SAML-D Baseline Model — Technical Report (v1)

1) Executive summary

- **Goal.** Build a leakage-safe, calibrated baseline to detect suspicious transactions in **SAML-D** (≈9.5M rows; prevalence ≈0.1%).
- **Approach.** Time-based split (70/15/15), engineered **past-only** behavioral features, gradient boosting (LightGBM) with heavy regularization + class weighting, and **post-training probability calibration** (isotonic) on validation.
- **Headline results (Test set).**
  - **Ranking:** ROC-AUC **0.9943** (pre-cal) / **0.9796** (post-cal).
  - **Precision-Recall:** PR-AUC **0.3043** (pre-cal) / **0.3006** (post-cal). With base rate ≈0.1%, that is roughly **~300× random**.
  - **Calibration:** ECE **0.0017** → **0.0001**; Brier **0.0037** → **0.00085** after isotonic calibration.
  - **Operational (Top-K):** At **top 0.50%** of transactions, **Precision ≈ 22%** with **Recall ≈ 92%**. At **top 0.05%**, **Precision ≈ 34%** for a high-priority queue.
- **Readiness.** Suitable as a **production baseline**. Use **calibrated** scores for thresholds and volumes; monitor drift and recalibrate periodically.

---

2) Data & preprocessing

Columns available (post initial drop):

- Numeric: `Sender_account`, `Receiver_account`, `Amount`, `Is_laundering`, `Year`, `Month`, `Day`, `Week`.
- Categorical: `Payment_currency`, `Received_currency`, `Sender_bank_location`, `Receiver_bank_location`, `Payment_type`, (`Laundering_type` dropped for leakage avoidance).

Initial steps

- Dropped `Time`, `Date`.
- Constructed `_date = to_datetime(Year, Month, Day)` and **sorted chronologically** (critical for leakage-safe features).
- Casted dtypes for memory/perf.

Train prevalence.

- Train size: **6,661,223** rows; positives **6,774** → **0.1017%**. Used `scale_pos_weight ≈ 982.4`.

---

3) Feature engineering (leakage-safe)

Domain features

- `log_amount = log1p(Amount)`
- `currency_mismatch = 1(Payment_currency ≠ Received_currency)`
- `country_mismatch` and `cross_border` from sender/receiver locations

Historical behavior (computed using past rows only)

- Activity: `sender_prev_txn_count`, `receiver_prev_txn_count`, `pair_prev_txn_count`
- Magnitude: cumulative and average amounts: `sender_prev_amount_sum/avg`, `receiver_prev_amount_sum/avg`
- Cadence: `sender_prev_gap_days`, `receiver_prev_gap_days`
- Relationship breadth: `sender_unique_receivers_prev`, `receiver_unique_senders_prev`
- Deviation: `amount_vs_sender_avg` (= Amount / sender_prev_amount_avg, clipped)

Categoricals used as categoricals: `Payment_currency`, `Received_currency`, `Sender_bank_location`, `Receiver_bank_location`, `Payment_type`.

Explicit protections against leakage

- **Raw entity IDs** (`Sender_account`, `Receiver_account`) **not used as model features**; only used to build *past-only* aggregates.
- `Laundering_type` **excluded** (not available at inference).

---

4) Train/Validation/Test split

- **Time-based** by `_date` quantiles: **70% Train**, **15% Val**, **15% Test**. This simulates forward-in-time deployment and prevents look-ahead leakage.

5) Model & training

Model. LightGBM (`LGBMClassifier`) with strong regularization:

- `learning_rate=0.03`, `n_estimators=5000` (with early stopping)
- `num_leaves=64`, `max_depth=8`, `min_child_samples=2000`
- `subsample=0.8`, `colsample_bytree=0.7`, `reg_alpha=1.0`, `reg_lambda=5.0`, `max_bin=127`
- `scale_pos_weight≈982.4`, `random_state=42`

Training dynamics.

- Early stopped at **3610** rounds.
- Validation AUC/PR steadily improved; no “1-tree” overfit pattern.

Calibration.

- **Isotonic** calibration fitted on **validation** predictions (`cv='prefit'`), then applied to test. Expect slight ROC dips (ties) but markedly better probability accuracy (ECE/Brier).

---

6) Evaluation results

6.1 Global metrics

Pre-calibration (LightGBM raw probabilities)

- **Validation:** ROC-AUC **0.9928**, PR-AUC **0.2708**, Brier **0.003322**, ECE **0.001489**
- **Test:** ROC-AUC **0.9943**, PR-AUC **0.3043**, Brier **0.003701**, ECE **0.001701**

Post-calibration (Isotonic on validation)

- **Validation:** ROC-AUC **0.9779**, PR-AUC **0.2678**, Brier **0.000737**, ECE **0.000002**
- **Test:** ROC-AUC **0.9796**, PR-AUC **0.3006**, Brier **0.000850**, ECE **0.000105**

Interpretation

- **Ranking** remains excellent (AUC ≈ 0.98–0.99).
- **PR-AUC ≈ 0.30** vs random baseline ≈ prevalence (~0.1%) ⇒ roughly **~300×** better than random in PR space.

