GROUP 2

Credit Card Fraud Detection: A Machine Learning Approach

Bradley Agwa - 673288
Candence Chumba - 673238
Tyrone Darren - 674100
Mishiel Nasambu Wakoli - 673012
Melissa Wachira - 672019
Joseph Kamau - 672753

Introduction

Problem:

- Credit card fraud costs $40B+ annually, with traditional rule-based systems failing to adapt to new fraud patterns.
- Limitations of traditional rule-based systems:

  High false positives (legitimate transactions blocked).

  High false negatives (fraudulent transactions missed).

Solution:

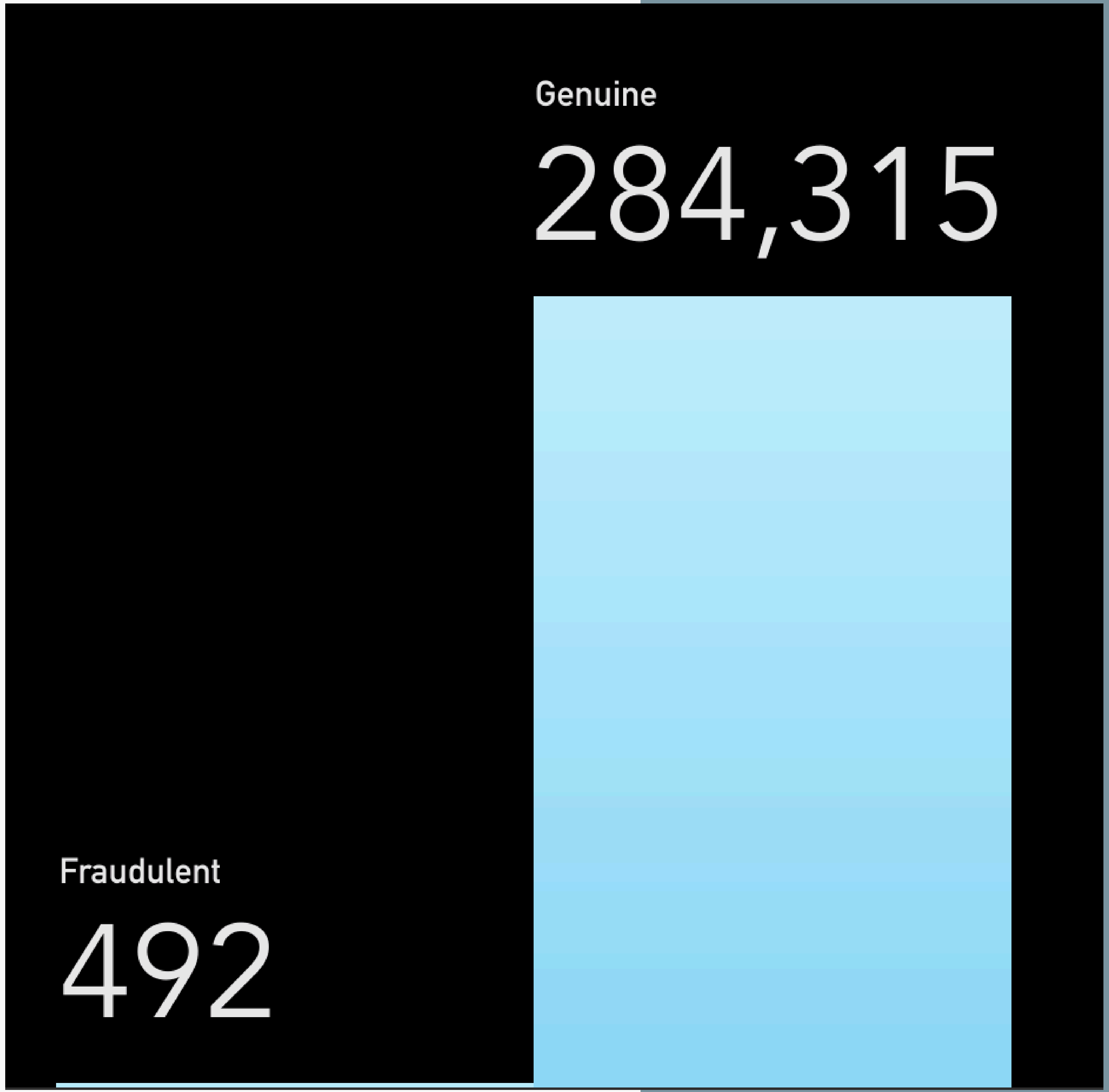
- An adaptive ML model to detect evolving fraud patterns in real time.

