

Model validation memo — Application Scorecard Challenger v4

Machine learning — application scorecard

SR 26-2 model validation memo

MODEL

Application Scorecard Challenger v4

ENGAGEMENT

Sample — application scorecard challenger
review

PREPARED

2026-10-07

PREPARED BY

AlphaIQ CertifyIQ, computed by the product

RECOMMENDATION

● **May be relied upon, with conditions**

12 sections · 10 traced figures · 2 finding(s) · 19 test(s) named as not performed

SYNTHETIC SAMPLE

Every figure in this report was computed by CertifyIQ from a synthetic book — A scored application book with outcomes, score bands, segments and a stability window. No real institution, model or customer is described. The book is `ml-scored-population.csv`; how it was generated, and the defects planted in it, are documented in `ml-scored-population-dgp.json`.

AT A GLANCE · SYNTHETIC SAMPLE

Application Scorecard Challenger v4

Machine learning — application scorecard · SR 26-2 model validation memo

May be relied upon, with conditions

1 material finding(s) require remediation or an accepted compensating control before unqualified reliance. Each is listed in section 9 with the evidence that produced it. 5 of 6 rows were not judged (no threshold declared) and are outside this conclusion.

12

SECTIONS

10TRACED
FIGURES**1**

MATERIAL

1

NOTABLE

0

OBSERVATIONS

19NOT
PERFORMED

Supervisory frame: this memo is built to "SR 26-2 Model Validation", published under US · Federal Reserve / OCC / FDIC. 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 US'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 (ml-scored-population.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 | 7 Stress and sensitivity, and why none was run |
| 2 Model description and provenance | 8 Ongoing monitoring: stability and performance across segments |
| 3 Independent replication, and why none was performed | 9 Findings |
| 4 The scored population and its integrity | 10 Third-party dependencies and residual risk |
| 5 Conceptual soundness: what the model relies on | 11 Not performed, and what each absence costs |
| 6 Outcomes analysis: discrimination and calibration | 12 Reproducibility |

1 Purpose and scope

This memo reports an independent review of Application Scorecard Challenger v4 against the ml metric pack, over 12,000 scored record(s), one credit application at decision time extracted as at 2026-08-31. The record unit is the institution's own: one credit application at decision time, cited to Model development document §3.1, generated fixture.

WHAT THIS REVIEW IS NOT, stated first because for this family it is the more surprising half. No replica of the model was built, none was executed, and no scenario was projected through it. Every figure below is computed from outcomes the model already produced. Section 3 says why that is permanent rather than pending.

Six metrics are declared by this pack. 7 further checks a reviewer might expect are named in section 11 with the reason each is not performed.

2 Model description and provenance

Application Scorecard Challenger is reviewed here as a binary classifier: one score per record and one realised outcome. That narrowing is the pack's, and it is declared rather than assumed — a regressor's error distribution and a ranker's ordering quality have nothing in common with what follows, and section 11 names them as out of scope.

THE BASIS ON WHICH ALPHAIQ WAS ABLE TO REVIEW IT AT ALL: a scored export and a declaration. No source code, no serialised artefact and no specification were supplied, and none was asked for. What a reviewer can say from that is bounded, and the boundary is drawn in the next section rather than left for a reader to find.

The observation dates run against the declared extraction date of 2026-08-31; a record observed after it is a defect in the export and is dropped with that reason rather than assigned to a window.

3 Independent replication, and why none was performed

NO REPLICA OF THIS MODEL WAS BUILT, AND NONE EVER WILL BE. This is not a step that was skipped, deferred, or priced out.

Machine-learning models in this family are gradient-boosted trees, random forests and neural networks. None of them is expressible as a declared specification — there are no coefficients to state — so no replica can be built and the replication fidelity gate has nothing to run. This is a property of how the model was built and no upload changes it. What this platform does perform on such a model is its outcomes, its calibration, its stability and its operating points, plus the evidence artefacts its owner can supply.

WHAT WAS DONE INSTEAD, which is the part this section exists to make legible: the model's realised outputs were read directly. Discrimination, calibration overall and by declared band, stability against a declared reference window, and performance across declared segments are all computable from what the model DID, and none of them needs a specification. What they cannot do is challenge the specification — a replica built from a declaration exposes a wrong declaration, and there is no declaration here to be wrong.

A REPLICA FITTED TO THIS MODEL'S BEHAVIOUR WOULD BE WORSE THAN NONE. It would agree with the model exactly where it was sampled and diverge everywhere else, and a fidelity gate would pass because the fixture is where it looks. That is a permanent refusal of this platform and not a limit of this engagement's evidence.

