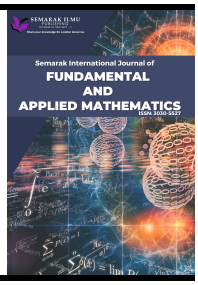
SEMARAK ILMU PUBLISHING

Semarak International Journal of Fundamental and Applied Mathematics

Journal homepage:
https://semarakilmu.my/index.php/sijfam
ISSN: 3030-5527



A Structural Calibration Framework for Tweedie Gradient Boosting in Zero-Inflated, Heavy-Tailed Regression

Jia He Lee[1], Srividhya Gunalan[2], Chee Nian Lee[1,*]

[1] School of Mathematics, Actuarial & Quantitative Studies, Asia Pacific University of Technology & Innovation, 57000, Kuala Lumpur, Malaysia
[2] School of Computing Science, KPR College of Arts Science and Research, 641407, Coimbatore, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 15 December 2025
Received in revised form 16 January 2026
Accepted 30 January 2026
Available online 4 February 2026

Keywords:
Gradient boosting; regression calibration; pure premium prediction; tweedie distribution

ABSTRACT

Accurate insurance pricing requires models that both rank risks effectively and produce well-calibrated loss estimates. Gradient boosting models trained with the Tweedie objective are widely used for pure premium modelling of insurance claim costs, particularly in health and other non-life portfolios. In highly skewed data with concentrated tail losses, these models often exhibit systematic miscalibration of aggregate and tail-level predictions despite strong discriminatory performance. This study investigates the structural sources of this behaviour, specifically the rigidity of the Tweedie mean-variance assumption in extreme skewness, and proposes a practical calibration framework for heavy-tailed insurance cost data. The empirical analysis is conducted in the context of U.S. medical insurance pricing, using person-level healthcare expenditure data from the Medical Expenditure Panel Survey (MEPS). The response variable represents annual aggregated insurer payments, reflecting pure premium estimation. We introduce PRISM, an additive correction architecture that combines a variance-stable regression model with a classifier-based certainty signal and a regularised meta-learner. Unlike global rescaling or monotonic post-processing, PRISM applies a localised residual adjustment that preserves ranking performance while improving absolute calibration. To evaluate calibration quality, we formalise the Root Mean Squared Calibration Error (RMSCE) and Mean Absolute Calibration Error (MACE) as bin-wise regression calibration diagnostics that summarise monetary miscalibration across the prediction range. Results show that PRISM consistently reduces exposure-weighted calibration errors and bias relative to standard Tweedie boosting and common post-hoc corrections, while maintaining comparable risk discrimination. Bootstrap confidence intervals confirm these improvements, indicating that the observed miscalibration under extreme skewness is primarily driven by structural modelling constraints.

* *Corresponding author.*
E-mail address: lee.cheenian@apu.edu.my

1. Introduction

1.1 Theoretical Background and Mathematical Challenges

Predictive models used in risk-sensitive applications need to satisfy two fundamental requirements: rank observations accurately, and produce well-calibrated predictions on the original outcome scale. Friedman [1] introduced Gradient Boosting Machines (GBMs), which provide exceptional discrimination by iteratively minimising a loss function, making them a dominant approach for tabular prediction problems. However, strong ranking performance does not guarantee accurate estimation of aggregate outcomes.

In insurance and cost-sensitive settings, calibration is commonly evaluated through the expected loss per unit of exposure. This quantity, known as the pure premium, is defined as the portion of an insurance premium specifically allocated to cover anticipated losses. Accurate pure premium estimation is essential for insurer pricing, reserve and aggregate risk assessment.

Insurance claims data exhibit a mixed distributional structure, with a point mass at zero and a continuous positive tail. Jørgensen [2] formalised the theory of dispersion models underlying this structure, and Smyth and Jørgensen [3] demonstrated that the Tweedie compound Poisson model provides a practical framework for insurance claims modelling. Yang *et al.,* [4] stated that GBMs are frequently trained using a Tweedie deviance objective.

Despite their widespread adoption, practitioners frequently observe systematic calibration deviations when applying Tweedie boosting models to highly skewed cost data. In particular, aggregate predictions often exhibit underestimation of tail losses, even when ranking performance remains strong. Lindholm and Wüthrich [5] showed that classical generalised linear models satisfy a balance property, ensuring that predicted aggregate premiums align with observed totals under standard assumptions. This discrepancy motivates a closer examination of the structural properties of the Tweedie objective in extreme skewness regimes. In the context of skewed cost distributions, modelling extreme right tails remains challenging; for example, Karlsson, Wang and Ziebarth [6] develop a specialised estimation approach for health expenditure data because conventional regression methods misrepresent the tail behaviour of such outcomes

1.2 The Boosting Flaw

The observed calibration limitation of Tweedie boosting models arises from a structural limitation of the modelling objective. Under the Tweedie framework, each observation is summarised by a single conditional mean parameter, which must simultaneously represent the probability of zero loss, the bulk of moderate claims, and extreme tail outcomes. Jørgensen [2] emphasised that the Tweedie distribution links these components through a single mean–variance relationship. In datasets with extreme loss concentration, a small fraction of observations contributes a disproportionate share of the total cost. Wüthrich and Merz [7] documented that such heavy-tailed insurance portfolios pose fundamental challenges for mean-based modelling approaches. When a single mean parameter is used, improvements in one region of the distribution often come at the expense of fit in other regions. As a result, the Tweedie objective may underestimate the tail or underestimate the bulk of the distribution, depending on where the likelihood optimisation concentrates its weight. This effect is further amplified by the log-link function. Manning [8] showed that nonlinear transformations introduce systematic bias when returning predictions to the original scale, a phenomenon commonly known as the retransformation problem. Basu and Rathouz [9] demonstrated that flexible link functions distort marginal effects under heteroscedasticity. In the context of Tweedie boosting, small errors on the link scale translate into multiplicative errors on the outcome scale, magnifying

underestimation when the fitted mean fails to align with the true cost concentration. These behaviours are not driven by implementation errors or numerical instability. They are predictable consequences of combining a single-mean Tweedie objective with a nonlinear link function in the presence of extreme skewness.

1.3 Limitations of Post-Hoc Calibration

As a baseline, we apply a global scalar smearing factor, estimated from out-of-fold predictions, to enforce mean alignment. We follow the standard post-hoc calibration practice described by Niculescu-Mizil and Caruana [10]. Isotonic Regression introduced by Vincze *et al.,* [11] minimises squared error subject to a monotonicity constraint. This method constructs a non-decreasing step map to minimise squared error. Isotonic regression is sensitive to small-sample fluctuations. Niculescu-Mizil and Caruana [10] showed it produces jump discontinuities (pricing cliffs), especially in sparse data regions. Dai *et al.,* [12] found that estimator bias depends on signal smoothness. This dependency leads to inconsistent error scaling and limits regression calibration.

Despite extensive use of Tweedie boosting models in insurance pricing, existing research primarily focuses on improving predictive accuracy or ranking performance, with limited attention to systematic aggregate miscalibration under extreme skewness. There is a lack of modelling frameworks that explicitly address the structural interaction between the single-mean Tweedie objective, nonlinear link functions, and concentrated tail losses. The objective of this study is to develop a calibration framework that corrects these effects without degrading risk discrimination. To this end, we propose the Probabilistic Regularisation Interpretable Stacking Model (PRISM), an additive ensemble architecture that incorporates a probability-based certainty signal to localise residual correction. The significance of this study lies in reframing calibration as a modelling design problem rather than a post-hoc adjustment, providing a practical and interpretable approach for improving pure premium estimation in highly skewed insurance cost data.

