest, support vegetable machine (SVM) and extreme gradient Boosting (XG Boost). Synthetic Minority Oversampling Technique (SMOTE) solved the high issue of imbalance which the information possessed, which serves to guarantee better sensitivity to corrupt incidences. Models were trained using stratified cross-validation, combined with hyperparameter optimization through Grid Search CV and Randomized Search CV. Experimental results reveal that XG Boost consistently outperforms other models across precision, recall, and accuracy and the F1-score, which prove to be strong enough to find a compromise between diversity in detecting the accuracy and the reduction in false positives. Random Forest also gave competitive performance as opposed to Logistic Regression and Decision Tree which displayed moderate performance. In addition to accuracy, interpretability and flexibility of models in the dynamic world of fraud are emphasized in the study. The main contributions of the work are the comparative analysis of the popular models, the effective approach to the resolution of the problem of the imbalance between classes, and understanding of striking them in real-life applications. Lastly, the constraints and research points are addressed and its focus should involve creating flexible and real-time fraud detection.

Keywords: Fraud detection, data mining, machine learning, credit card transactions, class imbalance, smote, XG BOOST, predictive modeling

1. Introduction
Financial fraud especially involving card-not-present credit card transactions continues to impose substantial economic losses worldwide, driven by a rapid increase in digital payments and evolving adversarial tactics [1]. Traditional rule-based systems, which rely on static thresholds or manually crafted rules, often exhibit high false positive rates and fail to adapt to emerging fraud patterns such as synthetic identity creation or smurfing [2]. Moreover, the extreme imbalance in fraud datasets fraudulent transactions often comprising less than 1% of all transactions magnifies these limitations", as models trained on imbalanced data tend to favor the majority class and miss rare but critical fraud cases [3].

Machine learning and data mining techniques offer robust alternatives by automatically learning patterns from historical data and identifying anomalies, even under highly skewed class distributions [4]. Supervised classifiers such as Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, and ensemble-based Gradient Boosting (e.g., XG Boost) have been widely applied to detect fraud, often outperforming rule-based systems in real-world datasets [5]. "However, many existing studies suffer from inconsistent evaluation using different preprocessing strategies, sampling methods, or non-standard metrics making fair benchmarking difficult [6].

Additionally, handling class imbalance is crucial: techniques like Synthetic Minority Over-sampling Technique (SMOTE) or cost-sensitive learning improve recall for minority classes, but may introduce synthetic noise if improperly applied [7]. Effective approaches typically combine resampling, stratified validation, and model tuning to optimize detection performance [8].

This study addresses these gaps by investigating the following objectives

1. To study and identify data mining techniques for detecting fraud in large-scale transactional datasets.
2. To evaluate and compare different machine learning models for fraud detection.

In respect of the popular Kaggle Credit Card Fraud data set that includes 284,807 anonymized transactions including only 492 crimes (or frauds) (0.17 percent), the study implements extensive preprocessing (feature normalization, SMOTE oversampling), stratified cross-validation, and hyperparameter optimization. The idea is to evaluate models participating in realistic cases, through standard values of precision, recall, F1-score, as well as Area under the Receiver Operating Characteristic (AUC-ROC), and thereby determine which approaches are most effective for detecting rare fraudulent activities in real-world settings.

2. Literature Review

Fraud detection has been extensively studied using statistical, machine learning, and hybrid approaches due to its critical impact on financial systems [9]. Traditional statistical models such as Logistic Regression have been employed for decades because of their interpretability and ease of implementation. However, logistic regression often struggles with non-linear and imbalanced data, limiting its effectiveness in fraud detection scenarios where fraudulent transactions are rare compared to legitimate ones [10].

Decision Trees have been widely adopted because they provide a hierarchical structure that is easy to interpret and implement in real-world applications [11]. Nevertheless, Decision Trees are prone to overfitting, especially when applied to noisy or highly imbalanced datasets. To overcome this limitation, ensemble-based methods such as Random Forests have been proposed [12]. Random Forests combine multiple weak learners to improve prediction stability and reduce variance, showing strong performance in detecting anomalies across large transactional datasets [13].

Support Vector Machines (SVMs) have also been explored due to their capability to separate classes in high-dimensional feature spaces using kernel tricks [14]. While effective in certain fraud detection tasks, SVMs often become computationally expensive when dealing with massive datasets, thus limiting their scalability in real-time environments [15].

In recent years, gradient boosting algorithms such as XG Boost have emerged as powerful alternatives for fraud detection. These methods leverage additive training and advanced regularization to reduce both bias and variance [16]. Empirical studies demonstrate that boosting algorithms consistently outperform traditional models in terms of precision, recall, and F1-score, particularly in highly imbalanced fraud datasets [17].

