JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

Interpretable AI for Telematics-Based Insurance Pricing: Nonlinear Risk Modeling, Feature Attribution, and Regulatory Transparency

Liam O'Sullivan[1, *]; Elena Marković[2]; Hendrik van der Berg[3]
[1] School of Business, Maynooth University, Maynooth, Co. Kildare, Ireland
[2] Faculty of Economics and Business, University of Rijeka, Rijeka, Croatia
[3] Faculty of Economic and Management Sciences, North-West University, Potchefstroom, South Africa
* Corresponding author: liam.osullivan@mu.ie

ARTICLE INFO
Received
October 18, 2022
Revised
December 21, 2022
Accepted
February 12, 2023
Available Online
March 30, 2023
DOI
10.63646/jaiaa.2023.010104

Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
Auto insurance ratemaking is undergoing a structural shift as telematics devices and usage-based insurance (UBI) programmes deliver granular data on individual driving behaviour. Generalised linear models (GLMs) remain the dominant pricing engine in the industry because of their statistical transparency and regulatory acceptance, yet they struggle to express the nonlinear and conditional structure that characterises telematics signals. This paper develops an interpretable analytics framework that combines generalised additive models (GAMs), gradient-boosted machines, and a low-dimensional clustering procedure for territorial risk design. Using a synthetic UBI portfolio that preserves the joint distribution of policy and behavioural variables, we model claim frequency and claim severity separately, evaluate spline-based interaction effects, and quantify feature contributions through partial dependence plots and Shapley values. The GAM-based frequency model lowers the empirical–predicted gap for the youngest age cohort by roughly seven percentage points and uncovers a U-shaped response for annual miles driven that is invisible in the GLM specification. An XGBoost benchmark ranks credit score, years without a claim, and car age as the dominant predictors, while annual mileage contributes mostly through interactions with car use and region. A penalised dispersion criterion stabilises the choice of cluster count for the territorial reduction step, and sensitivity tests confirm that the recommended cluster count varies within a narrow band as the penalty parameter changes. Taken together, the framework offers regulators and insurers a transparent path from raw telematics data to defensible rate relativities and parsimonious territorial classifications.

Keywords: usage-based insurance; generalised additive models; SHAP; interpretable clustering; telematics; ratemaking; xgboost; regulatory transparency.

I. INTRODUCTION

Pricing automobile insurance has historically depended on a tightly governed set of policy characteristics—driver age, marital status, vehicle age, primary use, credit score, and territory—combined inside generalised linear models (GLMs) to produce risk relativities that regulators can audit (Werner & Modlin, 2016; Frees, 2015). Three forces are now reshaping that practice. First, the diffusion of smartphone and in-vehicle telematics has produced rich behavioural records that complement traditional policy data with high-frequency measurements of speed, mileage, braking, cornering, and

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

time-of-day exposure (Husnjak et al., 2015; Eling & Kraft, 2020; Gilbert et al., 2019). Second, advances in flexible regression and machine learning have made it computationally inexpensive to fit nonlinear and interaction-rich models on millions of policy records (Henckaerts et al., 2021; Yang et al., 2018; Spedicato et al., 2018). Third, regulators and consumer-protection bodies have begun to scrutinise algorithmic pricing for transparency, stability, and fairness (Lindholm et al., 2022; Frees & Huang, 2023; Xin & Huang, 2024).

These forces create a methodological tension. Telematics signals deliver predictive lift, but their contribution to a tariff often runs through nonlinear thresholds, exposure–behaviour interactions, and spatially correlated risk patterns that linear specifications smooth away (Verbelen et al., 2018; Ayuso et al., 2019). Machine-learning ensembles capture this structure, yet their decisions are difficult to communicate to filing reviewers who need to verify that no protected attribute is being used and that recommended premiums do not change abruptly across adjacent rating cells (Rudin, 2019; Molnar, 2020; Adadi & Berrada, 2018). The frontier of usage-based insurance (UBI) pricing therefore lies not in maximising predictive accuracy alone, but in building models that combine flexibility with explanation infrastructure (Lu, 2019; Zhang & Lu, 2021).