NOT PERFORMED — AND WHY

Independent replication of the model, and every check resting on a replica

Machine-learning models in this family are gradient-boosted trees, random forests and neural networks. None of them is expressible as a declared specification — there are no coefficients to state — so no replica can be built and the replication fidelity gate has nothing to run. This is a property of how the model was built and no upload changes it. What this platform does perform on such a model is its outcomes, its calibration, its stability and its operating points, plus the evidence artefacts its owner can supply.

4 The scored population and its integrity

The export carried 12,000 record(s); 12,000 were usable. None was dropped.

Every one of 12,000 records carries a distinct "application_id", against the declared record unit: one credit application at decision time (Model development document §3.1, generated fixture).

The score range observed is 0.060593 to 0.835743. The scores are probabilities, so calibration is a meaningful question of them.

Records read from the export

MEASURED 12,000

Records read from the export: 12,000 measured on 12,000 scored records carrying 3,605 realised positive outcomes.

Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Records used

MEASURED 12,000

Records used: 12,000 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Records dropped, each with a reason

MEASURED 0

Records dropped, each with a reason: 0 measured on 12,000 scored records carrying 3,605 realised positive outcomes.

Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Realised positive outcome rate

MEASURED 30.04%

Realised positive outcome rate: 30.04% measured on 12,000 scored records carrying 3,605 realised positive outcomes.

Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

5 Conceptual soundness: what the model relies on

CONCEPTUAL SOUNDNESS IS ORDINARILY READ OFF A SPECIFICATION — these are the drivers, this is the sign of each coefficient, here is whether the theory holds. This model has no specification, so that window is closed, and the owner's own attribution artefact is the only other view of what the model relies on that exists.

What follows therefore grades nothing. It prints what the model uses, with the provenance of the artefact that says so, precisely so that a reviewer can form the judgement this platform will not: whether the model should be using it.

NO EXPLAINABILITY EVIDENCE HAS BEEN DECLARED for this engagement, and this pack expects one. A machine-learning model has NO DECLARABLE SPECIFICATION — there are no coefficients to transcribe — so the ordinary route to knowing what a model relies on is closed for this family. The owner's own attribution artefact is the only evidence of it that exists, which is why this pack expects one and names its absence rather than passing over it. Supply the attribution export your own tooling produces — one row per feature, one value per row — with the method, the population it was computed over, who produced it and the document it is cited to. This platform does not compute one in its place.

This is a NAMED ABSENCE and not a finding about the model. Nothing here says the model is unexplainable; it says that the evidence a reader would need in order to say anything about it was not supplied, and that this platform did not manufacture a substitute.

NOT PERFORMED — AND WHY

Explainability evidence for this model

NO EXPLAINABILITY EVIDENCE HAS BEEN DECLARED for this engagement, and this pack expects one. A machine-learning model has NO DECLARABLE SPECIFICATION — there are no coefficients to transcribe — so the ordinary route to knowing what a model relies on is closed for this family. The owner's own attribution artefact is the only evidence of it that exists, which is why this pack expects one and names its absence rather than passing over it. Supply the attribution export your own tooling produces — one row per feature, one value per row — with the method, the population it was computed over, who produced it and the document it is cited to. This platform does not compute one in its place.

6 Outcomes analysis: discrimination and calibration

Discrimination over the whole population is 0.763. NO ACCURACY CLAIM IS DECLARED for this model, so the figure is reported and not judged. EU AI Act Article 15 requires the accuracy metrics to be declared in the instructions for use; declaring one is what makes this row judgeable.

Band by band — [0, 0.1]: 1,124 record(s), +435 bp; [0.1, 0.25]: 4,690 record(s), +154 bp; [0.25, 0.5]: 4,749 record(s), -427 bp; [0.5, 1]: 1,437 record(s), -399 bp.

THE MEAN AND THE BANDS DISAGREE, AND THE BANDS ARE THE FINDING. Overall the model is under-predicting by 116 bp, which read alone is unremarkable; its worst band is out by 435 bp in the other direction. The two partly cancel in the mean, which is the failure mode of a model tuned on a rank metric and is invisible to a single calibration figure.

Out-of-sample discrimination (AUC)

MEASURED 0.763

Out-of-sample discrimination (AUC): 0.763 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Calibration bias (mean predicted – realised)

MEASURED -116 bp

Calibration bias (mean predicted – realised): -116 bp measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Worst declared band: absolute calibration bias

MEASURED 435 bp

Worst declared band: absolute calibration bias: 435 bp measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

MATERIAL ml-band-cancellation Calibration errors in opposite directions cancel in the overall bias

The overall calibration bias and the worst declared band's bias point in opposite directions, and the band's is more than twice the size. A reader of the headline figure alone would conclude the model is close to calibrated; band by band it is over-confident at one end of the score range and under at the other.

