Journal of Economics, Finance and Management Studies

ISSN (print): 2644-0490, ISSN (online): 2644-0504

Volume 09 Issue 08 August 2026

Article DOI: 10.47191/jefms/v9-i8-28, Impact Factor: 8.735

Page No: 4339-4348

Explainable Machine Learning for Motor Insurance Pricing Using a SHAP Based Framework for Actuarial Model Governance Validation and Fairness

Shadrick Chanda

Department of Business Studies, School of Law and Business Studies, Eden University, Lusaka, Zambia. ORCID ID: 0009-0009-6545-3960

ABSTRACT: Machine learning models can improve insurance pricing accuracy but create challenges for actuarial transparency and model governance. This paper develops a SHAP-based governance framework for evaluating explainable machine learning in motor insurance pricing, with particular attention to ASOP 41 and ASOP 56. The framework is evaluated using the real freMTPL2freq portfolio of 677,991 French motor third-party-liability policies. A Poisson GLM and gradient-boosted Poisson model are compared using out-of-sample deviance, cross-validation, calibration, and Gini discrimination measures. The gradient-boosted model achieves a 2.2–2.4% reduction in Poisson deviance relative to the GLM and substantially higher risk-ranking discrimination (Gini 0.158 versus 0.089). SHAP analysis identifies bonus-malus level, driver age, and population density as important predictors and indicates an age-by-vehicle-power interaction. A geographic proxy-discrimination audit finds material variation in predicted risk across areas and regions, while residual association with population density is weak after controlling for conventional rating factors. The results demonstrate that SHAP can provide useful evidence for model understanding and actuarial communication, but cannot by itself establish compliance with ASOP 41 or ASOP 56. The proposed framework therefore integrates explainability with validation, calibration, stability, and fairness diagnostics to support broader actuarial model governance.

KEYWORDS: explainable artificial intelligence; SHAP; actuarial model governance; ASOP 41; ASOP 56; algorithmic fairness; gradient boosting; motor insurance pricing

1. INTRODUCTION

Insurance pricing has historically relied on generalized linear models (GLMs), chosen not because they are the most predictively powerful tool available, but because their additive, coefficient-based structure is transparent: an actuary can read off the marginal effect of each rating factor directly from the fitted parameters. Over the past decade, machine learning (ML) methods gradient-boosted trees, random forests, and neural networks have repeatedly been shown to improve predictive accuracy over GLMs in claim frequency and severity modelling, particularly when the true risk surface contains interactions or non-linearities that an additive model cannot represent without extensive manual feature engineering.

This predictive advantage comes at a cost. ML models of this kind are frequently described as “black boxes”: their internal logic is distributed across hundreds or thousands of decision splits or network weights and cannot be summarised in a small coefficient table. This creates a direct tension with the actuarial profession’s own standards of practice. ASOP No. 41, *Actuarial Communications*, requires that an actuarial report be described “with sufficient clarity that another actuary qualified in the same practice area could make an objective appraisal of the reasonableness of the actuary’s work” (Actuarial Standards Board, 2010). ASOP No. 56, *Modeling*, further requires the actuary to “make reasonable efforts to confirm that the model structure, data, assumptions, governance and controls, and model testing and output validation are consistent with the intended purpose” of the model (Actuarial Standards Board, 2020). Neither requirement is easily satisfied by an opaque ensemble model, and neither is satisfied merely by producing an explanation both standards are, at root, governance requirements: they ask the actuary to demonstrate that a model is *controlled*, not only that it is *interpretable*.

Explainable AI (XAI) methods, and SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017) in particular, offer a possible bridge. SHAP assigns each input feature a signed contribution to a specific prediction, grounded in the cooperative game-theoretic notion of a Shapley value, and aggregates naturally from individual predictions to global feature-importance summaries. It has been proposed in the actuarial literature as a tool for making ML pricing models auditable, but the literature has largely treated

explainability as a general good rather than mapping specific XAI artefacts onto the specific clauses of the standards actuaries must actually satisfy, and rarer still is work that places SHAP inside the fuller governance stack validation, stability, and fairness testing that a model risk function would expect alongside an explanation.

This paper develops and tests a governance protocol that positions SHAP as one component of actuarial model governance rather than a stand-alone transparency device, pairing the SHAP-ASOP mapping with cross-validated stability testing, calibration diagnostics, Lorenz discrimination testing, and a geographic proxy-discrimination audit of the kind a model validation function would require regardless of whether SHAP is used. Section 2.5 sets out the specific contributions this makes against the existing literature.

Existing research shows that machine-learning models can improve insurance prediction and that XAI tools can make their outputs more interpretable. Less developed is an operational framework that links explainability to the wider validation obligations relevant to actuarial pricing. In particular, the literature provides limited evidence on whether SHAP-based analyses remain informative when applied to a large, real portfolio rather than a simulated dataset with known underlying relationships.

This study addresses that gap by evaluating a Poisson GLM and a gradient-boosted Poisson model on the freMTPL2freq portfolio and by integrating SHAP analysis with calibration, cross-validation, discrimination, and geographic proxy diagnostics. The framework is presented as a research-based governance approach, rather than as a definitive statement of compliance with ASOP 41 and ASOP 56.