This paper develops and evaluates such a framework using a synthetic UBI portfolio that preserves the joint statistical structure of a Canadian auto insurance book (So et al., 2021). We retain the classical separation between claim frequency and claim severity, which remains the actuarial standard because the two components respond to different risk drivers (Wuthrich & Buser, 2021; Quan & Valdez, 2018). We then layer three contributions onto that backbone. First, we fit generalised additive models (GAMs) with cubic spline bases on the continuous covariates and conditional spline interactions between annual mileage, marital status, car use, and region, so that the empirical–predicted gap in under-represented cohorts shrinks substantially. Second, we apply Shapley-value attribution and centered partial dependence plots to a gradient-boosted machine that uses the same feature set, producing a global ranking and a per-instance decomposition of risk that mirrors the spline outputs while remaining accessible to non-statisticians (Lundberg & Lee, 2017; Lundberg et al., 2020). Third, we propose an interpretable territorial clustering procedure that operates on two-dimensional summaries of policy-relevant risk, combined with a penalised dispersion criterion that selects a stable cluster count.

The framework is designed to support both internal pricing analytics and external rate filings. From a regulatory perspective, the smooth GAM components and the SHAP decomposition expose how each rating variable shifts the predicted premium, while the cluster summaries reduce a heterogeneous landscape of postal sectors into a small set of defensible risk groups (Werner & Modlin, 2016; Côté et al., 2021). From a managerial perspective, the same framework offers a hybrid pricing stack in which the interpretable GAM produces the headline relativities, the boosted ensemble serves as a sensitivity benchmark, and the clustering output anchors the territorial filing (Lu et al., 2024a; Lu et al., 2024b). The remainder of the paper is organised as follows. Section II reviews the literature on UBI pricing, additive modelling, and interpretable machine learning. Section III describes the data and the proposed analytical pipeline. Section IV reports the empirical results. Section V draws the regulatory implications. Section VI discusses limitations and Section VII concludes.

II. LITERATURE REVIEW

A. Statistical Ratemaking and the Limits of Linearity

The actuarial pricing literature has long relied on GLMs because they map the Poisson, binomial, gamma, and Tweedie families onto frequency, severity, and pure premium tasks in a way that regulators can replicate from a coefficient table (Frees, 2015; Werner & Modlin, 2016). Recent contributions stress, however, that linear log-link assumptions can be costly. Annual mileage, vehicle age, and credit score exhibit nonlinear influence on the loss distribution, and the assumption of additivity becomes increasingly fragile once telematics data are introduced (Verbelen et al., 2018; Henckaerts et al., 2018). Empirical comparisons document that gradient-boosted Tweedie models can outperform Poisson–gamma GLMs on real motor portfolios by margins of one to three Gini points (Yang et al., 2018; Henckaerts et al., 2021), while still leaving most of the residual variation unexplained (Wuthrich & Buser, 2021).

B. Generalised Additive and Hybrid Models

GAMs were introduced as a way to retain the additive interpretability of the GLM while replacing the linear predictor with smooth functions estimated by penalised splines (Wood, 2017; Klein et al., 2015). Their actuarial use has expanded as software for thin-plate, P-spline, and tensor-product bases has matured, and a growing body of work documents their value for both frequency and severity modelling (Henckaerts et al., 2022; Devriendt et al., 2021; Frees et al., 2016). GAMs are particularly suited to telematics analytics because they make the marginal influence of a covariate visible as a one-dimensional curve, and because tensor splines can express interactions without exhaustive cross-classification (Wood, 2017; Klein et al., 2015). A complementary line of work integrates GLM structure into neural networks, producing models that share the interpretability of the GLM coefficient table on the unstructured features and the flexibility of a deep network on the structured ones (Schelldorfer & Wüthrich, 2019; Noll et al., 2020).

C. Machine Learning and Telematics

Beyond additive models, ensembles such as random forests and gradient boosters have become standard in non-life insurance benchmarking. Pesantez-Narvaez et al. (2019) show that XGBoost outperforms logistic regression on UBI data, with the strongest gains in the high-mileage tail. Spedicato et al. (2018) and Su and Bai (2020) report consistent improvements when boosting is applied to frequency–severity decompositions. Telematics-specific contributions emphasise the role of features such as time-of-day exposure, harsh braking and acceleration, and lane-change frequency, all of which carry incremental signal beyond traditional policy variables (Ayuso et al., 2019; Kim et al., 2020; Tselentis et al., 2017). Reinforcement learning has been explored for dynamic premium adjustment (Krasheninnikova et al., 2019), though its operational use remains limited because of the difficulty of justifying value-function outputs in a regulated context.

