Group 5

Credit Card Fraud Analysis and Prediction

Chun-Yen Lin
Hsin-Yu Liao
Miteshkumar Dasharathbhai Patel
Vishnu Vaibhav Binde

Table of contents

01 Introduction & Background

02 Objective

03 Dataset Overview

04 Data Analytics & Finding

05 Managerial Implications

06 Idea Sharing

Introduction

- Credit card fraud is a major global issue, especially with the rise of digital payments.
- Traditional fraud detection methods are not keeping up with new tactics.
- In **2023**, fraud losses exceeded **$32 billion** worldwide (*Nilson Report*).
- Our project uses a real dataset from Kaggle with **55,000+ transactions**.

Background

- The dataset includes: transaction time, amount, customer demographics, location, and fraud labels.
- Goal: Identify fraud patterns and create dashboards for monitoring and risk analysis
- Fraud not only causes financial loss but also damages customer trust and brand reputation.
- Business analytics helps uncover hidden patterns that traditional systems often miss.
- By combining data analysis and visualization, we can make fraud detection more efficient and proactive.

Project objective

- Understand how fraudulent transactions differ from legitimate ones.
- Analyze fraud patterns based on:
  - Time of transaction
  - Location and region
  - Customer demographics like age and gender
- Engineer features such as transaction hour and customer age to uncover hidden patterns.
- Build a Tableau dashboard to visualize fraud trends across time and user segments.
- Provide clear insights to support fraud monitoring and faster decision-making.

Dataset Overview

Source:
Kaggle – “Credit Card Fraud Prediction”

Key Features:

- Transaction Time
- Transaction Category & Amount
- Gender, City, State
- Date of Birth
- is_fraud (Fraud = 1)

550,000

Total Records

2,200

Fraud Cases

0.4%

Fraud Rate(Highly Imbalanced)

Data Preprocessing

For Data Visualization (Tableau)

Convert Birthdate → Age

- Used as a feature to explore age–fraud relationship

Parse Transaction Time → Extract Hour

- Analyze fraud trends by time of day

Additional Processing

- Column selection
- Format standardization

For Prediction (Python)

Feature Engineering

- Extracted customer age from birthdate and created age_years.
- Derived time_period from transaction hour
- Created is_weekend feature to indicate if a transaction occurred on a weekend.

Categorical Encoding

- Applied LabelEncoder to convert categorical features into numerical values

Numerical Scaling

- Standardized numerical features with MinMaxScaler to range between 0 and 1: amt, city_pop, age_years