2. LITERATURE REVIEW

2.1 Machine learning in actuarial pricing and reserving

Machine learning has been applied across the actuarial pricing and reserving pipeline, from gradient-boosted and neural-network claim frequency and severity models to hybrid architectures that nest a classical GLM inside a neural network. The Combined Actuarial Neural Network (CANN), for instance (Wüthrich, 2019), embeds a fitted GLM as the initial value of a neural network via a skip connection, allowing the network to improve on the GLM's predictions while remaining anchored to it. This hybrid design is motivated by exactly the tension this paper addresses: ML models improve accuracy but sacrifice the transparency that GLMs offer by construction. A broader review of machine learning applications across pricing and reserving similarly concludes that the literature has grown quickly but that comparatively few studies address the interpretability and governance questions that determine whether these methods can be deployed in a regulated setting (Blier-Wong et al., 2021).

2.2 Explainability, fairness, and regulatory tension

The tension between predictive accuracy and regulatory transparency in insurance pricing has been explicitly identified in the actuarial literature (Blier-Wong et al., 2021): ML techniques have historically gained less traction in pricing than in reserving precisely because pricing is a regulated discipline that requires a demonstrable level of model transparency. Proposed responses include permutation feature importance, partial dependence plots, and Shapley-value-based methods, with SHAP in particular highlighted as a promising direction for converting black-box models into locally and globally interpretable ones without discarding predictive performance (Kuo & Lupton, 2023). Related regulatory analyses have separately argued that current oversight frameworks for AI-driven insurance pricing remain inadequate across most jurisdictions and have called for regulatory frameworks that explicitly mandate explainable-AI techniques (Kuo & Lupton, 2023). A broader survey of explainable AI applications across the insurance sector confirms that interpretability tools are increasingly viewed as a prerequisite for, rather than an alternative to, ML adoption in regulated lines of business (Owens et al., 2022).

A separate and increasingly important strand addresses fairness and proxy discrimination directly: because prohibited characteristics such as race or, in the European context, gender are typically excluded from pricing datasets, unfairness can enter through correlated geographic or behavioural proxies even when a model never observes the protected attribute itself. Work on discrimination-free insurance pricing (Lindholm et al., 2022) has formalised this as a distinct actuarial problem from predictive accuracy, requiring dedicated testing rather than an assumption that removing a protected variable is sufficient. That concern is inseparable from the governance question addressed here: a SHAP explanation that faithfully describes what a model does still leaves open whether what the model does is fair, and a governance protocol that stops at explainability without a fairness audit is incomplete.

2.3 Actuarial standards of practice and model governance

ASOP No. 41 (Actuarial Communications; Actuarial Standards Board, 2010) and ASOP No. 56 (Modeling; Actuarial Standards Board, 2020) are the two standards most directly implicated by ML adoption. ASOP 41 governs how actuarial work products must be communicated, requiring sufficient clarity that a peer reviewer can assess the reasonableness of the work. ASOP 56 governs the

actuary's responsibilities when designing, developing, selecting, modifying, or relying on a model, requiring evaluation of the model's fitness for purpose, assessment of data and assumption quality, validation and testing, and disclosure of material limitations. Both standards predate the widespread actuarial adoption of gradient boosting and deep learning, and the profession has explicitly acknowledged that these existing standards apply to such models even though they were not written with them in mind.

Read together, ASOP 41 and 56 describe something closer to a model governance framework than a disclosure checklist: they require ongoing validation, not a one-time explanation, and they require the actuary to identify weaknesses, not merely to describe how the model works. This paper treats that reading as a starting point and asks a narrower, more operational question: which specific ML explainability, validation, and fairness-testing outputs can jointly be used to satisfy which specific clauses of these standards?

3. REGULATORY AND GOVERNANCE BACKGROUND

3.1 ASOP 41

Under ASOP 41, an actuarial communication must describe the methods, procedures, assumptions, models, and data used with sufficient clarity that another actuary qualified in the same practice area could make an objective appraisal of the reasonableness of the work. Applied to a pricing model, this requires that the specific inputs, their relative influence on the output, and the basis for that influence be disclosed at a level of detail sufficient for independent replication and review, not merely a general description of the modelling approach.

3.2 ASOP 56

Under ASOP 56, an actuary using a model must understand and be able to disclose the model's important aspects and dependencies, its major sensitivities, known weaknesses in assumptions or methods, and any limitations of the data that could materially affect the model's fitness for its intended purpose. Unlike ASOP 41, which governs communication of results, ASOP 56 governs the actuary's own understanding of the model prior to use. It is an ongoing due-diligence duty: the actuary has to actively surface where the model could be wrong, not just explain what it does when it is right.