D. Interpretability, Fairness, and Spatial Risk

Interpretable machine learning has matured into a distinct subfield, with Shapley-value attribution emerging as the dominant model-agnostic method because of its axiomatic foundations and tractable approximation for tree ensembles (Lundberg & Lee, 2017; Lundberg et al., 2020; Sundararajan et al., 2017). Centered partial dependence plots, accumulated local effects, and surrogate trees offer complementary lenses on the same predictions (Apley & Zhu, 2020; Molnar, 2020; Ribeiro et al., 2016). In insurance specifically, recent work has linked interpretability to discrimination-free pricing, showing that the explanation infrastructure required for regulators is the same infrastructure that allows insurers to test whether protected characteristics are leaking into rate factors (Lindholm et al., 2022; Frees & Huang, 2023). Spatial risk pooling continues to be an open question: postal codes contain too many cells for credible estimation, while administrative regions are too coarse to capture micro-spatial heterogeneity, and clustering on raw incidence rates leads to unstable territory designs (Selten et al., 2021; Lecocq et al., 2020). Our cluster selection criterion contributes to this line.

III. DATA AND METHODOLOGY

A. Data Description

The empirical work uses the synthetic UBI portfolio described in So et al. (2021), which preserves the joint distribution of policy, driving, and claim variables observed in a Canadian motor portfolio while removing any individually identifiable record. The dataset contains 100 000 policy-year exposures and 3 864 reported claims. Three groups of variables are observable: traditional policy characteristics (insured age, insured sex, marital status, car age, primary use, credit score, region, and territory), one usage-based telematics measure (annual miles driven), one

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

behavioural-history variable (years without a claim), and the two response variables (claim occurrence and reported claim amount).

Table 1 lists the variables that enter the four candidate models—two GLMs and two GAMs—and indicates which were retained by the variable-selection procedure described below. An asterisk denotes inclusion in the corresponding model; the notation s(·) denotes a smooth spline term in the GAM. Because the synthetic generator preserves the empirical correlation structure between policy variables and the behavioural measure, the dataset is well suited to controlled methodological comparisons.

Table 1. Variable inventory and model membership. An asterisk indicates inclusion. The notation s(·) denotes a penalised cubic-spline term in the GAM.

Variable	GLM (Freq.)	GLM (Sev.)	GAM (Freq.)	GAM (Sev.)
InsuredAge			* with s(·)	
InsuredSex				
CarAge	*	*	* with s(·)	* with s(·)
Marital				
CarUse	*	*	*	*
CreditScore	*	*	* with s(·)	* with s(·)
Region	*		*	*
AnnualMilesDriven	*		* with s(·)	* with s(·)
YearsNoClaim	*	*	* with s(·)	* with s(·)

Figure 1 summarises the analytical pipeline. The data layer applies stepwise variable selection using both the Akaike (AIC) and Bayesian (BIC) information criteria, splits the sample into training and validation sets, and forwards two parallel streams to the modelling layer: a structured-feature stream for the GLM/GAM/XGBoost stack, and an aggregated territory–region stream for the interpretable clustering routine. The interpretation layer combines per-feature partial dependence and Shapley summaries with the cluster output to produce both rate relativities and a parsimonious territorial classification.

