FRAUD DETECTION - EXPLORATORY DATA ANALYSIS REPORT

DATASET OVERVIEW

Total Records: 10,000 transactions

Columns: transaction_id, amount, transaction_hour, merchant_category, foreign_transaction, location_mismatch, device_trust_score, velocity_last_24h, cardholder_age, is_fraud

Missing Values: None detected

Fraud Rate: 151 fraudulent cases (1.51%)

Legitimate Cases: 9,849 (98.49%)

KEY FINDINGS

Amount Distribution: Heavily right-skewed with long tail of high-value transactions. Fraudsters may use higher or lower amounts to avoid detection.

Merchant Categories: Transactions primarily in Grocery, Electronics, Travel, Clothing, and Food categories. Fraud patterns vary by category.

Class Imbalance: Only 1.51% fraud rate creates significant imbalance, requiring special handling during model development (SMOTE, class weights, stratified cross-validation).

Risk Indicators: Higher transaction velocity (multiple transactions in 24h) and foreign transactions combined with location mismatch show elevated fraud risk.

Device Trust Score: Fraudsters exhibit lower average device trust scores, indicating unrecognized devices or suspicious login patterns.

RECOMMENDED ANALYSES AND ACTIONS

Feature Engineering:

- Rolling velocity metrics (7-day, 30-day transaction counts)
- Merchant risk scoring and historical fraud rates by merchant
- Geospatial distance between billing address and transaction location
- Time-since-last-transaction by cardholder
- Cardholder spending pattern baseline (mean, std dev, categories)

Modeling Approach:

- Use class-weighted classifiers (Logistic Regression, Random Forest, XGBoost with scale_pos_weight)
- Alternative: SMOTE for synthetic oversampling of minority class
- Cross-validation: Stratified K-Fold to maintain class distribution in folds
- Optimization: Cost-sensitive learning with business cost matrix (false positive cost vs false negative cost)

Monitoring and Maintenance:

- Implement data drift detection using Evidently or similar tools
- Set up automated alerts when feature distributions shift significantly
- Schedule weekly EDA reports to catch seasonal patterns
- Monitor model performance (precision, recall, AUC-ROC) over time

BUSINESS IMPACT

Cost Analysis:

- Chargeback cost per fraudulent transaction: typically $25-100 (fraud + processing)
- Customer friction cost from false positives: account lockouts, declined transactions, support escalations
- Balance needed: minimize fraud while maintaining customer experience

Operational Strategy:

- Tune decision threshold based on cost matrix, not default 0.5 probability
- Use precision-recall trade-off tailored to business requirements
- Prioritize high-precision features (device trust, velocity, flags) for real-time scoring
- Reserve lower-precision detections for human review via case management system

Risk Reduction:

