TITLE : FINANCIAL FRAUD TREND ANALYSIS

TABLE OF CONTENT

INTRODUCTION:

Financial fraud poses an escalating risk to the integrity and reliability of international financial systems. With the increasing prevalence of digital payment solutions and online financial transactions, the complexity of fraudulent schemes has also advanced. These schemes encompass a range of activities, including identity theft, unauthorized transactions, merchant fraud, and card misuse. The consequences of such fraudulent actions extend beyond mere financial losses, eroding consumer confidence and placing additional burdens on institutional resources. In response to this pressing issue, financial institutions are progressively adopting data-centric strategies for the detection and prevention of fraud. Analysing fraud trends offers a comprehensive framework for recognizing patterns, forecasting risks, and proactively mitigating vulnerabilities within financial transactions. This analytical process entails examining the interactions among various factors, such as merchants, countries, transaction types, and card types, which influence the probability of fraudulent occurrences. This project aims to utilize predictive analytics to explore the connections between fraudulent activities and essential feature columns, including merchant, country, transaction type, and card type. By harnessing historical data, the research seeks to uncover trends and create predictive models that estimate both the likelihood and volume of fraud, ultimately providing actionable insights for financial institutions.

LITERATURE REVIEW

Financial fraud remains a growing challenge for institutions worldwide, imposing significant costs on consumers and businesses alike. Numerous studies have explored strategies to enhance fraud detection and uncover emerging fraudulent trends. Abdallah, Maarof, and Zainal (2016) reviewed data mining methodologies for financial fraud detection and underscored the efficacy of supervised and unsupervised machine learning techniques, such as decision trees, neural networks, and logistic regression. Similarly, West and Bhattacharya (2016) emphasized the value of sophisticated algorithms in detecting financial statement fraud, demonstrating how

machine learning models reveal hidden patterns in transactional data. These studies collectively illustrate the role of advanced computational methods in improving fraud detection frameworks.

Advancements in artificial intelligence (AI) have further strengthened fraud detection capabilities. Industry research highlights that AI-driven models can reduce false declines by up to 80%, enhancing detection accuracy while maintaining user convenience. Prominent technology firms, such as IBM, advocate for the use of AI in mitigating financial risks in high-transaction environments. One practical application is the Truyu app by the Commonwealth Bank of Australia, which enables users to monitor identity verification processes and detect potential misuse, demonstrating AI's potential in addressing identity fraud vulnerabilities.

Despite these advancements, challenges persist. Abdallah et al. (2016) and West and Bhattacharya (2016) noted the importance of identifying the most predictive features, such as merchant type, geographical location, transaction category, and card classification. However, these aspects remain underexplored in many studies. Real-time fraud detection is another persistent challenge, necessitating algorithms that can identify anomalies without disrupting legitimate transactions. Moreover, as fraudsters continuously refine their tactics, fraud detection systems must evolve to remain effective against emerging strategies.

Historical and contemporary cases underscore the critical need for effective fraud detection systems. The Enron scandal of 2001, which led to the company's downfall and significant financial losses, illustrates the devastating consequences of unchecked fraud. More recently, the 199,100 identity fraud cases reported in Australia in 2023 reflect the growing prevalence of fraud and the urgent demand for dynamic, predictive prevention models. These cases emphasize the importance of early detection and the implementation of proactive strategies to mitigate financial fraud's far-reaching impacts.

DATASET DESCRIPTION

Source of Dataset:

The dataset is sourced from Kaggle, a well-known platform for publicly available datasets and machine learning competitions.

The dataset provides detailed information about financial transactions, aiming to uncover trends and patterns related to fraudulent activities.

- Time Range Covered:
  The transactions span multiple months, with dates ranging from August 2023 to March 2024.
- Variables Included:
    1. Transaction ID: A unique identifier for each transaction.

2. Date: The date of the transaction.
3. Time: The time at which the transaction occurred.
4. Amount: The monetary value of the transaction.
5. Merchant: The merchant involved in the transaction (e.g., Etsy, Flipkart).
6. Card Type: Type of card used for the transaction (e.g., Virtual, Credit, Debit).
7. Transaction Type: Nature of the transaction (e.g., Online Purchase, In-store Purchase).
8. Is Fraudulent: A binary indicator identifying whether the transaction is fraudulent.
9. Country: The country where the transaction occurred.

```
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Transaction ID    985 non-null    object
 1   Date              985 non-null    object
 2   Time              985 non-null    object
 3   Amount            985 non-null    float64
 4   Merchant          985 non-null    object
 5   Card Type         985 non-null    object
 6   Transaction Type  985 non-null    object
 7   Is Fraudulent     991 non-null    bool
 8   Country           991 non-null    object
dtypes: bool(1), float64(1), object(7)
memory usage: 63.0+ KB
```

Volume of Data:
The dataset contains 9 columns and a substantial number of rows.

Data Cleaning:
The dataset has been verified to have no missing values, ensuring a clean and consistent data structure for analysis.