Background

Why This Matters:

- Fraud erodes customer trust and causes financial losses.
- Legacy systems fail to detect evolving fraud tactics

Our Data



Imbalanced Dataset:

The dataset comprises 284,807 credit card transactions. Within this dataset, the fraudulent transactions constitute 0.173% of all transactions, indicating a significant class imbalance.

Data Source:

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine

kaggle.com

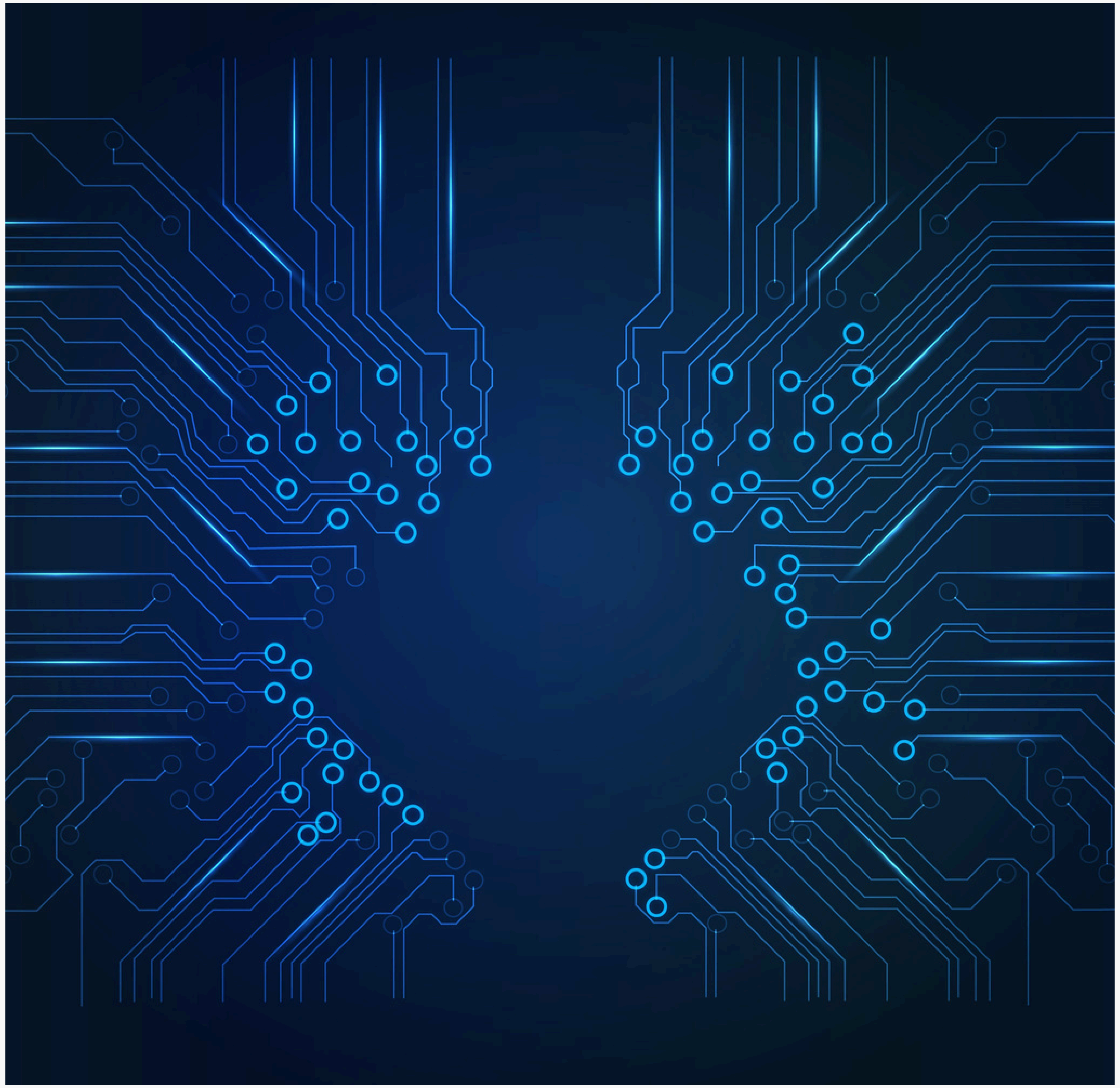
Data Preprocessing & Engineering

Preprocessing Steps:

Scaling: Normalization of Time and Amount features

Balancing: Application of SMOTE to counter class imbalance

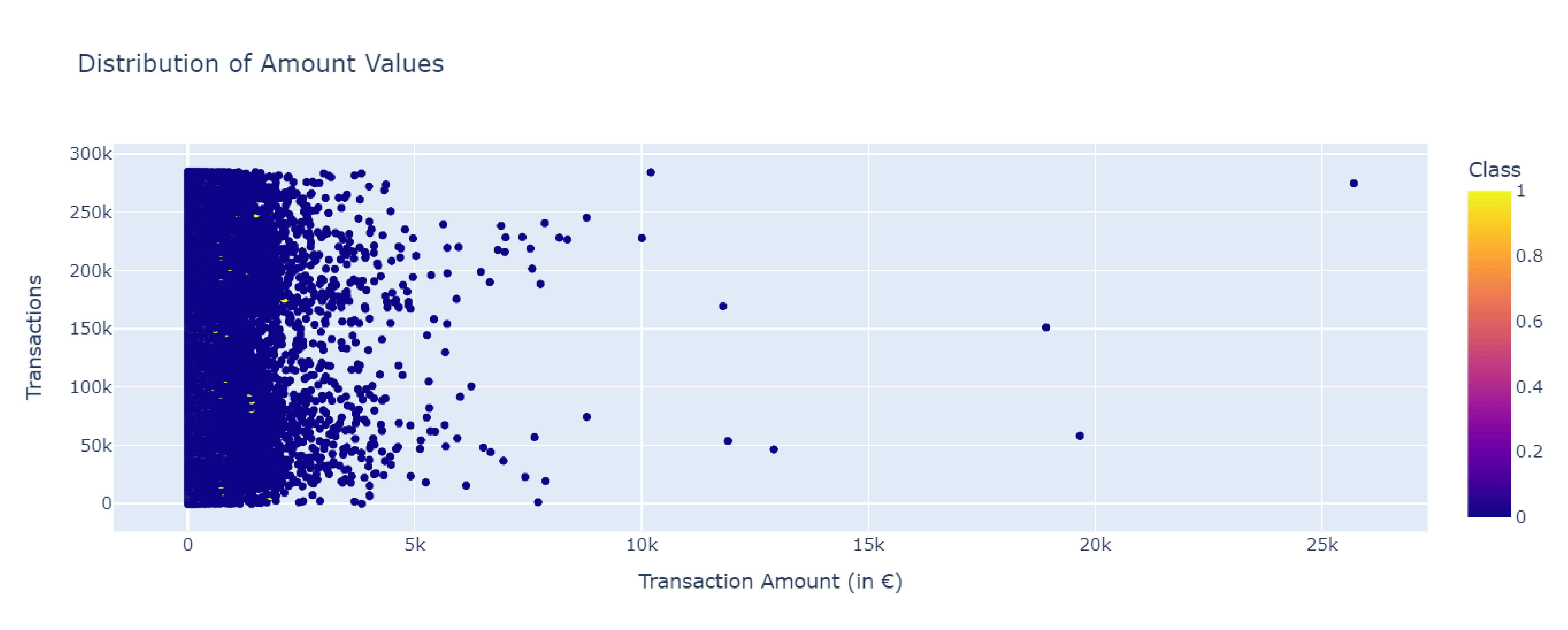
Splitting: Train-test split (80% training, 20% testing)