2. Methodology

2.1 Dataset and Preprocessing

We use data from the Medical Expenditure Panel Survey (MEPS) retrieved from the US Agency for Healthcare Research and Quality [13]. Total insurance payments serve as the target variable. As shown in Figure 1, the distribution is characterised by a discrete point mass at zero (13% zero-inflation) and a heavy-tailed positive component. The data follows a '5/50 rule,' exhibiting Pareto-like behaviour with extreme skewness in the tail. This dataset is an ideal testbed for Tweedie GBMs. The coexistence of a dense bulk and an extreme tail increases the risk that the Tweedie objective fails to represent the full distribution.

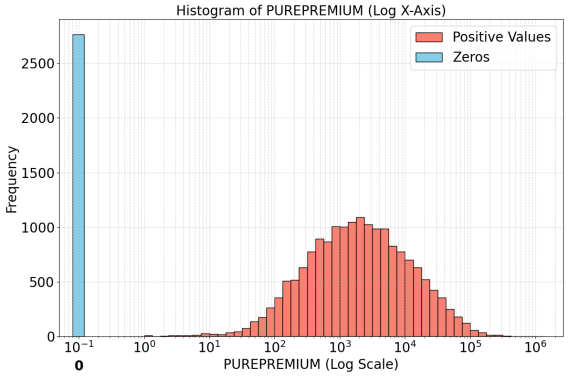


Fig. 1 Target variable distribution in logarithmic scale

2.2 Proposed Solution: The PRISM Architecture

PRISM (Probabilistic Regularisation Interpretable Stacking Model) adapts the additive stacking framework from Wolpert [14]. Wolpert defined the meta-learner's goal as "deducing the biases of the generalisers." The meta-learner "learns" the residual as a predictable error term, then applies a structural correction. Stacking provides a suitable topology for the PRISM framework. Instead of stacking several regression models, we decompose the conditional expectation into two orthogonal components. These are the Magnitude Anchor (biased Tweedie GBMs) and the Probability Corrector defined in Eq. (1).

$$F(x) = f(x) + \beta \cdot (g(x) - C) \tag{1}$$

2.2.1 The magnitude anchor

Formally, we define the base estimator $f(x)$ (the "Magnitude Anchor") as the gradient boosting regressor that minimises the exposure-weighted Tweedie Deviance over the training data. Specifically, the estimator approximates the conditional expectation $E[Y|X]$, by optimising the following objective function $J(f)$ shown in Eq. (2).

$$J(f) = \sum_{i=1}^{n} w_i \cdot d(y_i, f(x_i); p) \tag{2}$$

where w_i denotes the exposure weight and $d(y, \mu)$ represents the unit deviance for a Compound Poisson-Gamma process ($1 < p < 2$), defined by Jørgensen [2], shown in Eq. (3).

$$d(y, \mu) = 2\left(\frac{y^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right) \tag{3}$$

By minimising this objective $J(f)$, the model enforces the rigid mean-variance relationship $\mathrm{Var}(Y|X) = \phi(E[Y|X])^p$. This enables the model to capture complex non-linear covariate interactions (the "Magnitude" signal). The convexity of the deviance surface ensures that the

estimator structurally averages over the zero-mass and the heavy tail, leading to the systematic underestimation pattern described in Section 1.2.

2.2.2 The zero-threshold probability corrector

The corrective term $(g(x))$ formally designated as the Zero-Threshold Probability Corrector, is a standard LGBM binary classifier trained to predict the probability of a non-zero outcome $(P(Y > 0))$. We term this a "Zero-Threshold" estimator because it discriminates strictly between the point-mass at zero and the positive domain, providing a focused "Certainty Signal" orthogonal to the Anchor's magnitude estimate.

In heavy-tailed distributions, high severity often correlates with high event certainty. The Anchor model typically dilutes this signal by averaging the zero-mass into the mean estimate. Our Probability Corrector re-injects this information. We use the probability score as a continuous feature. We apply a `RandomizedSearchCV` to optimise the classifier specifically for `precision`. This ensures that surcharges are applied only when risk certainty is high. The optimisation converged on a `scale_pos_weight` of approximately 0.15, which aligns with the inverse class ratio (13% zero vs. 87% positive). This effectively down-weights the majority class to prevent the estimator from defaulting to a trivial "always-positive" prediction.

2.2.3 The pivot point (C)

We introduce a learnable parameter C to transform the probability score into a signed residual $R(x) = g(x) - C$. We posit a latent correction function $\mathcal{H}$ that maps risk certainty to the required magnitude adjustment. The PRISM formulation is derived via a First-Order Taylor Expansion of this function around a neutral pivot point C, expressed in Eq. (4).

$$H(P) \approx H(C) + H'(C)(P - C) + O((P - C)^2) \tag{4}$$

We enforce two structural constraints to simplify this expansion:

1. Vanishing Intercept $\mathcal{H}(C) = 0$: We define C as the root of the bias function (where the Anchor is unbiased), forcing the zeroth-order term to vanish by construction.
2. Linear Truncation: We discard higher-order curvature terms $\mathcal{O}((P - C)^2)$ to prevent overfitting, retaining only the first-order gradient $\beta = \mathcal{H}'(C)$.

The probability score P is centred around a learnable parameter C, transforming it into a signed residual $\beta(\mathrm{P} - \mathrm{C})$. If $P > C$, the term is positive. The model generates a positive magnitude correction to fill the underestimation gap. If $P < C$, the term is negative. The model generates a negative magnitude adjustment (correcting for zero-inflation).

2.2.4 The constrained meta-learner

The parameters are learned via Non-Negative ElasticNet optimisation as proposed by Zou and Hastie [16]. We use Gaussian (Least Squares) loss instead of Tweedie Deviance. This ensures numerical stability over the real domain $\mathbb{R}$ and allows for negative corrections. This approach enforces additive consistency. It prevents the reintroduction of multiplicative gradient cliffs.

β and C would be collinear in an unconstrained system. This collinearity arises because the correction term expands to $\beta P - \beta C$. The product $-\beta C$ acts as a dependent intercept. This allows the slopeβ to scale arbitrarily based on shifts in the pivot. The ElasticNet penalty $\lambda\left(\left|\left|\beta\right|\right|_1 + \left|\left|\beta\right|\right|_2^2\right)$ breaks this redundancy. It compels the learner to find the most efficient correction vector. The penalty taxes the magnitude of β. This forces the optimisation to converge on the specific centre. C that yields maximum covariance correction for minimum coefficient cost. We call this the "Stability plateau". Here, C is the unique pivot maximising explanatory power per unit of β.

We determine the optimal C^* via a 1D Grid Search on out-of-fold training set, we use the centre of the overlap area of the neutral bias range and the low error (RMSCE) range. The meta-learner is fully re-trained at each pivot point *C* to solve for the conditional optimal $\hat{\beta}$.

2.3 Benchmark Calibration Protocols

To rigorously evaluate the effectiveness of the proposed architecture, we compare it against standard post-hoc calibration techniques currently employed in actuarial practice to mitigate predictive bias.

2.3.1 Scalar Smearing (Global bias correction)

Scalar Smearing is a multiplicative correction method designed to enforce the global solvency constraint. We compute a single scalar factor ϕ from the ratio of observed to predicted aggregate losses, as defined in Eq. (5).

$$\phi = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} \hat{y}_i} \tag{5}$$

The calibrated prediction is $\hat{y}'_i = \phi \cdot \hat{y}_i$. This ensures the global first moment is unbiased ($E[\hat{Y}'] = E[Y]$) but imposes a uniform linear transformation. It fails to address local non-linearities in the bias manifold. This often leaves tail risks underpriced and overcorrects the distribution body.

