Mule Account Detection — Exploratory Data Analysis

Phase 1 · EDA Report · Financial Crime Detection
Dataset: Banking Transactions · 7.4M transactions · 5-year window (Jul 2020 – Jun 2025)

METRIC	VALUE
Total Accounts	40,038
Mule Accounts	263 (1.09% of training set)
Imbalance Ratio	90:1
Transactions	7.4M
Patterns Found	7 of 12 tested
Features Engineered	20+

1. Dataset Structure & Relationships

The dataset spans **six interrelated tables**. This is a 20% representative sample — class ratios and distributions are preserved.

TABLE	ROWS	DESCRIPTION	KEY
customers.csv	39,988	Demographics, KYC flags, banking registrations	customer_id
accounts.csv	40,038	Account attributes, balance metrics, status	account_id
transactions (×6 parts)	7,424,845	Every transaction — channel, amount, counterparty	account_id
customer_account_linkage.csv	40,038	Bridge: maps customers → accounts	customer_id, account_id

product_details.csv	39,988	Product holdings: loans, credit cards, overdraft	customer_id
train_labels.csv	24,023	Ground truth: is_mule flag, flag date, alert reason	account_id
test_accounts.csv	16,015	Accounts to predict on in Phase 2	account_id

Join path:

`customers` → *(customer_id)* → `linkage` → *(account_id)* → `accounts` → `transactions`
`customers` → *(customer_id)* → `product_details`
`accounts` → *(account_id)* → `train_labels` / `test_accounts`

> **Note:** There is no direct customer_id in accounts.csv — you must route through customer_account_linkage. After joining, the master training table has **24,023 rows × 60 columns**.

Key channels: UPC/UPD (UPI ~70%), IPM (IMPS), NTD (NEFT), FTD/FTC (Fund transfers), ATW (ATM withdrawal — cash extraction).

2. The Class Imbalance Problem

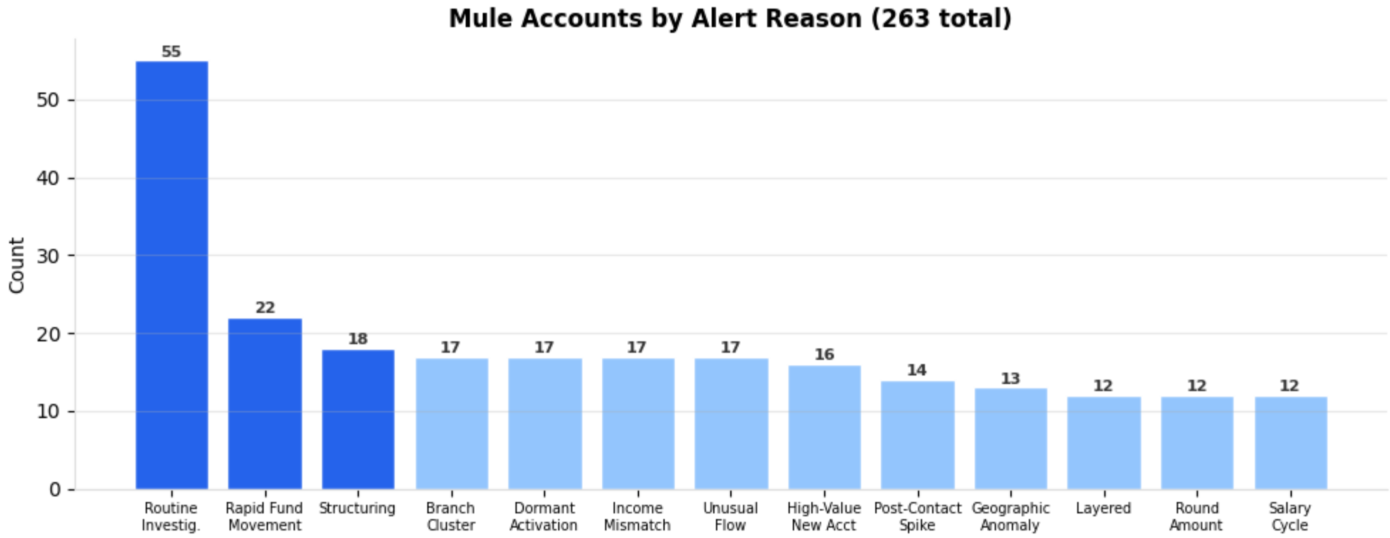
CLASS	COUNT	% OF TRAINING SET
Legitimate	23,760	98.91%
Mule	263	1.09%

Imbalance ratio: 90:1. A model always predicting "legit" gets 98.91% accuracy — but is completely useless. **AUC-ROC** is the correct evaluation metric.

