FRAUD TRANSACTION DETECTOR

Project Overview

This is a machine learning application designed to detect potentially fraudulent financial transactions. The system analyzes transaction patterns and flags suspicious activity based on historical data patterns.

What It Does

The system takes transaction details as input and predicts whether the transaction is likely to be fraudulent. It considers multiple factors including:

- Transaction type and amount
- Account balance changes
- Balance consistency checks
- Historical transaction patterns

How It Works

1. **Data Collection**: The model was trained on historical transaction data containing both legitimate and fraudulent transactions
2. **Feature Engineering**: The system calculates derived features like balance differences and consistency checks
3. **Model Prediction**: A Random Forest classifier analyzes the transaction against learned patterns
4. **Risk Assessment**: Returns a fraud probability score and classification

Technical Components

The project consists of three main parts:

1. **Training Script** (code1.PY):
    - Loads and prepares transaction data
    - Trains the machine learning model

- Saves the trained model and preprocessing objects

2. **Prediction Script** (code2.PY):
    - Loads the saved model
    - Takes user input for transaction details
    - Makes predictions on new transactions
    - Displays results in console
3. **Streamlit Web Application** (app.py):
    - Provides a user-friendly web interface
    - Allows interactive transaction checking
    - Displays results with clear formatting

MY ANALYSIS

DATA OVERVIEW INSIGHTS

1. **Massive dataset**: 6.36 million transactions
2. **Highly imbalanced**: Only 8,213 fraud cases (0.13% fraud rate)
3. **No missing values**: Clean dataset ready for analysis
4. isFlaggedFraud **is useless**: Only 16 flagged cases (0.00025%) - can be ignored

TRANSACTION TYPE INSIGHTS

Fraud rates by type:

CASH_IN: 0.000000% fraud

DEBIT: 0.000000% fraud

PAYMENT: 0.000000% fraud

CASH_OUT: 0.184000% fraud

TRANSFER: 0.768800% fraud

Key finding: Fraud only happens in **CASH_OUT** and **TRANSFER** transactions! This is a critical insight.

AMOUNT BINNING INSIGHTS

Amount bins and fraud rates:

