FINANCIAL FRAUD DETECTION & RISK SCORING

End-to-end SQL + Python + ML pipeline for detecting fraudulent financial transactions

1. Built on a 1,000,000-row curated dataset
2. Full data pipeline: SQL cleaning → feature engineering → ML
3. Combines rule-based, unsupervised, and supervised approaches
4. Outputs fraud probability + risk category for each transaction
5. Designed to mimic a real fintech fraud analytics workflow

Shivkanya Balamurugan

SQL DATA CLEANING & RULE FLAGS

- Loaded raw PaySim dataset into MySQL (2.5M rows)
- Applied rule-based anomaly detection entirely in SQL:
  - Missing/invalid amount checks
  - Negative balance detection
  - Ledger mismatch (sender/receiver balance rules)
  - Zero-change balance behavior
  - Overdraft anomaly (amount > balance)

https://www.kaggle.com/code/kartik2112/fraud-detection-on-paysim-dataset/input

FEATURE ENGINEERING & ANOMALY DETECTION

Behavioral Features (Python):

- time_since_last (velocity)
- is_velocity_anomaly (≤ 1 hour)
- is_burst (repeat amounts)
- Z-score & IQR amount outliers
- Sender & receiver profile metrics:
  - sender_txn_count
  - sender_amount_mean
  - receiver_amount_sum
  - receiver_txn_count

Unsupervised Methods:

- Isolation Forest → iso_anomaly
- KMeans distance threshold → kmeans_anomaly
- Voting system → is_anomaly_final

XGBOOST FRAUD MODEL

Model Details:

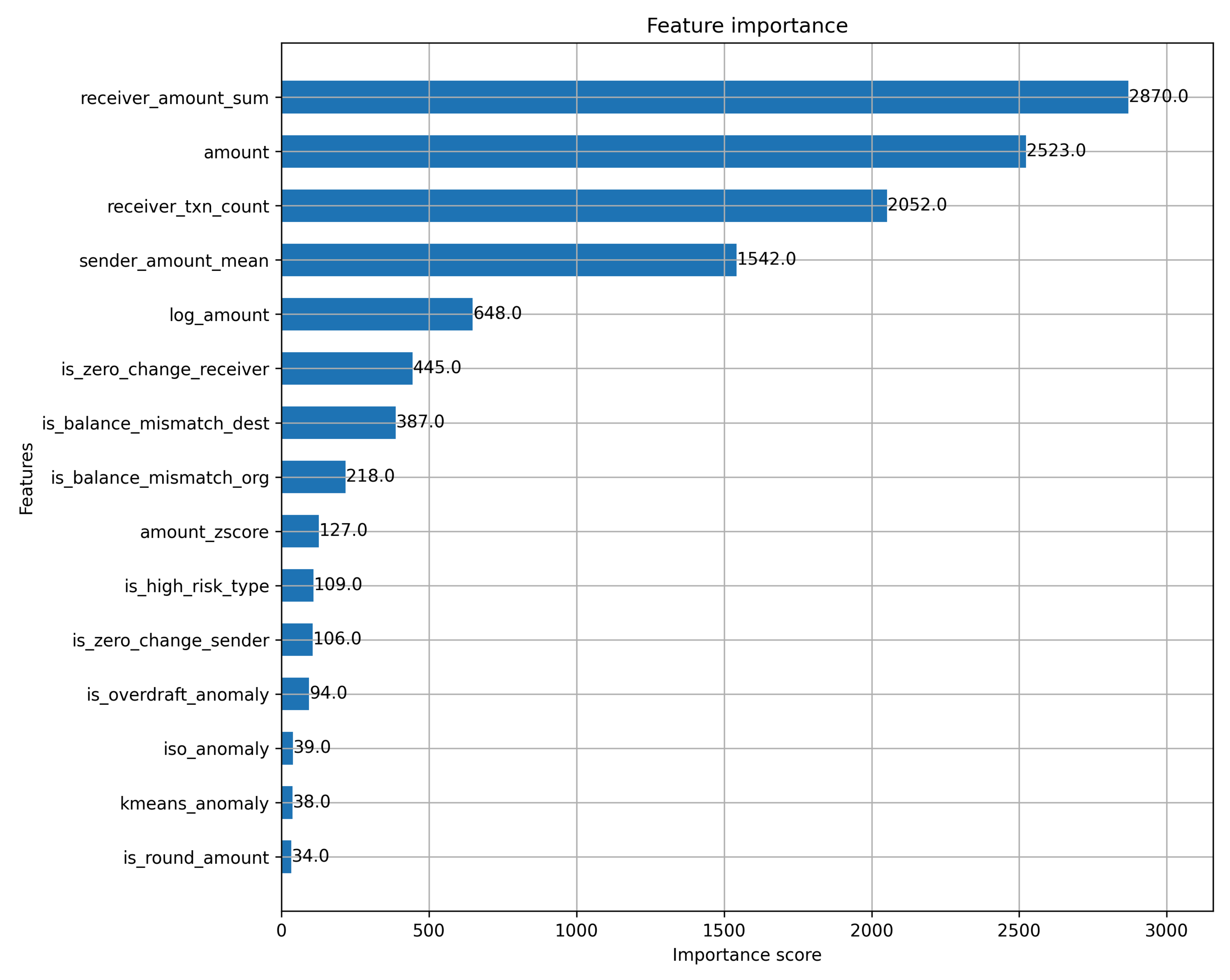
- XGBoost with scale_pos_weight=100 (handles imbalance)
- Trained on 80/20 split (stratified)
- Predicts:
  - fraud_probability
  - fraud_prediction (threshold = 0.5)

Inputs (Supervised Features):

- Rule flags
- Behavioral features
- Outlier signals
- Unsupervised anomaly flags

Performance:

- Recall: ~0.82–0.96 (varies by dataset sample)
- Confusion Matrix: Low false negatives
- Feature Importance:
  - receiver_amount_sum
  - amount
  - receiver_txn_count
  - sender_amount_mean
  - log_amount


Feature importance
receiver_amount_sum
amount
receiver_txn_count
sender_amount_mean
log_amount
is_zero_change_receiver
is_balance_mismatch_dest
is_balance_mismatch_org
amount_zscore
is_high_risk_type
is_zero_change_sender
is_overdraft_anomaly
iso_anomaly
kmeans_anomaly
is_round_amount
2870.0
2523.0
2052.0
1542.0
648.0
445.0
387.0
218.0
127.0
109.0
106.0
94.0
39.0
38.0
34.0
Features
0
500
1000
1500
2000
2500
3000
Importance score

FINAL OUTPUT: 1M-ROW RISK SCORED DATASET

- Sampled to 1,000,000 rows for dashboard & ML stability
- Added fraud probability (fraud_probability)
- Created interpretable risk categories:
  - High Risk (>0.85)
  - Medium Risk (>0.50)
  - Low Risk (>0.20)
  - Very Low Risk (≤0.20)

Final Deliverables:

- fraud_risk_output.csv (1M rows)
- SQL scripts
- Full Python notebook
- Feature importance image
- GitHub documentation