

Model validation memo — Motor & Property Severity GBM v6

Insurance — claim severity

Actuarial model validation

MODEL

Motor & Property Severity GBM v6

ENGAGEMENT

Sample — motor and property severity model review

PREPARED

2026-10-07

PREPARED BY

AlphaIQ CertifyIQ, computed by the product

RECOMMENDATION



Inconclusive — the evidence is insufficient

7 sections · 5 traced figures · 0 finding(s) · 8 test(s) named as not performed

SYNTHETIC SAMPLE

Every figure in this report was computed by CertifyIQ from a synthetic book — Closed motor and property claims, predicted against settled cost, by band and region. No real institution, model or customer is described. The book is `claims-severity-book.csv`; how it was generated, and the defects planted in it, are documented in `claims-severity-book-dgp.json`.

AT A GLANCE · SYNTHETIC SAMPLE

Motor & Property Severity GBM v6

Insurance — claim severity · Actuarial model validation

Inconclusive — the evidence is insufficient

No row in this pack was judged against a threshold: all 5 were reported without one, and each is listed as such in the results table. No defect was demonstrated and none was ruled out, so the honest conclusion is that the evidence is insufficient rather than that the model passed. Declaring the institution's own thresholds against this engagement is what turns these rows into judgements.

7

SECTIONS

5

TRACED
FIGURES

0

MATERIAL

0

NOTABLE

0

OBSERVATIONS

8

NOT
PERFORMED

Supervisory frame: this memo is built to "Actuarial Model Review (ASOP 56)", published under US · Actuarial Standards Board. 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 (claims-severity-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 | 5 Adequacy across declared segments |
| 2 The closed-claim book | 6 Obligations this memo does not establish, and checks not performed |
| 3 Predictive ability and discrimination | 7 Findings |
| 4 Calibration by band, and stability across development cohorts | |

1 Purpose and scope

This memo validates Motor & Property Severity GBM v6 — a claims severity model — against the closed-claim book extracted on 2026-06-30: 6,000 row(s) read and 6,000 closed claim(s) kept across 4 line(s) of business (commercial_property, household, motor_own_damage, motor_third_party), denominated in GBP. Every figure below is computed on that book and no other.

EVERY CLAIM IN IT IS CLOSED, and that is a refusal rather than a filter. An open claim carries a reserve, not a realised severity, and scoring a severity model against reserves measures the reserving process instead — so a

book carrying one is refused outright rather than quietly cut to its closed rows. A book cut silently is a book whose maturity nobody declared.

THE MATURITY WINDOW IS THE INSTITUTION'S: 365 day(s) from report. Claims that settle quickly are systematically smaller than claims that take years, so a book assembled without a declared window is a book about fast claims wearing the name of a book about claims. This platform ships no default window, because how long a line of business takes to settle is something the institution knows and this platform does not.

NOT ONE THRESHOLD IN THIS PACK IS THIS PLATFORM'S. ASOP No. 56 and the NAIC Model Bulletin both require that testing happen, be documented and be owned; neither publishes a number a figure passes against. Four of the five rows below are judged against a tolerance the institution declared, and any row with no declared tolerance is reported and NOT judged. The fifth — the adequacy ratio across declared segments — is reported and never judged at all.

THIS MEMO IS AN INPUT TO AN ACTUARIAL REPORT AND IS NOT ONE. ASOP No. 56's obligations fall on the actuary who issues that report: to understand the model, to disclose what it relies on, and to state its material limitations. Section 6 names each of those obligations with its section number and records that this memo does not establish them. It validates figures; it does not sign anything.

2 The closed-claim book

The book holds 6,000 closed claim(s). Realised severity: mean GBP 8,843, median GBP 3,350, largest GBP 1,303,402. The gap between the mean and the median is the shape of a severity distribution and is reported rather than smoothed: a handful of large losses move every mean-based figure in this memo, which is why the banded row in section 4 asks a different question from the pooled row in section 3.

169 claim(s) closed at nothing — reported, investigated, paid nil. They are realised outcomes and not missing ones, they are kept, and they are what the discrimination row's tie accounting in section 3 exists for.

0 row(s) were dropped as unreadable. Dropped rows are individual defects; the refusals that would have stopped this memo being written at all — an open claim, or a claim reported inside the maturity window — are refusals of the whole book and are named in section 1.