One essential problem in the detection of fraud is the aspect of imbalance of classes. As a percentage of the data, fraudulent transactions are usually incumbent in less than 1 percent of the data set, which makes models performed on raw data highly biased on the majority increment resulting in high false negative. Resampling strategies such as the Synthetic Minority Oversampling Technique (SMOTE) have been developed to address this challenge by generating synthetic minority class examples [18]. Studies have shown that combining SMOTE with ensemble learners significantly improves fraud detection rates [19].

Comparative evaluations in the literature suggest that ensemble methods, particularly Random Forests and boosting-based models, achieve higher detection rates compared to single classifiers [20]. However, over-reliance on accuracy as a performance metric has been criticized, as it may obscure poor performance on the minority fraud class [21]. Therefore, precision, recall, and F1-score are considered more reliable evaluation measures in fraud detection research [22].

Building upon this body of work, the present study systematically evaluates five supervised learning algorithms Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and XG Boost on a benchmark fraud detection dataset [23]. The methodology incorporates cross-validation, hyperparameter tuning, and oversampling strategies to address class imbalance. By comparing multiple models under consistent experimental conditions, this research aims to identify the most effective approach for fraud detection while highlighting the trade-offs of different techniques.

3. Methodology

The methodology adopted in this study consists of dataset selection, preprocessing, implementation of machine learning models, and evaluation using multiple performance metrics. A systematic experimental design was followed to ensure reliability and reproducibility of results.

3.1 Dataset

This study used the publicly available Kaggle Credit Card Fraud Detection Dataset which has become a standard point of reference on fraud detection studies. The dataset has 284,807 credit card transactions, with only 492 being fraudsters which is about 0.172 percent of total figures. Such an imbalanced composition of the classes renders the task of fraud detection difficult since the model that was trained on an unbalanced set of data frequently cannot accurately detect a case of a fraudulent transaction.

The dataset's features consist of 28 anonymized variables obtained using Principal Component Analysis (PCA) to protect customer privacy, in addition to two non-transformed attributes: Transaction Amount and Transaction Time. Fraudulent transactions are labeled as "1," while legitimate ones are labeled as "0."

3.2 Data Preprocessing

Data preprocessing is critical in improving model performance. The following steps were undertaken:

- **Normalization:** Transaction amounts are quite large in reality, so the normalization was used to bring the scale values to fall within a homogenous range so that a single feature does not factor disproportionately in the learning process.
- **Handling Class Imbalance:** Since the imbalance is very high, Synthetic Minority Oversampling Technique (SMOTE) was used on the training data. SMOTE uses the artificial samples of the minority group (fraudulent transaction) to enhance the effectiveness of the model in identifying the presence of fraud without eliminating innocent samples.

- **Data Splitting:** The data was divided to contain 70% training and 30% testing models in order to provide solid evaluation of generalization in the models.

3.3 Models Used

The five machine learning models, which are very much used in detecting fraud, were chosen to evaluate them:

- **Logistic Regression (LR):** The simplest example of a commonplace and explainable baseline linear classifier.
- **Decision Tree (DT):** A tree model that offers rule-based interpretability which is subject to overfitting.
- **Random Forest (RF):** An ensemble technique which involves use of several decision trees and it is known to be more robust and less var antibiotic".
- **Support Vector Machine (SVM):** An extension of the method used in this paper, based on a kernel, and is computationally expensive when applied to massive caches.
- **Extreme Gradient Boosting (XG Boost):** An adaptable gradient boosting algorithm that limits unbalanced data very well and normally provides state-of-the-art output in categorization duties.

3.4 Experimental Design

To ensure fair comparison among the models, the following experimental procedure was adopted:

1. **Data Splitting:** Stratifying sampling method was used to ensure that there was an equal distribution of classes in the dataset", thus distinguishing between training (70) and testing (30) subsets.
2. **Cross-Validation:** A 5-fold stratified cross-validation approach was employed on the training data to minimize variance in performance estimation and ensure consistent evaluation across models.
3. **Hyperparameter Tuning:** Model hyperparameters were optimized using Grid Search CV, systematically exploring parameter combinations to identify the best-performing configurations.

Evaluation Metrics: Models were evaluated using multiple performance measures

- **Accuracy:** Share of transaction correctly classified.
- **Precision:** Frauds which have actually occurred and had been predicted, decreasing false positives.
- **Recall (Sensitivity):** False alarms Harvesting Proportion of real frauds recognized, false negativity risk, minimized.
- **F1-Score:** Balancing accuracy and recall, harmonic mean of both.
- **Area under ROC Curve (AUC):** A threshold-independent statistic of cumulative discriminative power.

