Fraud Detection in Financial Transactions using Machine Learning

Author: Nifemi Kalejaiye-Matti

Date: August 2025

Financial fraud continues to pose a significant threat to banking and online transactions. Traditional rule-based detection systems struggle to adapt to new fraud techniques. This project demonstrates the use of machine learning models, Logistic Regression and Random Forest, to detect fraudulent financial transactions. By applying preprocessing, handling data imbalance, and evaluating model performance, this project highlights the potential of machine learning to improve fraud detection accuracy and reliability.

This report presents the dataset used, methodology, models implemented, results, and key insights, packaged in a portfolio-friendly format for practical demonstration.

Introduction

Fraudulent financial transactions are a growing concern for financial institutions, e-commerce platforms, and individuals. With billions lost annually to fraudulent activities, there is a pressing need for smarter and more adaptive fraud detection systems. Rule-based methods, while useful, are often rigid and fail to keep up with evolving fraud tactics. Machine learning offers a powerful alternative, capable of identifying hidden patterns in data and detecting fraud with greater accuracy.

The goal of this project is to develop and evaluate machine learning models for fraud detection using a real-world dataset. By comparing Logistic Regression and Random Forest, this project demonstrates the trade-off between interpretability and performance.

Dataset

The dataset used for this project is the Credit Card Fraud Detection dataset provided on Kaggle (by the Machine Learning Group, ULB). It contains anonymised financial transactions made by European cardholders in 2013.

Key characteristics:
- Total records: 284,807 transactions
- Fraudulent transactions: 492 (0.172%)
- Non-fraudulent transactions: 284,315

- Features: 30 input variables (mostly anonymised), transaction amount, and time
- Target variable: 'Class' (1 = Fraud, 0 = Genuine)

Methodology

The project followed a structured data science workflow:

1. Data Preprocessing
   - Handled missing values (if any)
   - Scaled/normalised numerical features
   - Addressed class imbalance using oversampling (SMOTE) and undersampling techniques

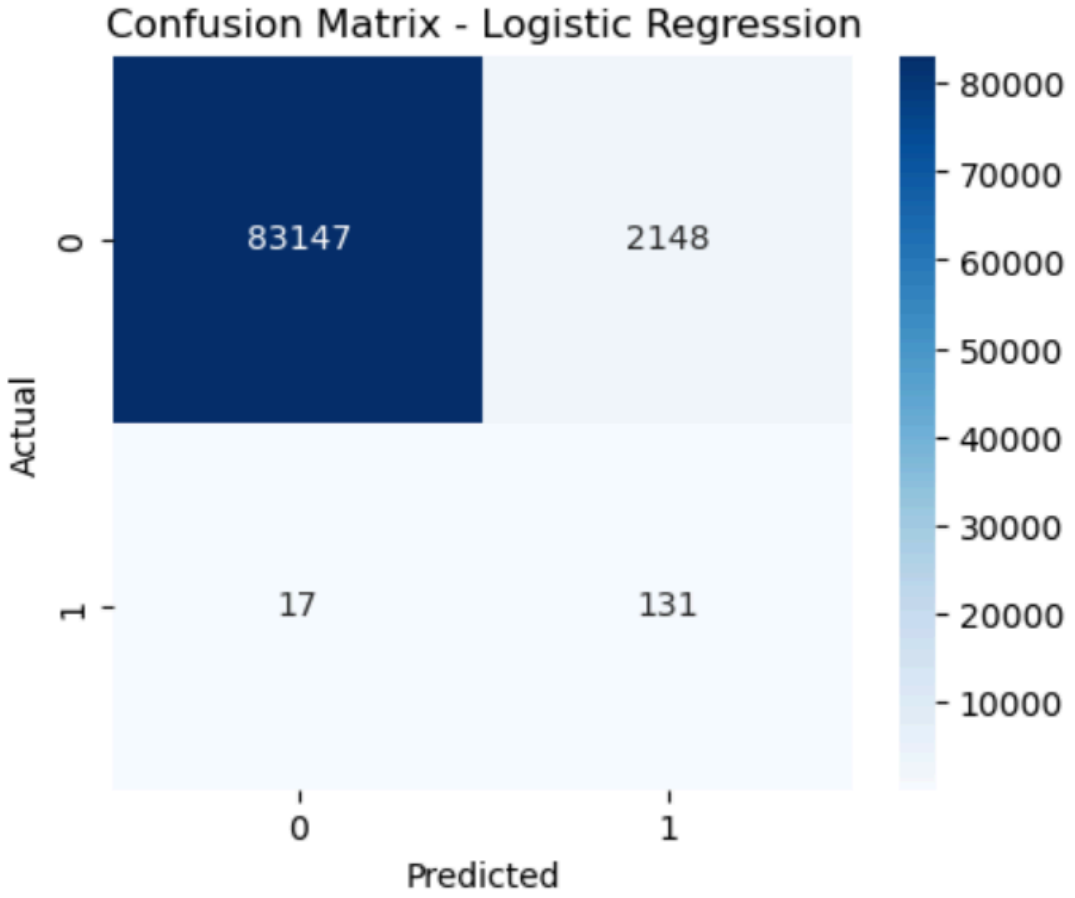
2. Model Selection
   - Logistic Regression: A linear, interpretable baseline model
   - Random Forest: An ensemble model combining decision trees for higher accuracy

