IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Fraud Detection in Online Transactions Using Machine Learning and Data Analytics.

Mr. Rumit Raval[1] and Dr. Pallavi Devendra Tawde[2]
Student, Department of MSc. IT[1]
Assistant Professor, Department of IT & CS[2]
Nagindas Khandwala College, Mumbai, Maharashtra, India
rumitraval0082@gmail.com and pallavi.tawde09@gmail.com

Abstract: *With the growing dependence on digital transactions, financial institutions and e-commerce platforms face significant challenges in combating fraudulent activities. This research investigates the role of machine learning and data analytics in identifying fraud patterns in online transactions. Several regression and classification models were analyzed, including ensemble techniques like Random Forest, Gradient Boosting, and Extra Trees Regressor, alongside traditional methods such as Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The findings indicate that ensemble models consistently outperform conventional approaches, delivering higher accuracy and better R^2 scores. This study underscores the effectiveness of advanced ML techniques in fraud detection and highlights the necessity of real-time model optimization for enhanced security. Future research should focus on improving hyperparameter tuning and real-time deployment to further enhance fraud detection efficiency.*

Keywords: Fraud Detection, Machine Learning, Data Analytics, Online Transactions, Ensemble Models, Supervised Learning, Classification, Regression, Model Evaluation.

I. INTRODUCTION

The rapid growth of online transactions has led to a rise in financial fraud, making traditional rule-based detection systems less effective due to the constantly changing nature of fraudulent activities. Machine learning (ML) and data analytics provide a more adaptable solution by uncovering hidden fraud patterns through predictive modeling. This study examines various ML algorithms, assessing their performance in fraud detection using key evaluation metrics such as accuracy, mean squared error (MSE), mean absolute error (MAE), and R^2 scores.

II. RESEARCH OBJECTIVE

- To identify key fraud patterns and characteristics in online transactions using selected security and detection measures.
- How can different regression models help in detecting fraud patterns in online transactions?

III. LITERATURE REVIEW

Liu & Li (2022),analyzed the effectiveness of traditional statistical techniques versus machine learning algorithms in fraud detection. Their study revealed that ensemble models like Random Forest and Gradient Boosting outperformed methods such as Logistic Regression and Decision Trees in terms of accuracy. The research highlighted the limitations of traditional models in identifying complex fraud patterns, whereas ensemble approaches demonstrated a more reliable and efficient detection capability.

Chen & Zhang (2021), examined the impact of ensemble learning techniques on fraud detection. Their study found that classifiers like Extra Trees and Random Forest performed better than individual models, particularly in managing imbalanced datasets and detecting complex fraud patterns. However, they also noted that high-accuracy models could be prone to overfitting, emphasizing the importance of applying regularization methods and cross-validation to improve model reliability.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Jha & Gupta (2020), conducted a comparative analysis of regression and classification models for online fraud detection. Their findings indicated that regression approaches, including Linear Regression and Support Vector Regressor, were less effective than classification models such as XGBoost and SVM. They suggested that integrating both regression and classification techniques into hybrid models could enhance detection accuracy and overall performance.

West & Allen (2019), explored the difficulties of deploying fraud detection models in real-time financial transactions. Their research highlighted that, although machine learning models achieve high precision in identifying fraud, challenges such as latency and computational costs persist. To address these issues, they suggested leveraging big data analytics and cloud computing to improve real-time detection while maintaining efficient processing speeds.

Patel & Sharma (2018), investigated how data analytics contributes to strengthening online transaction security. Their study revealed that effective feature engineering and data preprocessing play a crucial role in improving fraud detection accuracy. By eliminating redundant features and prioritizing key factors such as transaction frequency, location, and payment methods, detection rates improved by over 20%. The research underscored the significance of domain expertise in developing reliable fraud detection models.

