

Model validation memo — Card Fraud Scorer v7.1

Fraud — card transaction scoring

Fraud model validation, with the PSD2 reference rate

MODEL

Card Fraud Scorer v7.1

ENGAGEMENT

Sample — card fraud scoring annual validation

PREPARED

2026-10-07

PREPARED BY

AlphaIQ CertifyIQ, computed by the product

RECOMMENDATION



Inconclusive — the evidence is insufficient

10 sections · 5 traced figures · 0 finding(s) · 9 test(s) named as not performed

SYNTHETIC SAMPLE

Every figure in this report was computed by CertifyIQ from a synthetic book — 20,000 scored card transactions with matured fraud labels. No real institution, model or customer is described. The book is `fraud-transaction-book.csv`; how it was generated, and the defects planted in it, are documented in `fraud-transaction-book-dgp.json`.

AT A GLANCE · SYNTHETIC SAMPLE

Card Fraud Scorer v7.1

Fraud — card transaction scoring · Fraud model validation, with the PSD2 reference rate

Inconclusive — the evidence is insufficient

Only 2 of 6 rows in this pack were judged against a threshold; the other 4 were reported without one and are marked as such in the results table. A model cannot be called reliable when most of what the pack set out to measure went unmeasured, however well the remainder did. AlphaIQ requires at least half of a pack's rows to have been judged against a threshold before a reliance conclusion is available. No supervisor publishes such a proportion; this is a platform rule, stated so that a reader who disagrees can apply their own from the counts printed beside it.

10

SECTIONS

5TRACED
FIGURES**0**

MATERIAL

0

NOTABLE

0

OBSERVATIONS

9NOT
PERFORMED

Supervisory frame: this memo is built to "Payment Fraud Model Validation", published under EU · PSD2 RTS · SR 26-2. No supervisory frame is declared for this engagement, so the frame above was chosen by the model's family rather than by its supervisor. That is not a statement that the two agree: nobody has recorded who supervises this model, and a reader in another jurisdiction should treat the section structure, the citations and the recommendation vocabulary as EU's.

Evidence manifest: none sealed. The documents behind each figure are named where the figure is; the set of documents on record when this memo was written has not been attested by anyone, so it cannot be cited here.

How to read this sample. Every number below was computed by CertifyIQ from a synthetic book (fraud-transaction-book.csv), through the same resolver, metric pack and report builder a signed-in engagement uses. Each figure carries its own trace — where it came from and what it was judged against. Tests that could not be performed are named with the reason, never left out. Where no threshold was declared, a metric is reported but not judged, because this report does not invent an institution's tolerances.

CONTENTS

- | | |
|---|--|
| 1 Purpose and scope | 6 Score stability and disparate impact |
| 2 Label maturity as an evidence statement | 7 Feature-target association screen |
| 3 Discrimination under class imbalance | 8 Findings |
| 4 Operating point at the declared review capacity | 9 Third-party dependencies and residual risk |
| 5 PSD2 transaction-risk-analysis exemption rate | 10 Not performed |

1 Purpose and scope

This memo validates Card Fraud Scorer v7.1 against the transaction book extracted on 2026-09-30, spanning 2025-12-03 to 2026-05-31 (180 day(s)) and covering 20,000 usable transaction(s) of 20,000 read.

The book was declared, not defaulted: a label maturity window of 120 day(s), a review capacity of 1.00% of the book, a production cutoff of 0.02, and an intended PSD2 exemption claim under the card / EUR 100 band. Any of these left undeclared costs exactly the metric it unlocks a named not-computable row in section 8, rather than a default this platform chose on the institution's behalf.

This validation carries no replica of the model under test and no ongoing-monitoring reading: fraud figures below are measured directly from the declared book against the owner's own claims and against published or policy thresholds, so the recommendation stated above is not gated on a replication step this family's evidence chain has no equivalent of, and no monitoring section is stated because none is attached to this run.

2 Label maturity as an evidence statement

Every fraud label in this book was required to be at least 120 day(s) old as of the declared extraction date, 2026-09-30: a book whose most recent transaction fell inside that window would have been refused outright, because a "0" inside it means "no fraud reported yet", not "legitimate", and every recall, precision and PR-AUC computed over such rows is biased upward in a way that does not show in the file.

This book cleared that check: its latest usable transaction is dated 2026-05-31, at or before the 120-day cut-off measured back from 2026-09-30. Every figure in sections 3 through 6 is computed over labels this engagement is entitled to treat as settled, not as still arriving.

3 Discrimination under class imbalance

Precision-recall average precision, not ROC-AUC: at a fraud base rate under one per cent, ROC-AUC and Gini are dominated by the negative class and near-useless, and PR-AUC is the summary statistic that actually moves.

