# Impact of Transaction Characteristics on Fraud Detection

Edith Tache, Cali Hugelen, Jacob Brooks, Remington Webb

Summer, 2024

# TABLE OF CONTENTS

# Introduction 3-4

# Business Problem/Hypothesis 4-5

# Methods/Analysis 5-9

# Results 9-21

# Recommendations/Ethical Considerations 22-24

Limitations and Future Works 25-28

# Conclusion 29-31

# References 32

# Appendix 33-35

# Introduction

# In today’s digital age, fraud detection has become an essential aspect of financial transactions. With the increasing volume of transactions, it's crucial to develop robust methods for identifying fraudulent activities to protect both consumers and businesses. This project focuses on understanding the impact of transaction characteristics on fraud detection by leveraging machine learning techniques. By analyzing various transaction features, we aim to develop a predictive model that can accurately distinguish between fraudulent and non-fraudulent transactions. This will help financial institutions minimize losses and enhance the security of their transaction systems.

**Dataset Sourcing**

The dataset being examined is from Kaggle and is titled "Credit Card Fraud Transaction Data." The dataset includes various features that influence the likelihood of a transaction being fraudulent.

**Link to Dataset:** [Credit Card Fraud Transaction Data (kaggle.com)](https://www.kaggle.com/datasets/anurag629/credit-card-fraud-transaction-data)

**Description:**

* **Transaction ID:** Unique identifier for tracking transactions.
* **Date:** Date of the transaction.
* **Day of Week:** Day of the week when the transaction occurred.
* **Time:** Time of the transaction.
* **Type of Card:** Type of credit card used (e.g., Visa, MasterCard).
* **Entry Mode:** Mode of entry for the transaction (e.g., Tap, PIN, CVC).
* **Amount:** Transaction amount.
* **Type of Transaction:** Nature of the transaction (e.g., POS, Online, ATM).
* **Merchant Group:** Category of the merchant.
* **Country of Transaction:** Country where the transaction occurred.
* **Shipping Address:** Shipping address provided for the transaction.
* **Country of Residence:** Country of the cardholder's residence.
* **Gender:** Gender of the cardholder.
* **Age:** Age of the cardholder.
* **Bank:** Bank that issued the card.
* **Fraud:** Indicator of whether the transaction was fraudulent (1) or not (0).

# Business Problem/Hypothesis

The core of our project revolves around several fundamental research questions aimed at addressing the business problem of detecting fraudulent transactions through transaction analytics. Specifically, we seek to understand:

1. **Which transaction characteristics are the most indicative of fraudulent activity?**
2. **How do different entry modes (e.g., Tap, PIN, CVC) affect the likelihood of a transaction being fraudulent?**
3. **What role do demographic factors (e.g., age, gender, country of residence) play in predicting fraud?**
4. **Can a predictive model accurately predict fraud?**

By focusing on these questions, we aim to improve the accuracy and efficiency of fraud detection systems. Additionally, we seek to build a model that will help financial institutions effectively predict and prevent fraudulent transactions, thereby minimizing losses and enhancing transaction security.

**Hypothesis**

Based on our preliminary research and analysis of the dataset, we hypothesize that:

1. Transaction characteristics can be analyzed to determine the likelihood of fraud.
2. A well-developed predictive model can accurately distinguish between fraudulent and non-fraudulent transactions.

These hypotheses will guide our analysis and model development as we identify the critical factors influencing fraud detection and provide actionable recommendations for improving transaction security.

# Methods/Analysis

**Data Preparation**

The dataset was prepared by following these steps:

1. **Data Cleaning**: Removed leading and trailing whitespaces from column names, converted relevant columns to appropriate data types (e.g., numeric, datetime), and filled missing values with the mean for numerical columns and mode for categorical columns.
2. **Data Quality Assessment:**
   * A check for remaining missing values was conducted.
   * The number of duplicate rows was identified as 0.
   * Summary statistics of the dataset were displayed to provide an overview of the data distribution.

**Exploratory Data Analysis (EDA)**

EDA was conducted to understand the distributions and relationships within the data. Key visualizations included:

1. **Distribution of Fraudulent Vs Non-Fraudulent Transactions:** A bar chart was plotted to highlight the highly imbalanced distribution between fraudulent and non-fraudulent transactions. This imbalance underscores the necessity of employing specialized techniques in model training to ensure the effective identification of the minority class (fraudulent transactions).

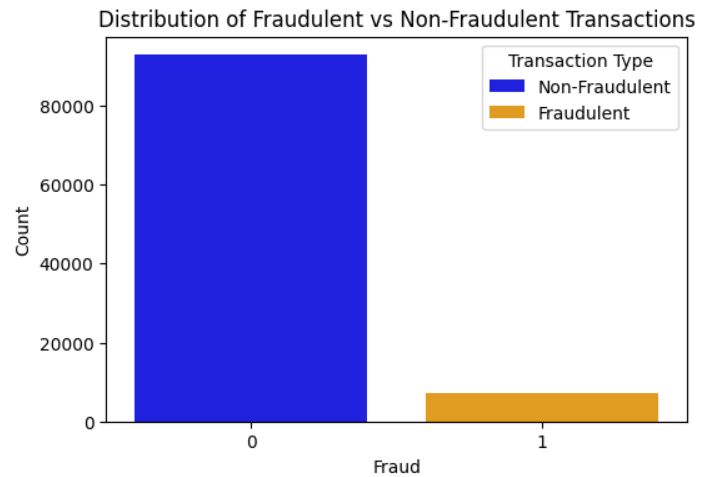


Figure 1 Distribution of Fraudulent Vs Non-Fraudulent Transactions

1. **Distribution of Fraud Across Transaction Characteristics:** Plots were created for various transaction characteristics such as type of card, entry mode, type of transaction, merchant group, country of transaction, and shipping address to visualize fraud distribution across these categories. This analysis helped identify which transaction characteristics were most indicative of fraudulent activity.
2. **Distribution of Fraud Across Demographic Factors:** Histograms were plotted to show fraud distribution across different age groups, genders, and countries of residence. These analyses provided valuable insights into the key characteristics and patterns associated with fraudulent transactions.

These insights guided the subsequent steps in model development and evaluation.

**Feature Engineering**

To prepare the data for machine learning models, several steps were taken:

* + Dropped unnecessary columns such as 'Transaction ID' and 'Date'.
  + Encoded categorical variables using one-hot encoding to make them suitable for machine learning algorithms.
  + Normalized numerical features like 'Amount' and 'Age' to standardize their scale.
  + Conducted correlation analysis to identify relationships between features and the target variable 'Fraud'.

A screenshot of a graph

Description automatically generatedFigure 2 Correlation of Features with Target Column (Fraud)

By conducting this feature engineering and correlation analysis, we identified the critical factors influencing fraud detection, guiding the subsequent model development and evaluation steps.

**Model Training and Evaluation**

**Model Selection**

We selected two machine learning models known for their effectiveness in handling imbalanced datasets, particularly with class weight adjustment:

1. **Logistic Regression**: Chosen for its simplicity and interpretability, logistic regression is suitable for binary classification tasks. Class weight adjustment helps manage the imbalanced nature of the dataset, ensuring the model effectively identifies fraudulent transactions.
2. **Random Forest**: This model is known for handling complex interactions between features and robustness against overfitting. Random Forest can manage imbalanced datasets through class weight adjustment and provides valuable feature importance scores, which help understand the impact of different variables on the prediction outcomes.

**Model Training**

Both Logistic Regression and Random Forest models were trained using the prepared dataset. The class weight adjustment was applied to handle the imbalanced data, ensuring the minority class (fraudulent transactions) was effectively identified.