Data Analytics

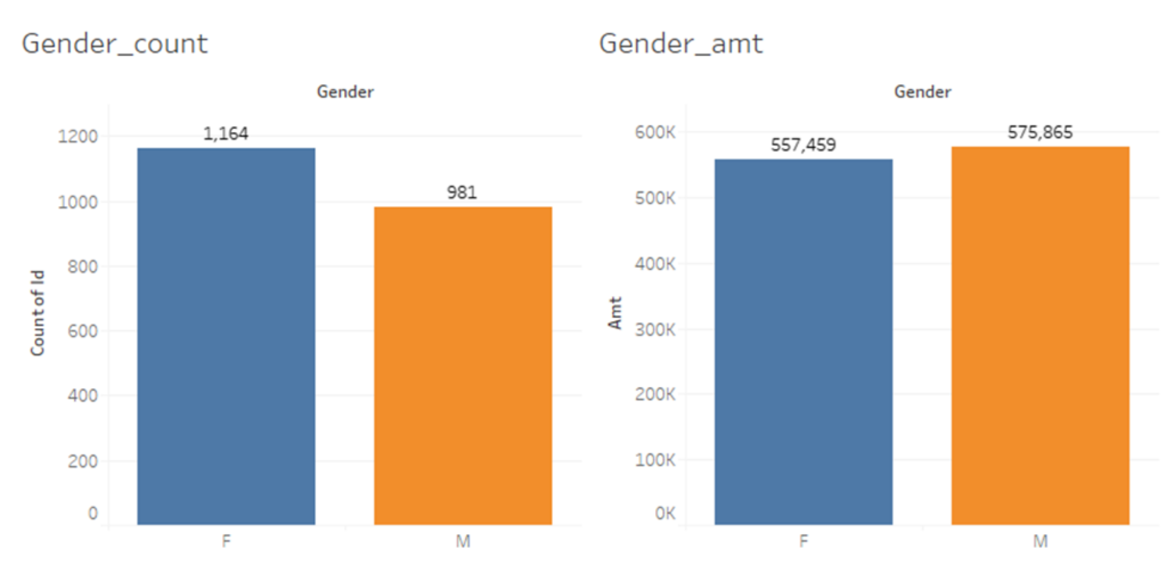

Figure 1. Comparison of Fraud Transaction Count and Total Loss Amount by Gender

Data Analytics