2.3.2 Isotonic Regression (Non-parametric mapping)

Isotonic Regression is a non-parametric approach [11]. It learns a monotonic mapping function $h(\cdot)$ to minimise the squared error between uncalibrated scores and targets, as given by Eq. (6).

$$\min \sum_{i=1}^{n} \left(y_i - h(\hat{y}_i)\right)^2 \quad \text{subject to} \quad \hat{y}_i \leq \hat{y}_j \Rightarrow h(\hat{y}_i) \leq h\left(\hat{y}_j\right) \tag{6}$$

Solved via the Pool Adjacent Violators Algorithm (PAVA). This method partitions the prediction space into bins of constant value to average out local noise. Niculescu-Mizil and Caruana [10] proved that Isotonic Regression is superior for probability calibration, yet it produces discrete "step artefacts" or plateaus in regression contexts with sparse data. These discontinuities introduce artificial volatility into the pricing structure. They violate the actuarial requirement for smooth risk differentiation.

2.4 Metric Definition
2.4.1 Conceptual foundation

Calibration in regression refers to the agreement between predicted conditional means and realised outcomes across the prediction range. Formally, a regression model is calibrated if $E\left[Y \mid \hat{Y} = \hat{y} \right]$, a definition that traces back to Dawid [17], early work on probabilistic forecast reliability and Murphy & Winkler [18] was later formalised within the broader framework of forecast calibration, Gneiting *et al.,* [19] introduced proper scoring rules.

Recent machine learning literature operationalises calibration through bin-wise reliability analysis, most notably, Guo *et al.,* [20] formalised the Expected Calibration Error (ECE) for classification. ECE is defined for Bernoulli outcomes and probability forecasts, its core principle is comparing empirical outcomes and model predictions within stratified bins. Kuleshov *et al.,* [21] showed that calibration can be defined for continuous predictions through conditional expectations. Chung *et al.,* [22] further demonstrated that minimising point-wise loss functions, such as pinball loss, does not guarantee calibrated predictive distributions, motivating explicit bin-wise calibration diagnostics. Levi *et al.,* [23] systematically evaluated calibration in regression tasks and emphasised that, unlike classification, no canonical scalar calibration error metric exists for continuous outcomes.

Building on this literature, we define two complementary regression calibration error metrics that summarise bin-wise deviations on a monetary scale: the Root Mean Squared Calibration Error (RMSCE) and the Mean Absolute Calibration Error (MACE).

2.4.2 Bin-wise regression calibration framework

Let $\{(y_i, \hat{y}_i, w_i)\}_{i=1}^{n}$ denote observed outcomes, model predictions, and exposure weights, respectively. Predictions are sorted and partitioned into *K* disjoint bins $\{B_k\}_{k=1}^{K}$ of approximately equal mass. For each bin *K*, we defined: $\overline{y_k} = \frac{\sum_{i \in B_k} w_i y_i}{\sum_{i \in B_k} w_i}$, $\overline{\hat{y}_k} = \frac{\sum_{i \in B_k} w_i \hat{y}_i}{\sum_{i \in B_k} w_i}$, $\omega_k = \frac{\sum_{i \in B_k} w_i}{\sum_{i=1}^{n} w_i}$. Here, $\overline{y_k}$and $\overline{\hat{y}_k}$ represent the empirical and predicted conditional means within bin k, ω_k denotes the exposure-normalised bin weight.

2.4.3 Root mean squared calibration error (RMSCE)

We defined the Root Mean Squared Calibration Error (RMSCE) as shown in Eq. (7).

$$RMSCE = \sqrt{\sum_{k=1}^{K} \omega_k \cdot (\overline{y_k} - \overline{y_k})^2} \tag{7}$$

We calculate RMSCE across 20 exposure-weighted quantiles. Each bin represents 5% of the portfolio. RMSCE corresponds to an L2 aggregation of bin-wise calibration deviations. Due to the squared error structure, RMSCE is highly sensitive to large deviations occurring in sparsely populated or high-severity regions of the prediction space. In heavy-tailed insurance loss distributions, this property makes RMSCE particularly suitable for assessing solvency-relevant tail miscalibration, where underestimation of extreme losses can have disproportionate financial consequences.

2.4.4 Absolute mean calibration error (MACE)

To complement RMSCE, we define the Mean Absolute Calibration Error (MACE) in Eq. (8).

$$\text{MACE} = \sum_{k=1}^{K} \omega_k \left| \overline{\hat{y}_k} - \overline{y_k} \right| \tag{8}$$

MACE applies an L1 aggregation to bin-wise calibration deviations. Unlike RMSCE, MACE is less sensitive to isolated extreme errors and instead reflects the average absolute monetary miscalibration across the portfolio. This makes MACE directly interpretable as an expected premium-level deviation and more robust to shock losses, aligning with traditional actuarial notions of average pricing adequacy.

Conceptually, MACE can be viewed as a regression analogue of the Expected Calibration Error (ECE) proposed by Guo *et al.,* [20], adapted from probability forecasts to continuous outcomes. While ECE measures absolute differences between predicted confidence and empirical frequency in classification, MACE replaces probabilities with conditional means and incorporates exposure weighting, consistent with Dawid [17] and Kuleshov *et al.,* [21] in regression calibration theory.

3. Experimental Results

3.1 Baseline Underestimation and Hyperparameter Sensitive Test

Table 1 confirms that GBMs significantly outperform the Tweedie Generalised Linear Models (GLMs) in risk segmentation, with LGBM achieving the highest Normalised Gini (0.6121) and Tweedie Deviance (0.2775) compared to the GLM baseline (0.5811; 0.2565). However, this superior ranking comes at the cost of severe calibration limitation: the GLM maintains aggregate neutrality (+1.17% bias), the Tweedie GBMs exhibit systematic underestimation, with LightBoost (LGBM) and XGBoost (XGBM) underestimating the portfolio by -15.16% and -24.99%, respectively.

Table 1
Comparison of Raw Tweedie GBMs with Tweedie GLM

Model Architecture	Norm. Gini (↑)	RMSE (↓)	Mean Tweedie Dev (↓)	Tweedie Dev (↑)	Aggregate Bias % (→0)
Tweedie GLM	0.5811	15,320	160.09	0.2565	1.17%
Tweedie LGBM	0.6121	15,128	155.56	0.2775	-15.16%
Tweedie XGBM	0.6069	15,232	161.37	0.2506	-24.99%

To verify that this deficit is structural rather than a result of regularisation artefacts, we conducted isolated stress tests (Table 2). Removing regularisation (λ=0) yielded negligible improvement (0.01% difference), and increasing the computational budget (n=1000, η=0.01) slightly exacerbated the bias to -15.49%. These results confirm that the underestimation is not a parameter tuning issue but a fundamental property of the Tweedie boosting objective in heavy-tailed regimes.

Table 2
Sensitive test on hyperparameter

Model Architecture	Regularisation	Hyperparameters	Norm. Gini (↑)	Aggregate Bias % (→0)
Tweedie LGBM	$\lambda = 1$	$n = 200, \eta = 0.05$	0.6121	-15.16%
	$\lambda = 0$	$n = 200, \eta = 0.05$	0.6121	-15.16%
	$\lambda = 1$	$n = 1000, \eta = 0.01$	0.6122	-15.49%
Tweedie XGBM	$\lambda = 1$	$n = 200, \eta = 0.05$	0.6069	-25.00%
	$\lambda = 0$	$n = 200, \eta = 0.05$	0.6069	-25.00%
	$\lambda = 1$	$n = 1000, \eta = 0.01$	0.6132	-25.64%

3.2 Select Pivot point C

To validate the well-posedness of the additive formulation, we mapped the objective function's error surface relative to the probability centring parameter (C) using a systematic grid search on the 5-Fold Out-of-Fold (OOF) training set. Figure 2 visualises this dual-objective landscape: the red dashed line tracks the Aggregate Bias (%), illustrating the linear sensitivity of aggregate calibration to the pivot point, the solid green line represents the Calibration Error (RMSCE), capturing the structural stability of the estimator.