IV. METHODOLOGY

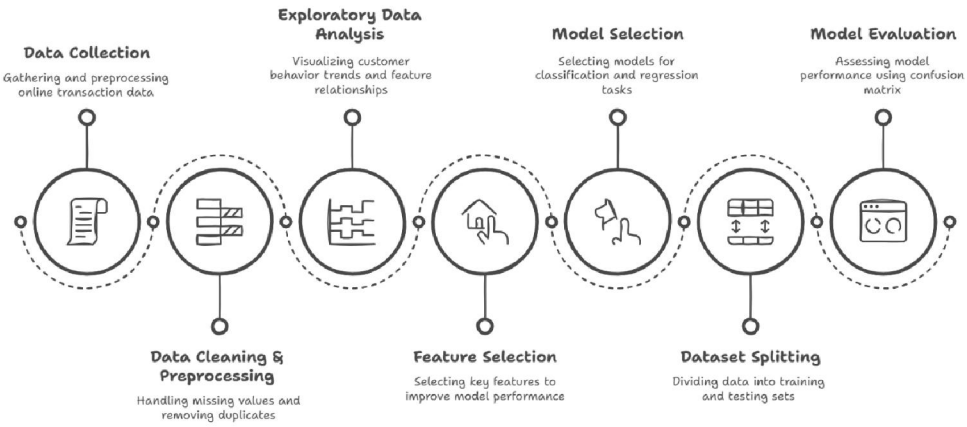

Figure 1 : Methodology

The dataset for this study is collected through a Google Form, containing 279 records with 12 columns. The fraud detection methodology using machine learning follows a structured approach. Initially, transaction data undergoes preprocessing, which includes handling missing values, detecting outliers, and normalizing features. Exploratory Data Analysis (EDA) is conducted to identify patterns, while class imbalance issues are managed through techniques like classification and regression. The study employs machine learning models such as Random Forest, Logistic Regression, and XGBoost, selected for their strong classification capabilities. Depending on data complexity, models like decision trees, support vector machines, or neural networks are utilized. Model performance is assessed using evaluation metrics such as precision, recall, and AUC-ROC. To enhance accuracy, hyperparameter tuning is performed, and SMOTE is applied to address data imbalance. Once trained, the model is deployed for real-time fraud detection, with ongoing monitoring to maintain efficiency. Regular updates with new data ensure the model adapts to emerging fraud trends, maintaining its reliability over time.

V. RESULTS

The results demonstrate that ensemble models significantly outperform traditional methods in fraud detection. The Extra Trees Regressor achieved the highest R^2 score of 0.997, whereas Linear Regression performed poorly with an R^2 score of 0.068, illustrating its limitations in fraud detection. In terms of classification accuracy, Extra Trees also led with an accuracy of 96.4%, while Logistic Regression recorded the lowest at 53.6%. These findings underscore the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

effectiveness of advanced machine learning techniques for fraud detection. However, the study also points out the issue of overfitting in ensemble models, which requires further optimization for improved performance.