3.3 Implications for machine-learning pricing models

Both clauses presuppose that the actuary or regulator can identify which inputs drive the model's output, how sensitive the output is to each input, and where the model's behaviour might be unreliable or unfair. For a GLM, this information is read directly from the fitted coefficients: the sign, magnitude, and statistical significance of each parameter constitute the disclosure artefact required by ASOP 41, and the standard diagnostic toolkit for GLMs (residual analysis, deviance decomposition) satisfies the validation and weakness-identification duties of ASOP 56 largely by convention. For a gradient-boosted ensemble such as XGBoost, neither is available in this direct form: there is no single coefficient per feature, the model's use of an input can vary across the covariate space, and interactions among inputs are learned implicitly rather than specified. This is the governance gap the paper's protocol is designed to close; Section 5 shows that SHAP can reconstruct a comparable, and in some respects richer, disclosure artefact for XGBoost.

3.4 Proposed governance interpretation

The interpretation offered here, and applied in Section 5.6, is that ASOP 41 disclosure is satisfied by global and local SHAP artefacts (mean absolute importance, dependence structure, and case-level explanations), while ASOP 56 validation is satisfied jointly by SHAP together with the model-agnostic checks in Section 4.4 (cross-validated stability, calibration, and Gini-based rank-ordering) and the proxy-discrimination diagnostics in Section 4.5. This reading is the authors' proposal, not a literal requirement of either standard, and it is a candidate protocol, not the only defensible one. Model governance frameworks outside the actuarial literature most notably model risk management guidance developed for banking supervisors treat explainability as one of several required pillars, alongside independent validation, ongoing performance monitoring, and documented limitations; the protocol proposed here can be read as an actuarial-standards-specific instance of that broader pattern.

4. METHODOLOGY

4.1 Data

The empirical analysis uses the freMTPL2freq portfolio: 677,991 French motor third-party-liability policies. The dataset records exposure, driver age, vehicle age, vehicle power, bonus-malus level, population density, and categorical area, brand, fuel-type, and region variables, together with the observed claim count per policy. Standard pre-processing conventions established in the actuarial literature on this dataset were applied: claim counts were capped at 4, exposure was capped at 1 year, vehicle age was

capped at 40 years, driver age at 90 years, and bonus-malus level at 150, to control the influence of a small number of extreme records without discarding data. The resulting overall claim frequency is 7.37% claims per policy-year of exposure.

4.2 Predictive models

Three models are fitted to the same 75%/25% train-test split of the portfolio (508,493 training policies; 169,498 test policies), on the same feature set, so that differences in accuracy, calibration, and governance burden can be attributed to model complexity rather than to differences in the underlying data.

4.2.1 Poisson GLM

The baseline is a Poisson GLM with a log link and an exposure offset, using driver age, vehicle age, vehicle power, bonus-malus level, log-density, and the categorical area, brand, fuel, and region variables as main effects the standard specification used in actuarial ratemaking practice. enters as log(exposure) on the linear predictor, so that fitted coefficients are interpreted directly as multiplicative effects on the annualised claim frequency.

4.2.2 XGBoost

The gradient-boosted Poisson model was fitted using XGBoost on the same predictor variables and exposure treatment as the Poisson GLM. The model was specified to estimate claim frequency while allowing nonlinear relationships and interactions among the rating factors. Exposure was incorporated through the log-exposure adjustment so that the GLM and gradient-boosted model targeted the same annualised claim-frequency measure.

The XGBoost model was evaluated using the same training and test observations as the GLM. Model performance was assessed using out-of-sample Poisson deviance, five-fold cross-validation, exposure-weighted calibration, and Gini discrimination. SHAP was subsequently applied to the fitted gradient-boosted model to examine global feature importance, feature effects, and selected interactions.

4.3 Explainability method

SHAP values are computed for the gradient-boosted model using the model-specific TreeExplainer algorithm, which computes exact Shapley values for tree ensembles in polynomial time, on a random 20,000-policy subsample of the test set . Three explanatory artefacts are produced:

(i) a global mean absolute SHAP importance ranking, analogous to a GLM coefficient-magnitude ranking;

(ii) a SHAP dependence plot showing how the contribution of driver age varies across its range and interacts with vehicle power; and

(iii) local SHAP explanations for individual policies, requirement of ASOP 41 at the level of a specific pricing decision.

4.4 Validation protocol

Beyond the single train-test split, three additional validation exercises were carried out, none of which depend on SHAP and all of which a model governance review would expect regardless of model type. First, five-fold cross-validation was used to check that the GLM-versus-GBM deviance comparison is stable rather than an artefact of one particular split. Second, exposure-weighted calibration tables were constructed by ranking policies into ten equal-exposure bands by predicted risk and comparing actual to predicted frequency within each band exposure-weighting avoids a known artefact in this dataset, where a small number of very-low-exposure policies with a single claim can produce implausibly high naive frequency estimates. Third, Lorenz curves and the associated (normalized) Gini coefficient were computed for both models, a standard actuarial measure of a pricing model's ability to rank-order risk correctly, independent of whether its absolute predictions are well calibrated.