#### **Hyperparameter Tuning**

We used GridSearchCV to find the best parameters for each model by exploring various combinations of hyperparameters. Despite extensive tuning, Logistic Regression did not outperform Random Forest. Therefore, the Random Forest model was selected as the preferred model for this task.

**Evaluation Metrics**

To evaluate the models, we used the following metrics:

1. **Precision**: Measures the proportion of true positive predictions out of all positive predictions. High precision indicates a low false positive rate.
2. **Recall**: Measures the proportion of true positive predictions out of all actual positives. High recall is crucial for capturing as many fraudulent transactions as possible.
3. **F1-Score**: The harmonic mean of precision and recall, providing a balance between the two.
4. **ROC-AUC**: Measures the model's ability to distinguish between classes. A high ROC-AUC score indicates strong performance.
5. **Confusion Matrix**: Provides a breakdown of true/false positives and negatives, helping to understand the model's performance in detail.

(See Appendix for Detailed Model Evaluation Metrics scores)

# Results

**Model Performance**

The Random Forest model outperformed Logistic Regression in terms of precision, recall, F1-score, and ROC-AUC. The confusion matrix for the Random Forest model indicated a high true negative rate, low false positive rate, moderate true positive rate, and moderate false negative rate. These results demonstrate the Random Forest model's effectiveness in identifying both non-fraudulent and fraudulent transactions, making it the preferred choice for fraud detection.

**Comparative Analysis: Random Forest vs. Logistic Regression**

Despite the slight improvement in Logistic Regression’s performance after hyperparameter tuning and the Random Forest model’s performance remaining unchanged, the Random Forest model consistently outperformed Logistic Regression. Here’s a detailed justification:

**1. Superior Precision and Recall:**

* **Precision for Non-Fraudulent Transactions (Class 0):**
  + **Random Forest:** Achieved a precision of 0.98, indicating that 98% of the transactions identified as non-fraudulent were indeed non-fraudulent.
  + **Logistic Regression:** Achieved a precision of 1.00, indicating perfect identification of non-fraudulent transactions, but this came at the cost of a lower precision for fraudulent transactions.
* **Recall for Non-Fraudulent Transactions (Class 0):**
  + **Random Forest:** With a recall of 1.00, the model correctly identified all non-fraudulent transactions, minimizing false positives.
  + **Logistic Regression:** Achieved a recall of 0.94, indicating some non-fraudulent transactions were incorrectly flagged as fraudulent.
* **Precision for Fraudulent Transactions (Class 1):**
  + **Random Forest:** The model achieved a precision of 0.97, meaning 97% of the transactions flagged as fraudulent were actually fraudulent.
  + **Logistic Regression:** Achieved a precision of 0.54, indicating a higher rate of false positives than Random Forest.
* **Recall for Fraudulent Transactions (Class 1):**
  + **Random Forest:** The recall was 0.81, indicating the model successfully identified 81% of all fraudulent transactions.
  + **Logistic Regression:** Achieved a higher recall of 0.95, indicating more fraudulent transactions were detected, but with a trade-off in precision.

**2. Balanced Performance Across Metrics:**

* **F1 Score:**
  + **Random Forest:** Achieved an F1 score of 0.88 for fraudulent transactions, demonstrating a good balance between precision and recall.
  + **Logistic Regression:** Achieved an F1 score of 0.69, indicating a less balanced performance.
* **Macro and Weighted Averages:**
  + **Random Forest:** The macro and weighted averages for precision, recall, and F1 score were higher, indicating robust performance across both classes.
  + **Logistic Regression:** Although it performed well for the majority class, it struggled with the minority class, affecting its overall performance metrics.

**3. High ROC-AUC Score:**

* **Random Forest:** Achieved an ROC-AUC score of 0.9922, reflecting the model’s excellent ability to distinguish between fraudulent and non-fraudulent transactions.
* **Logistic Regression:** Achieved an ROC-AUC score of 0.9819, slightly lower than Random Forest, indicating less effective discrimination between the classes.

Despite Logistic Regression achieving a slightly higher true positive rate, the Random Forest model significantly outperforms it in terms of precision and the overall balance between true positives and false positives. The high ROC-AUC score of the Random Forest model further underscores its superior ability to discriminate between fraudulent and non-fraudulent transactions, making it the more reliable and effective model for fraud detection.

**Random Forest Model Performance Accuracy: Confusion Matrix and ROC Curve**

**Confusion Matrix**

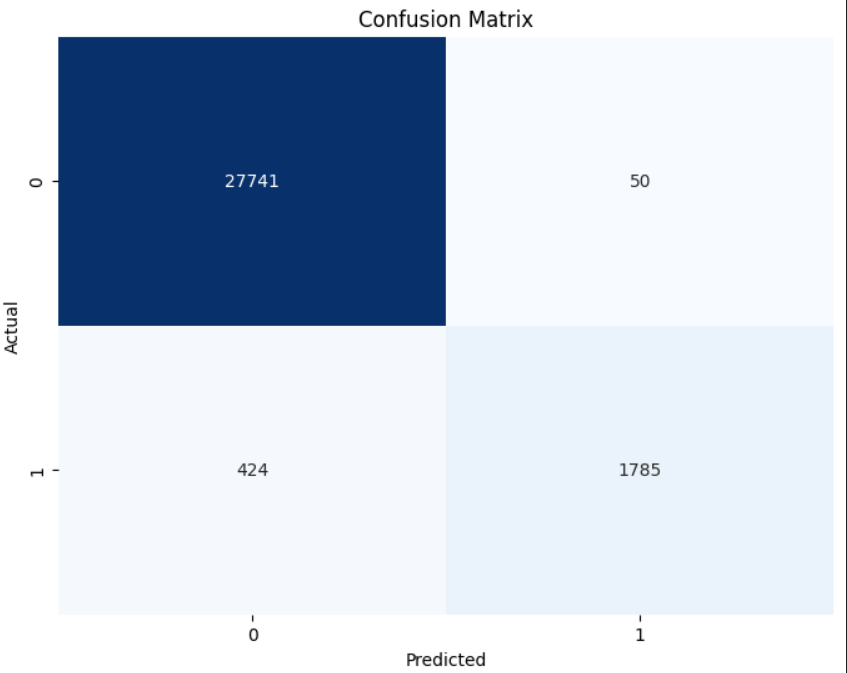
****

Figure 3 Confusion Matrix

The Confusion Matrix provides a comprehensive view of the model’s performance by showing the actual versus predicted classifications. It helps identify the number of correct and incorrect predictions made by the model.

For the Random Forest model, the Confusion Matrix is as follows:

* **True Negatives (TN):** 27,741
* **False Positives (FP):** 50
* **False Negatives (FN):** 424
* **True Positives (TP):** 1,785

**Interpretation:**

* **High True Negative Rate:** The model accurately identified 27,741 non-fraudulent transactions, indicating a solid ability to classify the majority class correctly.
* **Low False Positive Rate:** Only 50 non-fraudulent transactions were incorrectly classified as fraudulent, demonstrating high precision in identifying non-fraudulent transactions.
* **Moderate True Positive Rate:** The model correctly identified 1,785 fraudulent transactions, which is crucial for minimizing the risk of undetected fraud.
* **Moderate False Negative Rate:** There were 424 fraudulent transactions that were missed, which indicates an area for potential improvement.

The Confusion Matrix indicates that the Random Forest model effectively identifies non-fraudulent and fraudulent transactions, with a meager rate of false positives.