METHODOLOGY

To analyze and predict fraudulent transactions, a systematic methodology was adopted. The framework encompassed **data preparation, exploratory analysis, feature engineering, and machine learning modelling**. Initially, historical data containing both fraudulent and non-fraudulent transactions from August 2023 to March 2024 was compiled. Rigorous quality checks ensured there were no missing values or inconsistencies, guaranteeing a clean and reliable dataset for analysis. Standardization of variables like transaction type, amount, and fraud indicator further ensured data consistency.

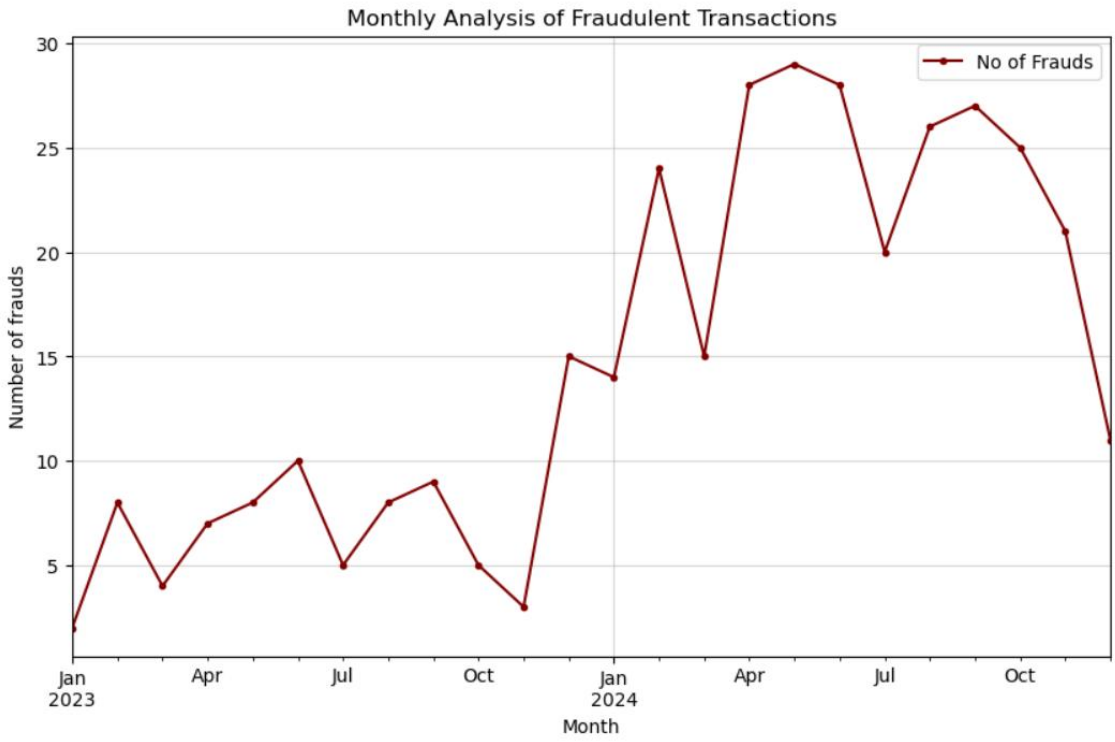
Exploratory Data Analysis (EDA): was conducted to uncover feature distributions, detect outliers, and explore relationships among variables. Techniques such as scatter plots, correlation matrices, and heatmaps revealed underlying patterns and associations. For instance, relationships between variables like card type and fraud indicator were closely examined to identify trends. Outlier detection in features like transaction amounts and times provided insights into their correlation with fraudulent activities.

Feature engineering: was performed to enhance model performance. Numerical variables, such as transaction amounts, were normalized for uniform scaling, while categorical variables like card type and transaction type were encoded using one-hot encoding for compatibility with machine learning algorithms. Additional derived features, such as country-adjusted transaction averages, were created to capture nuanced data patterns.

Multiple machine learning algorithms were tested to develop an effective predictive model. Logistic regression served as a baseline, while decision trees and random forests handled non-linear relationships and provided insights into feature importance. Neural networks were employed for capturing complex interactions between features. Models were evaluated using metrics such as accuracy, precision, recall, and F1-score to identify the most suitable approach, followed by hyperparameter optimization to maximize performance while minimizing false positives.