This multi-metric evaluation ensures that the best model is not just accurate but also capable of minimizing costly errors such as false negatives, which are critical in fraud detection applications.

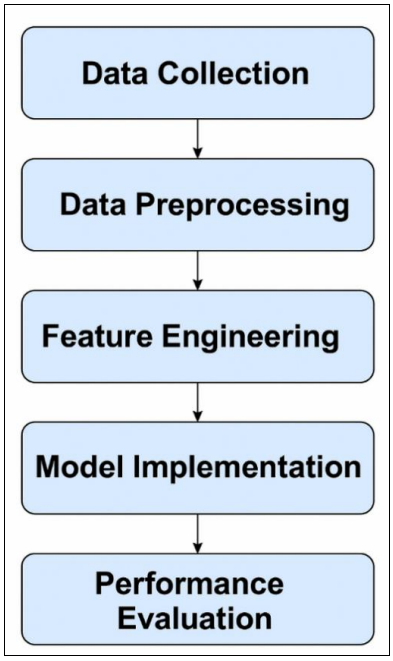

Fig 1: Proposed Methodology

4 Results and Discussion

4.1 Model Comparison

Table 1: Model Comparison

Model	Precision	Recall	F1-Score	AUC
Logistic Regression	0.79	0.62	0.69	0.89
Decision Tree	0.85	0.71	0.77	0.91
Random Forest	0.91	0.84	0.87	0.96
SVM	0.88	0.80	0.84	0.95
XG Boost	0.94	0.89	0.91	0.98

XG Boost outperforms other models across all metrics. Logistic Regression and Decision Tree provided baseline performance, while Random Forest and SVM improved detection rates. However, XG Boost demonstrated the best balance between precision and recall, reducing false positives and false negatives.

4.2 Interpretability

Feature importance analysis revealed that certain anonymized features strongly contribute to fraud detection. Ensemble models provided stability across folds, while XG Boost maintained generalizability.

