AML Risk Analysis & Modeling Report (based on the SAML-D study)

Executive summary

This report synthesizes the AML analytics performed on the SAML-D synthetic transaction-monitoring dataset. The work covers exploratory analysis, class imbalance profiling, typology patterns, payment-method risk, amount distributions, preprocessing, modeling with XGBoost, and performance assessment. Key takeaways: the dataset is extremely imbalanced (≈0.1% suspicious), suspicious activity clusters in specific typologies (notably structuring and cash-related flows), and suspicious transactions are materially larger on average. While the trained model scores a perfect 1.0 across metrics, that result is likely inflated by data leakage or synthetic identifiability and should not be treated as production-ready without stricter validation.

---

1) Dataset at a glance

- **Scope:** 9,504,852 transactions with 12 features; 28 typologies across 15 network structures. Suspicious prevalence ≈0.1039%.
- **Working sample for analysis:** Random sample of 2,000,000 transactions (reproducible seed).
- **Data quality & schema:** Mixed dtypes, engineered calendar features (year/month/day/week), and no missing values observed in the sample.
  These characteristics make the dataset ideal for technique exploration, with the important caveat that it is synthetic.

---

2) Exploratory insights

Class imbalance. In the 2M sample: ~1,997,967 normal vs. 2,033 suspicious (≈983:1). Accuracy is not a meaningful metric; focus should be on PR-AUC, precision/recall, and cost-sensitive thresholds.

Typologies. Normal traffic is dominated by fan-out/fan-in patterns; among suspicious typologies, **Structuring** is most frequent, followed by **Cash Withdrawal**, **Smurfing**, and

Deposit-Send—all consistent with real-world laundering playbooks (structuring to avoid thresholds, cash-heavy movement, rapid pass-through).

Payment-method risk.

- **Volume of suspicious**: Cross-border leads by count (520 within the sample).
- **Rate of suspicious**: Cash-intensive rails show the **highest relative risk** (e.g., Cash Deposit ≈0.63%, Cash Withdrawal ≈0.44%), outpacing Cross-border's ≈0.26%. In practice, prioritize **both** high-count and high-rate channels.

Amounts. Suspicious transactions are much larger on average (≈3.25× normal). Extremes are more pronounced: the suspicious max (~7.21M) far exceeds the normal max (~1.0M), suggesting "edge" behaviors beyond typical customer activity. Raw amounts are hugely right-skewed (skew ≈46); log transforms normalize them and improve modeling stability.

---

3) Preprocessing & feature engineering

- **Numerical pipeline:** Median imputation + RobustScaler to dampen outlier influence.
- **Categorical pipeline:** Ordinal encoding with unknown-category handling.
- **Temporal handling:** Raw timestamps removed after deriving calendar features to reduce leakage from direct time keys while retaining seasonality.
- **Split:** 80/20 train–test on the sampled data.
  This is a sensible, production-oriented preprocessor layout using a single `ColumnTransformer` for reproducibility.

---

4) Modeling approach

- **Imbalance strategy:** Use of `scale_pos_weight` to reflect the ≈983:1 skew (effective weight ~284 in the tuned configuration).
- **Estimator:** XGBoost with GPU acceleration; randomized search (200 trials), moderate CV (2-fold), and early stopping across 2,000 estimators and a sub-1e-1 learning rate.
- **Best hyperparameters (abridged):** `max_depth=7`, `min_child_weight=5`, `gamma≈9.5`, `subsample≈0.95`, `colsample_bytree≈0.69`, `learning_rate≈0.09`, `scale_pos_weight≈284`.
  The choices align with best practice for rare-event detection at scale.

---

5) Reported performance & interpretation

- **Headline metrics on the sample's test split:** ROC-AUC = 1.0, PR-AUC = 1.0, accuracy/precision/recall = 1.0.
- **Feature importance & SHAP** were used for interpretability; threshold curves and error analysis were also explored.
- **Reality check:** Perfect scores are a red flag. In AML, even excellent models rarely approach perfection; symptoms point to (a) **feature leakage** (e.g., including a "typology" field that encodes post-hoc labeling cues), (b) **train/test contamination**, or (c) **synthetic separability** not present in production. Treat these results as *diagnostic*, not deployable.

---

6) Risk patterns that warrant action

1. **Cross-border flows:** High suspicious **volume**; scrutinize corridors involving historically higher-risk geographies.
2. **Cash rails:** Highest suspicious **rates**; strengthen controls on deposit/withdrawal sequencing, branch/device geolocation, and cash-to-wire pivots.
3. **Amount dynamics:** Large values and rapid aggregation/dispersal are strong signals; use log-scaled features and percentile-based cutoffs to stabilize decisions.
   Structuring/Smurfing: Intensify pagination/sequence features (e.g., rolling counts above/below thresholds over short windows).