Figure 2 upper panel, reveals a distinct "Stability Region" in the RMSCE landscape. For the LGBM variant, the result is a wide, convex plain rather than a sharp local minimum. We identified a "Robust Solution Space" ($C \in [0.04, 0.27]$) where the model simultaneously satisfies the strict neutrality constraint ($\pm 1.5\%$) and the optimal error tolerance.

Figure 2 bottom panel illustrated critical divergence in the XGBM variant. Unlike the smooth convexity of LGBM, the XGBM landscape exhibits an "Optimisation Divergence": the global mathematical minimum for RMSCE corresponds to a biased state (OOF Bias: -2.5%). Consequently, strictly enforcing the original neutrality constraints resulted in an empty solution set for XGBM. To resolve this, we relaxed the neutrality constraint to $\pm 2.5\%$ and the optimal error tolerance to 10% (min +10%). This adjustment revealed a "Relaxed Solution Space" for XGBM ($C \in [0.04, 0.28]$) that overlaps almost entirely with the LGBM robust zone. To ensure architectural consistency, we selected the centroid of the solution space C=0.16 for both models.

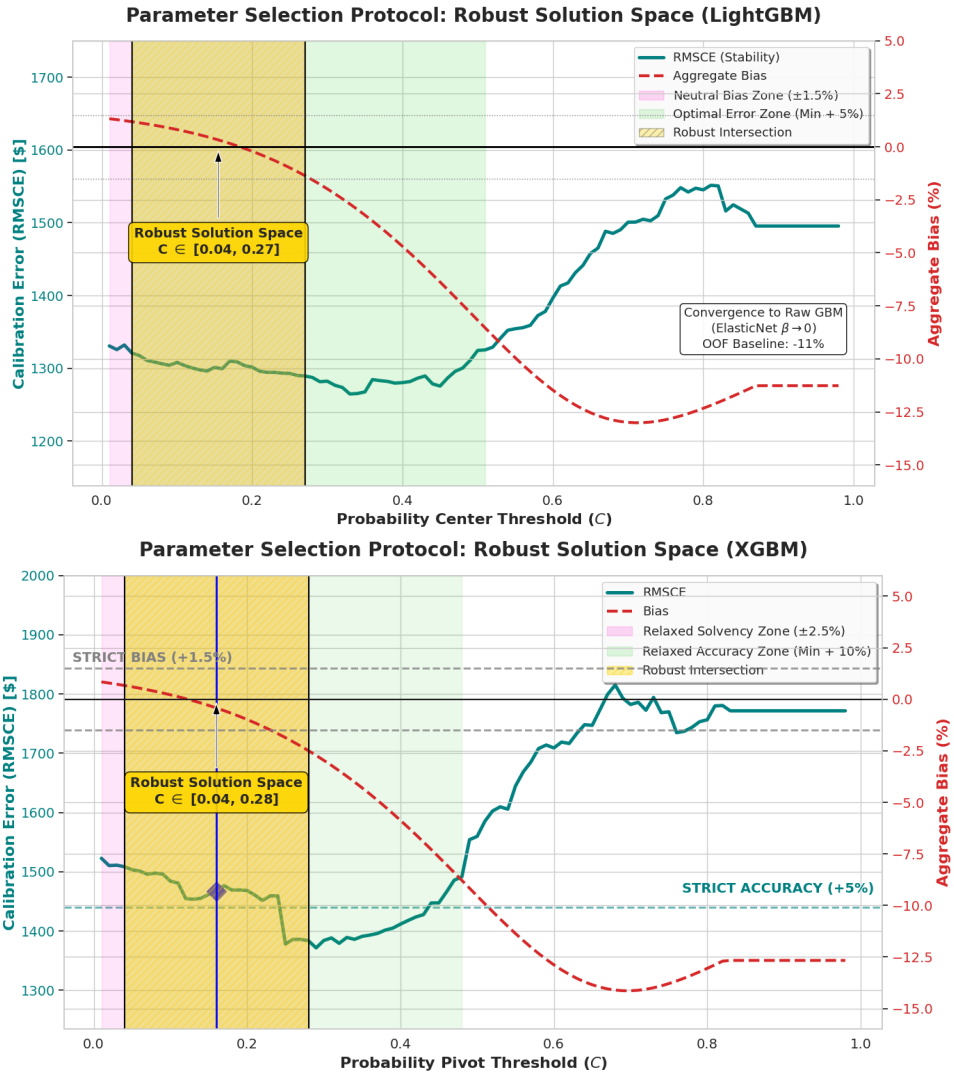


Fig. 2 LGBM robust solution space (upper panel); XGBM relaxed solution space (bottom panel)

3.2 Visual Forensic Analysis

3.2.1 Macro-calibration comparison

Figure 3 compares the marginal calibration curves on a log-log scale. The Raw Model (Blue) systematically sagged below the identity line, confirming structural underestimation in the distribution bulk. Isotonic Regression (Orange) corrects the mean but introduces severe "step artefacts," notably collapsing Bins 3–6 and Bins 13–16 into two static scalar clusters. This quantisation erases risk differentiation within these cohorts. In contrast, PRISM LGBM (Green) and Scaler Smearing have a uniform distance between each bin in the log scale visual plot.

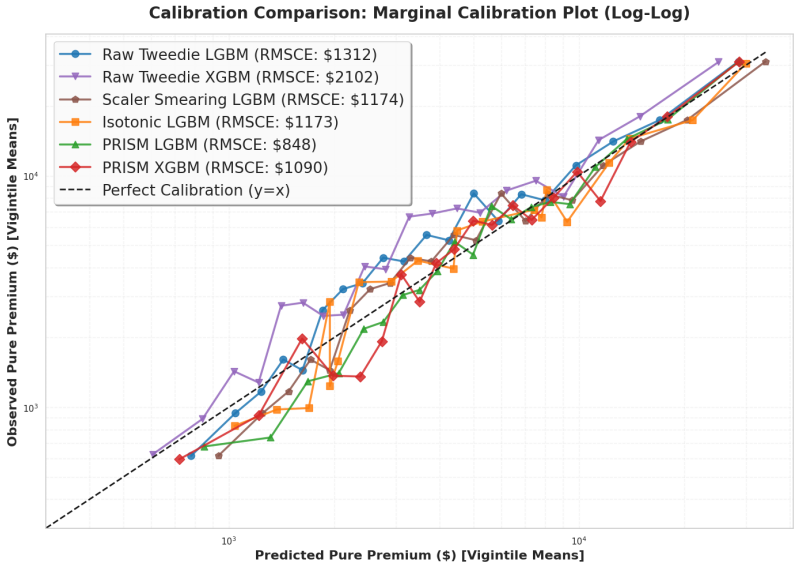


Fig. 3. Compare different calibration methods on Tweedie GBMs

3.2.3 Structural bias spectrum

The Bias Spectrum plot (Figure 4) deconstructs the aggregate error into 20 risk-stratified vigintiles. The Raw Tweedie LGBM exhibits a distinct "U-shaped" profile: underestimation peaks in the distribution bulk (Bin 13: -40.28%), the high-gradient tail approaches near-neutrality (Bin 19: -2.14%). Raw Tweedie XGBM shares this bulk underestimation but displays a more volatile "zig-zag" structure, suffering severe failure in the extreme tail (Bin 20: -19.12%).