> ⚠️ In Phase 2, SMOTE oversampling or class-weight balancing will be applied.

Mule accounts span **13 distinct alert reasons** — no single pattern dominates (as shown in the chart below):

With only 263 mule cases in this sample, differences smaller than ~5 percentage points should be interpreted with caution — they may not be statistically robust at this sample size.



3. Mule vs Legitimate — Statistical Comparison

All values are **medians** to reduce outlier sensitivity. Signal strength = separation between the two distributions.

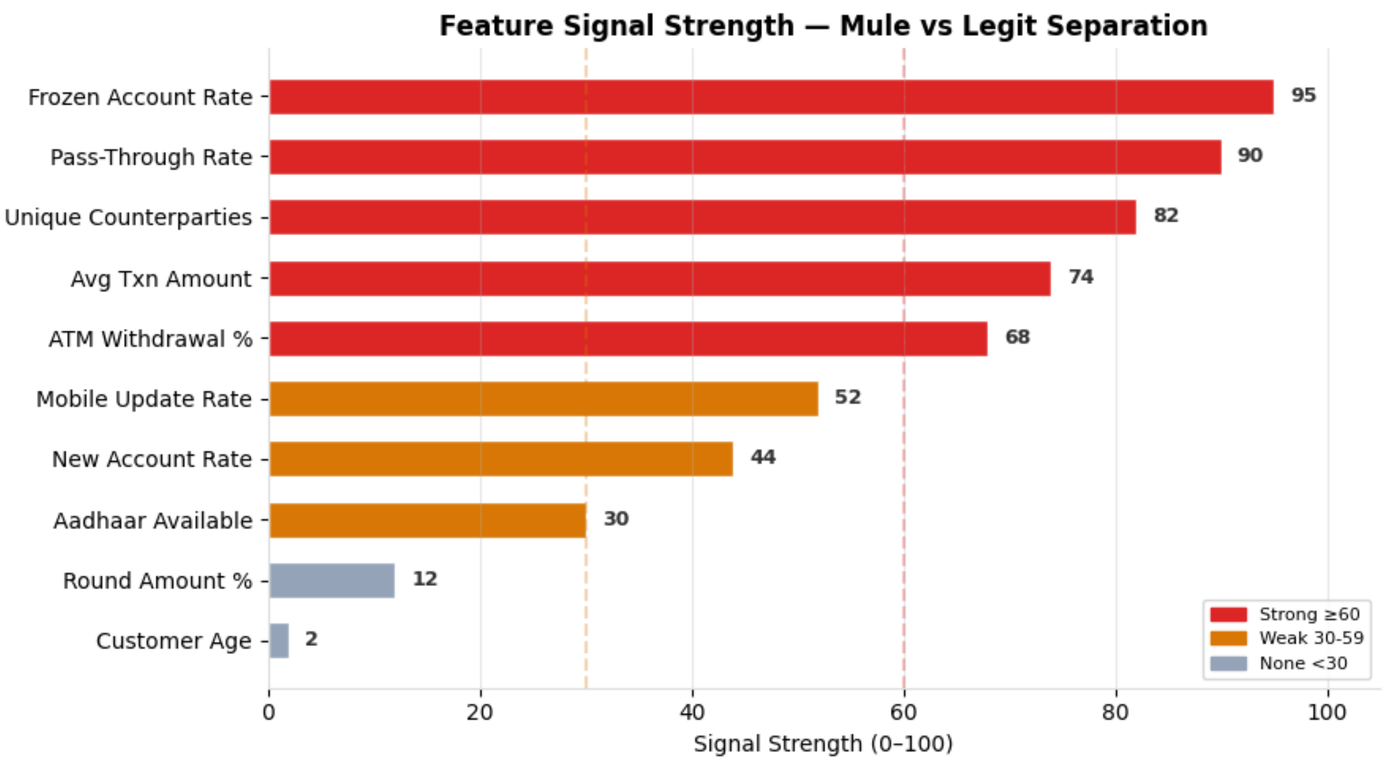
FEATURE	MULE	LEGIT	RATIO	SIGNAL
Frozen account rate	**39.92%**	2.04%	19.6×	🔴 Very Strong
Pass-through rate	**7.53%**	0.00%	∞	🔴 Very Strong
Unique counterparties	**30**	10	3.0×	🔴 Strong
Avg txn amount (₹)	**14,845**	7,343	2.0×	🔴 Strong
Total txn count	**67.5**	38.0	1.8×	🔵 Moderate
ATM withdrawal %	**1.69%**	0.00%	∞	🔵 Moderate
IMPS txns %	**6.59%**	4.17%	1.6×	🔵 Moderate
NEFT debit %	**4.40%**	1.93%	2.3×	🔵 Moderate
Mobile update rate	**20.53%**	14.75%	1.4×	🟡 Weak
Standing instruction %	0.00%	**1.10%**	inverse	🟡 Weak
Avg balance (₹)	3,561	**5,260**	–	🟡 Weak
Aadhaar available	38.0%	**47.1%**	–	🟡 Weak