**ROC Curve**

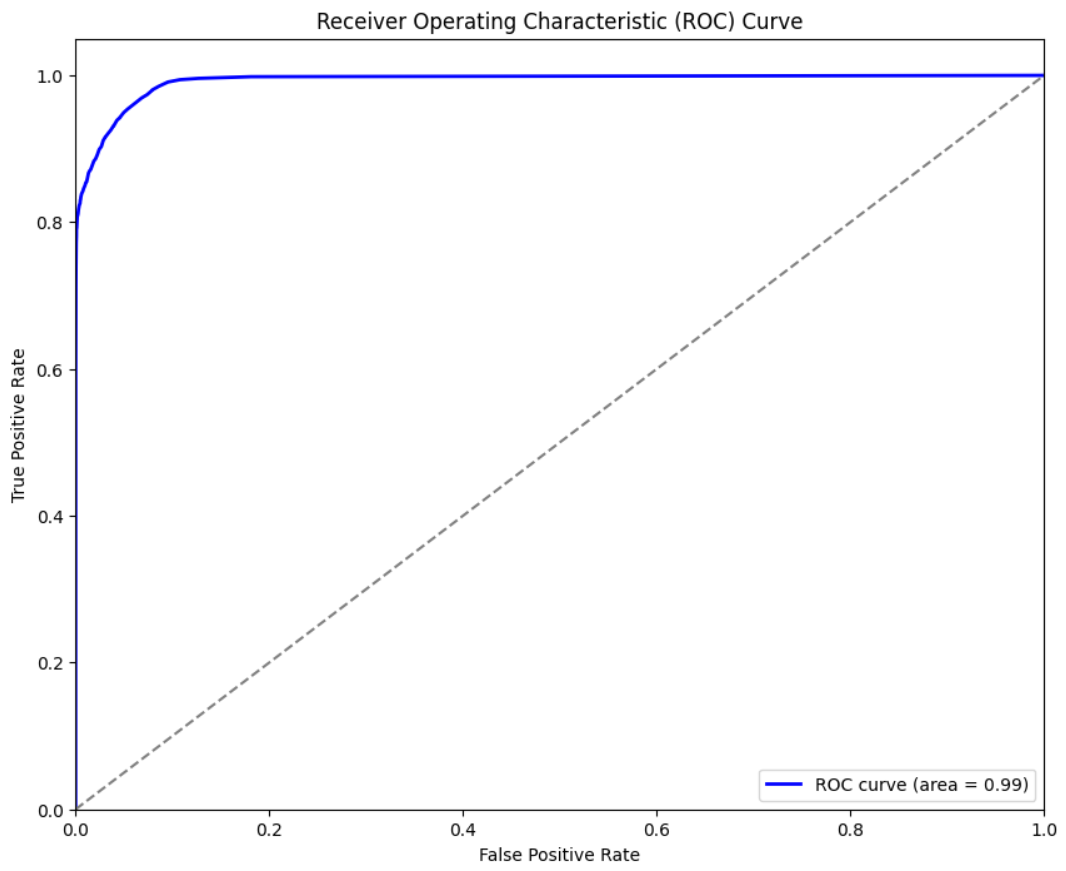
****

Figure 4: Receiver Operating Characteristic (ROC) Curve

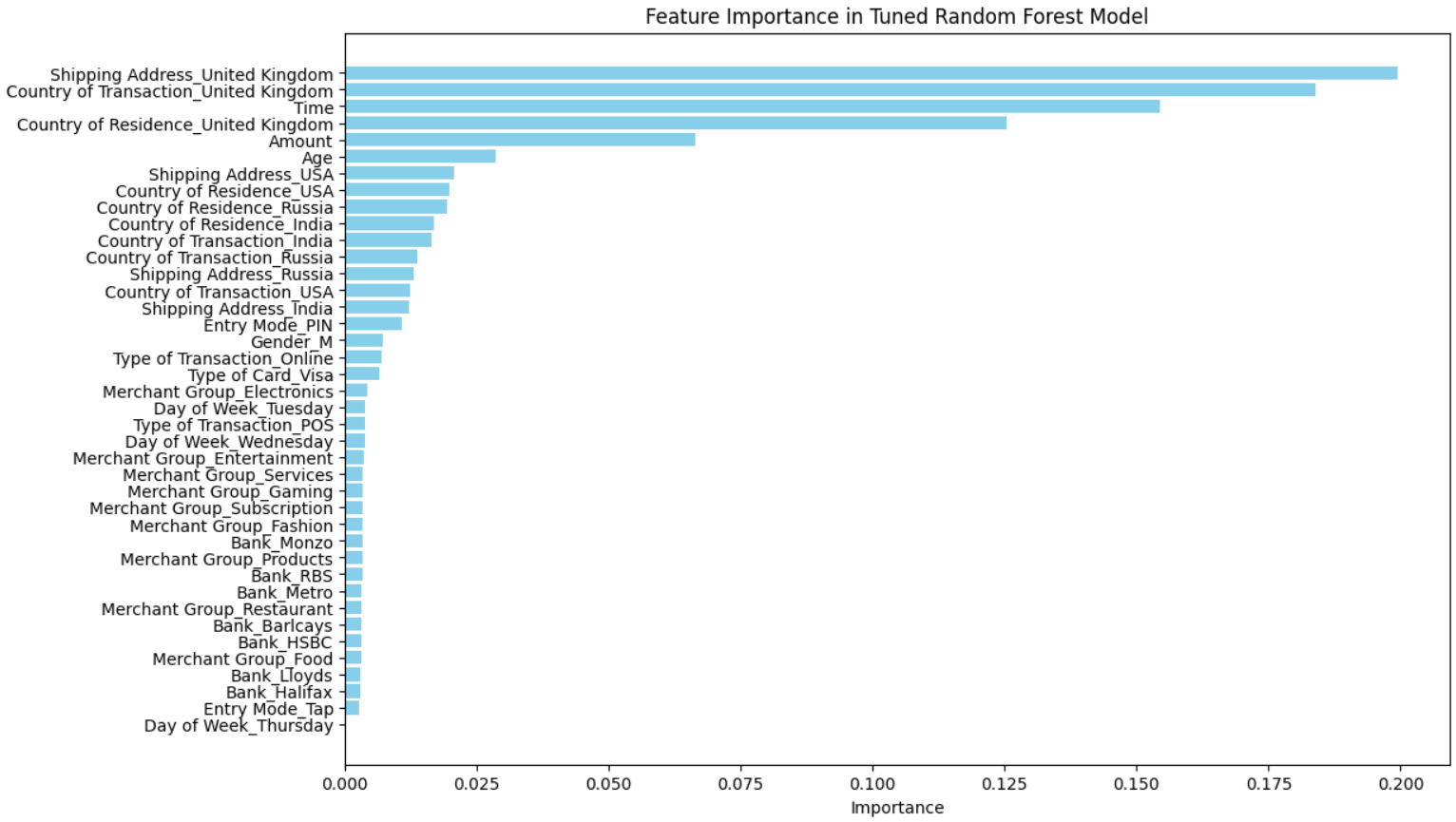
The ROC (Receiver Operating Characteristic) curve provides insight into the trade-offs between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various threshold settings.

**Interpretation:**

* **Curve Above Diagonal:** The ROC curve is above the diagonal line (representing random chance), indicating that the model performs better than random guessing.
* **High AUC Value:** The ROC curve and an AUC value of 0.99 suggest that the Random Forest model has excellent performance in distinguishing between fraudulent and non-fraudulent transactions.
* **Model Effectiveness:** The high area under the ROC curve demonstrates that the model is effective in differentiating between the two classes, making it a reliable tool for fraud detection.

The Random Forest model, supported by the Confusion Matrix and ROC Curve, demonstrates superior performance in fraud detection. The model’s high true negative rate, low false positive rate, and strong ROC-AUC score highlight its effectiveness in identifying fraudulent transactions while minimizing false alarms. This balance of precision and recall, coupled with robust overall accuracy, makes the Random Forest model the preferred choice for this application. This analysis addresses the research question: “Can a predictive Model accurately predict Fraud?”

