Business Analytics Approach to Credit Card Fraud Detection

Group 5

Chun-Yen Lin 989467253

Hsin-Yu Liao 989470611

Vishnu Vaibhav Binde 989486228

Miteshkumar Dasharathbhai Patel

Master Science of Business Analytics, University of the Pacific

MSBA 250 Applied Business Analytics

Dr. Xun Xu

Business Analytics Approach to Credit Card Fraud Detection

Abstract

This project investigates the pressing issue of credit card fraud, which poses a significant threat to the security and reliability of financial transactions in the digital era. Utilizing a publicly available dataset of anonymized credit card transactions, the study aims to identify patterns and behavioral indicators commonly associated with fraudulent activity. The analytical process involves comprehensive data preprocessing using Python, followed by the implementation of machine learning models to improve the accuracy of fraud prediction. In parallel, Tableau is employed for interactive data visualization, facilitating intuitive exploration of risk factors across variables such as transaction time, category, and user demographics. The final outcome includes a dynamic fraud monitoring dashboard designed to provide real-time insights and support informed decision-making. Overall, the project demonstrates the effective application of data analytics in combating fraud and strengthening the resilience of financial systems.

Keywords: Credit Card Fraud, Business Analytics, Fraud Detection, Machine Learning, Tableau Visualization

1. Introduction / Background

In an age where digital transactions have become the norm, credit card fraud has emerged as a significant global threat, impacting not just individual consumers but financial institutions and merchants as well. The sophistication of modern fraud schemes has outpaced traditional detection methods, demanding the implementation of advanced analytical tools.

According to the Nilson Report (2025), global losses due to credit card fraud have soared to over $33 billion, further emphasizing the need for data-driven preventive measures. This project, titled Credit Card Fraud Analysis, addresses these challenges by applying business analytics to a

dataset sourced from Kaggle. The dataset contains over 55,000 transaction records, complete with customer demographic information, transaction timestamps, locations, and fraud indicators.

This study focuses not only on identifying potential fraud patterns, but also on building a comprehensive understanding of customer risk profiles and merchant behaviors. The structure of the report follows a logical progression from background to objectives, data methods, analytics, conclusions, and practical recommendations, offering a complete journey from problem framing to solution implementation.

2. Project Objective

The primary objective of this study is to examine the characteristics of fraudulent credit card transactions and to understand how they differ from legitimate transactions in terms of timing, location, transaction amount, and user behavior. The aim is to identify patterns that allow financial institutions to proactively detect fraud while minimizing false positives that could disrupt genuine customer experiences.

To support this objective, the study analyzes trends in customer demographics, including age, gender, and other relevant attributes, and explores their correlation with fraud frequency. Merchant-related variables are also examined to identify categories and regions associated with elevated fraud risk. Furthermore, the project introduces derived features—such as transaction hour and customer age—to uncover behavioral patterns not immediately visible in the raw dataset.

One of the key deliverables of this study is an interactive Tableau dashboard that facilitates real-time monitoring of fraud trends, with the ability to assess risk by demographic segment and geographic region. Rather than focusing solely on technical modeling, this study emphasizes the development of tools that provide clear and actionable insights to support operational decision-making. Overall, the project illustrates how business analytics can be effectively applied to real-

world challenges, contributing to both improved fraud prevention strategies and enhanced customer confidence.

3. Data / Problem Analytics

3.1 Data

The dataset used in this study is sourced from the "Credit Card Fraud Prediction" project on the Kaggle platform, comprising approximately 550,000 transaction records. Each transaction includes multiple fields such as transaction time, merchant name, transaction category and amount, as well as basic consumer information including gender, city and state of residence, occupation, and date of birth. The is_fraud field indicates whether a given transaction is fraudulent. Among the entire dataset, only about 2,200 transactions are labeled as fraud, accounting for roughly 0.4% of the total sample. This indicates a highly imbalanced classification problem, which poses challenges for model training and prediction accuracy. Therefore, this study also implements corresponding techniques to address the class imbalance issue.

3.2 Methods

Before training the model, this study first conducted data cleaning and transformation using Python to enhance the accuracy of subsequent analysis and modeling. First, the original birthdate field was converted into age, which was used as a model feature to facilitate the exploration of the relationship between different age groups and fraudulent behavior. Second, the transaction time field was parsed to extract the hour of transaction, allowing for analysis of whether fraudulent transactions tend to occur during specific time periods. In addition, the dataset underwent column selection, missing value handling, and format standardization to ensure data consistency and usability, thereby laying a solid foundation for model development.

In the feature engineering phase, categorical variables such as gender, transaction type, state, and time period were transformed using LabelEncoder to convert text data into numerical formats

acceptable to the model. Numerical variables such as transaction amount, city population, and age were standardized using MinMaxScaler, which scaled all values between 0 and 1 to prevent bias caused by differences in scale. Given that only about 0.4 percent of the original dataset consisted of fraudulent transactions, resulting in a highly imbalanced class distribution, the study applied the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class in the training data. This approach improved the model's ability to learn from fraudulent patterns and enhanced its overall classification performance.

This study further developed and compared two supervised classification models: Logistic Regression and Random Forest. Logistic Regression offers strong interpretability, allowing for an understanding of the direction and magnitude of each variable's influence on the likelihood of fraud. In contrast, Random Forest leverages an ensemble of decision trees, providing high tolerance for nonlinear relationships and outliers. After training the models, their performance was evaluated using the test dataset. Evaluation metrics included Accuracy, Precision, Recall, F1-score, and the Confusion Matrix.

Finally, the study includes a visual analysis of the model prediction results. Through the use of probability distribution plots and model probability comparison charts, such as scatter plots and ROC curves, the differences and classification stability between the two models were examined. These visualizations provide valuable insights that support subsequent model selection and optimization decisions.