- **Calibration** improves dramatically post-isotonic (ECE ~10^{-4}; Brier substantially lower). Use **calibrated scores** for thresholding and alert volume management.

6.2 Operational (Top-K) performance — Test set

Fractions are of all test transactions; k is the number of alerts. Metrics shown as **Precision / Recall**.

Slice	**k**	**Pre-cal P / R**	**Post-cal P / R**
0.05%	706	0.2890 / 0.1212	**0.3399 / 0.1426**
0.10%	1,412	**0.3215 / 0.2698**	0.3088 / 0.2591
0.25%	3,532	0.3199 / 0.6714	**0.3213 / 0.6744**
0.50%	7,064	**0.2193 / 0.9204**	**0.2193 / 0.9204**
1.00%	14,128	0.1110 / 0.9317	0.1111 / 0.9323

Workload example. At **top 0.50%** (k=7,064), **Precision ≈ 21.9%** ⇒ **~1,550 TPs** and **~5,514 FPs**; **Recall ≈ 92%** implies ~**1,680** positives in the test split (≈0.12% prevalence). This is a strong primary work-queue operating point. At **top 0.05%**, **Precision ≈ 34%** yields a compact "urgent" queue.

---

7) Recommended operating strategy

1. **Primary queue:** Review **top 0.50%** of scores daily/weekly. Expect ≈22% precision with ≈92% recall.
2. **High-priority queue:** Review **top 0.05%** for faster SLA; ~34% precision.
3. **Thresholds:** Use the **calibrated** score distribution and pick thresholds by fraction on **validation** (the code includes a helper to compute the cut-off for any target fraction).

4. **Analyst capacity guardrails:** Track alert volume and re-tune the fraction to keep queues stable.

---

8) Model risk & leakage controls

- **Entity-ID memorization avoided:** Raw `Sender_account`/`Receiver_account` excluded from features; only used to compute **past-only** aggregates.
- **`Laundering_type` excluded.**
- **Time-based split** prevents look-ahead leakage.
- **Class-weighting** (not oversampling) preserves the negative class distribution for validation/test.

---

9) Monitoring & maintenance

- **Score calibration drift:** Re-fit isotonic on the latest month/quarter; watch **ECE** and **Brier**.
- **Volume stability:** Monitor % of transactions above the operating threshold; keep it within capacity bands.
- **Quality:** Track **PR-AUC**, **precision@K/recall@K**, and case-closure outcomes (confirmed SARs, etc.).
- **Data drift:** Monitor base rate, amount distributions, new corridors/currencies.
- **Retraining trigger:** Significant ECE increase, precision@K drop, or base-rate shift.

---

10) Limitations & next steps

Limitations

- No explicit **time-window** features (e.g., 7d/30d counts) beyond cumulative stats.
- No **graph/network** features beyond simple pair counts.
- Calibration uses isotonic (piecewise constant), which can reduce ROC marginally via ties.

Next steps (low → medium effort)

1. **Windowed features** (7/30-day sender/receiver/pair volume and amount; distinct counterparties per window).
2. **Cross-border nuances** (region-risk embeddings; currency volatility flags).

3. **Graph signals** (in/out degree, PageRank, triadic closure) aggregated per account; feed into the same model.
4. **Alternative calibration** (Platt/sigmoid) if preserving ROC is desirable while keeping good calibration.
5. **Periodic recalibration** and **champion/challenger** with CatBoost (ordered target encoding for categoricals).

---

11) Reproducibility notes

- **Split:** Time-based (70/15/15 by `_date`).
- **Model:** LightGBM `LGBMClassifier` with parameters listed in §5; early stopping at **3610** rounds.
- **Imbalance:** `scale_pos_weight ≈ 982.4` computed from train class counts.
- **Calibration:** `CalibratedClassifierCV(method='isotonic', cv='prefit')` trained on validation predictions.
- **Key features:** See §3 list; raw IDs and `Laundering_type` excluded from features.
- **Random seed:** `random_state=42`.

---

12) Glossary (quick)

- **ROC-AUC:** Ranking quality over all thresholds.
- **PR-AUC:** Precision/Recall trade-off; more informative at extreme imbalance.
- **Brier score:** Mean squared error of probabilities; lower is better.
- **ECE (Expected Calibration Error):** Average gap between predicted probability and observed frequency; lower is better.