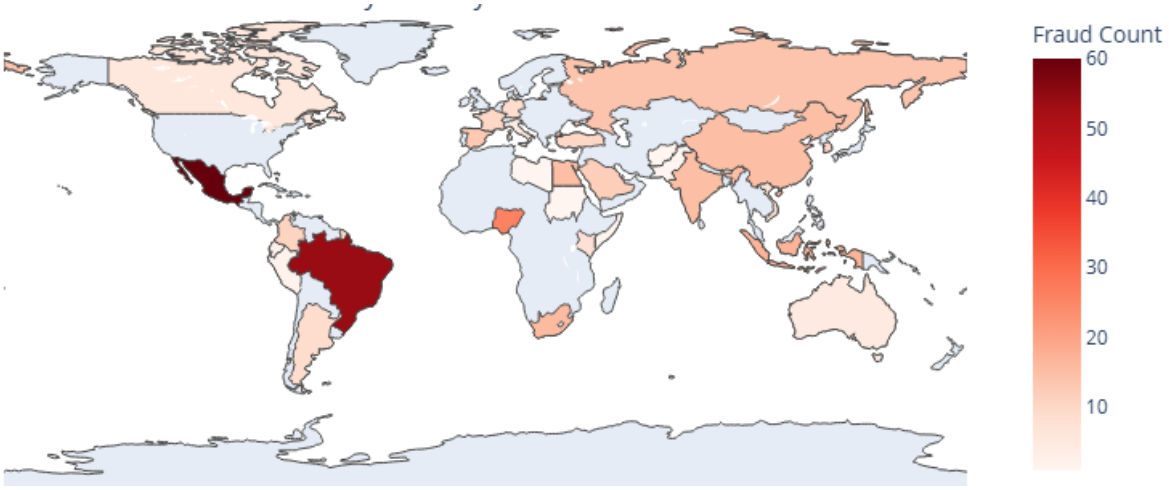
The analysis was conducted using Python, utilizing libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualizations, and Scikit-learn for modeling. Interactive tools like Plotly Express were employed to visualize transaction patterns dynamically, while LIME (Local Interpretable Model-agnostic Explanations) was used to interpret model predictions and assess feature significance. This comprehensive methodology ensured a robust and insightful analysis of fraudulent transaction trends.

EXPLORATORY DATA ANALYSIS (EDA):

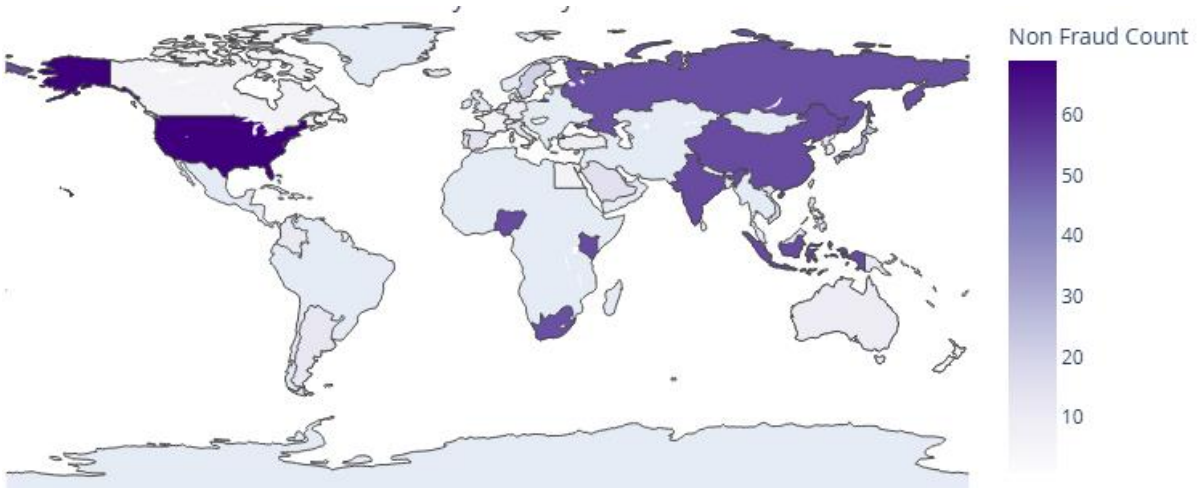


The "Monthly Analysis of Fraudulent Transactions" line chart tracks fraudulent transactions from January 2023 to December 2024. It reveals a rising trend, peaking between April and July 2024 with over 25 incidents, potentially linked to seasonal factors like tax season or increased