$0-$5,711: 0.0216% fraud

$5,711-$13,389: 0.0204% fraud

$13,389-$30,078: 0.0431% fraud

$30,078-$74,871: 0.0917% fraud

$74,871-$135K: 0.0938% fraud

$135K-$208K: 0.0826% fraud

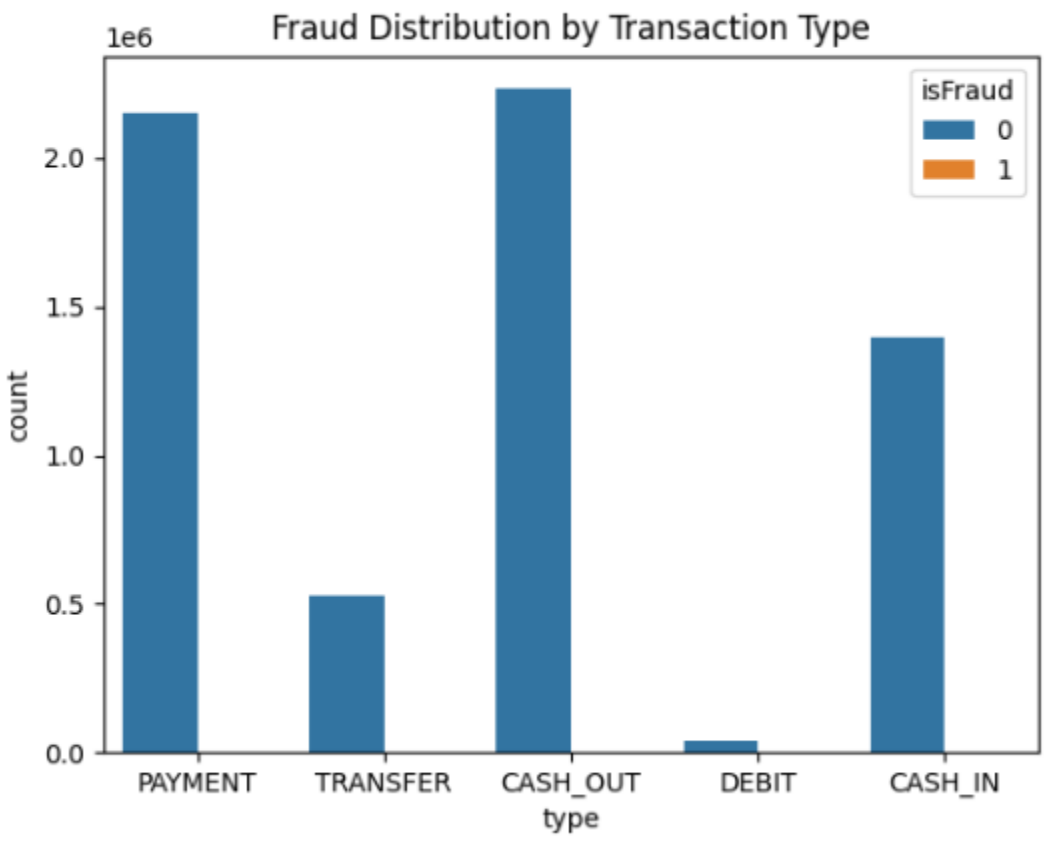
$208K-$325K: 0.0928% fraud

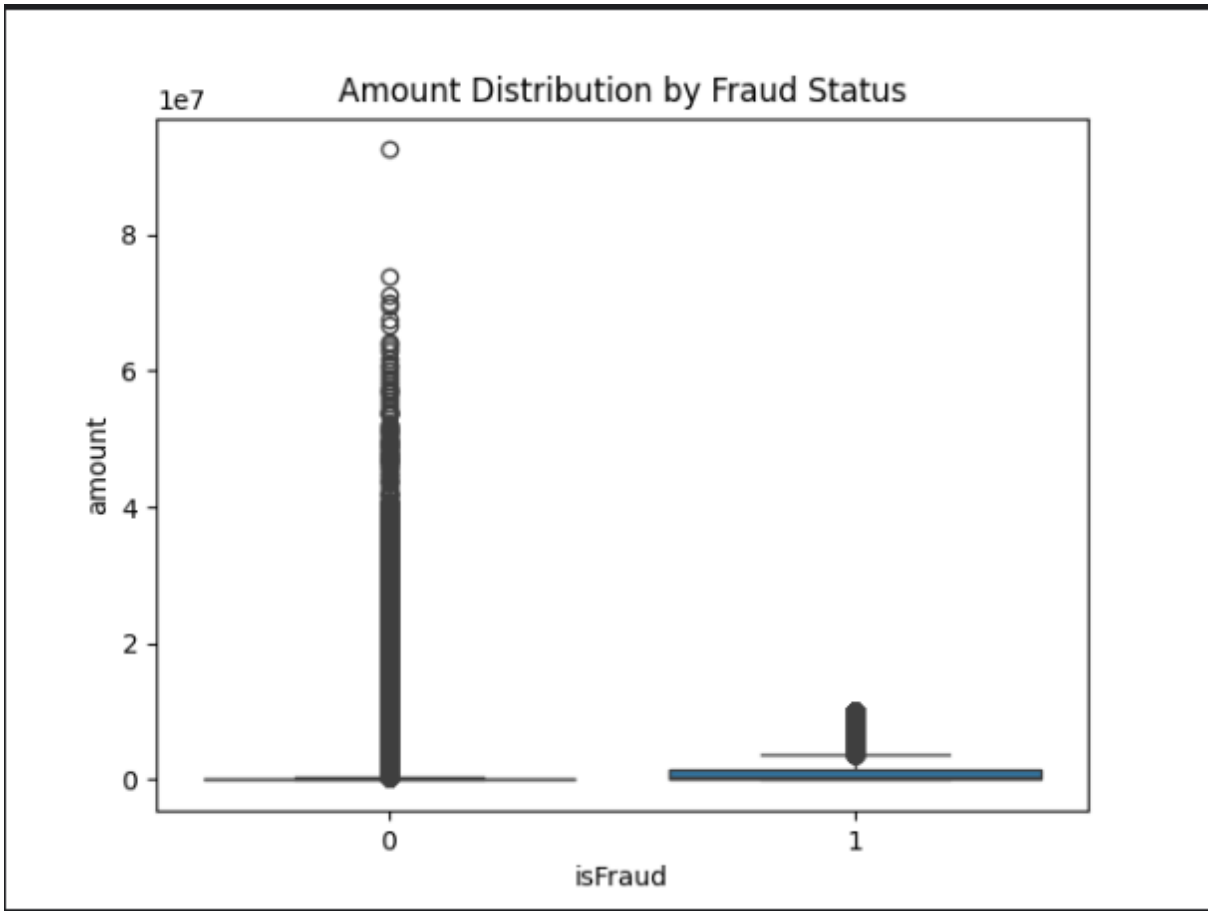
$325K-$92M: 0.5867% fraud ← HIGHEST RISK