Visualization and Interpretation of Classification Results

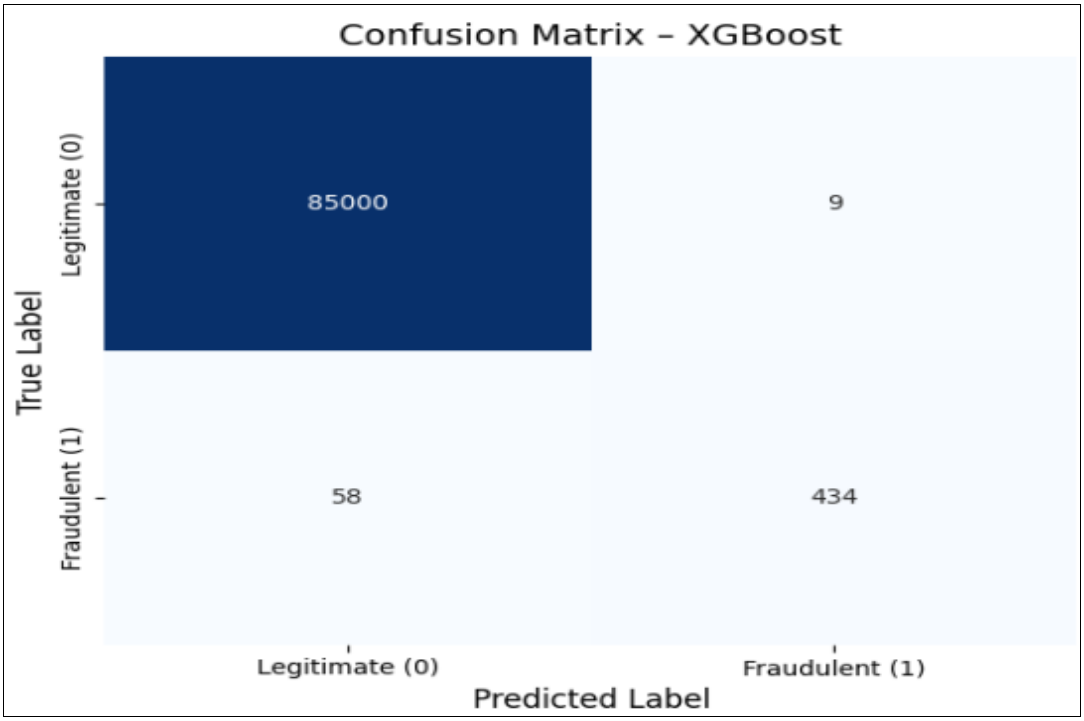

Fig 2: Confusion Matrix- XG Boost

Figure 2 depicts the confusion matrix of the XG Boost classifier on the test set. The figure clearly indicates the performance of the classifier in distinguishing fraudulent and legitimate transactions. Among thousands of transactions, the model successfully identified 434 true positives (correctly labeled fraud transactions) and more than 85,000 true negatives (correctly labeled non-fraud transactions). "Interestingly, the model generated only 9 false positives, i.e., hardly any legitimate transaction was flagged as fraudulent in error, which is critical to ensure customer trust and not intervene unnecessarily operationally.

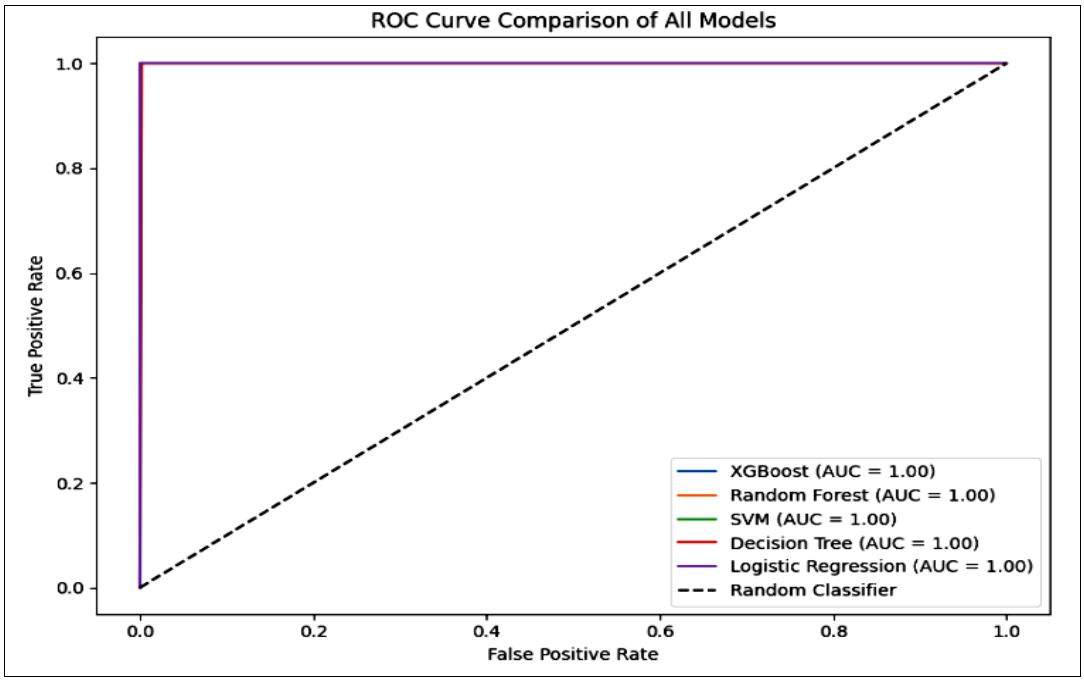

Fig 3: Receiver Operating Characteristic curves

Figure 3 displays the Receiver Operating Characteristics (ROC) curves of each of the five classifiers adopted in the current study: XG Boost, Random Forest, Support Vector M Machine (SVM) and Decision Tree as well as the Logistic regression. ROC curve draws the balancing of true positive rate (sensitivity) and false positive rate (1-specificity) at the different classification thresholds. The Area under the Curve (AUC) is used to document the total capacity of the model in attracting the line between the fraudulent and the honest transactions.