3.3 Data / Problem Analytics

To analyze the underlying trends and characteristics of fraudulent behavior, this study utilized Tableau for visual analysis to help understand the relationships between various variables and fraudulent transactions.

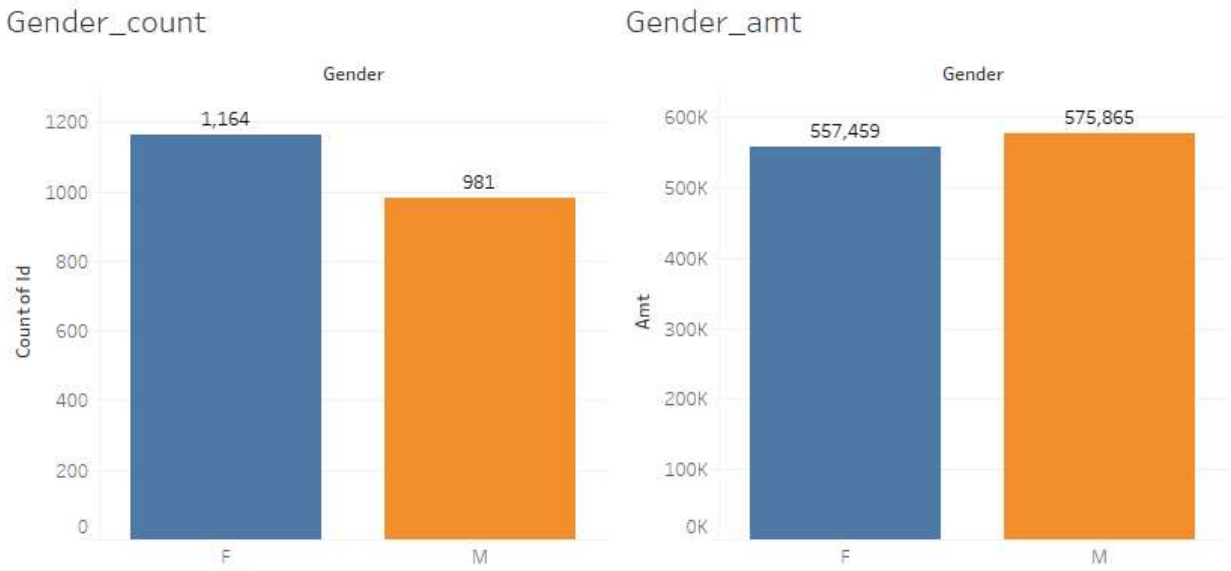

Figure 1. *Comparison of Fraud Transaction Count and Total Loss Amount by Gender*

First, an analysis was conducted based on gender, with two bar charts created: the left chart shows the number of fraudulent transactions (Count) by gender, while the right chart displays the total amount of fraud (Total Amount) for each gender.

From the figure 1, it can be observed that the number of fraudulent transactions involving female customers is 1,164, which is higher than the 981 cases involving male customers. However, in terms of total loss amount, male customers incurred a slightly higher total loss of 575,865 compared to 557,459 for female customers. This result suggests that while fraudulent transactions occur more frequently among females, the loss per transaction may be higher for males.

Next, an analysis was conducted based on transaction category. A bar chart was created to display the number of fraudulent transactions (Count of Id) across different categories.

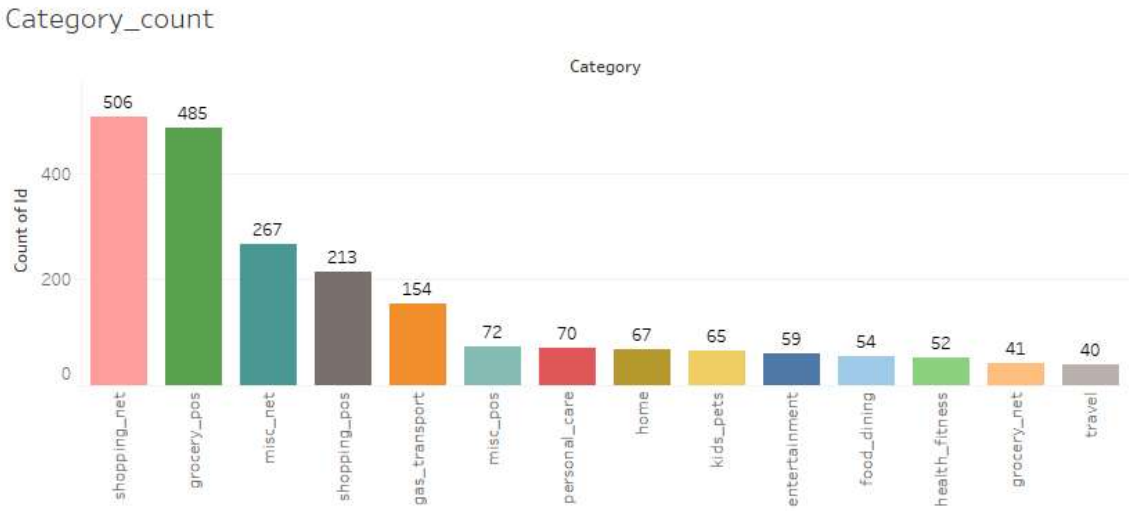

Figure 2. *Number of Fraud Transactions by Category*

From figure 2, it can be observed that fraudulent transactions occur most frequently in the shopping_net (online shopping) category, with a total of 506 cases. This is followed by grocery_pos (in-store grocery purchases) with 485 cases, and misc_net (other online transactions) with 267 cases. Other notable categories include shopping_pos (213 cases) and gas_transport (154 cases), which also show significant instances of fraud. These results suggest that transaction types related to everyday spending, especially online shopping, are primary targets for fraudsters. The high risk associated with online transactions may be linked to factors such as anonymity and automated payment processes. In-store transactions, such as those in grocery stores, may present opportunities for fraud due to their high frequency. Overall, while fraud is distributed across various categories, the top few with higher concentrations warrant special attention and enhanced security measures.