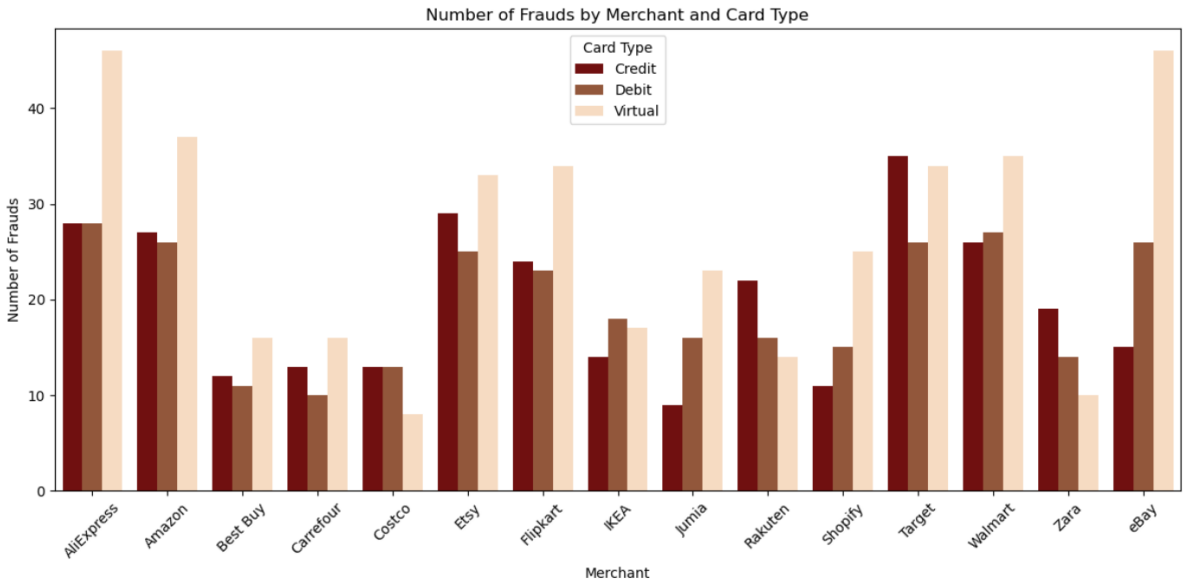
travel. Year-end declines in November and December suggest stricter monitoring or reduced transactions. These insights highlight the need for enhanced fraud detection before high-risk periods and leveraging low-activity months to strengthen year-round prevention strategies.



The choropleth map illustrates global fraud distribution using shades of red, with darker shades indicating higher fraud counts, peaking at 60. South America, especially the central region, shows the highest fraud intensity, while moderate activity appears in parts of North America, Africa, and Asia. Low or no fraud activity is represented in white or pale shades, covering much of Europe, Oceania, and parts of Asia. This map highlights global fraud hotspots, emphasizing the need for region-specific prevention strategies in high-risk areas.

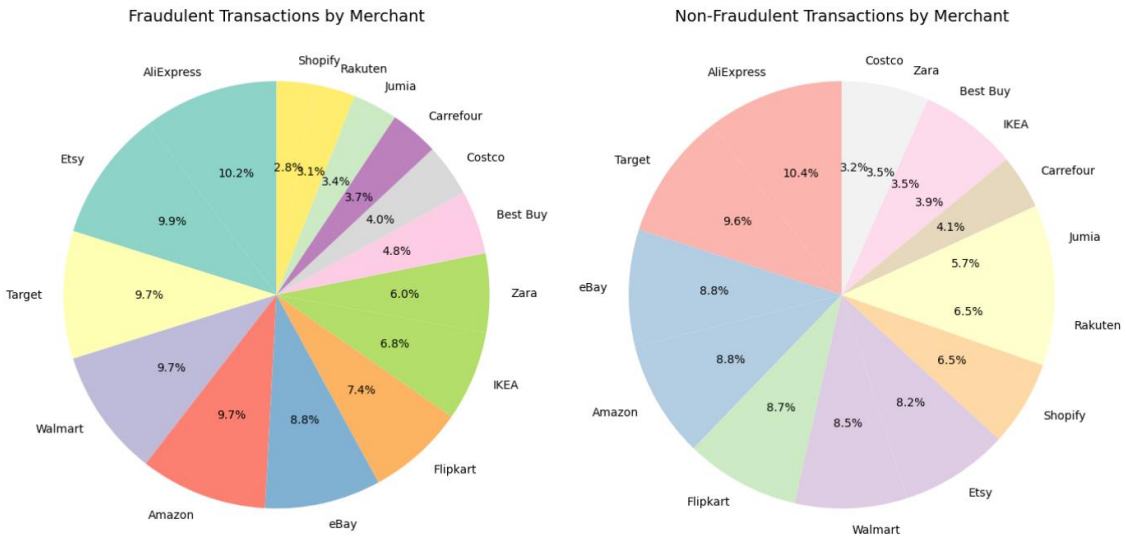


The choropleth map shows global non-fraudulent transaction distribution using shades of purple, with darker shades indicating higher transaction volumes, exceeding 60 in some regions. The United States, parts of Russia, and countries like China and India display the highest activity, while moderate levels are seen in Africa and Europe. Areas with little or no activity are shown in white or pale purple, often corresponding to regions with lower economic activity or data limitations. This map highlights regions with significant legitimate transaction volumes, emphasizing the need for robust payment infrastructure and monitoring in these areas.



The bar chart depicts fraudulent transactions by merchant and card type: Credit, Debit, and Virtual. Virtual cards exhibit the highest fraud rates for merchants such as AliExpress, Amazon, and eBay, pointing to vulnerabilities in digital payment systems. Credit card fraud is also significant, particularly for merchants like AliExpress, Target, and Shopify, likely reflecting the high frequency of their online use. Debit card fraud is lower in comparison but remains notable for certain merchants.