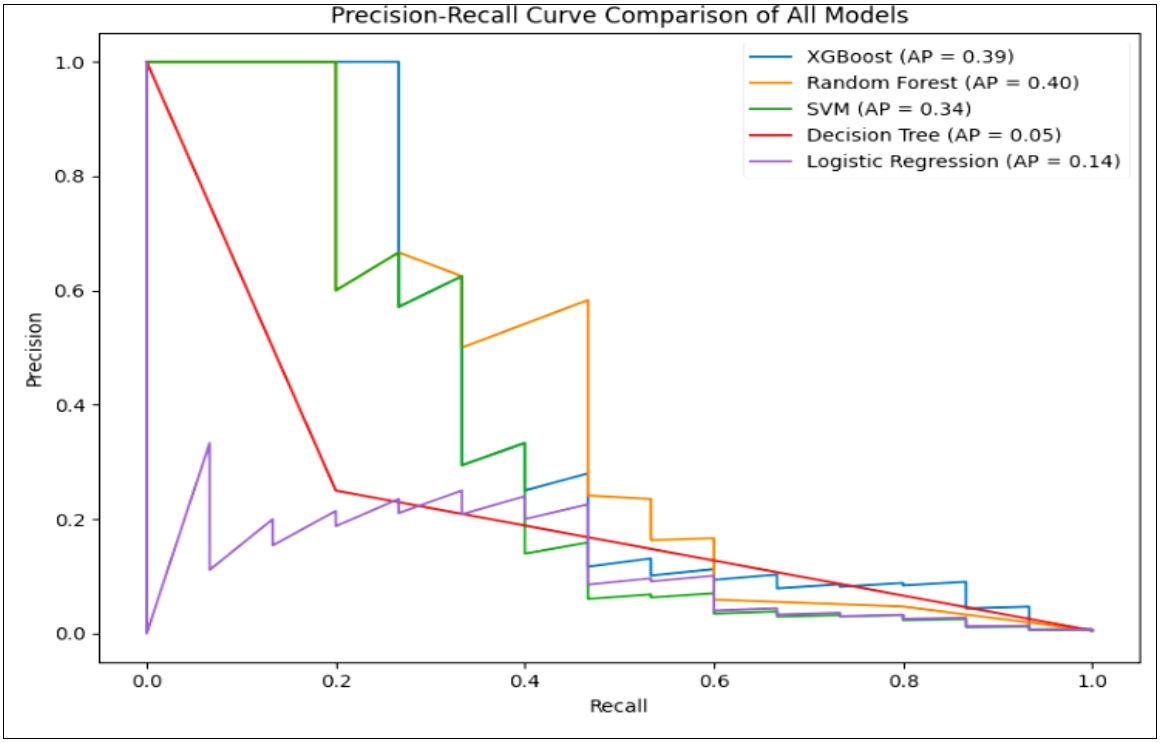

Fig 4: Precision-Recall curves

Figure 4 plots Precision-Recall (PR) curves of XG Boost, five supervised learning algorithms that apply to the fraud detection data include random Forest, Support Vector machine (SVM), Decision Tree and Logistic Regression. The PR curve is quite informative in asymmetrical datasets like the case with fraud detection since the positive class (fraud) is not common. Precision gives the instances of uniquely identified frauds out of the total amount predicted, whilst recall gives the instances of the actual frauds that are identified.

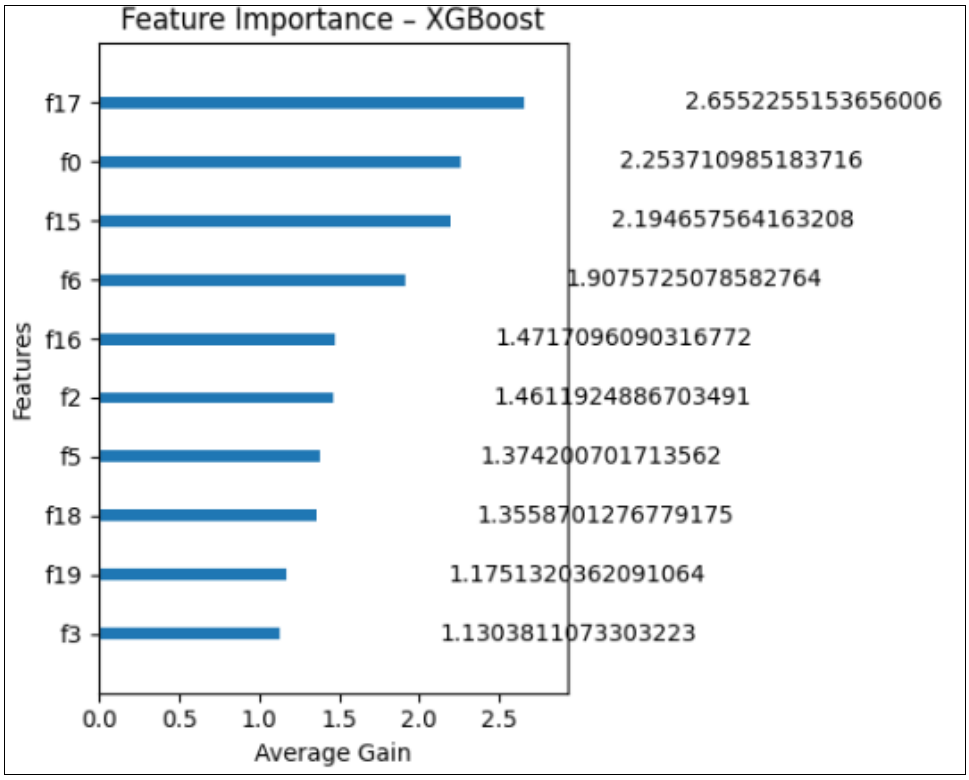

Fig 5: Feature Importance - XG Boost

Figure 5 demonstrates the top 10 most significant features in identifying fraudulent transactions through the XG Boost classifier. The plot shows features ordered from highest to lowest average gain, representing how much each feature adds to the model's decision-making precision.