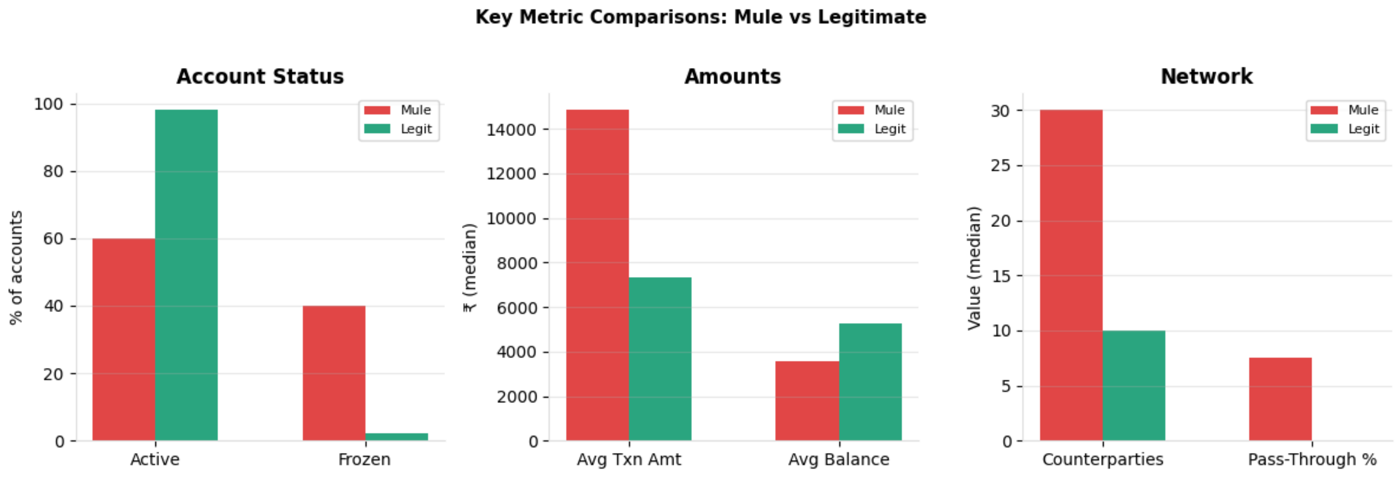
Round amount %	11.5%	16.78%	–	None
Customer age	49.9	49.5	–	None
Relationship tenure	15.5 yrs	15.4 yrs	–	None

> *Note: ∞ in the Ratio column indicates legit = 0, ratio undefined.*

Key interpretations:

- *Pass-through rate (7.53% vs 0%): Mule accounts function as temporary fund conduits, not personal savings accounts.*
- *Unique counterparties (30 vs 10): Consistent with fan-in/fan-out laundering — aggregating from many or distributing to many.*
- *ATM withdrawal (1.69% vs 0%): Physical cash extraction is the final step — converting digital funds to untraceable cash.*
- *Standing instruction (0% vs 1.10%): Absence of recurring payments signals no stable financial life — accounts exist solely to move money.*
- *Round amount % (11.5% vs 16.78%): Mules move precise amounts received from others — round numbers are a normal-user behaviour.*
- *Customer age (49.9 vs 49.5): Demographics offer no signal — mule recruitment cuts across all age groups equally.*