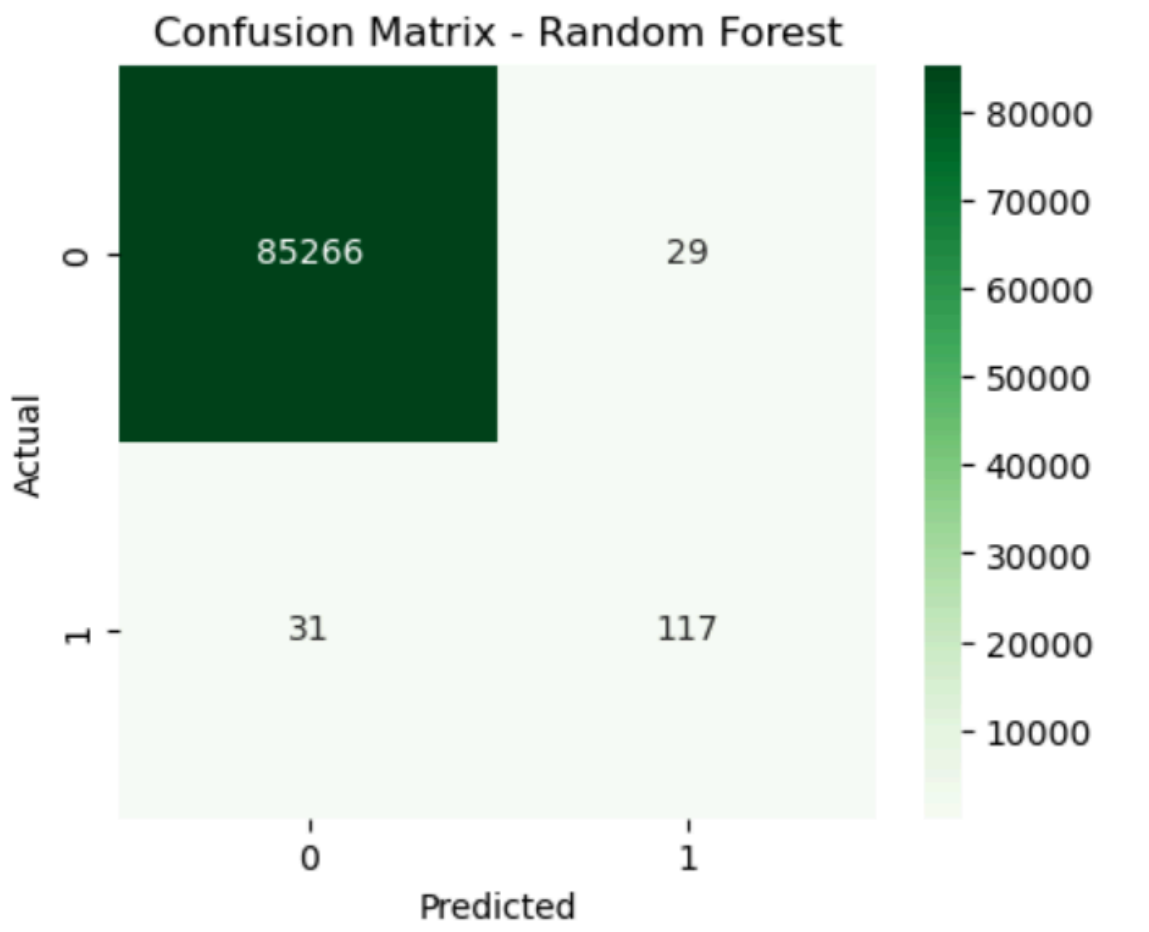
3. Evaluation Strategy
   - Accuracy
   - Precision, Recall, and F1-Score
   - Confusion Matrix
   - ROC-AUC Curve