Next, an analysis was conducted based on transaction location (State), using both a map and a bar chart to present the distribution of fraudulent transaction amounts and counts. The map at the top visualizes the total fraud amount in each state using color intensity, while the bar chart below displays the number of fraudulent transactions (Count of Id) per state.

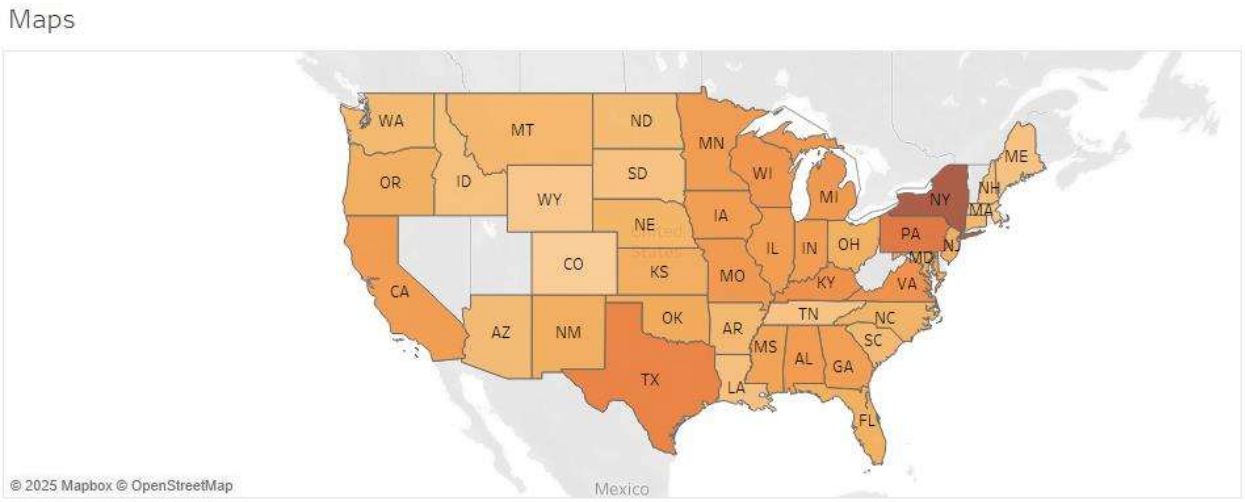

Figure 3 *Geographic Distribution of Fraud Losses by U.S. State*

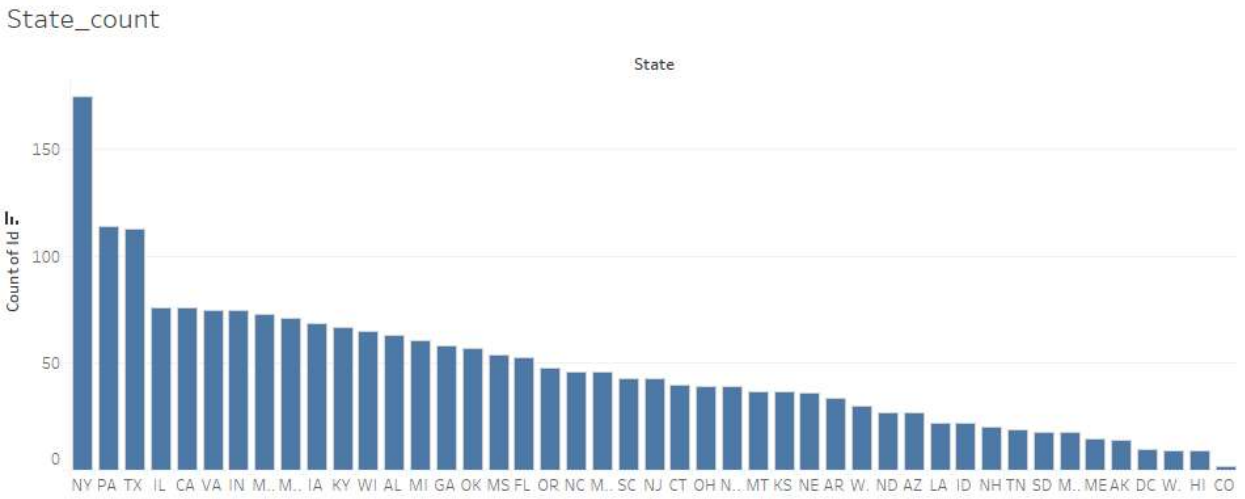

Figure 4 *Number of Fraud Cases by U.S. State*

From figure 3, it is evident that New York (NY) has the highest number of fraudulent transactions, totaling 180 cases, followed by Pennsylvania (PA) and Texas (TX), each with over 100 cases. These states are among the most populous in the U.S., and the higher volume of transactions may contribute to the increased number of fraud cases. Figure 4 also shows that New York, Pennsylvania, and Texas have relatively high total fraud amounts, indicating that these states not only experience frequent fraud but also significant financial losses. These findings highlight the geographic distribution of fraud cases across the United States, which appears to be related to

factors such as population density, consumer behavior, and the prevalence of online transactions. This information serves as a valuable reference for risk assessment and fraud prevention strategies.