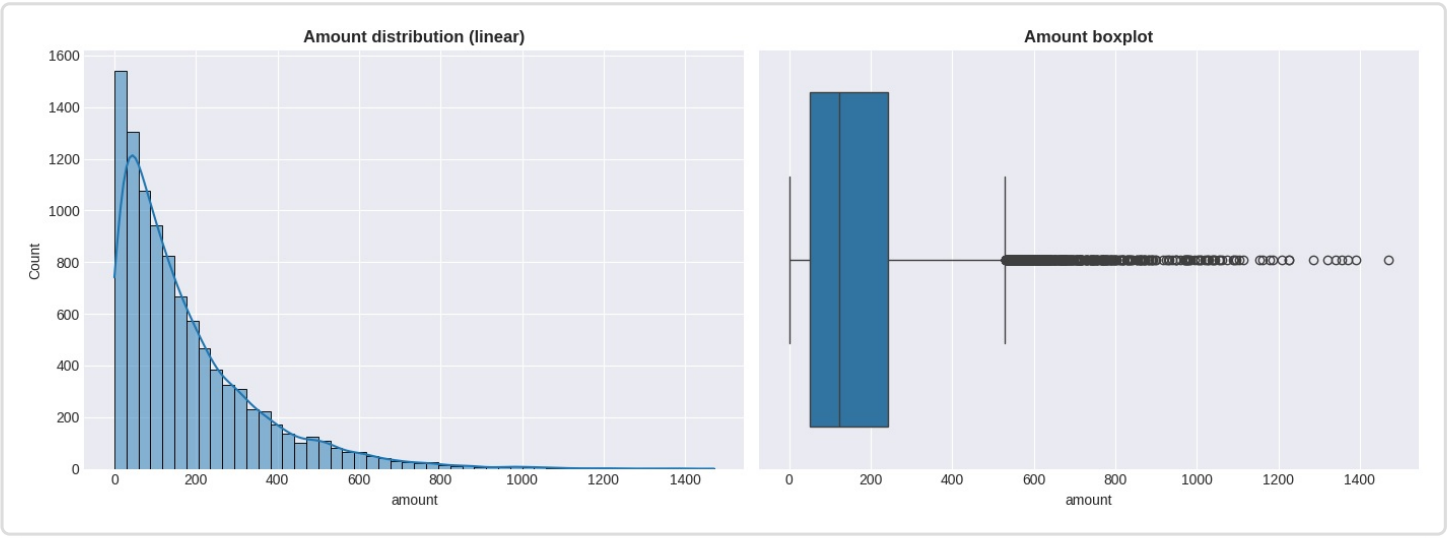
- Implement multi-layered detection: rule-based alerts + ML model + human review
- Focus on velocity and device anomalies as primary signals
- Flag geographic anomalies (location mismatch) for enhanced authentication

NEXT STEPS

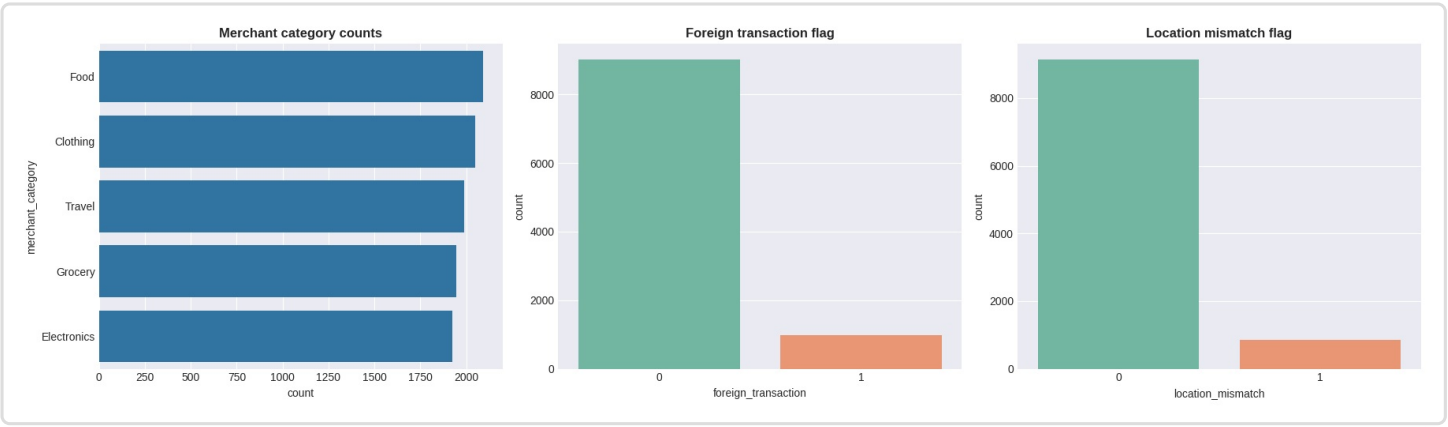
1. Build baseline models (logistic regression, random forest) and compute cost-weighted metrics
2. Perform statistical significance testing on feature-fraud associations
3. Generate SHAP explanations for high-impact features to improve interpretability
4. Develop feature importance rankings and domain validation
5. Create monitoring dashboards and automated alerting system
6. Establish feedback loop to retrain models monthly with new transaction patterns