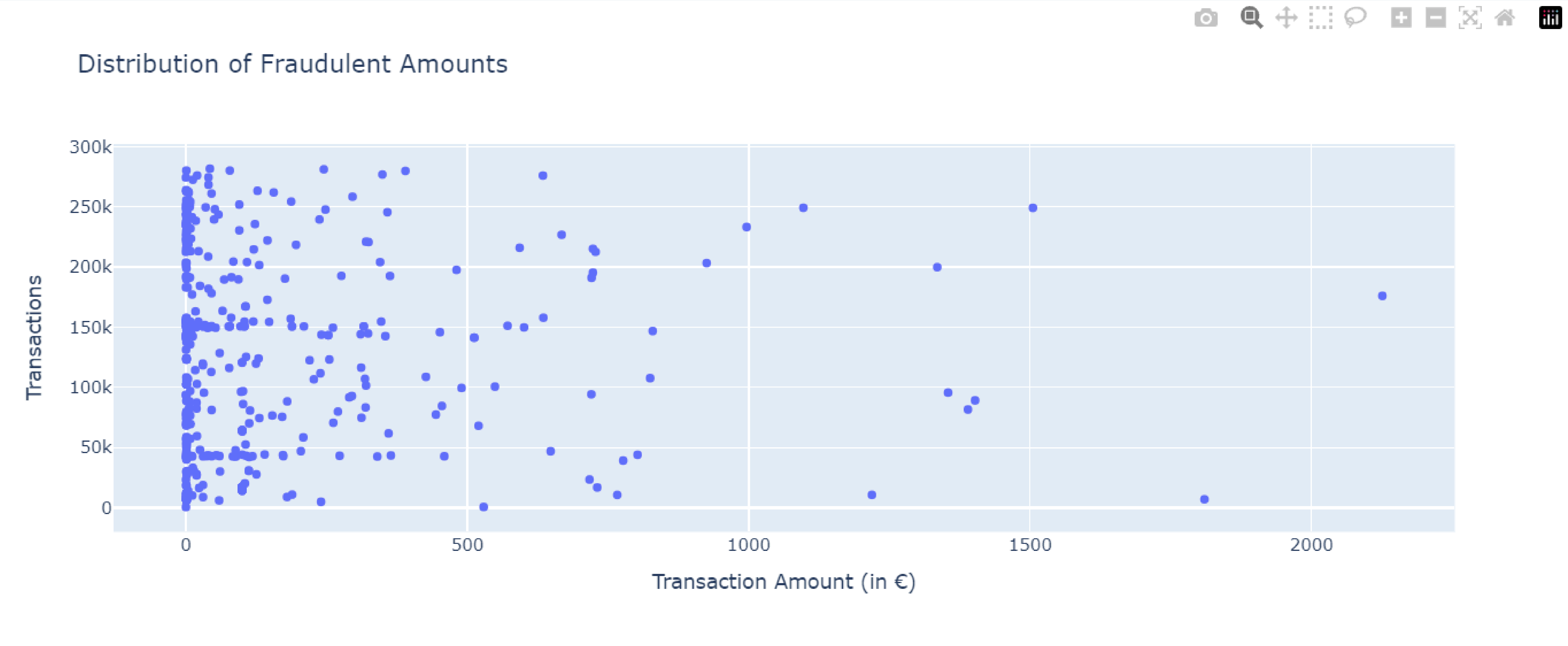
Feature Engineering:

Potential creation of new features from transaction metadata

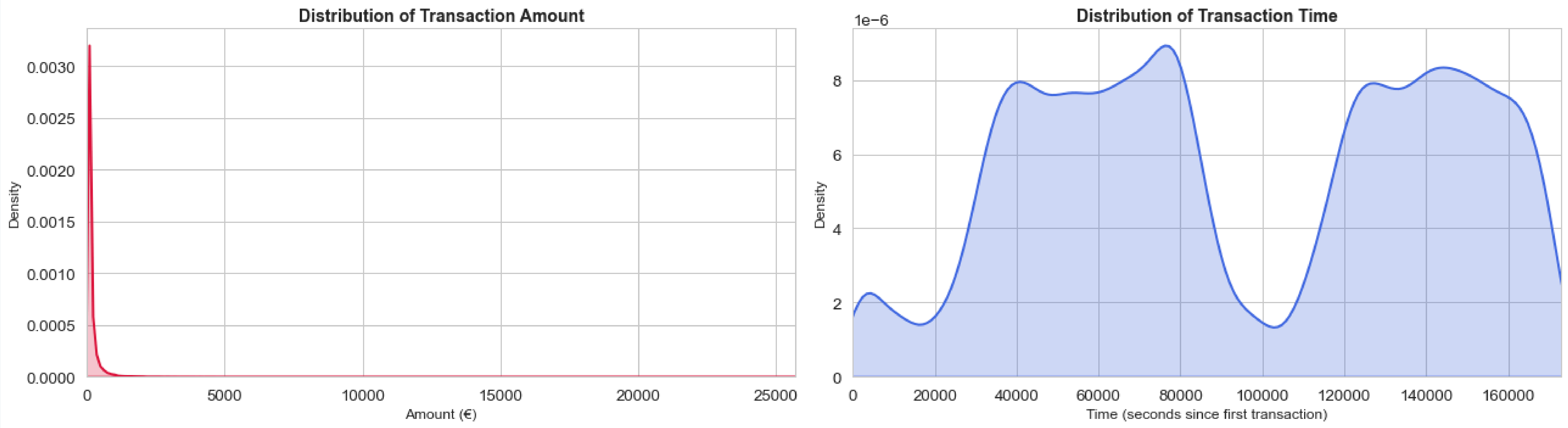
Data Analysis



Data Analysis



Data Analysis



Model Selection & Rationale

Models Evaluated:

- Logistic Regression: Serves as a baseline
- Random Forest: Delivers the highest AUC performance
- XGBoost: Provides flexible decision thresholds

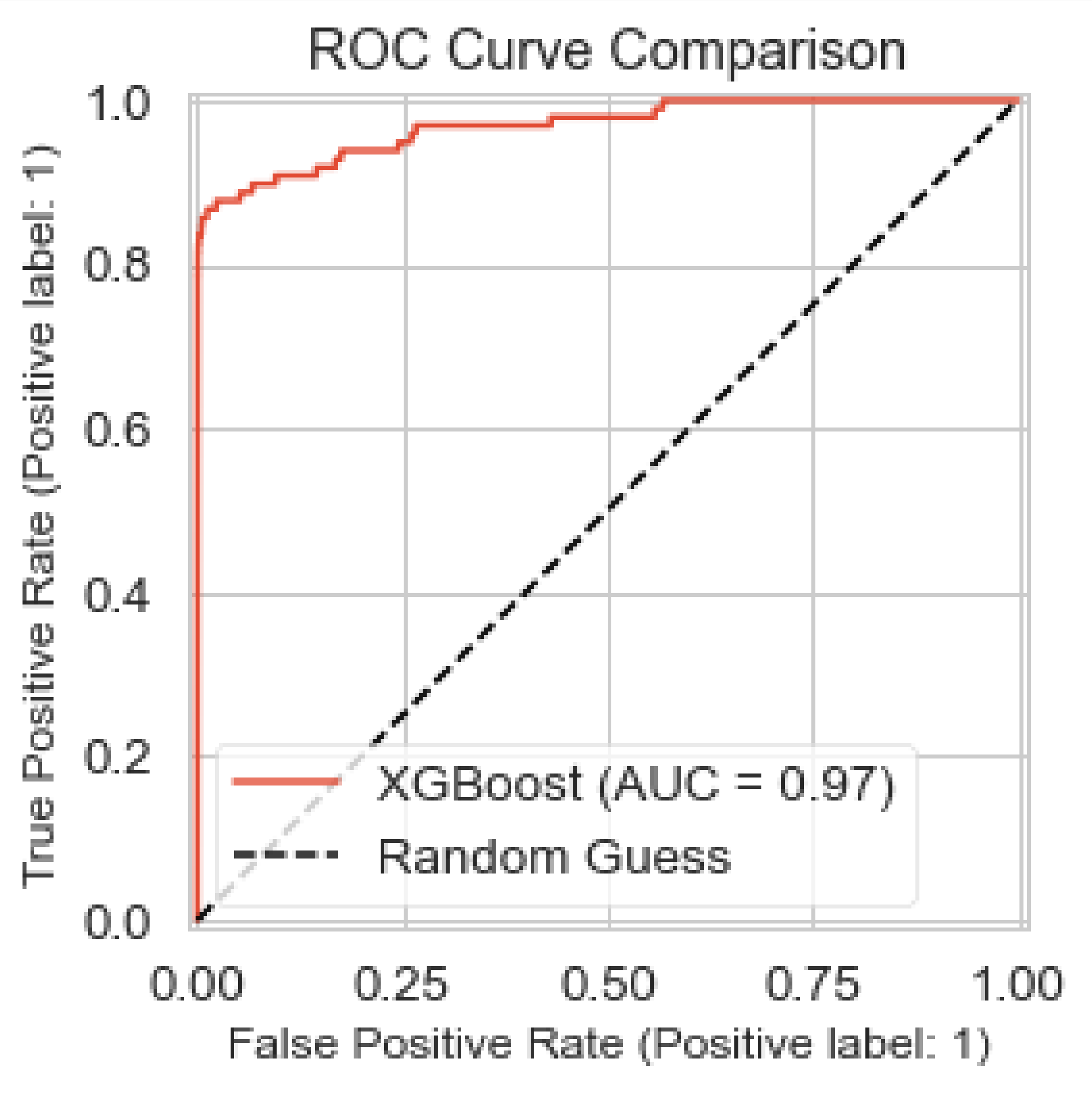
Rationale:

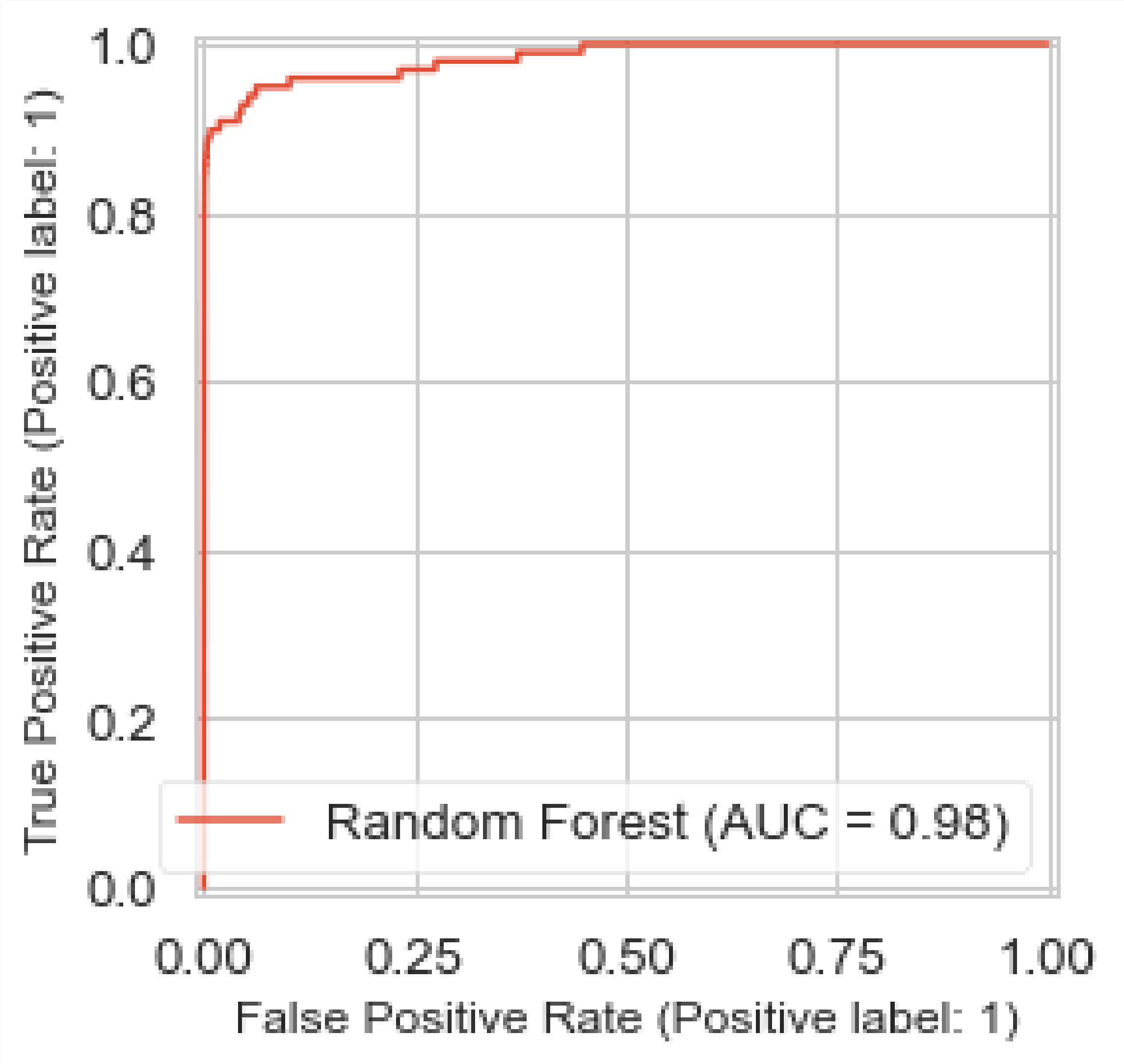
- Each model is chosen for its ability to handle imbalanced data (using class weights, SMOTE adjustments) and to provide interpretability (e.g., feature importance analysis)