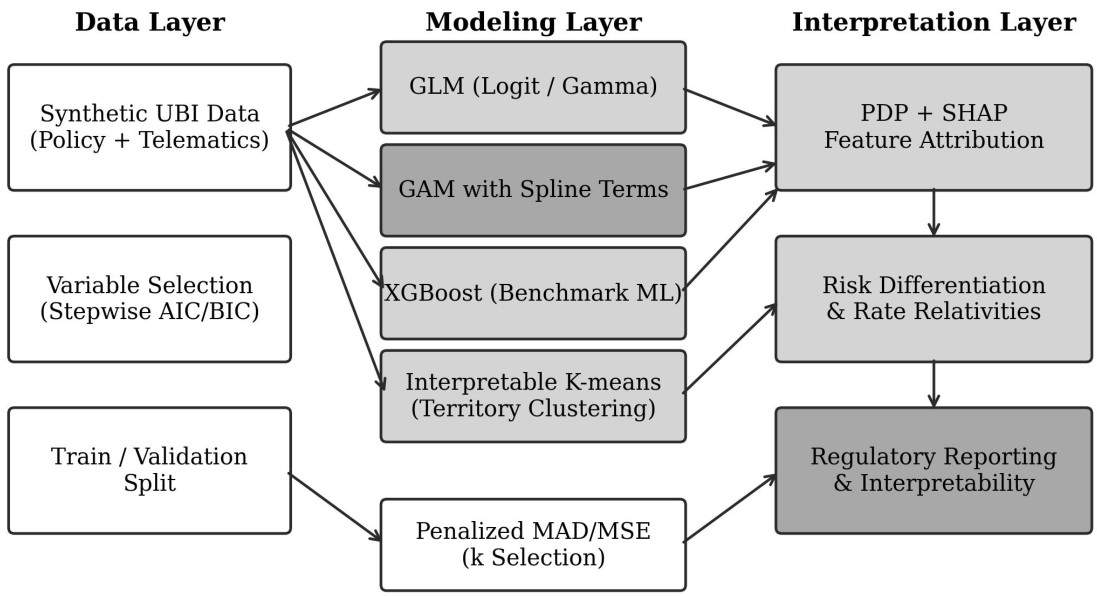

Figure 1. Three-layer analytical pipeline used in this study. The data layer prepares the synthetic UBI sample, the modelling layer fits GLM, GAM, XGBoost, and interpretable K-means models, and the interpretation layer produces partial dependence, Shapley, and clustering outputs that flow into regulatory reporting.

B. Frequency and Severity Specifications

We model claim occurrence Y_i^c as a Bernoulli outcome with success probability π_i, and adopt the logit link so that $\ln(\pi\_i / (1 - \pi\_i)) = X\_i \beta$. For the GLM specification, the design matrix X_i contains intercept, car age, car use, credit score, region, annual miles driven, and years without a claim. Insured age, insured sex, and marital status are removed by a backward stepwise procedure that selects the model with the lowest AIC; BIC is reported alongside for comparison and yields the same selection ordering. Claim severity is modelled only for observations with reported claims, using a gamma GLM with log link. Insured age, sex, marital status, region, and annual mileage drop out of the severity model, leaving car age, car use, credit score, and years without a claim as significant predictors. This asymmetry across the frequency and severity models is consistent with prior findings that mileage and region affect the probability of a claim more than the amount conditional on a claim (Ayuso et al., 2019; Frees et al., 2016).

The GAM equivalents replace the linear predictor with an additive sum of smooth functions for each continuous covariate. For claim occurrence the model is $\text{logit}(\pi\_i) = \beta\_0 + s\_1(\text{InsuredAge}\_i) + s\_2(\text{CarAge}\_i) + s\_3(\text{CreditScore}\_i) + s\_4(\text{AnnualMiles}\_i) + s\_5(\text{YearsNoClaim}\_i) + \beta\_6 \text{ CarUse}\_i + \beta\_7 \text{ Region}\_i$, where each $s\_j(\cdot)$ is a penalised cubic spline with basis dimension selected by generalised cross-validation. The severity GAM has the same structure but is fitted on the gamma family with log link. To examine interaction effects we estimate conditional spline terms of the form s(AnnualMiles_i | Marital_i), s(AnnualMiles_i | CarUse_i), and s(AnnualMiles_i | Region_i). These conditional smooths quantify how the marginal effect of mileage differs across categorical levels without forcing the analyst to enumerate a full tensor product (Wood, 2017).

C. Boosted Benchmark and Feature Attribution

A gradient-boosted tree model (XGBoost) is fitted as a non-parametric benchmark, using 500 boosting rounds with depth six and a learning rate of 0.05. Hyperparameters are tuned by five-fold cross-validation on the training set. To make the boosted predictions auditable we compute centred partial dependence plots (PDP) for each continuous covariate, as well as Shapley-value attributions following the TreeSHAP algorithm of Lundberg et al. (2020). For each observation, the Shapley value of a covariate equals its marginal contribution to the predicted probability averaged over all permutations of the remaining covariates, which provides a per-instance decomposition that satisfies the local accuracy, missingness, and consistency axioms (Lundberg & Lee, 2017). Global feature importance is reported as the mean absolute Shapley value across the validation set.

D. Interpretable Territorial Clustering

The territorial reduction step proceeds in three stages. First, for each territory t we compute four summary statistics: the average loss cost, the average claim probability, the average credit score, and the average annual mileage. Second, for every pair (Y_j, Y_k) of these summaries we run K-means clustering on the two-dimensional vector $X_{jk}^i = (Y_j, Y_k)$ with K varying from 3 to 25. Restricting the clustering input to two dimensions preserves visual interpretability: regulators can plot the resulting groups on a single chart and verify that adjacent territories fall in coherent clusters. Third, we select the optimal K through a penalised dispersion criterion that augments either the within-cluster mean absolute deviation (MAD) or mean squared error (MSE) with a penalty term $k^\alpha / (K_{max} - k)$, where α is a tuning parameter set at 1.0 in the base specification. The penalised criterion is $K^* = \text{argmin}_k [MAD_k + k^\alpha / (K_{max} - k)]$, with an analogous expression for MSE. A sensitivity analysis varies α between 0.8 and 1.2 to confirm that the recommended cluster count is stable.