**Feature Importance**

The feature importance analysis from the Random Forest model revealed that the top predictors of fraudulent transactions included temporal, geographic, and transaction amount features. Understanding these key predictors can help improve fraud detection strategies.****Figure 5: Feature Importance in Tuned Random Forest Model

**Analysis of Fraud Across Transaction Characteristics and Entry Modes**

This analysis addresses the research questions: "Which transaction characteristics are the most indicative of fraudulent activity?" and "How do different entry modes (e.g., Tap, PIN, CVC) affect the likelihood of a transaction being fraudulent?"

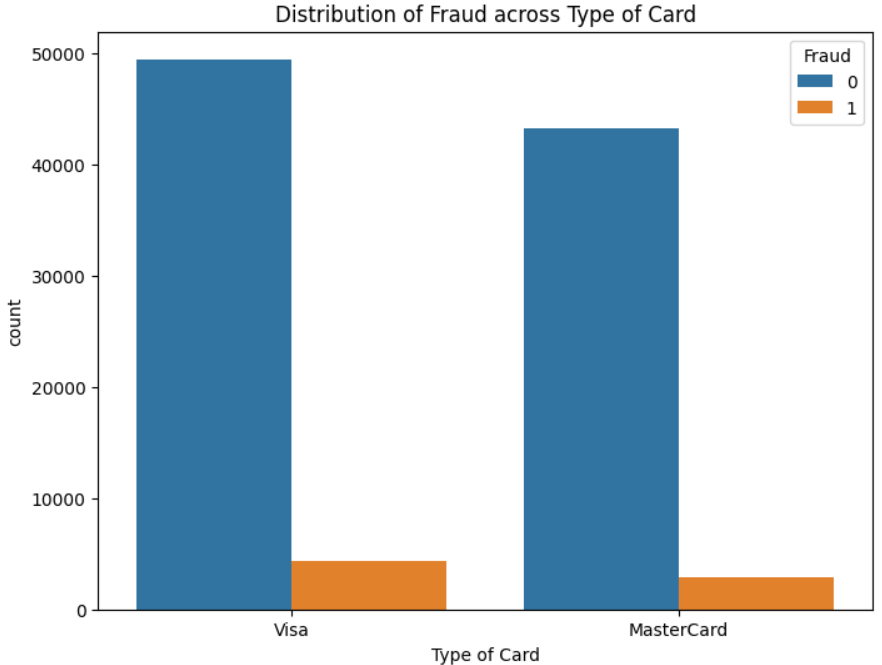
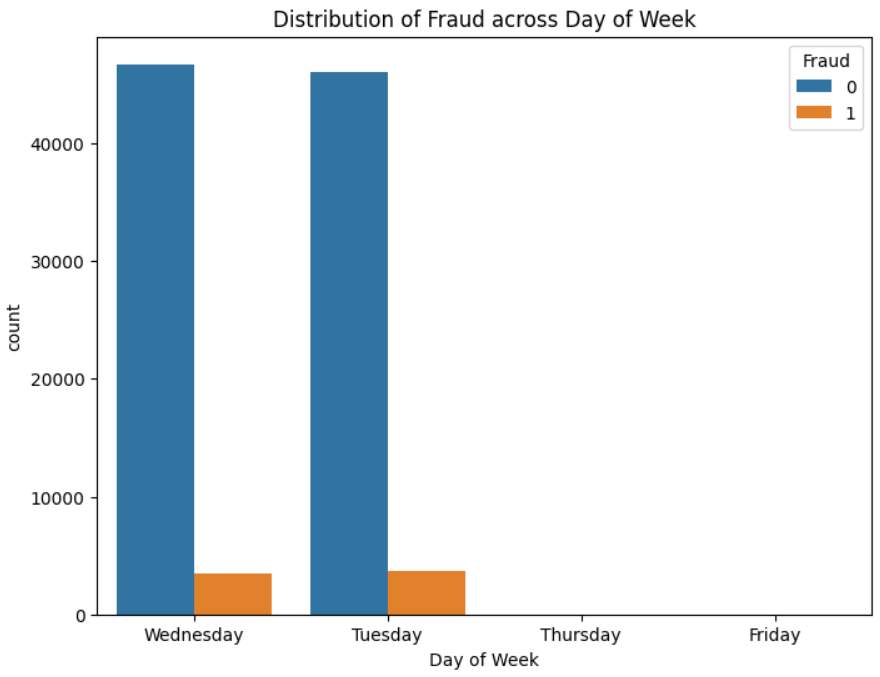
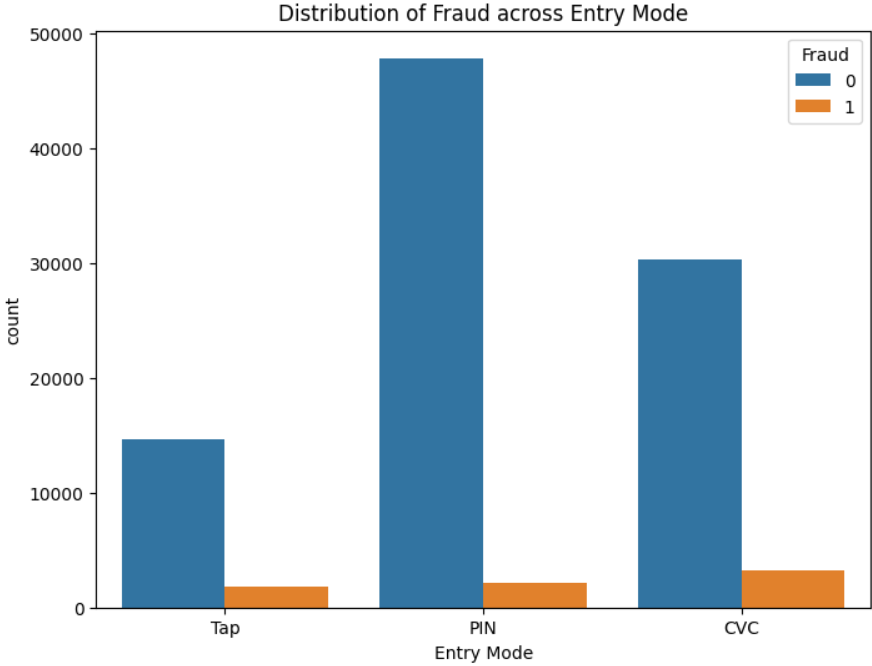
****

Figure 6: Distribution of Fraud Across Day of Week Figure 7: Distribution of Fraud Across Type of Card

