SiriusAI

ANOMALY DETECTION IN AML

WITH ISOLATION FOREST

By :- Parth Tyagi

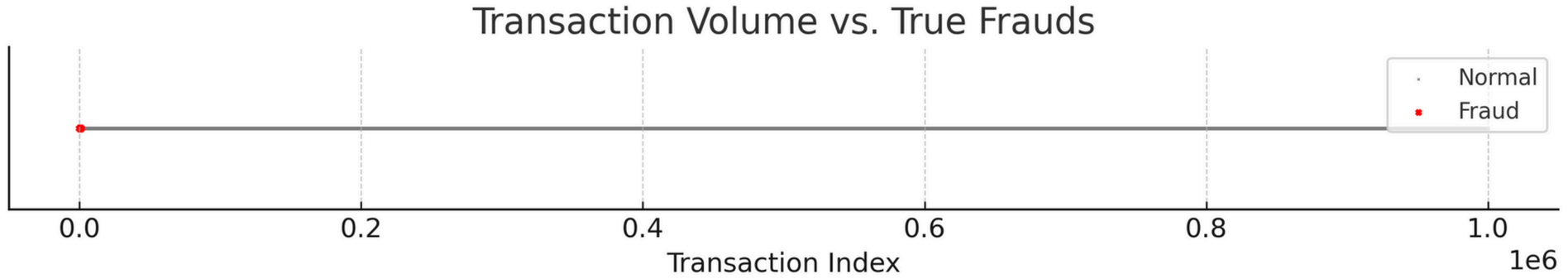
June 18 , 2025

AML & the Role of Anomaly Detection

AML (Anti-Money Laundering) aims to detect illegal movement of funds (e.g., drug money, terrorism).

Key challenge:

- Real suspicious transactions are extremely rare (~0.1%).
- Labels are scarce or delayed → traditional supervised ML underperforms.
- Anomaly detection flags unusual patterns without needing labeled fraud.



Fraudulent transactions (red) are just a tiny minority in a sea of normal ones.

Highlights extreme class imbalance—a major challenge in AML.

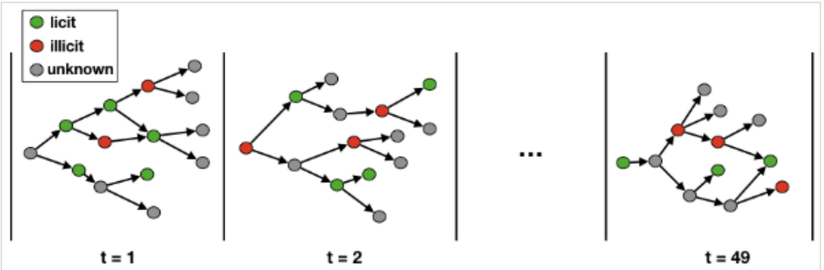
Key Public Datasets for AML & Fraud Detection

PaySim

Synthetic Financial Datasets For Fraud Detection

Synthetic datasets generated by the PaySim mobile money simulator

kaggle.com



The Elliptic Data Set: Working With the Community to Combat Financial Crime in Cryptocurrencies

The Elliptic Data Set, the world's largest labeled transaction dataset publicly available in any cryptocurrency with 200,000 transactions valued at $6 billion.

elliptic.co /

Elliptic Bitcoin Dataset

AMLSim (IBM Research)

IBM/**AMLSim**

IBM

The AMLSim project is intended to provide a multi-agent based simulator that generates synthetic banking transaction data together with a...

4 Contributors · 45 Issues · 287 Stars · 86 Forks

IBM/AMLSim: The AMLSim project is intended to provide a multi-agent based simulator that generates synthetic...

The AMLSim project is intended to provide a multi-agent based simulator that generates synthetic banking transaction data together with a set of known money laundering patterns - mainly for the pur...