4. Pattern Identification

12 known mule patterns were tested against the data. **7 confirmed, 2 not found, 1 counterintuitive, 2 untestable.**

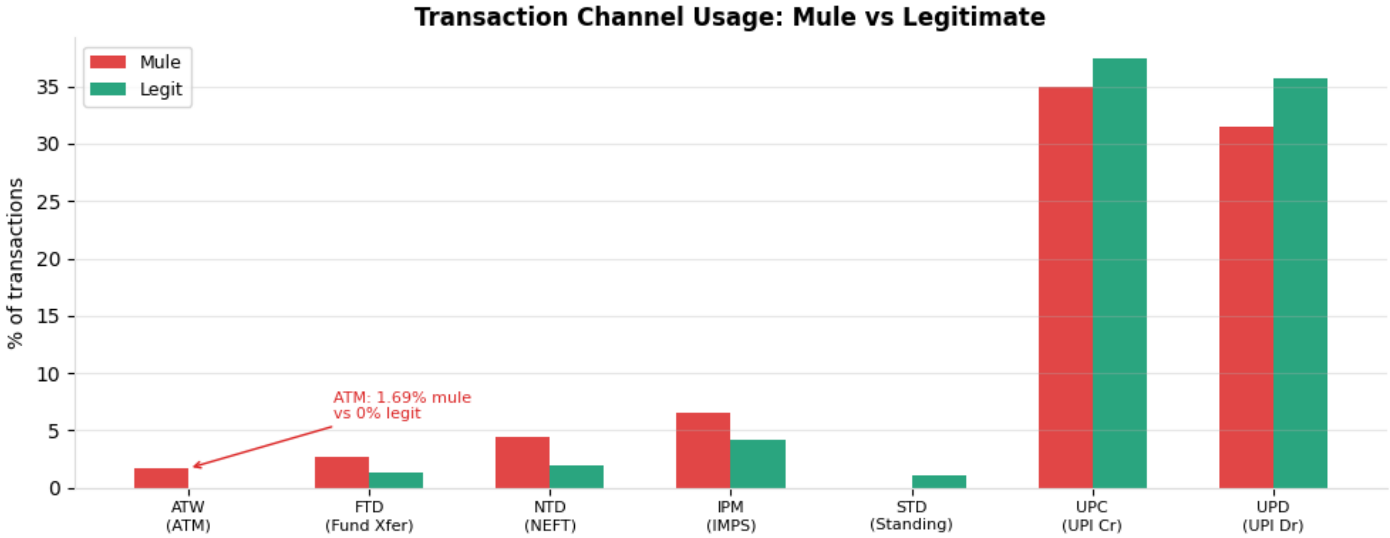
Confirmed Patterns

PATTERN	MULE	LEGIT	STRENGTH
Rapid Pass-Through — money in & out same day	7.53% of days	0.00%	🔴 Very Strong
Fan-In/Fan-Out — wide counterparty network	30 counterparties	10	🔴 Strong
Income Mismatch — high txn amt, low balance	₹14,845 txn / ₹3,561 bal	₹7,343 / ₹5,260	🔴 Strong
Post-Mobile-Change Spike — account takeover signal	20.53% updated	14.75%	🟡 Weak
New Account High Value — new accts have 2× mule rate	2.15%	1.02%	🟡 Weak
Branch Collusion — cluster investigation	17/263 mules (6.5%)	—	🔵 Indirect
Structuring — near-₹50K transactions	Median: 1 txn	0	🟡 Partial

Not Found / Counterintuitive

PATTERN	FINDING
Dormant Activation	Max gap: 81 days (mule) vs 86 days (legit) — no difference
Round Amounts	Mules use *fewer* round amounts (11.5% vs 16.78%) — opposite of expected
Geographic Anomaly	No transaction-level location data — untestable

✅ **Novel finding:** ATM withdrawals (ATW) appear in **1.69% of mule txns vs 0% for legit**. Physical cash extraction = final laundering step. Not listed as a known pattern in the README — discovered independently through channel-level analysis.



5. Feature Engineering Plan

Every feature justified by EDA evidence and backed by specific numbers from Section 3, ordered by predictive strength.