The Scalar Smearing spectrum retains the "U-shaped" error profile of the base learner. It indiscriminately lifts the predictions, over-correcting the high-risk ends (Bin 16-20), leaving the core portfolio (Bins 6–13) systematically underpriced by -10% to -25%. The Isotonic Regression spectrum exhibits a "Zig-Zag" oscillation characteristic of non-parametric step functions. The bias flips unpredictably from positive to negative in adjacent groups (e.g., Bin 4 to 5, and Bin 6 to 10).

PRISM substantially improves calibration in the distribution bulk without introducing artificial volatility. In the critical Bin 13, PRISM restores calibration from the raw baseline of -40.28% to a near-perfect -1.23%, Bin 14 is corrected to -0.08%. However, the current zero-threshold corrector reaches its topological limit in the extreme tail, where Bin 20 retains a -8.14% deficit.

The Raw Tweedie XGBM displays a highly erratic ("zig-zag") error profile compared to the smoother LGBM baseline. The model exhibits a severe collapse in the distribution bulk, where underestimation plunges to -50.76% in Bin 10 (average actual: $6,652). This deficit persists into the extreme tail, with the top 5% of risks (Bin 20, average actual: $31,006) remaining underpriced by -19.12%. PRISM XGBM effectively neutralised this structural underestimate, restoring the aggregate bias to +0.85%. In the extreme tail (Bin 20), the architecture reduces the underestimation from -19.12% to -7.64%, converging to a similar structural limit as the PRISM LGBM variant. Crucially, PRISM preserves the inherent volatility of the base learner: the localised "zig-zag" pattern persists in lower risk bands (Bin 5: +75.04%), confirming that PRISM acts as a magnitude corrector but does not artificially smooth the local variance of unstable base estimators.

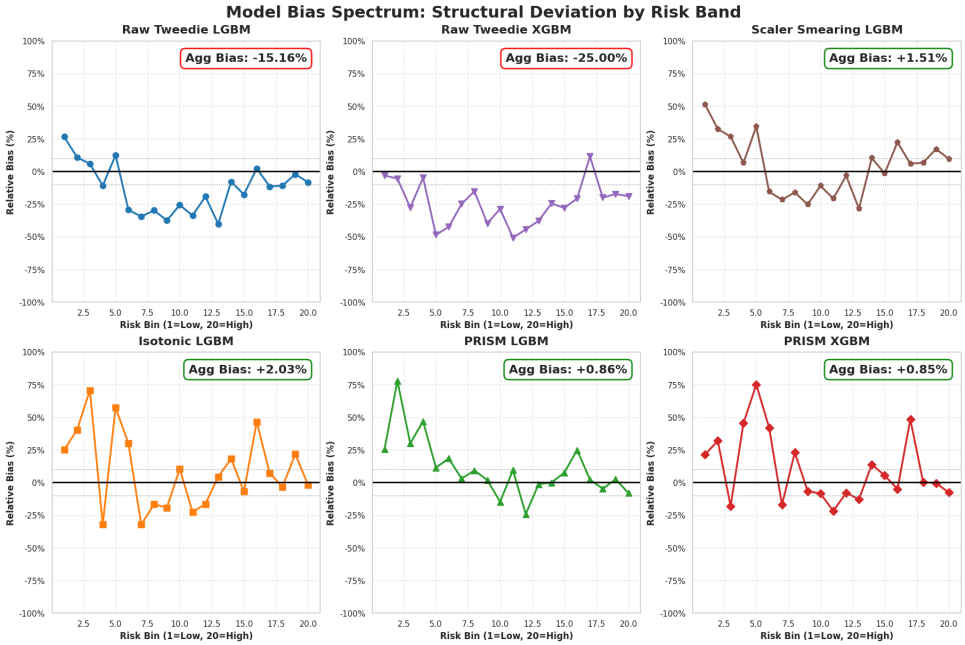


Fig. 4. Relative bias spectrum plot

3.3 Quantitative Analysis

3.3.1 Evaluation metrics result

Table 1 isolates the impact of topological choices versus post-hoc calibration. Scalar Smearing acts only as a global patch. It reduced global bias to 1.506%, yet its high RMSCE and MACE ($1173.65; $845.44) compared to PRISM ($853.28; $577.57). Isotonic Regression underperformed in this high-sparsity regime. It corrected global bias to +2.03% but yielded the highest calibration error (RMSCE: $1173.09). PRISM (LGBM) achieved the lowest RMSCE ($853.28), 27% reduction relative to Isotonic and Scaler Smearing; lowest MACE ($577.57), 36% and 32% reduction to Isotonic and scaler smearing. The lowest RMSE ($15,100) and Mean Tweedie Deviance (0.2907). PRISM retained 99.66% of the base learner's ranking power, delivering accurate financial predictions. PRISM (XGB) results validate the architecture's generalisability. The raw Tweedie XGBM exhibited a severe -25% structural bias. The PRISM meta-learner successfully normalised this to +1.7461%.

Table 3
Comparison of calibration methods

Model	RMSCE ($)	MACE ($)	RMSE ($)	Bias (%)	Mean Tweedie Dev.	Tweedie D2	Norm Gini
PRISM LGBM	847.62	578.74	15100	0.864	152.74	0.2907	0.6100
PRISM XGBM	1089.93	771.70	15126	0.851	153.92	0.285	0.6037
Isotonic LGBM	1173.09	903.97	15074	2.033	153.62	0.2866	0.6086
Scaler Smearing LGBM	1173.65	845.44	15124	1.506	153.00	0.2894	0.6121
Raw Tweedie LGBM	1311.81	990.96	15128	-15.164	155.56	0.2775	0.6121
Raw Tweedie XGBM	2101.99	1590.04	15232	-25.00	161.37	0.2506	0.6069

3.3.2 Bootstrap confidence intervals

Bootstrap-based RMSCE estimates are systematically higher than plug-in test-set values due to the extreme tail sensitivity of squared calibration loss under heavy-tailed resampling. Figure 5 reports the absolute calibration levels with uncertainty. PRISM achieves an RMSCE of $1349.53 (95% CI: [844, 2030]) and an MACE of $949.63 (95% CI: [653, 1304]), substantially outperforming Isotonic Regression (RMSCE: $1438, [727, 1326]; MACE: $996.65, [714, 1342]) and scaler smearing (RMSCE: $1608.15 [1050, 2337]; MACE: $1084.672499 [751, 1497]). Aggregate bias is statistically indistinguishable from zero.

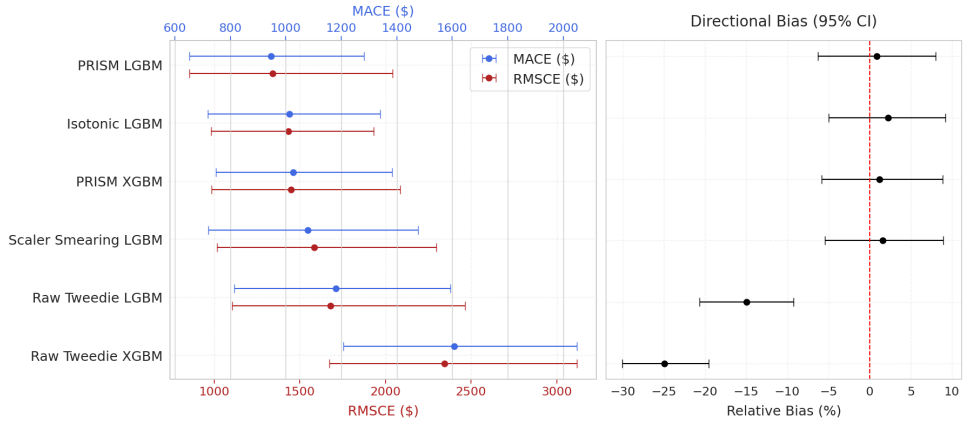


Fig. 5 95% Confidence interval of 1000 iteration Bootstrap for MACE, RMSCE, Bias

3.3.3 Response granularity and continuity

Figure 6 shows the smoothness of the calibrated estimators by measuring the cardinality of the prediction space and the magnitude of local discontinuities (step sizes) in the sorted response curve.