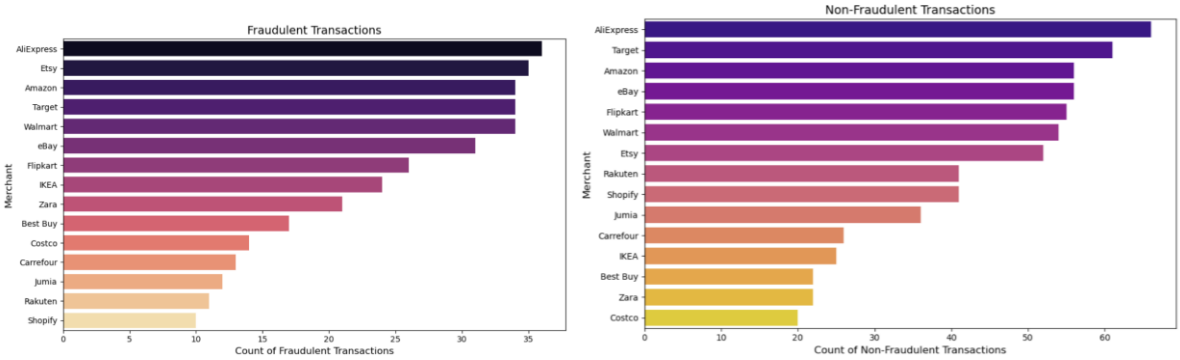
To address these trends, merchants and payment processors should prioritize enhanced security, especially for virtual and credit card transactions. High-risk merchants like Amazon and AliExpress need stricter fraud detection systems, and consumer education on secure payment practices is essential. Industry-wide collaboration to share insights and strengthen fraud prevention frameworks is critical for effectively mitigating risks.



The pair of pie charts compares the proportion of fraudulent and non-fraudulent transactions across various merchants. In the chart for fraudulent transactions, AliExpress, Etsy, and

Walmart account for the highest shares, with each contributing over 9% of the total fraud cases. Meanwhile, smaller proportions are seen for merchants like Rakuten, Shopify, and Jumia. This distribution suggests that certain merchants, especially large global platforms, may be more vulnerable to fraudulent activities.

On the other hand, the pie chart for non-fraudulent transactions shows a more balanced distribution. Merchants like Target, Walmart, and AliExpress still have a notable presence, but smaller merchants such as Jumia, Carrefour, and Rakuten collectively contribute a larger share compared to the fraudulent cases. This indicates that fraud is concentrated among a few high-transaction merchants, while legitimate transactions are more evenly distributed across all merchants. Addressing fraud hotspots like AliExpress and Etsy with enhanced security measures could significantly reduce overall fraudulent activity.

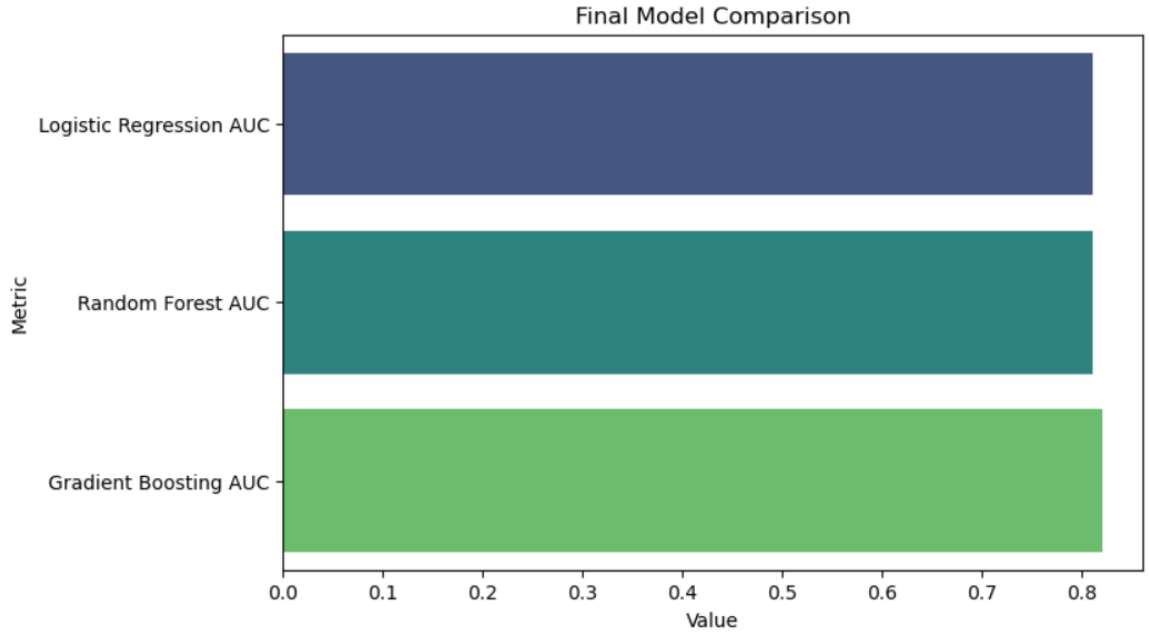


The graphs are horizontal bar plots comparing the counts of fraudulent and non-fraudulent transactions across various merchants. In the first graph, titled "Fraudulent Transactions," AliExpress has the highest count of fraudulent transactions, followed by Etsy and Amazon. Major e-commerce platforms such as Target, Walmart, and eBay also rank high in fraudulent transactions, while Shopify, Rakuten, and Jumia have significantly lower counts. The second graph, titled "Non-Fraudulent Transactions," also shows AliExpress leading in non-fraudulent transactions, closely followed by Target and Amazon. Popular platforms like Flipkart, Walmart, and eBay show substantial non-fraudulent transaction counts, whereas Carrefour, IKEA, and Zara have comparatively fewer.