No interval is reported on this figure, and its width is therefore unknown rather than narrow: the stratified bootstrap needs a declared resample count and a declared root seed, and this engagement has declared neither. Both are declared on the engagement workspace, under Declare a computation input, and each carries the same document, page, declarer, date and reason a threshold claim does. The count is declared rather than shipped because how many resamples an interval needs is a methodological appetite, and the seed is declared so a third party re-derives the same interval byte for byte.

PR-AUC (average precision)

MEASURED 0.146

PR-AUC (average precision): 0.146 measured on 20,000 transactions carrying 66 confirmed fraud. Benchmark source: Average precision over the scored transaction book, computed as scikit-learn's `average_precision_score` does — the rectangular sum $\sum (R_n - R_{\{n-1\}}) P_n$, not a trapezoid. At a fraud base rate below one per cent, AUC and Gini are dominated by the negative class and near-useless; PR-AUC is the summary that moves. No authority publishes a floor, so the figure is judged against the model owner's own documented value.

4 Operating point at the declared review capacity

Recall at the declared review capacity states what the queue this institution can actually work catches; precision at the declared production cutoff states what the rule that actually fires in production catches. Neither is computed at a cutoff or capacity this platform chose.

No interval is reported on these figures either, for the reason stated above: the engagement has declared neither the resample count nor the root seed the stratified bootstrap needs.

Recall in the queue the declared review capacity buys**MEASURED 0.379**

Recall in the queue the declared review capacity buys: 0.379 measured on 20,000 transactions carrying 66 confirmed fraud. Benchmark source: Share of fraudulent transactions falling in the top-k by score, where k is the alert volume this institution's review function can actually work. The capacity is declared per engagement — shipping one or five per cent would be this pack inventing an operations budget — and the recall is judged against the owner's own claim at that capacity.

Precision at the live production cutoff**MEASURED 0.054**

Precision at the live production cutoff: 0.054 measured on 20,000 transactions carrying 66 confirmed fraud. Benchmark source: True positives ÷ alerts raised at the score cutoff the rule actually fires at in production, with the alert volume that cutoff creates reported beside it. Both the cutoff and the precision claimed for it are the model owner's own figures.

5 PSD2 transaction-risk-analysis exemption rate

Measured over the rolling 90-day window 2026-03-03 to 2026-05-31, on 7,131 remote transaction(s) totalling 431,101.58: the value-weighted rate — Article 19(1)'s own measure — is 11.45 bp. The same window counted by transaction instead of by value is 43.47 bp. The two are printed together because they differ, and only the value-weighted figure is ever graded against the Annex.

PSD2 transaction-risk-analysis fraud rate (value-weighted, remote, rolling 90 days)**MEASURED 11 bp**

PSD2 transaction-risk-analysis fraud rate (value-weighted, remote, rolling 90 days): 11 bp measured on 20,000 transactions carrying 66 confirmed fraud. Benchmark source: Commission Delegated Regulation (EU) 2018/389, Annex — reference fraud rate for remote electronic card-based payments at an exemption threshold value of EUR 100: 0.130% (13.0 bp). The band is the one this engagement has declared it intends to claim the Article 18 transaction-risk-analysis exemption under; the rate is not this platform's number.

6 Score stability and disparate impact

Score stability (PSI) compares the declared reference window against everything after it; disparate impact is the selection-rate ratio between declared groups at the production cutoff. Neither carries a threshold this platform picked — PSI's 0.10 is an industry convention cited as one, and disparate impact is judged against this institution's own fair-lending policy, never the four-fifths rule, which has no standing in fraud model supervision.

Score stability (PSI) between the declared reference and recent windows**MEASURED 0.0014**

Score stability (PSI) between the declared reference and recent windows: 0.0014 measured on 20,000 transactions carrying 66 confirmed fraud. Benchmark source: Population stability index over the scores, between the reference window this engagement declares and everything after it. The 0.10 is an industry convention rather than a supervisory limit and is cited as one; the window is declared rather than derived, because a split this pack chose would decide which period counts as normal.

NOT PERFORMED — AND WHY**Disparate impact — selection-rate ratio between declared groups**

No group column has been declared, so no selection-rate ratio is computed. This is the usual outcome and the right one: protected-class attributes are personal data, the personal-data scan refuses them at intake, and the only evidence that can reach this check is a group label the institution has already derived and supplied deliberately. This is a permanent gate and not a backlog item: the only route to a figure is demographic tokenisation performed OUTSIDE the client boundary, producing a label the institution derives, owns and declares. Declare that column if this engagement has one; otherwise this check is not performed on the evidence axis and the memo says so in words.

7 Feature–target association screen

No feature–target association screen was performed on this book, and the reason is stated rather than left to inference: This evidence declares no feature columns, so there is nothing to screen. The columns are DECLARED, never suggested: a feature is named whatever the model's owner called it, and a column bound by sniffing would be printed in the memo as an association about a feature nobody named (this platform's rule, with more force here). Declare the feature columns on the mapping.