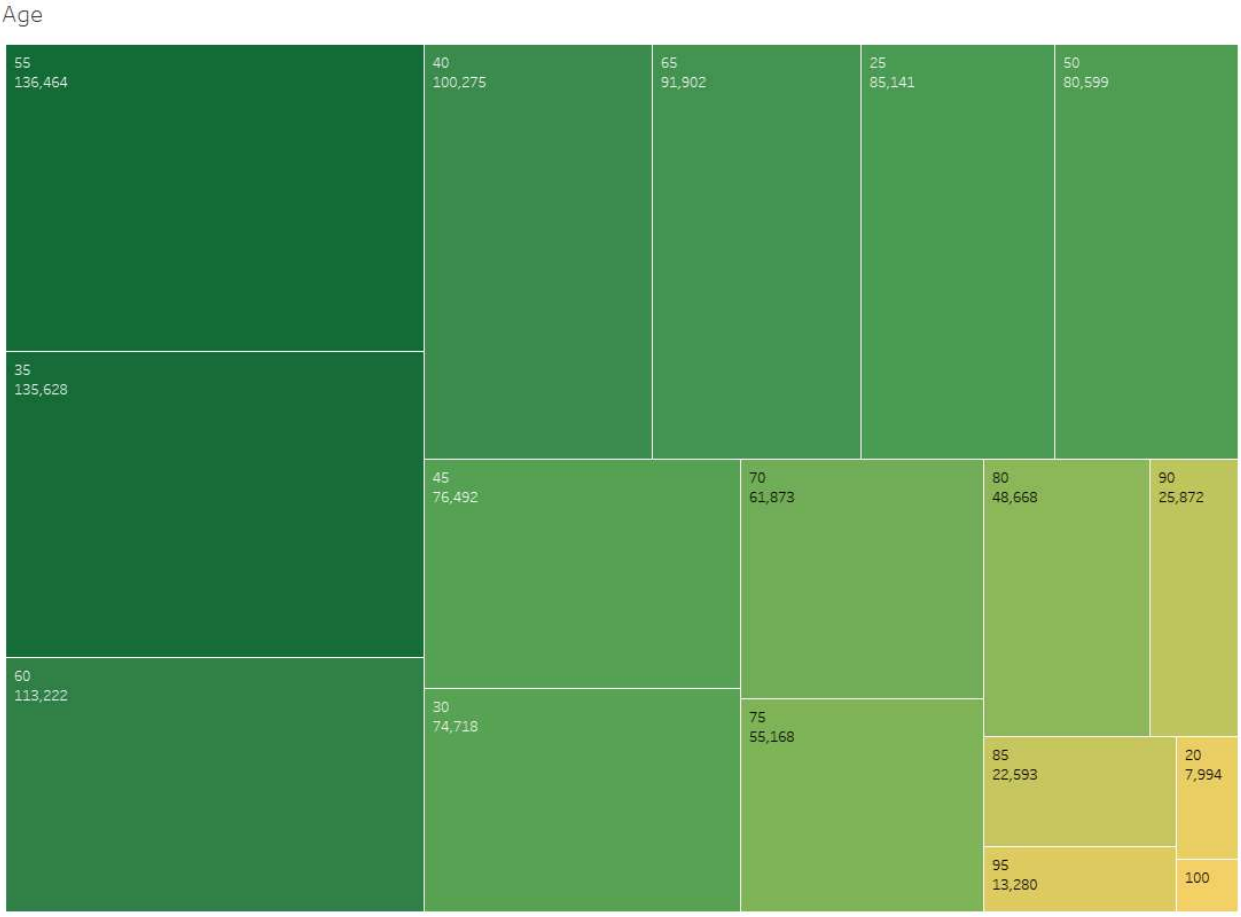

Figure 5 *Treemap of Fraud Losses by Customer Age*

Next, an analysis was conducted based on age, with a treemap created to display the total fraud amount (Total Amount) accumulated by different age groups in fraudulent transactions. Figure 5 shows that individuals aged 55 experienced the highest total fraud loss, reaching 136,464, followed by those aged 35 (135,628) and 60 (113,222), indicating that these three age groups account for the largest portions of fraudulent transaction amounts. In addition, the age groups of 40, 65, and 25 also recorded losses exceeding 85,000, marking them as high-risk segments. In contrast, younger individuals (such as those aged 20) and older adults (such as those aged 85, 90, and 95) showed relatively lower fraud amounts, ranging between 10,000 and 25,000. These results suggest that middle-aged to early senior individuals (approximately ages 35 to 65) not only appear more

frequently in fraud cases but also incur higher average losses. This may be attributed to their higher level of economic activity and greater asset holding. Therefore, enhanced financial security education and fraud prevention efforts should be prioritized for this demographic group.

Finally, an analysis was conducted based on transaction time (Hour), with a bar chart created to display the number of fraudulent transactions (Count of Id) across different time periods within a 24-hour span.

