Statistical Findings in Fraud Detection

Transaction Amounts Are Skewed

Most transaction amounts are small, but a few are extremely large. This skew required log transformation to reveal hidden structure.

Fraudulent Transactions Concentrate in Specific Types

The majority of fraud cases occurred in 'TRANSFER' and 'CASH_OUT' transactions. Other types had very few or no frauds.

Engineered Features Capture Fraud Patterns

We created 'orig_diff' and 'dest_diff' as balance difference features. Fraudulent transactions showed higher values in these columns.

Log Transformation Improved Feature Distributions

Applying np.log1p() to skewed columns like amount, orig_diff, and dest_diff helped normalize them and improved model interpretability.

Imbalance in Target Labels

Only about 0.1 percent of transactions were fraud. Accuracy was misleading, so we focused on precision, recall, and F1-score.

Mismatch Between isFlaggedFraud and isFraud

The built-in fraud flag rarely aligned with actual fraud. This supported the need for a smarter ML-based model.