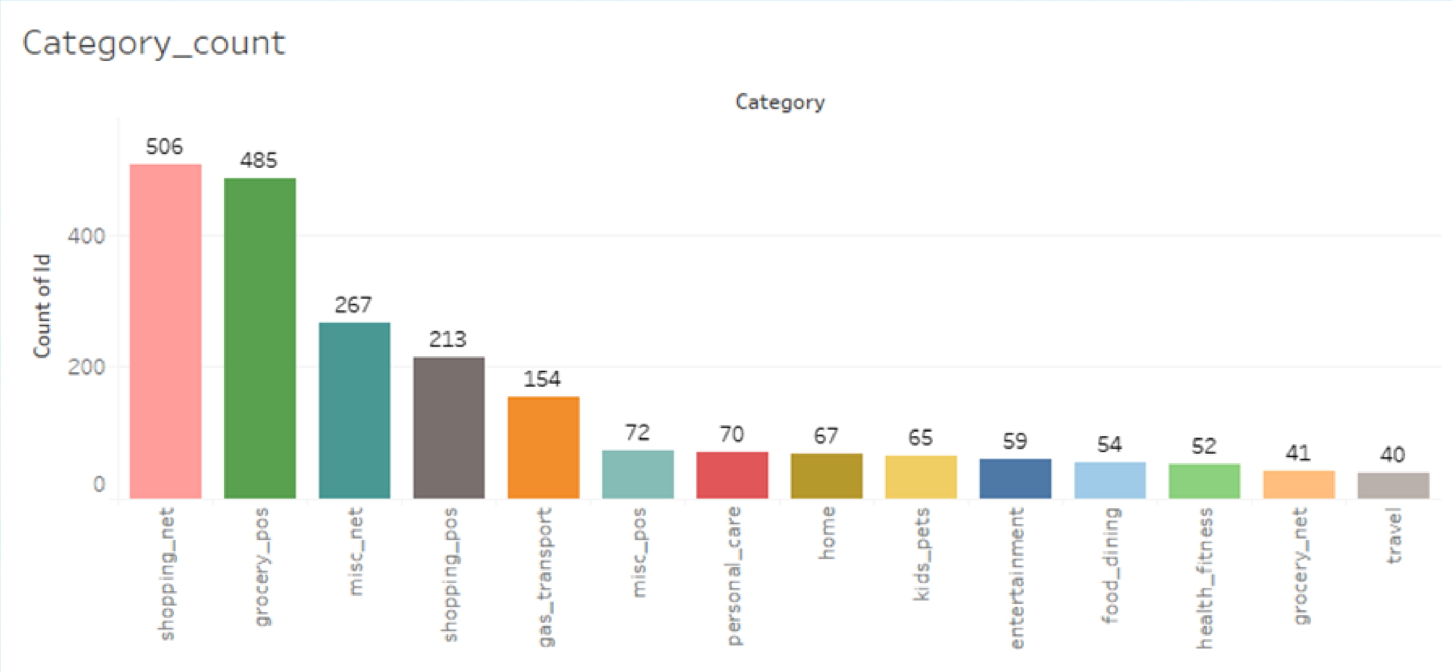

Figure 2. Number of Fraud Transactions by Category

Data Analytics

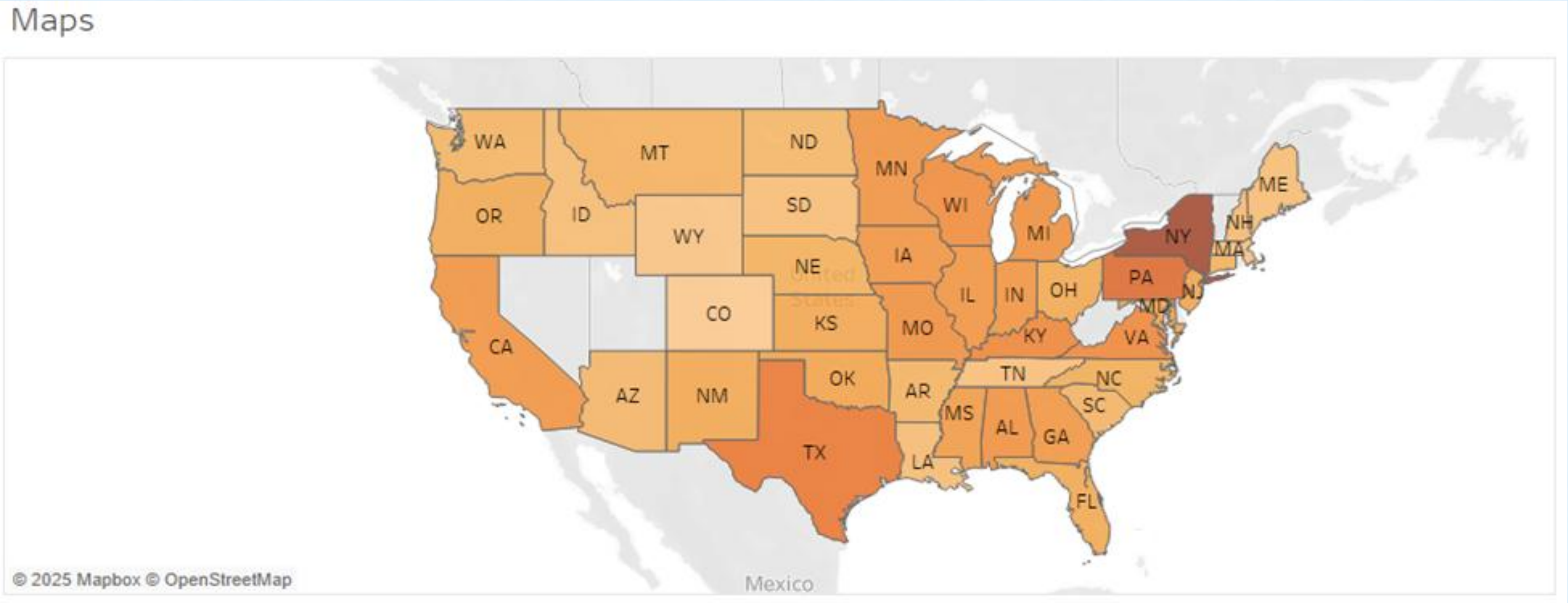

Figure 3. Geographic Distribution