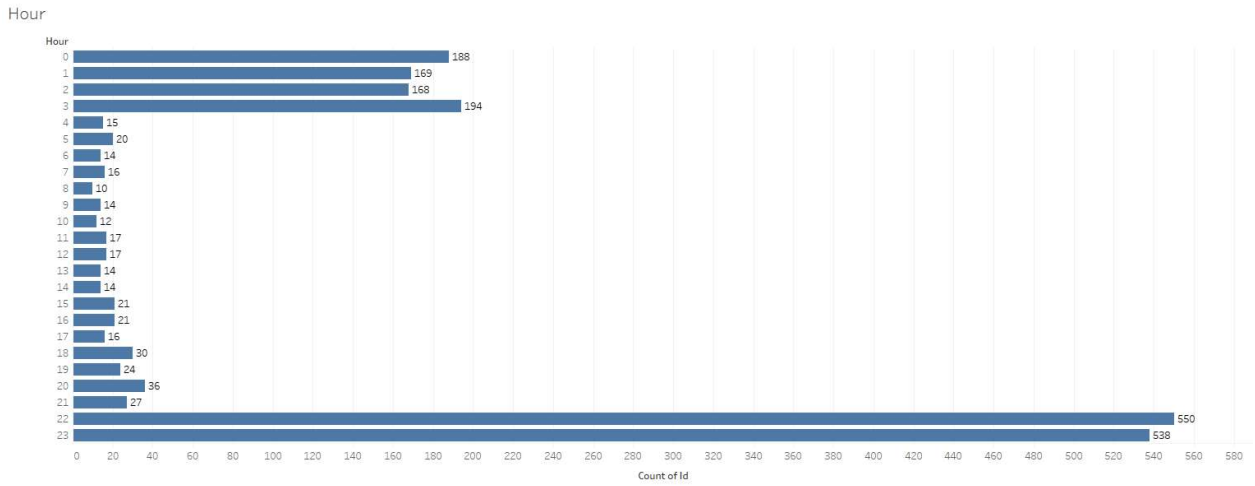

Figure 6 *Distribution of Fraudulent Transactions by Hour*

Figure 6 clearly shows that fraudulent transactions occur most frequently at 10 PM (550 cases) and 11 PM (538 cases), followed by 3 AM (194 cases), 12 AM (188 cases), and 1 AM (169 cases), all of which fall within the late-night to early-morning hours. Overall, there is a clear trend of fraud being concentrated during night-time and early-morning periods, while daytime transactions (from 6 AM to 5 PM) are significantly lower, typically ranging between 10 and 30 cases per hour. This pattern may be related to reduced user alertness at night, fraudsters intentionally avoiding peak monitoring hours, or increased mobile device usage by victims during nighttime. These findings suggest that the time of transaction is a critical factor in fraud risk assessment, and that enhanced real-time detection mechanisms and transaction risk controls are especially necessary during late-night hours.

In addition to conducting preliminary exploratory data analysis using Tableau, this study also developed fraud prediction models using two classification algorithms: Logistic Regression and Random Forest. The performance of these models was compared using various evaluation metrics and visualizations.

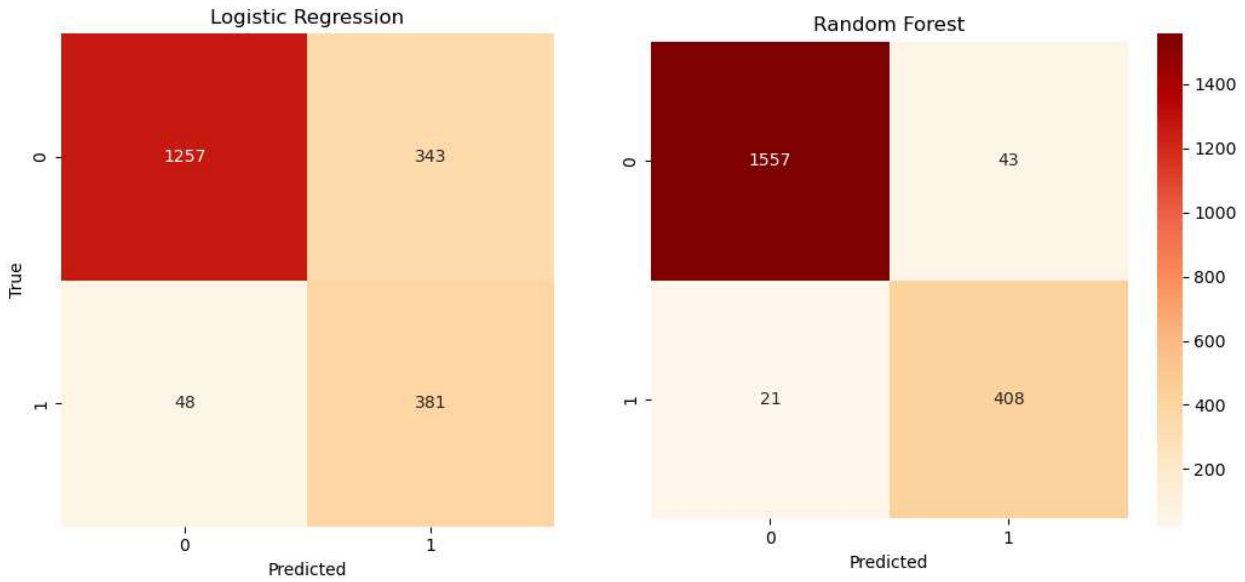

Figure 7. *Confusion Matrix Comparison between Logistic Regression and Random Forest*

First, the confusion matrix (Figure 7) provides insight into the models' performance in actual predictions. For instance, with the Random Forest model, out of 2,029 test samples, the model successfully identified 408 fraudulent cases (True Positives) and missed only 21 (False Negatives), demonstrating high accuracy. In contrast, while the Logistic Regression model also achieved a respectable recall of 0.89, it exhibited a higher false positive rate, misclassifying 343 legitimate transactions as fraudulent (False Positives), which could lead to excessive disruptions.