GitHub

SAML-D

https://arxiv.org/pdf/2404.14746

Table

Dataset	Type	Size	Fraud %	Format
PaySim	Mobile	6M+	~0.13%	Tabular
AMLSim	Bank	1M+	~0.2%	Tabular + Graph
Credit Card	Card	284K	0.17%	Tabular
Elliptic	Crypto	200K	Tagged	Graph
SAML-D	Cross-bank	9.5M	0.10%	Tabular

Isolation Forest: How It Detects Anomalies

Unsupervised Algorithm (Liu et al., 2008)

→ No need for labeled data – perfect for AML where labels are scarce

Works by Isolation

→ Randomly partitions data using binary trees

→ Outliers = isolated quickly (short path in tree)

→ Normals = need many splits to isolate

Scoring Mechanism

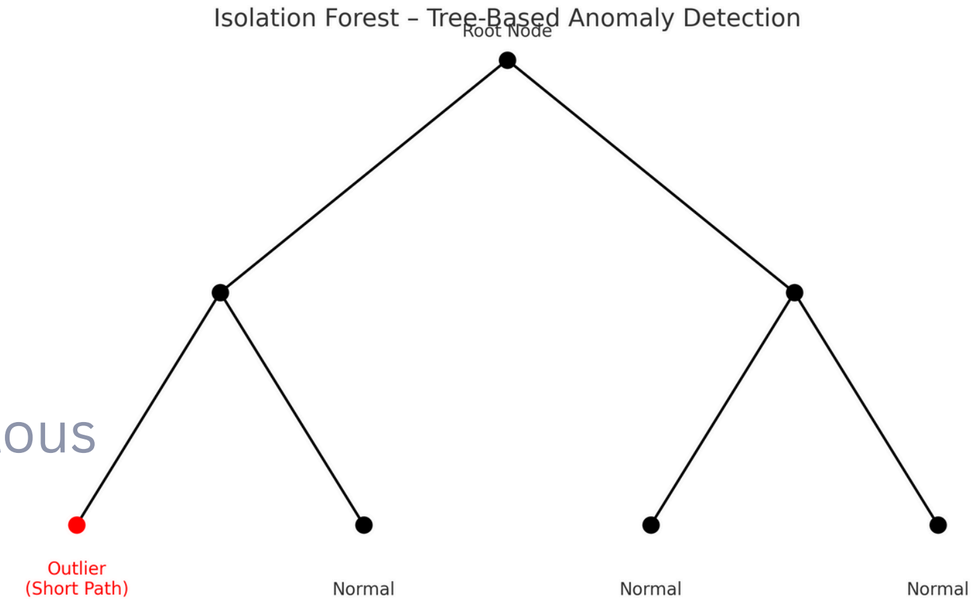
→ Each data point gets an anomaly score

→ Shorter average path length across trees = more anomalous

Fast & Scalable

→ Linear time complexity: O(n log n)

→ Handles high-dimensional financial datasets efficiently



Why Isolation Forest is Ideal for AML

SiriusAI

No Labels Required

→ Unsupervised approach fits AML, where labeled laundering cases are extremely rare