of Fraud Losses by U.S. State

Data Analytics

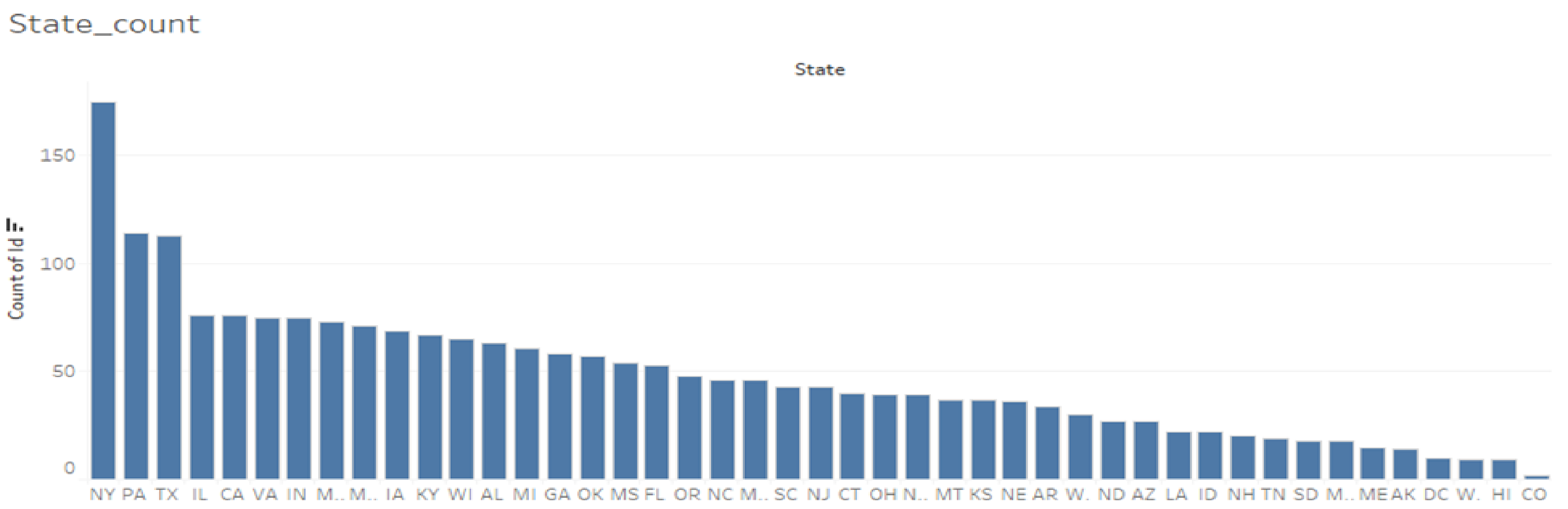

Figure 4. Number of Fraud Cases by U.S. State

Data Analytics