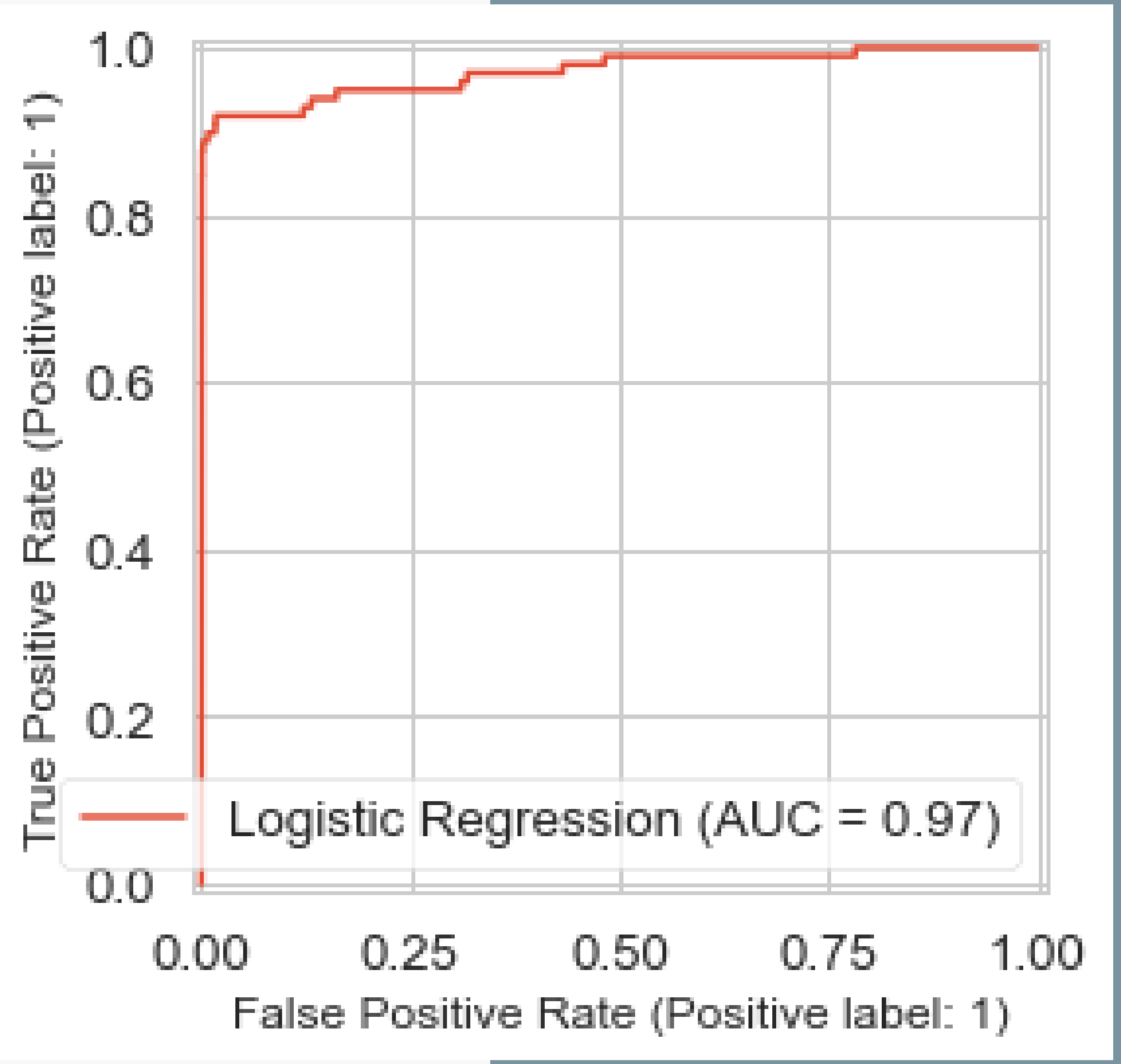
Performance Evaluation - ROC Analysis

Key Performance Metric: AUC (Area Under the ROC Curve)