Catches Rare, Unknown Patterns

→ Flags new, never-before-seen behavior

→ Useful for dynamic criminal tactics that bypass static rules

Efficient for Real-Time Monitoring

→ Linear time complexity enables live fraud scanning

→ Used by Hawk AI, ING, and other financial institutions

Works Well on High-Dimensional Data

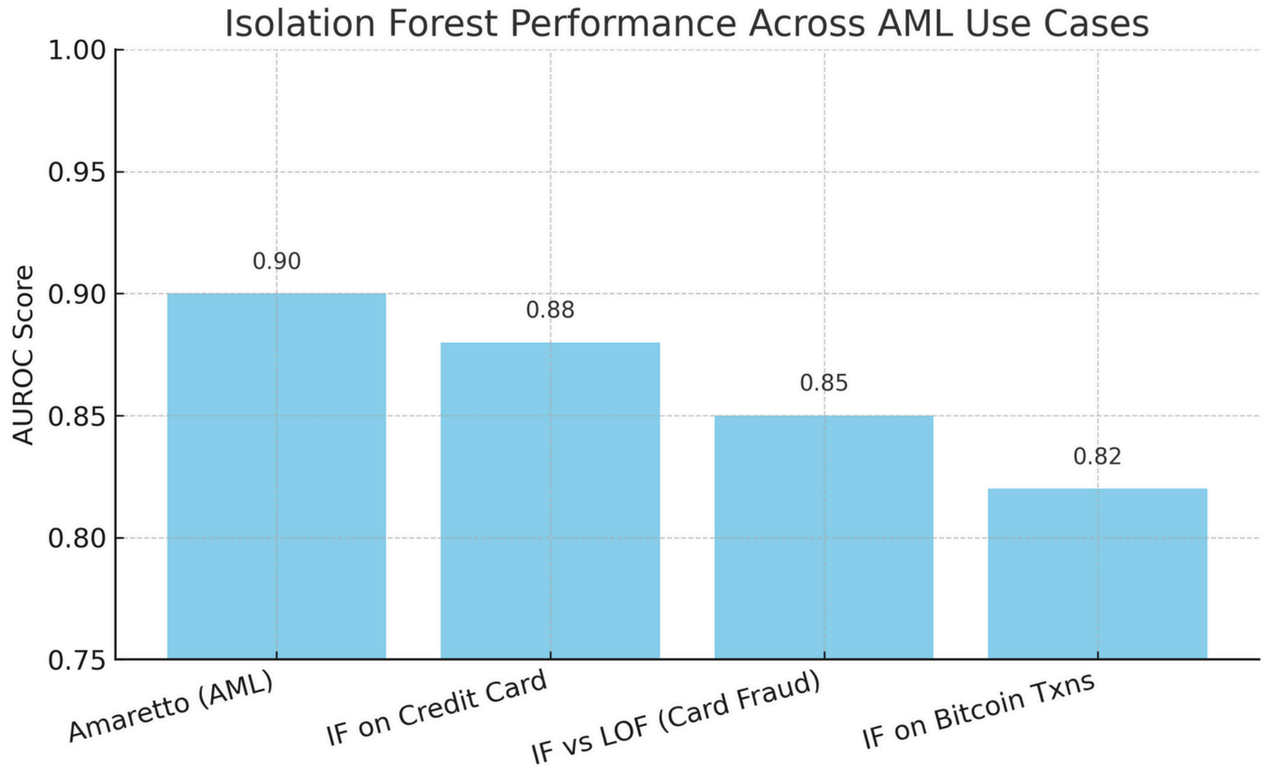
→ Robust against irrelevant/noisy features

→ Handles transaction-level and account-level features easily

Scalable to Big Data

→ Handles millions of records without requiring deep neural networks or large infrastructure

Real-World Use Cases of Isolation Forest in AML



Amaretto Framework (Labanca et al., 2022)

Combines Isolation Forest + Random Forest

→ Human feedback loop → AUROC ~0.90

→ Detected complex laundering with minimal labels

Bitcoin AML Detection (Pham & Lee, 2016)

₿ Used Isolation Forest to flag suspicious wallets

→ Outliers = unusual amounts, frequency, regions

→ Aided early detection of darknet transactions

Credit Card Fraud Studies

IF tested on 284K Kaggle dataset

→ Outperformed LOF and One-Class SVM

→ Detected ~0.17% fraud with strong precision

Hawk AI – Industry Deployment

Isolation Forest used in real-time AML platform

→ Robust across datasets, adaptable to evolving risks

→ Reduced false negatives and manual effort

Root Cause Analysis (RCA) for AML Anomalies

SiriusAI

Why RCA Matters

→ Anomaly score ≠ explanation

→ Analysts need to know why a transaction was flagged

→ Builds trust in AI systems and helps refine alerts

Methods for RCA in Isolation Forest

Use feature importance tools: e.g., SHAP values, feature drop analysis

Combine with expert rules to validate causes (e.g., high frequency + offshore)