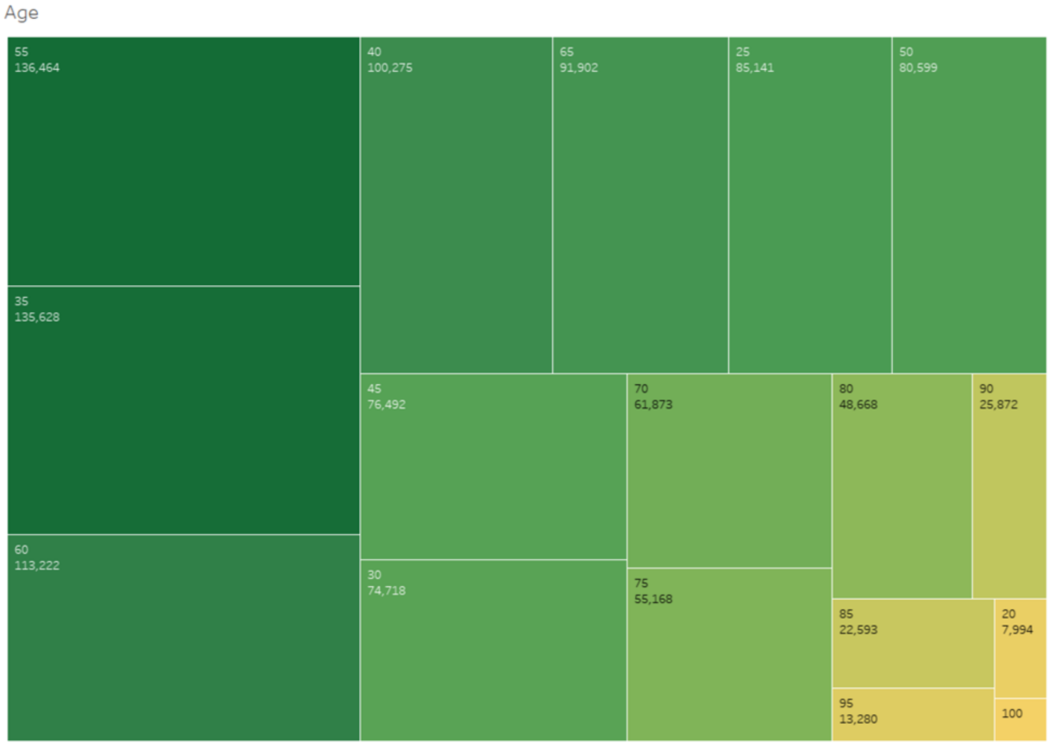

Figure 5. Treemap of Fraud Losses by Customer Age

Data Analytics

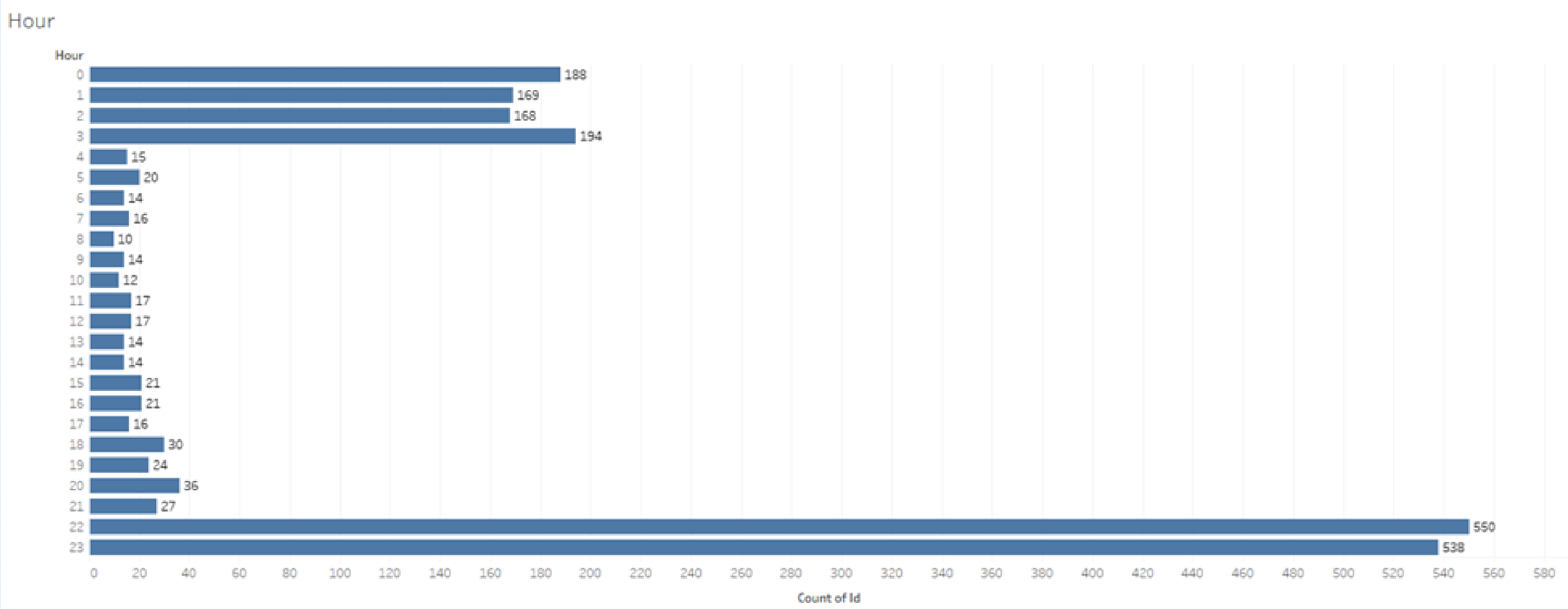

Figure 6. Distribution of Fraudulent