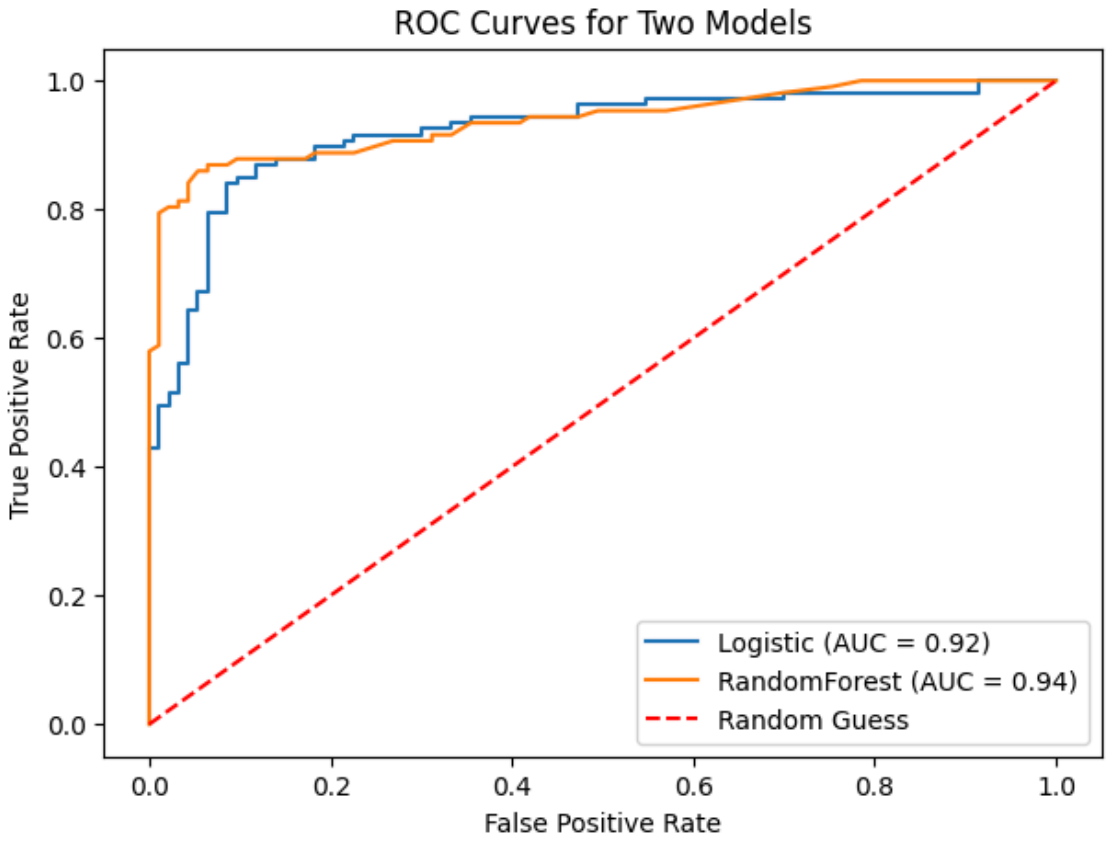
Results

Both models were trained and evaluated on the dataset. Key findings include:

- Logistic Regression: Achieved high accuracy but lower recall on the fraud class, meaning it struggled to detect some fraudulent cases.
- Random Forest: Outperformed Logistic Regression with better recall and F1-score, making it more effective for fraud detection.


Confusion Matrix - Logistic Regression
83147
2148
17
131
0
1
Actual
0
1
Predicted
80000
70000
60000
50000
40000
30000
20000
10000


Confusion Matrix - Random Forest
85266
29
31
117
0
1
Actual
0
1
Predicted
80000
70000
60000
50000
40000
30000
20000
10000



Discussion

The results highlight the importance of choosing the right model depending on business needs. While Logistic Regression is simple and easy to interpret, its lower recall may result in missed fraud cases, potentially costly for financial institutions. Random Forest, on the other hand, strikes a better balance between precision and recall, making it more suitable for high-stakes scenarios. However, it is more complex and less interpretable.

The dataset's imbalance was a key challenge. Without resampling, models were biased towards predicting 'non-fraud'. Techniques like SMOTE significantly improved the detection of minority fraud cases.

Conclusion

This project successfully demonstrated the application of machine learning in fraud detection. Random Forest emerged as the stronger model, offering improved detection of fraudulent transactions compared to Logistic Regression. The project emphasises the potential of machine learning to complement existing fraud detection systems.

Future improvements could include:
- Testing advanced algorithms such as XGBoost and LightGBM
- Deploying the model as a REST API for real-time fraud detection
- Applying explainable AI techniques to improve transparency for stakeholders
- Scaling solutions to handle big data in production environments

References

- Kaggle: Credit Card Fraud Detection Dataset
(https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud)
- Scikit-learn Documentation (https://scikit-learn.org/)
- Imbalanced-learn Documentation (https://imbalanced-learn.org/)