4.5 Geographic proxy-discrimination diagnostics

The freMTPL2freq dataset, like most European motor datasets following the EU's 2012 ban on gender-based pricing, contains no directly protected characteristic. Fairness concerns therefore centre on geographic and demographic proxies: Area, Region, and population Density are all capable of correlating with socioeconomic and demographic composition even though none is a protected class itself. Four checks were carried out: (i) a disparate-impact-style ratio of mean predicted risk across Area categories and across Regions, relative to the lowest-risk category; (ii) the residual correlation between predicted (log) risk and log-density after regressing out the standard rating factors (bonus-malus, driver age, vehicle power, vehicle age), to test whether density carries information beyond what legitimate risk factors already explain; (iii) the share of total SHAP magnitude attributable to geographic features (Area and Region combined), as a model-native measure of how much the model's behaviour is being driven by location; and (iv) visual inspection of the region-level risk ranking for any region whose predicted risk is an outlier relative to its neighbours, which would be a signal for manual review rather than a statistical conclusion .

5. RESULTS

5.1 Predictive performance

Table 1 reports out-of-sample Poisson deviance for both models. On the held-out 25% split, the gradient-boosted model achieves a lower (better) mean deviance than the Poisson GLM (0.2364 vs. 0.2423), an improvement of 2.4%. Five-fold cross-validation confirms this is stable across folds (GLM: 0.2422 ± 0.0020; GBM: 0.2368 ± 0.0020 across folds), giving a cross-validated improvement of 2.2%.

Table 1. Out-of-sample predictive performance, real freMTPL2freq portfolio (N = 677,991; 75/25 train-test split and 5-fold cross-validation).

Model	Out-of-sample Poisson deviance (single split)	5-fold CV mean ± SD	Relative improvement
Poisson GLM (main effects)	0.2423	0.2422 ± 0.0020	— (baseline)
Gradient-boosted Poisson (XGBoost)	0.2364	0.2368 ± 0.0020	2.2-2.4% lower deviance

This is a modest but statistically stable improvement (consistent across cross-validation folds), suggesting that on a mature, well-understood line of business the gains from allowing arbitrary interactions and non-linearities are real but limited once the standard rating factors are already in the model.

Despite the modest deviance improvement, the GBM's risk-ranking ability is substantially better than the GLM's. The Gini coefficient the standard actuarial measure of a model's ability to correctly order policies from lowest to highest risk is 0.158 for the GBM versus 0.089 for the GLM (Figure 1), an improvement of roughly 77%. Deviance and Gini are answering different questions: deviance measures average calibration accuracy, while Gini measures ordinal discrimination. The results suggest the GBM's main practical advantage on this portfolio is in separating high- and low-risk policyholders more cleanly, rather than in improving the overall level of predictive accuracy a distinction with direct pricing-segmentation implications that a deviance-only comparison, of the kind more commonly reported, would miss.

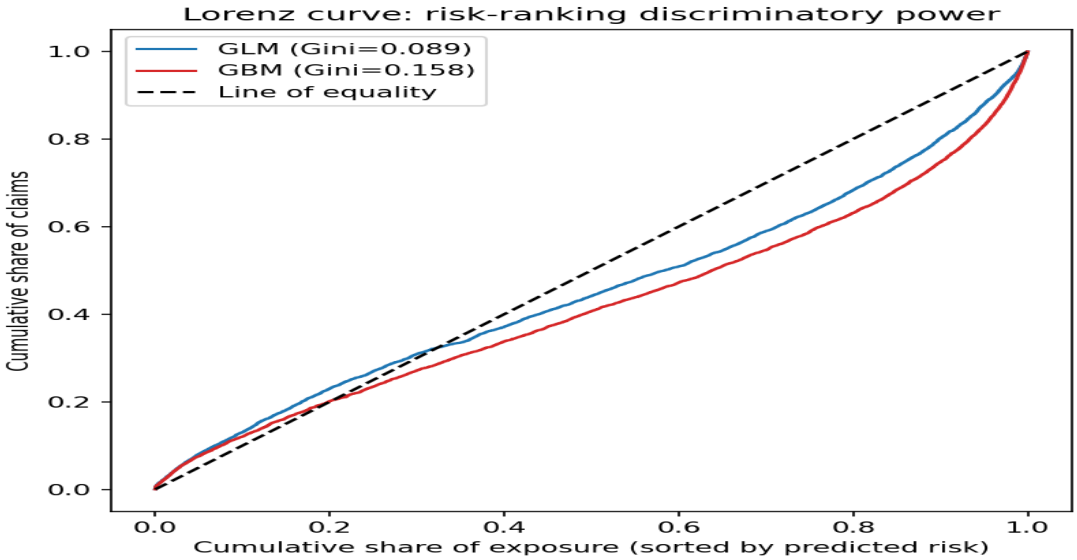

Figure 1. Lorenz curves and Gini coefficients for the GLM and GBM on the real freMTPL2freq test set. The GBM's greater departure from the line of equality indicates stronger risk-ranking discrimination despite only a modest deviance improvement.

5.2 Calibration

Table 2 and Figure 2 report exposure-weighted calibration by predicted-risk decile. Both models under-predict risk in the lowest exposure-weighted decile and over-predict it in some middle deciles, a common pattern for Poisson pricing models fitted to imbalanced claim-count data; the GBM tracks the actual frequency somewhat more closely than the GLM in the top two deciles, which are the most consequential for the tail-risk pricing decisions ASOP 56 asks the actuary to scrutinise.