**** **** Figure 8: Distribution of Fraud Across Entry Mode Figure 9: Distribution of Fraud Across Type of Transaction

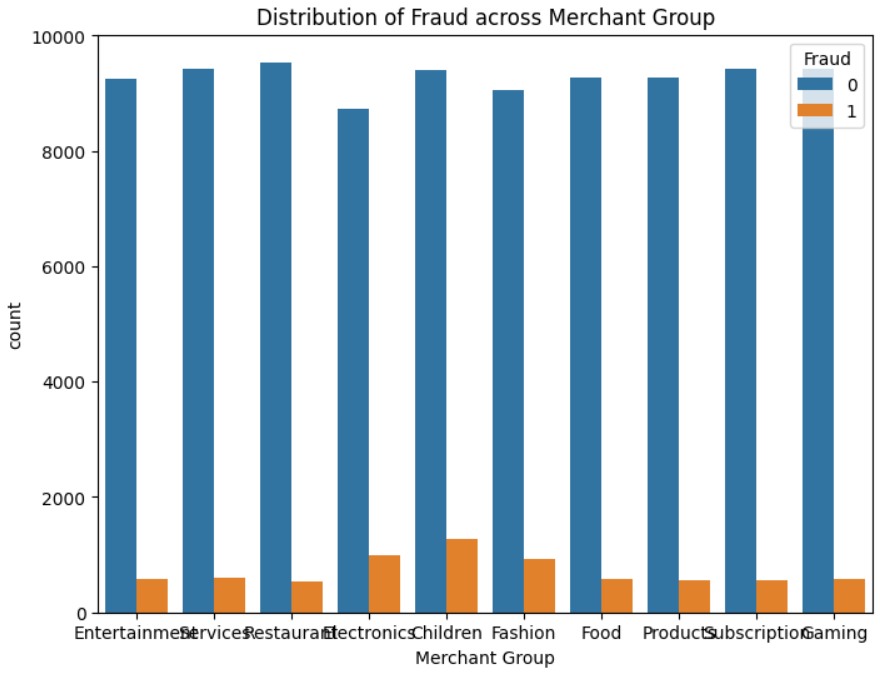
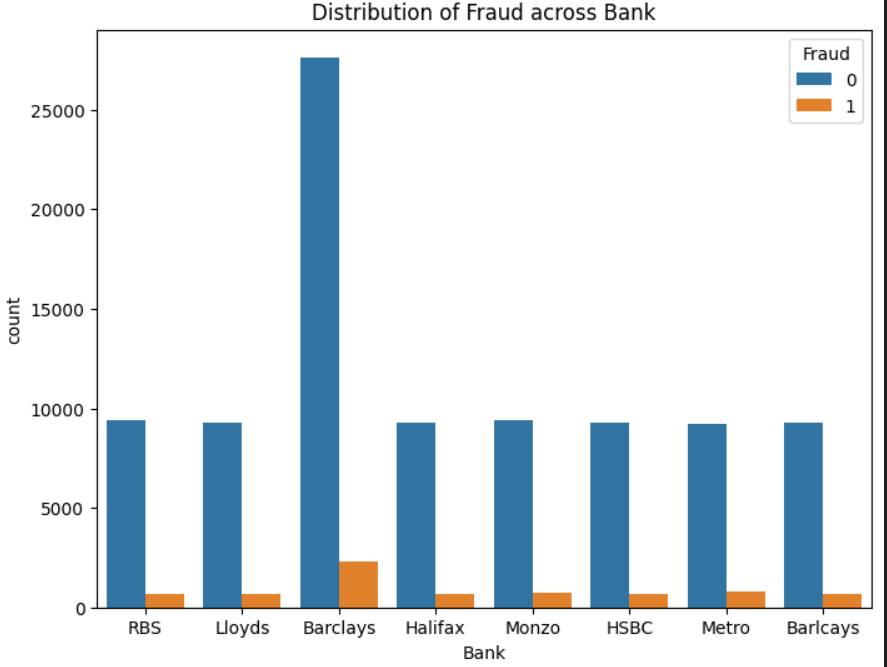
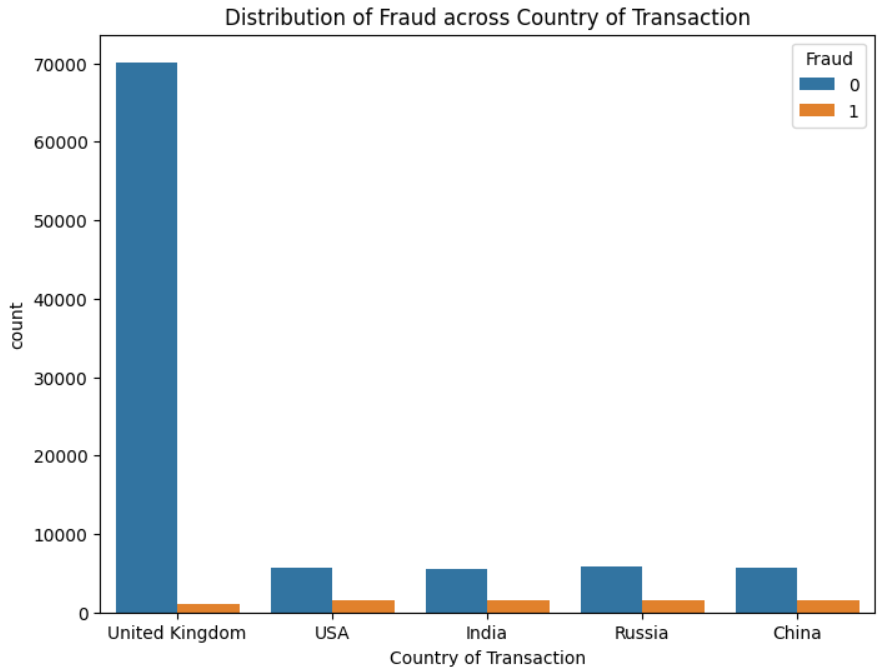
 

Figure 10: Distribution of Fraud Across Merchant Groups Figure 11: Distribution of Fraud Across Banks

**** **** Figure 12: Distribution of Fraud Across Country of Transaction Figure 13: Distribution of Fraud Across Shipping Address

**Fraud Distribution Across Days of the Week (Figure 6)**

* **Observation:** Fraudulent transactions occur most frequently on Tuesdays and Wednesdays.
* **Insight:** These days may represent peak periods for fraudulent activities, possibly due to transaction volumes or specific fraudulent strategies targeting these days.

**Fraud Distribution by Type of Cards (Figure 7)**

* **Observation:** Both Visa and MasterCard show instances of fraud, with Visa having a slightly higher number of fraudulent transactions.
* **Insight:** Visa cards might be more targeted or have higher transaction volumes, making them more susceptible to fraud.

**Fraud Distribution Across Entry Modes (Figure 8)**

* **Observation:** The PIN entry mode has the highest number of transactions and a notable amount of fraud. Despite having fewer total transactions than PIN, the CVC entry mode shows a relatively higher proportion of fraudulent transactions. The Tap entry mode also exhibits fraudulent activities but to a lesser extent than PIN and CVC.
* **Insight:** While the most common and generally secure, PIN-based transactions still show significant fraud, potentially due to PIN theft or skimming. CVC transactions, though less frequent, have the highest proportion of fraud, indicating a vulnerability in card-not-present transactions. Tap transactions, with the lowest volume and fraud instances, should still be monitored and secured due to their increasing usage and potential risks.

**Fraud Distribution by Type of Transaction (Figure 9)**

* **Observation:** Fraudulent transactions are distributed across POS, online, and ATM transactions, with online transactions showing a higher count of fraud.
* **Insight:** Online transactions, being more prone to security breaches, show higher fraud rates, highlighting the need for enhanced security measures in e-commerce.

**Fraud Distribution by Merchant Group (Figure 10)**

* **Observation:** Fraudulent transactions are spread across various merchant groups, with children, electronics, and fashion showing higher counts.
* **Insight:** High-value items and frequent transactions in these categories make them attractive targets for fraudsters.

**Fraud Distribution by Bank (Figure 11)**

* **Observation:** Fraudulent transactions are present across various banks, with Barclays showing the highest count.
* **Insight:** The higher volume of transactions at Barclays might correlate with higher fraud occurrences, suggesting the need for more stringent fraud prevention measures at this bank.

**Fraud Distribution by Country of Transaction (Figure 12)**

* **Observation:** Fraudulent transactions are distributed across multiple countries. The United Kingdom shows the highest total transactions but the lowest number of frauds. Other countries like the USA, India, Russia, and China have fewer total transactions but a higher proportion of fraud.
* **Insight:** While the United Kingdom effectively mitigates fraud despite its high transaction volume, other countries should strengthen their fraud detection and prevention mechanisms to reduce the higher proportion of fraudulent activities.

**Fraud Distribution by Shipping Address (Figure 13)**

* **Observation:** The United Kingdom has the highest number of transactions based on shipping addresses, but it shows the least amount of fraud. The USA, India, Russia, and China have fewer total transactions but a higher proportion of fraud.
* **Insight:** Shipping address data confirms the trend seen with transaction locations, suggesting that fraud detection systems should closely monitor transactions involving these high-risk regions.