Evidence to re-run: Band table over 12,000 scored record(s), one credit application at decision time, against the score bands declared for this engagement.

7 Stress and sensitivity, and why none was run

NO STRESS TEST, PERTURBATION OR COUNTERFACTUAL SWEEP WAS RUN. Every one of them asks what the model does under conditions that did not happen, and answering that requires the model to be EXECUTED on inputs this platform generates — the runtime capability, which is permanently out of scope rather than pending.

Perturbing the SCORED POPULATION instead is available, cheap, and would be misleading: it measures the arithmetic of this pack under noise, not the stability of the model. This memo does not do it and does not report a figure that could be mistaken for it.

NOT PERFORMED — AND WHY

layer1.macro-stress

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

layer2.ldp

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

layer3.counterfactual

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

layer4.migration

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

8 Ongoing monitoring: stability and performance across segments

The score distribution after the declared reference window moved by 0.1334 on the population stability index. The 0.10 compared against is an INDUSTRY CONVENTION and not a supervisory limit, and it is the only number this pack ships. It is exceeded here, which is a statement that the distribution moved — and this memo does not say WHY. Population change, an upstream data change and genuine regime change are not separable from a scored population, and naming one would be a guess under a heading.

Across the declared segments — near-prime: 3,994 record(s), AUC 0.764; prime: 6,041 record(s), AUC 0.793; thin-file: 1,965 record(s), AUC 0.648. The model discriminates best in "prime" and worst in "thin-file".

The lowest-to-highest selection rate across declared segments at the declared operating cutoff is 0.45. IT IS REPORTED AND PERMANENTLY UNJUDGED. The four-fifths rule a reader may recognise is 29 CFR 1607.4(D), the EEOC's Uniform Guidelines on Employee Selection Procedures — a US employment-selection rule of thumb with no standing in model supervision — and shipping it as a bar would be this platform asserting a legal standard in a domain it does not come from. The ratio a model must hold is the institution's own policy.

No tolerance for segment spread is declared for this engagement, so the figure above is reported and not judged. No supervisor publishes one.

Score stability (PSI) against the declared reference window**MEASURED 0.1334**

Score stability (PSI) against the declared reference window: 0.1334 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Segment performance spread (best – worst AUC)**MEASURED 0.145**

Segment performance spread (best – worst AUC): 0.145 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

Selection-rate ratio at the declared cutoff**MEASURED 0.45**

Selection-rate ratio at the declared cutoff: 0.45 measured on 12,000 scored records carrying 3,605 realised positive outcomes. Benchmark source: one credit application at decision time (Model development document §3.1, generated fixture), as at 2026-08-31.

NOTABLE ml-segment-spread One declared segment is served materially worse than the population as a whole

Discrimination in "thin-file" is below the population figure by more than five points of AUC. This is a finding about the MODEL and not a fairness conclusion: whether a gap of this size is lawful or acceptable is a judgement about the population and the decision it feeds, which this review does not make.

Evidence to re-run: Per-segment AUC over the declared segment column, 12,000 scored record(s), one credit application at decision time.

9 Findings

2 finding(s) were raised, 1 of them material. Each carries the evidence a reviewer would re-run to check it.

Against the findings register: 2 finding(s) open, 1 at grade 1, 2 carrying a grade the engine PROPOSED and nobody has owned, 2 newly raised. Oldest open finding is 0 day(s) old.

2 live finding(s) have no owner assigned. An unowned finding has no one accountable for closing it and will age without anyone noticing.

MATERIAL ml-band-cancellation Calibration errors in opposite directions cancel in the overall bias

The overall calibration bias and the worst declared band's bias point in opposite directions, and the band's is more than twice the size. A reader of the headline figure alone would conclude the model is close to calibrated; band by band it is over-confident at one end of the score range and under at the other.

Evidence to re-run: Band table over 12,000 scored record(s), one credit application at decision time, against the score bands declared for this engagement.

NOTABLE `ml-segment-spread` **One declared segment is served materially worse than the population as a whole**

Discrimination in "thin-file" is below the population figure by more than five points of AUC. This is a finding about the MODEL and not a fairness conclusion: whether a gap of this size is lawful or acceptable is a judgement about the population and the decision it feeds, which this review does not make.

Evidence to re-run: Per-segment AUC over the declared segment column, 12,000 scored record(s), one credit application at decision time.

10 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.

11 Not performed, and what each absence costs

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 — and for this family it is longer than usual, because several of its entries are permanent rather than pending.