Group A — Transaction Behaviour (HIGH priority)

FEATURE	DESCRIPTION
`pass_through_rate`	% of active days with both credit AND debit (7.53% vs 0%)
`unique_counterparties`	Distinct counterparty count (30 vs 10)
`avg_txn_amount`	Mean transaction amount (₹14,845 vs ₹7,343)
`atm_withdrawal_pct`	% via ATW channel (1.69% vs 0%)

`imps_neft_pct`	% via IPM + NTD (11% vs 6.1%)
`net_flow_ratio`	(credits − debits) / credits — near zero for pass-through
`txn_velocity_per_month`	Txn count / active months (1.8× higher for mules)
`near_50k_count`	Transactions ₹45K–₹50K — structuring signal
`burst_ratio`	Max 7-day txns / avg weekly rate
`std_instruction_pct`	% via STD — absence = no stable financial routine

Group B — Account-Level (HIGH priority)

FEATURE	DESCRIPTION
`is_frozen`	account_status == 'frozen' (40% vs 2%) — **leakage caution**
`balance_to_txn_ratio`	avg_balance / avg_txn_amount — low = pass-through
`account_age_days`	Days since opening — new accounts have 2× mule rate
`had_mobile_update`	Mobile update exists — account takeover signal

Group C — Customer Identity (MEDIUM priority)

FEATURE	DESCRIPTION
`kyc_id_score`	Sum of KYC documents on file (PAN + Aadhaar + Passport)
`digital_banking_score`	Sum of digital banking flags

Group D — Network / Branch (MEDIUM priority)

FEATURE	DESCRIPTION
`branch_mule_concentration`	% of training mules at same branch
`shared_mule_counterparty`	Counterparties shared with known mules

> ⚠️ Group D features must be computed from training labels only — never test labels.

6. Data Quality & Leakage Risks

Columns to exclude (data leakage)

COLUMN	REASON
`alert_reason`	Empty for all legit accounts — directly encodes label
`mule_flag_date`	Only exists post-detection
`flagged_by_branch`	Exists only because account was caught
`freeze_date` / `unfreeze_date`	Consequence of fraud detection, not predictor
`account_status` (frozen)	**Grey area** — strongest predictor but may be post-detection

Missing values

COLUMN	MISSING	ACTION
alert_reason, flag_date, flagged_by	~23,760	🔴 Exclude — structurally empty for legit
freeze/unfreeze_date	~23,750	🔴 Exclude — leakage
last_mobile_update_date	20,465	🔵 Keep — presence is itself a signal
cc_sum / loan_sum	18,904–20,233	🟡 Treat as 0 (no credit products)
aadhaar / pan_available	5,790 / 3,435	🔵 Keep — missing KYC = risk signal
avg_balance	725	⚪ Impute with median — negligible missing rate, low impact on model

Noisy labels

> ⚠️ Labels may contain noise (per README). With only 263 mules, ~13 mislabelled accounts (5%) could meaningfully shift model behaviour. Use label smoothing and robust ensembles in Phase 2.
>
> ⚠️ Since this is a 20% sample, patterns requiring large numbers to emerge (e.g.,structuring, dormancy bursts) may become clearer in the full dataset. Feature importance rankings may also shift when trained on 5× more data.

7. Key Takeaways & Phase 2 Direction

FINDING	DETAIL
Top behavioural signal	Pass-through — money in/out same day (7.53% vs 0%)
Strongest predictor	Frozen status (40% vs 2%) — use with leakage caution
Novel finding	ATM withdrawals (1.69% vs 0%) — cash extraction endpoint
Null findings	Age, tenure, dormancy, round amounts — no signal

The mule account profile: High transaction velocity, large amounts, wide counterparty networks, same-day fund cycling — a **financial pipeline**, not a personal account.

Phase 2 approach: Gradient boosted tree (XGBoost/LightGBM) on 20+ features with class-weight balancing. Metric: AUC-ROC.

> ✅ **Bottom line:** Features capturing pipeline behaviour — pass-through rate, counterparty breadth, channel mix, and balance-to-transaction ratio — will drive a high-AUC model.

Phase 1 EDA Report · 7.4M transactions · 40,038 accounts · Jul 2020–Jun 2025 · 20% sample