Transactions by Hour

Data Analytics

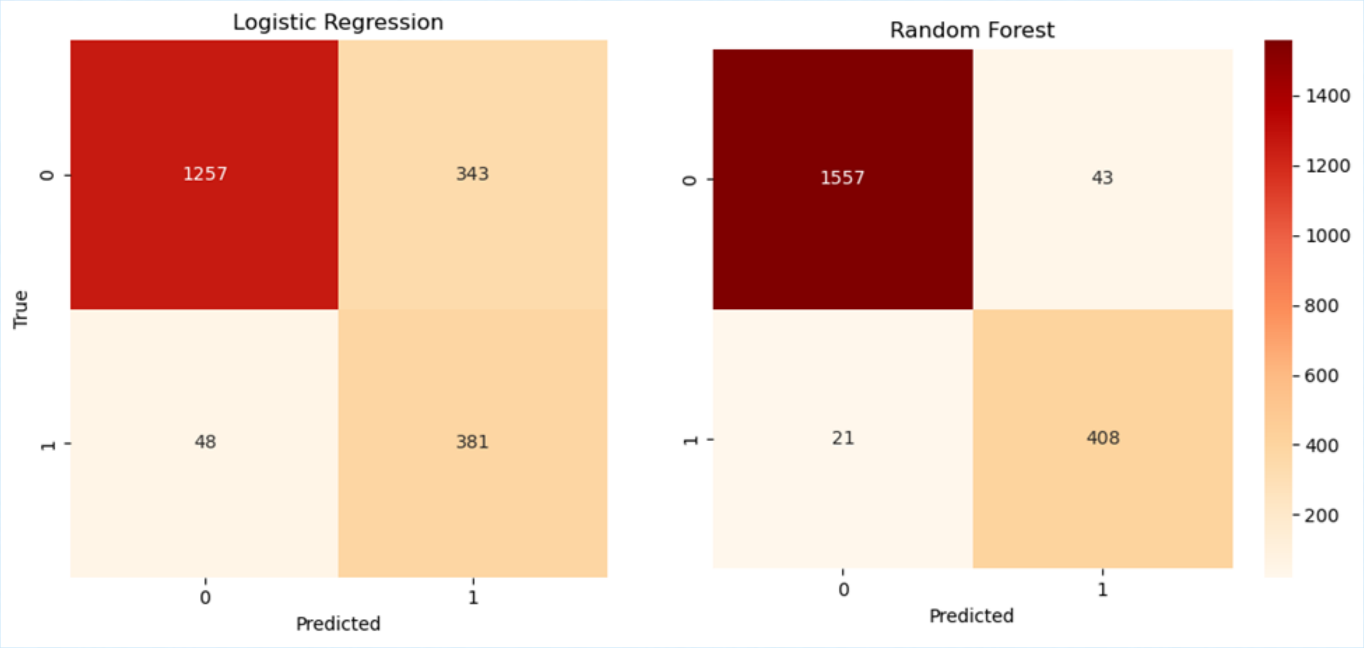

Figure 7. Confusion Matrix Comparison between Logistic Regression and Random Forest