VISUALIZATIONS

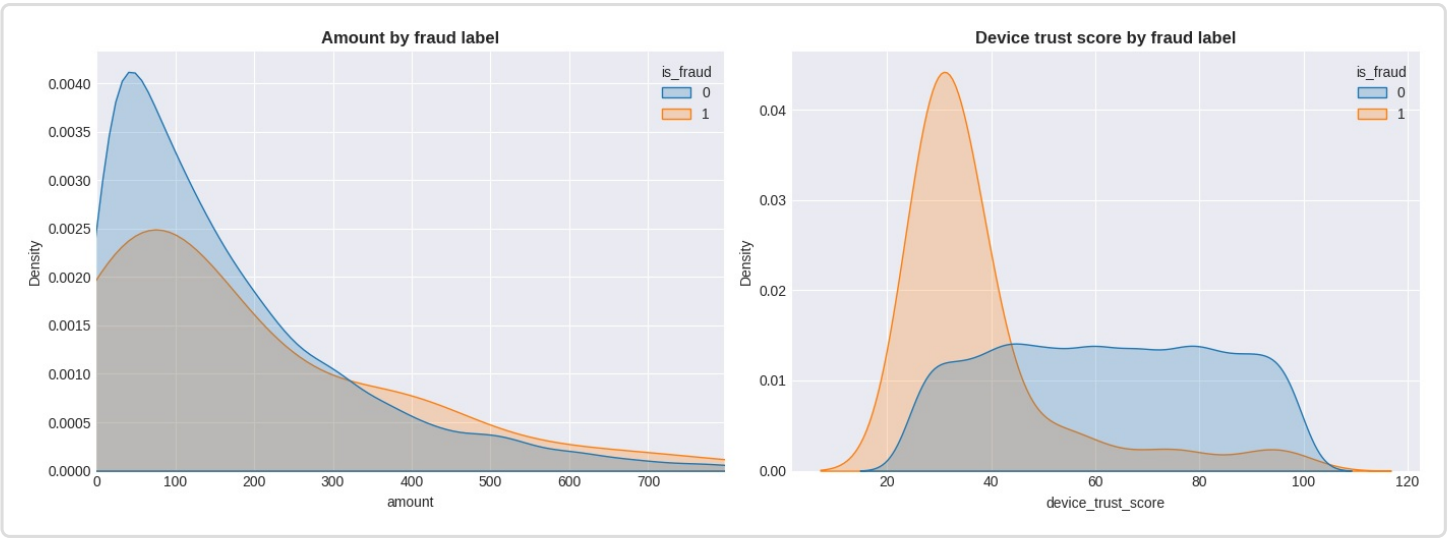
Amount Dist



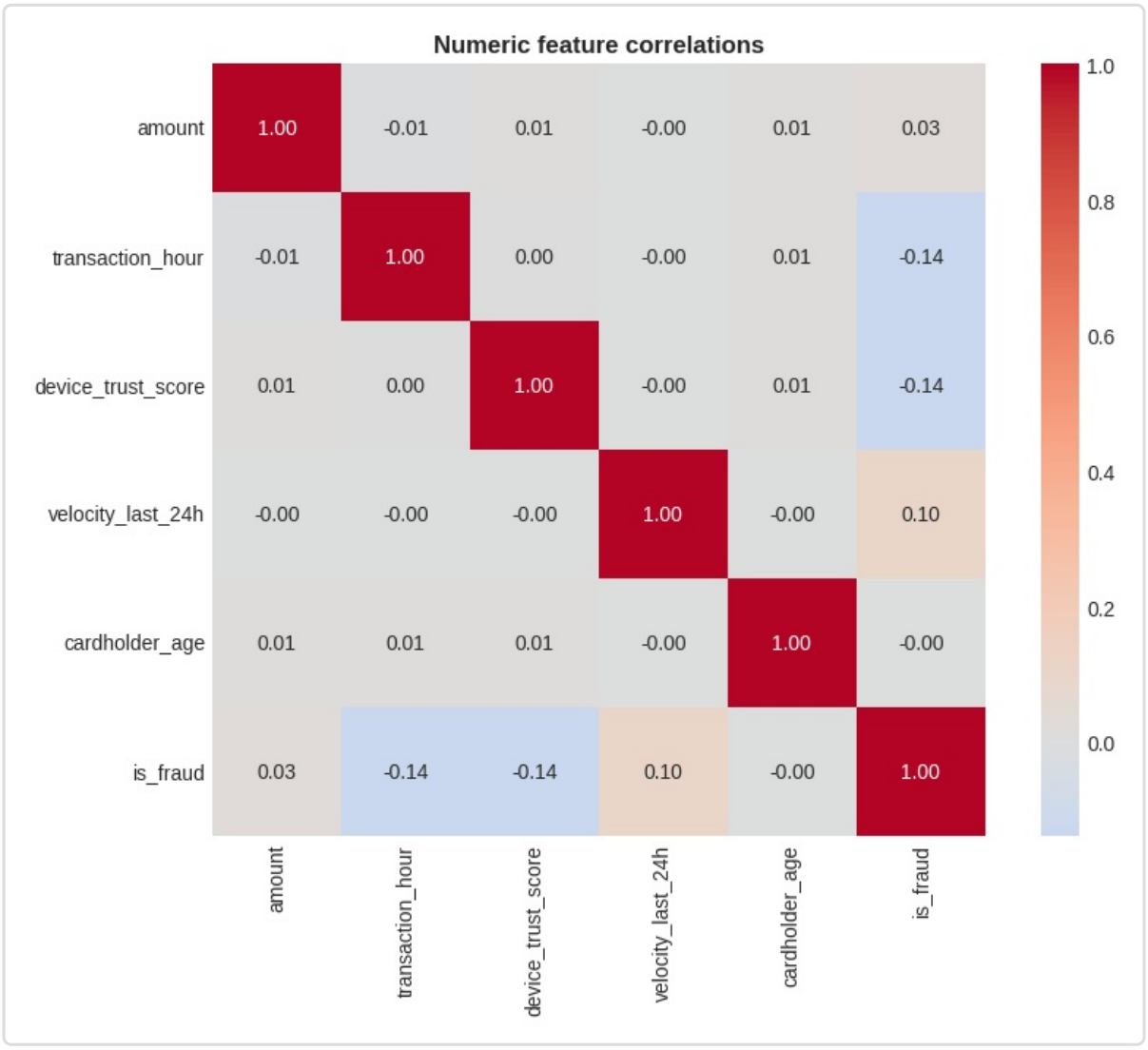
Categorical Dist



Fraud Features



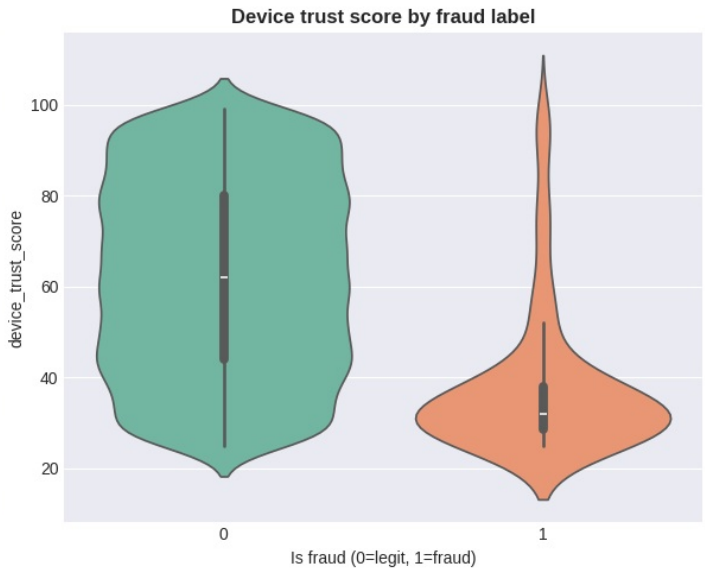
Correlation Heatmap



Hourly Fraud Rate

Velocity Device Analysis


Velocity (last 24h) by fraud label
velocity_last_24h
0
2
4
6
8
0
1
Is fraud (0=legit, 1=fraud)


Device trust score by fraud label
device_trust_score
20
40
60
80
100
0
1
Is fraud (0=legit, 1=fraud)