The analysis reveals that the most indicative transaction characteristics of fraudulent activity include the day of the week, with Tuesdays and Wednesdays showing the highest frequency of fraud, and the type of card used, with Visa being more susceptible. Online transactions exhibit a higher count of fraud, indicating a vulnerability to security breaches. Among merchant groups, children, electronics, and fashion show higher fraud counts. Specific banks like Barclays have a higher incidence of fraud, potentially due to higher transaction volumes.

Regarding entry modes, PIN-based transactions display significant fraud, likely due to PIN theft or skimming. Despite being less frequent, CVC transactions have the highest proportion of fraud, suggesting vulnerabilities in card-not-present transactions. Tap transactions, with fewer instances of fraud, still require monitoring due to their growing usage and potential risks.

This analysis underscores the need for targeted fraud prevention strategies focusing on high-risk regions, transaction types, and entry modes.

**Analysis of Fraud Across Demographic Factors**

This analysis addresses our research question: How do demographic factors influence the occurrence of fraudulent transactions?

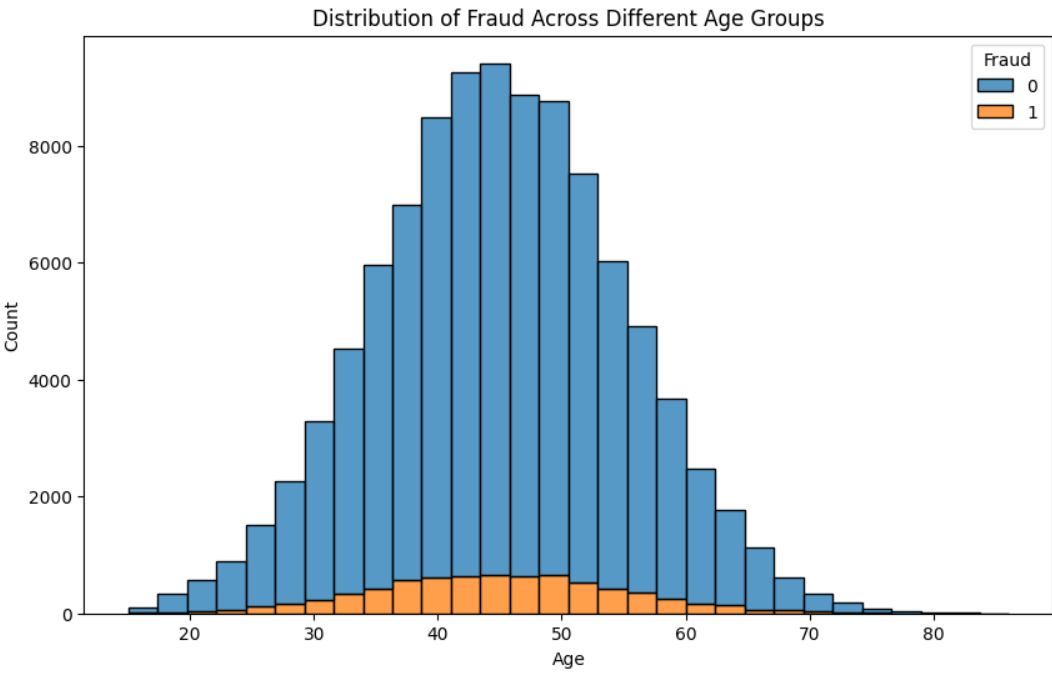
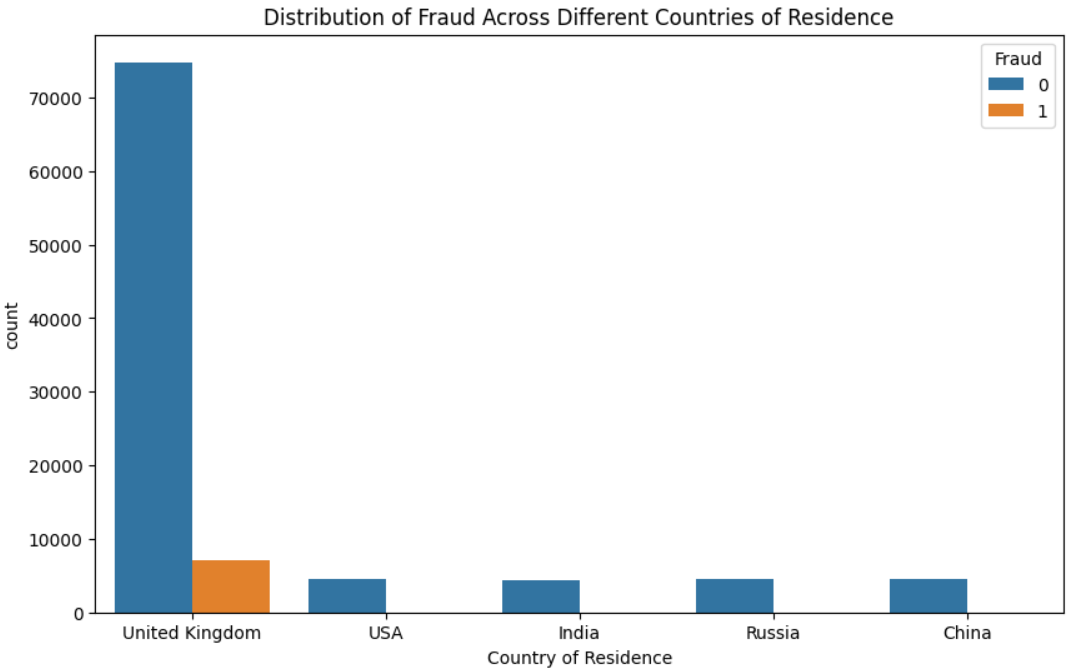
**** ****Figure 14: Distribution of Fraud Across Different Age Groups Figure 15: Distribution of Fraud Across Different Genders 

Figure 16: Distribution of Fraud Across Day of Week

**Fraud Distribution by Age Group (Figure 14)**

* **Observation:** Fraudulent transactions are most frequent among individuals aged 30 to 60. This age group shows a higher concentration of fraud cases than younger and older.
* **Insight:** The age group 30-60 might be more targeted or susceptible to fraud due to their higher transaction volumes or risk behaviors.

**Fraud Distribution by Gender (Figure 15)**

* **Observation:** Both males (M) and females (F) show instances of fraudulent transactions, with a slightly higher count among males.
* **Insight:** Gender does not appear to significantly influence the likelihood of fraud, although there is a marginally higher occurrence among males.

**Fraud Distribution by Country of Residence (Figure 16)**

* **Observation:** The United Kingdom has the highest total transactions yet exhibits a relatively low proportion of fraudulent transactions.
* **Insight:** With its high transaction volume, this indicates that the UK maintains effective fraud prevention measures, while the other countries need to strengthen their fraud detection and prevention mechanisms.

The analysis reveals that demographic factors do influence the occurrence of fraudulent transactions. Individuals aged 30 to 60 exhibit the highest frequency of fraudulent transactions. Gender has a marginal impact, with males showing slightly higher fraud occurrences. The country of residence is a crucial factor, with the United Kingdom maintaining a high volume of transactions but a relatively low proportion of fraud, indicating effective fraud prevention measures.