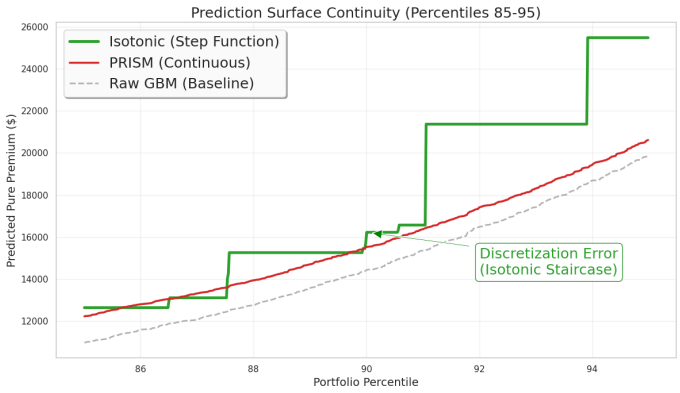


Fig. 6 Step effect on the Isotonic LGBM prediction

Table 4 shows Isotonic Regression baseline collapsed the continuous output space into a coarse piecewise constant function with a cardinality of only 59 unique values. This extreme quantisation introduced massive local discontinuities, with a maximum step of $12,119 and an average step of $732, effectively reducing the regression to a low-resolution binning exercise. In contrast, PRISM preserved the high-resolution topology of the base learner, generating 6,080 unique values with a negligible mean step size of $9.77.

Table 4
Prediction smoothness

Model Name	Unique Predicted Values (Granularity)	Max Prediction Cliff (Max Step)	Avg Step Size
Isotonic Regression LGBM	59	$12,119.03	$732.82
PRISM LGBM	6080	$3,679.60	$9.77
Raw Tweedie LGBM	6065	$3,715.98	$10.11
Smearing LGBM	6065	$4,446.14	$12.10

3.3.3 Structural Stability Analysis

Table 5 verify that the PRISM architecture is robust across different GBMs implementations, we conducted a 5-fold cross-validation of the meta-learner weights for both the LGBM and XGBM backbones (Table 5). The results confirm a highly stable additive topology in both cases. The PRISM XGBM variant demonstrated exceptional rigidity, with the Tweedie Anchor retaining a mean coefficient of 0.993 and a near-zero coefficient of variation (1.15%). This near-unity weight indicates that the meta-learner preserves the base predictions entirely, utilising the Probability Corrector as a pure, additive surcharge.

The PRISM LGBM variant exhibited similar stability, with a Tweedie Anchor coefficient of 0.915 and a low variation of 2.10%. The Probability Corrector weights were consistent across both architectures (CV < 7%), confirming that the mechanism for recovering tail mass is statistically reliable regardless of the underlying base learner.

The consistent low variance across both architectures demonstrates that the PRISM framework effectively stabilises the risk signal for different gradient boosting implementations (LGBM and XGBM) without structural modification.

Table 5
PRISM structural stability analysis

Model	Component	Mean Weight	Std Dev	Coefficient of Variance
PRISM LGBM	Tweedie LGBM	0.9146	0.0192	2.10%
	Prob Corrector	3,412.57	236.96	6.94%
PRISM XGBM	Tweedie XGBM	0.9928	0.0114	1.15%
	Prob Corrector	3,670.57	163.95	4.47%

3.4 Post Hoc Analysis

To verify that the performance of the PRISM architecture was not dependent on hyper-specific parameter tuning, we conducted a post-hoc sensitivity analysis on the Test Set by varying the probability pivot C from 0.04 to 0.28 (Table 6).

Table 6
Sensitive test on pivot point C

Pivot Point (C)	Aggregate Bias (%)	RMSCE ($)	Norm Gini	Status
0.04	1.84	846.13	0.6112	(Strongest Ranking)
0.08	1.59	812.67	0.6108	
0.12	1.27	845.74	0.6104	
0.16	0.86	847.62	0.6100	Selected in Result (Robust Centre)
0.2	0.36	826.52	0.6095	
0.24	-0.25	830.27	0.6092	
0.28	-1.00	788.12	0.6087	(Lowest Error)

The aggregate bias decreases linearly as the pivot point C increases. The model transitions from a conservative surplus (+1.84% at C=0.04) to a slight deficit (-1.00% at C=0.28). The selected parameter ($C = 0.16$) sits comfortably in the "Safe Zone" (+0.86%), prioritising solvency over aggressive price cutting.

The calibration error (RMSCE) exhibits a robust plateau, remaining consistently low (range: 788–853) across the entire spectrum. Notably, while an aggressive threshold of C=0.28 yielded the lowest absolute error ($788), it resulted in a negative bias (-1.0%), which poses an inadequate risk.

3.4.2 Sensitive test of the L1/L2 regularisation of the Elasticnet

A low regularisation strength (α = 0.1) was selected because the ensemble consists of only two base learners, requiring only a light touch to prevent over-fitting without distorting the underlying actuarial signals. The L_1/L_2 ratio was set to 0.9 (mostly Lasso) to ensure the meta-learner maintains a high degree of sparsity, effectively isolating the probability corrector's signal from the regression noise. As shown in Table X, reducing the L_1/L_2 ratio leads to a systematic underestimation of aggregate losses (up to 2.3%), confirming that a strong Lasso penalty is mathematically necessary to maintain the 'corrective' power of the PRISM architecture.

Table 7
Meta-learner Sensitivity Analysis on $\boldsymbol{L_1}$ Ratio ($\boldsymbol{\alpha}$ = 0.1)

L1_ratio	Aggregate Bias (%)	RMSCE ($)	Norm Gini	Tweedie LGBM Coefficient	Prob Corrector Coefficient (β)	Status
0.1	-2.36	797.53	0.6124	0.976441	1772.02	Strongest Ranking
0.5	-1.06	778.75	0.6115	0.958136	2113.94	
0.9	0.86	847.62	0.6100	0.931068	2619.52	Selected in Result

4. Discussion

4.1 The limitation of Post-Hoc Calibration

Scalar Smearing and Isotonic Regression are standard actuarial techniques for correcting global bias, our empirical results compel their rejection in the context of heavy-tailed Tweedie GBM. The limitations of these methods are not merely statistical, but structural.

4.1.1 The limitations of Scalar Smearing

Scalar Smearing is not well suited to this setting because it addresses a complex distributional bias using a single linear multiplier. While the log-link function serves as an amplifier, the fundamental underestimation arises from the rigidity of the Tweedie objective function

The Tweedie loss minimises deviance based on a fixed variance power parameter ϕ. This imposes a strict assumption that a single mean-variance relationship must hold across the entire manifold. However, insurance data is characterised by extreme skewness and zero-inflation. Faced with this tension, the Gradient Boosting machine fails to capture the aggregate tail cost, leading to a systematic underestimation of the distribution bulk. Our results show the best-performing LGBM model exhibits a -15.16% aggregate deficit.

This confirms that the bias structure is non-linear (convex). A single scalar ϕ can shift the bias curve vertically to satisfy the global mean condition ($E[\hat{y}] = E[y]$), but it cannot flatten the curve itself. Consequently, the model remains miscalibrated in the core portfolio segments, leaving significant volume systematically underpriced.