NOT PERFORMED — AND WHY

Independent replication of the model, and every check resting on a replica

Machine-learning models in this family are gradient-boosted trees, random forests and neural networks. None of them is expressible as a declared specification — there are no coefficients to state — so no replica can be built and the replication fidelity gate has nothing to run. This is a property of how the model was built and no upload changes it. What this platform does perform on such a model is its outcomes, its calibration, its stability and its operating points, plus the evidence artefacts its owner can supply.

Explainability evidence for this model

NO EXPLAINABILITY EVIDENCE HAS BEEN DECLARED for this engagement, and this pack expects one. A machine-learning model has NO DECLARABLE SPECIFICATION — there are no coefficients to transcribe — so the ordinary route to knowing what a model relies on is closed for this family. The owner's own attribution artefact is the only evidence of it that exists, which is why this pack expects one and names its absence rather than passing over it. Supply the attribution export your own tooling produces — one row per feature, one value per row — with the method, the population it was computed over, who produced it and the document it is cited to. This platform does not compute one in its place.

layer1.macro-stress

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

layer2.ldp

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

layer3.counterfactual

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

layer4.migration

The four simulation layers act on a replica, and this family has none — not because one has not been built, but because a model with no declarable specification cannot have one. The layer is declared out of scope with its ground, rather than reported as not yet run.

Independent replication of the model, and every check resting on a replica

PERMANENT, on this platform's ground. A gradient-boosted machine, a random forest and a neural network are not expressible as a declared specification — there are no coefficients to state — so no replica can be built and the replication fidelity gate has nothing to run. This is a property of how the model was built and NO UPLOAD CHANGES IT; the capability ruling says so with a permanent gate rather than a pending one. A replica fitted to the model's input/output behaviour instead would agree with it exactly where it was sampled and diverge everywhere else, and the fidelity gate would pass because the fixture is where it looks

Anything requiring the model to be re-run: robustness under perturbation, adversarial probing, counterfactual sweeps

These need the model executed on inputs this platform generates, which is the runtime capability — permanently gated, not pending. Every figure in this pack is computed from outputs the model already produced. Perturbation robustness is the most-requested of the three and the most misleading to approximate: perturbing the SCORED POPULATION rather than the model measures the arithmetic of this pack, not the stability of the model

A fairness verdict — whether the selection-rate ratio between segments is acceptable

The ratio is REPORTED at the declared cutoff and permanently unjudged. The four-fifths rule is 29 CFR 1607.4(D), a US employment-selection guideline with no standing in model supervision, and shipping it as a bar would be this product asserting a legal standard in a domain it does not come from. Fraud's decision, unchanged. What a model must hold is the institution's own policy, declared as a claim

Explainability computed by this platform — SHAP values, attributions, partial dependence

Explainability artefacts are DECLARED EVIDENCE: supplied by the owner, hashed, tenant-scoped and re-verified on read, exactly like every other artefact here. This product will not compute them from a replica it authored, and that is the load-bearing refusal — computing an explanation of a model we reconstructed and then validating the reconstruction against itself is the independence hole this platform closed for specifications, arriving one artefact later. It would also be impossible here: there is no replica, which is the premise of this whole family

Any reading of feature importance as a statement about the world

An attribution is a statement about a MODEL and never about the world. A memo that let one read as the second would be making a causal claim nobody evidenced — that a feature DRIVES an outcome rather than that a model USES it — and the two are routinely confused in exactly the documents this platform reads

Why the score distribution moved

That it moved is measurable and is this pack's PSI row. WHY — population change, an upstream data change, or genuine regime change — is not recoverable from a scored population, and the three call for different action. A pack that named a cause would be guessing under a heading, and the guess would be acted on

Multi-class, ranking and regression models

This pack reads a SCORED POPULATION — a score per record and a realised binary outcome — which is what an adverse-action or eligibility model produces. A regressor's error distribution, a ranker's ordering quality and a multi-class model's per-class confusion have nothing in common with it or with each other, and a pack covering all of them would compute their intersection, which is almost nothing. Named here rather than silently mis-served: a second pack, with its own scoping

12 Reproducibility

Engine ml-metrics/1.0.0, metric pack "ml", thresholds from AlphaIQ machine-learning pack — each threshold cited to its own instrument.

The scored population is stored as supplied and content-hashed over its raw bytes; it is re-hashed on every read and a disagreement refuses the run rather than repairing it. The declaration is stored beside it and validated on resolution: the same bytes read against a different banding, a different reference window or a different segment column are a different run, not a better-formatted one.

Every figure in this memo is a pure function of the stored book and the stored declaration. No model was executed, no replica exists and no random draw was made, so a rerun over the same two artefacts reproduces this document exactly.



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 c71ca71ff0e5bfe3119529fe5c3f356f746a8d1503ff2a3701a97722ef7793b9

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