The institution declared a segment for the ratio in section 5: Claimant region, as the insurer classifies it, taking 3 value(s) in this book. It is DECLARED — this platform does not estimate a claimant's group from their surname and address, because whether to do that is a decision an institution takes under its own state's regulation with its own legal advice.

3 Predictive ability and discrimination

Across the whole book the model predicted a mean of GBP 7,541 against a mean realised severity of GBP 8,843 — an UNDER-estimate of 14.7%. The judged figure is the absolute value, because a tolerance cannot be compared to a signed quantity without deciding a sign; the DIRECTION is stated here because over by that much and under by that much fail the same bar and call for different conversations — an over-reserved book ties up capital and an under-reserved one fails to.

The paired t on the per-claim differences is -4.24 on 5,999 degrees of freedom, two-sided. Two-sided because over-estimating a severity and under-estimating one are both mis-estimating it, and which is worse is the institution's judgement rather than this platform's.

Ordering: of 17,982,792 comparable claim pair(s), the model ranked 14,214,971 in the right order and 3,767,802 in the wrong one, giving Somers' D of 0.5810 — 0.7905 on the AUC scale, where 0.5 is no ordering at all. Pairs tied on the REALISED amount are excluded from that denominator rather than counted as half-concordant: two claims that both closed at nothing are not a pair the model could have ordered rightly or

wrongly, and counting them would drag every model towards 0.5 in proportion to how many claims settled nil. The 19 pair(s) tied on the PREDICTION are counted, because those the model could have separated and did not.

These two rows answer different questions and a model can pass either while failing the other. A model that is right on average can order claims no better than chance; a model that orders claims perfectly can be wrong on every amount by the same factor. Neither figure substitutes for the other, and a reader carrying one away as the model's quality would be carrying half of it.

Severity bias (absolute, relative to mean realised)

MEASURED 14.73%

Severity bias (absolute, relative to mean realised): 14.73% measured on 6,000 closed claims carrying 6,000 claims with a realised severity. Benchmark source: $|\text{mean}(\text{predicted}) - \text{mean}(\text{realised})| \div \text{mean}(\text{realised})$ over the closed book, with the paired t on the per-claim differences reported beside it. TWO-SIDED, because over-estimating a severity and under-estimating one are both mis-estimating it: an over-reserved book ties up capital and an under-reserved book fails to, and which is worse is the institution's judgement rather than this platform's. NO THRESHOLD IS SHIPPED: ASOP No. 56 s. 3.6.1 requires sufficient testing that the model reasonably represents what is modelled and publishes no tolerance, so the bar is the institution's declared one. The SIGN is printed beside the figure, because over by 8% and under by 8% fail the same bar and call for different conversations.

Severity discrimination (Somers' D on [0, 1])

MEASURED 0.790

Severity discrimination (Somers' D on [0, 1]): 0.790 measured on 6,000 closed claims carrying 6,000 claims with a realised severity. Benchmark source: Somers' D over the claim pairs, mapped onto [0, 1] so it reads like an AUC: does a claim this model rated more expensive actually cost more. Pairs TIED ON THE REALISED AMOUNT are excluded from the denominator rather than counted as half-concordant — two claims that both closed at zero are not a pair the model could have ordered rightly or wrongly, and in a book with a mass at zero counting them would drag every model towards 0.5 in proportion to how many claims closed at nothing. Pairs tied on the PREDICTION are counted, because those the model could have separated and did not. NO THRESHOLD IS SHIPPED: no instrument in this lane publishes a discrimination floor for a severity model.

4 Calibration by band, and stability across development cohorts

Calibration, band by band, over the institution's DECLARED bands. Banded on the PREDICTION and never on the realised amount: banding on the outcome sorts claims by the very thing being predicted, so the mean realised amount inside a band is bounded by the band and the comparison would look good at the bottom and bad at the top whatever the model did.

GBP 0–1,000: 674 claim(s), predicted mean GBP 682 against realised mean GBP 1,200 — 43.2% out.

GBP 1,000–5,000: 3,068 claim(s), predicted mean GBP 2,571 against realised mean GBP 3,716 — 30.8% out.

GBP 5,000–20,000: 1,767 claim(s), predicted mean GBP 9,861 against realised mean GBP 11,566 — 14.7% out.

GBP 20,000–100,000: 473 claim(s), predicted mean GBP 35,982 against realised mean GBP 39,884 — 9.8% out.

GBP 100,000 and above: 18 claim(s), predicted mean GBP 136,274 against realised mean GBP 85,977 — 58.5% out.