4.1.2 The limitations of Isotonic Regression

We find limited suitability for Isotonic Regression due to its topological destruction of the risk surface. While it successfully creates a non-decreasing mapping that minimises squared error, it achieves this by collapsing the continuous prediction space into a discrete step function.

Our granularity analysis revealed that the Isotonic estimator reduced the prediction cardinality from 6,065 unique values (Raw Model) to just 59 unique bins. This extreme quantisation introduces Information Loss and Pricing Instability.

High-resolution discrimination signals from the Gradient Boosting machine are erased within the flat "plateaus" of the Isotonic steps. The transitions between bins create artificial "Pricing Cliffs," with discontinuities reaching as high as $12,119.

In an actuarial context, such volatility is unacceptable. A robust pricing model must be differentiable, ensuring that small changes in risk characteristics result in proportionate, not discrete, changes in premium. PRISM preserves prediction continuity, whereas Isotonic Regression introduces substantial discretisation.

4.2 Structural Forensics

The superior calibration of PRISM stems from topological repair. Theoretically, Tweedie GBMs enforce a rigid covariance assumption on $Var(Y) = \phi\ \mu^p$)across the entire manifold. This implies a uniform relationship between event probability (*P*) and severity (*S*). This assumption fails when risk structures invert (Table 8). "Chronic" risks exhibit positive covariance (high certainty correlates with high cost), whereas "Traumatic" risks exhibit negative covariance (high cost is an outlier for a low-

risk profile). A single-distribution Tweedie estimator cannot distinguish these opposing structures, often averaging them into a biased mean.

Table 8
The covariance between utilisation and intensity

Policyholder Group	Utilization Probability (P)	Event Intensity (S)	Covariance $Cov(P,S)$	Rationale
Chronic / Complex (Worse Health)	High & Predictable	High & Variable	Positive	Systemic Risk: High certainty of a claim (P) correlates directly with high medical complexity (S). The model must surcharge for this compounded risk.
Acute / Traumatic (Healthy)	Low & Rare	Low & Catastrophic	Negative	Shock Risk: The event is severe (S), but it is an unpredicted outlier for a low-risk profile (P). High severity is inversely related to the *ex-ante* probability.

PRISM resolves these limitations by using the Probability Corrector as a bi-directional covariance proxy. The meta-learner explicitly decouples the certainty signal from the magnitude estimate. It then uses the pivot parameter C to dynamically switch between regimes. In the heavy tail$(P > C)$, the model applies a positive correction to capture systemic risk. Negative Regime $(P < C)$ the risk structure inverts (e.g., healthy profiles), and it applies a negative adjustment. Here, the model applies a negative adjustment. This effectively relaxes the fixed power law assumption, ensuring the model adapts to contradictory risk structures rather than averaging them.

4.3 Theoretical Interpretation

From a systems engineering perspective, the PRISM architecture functions as an adaptive error-correction mechanism. The meta-learner enforces a Zero-State Consistency (Zero-Intercept Constraint), ensuring that the probability corrector acts solely as an additive adjustment signal rather than introducing a baseline bias. This architecture aligns theoretically with Additive Bühlmann Credibility Mahler & Dean [24] but applies it through a dimensional inversion. In traditional credibility theory, a unitless weight (Z) scales a financial variable. PRISM inverts this by learning a dollar-value correction gain (β) to scale the unitless probability signal (P).
Consequently, β represents the 'Marginal Price of Certainty', a linear reliability margin applied only when the system detects high-probability loss events. This explicit coefficient allows for the precise isolation of the safety factor applied to the portfolio, rendering the model transparent and verifiable. This satisfies regulatory requirements such as ASB [25], which requires statistically defensible rate differentials by proving that pricing gradients correspond directly to empirical risk signals rather than 'Black Box' artefacts.

4.3 Methodological Limitations

The effectiveness of the PRISM framework depends on the variance stability of the regression learner. In the proposed architecture, the Magnitude Anchor is trained using a Tweedie objective, which imposes a coherent mean–variance relationship and yields a stable risk signal under heavy-tailed loss distributions. This property makes Tweedie gradient boosting particularly well-suited as the anchor component. However, extending PRISM to highly volatile regression estimators, such as

ordinary least squares applied to extreme-tailed outcomes. This could introduce a different failure mode. In this regime, excessive variance in the anchor prediction inflates estimation uncertainty, prompting the ElasticNet meta-learner to aggressively shrink the regression coefficient. As a result, weight may shift disproportionately toward the classifier-based correction, yielding a degenerate decomposition in which the prediction is dominated by the certainty signal rather than the intended magnitude anchor.

PRISM introduces a non-trivial computational trade-off. The architecture maintains three coupled estimators: the Magnitude Anchor, Certainty Signal, and meta-learner. Which collectively increase training time by approximately a factor of three and raise inference latency. While acceptable for offline pricing and portfolio optimisation, this overhead may constrain deployment in sub-millisecond or real-time quoting environments.

The framework relies on an accurate classifier "Certainty Signal." Unlike traditional multiplicative hurdle formulations, $(E[Y] = P(Y > 0) \times E[Y|Y > 0])$. Where probability mis-specification propagates directly into the expected value. PRISM employs an additive topology, $\big(F(x) = f(x) + \beta \times g(x)\big)$, which treats the probability score as an orthogonal residual correction rather than a scaling factor. This structure provides greater resilience: if the probability estimator generalises poorly, the ElasticNet meta-learner suppresses the correction term ($\beta \to 0$), effectively reverting the system to the stable Magnitude Anchor.

A sensitivity analysis substituting the LGBM probability corrector with a Logistic Regression confirmed this resilience: while linear classifiers restored aggregate neutrality (|Bias| < 1%). It incurred an interpretability cost in the form of reduced discriminatory power (Gini decline from 0.612 to ~0.604). This quantifies the trade-off between architectural transparency and the non-linear resolution required to preserve ranking efficiency.

4.4 Future Directions

This study utilised a zero-threshold LGBM Classifier corrector ($P(Y > 0)$) to address convexity bias in the distribution's bulk, the architecture is theoretically extensible to other error profiles. We hypothesise that an Adaptive Thresholding allow the optimal threshold τ depends on the locus of the structural bias. For positive-continuous distributions (e.g., Gamma) where underestimation is concentrated in tail suppression, the corrector should target high-quantile events (e.g., $P(Y > E[Y])$) or $P(Y > q_{75})$). In such tail-correction regimes, the pivot mechanism may require adaptation to an asymmetric correction strategy. Applying additive adjustments only to extreme tail events while neutralising negative corrections for the distribution body to prevent degrading the calibration of the already-unbiased bulk. However, a high-quantile threshold introduces a signal-to-noise trade-off: shifting the threshold to extreme quantiles (e.g., $P(Y > q_{90})$) isolates the correction signal to sparse, high-variance regions dominated by aleatoric noise. Consequently, an excessively high threshold may lead to stochastic overfitting, where the corrector chases random outliers rather than systematic bias. Future work will therefore treat τ as a constrained hyperparameter, optimising the balance between tail-sensitivity and estimator stability. Finally, given the temporal nature of insurance liabilities, future research should extend this framework to Temporal Validation (Out-of-Time testing) to assess the structural stability of the correction parameters β and *C* against claims inflation and shifting risk landscapes.

5. Conclusions

This study synthesises theoretical analysis with empirical validation to explain and resolve the calibration failures of Tweedie GBMs. Theoretically, we demonstrated that the systematic underestimation is a predictable artefact of the objective's rigid mean-variance assumption ($\text{Var}(Y) = \phi\mu^p$). The optimisation places disproportionate weight on high-gradient tail events. Thereby creating a structural convexity bias in the lower-variance bulk.