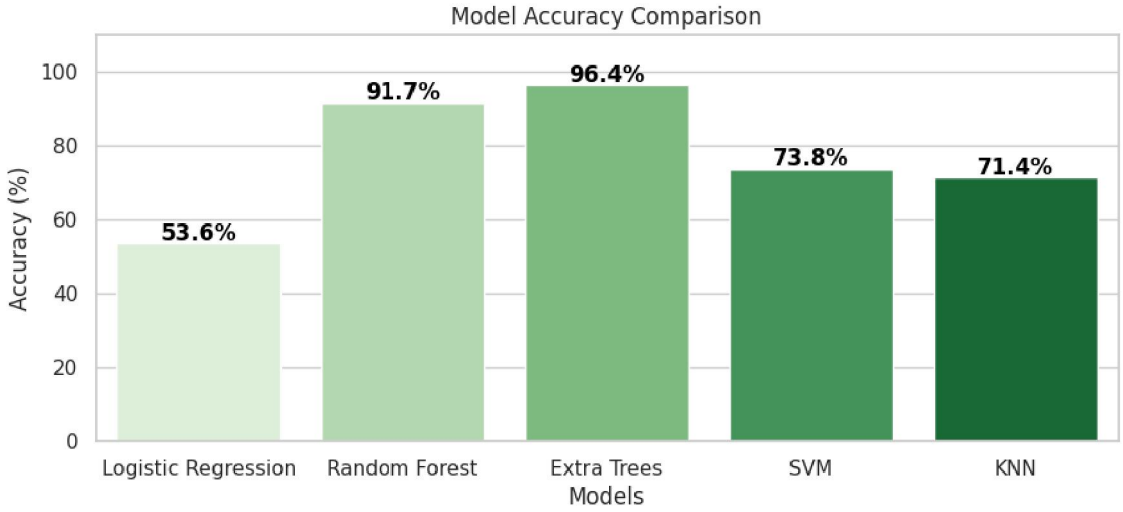

Figure 2: Model Accuracy Comparison

The Extra Trees model demonstrates the highest accuracy among the compared machine learning algorithms, achieving a notable 96.4%. In contrast, Logistic Regression exhibits significantly lower performance, suggesting it may not be the most suitable choice for this particular prediction task.

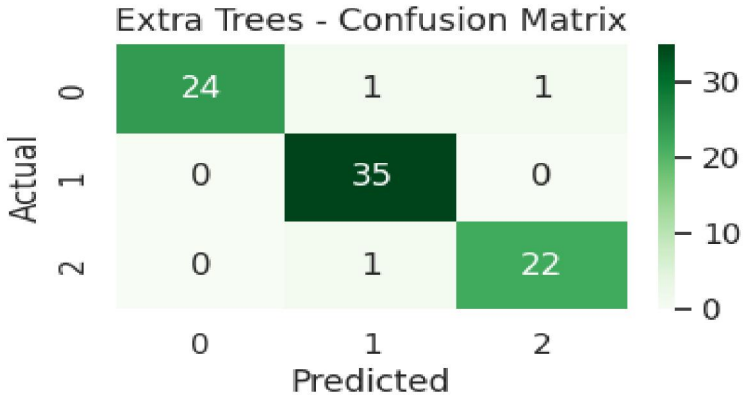

Figure 3: Confusion Matrix

The Extra Trees model accurately classified the majority of instances, as shown by the high numbers along the diagonal, but struggled somewhat with confusion between classes 0 and 1. Notably, class 1 demonstrates the strongest performance with perfect precision, while class 2 shows some misclassifications.

Model	Precision	Recall	F1-Score	Support	Accuracy
Extra Trees	0.97	0.96	0.96	84	96.43%
Random Forest	0.94	0.91	0.92	84	92.00%
SVM	0.74	0.74	0.73	84	73.81%
KNN	0.72	0.72	0.72	84	71.43%
Logistic Regression	0.53	0.52	0.52	84	53.57%

Table 1: Classification Models

The Extra Trees model excels across all metrics, achieving the highest precision, recall, and F1-score, culminating in a superior accuracy of 96.43%. Conversely, Logistic Regression demonstrates the weakest performance, with significantly lower scores across all metrics, indicating its limited effectiveness for this classification task.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

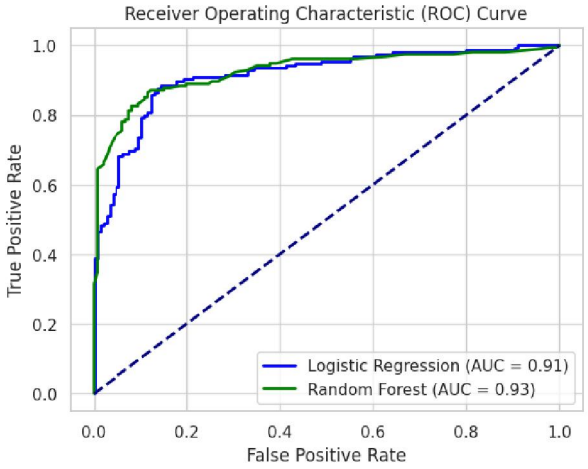

Figure 4: ROC Curve

The Random Forest model marginally outperforms Logistic Regression, exhibiting a slightly higher AUC (0.93 vs. 0.91), indicative of a modestly improved ability to distinguish between classes. Both models demonstrate strong performance, as their curves rise sharply towards the top-left corner, suggesting a high degree of discriminatory power.