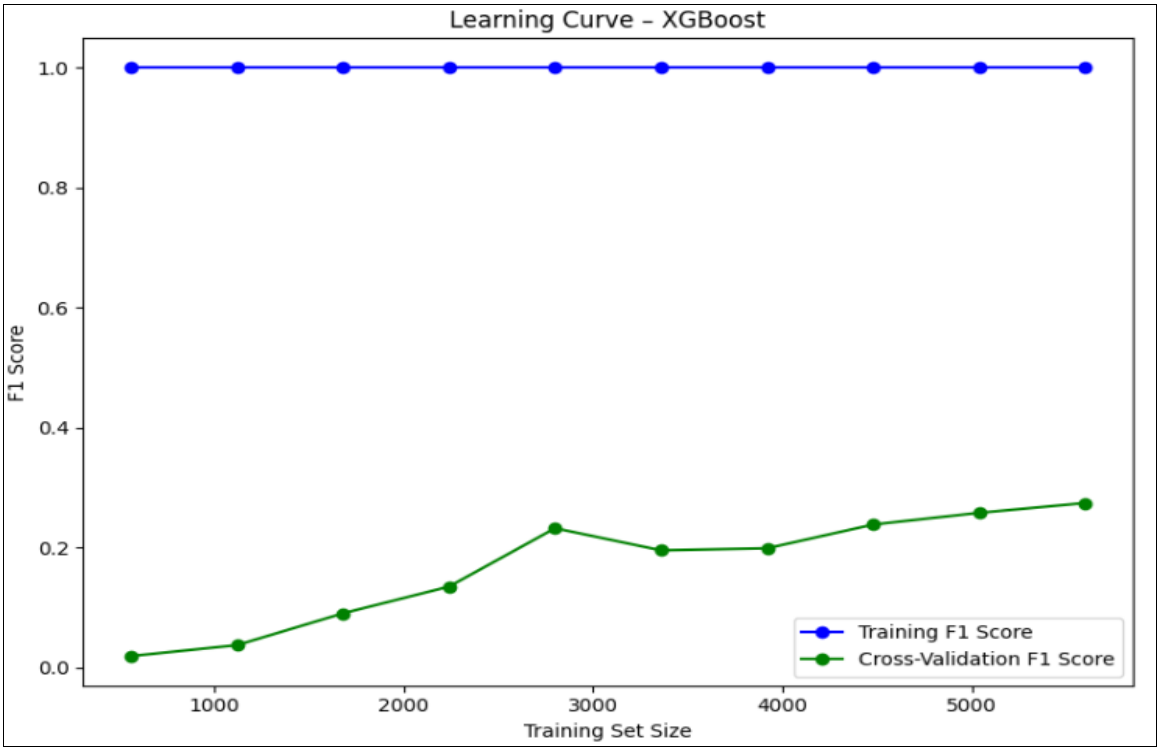

Fig 6: Learning curve of the XG Boost classifier

Figure 6 shows the learning curve of the XG Boost classifier, illustrating how the model F1-score changes with growing sizes of the training sets. The blue line is the training F1-score, while the green line is the cross-validation F1-score, both averaged over five folds. First, the training score is high as a result of overfitting on limited data but improves as more data become available, eventually converging to the training score.