NOT PERFORMED — AND WHY

Feature–target association screen

This evidence declares no feature columns, so there is nothing to screen. The columns are DECLARED, never suggested: a feature is named whatever the model's owner called it, and a column bound by sniffing would be printed in the memo as an association about a feature nobody named (this platform's rule, with more force here). Declare the feature columns on the mapping.

8 Findings

No findings were raised. Every metric listed in section 9 as performed returned within its stated limits.

Against the findings register: 0 finding(s) open.

9 Third-party dependencies and residual risk

No third-party dependency is declared for this engagement. That is a statement by the institution that the model embeds no vendor or open-source component it cannot validate inside — not an absence of the question.

10 Not performed

The following checks were not performed, each with the reason. A validation report's most dangerous property is the test it quietly omitted, so this register is part of the memo rather than an appendix to it.

NOT PERFORMED — AND WHY

Disparate impact — selection-rate ratio between declared groups

No group column has been declared, so no selection-rate ratio is computed. This is the usual outcome and the right one: protected-class attributes are personal data, the personal-data scan refuses them at intake, and the only evidence that can reach this check is a group label the institution has already derived and supplied deliberately. This is a permanent gate and not a backlog item: the only route to a figure is demographic tokenisation performed OUTSIDE the client boundary, producing a label the institution derives, owns and declares. Declare that column if this engagement has one; otherwise this check is not performed on the evidence axis and the memo says so in words.

Feature–target association screen

This evidence declares no feature columns, so there is nothing to screen. The columns are DECLARED, never suggested: a feature is named whatever the model's owner called it, and a column bound by sniffing would be printed in the memo as an association about a feature nobody named (this platform's rule, with more force here). Declare the feature columns on the mapping.

Scoring latency at the 99th percentile

Measured on the production scoring path by an APM agent, under production load. No run this platform performs produces it, so a figure here could never satisfy the provenance rule that every number in a report be reproducible from its manifest. A latency printed from a validation run would be the latency of the validation run.

Real-time feature availability at score time

Whether each feature was actually present when the transaction was scored is a property of the feature store's point-in-time state. It is a data-lineage audit of a running system, not a statistic over a book, and a book exported after the fact cannot evidence what was available before it.

Feature leakage, as a pass/fail

Turning a feature-target association into a pass or a fail is not this pack's to do: whether a feature that correlates with the outcome is leakage or is simply predictive is a judgement about how the feature is produced — about WHEN its value became knowable relative to the decision — and no correlation threshold separates the two. That stays permanently not performed; no story closes it. THE SCREEN ITSELF IS DELIVERED: this pack reports, per DECLARED feature column, a point-biserial correlation for numeric columns and per-category outcome rates with their counts for categorical ones — never blended into one ranking, since an unordered category has no correlation and any number that looked like one would depend on an ordering nobody declared. Every figure carries the sample it was computed on; columns that cannot be screened (constant, or one distinct value per row) are NAMED with the reason rather than dropped; and the memo prints the ranking with no cut-off, no verdict and no count of anything flagged. This entry stays in the register because the GRADE is what is missing, not the evidence — an entry that vanished the day the screen landed would read as the pack having grown the verdict it deliberately refuses to issue.

The PSD3 / Payment Services Regulation successor to the Annex rates

The six reference fraud rates this pack grades against come from the ANNEX to Commission Delegated Regulation (EU) 2018/389, which is the applicable RTS today and remains so. PSD3 and the Payment Services Regulation reached provisional political agreement in November 2025; publication in the Official Journal was anticipated for mid-2026 and had not happened as at 11 September 2026. The PSR applies twenty days after publication and PSD3 is transposed nationally within eighteen months of entry into force. THE FIGURES ARE NOT IN EITHER TEXT: the exemptions are retained and refined at Level 1, and the thresholds sit in regulatory technical standards the EBA must draft after the PSR is in force — so the successor to this Annex is a later act than the regime that authorises it, and whether it keeps these six figures, changes them, or removes the transaction-risk-analysis exemption is not settled, and this pack does not guess. Named here so the citation cannot go stale silently, and dated so a reader can tell a fact that was checked from one that was assumed.

Sanctions and watchlist screening interactions

A different control with a different evidence shape — name matching against a list, not a score against an outcome. Nothing in a scored transaction book evidences it, and a fraud pack that appeared to cover it would leave the gap unexamined.



AlphaIQ CertifyIQ — independent validation of models and the decisions they drive. This sample was produced by the product itself; nothing in it was written by hand.

Memo content hash d5ea6546ec78795afc284138f81d6ee69bf1bb803e8c2af210c93e990e8e9451

This document is a recommendation, not a certification. A certificate is issued only on attested evidence, which a synthetic sample does not have.