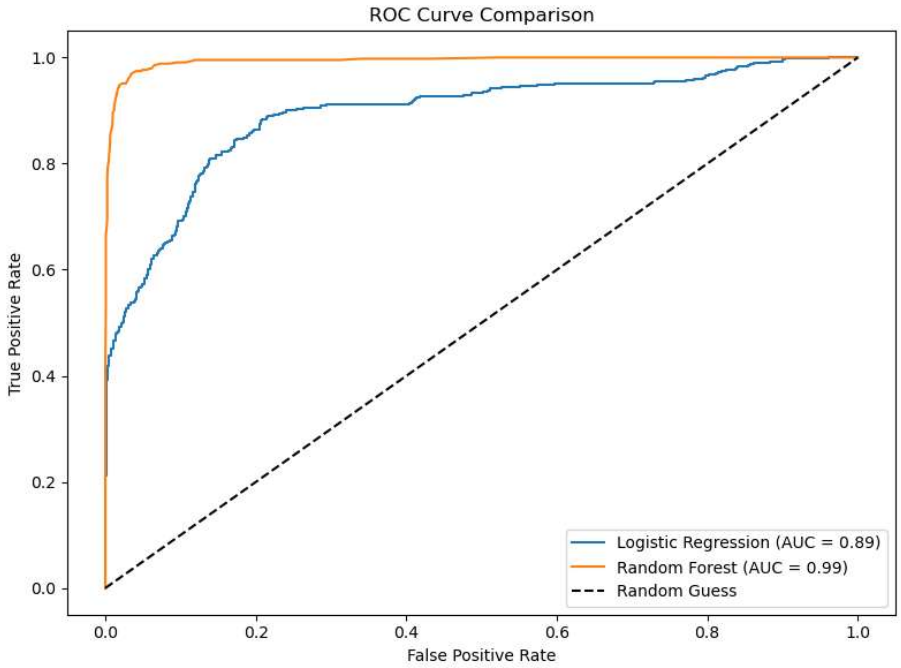

Figure 8. *ROC Curve Comparison Between Logistic Regression and Random Forest*

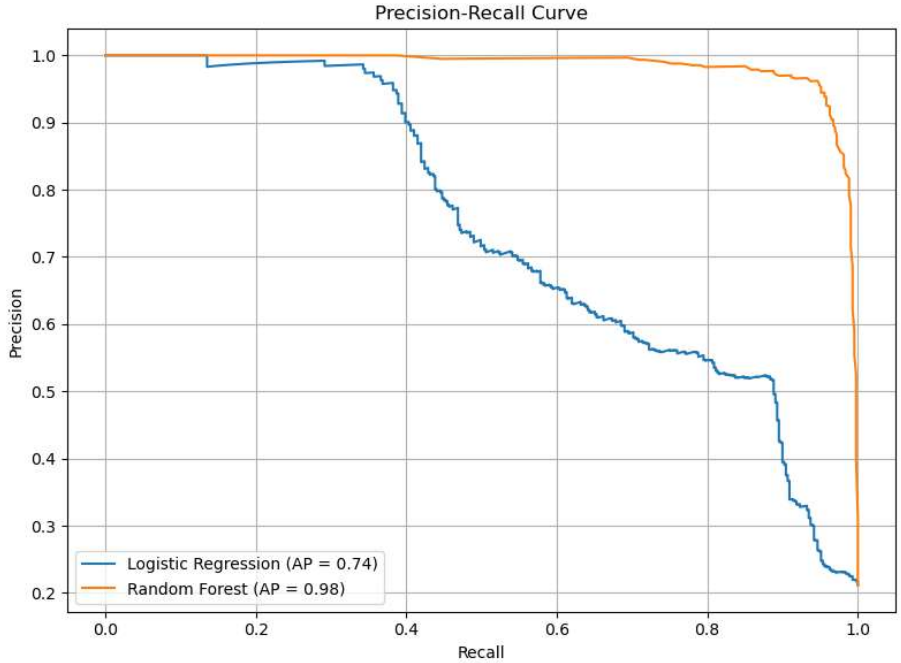

Figure 9. *Precision-Recall Curve Comparison Between Logistic Regression and Random Forest*

Further analysis using the ROC curve (Figure 8) and Precision-Recall curve (Figure 9) reveals that the Random Forest model outperforms Logistic Regression in overall performance. The Random Forest achieved an AUC of 0.99, significantly higher than Logistic Regression's

0.89. In terms of average precision (AP), Random Forest also demonstrated superior results, scoring 0.98 compared to 0.74 for Logistic Regression. These findings indicate that Random Forest is more effective in detecting fraudulent transactions within highly imbalanced datasets.

In addition, a similar trend was observed in the Precision-Recall curve. The Random Forest model achieved an average precision (AP) of 0.98, significantly outperforming Logistic Regression, which scored 0.74. This indicates that Random Forest demonstrates more robust performance when handling highly imbalanced data.

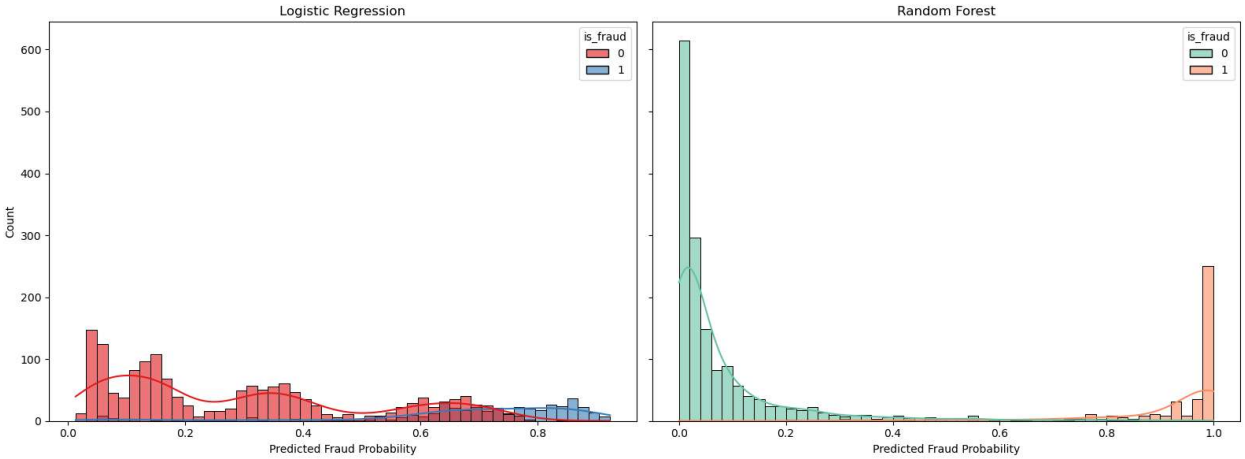

Figure 10. *Probability Distribution Comparison between Logistic Regression & Random Forest*

To further examine the distribution of model predictions, this study also show the predicted fraud probability distributions for both models (Figure 10). The results show that the Logistic Regression model produces a more dispersed distribution, with significant overlap between fraudulent and non-fraudulent samples, making the model more prone to confusion. In contrast, the Random Forest model is able to concentrate most fraudulent samples in the high-probability range, demonstrating greater prediction confidence.