Common Anomaly Drivers in AML

- Sudden transaction amount spikes
- Transfers to/from high-risk jurisdictions
- Round-number payments or structuring behavior
- Use of previously inactive accounts

Benefits of RCA

→ Filters false positives

→ Enables model feedback for improvement

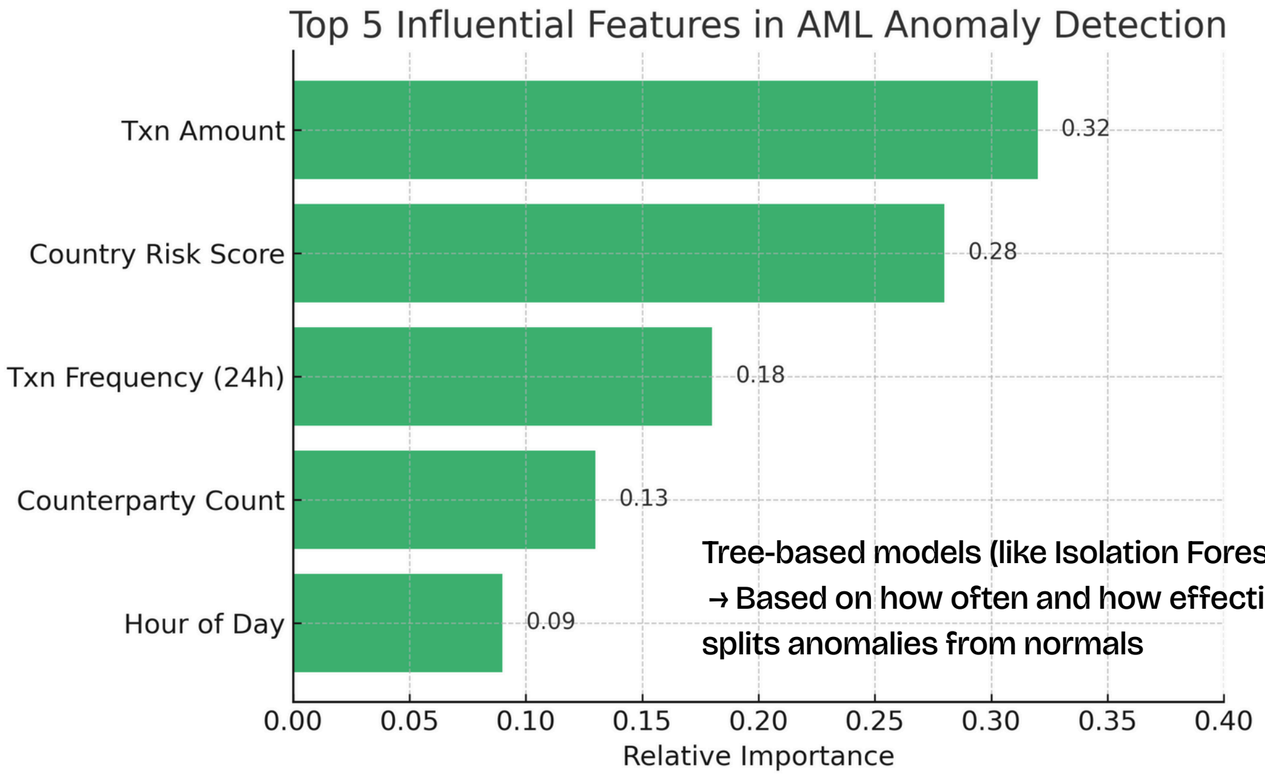
Correlated & Influential Features for AML Models