# Recommendations/Ethical Considerations

Based on the analysis, targeted strategies are essential to mitigate fraudulent activities effectively. Here are some recommendations:

1. **Strengthen Security Measures for Online Transactions**:
   * Implement multi-factor authentication (MFA) for online transactions to add an extra layer of security.
   * Utilize advanced machine learning algorithms to identify and block suspicious activities in real-time.
2. **Focus on High-Risk Merchant Groups**:
   * Increase transaction monitoring within high-risk merchant groups such as electronics, fashion, and children's products.
   * Regular training should be conducted for merchants in these sectors to recognize and prevent fraudulent activities.
3. **Strengthen Regional Fraud Prevention**:
   * Conduct a detailed regional analysis to understand the specific fraud patterns in different countries.
   * Develop region-specific fraud prevention measures, particularly for countries like the USA, India, Russia, and China, where higher proportions of fraud are observed.
4. **Monitor and Secure Entry Modes**:
   * Enhance security measures for PIN-based transactions, such as encryption and secure PIN entry devices.
   * Implement advanced fraud detection algorithms for CVC and Tap transactions, which are shown to have higher fraud rates.
   * Introduce secure protocol for CVC-based transactions, such as tokenization and dynamic CVC codes.
   * Educate merchants about best practices for handling CVC transactions to reduce the risk of fraud.
5. **Target High-Risk Age Groups**:

* Develop targeted awareness campaigns for individuals aged 30 to 60 who show higher instances of fraudulent transactions.
* Implement customized security protocols for transactions involving this age group, such as additional verification steps.

#### **Ethical Considerations**

1. **Data Privacy**:
   * Ensure that all data used in fraud detection and analysis is handled in compliance with data privacy regulations such as GDPR and CCPA.
   * Implement robust data anonymization techniques to protect customer identities.
2. **Fairness and Bias**:
   * Regularly audit fraud detection algorithms to ensure they do not unfairly target specific demographic groups or regions.
   * Ensure transparency in fraud detection processes to build trust with customers.
3. **Customer Communication**:
   * Provide clear and transparent communication to customers about the steps being taken to protect their transactions and personal information.
   * Offer support and guidance to customers who fall victim to fraud, including easy access to reporting and resolution mechanisms.

By implementing these recommendations and considering the ethical implications, financial institutions can enhance their fraud detection capabilities and provide a safer transaction environment for their customers.

### Limitations and Future Work

# While this study provides valuable insights into fraud detection using machine learning techniques, several limitations should be acknowledged. Addressing these limitations in future research could enhance the robustness and applicability of the findings.

# **Limitations:**

# **Data Quality and Completeness**:

# **Impact on Results**: The dataset used in this study may have data quality and completeness limitations. Missing values, incorrect entries, or incomplete records can affect the accuracy and reliability of the analysis and the predictive models.

# **Future Work**: Future research should focus on obtaining more comprehensive and high-quality datasets. Collaborations with financial institutions could provide access to more detailed and accurate transaction data.

# **Generalizability**:

# **Impact on Results**: The findings of this study are based on a specific dataset and may not be generalizable to all organizations or industries. The patterns and predictors identified may vary significantly across different contexts.

# **Future Work**: Replicating the study with data from various industries and organizational contexts can validate the findings and enhance generalizability. Including datasets from different regions, transaction types, and financial institutions will help determine if the identified predictors and model performance hold across diverse settings.

# **Class Imbalance**:

# **Impact on Results**: The dataset used in this study exhibited a significant class imbalance, with far fewer fraudulent transactions compared to non-fraudulent ones. This imbalance can affect the model's ability to learn and predict fraud accurately.

# **Future Work**: Future research should explore advanced techniques for handling class imbalance, such as synthetic data generation (e.g., SMOTE), cost-sensitive learning, or ensemble methods to mitigate imbalance effects. These techniques can help improve the model's ability to detect fraudulent transactions without compromising accuracy.

# **Feature Selection and Engineering**:

# **Impact on Results**: While this study conducted extensive feature engineering and selection, there may still be important features not captured in the dataset or newly emerging transaction characteristics influencing fraud detection.

# **Future Work**: Future studies should explore additional feature engineering techniques and include new data sources, such as behavioral data or real-time transaction monitoring systems. Incorporating domain expertise and advanced feature selection algorithms can further enhance the model's predictive power.

# **Temporal Dynamics**:

# **Impact on Results**: The study provides a snapshot of fraudulent transactions at a given time, but fraud patterns and tactics can evolve. The static nature of the dataset limits the ability to capture these temporal dynamics.

# **Future Work**: Conducting longitudinal analyses to understand how fraud patterns change over time is essential. Future research should focus on developing models that can adapt to new types of fraud and evolving transaction characteristics. This may involve using real-time data and continuous model updates.

# **Future Work:**

# **Integration with Real-Time Systems**:

# Developing models that can be integrated into real-time transaction processing systems will enhance the practicality and utility of fraud detection. Future research should focus on creating scalable, efficient algorithms capable of processing large volumes of data in real-time.

# **Cross-Industry and Cross-Regional Studies**:

# Expanding the scope of the research to include cross-industry and cross-regional studies will provide a broader understanding of fraud patterns and improve the generalizability of the findings. This approach will help identify unique challenges and opportunities in different contexts.

# **User Behavior Analysis**:

# Incorporating user behavior analysis, such as transaction patterns, device usage, and geolocation data, can provide deeper insights into fraud detection. Future research should explore how behavioral analytics can complement traditional transaction characteristics to enhance model performance.

# **Advanced Fraud Detection Techniques**:

# Exploring advanced machine learning and artificial intelligence techniques, such as deep learning, anomaly detection, and graph-based methods, can improve the detection of complex and subtle fraud patterns. Future studies should investigate the applicability and effectiveness of these techniques in fraud detection.

# By addressing these limitations and focusing on the suggested areas for future research, financial institutions can further enhance their fraud detection capabilities. This will lead to more accurate, efficient, and robust systems that protect consumers and businesses from fraudulent activities.

# Conclusion

This study aimed to understand the impact of transaction characteristics and demographics on fraud detection using machine learning techniques. By analyzing a comprehensive dataset from Kaggle, we developed a predictive model that accurately distinguishes between fraudulent and non-fraudulent transactions. Here are the key findings and contributions of the study:

#### Significant Predictors of Fraud:

1. **Transaction Characteristics**
   * **Days of the Week**: Transactions occurring on Tuesdays and Wednesdays are more indicative of fraud.
   * **Type of Card**: Visa card usage is more susceptible to fraud.
   * **Entry Modes**: PIN-based transactions show significant fraud due to potential PIN theft or skimming. Despite being less frequent, CVC transactions have the highest proportion of fraud, indicating vulnerabilities in card-not-present transactions.
   * **Type of Transactions**: Online transactions exhibit a higher count of fraud, highlighting their vulnerability to security breaches.
   * **Merchant Groups**: Fraud is more prevalent in categories like children’s products, electronics, and fashion due to high-value items and frequent transactions.
   * **Banks**: Barclays exhibits the highest count of fraudulent transactions, likely due to higher transaction volumes.
   * **Countries**: Despite having the highest transaction volume, the UK shows effective fraud prevention with a relatively low proportion of fraud, while other countries like the USA, India, Russia, and China have higher proportions of fraud.
2. **Demographic Factors**:
   * **Age Group**: Individuals aged 30 to 60 exhibit the highest frequency of fraudulent transactions, possibly due to higher transaction volumes or risk behaviors.
   * **Gender**: Both males and females are susceptible to fraud, with a slightly higher count among males.
   * **Country of Residence**: The UK shows a high volume of transactions but maintains a low proportion of fraud, indicating effective fraud prevention measures. Other countries like the USA, India, Russia, and China need to strengthen their fraud detection mechanisms.