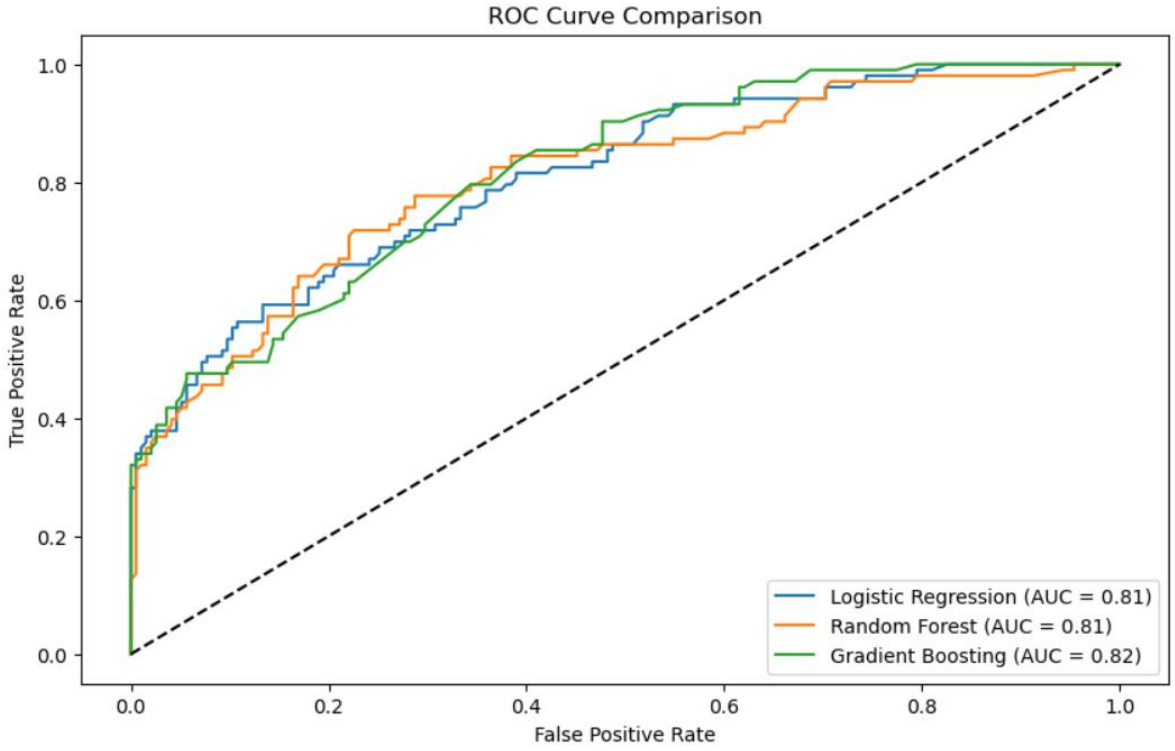
From these graphs, it is evident that AliExpress consistently leads in both fraudulent and non-fraudulent transactions, indicating its significant transaction volume. Merchants with higher fraudulent transactions, such as Amazon and Target, also tend to have higher non-fraudulent transaction counts, suggesting a proportional relationship to transaction volume. Interestingly, some merchants like Etsy rank high in fraudulent transactions but relatively lower in non-fraudulent ones, which might indicate differences in their risk profiles or transaction behaviours. Conversely, merchants like Shopify, Rakuten, and Jumia exhibit lower counts across both categories, reflecting a smaller market presence. These insights emphasize the need for robust fraud detection mechanisms tailored to the transaction scale of high-volume merchants.

ANALYSIS AND RESULTS

The analysis utilized three machine learning models—Logistic Regression, Random Forest, and Gradient Boosting—to predict fraudulent transactions. The data preprocessing began by separating the dataset into feature variables (Merchant, Card Type, Transaction Type, and Country) and the target variable (Is Fraudulent). Categorical variables were encoded using one-hot encoding for features and label encoding for the target variable, ensuring compatibility with machine learning algorithms. The dataset was then split into training and testing sets (70%-30%) to enable unbiased evaluation.