Data Analytics

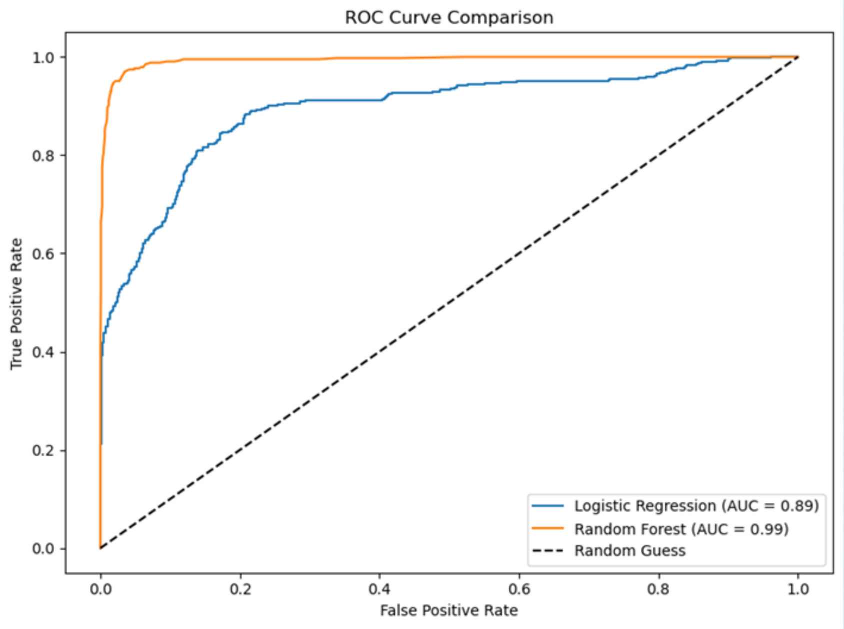

Figure 8. ROC Curve Comparison

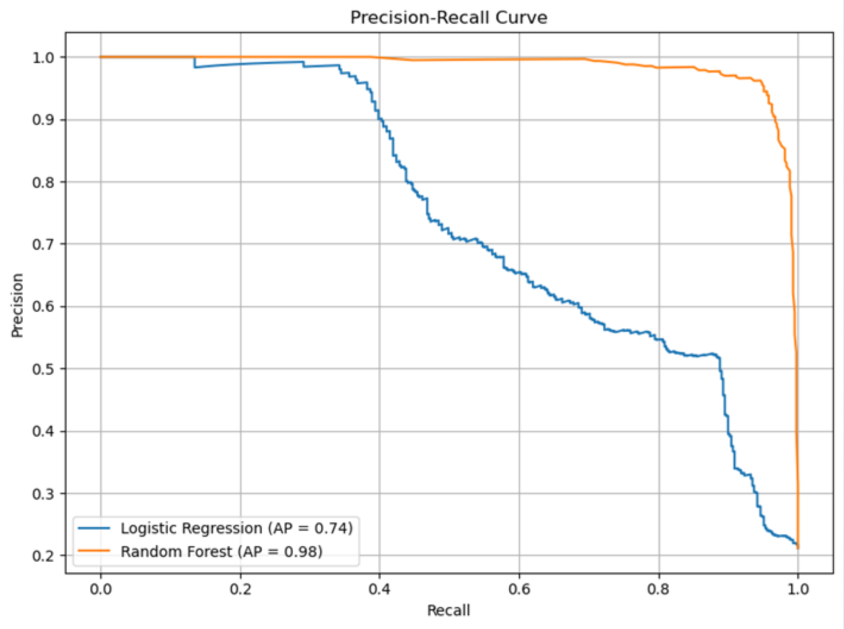

Figure 9. Precision-Recall Curve Comparison

Data Analytics

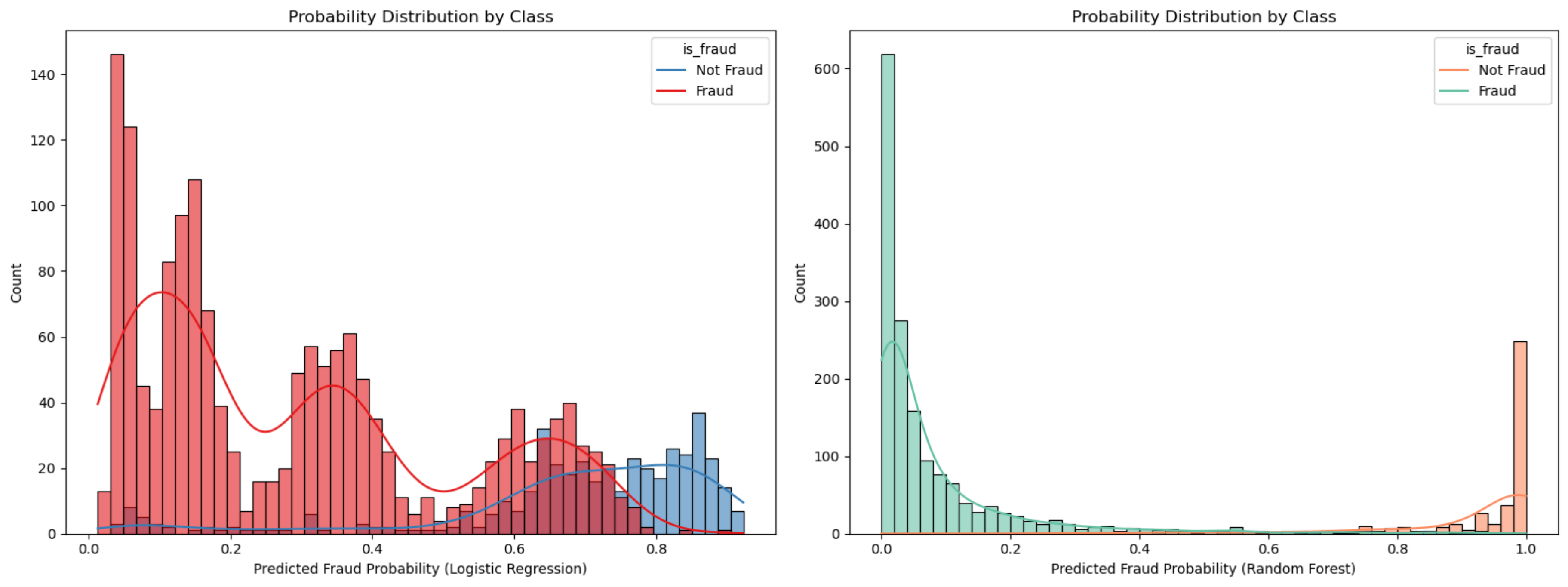

Figure 10. Probability Distribution Comparison between Logistic Regression & Random Forest

