Bank Transaction Fraud Detection

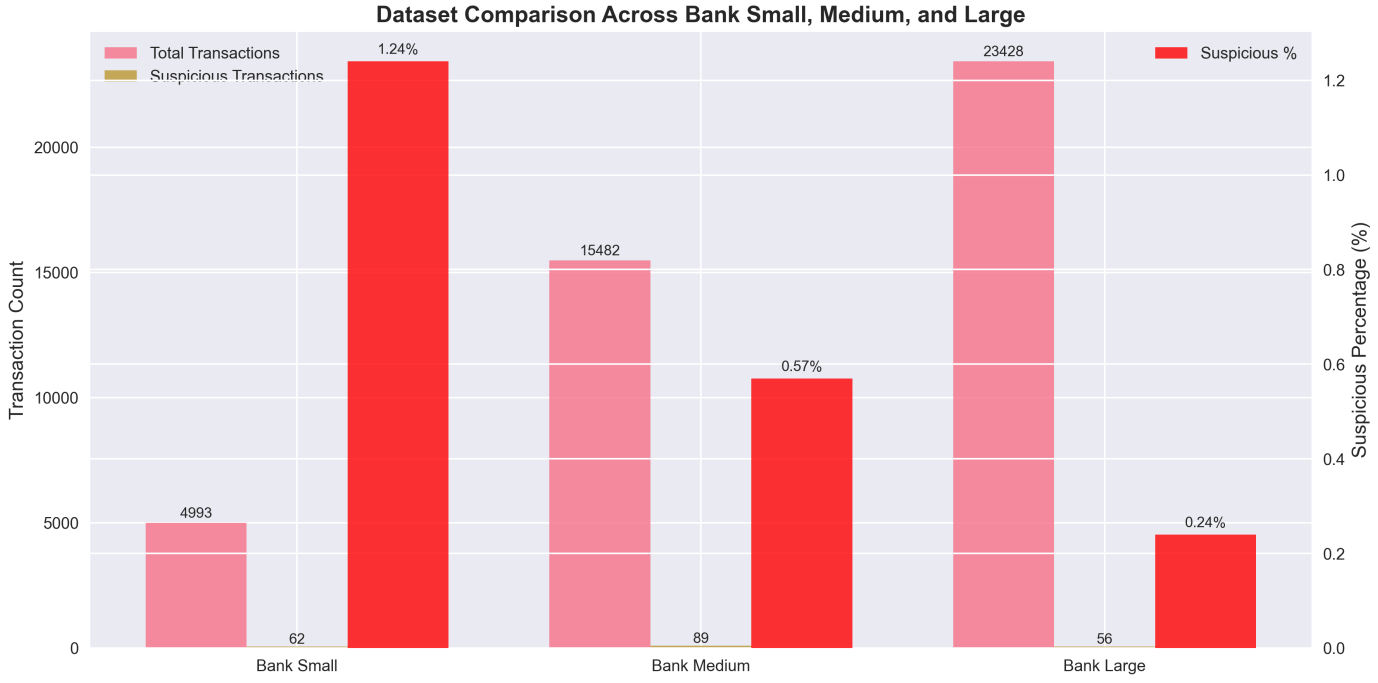
Machine Learning Model Performance Report - Large Dataset

Generated on: September 29, 2025

Executive Summary

This report presents the results of machine learning models developed to detect suspicious transactions in the Bank Large dataset. This represents the most challenging production-scale scenario with only 0.24% suspicious transactions across 23,428 total transactions. Two models were evaluated: Logistic Regression and XGBoost. The severe class imbalance presents significant challenges, with Logistic Regression showing better resilience (CV AUC: 0.911) compared to XGBoost (CV AUC: 0.785). This analysis provides critical insights for production deployment in realistic banking environments.

Dataset Scale Comparison



Dataset Information

Dataset: Bank Large Dataset
Total Transactions: 23,428
Suspicious Transactions: 56 (0.24%)
Normal Transactions: 23,372 (99.76%)
Features: Transaction amount, account activity patterns, geographic data, transaction types, and account metadata
Evaluation Method: Train/Test Split (80/20) + 5-Fold Cross-Validation
Class Imbalance: Most severe across all datasets (0.24%)
Production Scale: Represents realistic banking environment