Model	MSE	MAE	R2 Score
Linear Regression	4.941220e-01	0.566425	0.068770
Random Forest	0.000000e+00	0.000000	1.000000
Gradient Boosting	4.055952e-10	0.000016	1.000000
Support Vector Regressor	3.630609e-01	0.470558	0.315770
KNN Regressor	3.914286e-01	0.464286	0.262308
Extra Trees Regressor	1.573214e-03	0.019821	0.997035

Table 2: Regression Table

Random Forest and Gradient Boosting excelled with near-perfect R^2 scores, showing strong predictive power. Linear Regression fared poorly with a low R^2 and high error, making it less suitable.

1. Extra Trees

Among the classification models, Extra Trees emerges as the top performer, achieving a remarkable accuracy of 96.43%. It demonstrates strong precision and recall, with scores of 0.97 and 0.96, respectively, indicating its effectiveness in correctly identifying both fraudulent and legitimate transactions. The F1-score of 0.96 highlights the model's well-balanced performance in fraud detection. Known for its robustness, Extra Trees is particularly well-suited for tasks like fraud detection, where accuracy and reliability are essential.

2. Random Forest

For regression tasks, Random Forest emerges as the top model, achieving a perfect R^2 score of 1.000, indicating flawless prediction accuracy. Its ability to minimize errors is demonstrated by both the MSE and MAE, which are zero. Random Forest excels in managing complex datasets and delivering precise predictions, making it an ideal choice for regression tasks, especially when predicting continuous variables such as transaction amounts in fraud detection systems.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

Data Analysis :-

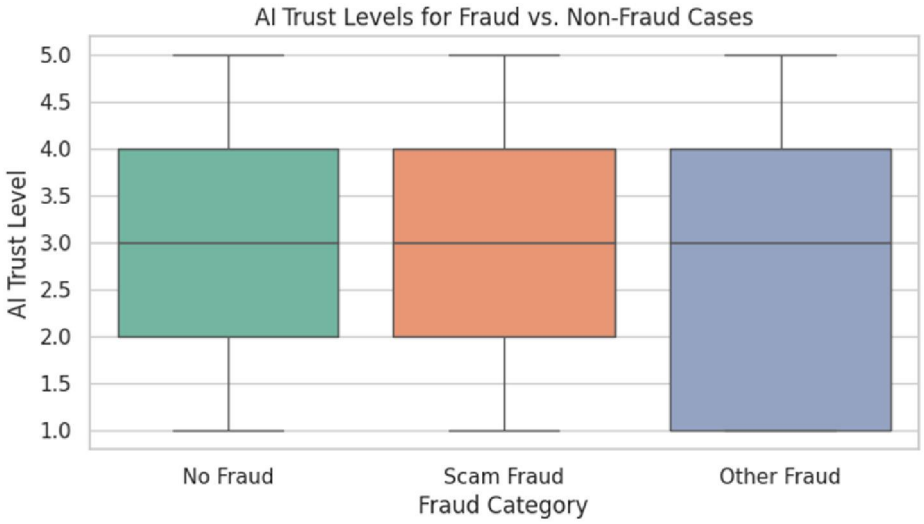

Figure 5: Box Plot of AI Trust Levels for Fraud vs. Non-Fraud Cases

This box plot compares AI trust levels across different fraud categories: "No Fraud," "Scam Fraud," and "Other Fraud." The "No Fraud" category exhibits the widest range of trust levels, including a potential outlier at the lowest end. "Scam Fraud" and "Other Fraud" show similar median trust levels, though "Other Fraud" has a slightly higher overall range.