Why It Matters

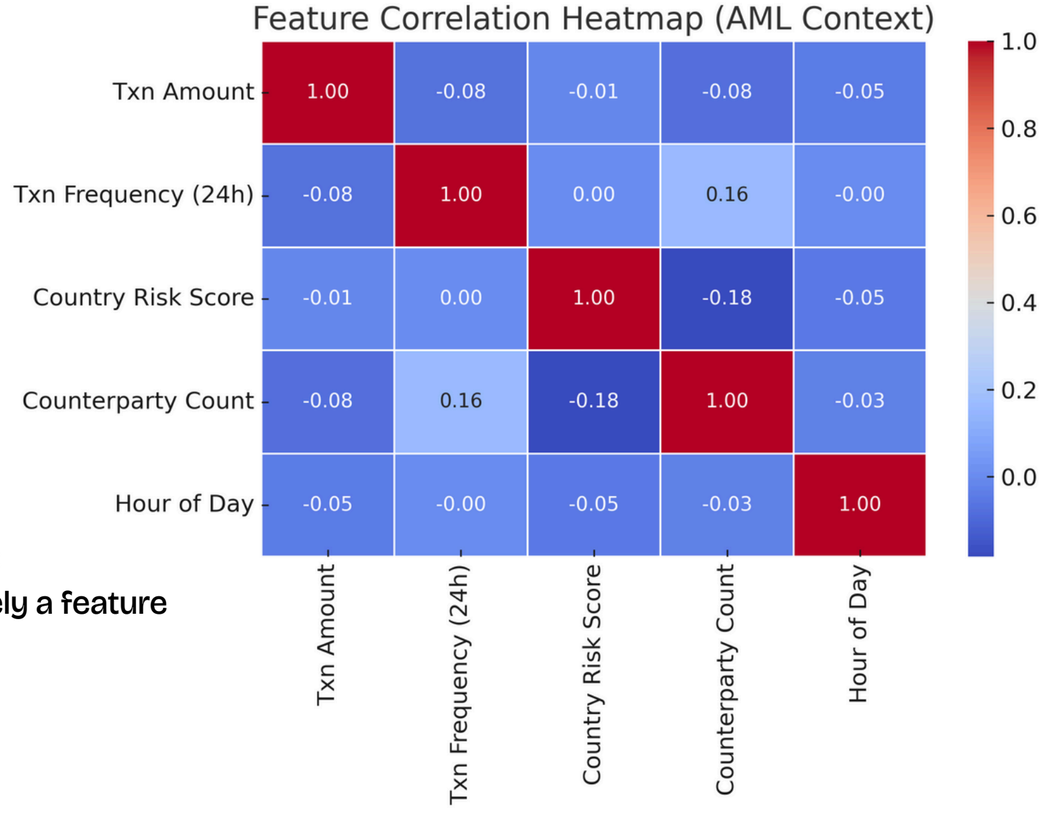
→ Financial data is high-dimensional

→ Not all features contribute equally to anomaly detection

→ Understanding correlations improves model accuracy and interpretability



Tree-based models (like Isolation Forest):
→ Based on how often and how effectively a feature splits anomalies from normals