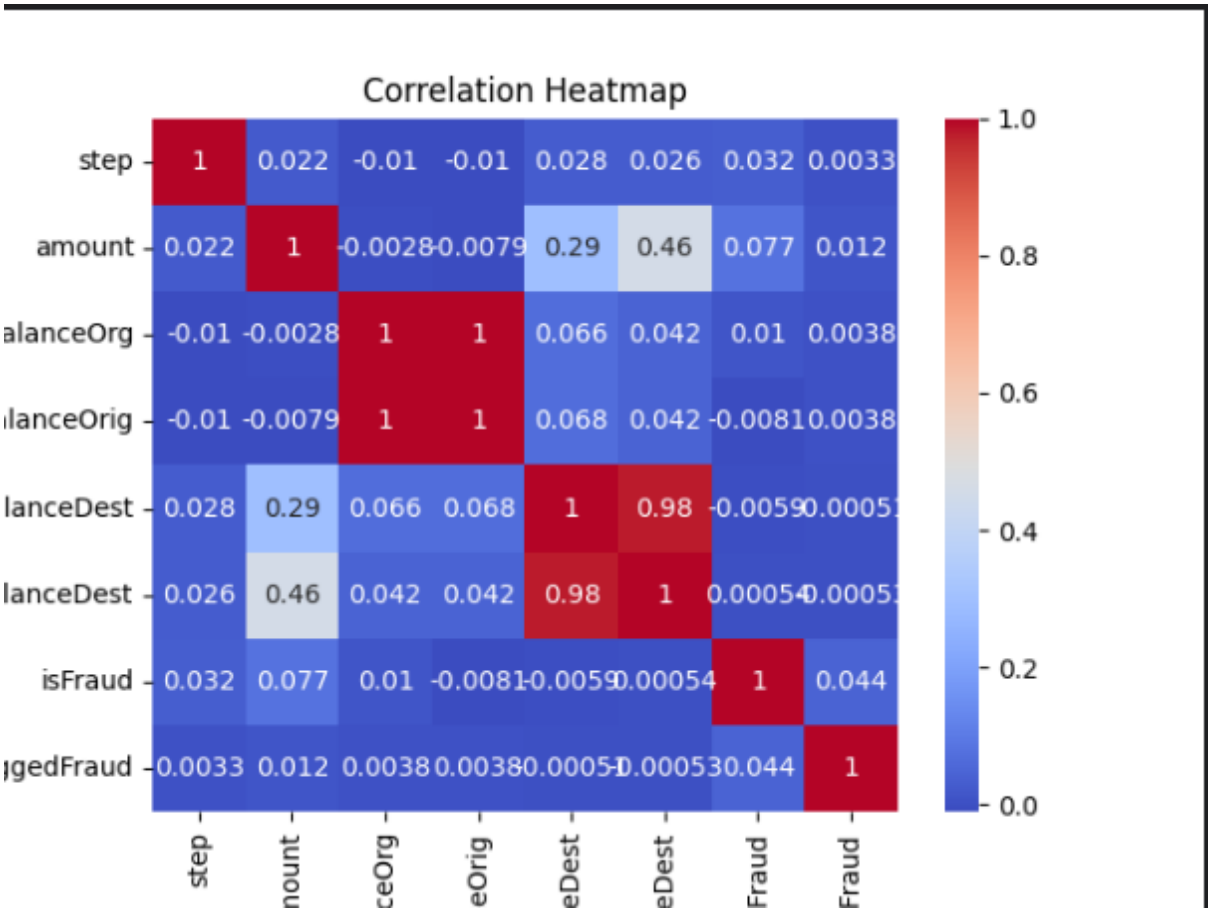
Key findings:

1. Fraud increases with transaction amount
2. Very high amounts (>$325K) have **27x more fraud** than small amounts
3. Medium amounts have consistent low fraud rates.

GRAPHS:




Amount Distribution by Fraud Status
1e7
8
6
4
2
0
amount
0
1
isFraud



FEATURE ENGINEERING INSIGHTS

1. Balance Error Analysis

balance_error vs fraud:

Small errors (False): 0.8524% fraud rate

Large errors (True): 0.0023% fraud rate ← Counterintuitive!

Insight: Transactions with small balance errors are MORE LIKELY to be fraudulent. Large errors might be data issues rather than fraud.

2. Negative Balance Feature

neg_balance_orig distribution:

True: 4,079,080 transactions (64%)

False: 2,283,540 transactions (36%)

Fraud rates:

neg_balance_orig=False: 35.84% fraud rate ← EXTREMELY HIGH

neg_balance_orig=True: 0.07% fraud rate

CRITICAL INSIGHT: When neg_balance_orig=False (account has enough money), fraud rate is **35.84%**! This is 275x higher than the average!

Key patterns:

1. **Fraudulent transactions**:
    - Much larger orig_diff ($1.46M vs -$23K)
    - Smaller balance_error (-$10K vs -$201K)
    - 8x larger transaction amounts

MODEL PERFORMANCE

MODEL ALGORITHM USED

Random Forest Classifier

MODEL PERFORMANCE INSIGHTS

Confusion Matrix:

Non-Fraud (0): 1,270,424 correct, 457 false positives

Fraud (1): 1,197 correct, 446 false negatives

Performance Metrics:

Precision (fraud): 72% (72% of predicted fraud is actual fraud)

Recall (fraud): 73% (73% of actual fraud is caught)

F1-Score (fraud): 73%

ROC AUC: 88.8%

Model strengths:

1. **Excellent at identifying non-fraud**: 99.96% accuracy
2. **Good fraud detection**: Catches 73% of fraud cases
3. **Low false positives**: Only 457 normal transactions flagged as fraud