Appendix A: Data at a glance

- **Rows / columns:** 9,504,852 rows × 12 columns (≈870 MB in memory).
- **Base rate:** 9,873 suspicious / 9,504,852 total ⇒ **0.1039%** (≈ **1 in 963** transactions).
- **Key fields (raw):** `Time`, `Date`, `Sender_account`, `Receiver_account`, `Amount`, `Payment_currency`, `Received_currency`, `Sender_bank_location`, `Receiver_bank_location`, `Payment_type`, `Is_laundering`, `Laundering_type`.
- **Split-ready time index:** `_date = to_datetime(Year, Month, Day)` used for chronological splitting and leakage-safe history features.

Payment type volume (all transactions)

- ACH **2,008,807** · Credit card **2,012,909** · Debit card **2,012,103** · Cheque **2,011,419** · Cross-border **933,931** · Cash Withdrawal **300,477** · Cash Deposit **225,206**.

Suspicious rate by payment type (share of transactions flagged suspicious; **relative risk vs baseline** in parentheses)

- **Cash Deposit: 0.6239%** (**6.01×**)
- **Cash Withdrawal: 0.4440%** (**4.27×**)
- **Cross-border: 0.2814%** (**2.71×**)
- **ACH: 0.0577%** (**0.56×**)
- **Credit card: 0.0564%** (**0.54×**)
- **Debit card: 0.0559%** (**0.54×**)
- **Cheque: 0.0540%** (**0.52×**)

Suspicious typologies (share within suspicious class)

- **Structuring** 18.94% · **Cash_Withdrawal** 13.51% · **Deposit-Send** 9.57% · **Smurfing** 9.44% · **Layered_Fan_In** 6.64% · **Layered_Fan_Out** 5.36% · **Stacked Bipartite** 5.13% · **Behavioural_Change_1** 3.99% · **Bipartite** 3.88% · **Cycle** 3.87% · **Fan_In** 3.69% · **Gather-Scatter** 3.59% · **Behavioural_Change_2** 3.49% · **Scatter-Gather** 3.42% · **Single_large** 2.53% · **Fan_Out** 2.40% · **Over-Invoicing** 0.55%.

Amount distribution

- **Suspicious:** max **12,618,500**, mean **40,588**, min **15.82**.
- **Normal:** max **999,962.19**, mean **8,729.88**, min **3.73**.
- Suspicious mean is **~4.65×** normal; suspicious maximum is >**12×** the normal maximum.
- **Skewness:** raw `Amount` skew **102.16** (extremely heavy-tailed); **log1p(Amount)** skew **−1.01**, supporting our use of `log_amount`.

Appendix B: EDA highlights & modeling implications

1. **Extreme imbalance (0.1039%)** → Evaluate with **PR-AUC** and **top-K** metrics; use class weighting (as implemented).
2. **Channel risk heterogeneity.** Cash-related and cross-border channels carry **2.7–6.0×** higher suspicious rates; ACH/cards/cheques are **~0.5×** baseline. **Implication:** interactions like *(channel × amount deviation)*, *(channel × cross-border)* can add lift.
3. **Magnitude & deviations matter.** Suspicious amounts are much larger on average; deviation from an account's historical average (`amount_vs_sender_avg`) is informative.

4. **Cadence & relationships.** History counts, time gaps, and unique counterparties capture typological behaviors (e.g., fan-in/out, structuring, layering). Our past-only aggregates directly target these patterns.
5. **Typology mix.** A handful of patterns (e.g., **Structuring**, **Cash_Withdrawal**, **Deposit-Send**, **Smurfing**) account for >50% of suspicious cases; features tied to bursts, splitting, and network breadth are particularly valuable.
6. **Probability calibration.** Raw boosting scores ranked well but required calibration; isotonic reduced ECE to **~1e-4**, stabilizing thresholds and alert volumes.
7. **Data quality notes.** `Time`/`Date` appear as strings in raw data; we derive a proper `_date` index. If hour-level ordering becomes available, add intra-day features (e.g., time-of-day spikiness) and use **temporal windows** (7/30-day) for finer control.

Appendix C: Quick reference table — payment types

Payment type	Total Tx	Suspicious Tx	Suspicious rate	Relative risk vs baseline
Cash Deposit	225,206	1,405	**0.6239%**	**6.01×**
Cash Withdrawal	300,477	1,334	**0.4440%**	**4.27×**
Cross-border	933,931	2,628	**0.2814%**	**2.71×**
ACH	2,008,807	1,159	0.0577%	0.56×
Credit card	2,012,909	1,136	0.0564%	0.54×
Debit card	2,012,103	1,124	0.0559%	0.54×

Cheque	2,011,419	1,087	0.0540%	0.52×

Conclusion. The baseline delivers **excellent ranking**, **strong lift in PR space**, and **near‑perfect calibration** after isotonic scaling. It is suitable for deployment with a recommended operating band around **top 0.5%** (primary queue) and **top 0.05%** (urgent queue), with routine monitoring and periodic recalibration.