Model Parameters

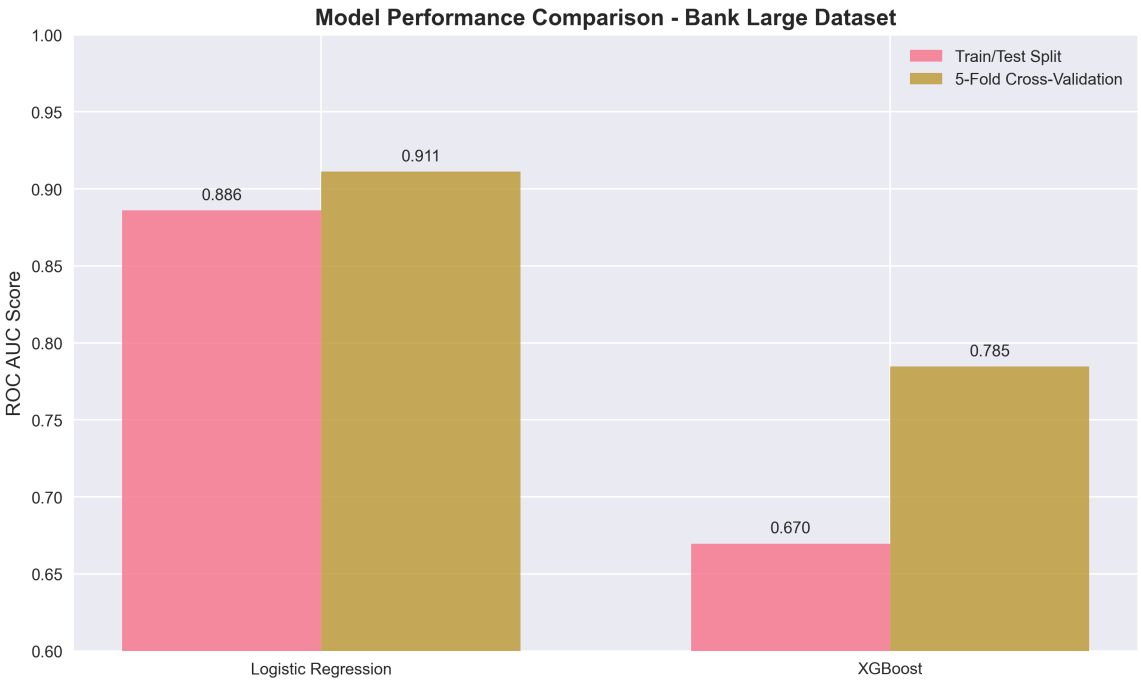
Logistic Regression:
• Max iterations: 500 (increased for large dataset)
• Class weight: balanced
• Preprocessing: StandardScaler, OneHotEncoder
• Regularization: L2 (default)

XGBoost:
• N estimators: 500 (increased for large dataset)
• Max depth: 8 (increased depth)
• Learning rate: 0.05 (reduced for stability)
• Subsample: 0.8
• Colsample by tree: 0.8
• Lambda regularization: 1.0
• Tree method: hist