The judged figure is the worst band: 58.5% in "GBP 100,000 and above", which holds 18 claim(s). No minimum claim count is applied to it, and that is the opposite of the rule applied to cohorts below. A thin cohort is an accident of when claims happened to be reported; a thin TOP band is the large-loss tail, which is the thing a severity model exists for — excluding it behind a sample-size rule would hide exactly what a reader came for. The count is printed instead, so the figure is judged with the evidence in view.

Stability across report cohorts, by calendar quarter of report. A model that has drifted with claims inflation shows a bias that WALKS in one direction across cohorts; a model that is simply biased shows the same bias in every

one. A single pooled figure cannot tell those apart, which is why the table is here and the spread is only its summary:

2023 Q1: 664 claim(s), under by 7.7%.

2023 Q2: 737 claim(s), under by 17.9%.

2023 Q3: 741 claim(s), under by 30.1%.

2023 Q4: 803 claim(s), under by 10.6%.

2024 Q1: 796 claim(s), under by 8.0%.

2024 Q2: 746 claim(s), under by 12.8%.

2024 Q3: 733 claim(s), under by 10.4%.

2024 Q4: 740 claim(s), under by 16.0%.

2025 Q1: 40 claim(s), under by 20.3%.

The spread between the widest and narrowest included cohort is 22.4%. It is TAIL-SENSITIVE and a reader should know it: a quarter that happens to catch a catastrophe carries a large bias on its own, so this number can be wide in a book with no drift and can move only a little in one with drift. The table above is the evidence; this figure is the summary of it. Excluding large losses would need an attachment point, and where a large loss begins is a decision the institution takes per line of business — which is why a dedicated tail treatment is a declared omission in section 6 rather than a default taken here.

Worst-band calibration error

MEASURED 58.50%

Worst-band calibration error: 58.50% measured on 6,000 closed claims carrying 6,000 claims with a realised severity. Benchmark source: The largest $|\text{mean}(\text{predicted}) - \text{mean}(\text{realised})| \div \text{mean}(\text{realised})$ across the institution's DECLARED severity bands. Banded on the PREDICTION and never on the realised amount: banding on the outcome sorts claims by the thing being predicted, so the mean realised amount inside a band is bounded by the band and the comparison is guaranteed to look good at the bottom and bad at the top whatever the model does. With no bands declared the figure is NOT COMPUTED and says so — a band structure decides what calibrated means here, and bands this platform chose would be it making that decision. An empty declared band is reported with its zero count rather than dropped, because a band the book does not occupy is a range the model was never tested in. NO MINIMUM CLAIM COUNT IS APPLIED to the worst band, which is the opposite of the cohort rule beside it and deliberate: a thin cohort is an accident of when claims were reported, whereas a thin TOP band is the large-loss tail, which is the thing a severity model is for — so the band's claim count is REPORTED beside the figure rather than used to exclude it.

Development-cohort bias spread

MEASURED 22.36%

Development-cohort bias spread: 22.36% measured on 6,000 closed claims carrying 6,000 claims with a realised severity. Benchmark source: $\max - \min$ of the signed relative bias across report-quarter cohorts. A model that has drifted with claims inflation shows a bias that walks in one direction across cohorts; a model that is simply biased shows the same bias in every one, and a single pooled figure cannot tell the two apart. Cohorts are CALENDAR QUARTERS OF REPORT — the one partition this pack chooses rather than reads, because a quarter is the period an insurance reserving cycle already runs on, so a cohort here lines up with a cohort in the institution's own reporting. A quarter carrying fewer than 30 claims is listed and EXCLUDED from the spread by name, because a spread driven by a quarter of four claims is a statement about one large loss. THE SPREAD IS TAIL-SENSITIVE EVEN SO: a quarter that catches a catastrophe carries a large bias on its own, so this figure can be wide in a book with no drift and move only a little in one with drift. The COHORT TABLE beside it is the evidence — drift reads as a bias that walks in one direction across cohorts — and this number is the summary of it.

5 Adequacy across declared segments

Adequacy — mean predicted over mean realised — within each segment the institution declared (Claimant region, as the insurer classifies it). Adequacy rather than a selection rate, because a severity model does not select anybody; what it can do unevenly is under-estimate one group's claims and not another's:

region_central: 2,020 claim(s), mean realised GBP 8,399, adequacy 0.852.

region_north: 2,012 claim(s), mean realised GBP 9,767, adequacy 0.788.

region_south: 1,968 claim(s), mean realised GBP 8,355, adequacy 0.931.