Data Analytics

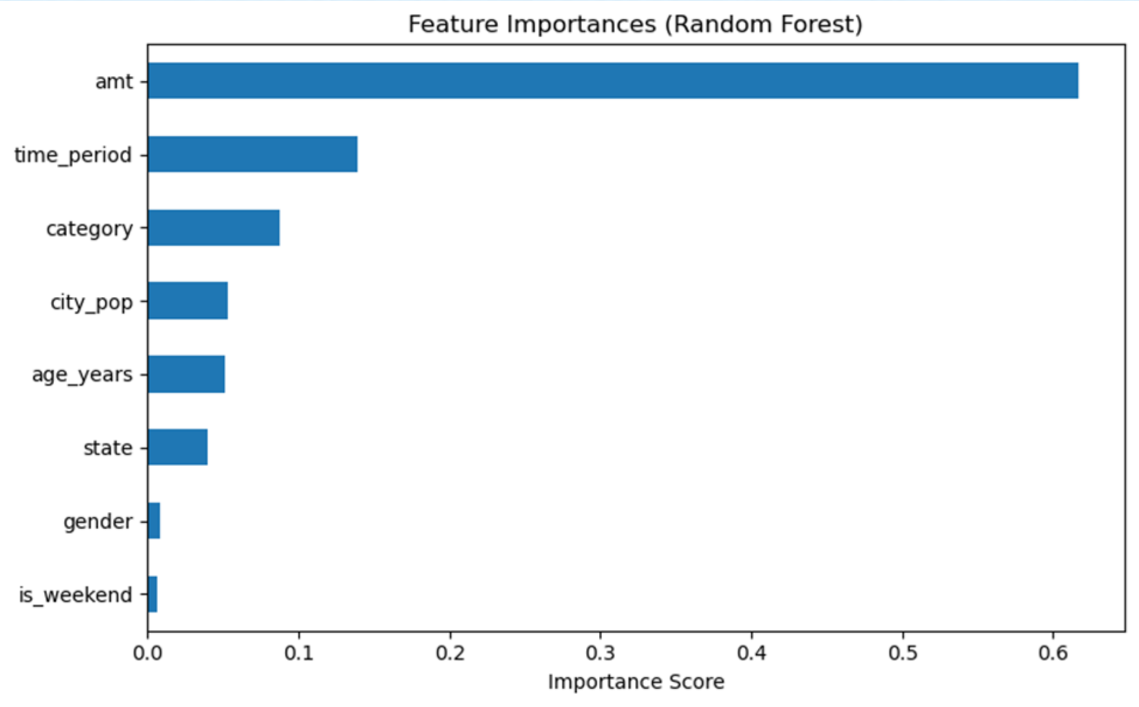

Figure 11. Feature Importances from Random Forest Model

Key Finding

Fraud Patterns by Gender

- Female users encountered more fraud cases.
- Male users lost more per transaction.

High-Risk Transaction Categories

- Fraud mainly occurred in **online shopping** and **in-store grocery.**

Geographic Risk Hotspots

- New York, Pennsylvania, and Texas showed the highest fraud frequency and total loss.

Age-Based Risk Groups

- Users aged **35–60** had the highest fraud loss, especially ages **55** and **35**.

High-Risk Time Periods

- Most fraud occurred during late night (10 PM – 12 AM).

Random Forest had better overall performance in fraud detection.

- Important Features: Transaction Amount, Time Period, and Category.

Managerial Implications

- Increase real-time monitoring during peak fraud times (10 PM–11 PM).
- Targeted educational campaigns for vulnerable groups (ages 35–60).
- Focused security resources in high-risk states (NY, TX, PA).
- Improve internal controls, enhance customer trust, and protect company reputation.

Idea Sharing & Project Experience

- Gained practical skills managing imbalanced datasets (SMOTE).
- Enhanced understanding using visualization tools (Tableau).
- Learned strategic selection between models (Random Forest vs. Logistic Regression).
- Strengthened appreciation for analytics in solving real-world business problems.

Reference

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

Thanks!

Do you have any questions?