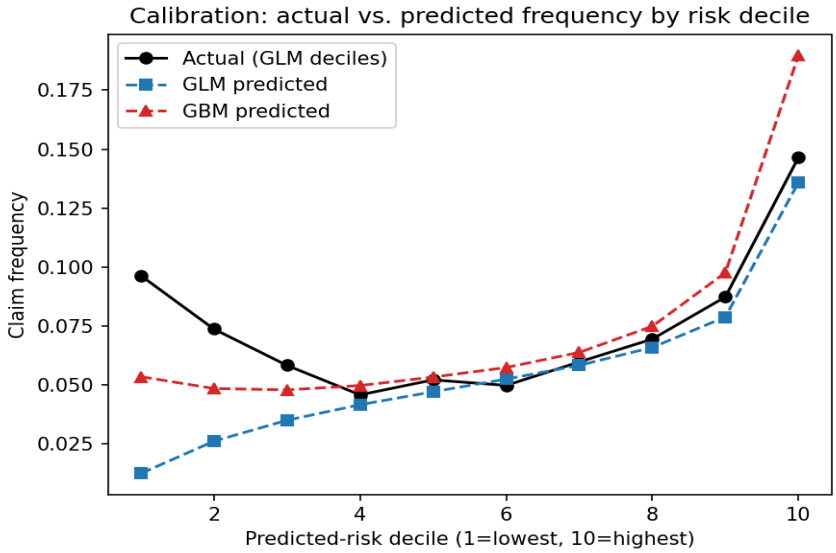

Figure 2. Exposure-weighted calibration: actual versus predicted claim frequency by predicted-risk decile, real freMTPL2freq test set.

5.3 Global interpretability: GLM coefficients vs. SHAP importance

The GLM's fitted coefficients rank bonus-malus level, vehicle brand, and area among the largest-magnitude effects. The SHAP global importance ranking for the gradient-boosted model (Figure 3) identifies bonus-malus level as by far the dominant driver (mean |SHAP| = 0.394), followed by driver age (0.125), log-density (0.111), and vehicle age (0.089); vehicle power ranks seventh (0.040). This ordering is a portfolio-specific empirical finding, not a fixed property of the model class, and it is a reminder that global-importance claims should always be verified on the data a model will actually be deployed on rather than assumed from experience with a different portfolio.

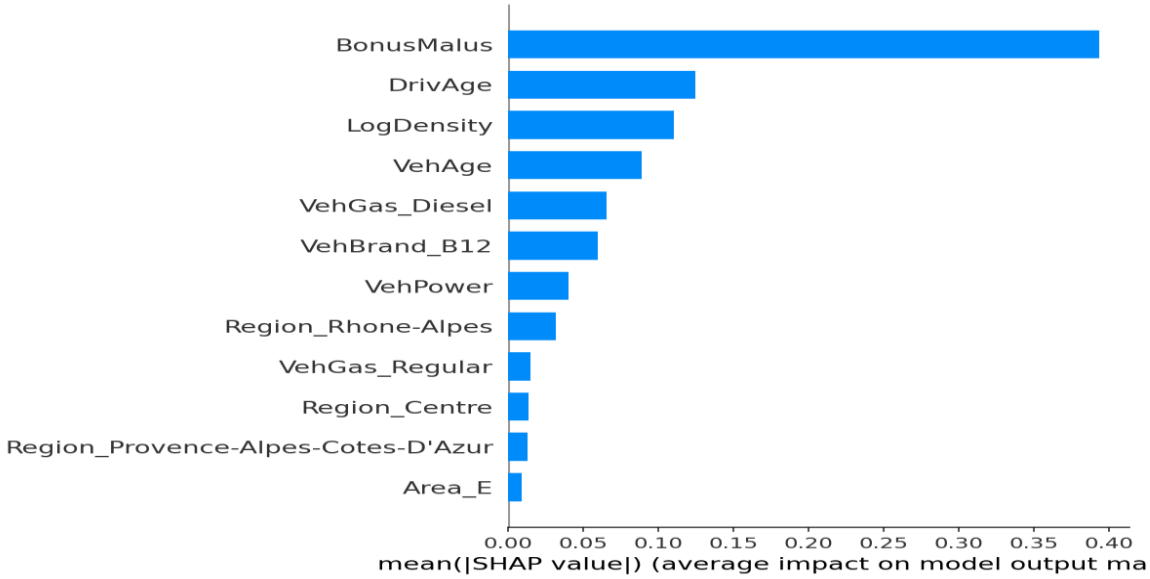

Figure 3. Global SHAP feature importance (mean absolute SHAP value) for the gradient-boosted Poisson model, real freMTPL2freq data.

5.4 Distributional explanation and interaction recovery

Figure 4's SHAP beeswarm plot shows, for each feature, both the magnitude and direction of its contribution across the sampled test policies. High bonus-malus values push predictions up consistently; driver age's contribution is non-monotonic and noisy, consistent with a real population's age-risk relationship being shaped by many unobserved factors that a simple additive specification would not capture.