- Random Forest: 0.98
- XGBoost: 0.97
- Logistic Regression: 0.97

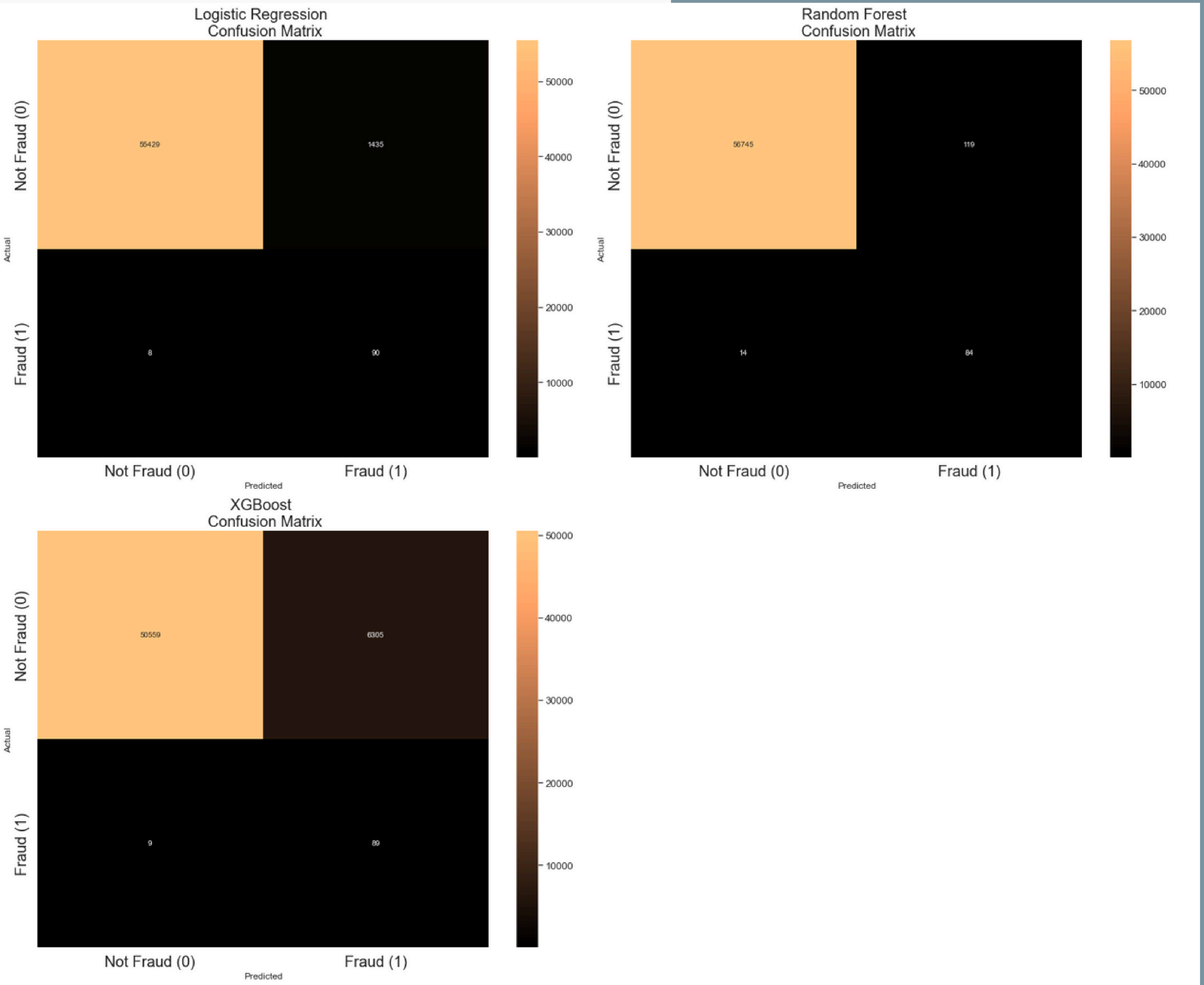






Performance Evaluation - Confusion Matrices

- Actual / Predicted → Fraud (1) | Not Fraud (0)
- Fraud (1) → True Positive (TP) | False Negative (FN)
- Not Fraud (0) → False Positive (FP) | True Negative (TN)



Conclusion

This project underscores the importance of addressing class imbalance and selecting context-appropriate metrics (AUC, recall) over accuracy in fraud detection. By deploying the Random Forest model, financial institutions can significantly reduce fraud-related losses while maintaining customer trust.

Thank you