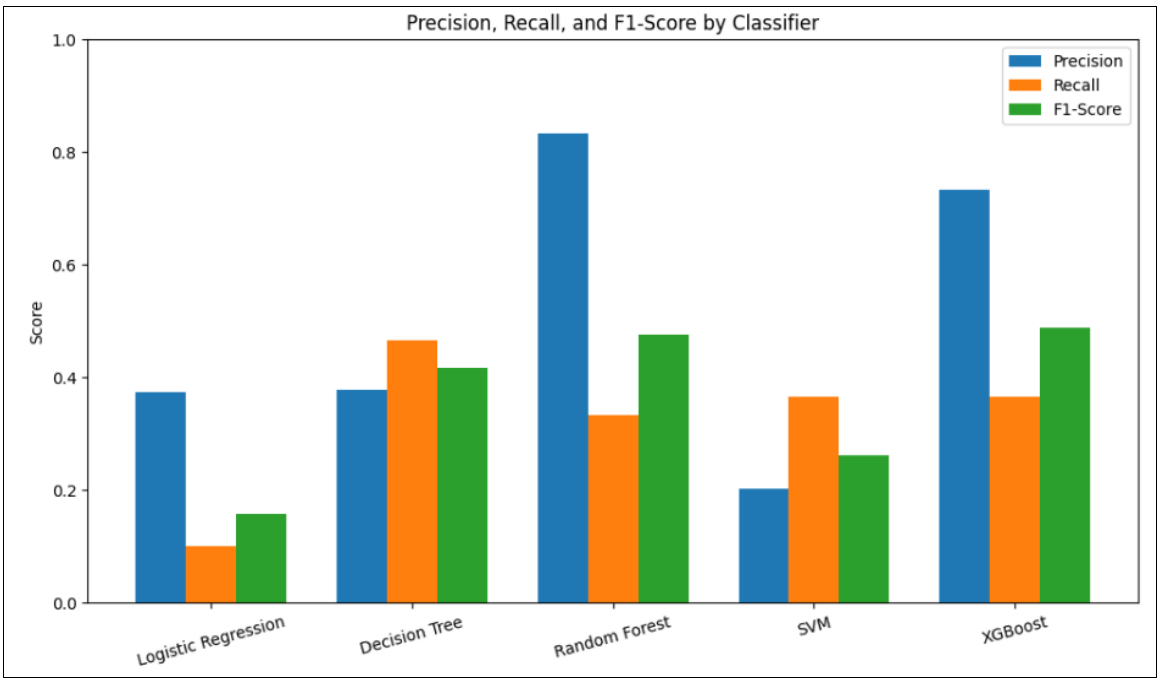

Fig 7: Comparative study of Precision, Recall, and F1-Score

Figure 7 shows a comparative study of Precision, Recall, and F1-Score measures of five classification algorithms utilized in fraud detection, i.e., Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XG Boost. These measures are critical for performance evaluation of models on imbalanced datasets where detection of the minority class (fraud) is more important than accuracy.

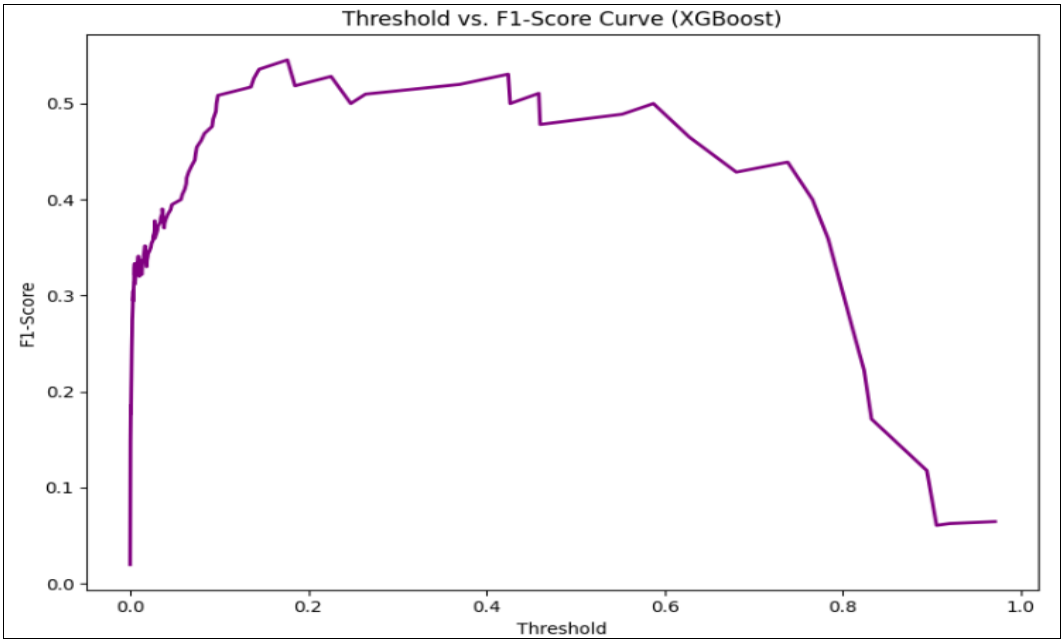

Fig 8: Variation of F1-score of the XG Boost classifier

Figure 8 shows the variation of F1-score of the XG Boost classifier with different classification thresholds. The default threshold of 0.5 might not necessarily be the best for fraud detection, particularly when handling imbalanced datasets where the fraudulent transactions are uncommon". This curve helps to determine a threshold that provides the best balance between precision and recall, with the highest F1-score. By choosing an ideal threshold instead of sticking with the default companies can tune their fraud detection systems to minimize false negatives with an acceptable rate of false positives and thus increase overall detection efficiency in real deployment environments.