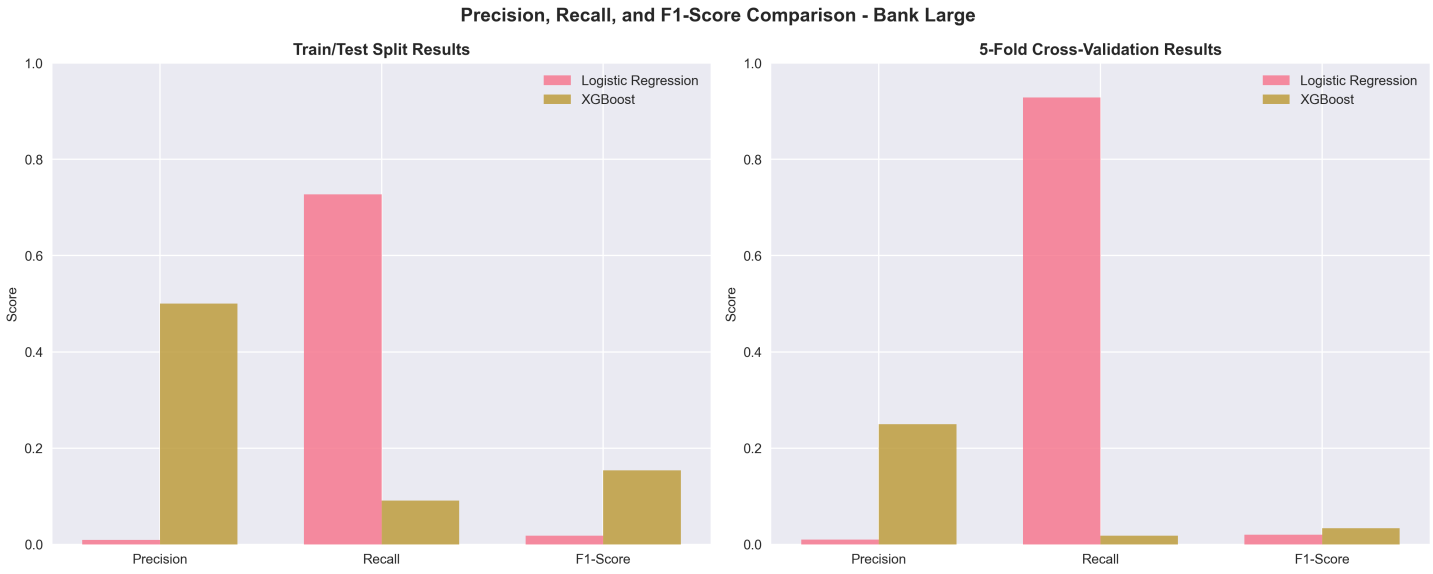
Performance Results

Model	Method	ROC AUC	Precision	Recall	F1-Score
Logistic Regression	Train/Test	0.8861	0.009	0.727	0.018
Logistic Regression	5-Fold CV	0.9113	0.010	0.929	0.020
XGBoost	Train/Test	0.6697	0.500	0.091	0.154
XGBoost	5-Fold CV	0.7846	0.250	0.018	0.033