Finally, to understand the contribution of each feature to the model's predictions, this study utilized the Feature Importance (Figure 11) provided by the Random Forest model to assess variable significance. The results indicate that transaction amount (amt) is the most critical factor,

followed by time period and transaction category. These variables demonstrate strong explanatory power in identifying fraudulent behavior.

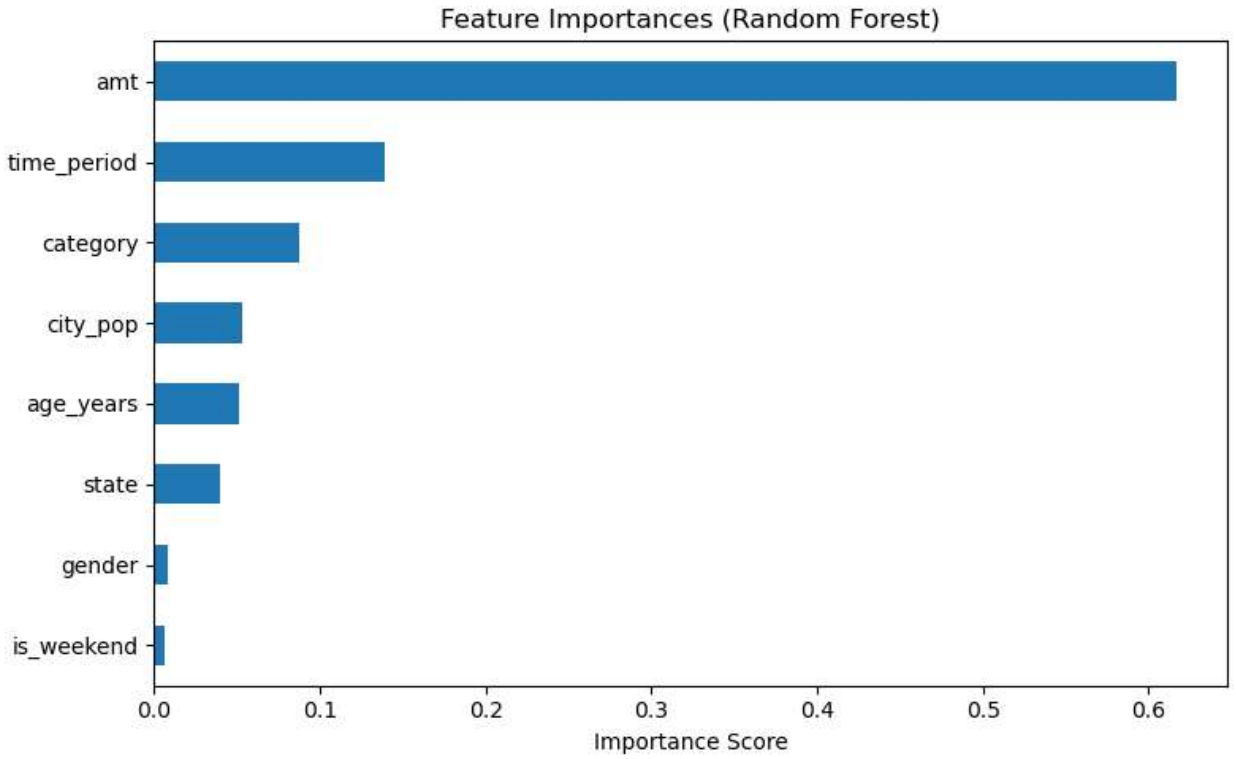

Figure 11. *Feature Importances from Random Forest Model*

The detailed classification report further reveals the overall performance of both models in terms of Precision, Recall, and F1-score. For specific figures, please refer to Table 1.

	Logistic Regression			**Random Forest**			
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Support
0	0.96	0.79	0.87	0.99	0.97	0.98	1600
1	0.53	0.89	0.66	0.90	0.95	0.93	429
Accuracy			0.81			0.97	2066
Marco Avg	0.74	0.84	0.76	0.95	0.96	0.95	2029
Weighted Avg	0.87	0.81	0.82	0.97	0.97	0.97	2029

Table 1. Classification Report of Logistic Regression and Random Forest Models

4. Findings / Conclusions

Through the visual analysis of fraudulent transaction data in this study, several key findings were identified:

First, regarding gender, although the number of fraudulent transactions involving female customers is higher than that of male customers, the total loss amount is slightly greater for males. This suggests that male victims may experience higher losses per transaction, while females may be more susceptible to fraud in terms of frequency.

Second, in terms of transaction categories, fraud occurs most frequently in those related to everyday purchases, particularly shopping_net (online shopping) and grocery_pos (in-store grocery shopping). This indicates that fraudsters tend to target high-frequency, widely used transaction types, with online channels being the primary focus due to their anonymity and ease of exploitation.