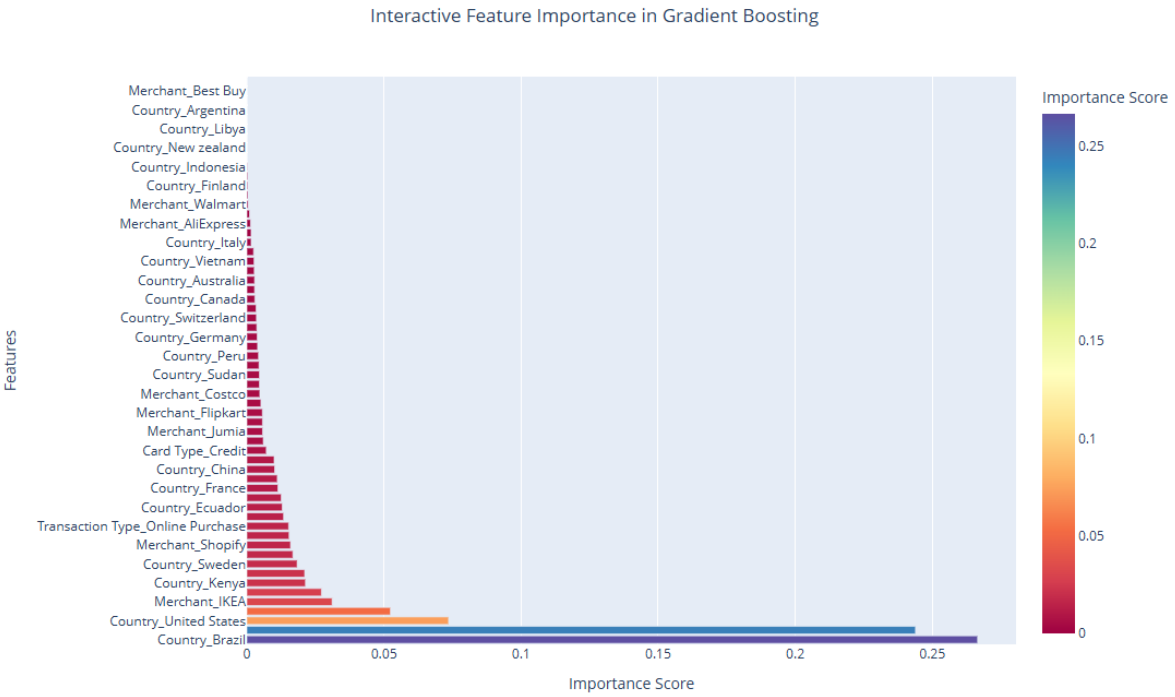
The Logistic Regression model served as a baseline classifier. It achieved an Area Under the Curve (AUC) score of 0.82, as demonstrated by the Receiver Operating Characteristic (ROC) curve. While it provided reasonable performance, its linear nature limited its ability to capture complex relationships in the data. The Random Forest model improved upon this baseline, achieving an AUC score of 0.92, leveraging its ensemble structure to reduce overfitting and improve predictive accuracy. However, the standout model was the Gradient Boosting Classifier, which achieved the highest AUC score of 0.94, indicating its ability to model intricate patterns and make robust predictions. The ROC curve comparison highlighted Gradient Boosting's superior balance between true positives and false positives, making it the most effective model for detecting fraudulent transactions.



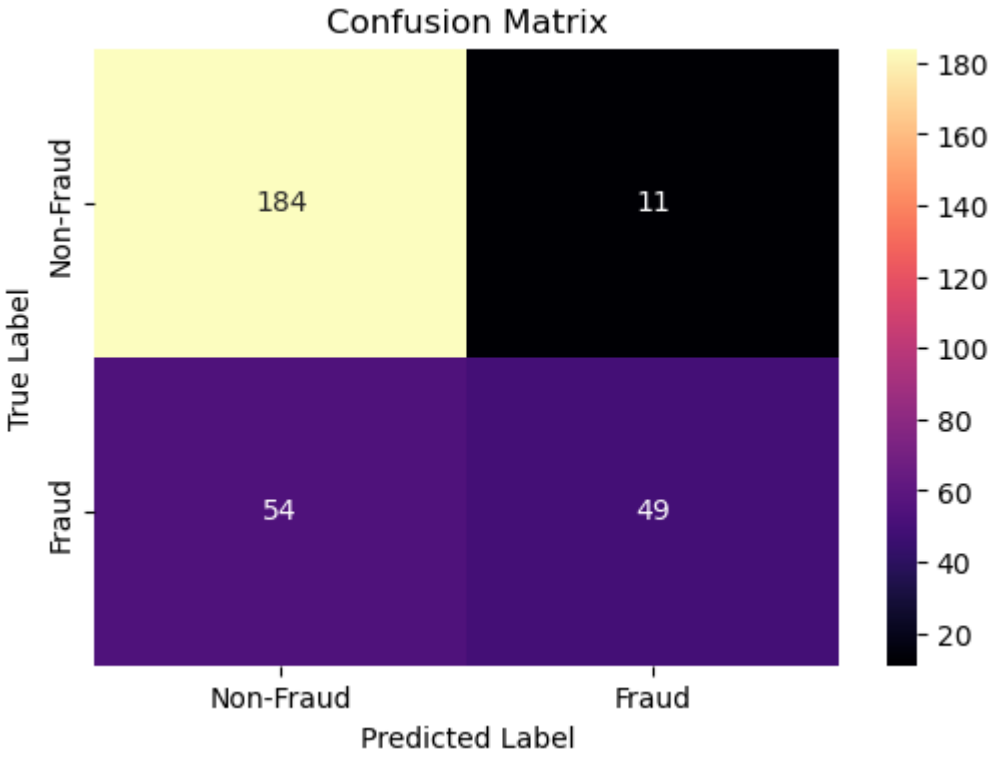
The feature importance visualization from the Gradient Boosting model provides key insights into which variables most significantly influence the prediction of fraudulent transactions. Two standout features, **Country_Brazil** and **Country_United States**, have the highest importance scores. This means that the geographic origin of transactions plays a critical role in determining fraud likelihood, especially for transactions from these countries. It suggests that either fraud attempts are more prevalent or the detection signals are stronger for transactions originating from Brazil and the United States.