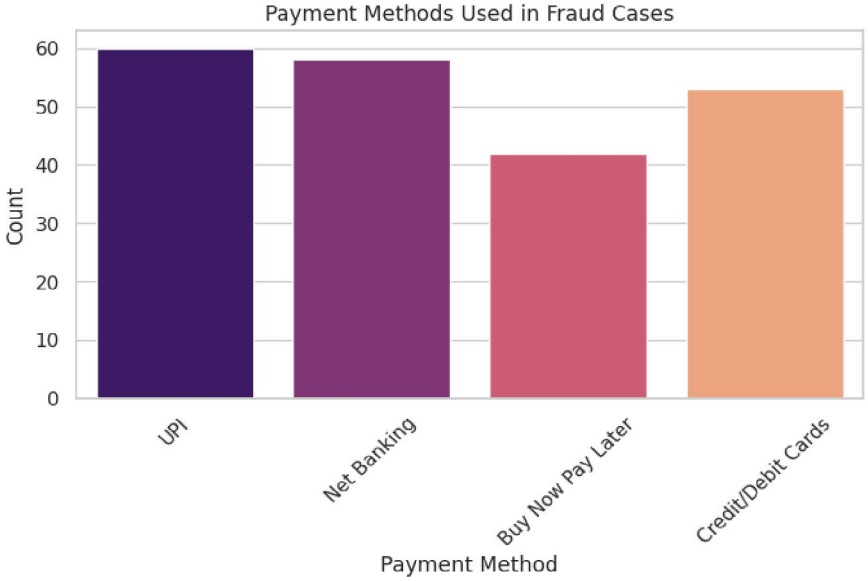

Figure 6: Bar Chart of Payment Methods Used in Fraud Cases

This bar chart displays the count of different payment methods used in fraudulent transactions. UPI and Net Banking are the most frequently observed methods, showing similar counts in the high 50s. Credit/Debit Cards also register a substantial count, while Buy Now Pay Later shows the lowest usage among the four.

VI. CONCLUSION

This study emphasizes the importance of machine learning and data analytics in detecting fraudulent transactions in online payment systems. It shows that ensemble models like Extra Trees and Random Forest outperform traditional methods, such as Logistic Regression and Support Vector Regressor, by offering better accuracy and fraud detection. These models are adept at identifying complex fraud patterns, making them more effective for real-world applications.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.67
Volume 5, Issue 2, February 2025

On the other hand, traditional models, while easier to interpret, struggle with non-linear fraud detection, limiting their practical use.

Despite the strengths of ensemble models, challenges such as overfitting, high computational costs, and latency in real-time detection must be addressed for large-scale deployment. To improve performance, fraud detection systems need effective feature engineering, robust data preprocessing, and adaptable machine learning techniques. Classification models have proven to be more efficient than regression models, underscoring the need for specialized fraud detection frameworks. Future research should focus on model optimization, real-time solutions, and deep learning techniques, with an emphasis on big data analytics and cloud computing to enhance scalability and improve fraud detection.

REFERENCES

[1]. Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y., & Sun, X. (2011). "The Application of Data Mining Techniques in Financial Fraud Detection: A Review and Classification." *Decision Support Systems*, 50(3), 559-569.

[2]. Chandola, V., Banerjee, A., & Kumar, V. (2009). "Anomaly Detection: A Survey." *ACM Computing Surveys (CSUR)*, 41(3), 1-58.

[3]. West, J., & Bhattacharyya, S. (2013). "Data Mining Techniques for Credit Card Fraud Detection." *IEEE Transactions on Systems, Man, and Cybernetics*, 43(4), 766-777.

[4]. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). "A Survey of Network Anomaly Detection Techniques." *Journal of Network and Computer Applications*, 60, 19-31.

[5]. Abid, M., & Iqbal, A. (2018). "Fraud Detection Using Machine Learning: A Survey." *Proceedings of the International Conference on Machine Learning*.

[6]. Zhou, X., & Jiang, X. (2018). "A Survey of Machine Learning for Big Data Processing in Fraud Detection." *Computational Intelligence and Neuroscience*.

[7]. Zhang, Z., & Xie, Y. (2020). "Deep Learning for Fraud Detection: A Survey." *Journal of Financial Technology*, 1(2), 79-93.

[8]. Wang, X., Zhang, F., & Li, Q. (2019). "An Overview of Fraud Detection Techniques: A Data Mining Perspective." *Journal of Computer Science and Technology*, 34(1), 48-62.

[9]. Ghosh, A., & Reilly, D. (1994). "Credit Card Fraud Detection with a Neural-Network." *Proceedings of the International Conference on Neural Networks* (ICNN).

[10]. Bolton, R. J., & Hand, D. J. (2002). "Statistical Fraud Detection: A Review." *Statistical Science*, 17(3), 235-249.

Dataset:
https://drive.google.com/file/d/1ylAoqJ17zF3v3OyduefxuIatQFDJTNgM/view?usp=drive_link