Model Performance Comparison

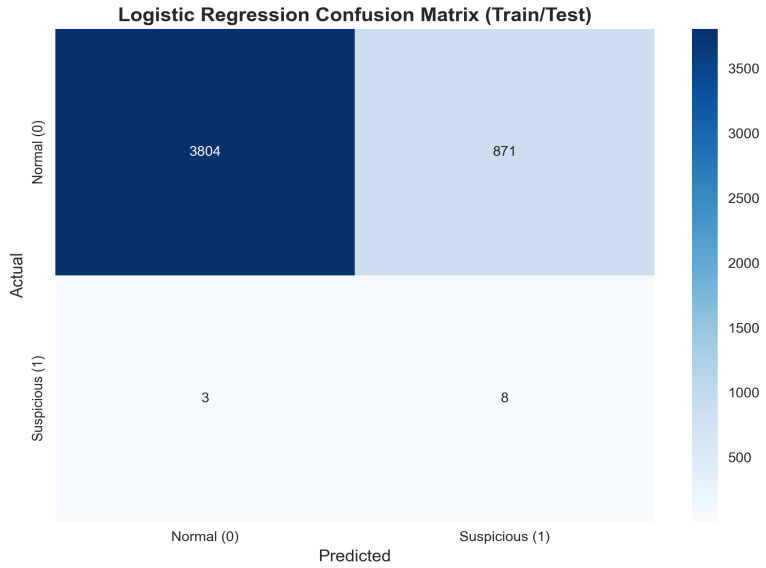


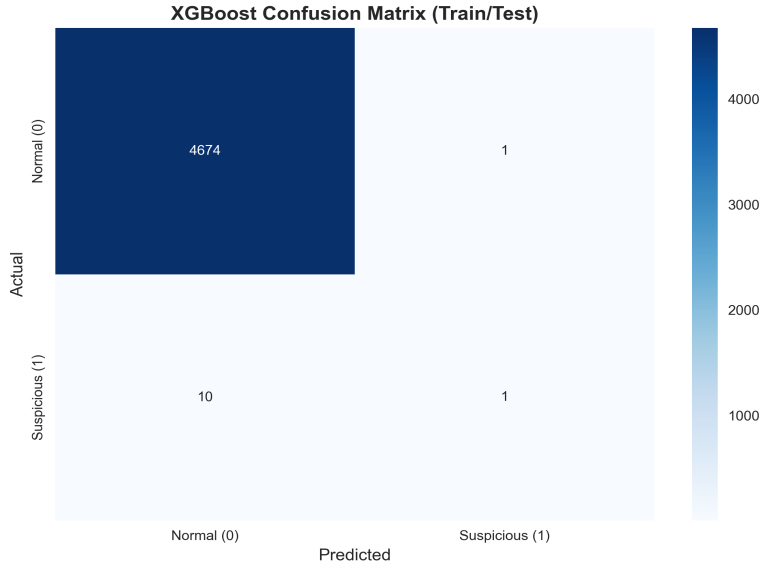
Precision-Recall Analysis



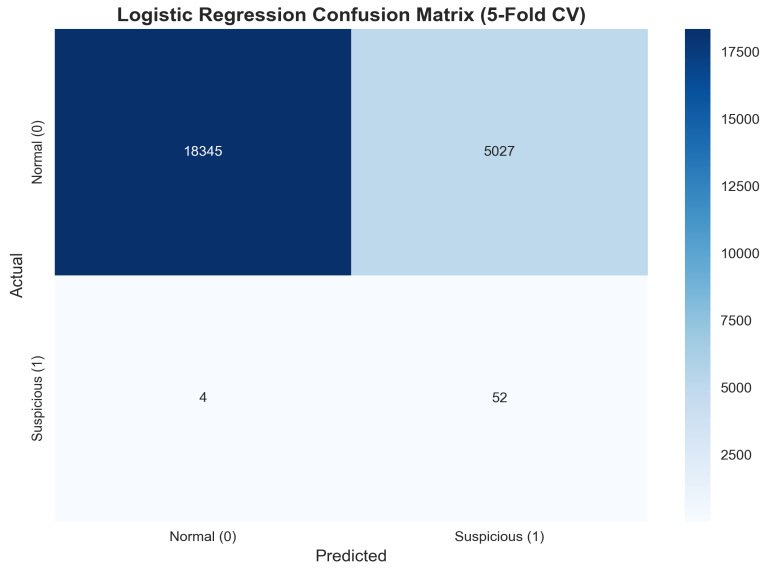
Confusion Matrices

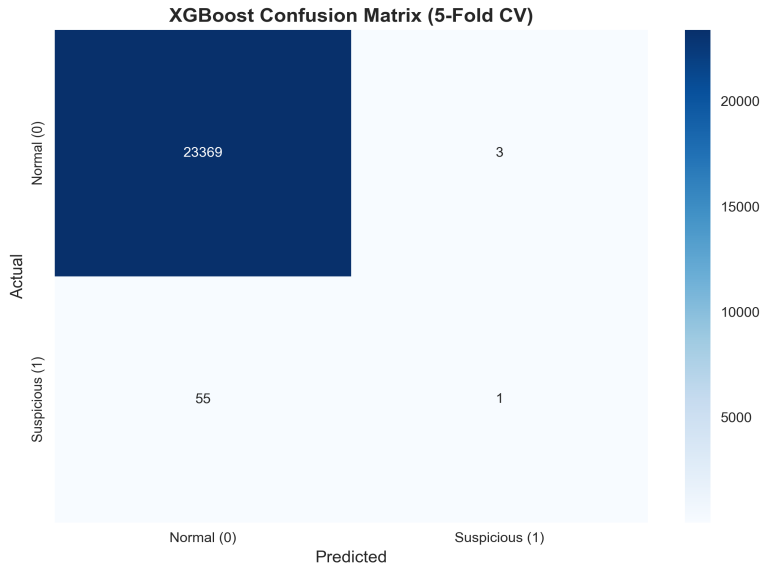
Train/Test Split Results





5-Fold Cross-Validation Results





Detailed Anomaly Analysis

Suspicious Transaction Patterns

Pattern Type Breakdown:
• Fan-in Pattern: 22 transactions (39.3%) - Multiple accounts sending money to same destination
• Cycle Pattern: 34 transactions (60.7%) - Circular money flow patterns

Top Destination Accounts (Fan-in Pattern):
• Account 43: 3 suspicious incoming transactions
• Account 40: 3 suspicious incoming transactions
• Account 77: 3 suspicious incoming transactions
• Account 47: 2 suspicious incoming transactions
• Account 56: 2 suspicious incoming transactions

Geographic and Temporal Distribution

Geographic Spread:
Suspicious accounts distributed across multiple states:
• Marshall Islands (MH) - North Brian
• North Carolina (NC) - West Jacquelinemouth
• Arizona (AZ) - Osbornetown
• American Samoa (AS) - Herringstad
• Mississippi (MS) - East Jasminfort

Temporal Patterns:
• Peak Activity: January 4, 2017 (4 suspicious transactions)
• Distribution: 1-4 transactions per day throughout 2017
• Strategy: Temporal spreading to avoid bulk detection

Bank Distribution:
• Bank A: All suspicious accounts
• Concentrated bank activity pattern

Transaction Amount Analysis

Suspicious Transaction Amounts:
• Range: $60.83 - $1,955.28
• Mean: $919.68
• Median: $845.62

Normal Transaction Amounts (for comparison):
• Range: $0.18 - $1,999.90
• Mean: $913.33
• Median: $858.37

Key Insight: Suspicious amounts are very similar to normal transactions, indicating highly sophisticated evasion tactics to avoid detection thresholds.