In terms of geographic distribution, fraud cases and amounts are highly concentrated in large states such as New York, Pennsylvania, and Texas. This suggests a potential positive correlation between fraud risk and factors like population density and economic activity. These regions may require enhanced fraud alerts and risk control mechanisms from local financial institutions and platforms.

The age group analysis reveals that individuals between the ages of 35 and 60—classified as middle-aged and working adults—experience the highest total fraud losses. Notably, those aged 55 and 35 each recorded losses exceeding 130,000. This may be due to their stronger financial capacity and higher engagement in online or financial transactions, making them prime targets for fraudsters.

Lastly, the analysis of transaction time indicates that fraudulent activities are concentrated during the late evening and early morning hours, with 10 PM and 11 PM being peak times—

significantly higher than other periods. This suggests that fraudsters may be exploiting reduced vigilance among victims during nighttime hours to carry out their attacks.

Based on the above analysis, it is recommended that businesses and institutions enhance their fraud prevention mechanisms by focusing on high-risk user groups, such as middle-aged individuals, high-risk time periods, particularly nighttime hours, and high-risk transaction categories, such as online shopping. The implementation of real-time monitoring and alert systems can further assist in reducing potential risks and strengthening overall fraud detection efforts.

Additionally, when comparing the performance of the two classification models—Logistic Regression and Random Forest—the Random Forest model clearly outperforms Logistic Regression. It achieved an overall accuracy of 97%, compared to just 81% for Logistic Regression. In terms of precision and recall, Random Forest also demonstrated better balance, particularly in detecting the minority class (fraudulent transactions), with a recall rate as high as 95%, significantly improving the model's ability to identify anomalous cases.

The confusion matrix further highlights the performance gap between the two models in real-world predictions. For Logistic Regression, although it accurately predicts the majority of normal transactions, it shows a relatively high false positive rate when identifying fraudulent cases. In contrast, the Random Forest model effectively detects most fraudulent transactions with fewer misclassifications, demonstrating greater feasibility and reliability in practical applications.

In the feature importance analysis, both models identified transaction amount (amt) as the most significant variable influencing prediction outcomes, followed by time period (time_period) and transaction category (category). This indicates that fraudulent transactions often occur within specific amount ranges and time windows, and are closely associated with certain transaction types—findings that align with the earlier visual exploration results.

Further examining model stability and performance through ROC and Precision-Recall curves, the Random Forest model achieved an AUC of 0.99 and an Average Precision (AP) of 0.98—significantly outperforming Logistic Regression, which recorded an AUC of 0.89 and an AP of 0.74. These results once again confirm the superiority of Random Forest in fraud detection tasks.

Based on the evaluation of machine learning models and visual analysis, this study recommends the use of ensemble learning methods such as Random Forest for handling highly imbalanced fraud datasets. When combined with oversampling techniques like SMOTE, these methods can significantly improve the model's ability to identify minority classes. Looking ahead, integrating deep learning models with real-time anomaly detection mechanisms holds great promise for developing more timely and accurate fraud prevention systems in practical applications.

5. Managerial Implications

The findings of this study offer actionable insights that managers can leverage to strengthen fraud prevention strategies and enhance overall operational performance. These analytics-driven results can inform strategic decision-making across various organizational functions.

For instance, the discovery that fraudulent transactions peak between 10 PM and 11 PM highlights the need for enhanced real-time monitoring during late-night hours. Additionally, identifying middle-aged consumers (ages 35–60) as the most vulnerable group calls for targeted financial education campaigns and increased security measures tailored to this demographic. Geographic insights also pinpoint high-risk states such as New York, Texas, and Pennsylvania, suggesting that resource allocation for fraud detection systems should be prioritized in these areas to reduce financial losses and boost consumer trust.

By implementing these findings, organizations can improve internal control mechanisms, safeguard customer assets, and reinforce their reputation in an increasingly complex digital

environment. Ultimately, applying the results of this project enables firms to proactively address fraud risk while driving smarter, data-informed management decisions.

6. Idea Sharing

This project provided a valuable opportunity to apply business analytics concepts in a real-world context, revealing both the technical challenges and practical potential of data-driven problem solving. One key learning outcome was the importance of managing imbalanced data, which is often encountered in fraud detection. The use of oversampling techniques such as SMOTE significantly improved the model's ability to detect rare but critical events.

Another important takeaway was the role of data visualization in translating complex analytical findings into clear, actionable insights. Tools like Tableau enabled the team to discover patterns that might have been missed through raw numerical analysis alone, reinforcing the importance of visual storytelling in communicating with diverse stakeholders.

The comparison between Random Forest and Logistic Regression models emphasized that model selection should consider not only predictive performance, but also the practical context in which the model will be deployed. This reinforced the idea that business analytics is not just about technical methods, but about applying analytical thinking to support strategic decision-making.

Ultimately, the project deepened our understanding of how analytics can bridge the gap between data and decisions. It reaffirmed the value of thoughtful analysis, careful model evaluation, and clear communication in delivering insights that can truly support business performance and customer protection.

7. References

Kelue, K. (2024, March 11). *Credit Card Fraud Prediction*. Kaggle. https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction

Report, N. (2025, January 6). Payment Card Fraud Losses Approach $34 Billion. *GlobeNewswire News Room*. https://www.globenewswire.com/news release/2025/01/06/3004931/0/en/Payment-Card-Fraud-Losses-Approach-34-Billion.html

8. Appendix

Time	Contents
Saturday January 25, 2025	Project Group Forms
Friday February 24, 2025	Defineing the issue and Topic
Monday February 24, 2025	Project Proposal Presentation
Monday March 24, 2025	Data preprocessing
Monday April 7, 2025	Training the model
Friday April 11, 2025	Visual Exploration of the Data
Monday April 21, 2025	Final Project Paper Submission and Final Project Presentation Slides Submission