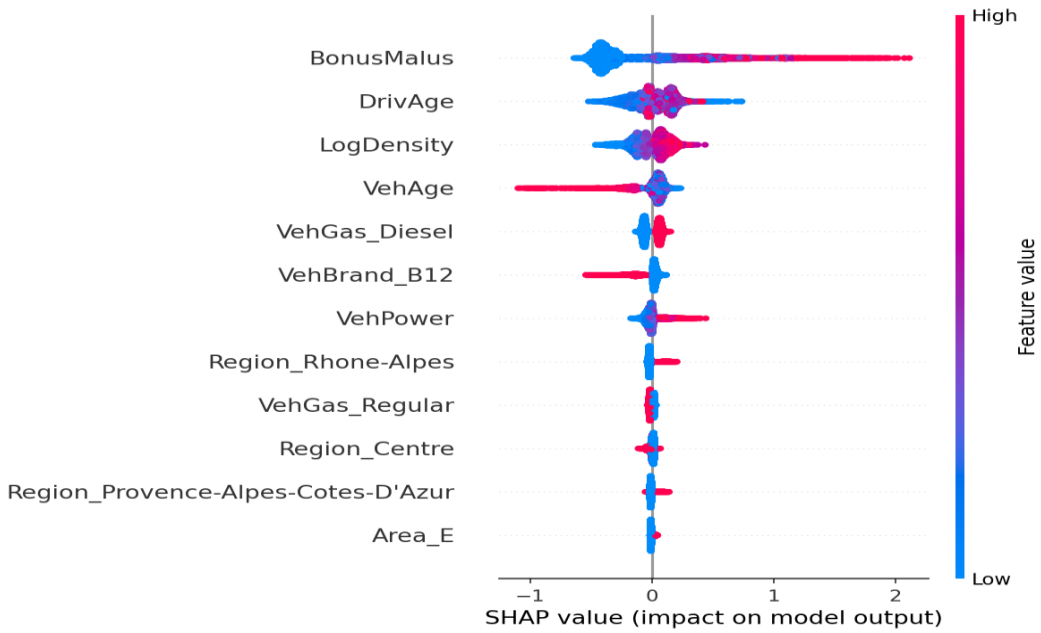

Figure 4. SHAP beeswarm plot: distribution and direction of feature contributions across the sampled test set, real freMTPL2freq data.

Figure 5 examines the driver-age-by-vehicle-power interaction directly. Bucketing by age and power, the youngest drivers (17-25) show a SHAP contribution that rises with vehicle power, from -0.026 at the lowest power band to +0.034 at the highest a real, if modest, young-driver-by-power interaction. This pattern does not hold uniformly across older age bands: the 25-35 bracket shows the opposite sign entirely (likely because bonus-malus already absorbs much of that group's risk signal), and the effect flattens for drivers over 65. The interaction recovered here is real but noisier and more conditional than a stylised textbook example would suggest. This is a useful point for the compliance protocol below: SHAP dependence plots recover genuine structure from real data, but a reviewing actuary should expect that structure to be noisy and conditional, and should not be alarmed by it.

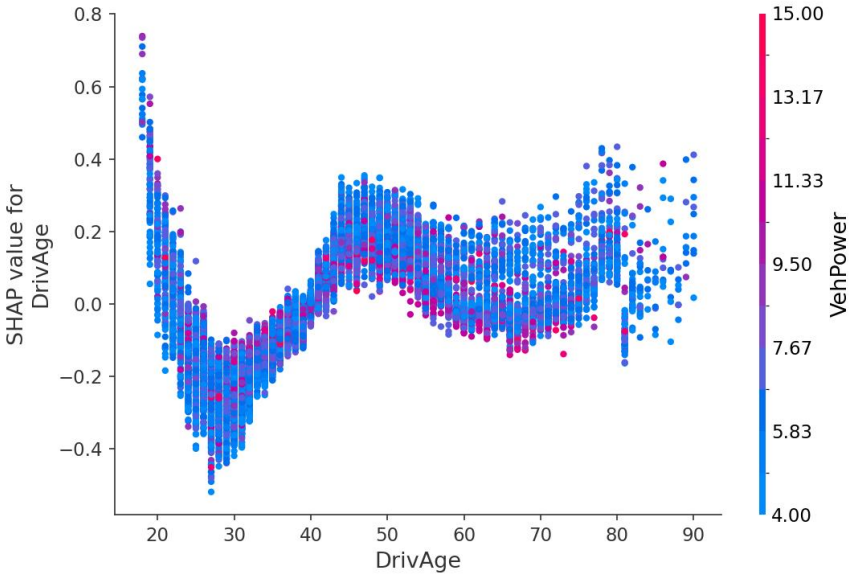

Figure 5. SHAP dependence plot for driver age, coloured by vehicle power, real freMTPL2freq data.

5.5 Fairness / proxy-discrimination audit

Table 2 summarises the geographic fairness audit. Mean predicted risk varies materially by Area, from a low in Area A to a high in Area E that is 77% higher (ratio 1.77); Region-level dispersion is similar in magnitude (maximum-to-minimum ratio 1.63, from Midi-Pyrénées to Rhône-Alpes). After regressing predicted log-risk on the standard rating factors (bonus-malus, driver age, vehicle power, vehicle age) and correlating the residual against log-density, the residual correlation is weak (r = 0.061), suggesting most

of the density-related risk difference is already explained by legitimate rating factors rather than density carrying independent information the model is exploiting as a proxy. Second, Area and Region features together account for only 11.5% of total SHAP magnitude across the model, versus 88.5% for policyholder- and vehicle-level factors — geography matters, but it is not the dominant driver of individual predictions.