Empirically, our analysis of the MEPS dataset validated this mechanism. In our setting, the top 5% of claims drive 50% of the total cost. The Tweedie loss function optimisation is dominated by extreme tail gradients, structurally prioritising the high-risk segments. Consequently, the model achieves near-neutrality in the high-gradient tail (Bin 19: -2.14% bias) at the expense of the distribution bulk, which suffers a severe -40.28% collapse (Bin 13). This mid-range limitation, compounded by the log-link's convexity bias, results in a -15.16% aggregate underestimation despite the extreme tail retaining a comparatively moderate -8.34% error.

To address this, we introduced PRISM, an additive ensemble architecture that decouples risk discrimination from magnitude calibration. Empirical validation on the MEPS dataset confirms PRISM's statistical superiority over standard post-hoc methods. PRISM LGBM restored aggregate bias to +0.86% bias, achieved the lowest Root Mean Squared Calibration Error (RMSCE) of $847.62, representing a 27.7% reduction relative to both Isotonic Regression LGBM ($1,173.09) and Scalar Smearing LGBM ($1,173.65). Bootstrap validation (1,000 iterations) further confirms the stability of the PRISM architecture. PRISM XGBM successful recover the -25% aggregate underestimate to +0.85%. PRISM LGBM also has the tightest 95% Confidence Interval [$873, $1,994], outperforming Scalar Smearing LGBM [$1,042, $2,323]. This indicates that PRISM's additive correction is statistically more reliable across varying data subsamples than multiplicative scaling methods.

Beyond the architectural novelty, this study formalises the RMSCE and MACE as robust diagnostics for heavy-tailed regression. Unlike point-wise error metrics, which are sensitive to individual outliers, these bin-wise metrics leverage the Law of Large Numbers to neutralise aleatoric noise within stratified cohorts. This granular aggregation allows future research to distinguish between genuine structural bias and irreducible stochastic volatility, providing a stable standard for benchmarking solvency models in high-variance domains.

Topologically, PRISM eliminates the "pricing cliffs" characteristic of non-parametric methods. Granularity analysis revealed that while Isotonic Regression collapsed the prediction space into only 59 unique values with discontinuities up to $12,119, PRISM generated 6,080 unique values with a negligible average step size of $9.77. This confirms that the architecture promotes the functional continuity required for regulatory compliance.

However, these performance gains involve specific trade-offs. The PRISM architecture increases computational complexity, effectively tripling the training time and inference latency. PRISM introduces a dependency on the classifier's precision; while the ElasticNet meta-learner acts as a fail-safe, severe mis-specification of the probability estimator could theoretically dilute the correction signal. Furthermore, the framework's efficacy relies on the variance stability of the base learner; sensitivity analysis confirms that while simpler linear estimators can maintain calibration, non-linear estimators are necessary to preserve discriminatory efficiency. Future work will extend this topology to tail-specific thresholding for positive-only distributions, offering a generalised solution for heavy-tailed regression calibration.

Acknowledgement

This research was funded by a grant from Asia Pacific University of Technology & Innovation (RDIG Grant RDIG/05/2023).

References

[1] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." *Annals of statistics* (2001): 1189-1232. https://doi.org/10.1214/aos/1013203451

[2] Tweedie, Maurice CK. "An index which distinguishes between some important exponential families." In *Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference*, vol. 579, pp. 579-604. 1984.

[3] Jørgensen, Bent. "Exponential dispersion models." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 49, no. 2 (1987): 127-145. https://doi.org/10.1111/j.2517-6161.1987.tb01685.x

[4] Yang, Yi, Wei Qian, and Hui Zou. "Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models." *Journal of Business & Economic Statistics* 36, no. 3 (2018): 456-470. https://doi.org/10.1080/07350015.2016.1200981

[5] Lindholm, M., & Wüthrich, M. V. (2025). The balance property in insurance pricing. Scandinavian Actuarial Journal, 1-38.

[6] Karlsson, Martin, Yulong Wang, and Nicolas R. Ziebarth. "Getting the right tail right: Modeling tails of health expenditure distributions." Journal of Health Economics 97 (2024): 102912. https://doi.org/10.1016/j.jhealeco.2024.102912

[7] Wüthrich, Mario V., and Michael Merz. *Stochastic claims reserving methods in insurance*. John Wiley & Sons, 2008.

[8] Manning, W. G. (1998). "The Logged Dependent Variable, Heteroscedasticity, and the Retransformation Problem." Journal of Health Economics 17 (3): 283–95. https://doi.org/10.1016/S0167-6296(98)00025-3

[9] Basu, Anirban, and Paul J. Rathouz. "Estimating marginal and incremental effects on health outcomes using flexible link and variance function models." *Biostatistics* 6, no. 1 (2005): 93-109. https://doi.org/10.1093/biostatistics/kxh020

[10] Niculescu-Mizil, Alexandru, and Rich Caruana. "Predicting good probabilities with supervised learning." In *Proceedings of the 22nd international conference on Machine learning*, pp. 625-632. 2005. https://doi.org/10.1145/1102351.1102430

[11] Vincze, I., Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1973). Statistical Inference under Order Restrictions (The Theory and Application of Isotonic Regression). International Statistical Review, 41(3), 395–395. https://doi.org/10.2307/1402630OLS

[12] Dai, Ran, Hyebin Song, Rina Foygel Barber, and Garvesh Raskutti. "The bias of isotonic regression." *Electronic journal of statistics* 14, no. 1 (2020): 801. https://doi.org/10.1214/20-EJS1677

[13] Agency for Healthcare Research and Quality. MEPS HC-243: 2022 Full Year Consolidated Data File Documentation. U.S. Department of Health and Human Services, 2024.

[14] Wolpert, David H. "Stacked generalization." *Neural networks* 5, no. 2 (1992): 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

[15] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* 30 (2017).

[16] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67, no. 2 (2005): 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[17] Dawid, A. Philip. "The well-calibrated Bayesian." *Journal of the American statistical Association* 77, no. 379 (1982): 605-610.

[18] Murphy, Allan H., and Robert L. Winkler. "A general framework for forecast verification." *Monthly weather review* 115, no. 7 (1987): 1330-1338.

[19] Gneiting, Tilmann, and Adrian E. Raftery. "Strictly proper scoring rules, prediction, and estimation." *Journal of the American statistical Association* 102, no. 477 (2007): 359-378. https://doi.org/10.1198/016214506000001437

[20] Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On calibration of modern neural networks." In *International conference on machine learning*, pp. 1321-1330. PMLR, 2017. https://doi.org/10.48550/arXiv.1706.04599

[21] Kuleshov, Volodymyr, Nathan Fenner, and Stefano Ermon. "Accurate uncertainties for deep learning using calibrated regression." In *International conference on machine learning*, pp. 2796-2804. PMLR, 2018.

[22] Chung, Youngseog, Willie Neiswanger, Ian Char, and Jeff Schneider. "Beyond pinball loss: Quantile methods for calibrated uncertainty quantification." *Advances in Neural Information Processing Systems* 34 (2021): 10971-10984. https://doi.org/10.48550/arXiv.2011.09588

[23] Levi, Dan, Ofer Amir, and Tal Schwing. 2022. "Evaluating and Calibrating Uncertainty Prediction in Regression Tasks." *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (7): 7461–69. https://doi.org/10.3390/s22155540

[24] Mahler, H. C., & Dean, C. G. (2001). *Chapter 8: Credibility. In Foundations of Casualty Actuarial Science* (4th ed.). Casualty Actuarial Society.

[25] Actuarial Standards Board (ASB). (2005). Actuarial Standard of Practice No. 12: Risk Classification (for All Practice Areas). Washington, DC: American Academy of Actuaries.