Lowest over highest is 0.85. IT IS REPORTED AND IT IS NOT JUDGED, and it never will be. The four-fifths rule familiar from this shape 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 insurance supervision — and shipping it here as a bar would be this platform asserting a legal standard in a domain it does not come from. What this ratio must hold is the institution's own policy, and this memo prints the number for the people who own that policy to read.

Two things this section deliberately does not do. It does not say whether a permitted rating variable is acting as a PROXY for a protected one: whether a variable is a proxy is a judgement about how the variable is produced, and no association threshold separates a proxy from a legitimately predictive factor. And it does not estimate the protected class. A ratio over a group this platform inferred from a surname and an address would be a decision taken on the institution's behalf, under a regulation that is theirs and not ours.

Adequacy ratio across declared segments

MEASURED 0.85

Adequacy ratio across declared segments: 0.85 measured on 6,000 closed claims carrying 6,000 claims with a realised severity. Benchmark source: Lowest ÷ highest ADEQUACY — mean predicted over mean realised — across the segments the institution declares. Adequacy rather than a selection rate, because a severity model does not select anybody; what it can do unevenly is under-estimate one group's claims and not another's. NO THRESHOLD IS SHIPPED AND NONE EVER WILL BE: the four-fifths rule familiar from this shape 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 insurance 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 and ML's, unchanged. AND THE SEGMENT IS DECLARED, NEVER ESTIMATED: this platform will not infer a claimant's group from surname and address, because whether to do that is a decision an institution takes under its own state's regulation with its own legal advice.

6 Obligations this memo does not establish, and checks not performed

Two different kinds of absence are recorded below, and a reader must not read them as one. The first is a set of OBLIGATIONS: things ASOP No. 56 and the NAIC Model Bulletin require of the actuary and of the institution, which no figure can establish and which this memo does not discharge. The second is a set of CHECKS this pack does not run, each with the reason it does not.

THE OBLIGATIONS, with the section that imposes each. Every one is established by evidence the institution supplies and records in the engagement's completeness checklist — a document, a signature, a retained record — and by nothing in this memo:

ASOP No. 56, s. 3.1 — Model Meeting the Intended Purpose — The model's intended purpose is stated, and the model is capable of it. Established by: the model documentation naming the purpose this severity model was built for, together with the actuary's assessment that its structure is capable of it, recorded as evidence. No figure establishes this: a model can be accurate at a question nobody asked

ASOP No. 56, s. 3.2 — Understanding the Model — The actuary understands the model's operation, dependencies and sensitivities. Established by: a documented walk-through of the model's basic operation, its important dependencies and its major sensitivities, signed by the actuary relying on it and recorded as evidence. This is the obligation a validation memo most often looks like evidence of and is not: a memo reporting figures is the reviewer's understanding, not the actuary's

ASOP No. 56, s. 3.3 — Reliance on Data or Other Information Supplied by Others — Reliance on data supplied by others is disclosed. Established by: the declared provenance of the claim book and of any exposure or rating data behind the model, with the extent of reliance stated, recorded as evidence. The book's own content hash

and supply record carry the provenance; the DISCLOSURE is a sentence in the actuarial report and is not produced here

ASOP No. 56, s. 3.4 — Reliance on Models Developed by Others — Reliance on a model developed by others is disclosed, with the extent stated. Established by: a statement of what the actuary relied on another party for and what they did to understand its general operation and major sensitivities, recorded as evidence alongside the vendor documentation. Bears directly on a vendor severity model and is never established by a run of this pack

ASOP No. 56, s. 3.6.4 — Reasonable Governance and Controls — Governance and controls over the model are in place. Established by: the institution's model governance record — change control, access, review cadence and who owns the model — recorded as evidence. This pack computes figures about one book on one day; whether anything governs the model between those days is a documentary fact

ASOP No. 56, s. 3.7 — Documentation — Documentation supporting the modelling work is prepared and retained. Established by: the retained modelling documentation itself, recorded as evidence with its hash. Named here because a pack that reported five figures and said nothing about retention would let a review pass with the one obligation nobody can reconstruct later left unmet

ASOP No. 56, s. 4.1 — Required Disclosures in an Actuarial Report — The actuarial report discloses the intended purpose, material limitations and reliances. Established by: the actuarial report itself, recorded as evidence once issued. This memo is an input to it and is not it: the disclosure obligation sits with the actuary issuing the report, and a validation memo that presented itself as discharging it would be this platform signing somebody else's work