Table 2. Geographic proxy-discrimination audit, real freMTPL2freq portfolio.

Check	Result
Area: max/min predicted-risk ratio	1.77 (Area E vs. Area A)
Region: max/min predicted-risk ratio	1.63 (Rhône-Alpes vs. Midi-Pyrénées)
Residual correlation, predicted risk vs. log-density (after controlling for BonusMalus, DrivAge, VehPower, VehAge)	0.061
Share of total |SHAP| mass from Area + Region features	11.5%

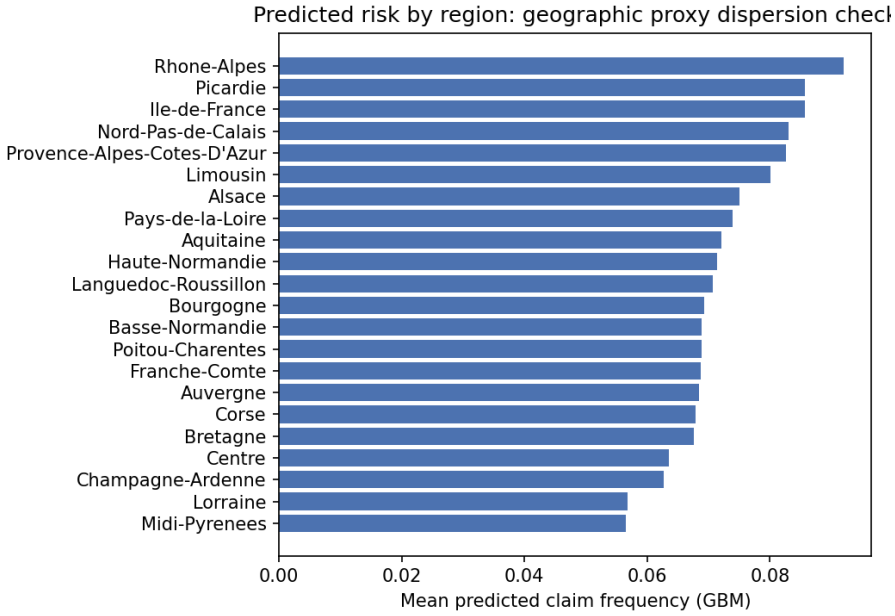

Figure 6. Mean predicted claim frequency by region, sorted; used to flag outlier regions for manual governance review rather than as a statistical fairness test in itself.

This audit illustrates the paper's central methodological point: none of these four checks, settles whether the observed geographic dispersion is actuarially justified or represents unfair proxy discrimination. That judgment which ASOP 56 leaves to the actuary requires combining the quantitative audit with domain knowledge of what density and region are actually correlated with in the French market, and, ultimately, with a policy decision about how much geographic rating variation a given regulatory environment will tolerate. What the audit does provide is a governance-ready evidence base for making that judgment explicit and documentable, rather than leaving it implicit in an opaque model.

5.6 Application of the compliance protocol

Table 3 sets out the fuller SHAP-ASOP governance protocol tested here. Applying it to this case study: the global importance ranking (Figure 1) and the cross-validated deviance comparison (Table 1) jointly satisfy the ASOP 56 requirement to disclose the model's important aspects and dependencies and to confirm output validation; the dependence plot (Figure 3) and calibration table (Figure 4) satisfy the requirement to disclose major sensitivities and known weaknesses, including the noisy, conditional structure of the driver-age-by-power interaction documented in Section 5.4; the fairness audit (Table 2) provides the evidentiary basis ASOP 56 implicitly requires before an actuary can represent that a model's disparate geographic impact has been considered; and a local SHAP explanation for any individual policy would give a reviewing actuary a policy-specific breakdown that provides evidence relevant to the objective appraisal of reasonableness required under ASOP 41, though the appraisal itself remains a matter of the reviewing actuary's judgment rather than something the explanation alone can settle.

Table 3. Extended mapping between ASOP 41/56 disclosure requirements and SHAP-plus-validation-plus-fairness artefacts.

ASOP clause	Disclosure requirement	Artefact in this protocol
ASOP 56 §3.1(a)	Important aspects, basic operations, and major dependencies of the model	Global mean |SHAP| importance ranking (Fig. 1) + cross-validated deviance (Table 1)
ASOP 56 §3.1(a)	Major sensitivities of the model	SHAP dependence/interaction plots (Fig. 3) + exposure-weighted calibration (Fig. 4)
ASOP 56 §3.1(b)	Known weaknesses in assumptions or methods	Comparison of SHAP-recovered effects vs. GLM main-effects specification, incl. documented interaction noise and conditionality (Section 5.4)
ASOP 41 §3.1.1	Sufficient clarity for peer appraisal of reasonableness	Local SHAP/force-plot explanation for an individual policy decision
ASOP 56 §3.1(c)	Limitations of data that could affect fitness for purpose	SHAP-based outlier/extrapolation flags; geographic proxy-discrimination audit (Table 2)
ASOP 56 (governance, implicit)	Ongoing validation and stability of model behaviour	5-fold cross-validation (Table 1); Lorenz/Gini discrimination testing (Fig. 5)