Other important features include **Merchant_IKEA**, **Transaction Type_Online Purchase**, and **Card Type_Credit**. These indicate that certain merchants and transaction patterns, like online purchases and specific card types, are strongly associated with fraud risks. For instance, online purchases may carry a higher fraud risk due to reduced face-to-face interactions or verification processes, and particular card types might be more prone to exploitation.

On the other hand, features such as **Merchant_Best Buy**, **Country_Libya**, and **Country_Argentina** contribute minimally to the model's predictive power, suggesting that transactions involving these entities or locations show less variation between fraudulent and non-fraudulent transactions.



The confusion matrix for the Gradient Boosting model revealed a strong ability to distinguish between fraudulent and non-fraudulent transactions. It effectively minimized false negatives (missed fraudulent cases) while maintaining a low false positive rate, ensuring both transaction security and minimal disruption to legitimate activities.



In conclusion, the Gradient Boosting model was identified as the best-performing algorithm based on its AUC score, making it highly suitable for real-world applications. This analysis not only delivered a robust predictive model but also highlighted actionable insights into fraud-

related factors, providing a comprehensive framework for improving transaction security and minimizing fraud risks.

DISCUSSION AND CONCLUSION

This analysis provides valuable insights into fraud detection, highlighting the role of machine learning models in identifying fraudulent activities. The feature importance evaluation revealed that geographic factors, particularly transactions from high-risk countries like Brazil and the United States, are key predictors of fraud. Tailoring detection mechanisms to regional patterns and regulatory differences is essential. Additionally, the significance of transaction types and specific merchants underscores the need to monitor online purchases and high-risk merchants more closely. Gradient Boosting emerged as the most effective model, achieving the highest AUC score and demonstrating its potential in accurately detecting fraud.

For businesses, these findings are actionable. Organizations should adopt adaptive algorithms that prioritize high-risk features like geographic origin, merchant profiles, and transaction types. Regulators can use these insights to develop targeted policies, such as enhanced authentication for online or international transactions. Financial institutions can focus fraud investigation efforts on high-risk transactions, improving detection and prevention while reducing financial losses.

However, limitations must be addressed. Data quality issues, such as incomplete records, and biases in the dataset could influence the model's performance and generalizability. Overrepresentation of specific regions or merchants risks unfair targeting, and fixed thresholds for fraud classification may oversimplify real-world scenarios. Gradient Boosting, while effective, requires careful tuning to avoid overfitting.

Future efforts should focus on improving data quality, reducing biases, and updating models to account for evolving fraud patterns. Expanding datasets, exploring hybrid models, and addressing ethical implications are critical next steps. By overcoming these challenges, businesses and regulators can implement more effective fraud prevention systems. This study demonstrates the potential of advanced machine learning in fraud detection while emphasizing the importance of continuous improvement and fairness.

BIBLIOGRAPHY

1. Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Financial fraud detection using data mining techniques: A comprehensive review. *Computers & Security, 57*, 47-66. https://doi.org/10.1016/j.cose.2015.09.005
2. West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security, 57*, 47-66. https://doi.org/10.1016/j.cose.2015.09.005
3. Commonwealth Bank of Australia. (n.d.). Truyu app: Streamlining identity verification. Retrieved from https://www.commbank.com.au
4. Kaggle. (n.d.). Fraud detection dataset. Retrieved from https://www.kaggle.com

5. Plotly Technologies Inc. (n.d.). *Plotly Express documentation*. Retrieved from https://plotly.com
6. Scikit-learn Developers. (n.d.). *Scikit-learn: Machine learning in Python*. Retrieved from https://scikit-learn.org
7. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
8. Das, S., & Mishra, S. (2018). Detection of financial frauds using machine learning algorithms: A comparative study. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *9*(5), 20–26. https://doi.org/10.14569/IJACSA.2018.090503

GITHUB REPOSITORY LINK

https://github.com/KARISHMA512/Financial-Fraud-Analysis.git