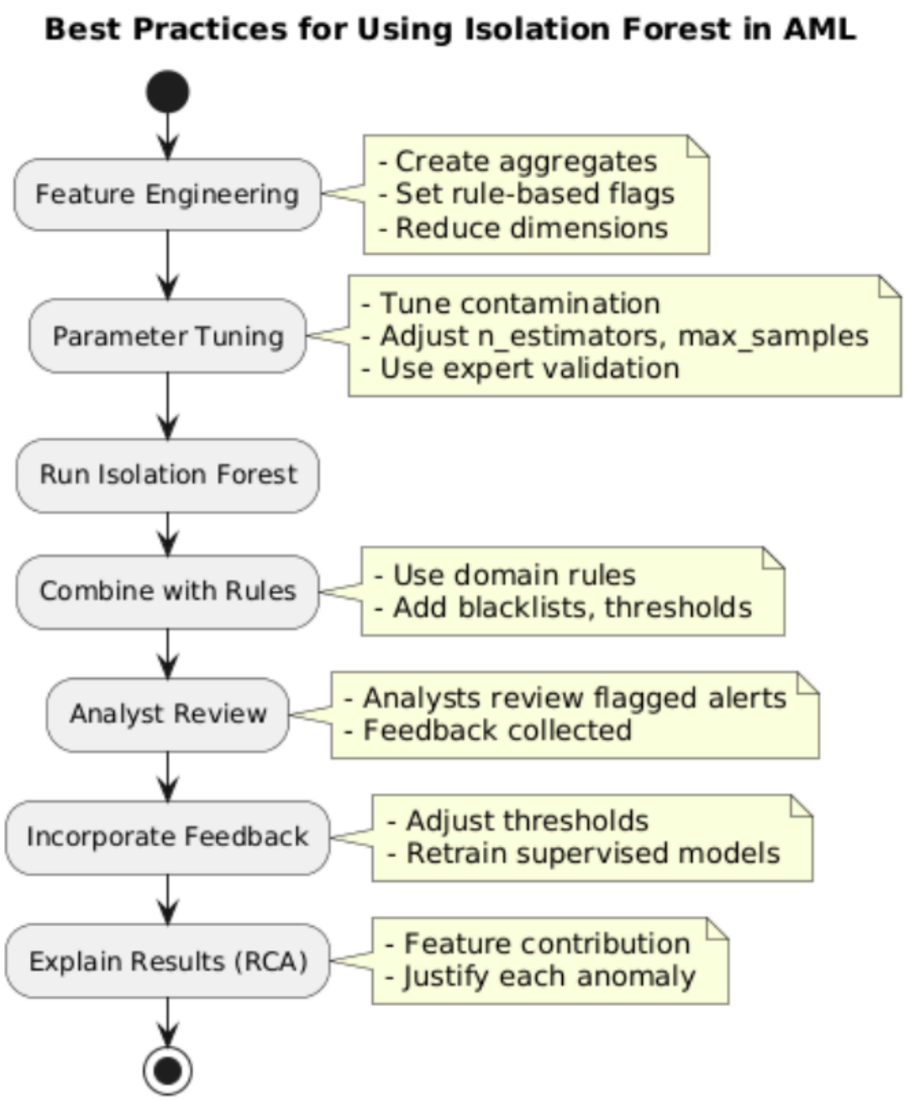


Isolation Forest vs Other Detection Techniques

SiriusAI

Method	Label-Free	Scalable	Interpretable	Best Use Case
Isolation Forest	✅	✅	⚠️ Moderate	Fast unsupervised detection
LOF	✅	❌	✅	Local density anomalies
One-Class SVM	✅	❌	✅	Small, clean datasets
Autoencoder	❌	⚠️ Medium	❌	Deep complex anomalies

Best Practices for Using Isolation Forest in AML



🔧 1. Feature Engineering –Examples

total_sent_last_24h → Captures structuring or layering behavior

is_high_risk_country = 1 if country_score > 7 → Flags geographic risk

⚙️ 2. Parameter Tuning - Examples

contamination = 0.001 → Realistic setting for AML where fraud is <0.1%

Use historical SAR-labeled alerts to calibrate threshold and evaluate precision

🧠 3. Rule-Based Filtering – Examples

if anomaly_score > 0.9 AND is_blacklisted_recipient → High precision trigger

if txn_amount > ₹500,000 AND txn_time in [2am–4am] → Filters nocturnal high-value anomalies

🔄 4. Analyst Feedback - Examples

Analyst validates top 50 alerts → use as labeled positives to train a supervised classifier

Analyst flags a false positive → adjust contamination threshold to reduce such noise

🔍 5. RCA / Explainability – Examples

SHAP Output: Txn Amount = ₹9,90,000 and Risk Score = 9.2 contributed 70% of anomaly score

Tree Path Logic: “This transaction was isolated early due to rare combo of: offshore account + rapid frequency”

SiriusAI

"In AML, you don't need to catch everything upfront — you just need to catch what no one else can see."

Thank you