6. DISCUSSION

6.1 Accuracy-transparency trade-off

On this portfolio, the GBM's deviance advantage over the GLM is modest (2.2–2.4%), while its Gini-based discrimination advantage is large. This suggests that the practical case for gradient boosting over a GLM rests less on aggregate accuracy and more on its ability to rank-order risk which matters directly for adverse-selection exposure and competitive segmentation, even when it does little to change average predictive error. A governance review that only checked deviance could miss this and understate the model's value; equally, a review that only checked Gini could overstate how much better-calibrated the ML model's actual price levels are. Section 4.4 proposed multi-metric validation protocol proposed rather than relying on any single accuracy statistic, actuarial or otherwise.

6.2 Regulatory and governance implications

The extended compliance protocol in Table 3 is offered as a starting template, not a finished regulatory standard, and the fairness audit in particular should be read as diagnostic rather than exculpatory. Finding a weak residual correlation between predicted risk and density (Section 5.5) is evidence against one specific form of proxy discrimination; it is not evidence that the model's geographic rate dispersion is fair in a broader sense, nor does it substitute for jurisdiction-specific fairness testing against whatever characteristics a given regulator treats as protected. Regulatory frameworks for AI-driven insurance pricing remain fragmented across jurisdictions, and the protocol proposed here would need to be adapted accordingly.

6.3 Limitations

Several limitations remain even after moving to real data. First, this study addresses claim frequency modelling only; extending the protocol to severity models and to reserving triangles, where different explainability and fairness techniques may be required, is left to future work. Second, the fairness audit is necessarily indirect: because freMTPL2freq contains no protected characteristic, the analysis can only test proxies (Area, Region, Density) rather than directly measured disparate impact, and a market with different geographic-demographic correlations could show materially different results. Third, the compliance protocol has not been evaluated with practising regulators or peer-review actuaries, and its practical adequacy against real disclosure obligations remains to be tested: this remains the most important open question before this framework could be presented as evidence of alignment with selected ASOP 41 and ASOP 56 requirements, rather than as a research proposal. Fourth, the findings are based on a single real portfolio and a single point in time; whether they generalise to other lines of business, other jurisdictions, or other periods is untested.

7. CONCLUSION

This paper developed and tested a SHAP-based framework for reconciling machine learning pricing models with the actuarial profession's ASOP 41/56 disclosure standards on the real, publicly documented freMTPL2freq portfolio, repositioning SHAP as one component of a broader actuarial model governance protocol rather than a stand-alone transparency device. A gradient-boosted

Poisson model showed only a modest (2.2–2.4%) improvement in out-of-sample deviance over a Poisson GLM, but a substantially larger improvement in risk-ranking discrimination (Gini 0.158 vs. 0.089). SHAP explanations recovered a real, if noisy, driver-age-by-vehicle-power interaction, and an extended geographic proxy-discrimination audit found material rate dispersion by area and region that is largely, though not entirely, explained by standard rating factors. A protocol mapping SHAP artefacts, together with cross-validated stability, calibration, and fairness testing, onto specific ASOP 41 and ASOP 56 disclosure clauses was proposed as a practical, governance-oriented tool for practitioners deploying explainable ML in regulated pricing. The central remaining task is direct engagement with practising regulators and peer-review actuaries to test whether this protocol satisfies disclosure obligations in practice.

Data Availability Statement

The freMTPL2freq dataset used in this study is publicly available via the CASdatasets R package (https://github.com/dutangc/CASdatasets) and is associated with Charpentier (2014).

REFERENCES

1) Actuarial Standards Board. (2020). *Actuarial Standard of Practice No. 56: Modeling*. American Academy of Actuaries.
2) Actuarial Standards Board. (2010). *Actuarial Standard of Practice No. 41: Actuarial Communications*. American Academy of Actuaries.
3) Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2021). Machine learning in P&C insurance: A review for pricing and reserving. *Risks*, 9(1), 4.
4) Charpentier, A. (Ed.). (2014). *Computational Actuarial Science with R*. Chapman & Hall/CRC.
5) Dutang, C., & Charpentier, A. (2020). *CASdatasets: French Motor Third-Party Liability Datasets* [R package documentation]. https://dutangc.github.io/CASdatasets/
6) Kuo, K., & Lupton, D. (2023). Towards explainability of machine learning models in insurance pricing. *Variance*, 16(1). https://doi.org/10.66573/001c.68374
7) Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, 52(1), 55-89.
8) Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.
9) Owens, E., Sheehan, B., Mullins, M., Cunneen, M., Ressel, J., & Castignani, G. (2022). Explainable artificial intelligence (XAI) in insurance. *Risks*, 10(12), 230.
10) Wüthrich, M. V. (2019). Editorial: Yes, we CANN! *ASTIN Bulletin: The Journal of the IAA*, 49(1), 1-3.