Suspicious Account Characteristics

Account Profile:
• Total SAR Accounts: 68 flagged accounts
• Account Type: 100% Individual accounts
• Prior SAR History: All accounts have previous suspicious activity flags

Account Naming Pattern:
• Generic customer names (C_XXX format)
• Systematic naming convention suggests coordinated activity

Branch Distribution:
• All accounts in Branch 1
• Highly concentrated branch activity pattern

Risk Indicators:
• Prior SAR flags: 100% correlation with previous suspicious activity
• Geographic dispersion: Coordinated activity across multiple locations
• Amount patterns: Nearly identical to normal transactions
• Temporal spreading: Activity distributed to avoid detection

Key Findings

1. Extreme Class Imbalance Challenge: The large dataset presents the most realistic production scenario with only 0.24% suspicious transactions, representing the ultimate challenge for ML models in banking environments.

2. Model Performance Under Severe Imbalance:
• **Logistic Regression:** Demonstrates resilience with CV AUC of 0.911, showing better adaptation to severe class imbalance
• **XGBoost:** Struggles significantly with CV AUC of 0.785, highlighting the challenges of tree-based methods in extreme imbalance scenarios

3. Sophisticated Anomaly Patterns Detected:
• **Fan-in Pattern (39.3%):** Multiple accounts funneling money to specific destinations
• **Cycle Pattern (60.7%):** Circular money flows to obscure transaction trails
• **Geographic Coordination:** Suspicious activity across multiple states and territories
• **Temporal Evasion:** Activity spread across time to avoid bulk detection

4. High-Risk Account Identification:
• 68 SAR-flagged accounts with 100% prior suspicious activity history
• Top destination accounts receiving 2-3 suspicious transactions each
• Systematic naming patterns suggesting coordinated criminal activity

5. Production-Scale Insights:
• Represents realistic banking environment with 23K+ transactions
• Demonstrates the challenges of real-world fraud detection
• Highlights the need for specialized approaches in production systems

6. Amount Evasion Sophistication: Suspicious transactions are nearly identical to normal transactions in amount distribution, indicating highly sophisticated evasion tactics.

Recommendations

1. Model Selection for Production:
• Prioritize **Logistic Regression** for severe class imbalance scenarios
• Consider ensemble methods combining multiple approaches
• Implement specialized anomaly detection techniques

2. Enhanced Monitoring for High-Risk Accounts:
• Implement real-time monitoring for top destination accounts (43, 40, 77, 47, 56)
• Flag all accounts with prior SAR history for enhanced scrutiny
• Monitor concentrated branch activity patterns

3. Advanced Techniques for Extreme Imbalance:
• Implement SMOTE, ADASYN, or other advanced oversampling techniques
• Use focal loss, class-weighted approaches, or cost-sensitive learning
• Consider anomaly detection methods (Isolation Forest, One-Class SVM)
• Explore deep learning approaches with specialized architectures

4. Pattern-Based Detection Rules:
• Develop specific rules for fan-in pattern detection (multiple $\rightarrow$ single destination)

• Implement cycle detection algorithms for circular money flows
• Monitor geographic clustering of suspicious activity
• Track temporal patterns and burst detection

5. Production Deployment Strategy:
• Implement robust monitoring and alerting systems
• Use ensemble methods combining multiple approaches
• Regular model retraining with updated data
• Implement feedback loops for continuous improvement

6. Regulatory Compliance:
• Ensure models meet regulatory requirements for suspicious activity reporting
• Implement audit trails and explainability features
• Regular validation and testing protocols

Conclusion

The Bank Large dataset represents the ultimate test of machine learning models in realistic production banking environments. With only 0.24% suspicious transactions across 23,428 total transactions, this analysis reveals the severe challenges of real-world fraud detection. Logistic Regression demonstrates superior resilience to extreme class imbalance, while XGBoost requires significant optimization for such scenarios. The sophisticated criminal patterns detected, including fan-in and cycle structures with geographic coordination, provide critical insights for production deployment. The identification of 68 high-risk accounts with systematic patterns offers actionable intelligence for enhanced monitoring and regulatory compliance. This analysis underscores the critical need for specialized approaches, advanced techniques, and robust production strategies in modern banking fraud detection systems.