NAIC Model Bulletin on the Use of Artificial Intelligence Systems by Insurers, adopted by the NAIC on 4 December 2023 — AS ADOPTED BY THE INSURER'S OWN STATE — A written AI systems programme, with governance, risk management, testing and documentation available to the regulator on request. Established by: the institution's written AI systems programme, recorded as evidence, together with a statement of which state's adoption of the bulletin binds it. A MODEL bulletin is not law anywhere until a state adopts it: as at the NAIC's implementation map of 1 April 2026 roughly half the jurisdictions had, and several states regulate insurer AI under their own frameworks instead. This pack therefore does not assert that the bulletin applies — the institution names its state, and the current adoption list is the NAIC's map rather than a constant in this product

ONE OF THOSE IS ABOUT THIS DOCUMENT AND IS WORTH RESTATING ON ITS OWN. ASOP No. 56, s. 4.1 requires an actuarial report to disclose the model's intended purpose, its material limitations and what it relies on. This memo is an INPUT to such a report and is not one: the obligation sits with the actuary who issues it and who is answerable for it, and a validation memo presented as discharging it would be this platform signing somebody else's work.

AND THE NAIC BULLETIN IS A MODEL BULLETIN, which binds an insurer only through their own state's adoption of it. This memo does not assert that it applies: the institution names the state whose adoption binds it, and the current adoption list is the NAIC's own implementation map rather than a constant inside this product — a count recorded here would go stale silently, which is the failure mode this platform exists to notice in other people's documents.

THE CHECKS THIS PACK DOES NOT RUN follow, each with the reason. They were written down before the pack was built rather than collected after it, so this register is a set of decisions and not a list of things that did not get done.

NOT PERFORMED — AND WHY

Anything about open claims

An open claim carries a RESERVE, not a realised severity, and scoring a severity model against reserves measures the reserving process rather than the model — a second model's validation wearing the first one's name. The parser REFUSES a book carrying open claims rather than dropping them, because a book quietly cut to its closed rows is a book whose maturity nobody declared

Whether the reserves were adequate

A triangle question, and a development triangle's cells are not independent — every test this platform performs over independent rows is invalid over one. Different family, different machinery, and naming it here is cheaper than a reader assuming a severity pack covered it

Rate adequacy and rate filing

A filed rate is a decision about the future, and no book of closed claims can be tested against it. The severity model feeds a rate; the rate is not what this pack reads

A fairness verdict on the adequacy ratio between segments

The ratio is REPORTED and permanently adjudged. The four-fifths rule is 29 CFR 1607.4(D), a US employment-selection guideline with no standing in insurance 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 and ML's, unchanged. What a model must hold is the institution's own policy

Whether a permitted variable is acting as a proxy for a protected one

Whether a variable is a proxy is a judgement about how the variable is PRODUCED — the leakage screen's argument one domain over — and no association threshold separates a proxy from a legitimately predictive rating factor. The association can be reported; the grade is refused

Anything requiring an estimated protected class

Estimating a claimant's race from their surname and address — BISG or any equivalent — to test a model is a decision an institution makes under its own state's regulation with its own legal advice. This platform computes a ratio over a segment the institution DECLARES and will not manufacture the protected class itself, because doing so would take that decision on their behalf

A separate treatment of the large-loss tail

Every figure here is computed over the whole book, and a severity model is usually judged on its tail. A tail analysis needs an ATTACHMENT POINT, and where a large loss starts is a decision an institution takes per line of business — the same shape as the band boundaries, which are declared for the same reason. The bands are the route to it today: declare a top band and its calibration row IS the tail row. A dedicated tail treatment with its own excess-of-loss reasoning is a separate story

Colorado's quantitative unfair-discrimination testing for external consumer data

3 CCR 702-10 is the closest thing to a computable requirement anywhere in insurance and is deliberately not cited here: it governs LIFE UNDERWRITING using external consumer data, which is neither this family nor this model type. Citing a real instrument at the wrong thing is worse than citing nothing, and a memo that named Colorado at a claims-severity model would be doing exactly that

7 Findings

No findings were raised. Every metric judged against a declared tolerance returned within it; rows reported without a declared tolerance were not judged and raise nothing, and the adequacy ratio is never judged.

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



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 618e183827bb24a46c7ac2c85ad2275c174b2b876b5e8877800cd684072d4031

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