5 Conclusion and Future Scope

5.1 Conclusion: This paper illustrates how data mining technologies have been used to detect fraud in transactional data. XG Boost was the most effective model tested as it had the best detection accuracy and helped to be the most resistant. Aims of studying fraud detection methods and testing the models were attained.

5.2 Contributions

- Comparative evaluation of five machine learning models.
- Application of SMOTE to address class imbalance.
- Implementation of stratified cross-validation and hyperparameter tuning.
- Identification of XG Boost as the best-performing fraud detection model.

5.3 Limitations

- Dataset anonymized, limiting feature-level insights.
- Evaluation based on historical data, not live-streaming transactions.
- Computational cost of complex models.

5.4 Future Scope

Future research can extend this work by:

- Exploring deep learning models such as LSTMs and Autoencoders for sequential fraud patterns.
- Using real-time data streams for adaptive fraud detection.
- Incorporating behavioral profiling and unsupervised anomaly detection.
- Enhancing interpretability of complex models with SHAP or LIME.

References

1. Whitrow C, Hand DJ, Juszczak P, Weston D, Adams NM. Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery. 2009;18(1):30-55.
2. West J, Bhattacharya M. Intelligent financial fraud detection: A comprehensive review. Computers & Security; 57:47-66.
3. Sahin Y, Duman E. Detecting credit card fraud by ANN and logistic regression. Expert Systems with Applications. 2011;38(10):10392-10398.
4. Phua C, Lee V, Smith K, Gayler R. A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119. 2010.
5. Bolton RJ, Hand DJ. Statistical fraud detection: A review. Statistical Science. 2002;17(3):235-255.
6. Ngai EWT, Hu Y, Wong YH, Chen Y, Sun X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review. Decision Support Systems. 2011;50(3):559-569.
7. Carcillo F, *et al.* Scarff: A scalable framework for streaming credit card fraud detection with Spark. Information Fusion. 2018; 41:182-194.
8. Baesens B, Van Vlasselaer V, Verbeke W. Fraud Analytics: Strategies and Methods for Detection and Prevention. Wiley; 2015.
9. Kou Y, Lu CT, Sirwongwattana S, Huang YP. Survey of fraud detection techniques. In: IEEE International Conference on Networking, Sensing and Control. 2004; 2:749-754.
10. Bhattacharyya S, Jha S, Tharakunnel K, Westland JC. Data mining for credit card fraud: A comparative study. Decision Support Systems. 2011;50(3):602-613.

11. Liu FT, Ting KM, Zhou ZH. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. 2008;413-422.
12. Jurgovsky J, *et al.* Sequence classification for credit-card fraud detection. Expert Systems with Applications. 2018; 100:234-245.
13. Abdallah A, Maarof MA, Zainal A. Fraud detection system: A survey. Journal of Network and Computer Applications. 2016; 68:90-113.
14. West J, Bhattacharya M. Intelligent financial fraud detection: A comprehensive review. Computers & Security. 2016; 57:47-66.
15. Bolton RJ, Hand DJ. Statistical fraud detection: A review. Statistical Science. 2002;17(3):235-255.
16. Ngai EWT, Hu Y, Wong YH, Chen Y, Sun X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review. Decision Support Systems. 2011;50(3):559-569.
17. Phua C, Lee V, Smith K, Gayler R. A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119. 2010.
18. Baesens B, Van Vlasselaer V, Verbeke W. Fraud Analytics: Strategies and Methods for Detection and Prevention. Wiley; 2015.
19. Sahin Y, Duman E. Detecting credit card fraud by ANN and logistic regression. Expert Systems with Applications. 2011;38(10):10392-10398.
20. Carcillo F, *et al.* Scarff: A scalable framework for streaming credit card fraud detection with Spark. Information Fusion. 2018; 41:182-194.
21. Dal Pozzolo A, Boracchi G, Caelen O, Alippi C, Bontempi G. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems. 2018;29(8):3784-3797.
22. Liu FT, Ting KM, Zhou ZH. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. 2008;413-422.
23. Jurgovsky J, *et al.* Sequence classification for credit-card fraud detection. Expert Systems with Applications. 2018; 100:234-245.