PAYMENT FRAUD (TRANSACTION LEVEL)

ONLINE METRICS: ✓ i) Recall / Detection Rate: % of actual fraud caught
✓ ii) False +ve rate: % of legitimate transactions incorrectly flagged as fraudulent
iii) Precision

Trade-off: FPR vs Recall: Lower FPR reduces friction but may miss fraud; Higher recall catches more fraud but increases false +ve

iv) Revenue loss
✓ v) Manual review queue size
vi) Flag rate: (until labels are available)

LABEL CRITERIA: i) Initial labels: chargebacks, refund reason codes
ii) Due to label delay (45-60 days) and evolving "fraud" defn/tactics:
- Manual review on rule based & ML model flagged + customer reported fraud
iii) Feedback loop issue: Transactions blocked by model never get true labels ⇒ even blocked by model = manual review of a sample

TRAINING DATA: Imbalanced dataset

FEATURES:
- Data leakage: ensuring model doesn't learn from post-transaction signals
- Privacy concerns: PII/sensitive data
- Feature freshness (due to evolving fraud tactics): real time features valuable but has cost attached (infra, simple feat engg.)
- Feature crosses: (transaction-level so account-level info may not be used)
  i) Velocity features {Trans. history based on credit card, email etc}: ratio of purchases in last hr to last 30 days. ratio of amt spent in last hr to last 30 days.
  ii) Affinity features on top of ratio counter based features: ratio of purchases in last 1 hrs to last 24 hr in a) location b) credit card
  iii) Reputation features: "email domain", "IP", "phone no.", "device id" (emailage, telesign, maxmind, rNdata)
  iv) Account data (if it can be linked) + Profile: account-level transaction history, account age, gender, age etc.

MODELING: - Boosting based: LightGBM (quick training on large dataset)
Fast inference
Interpretable
Re-training every week or when metrics deteriorate

DEPLOYMENT:

Shadow mode testing →

i) Blue green deployment (instanteous cutover | but requires duplicate infra)

ii) Canary deployment (gradual rollout to limit risk)

A/B Test : (allows some fraud)

(only) ↙ Rule-based flagged | ML Model ↘ flagged (only) → cannot wait for 45-60 days to get labels.

Manual Review | Manual Review

Aspect	Reserve Abuse	Fraud	WF
Product metrics	In-cart conversion	Fraud found rate	Penetration metrics
ML metrics	P ✓ R	P R ✓	P R
Eng.	real-time/online training real-time prediction	online prediction offline training	batch training batch prediction
Labels	ratio of reserve to purchase	1mm → Fraud Analysts/ Accertify (rule-based engine) Real labels → 45 day lag	1mm → Listings (in secondary Market) Real labels → Presence Real labels → only after the event has played

Training Data → ACCOUNT DATA + DATA FROM THIRD PARTY ←

ACCOUNT DATA

1) Historical purchase data
2) Historical user-system interaction data (repeat offenders like ticket limit abuse etc.)

e.g. # of times email changed, phone number change how many diff IPs seen for this account

DATA FROM THIRD PARTY

a) emailage
b) Telesign
c) maxmind
d) nudata

Aspect	Reserve Abuse	Fraud	WF
Data split	Time-based	Time-based	(IID) CV → no need to login originally, now includes account data.
FEATURES (unique ones)	Velocity-based (No credit card data)	Credit card	Actor-Artist affinity (No credit card data)

MODELS

Vowpal wabbit

→ online learning

→ feature-eng. on the fly (interaction terms)

→ categorical features: : hashing ~~to~~ → int using murmur-hash (no hash table storage)

→ Scalable

→ Interpretable

Disad: ~~1) Not [illegible]~~

1) Technically linear

→ a good amt of effort in data cleaning

LightGBM

+ ~~Lambda~~ LambdaRank Ranking for analyst review

Non-linear (very noisy data)

→ Rule-based system had been in place: moderate level of understanding of patterns

+

Tree-based models are interpretable to an extent.

Vowpal wabbit

→ Roadmap for online learning

→ behavior of bad actors: not enough info on our problem space : hence interpretable model

1) Adv. & Disadv of Vowpal wabbit & lightGBM

2) Some case studies in error-analysis / debugging