Executive Summary

Proactive Fraud Detection Using Machine Learning

Objective

The objective of this project is to design a proactive fraud detection system capable of identifying fraudulent financial transactions in a highly imbalanced, real-world dataset. The primary focus is to maximize fraud detection (recall) while maintaining operational feasibility, rather than relying on misleading accuracy metrics.

Dataset Overview

• Total transactions analyzed: ~6.3M+
• Fraud rate: ~0.13%
• Target variable: isFraud
• Data characteristics: large-scale, imbalanced, and transaction-level financial records

The dataset represents realistic challenges faced in banking and digital payment systems.

Data Preparation & Feature Engineering

• Verified no missing values, eliminating the need for imputation.
• Applied log transformation to transaction amounts to handle extreme outliers without removing legitimate fraud cases.
• Reduced multicollinearity by engineering balance change features instead of using raw account balances.
• Extracted time-based information to capture behavioral patterns in fraudulent activity.

These steps improved model stability, interpretability, and predictive power.

Modeling Approach

Three models were evaluated: Logistic Regression (baseline), Random Forest, and XGBoost.

XGBoost was selected due to its superior ability to handle severe class imbalance, capture complex non-linear fraud patterns, and optimize recall using class weighting (scale_pos_weight). A time-ordered train-validation split was used to simulate real-world deployment and prevent data leakage.

Model Performance

Evaluation focused on business-relevant metrics:
• Fraud Recall: ~90%
• Fraud Precision: ~20%
• ROC-AUC: ~0.997

Accuracy was intentionally deprioritized due to its ineffectiveness in imbalanced fraud detection problems.

Key Fraud Indicators

• Abnormal origin and destination balance changes
• High-risk transaction types (TRANSFER, CASH_OUT)
• Transaction amount anomalies
• Temporal transaction patterns

These indicators align with known financial fraud behaviors and confirm the model’s practical validity.

Business Impact & Prevention Strategy

The model enables early fraud detection and supports real-time transaction monitoring, risk-based authentication, and reduced financial loss through early intervention. Performance can be continuously monitored using fraud loss rate, recall stability, and false positive trends to ensure long-term effectiveness.

Conclusion

This project demonstrates a production-aligned, explainable, and scalable fraud detection framework. The approach balances technical rigor with business practicality and reflects industry-standard methodologies for handling large, imbalanced financial datasets.