#### **Practical Applications**

* **Enhancing Security Measures**: Focus on strengthening security for online transactions, Visa card transactions, and high-risk merchant categories like electronics and fashion.
* **Targeted Fraud Prevention**: Deploy targeted interventions and enhanced security protocols for transactions on Tuesdays and Wednesdays.
* **Improving Bank-Specific Measures**: Banks like Barclays with higher fraud counts should implement more stringent fraud prevention measures and continuously update their fraud detection systems.
* **Leveraging Predictive Models**: Utilize the Random Forest model to predict fraudulent transactions effectively. Integrate the model into transaction processing systems for real-time fraud detection and prevention. Regularly update and retrain the predictive model with new data to maintain accuracy and relevance.

#### **Impact and Contributions**

This study demonstrates the effectiveness of machine learning techniques in identifying and predicting fraudulent transactions based on transaction characteristics and demographics. The Random Forest model showed superior performance with high precision, recall, and ROC-AUC scores, making it a reliable tool for fraud detection. By providing actionable insights and practical recommendations, this study contributes to enhancing the accuracy and efficiency of fraud detection systems, ultimately minimizing losses and ensuring secure transaction environments.

#### **Future Work**

To build on this research, future studies should address identified limitations, such as data quality and completeness, generalizability, class imbalance, feature selection and temporal dynamics. Conducting longitudinal analyses, integrating models with real-time systems, expanding cross-industry and cross-regional studies, and incorporating user behavior analysis will further improve the predictive accuracy and effectiveness of fraud detection mechanisms.

By addressing these limitations and focusing on the suggested areas for future research, financial institutions can enhance their fraud detection capabilities, leading to more robust and reliable systems that protect consumers and businesses from fraudulent activities.

**References**

1. Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602-613. [**Data mining for credit card fraud: A comparative study - ScienceDirect**](https://www.sciencedirect.com/science/article/abs/pii/S0167923610001326)
2. Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review, 34(4), 267-308. [**[1009.6119] A Comprehensive Survey of Data Mining-based Fraud Detection Research (arxiv.org)**](https://arxiv.org/abs/1009.6119)
3. Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235-249.[**Statistical Fraud Detection: A Review (projecteuclid.org)**](https://projecteuclid.org/journals/statistical-science/volume-17/issue-3/Statistical-Fraud-Detection-A-Review/10.1214/ss/1042727940.full)
4. Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569.[**The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature - ScienceDirect**](https://www.sciencedirect.com/science/article/abs/pii/S0167923610001302)
5. Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines. Proceedings of the International MultiConference of Engineers and Computer Scientists, Vol. 1.[**Microsoft Word - IMECS2011\_ysahin\_eduman (iaeng.org)**](https://www.iaeng.org/publication/IMECS2011/IMECS2011_pp442-447.pdf)
6. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. Data Mining and Knowledge Discovery, 18(1), 30-55.
7. Carcillo, F., Le Borgne, Y. A., Caelen, O., & Bontempi, G. (2018). SCARFF: A scalable framework for streaming credit card fraud detection with spark. Information Fusion, 41, 182-194. [**SCARFF: A scalable framework for streaming credit card fraud detection with spark - ScienceDirect**](https://www.sciencedirect.com/science/article/abs/pii/S1566253517305444)
8. Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B., & Lemahieu, W. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75, 38-48. [**APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions - ScienceDirect**](https://www.sciencedirect.com/science/article/abs/pii/S0167923615000846)
9. Bauder, R. A., & Khoshgoftaar, T. M. (2018). A probabilistic programming approach for outlier detection in imbalanced datasets. Journal of Big Data, 5(1), 1-20.[**A Probabilistic Programming Approach for Outlier Detection in Healthcare Claims | Semantic Scholar**](https://www.semanticscholar.org/paper/A-Probabilistic-Programming-Approach-for-Outlier-in-Bauder-Khoshgoftaar/e882c347036629e7f3e87af7e52800f4c9d77959)
10. Mena, J. (2003). Investigative Data Mining for Security and Criminal Detection. Butterworth-Heinemann.

# Appendix

**Detailed Model Evaluation**

#### **Logistic Regression with Class Weight Adjustment**

**Classification Report:**

* **Class 0 (Non-Fraudulent):**
  + Precision: 1.00
  + Recall: 0.93
  + F1-Score: 0.96
  + Support: 27,791
* **Class 1 (Fraudulent):**
  + Precision: 0.54
  + Recall: 0.95
  + F1-Score: 0.68
  + Support: 2,209
* **Overall:**
  + Accuracy: 0.94
  + Macro Avg Precision: 0.77
  + Macro Avg Recall: 0.94
  + Macro Avg F1-Score: 0.82
  + Weighted Avg Precision: 0.96
  + Weighted Avg Recall: 0.94
  + Weighted Avg F1-Score: 0.94

**Confusion Matrix:**

* True Negatives (TN): 25,972
* False Positives (FP): 1,819
* False Negatives (FN): 115
* True Positives (TP): 2,094

**ROC-AUC:** 0.9822

#### **Logistic Regression with Hyperparameter Tuning and Class Weight Adjustment**

**Classification Report:**

* **Class 0 (Non-Fraudulent):**
  + Precision: 1.00
  + Recall: 0.94
  + F1-Score: 0.96
  + Support: 27,791
* **Class 1 (Fraudulent):**
  + Precision: 0.54
  + Recall: 0.95
  + F1-Score: 0.69
  + Support: 2,209
* **Overall:**
  + Accuracy: 0.94
  + Macro Avg Precision: 0.77
  + Macro Avg Recall: 0.94
  + Macro Avg F1-Score: 0.83
  + Weighted Avg Precision: 0.96
  + Weighted Avg Recall: 0.94
  + Weighted Avg F1-Score: 0.94

**Best Parameters:**

* {'C': 0.1, 'class\_weight': 'balanced', 'penalty': 'l1', 'solver': 'liblinear'}

**Confusion Matrix:**

* True Negatives (TN): 25,988
* False Positives (FP): 1,803
* False Negatives (FN): 116
* True Positives (TP): 2,093

**ROC-AUC:** 0.9819

#### **Random Forest with Class Weight Adjustment**

**Classification Report:**

* **Class 0 (Non-Fraudulent):**
  + Precision: 0.98
  + Recall: 1.00
  + F1-Score: 0.99
  + Support: 27,791
* **Class 1 (Fraudulent):**
  + Precision: 0.97
  + Recall: 0.81
  + F1-Score: 0.88
  + Support: 2,209
* **Overall:**
  + Accuracy: 0.98
  + Macro Avg Precision: 0.98
  + Macro Avg Recall: 0.90
  + Macro Avg F1-Score: 0.94
  + Weighted Avg Precision: 0.98
  + Weighted Avg Recall: 0.98
  + Weighted Avg F1-Score: 0.98

**Confusion Matrix:**

* True Negatives (TN): 27,741
* False Positives (FP): 50
* False Negatives (FN): 424
* True Positives (TP): 1,785

**ROC-AUC:** 0.9922

#### **Random Forest with Hyperparameter Tuning and Class Weight Adjustment**

**Classification Report:**

* **Class 0 (Non-Fraudulent):**
  + Precision: 0.98
  + Recall: 1.00
  + F1-Score: 0.99
  + Support: 27,791
* **Class 1 (Fraudulent):**
  + Precision: 0.97
  + Recall: 0.81
  + F1-Score: 0.88
  + Support: 2,209
* **Overall:**
  + Accuracy: 0.98
  + Macro Avg Precision: 0.98
  + Macro Avg Recall: 0.90
  + Macro Avg F1-Score: 0.94
  + Weighted Avg Precision: 0.98
  + Weighted Avg Recall: 0.98
  + Weighted Avg F1-Score: 0.98

**Best Parameters:**

* {'class\_weight': 'balanced', 'max\_depth': None, 'max\_features': 'sqrt', 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 100}

**Confusion Matrix:**

* True Negatives (TN): 27,741
* False Positives (FP): 50
* False Negatives (FN): 424
* True Positives (TP): 1,785

**ROC-AUC:** 0.9922