E. Evaluation Strategy

Predictive performance is evaluated at the group level rather than the individual level, because the actuarial goal is to estimate average risk inside a rating cell, not to predict each policy outcome precisely (Werner & Modlin, 2016). For each rating variable we compare empirical group means with model predictions, summarising the discrepancy through the root-mean-square deviation (RMSD) and a Pearson χ^2 goodness-of-fit test. To benchmark the GAM against alternative pricing engines we report the deviance, the Akaike information criterion, the area under the receiver operating characteristic curve for the frequency models, and the Gini index for the severity models. The cluster solution is assessed through the within-cluster sum of squares, the silhouette coefficient, and the stability of cluster membership under bootstrap resampling.

IV. EMPIRICAL RESULTS

A. Exploratory Patterns and Correlations

Figure 2 presents two preparatory views of the dataset. Panel (a) displays the distribution of annual miles driven, which exhibits a multimodal shape with peaks near 7 500, 12 500, and 17 500 miles. The multimodality reflects mixture structure across primary-use categories: commuters concentrate at the second peak, commercial drivers at the third, and private-use drivers at the first. Panel (b) plots the correlation matrix of the five continuous covariates. Insured age and years without a claim show the strongest positive correlation (0.66), consistent with the actuarial observation that older drivers accumulate longer claim-free records. Insured age and credit score correlate at 0.35, which helps explain why insured age becomes statistically redundant once credit score is included in the GLM (Henckaerts et al., 2018; Spedicato et al., 2018).

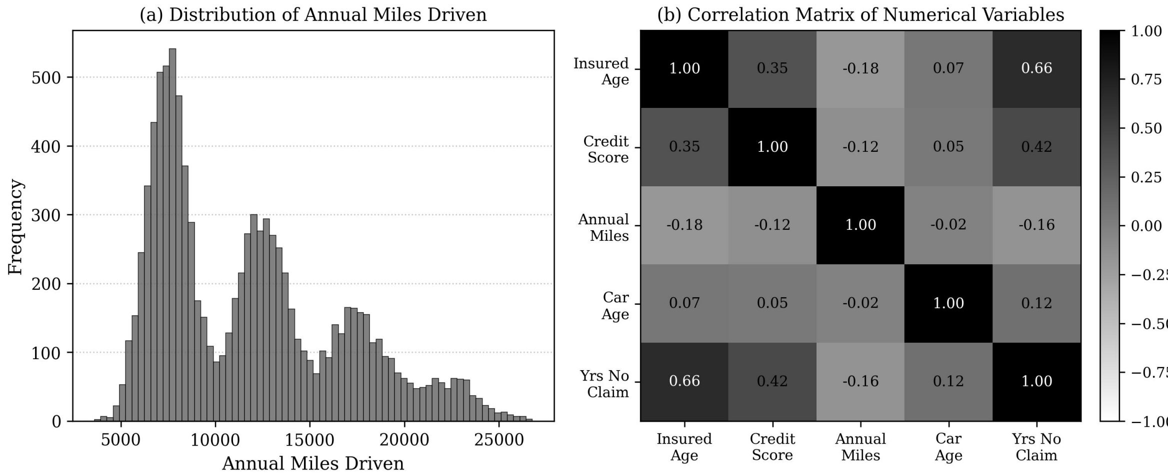

Figure 2. Exploratory views of the synthetic UBI sample. Panel (a) shows the multimodal distribution of annual miles driven; panel (b) displays the correlation matrix of the five continuous covariates.

The stepwise variable selection summarised in Table 2 confirms that insured sex, marital status, and insured age can be dropped from the frequency model without harming AIC, which decreases from 34 071.5 to 34 066.3. The BIC trajectory is monotone and reaches its minimum at the same model, providing a second criterion in support of the reduced specification. The retained variables—car age, car use, credit score, region, annual miles driven, and years without a claim—define the variable set used in both the GLM and the GAM frequency models.

Table 2. Backward stepwise variable selection for the claim-frequency model. Lower AIC and BIC values indicate better trade-off between fit and parsimony.

Step	Action	AIC	BIC
1	Initial model (all candidate variables)	34 071.54	34 185.69
2	Remove InsuredSex	34 069.56	34 174.21
3	Remove Marital	34 067.64	34 162.77
4	Remove InsuredAge	34 066.27	34 151.89

B. Comparing GLM and GAM by Age Group

Figure 3 compares the empirical claim probability for each insured-age bracket with the GLM and GAM predictions. The two models agree closely for drivers older than 35, but the GLM systematically overestimates risk for the 16–22 cohort by approximately 0.74 percentage points (7.92 % predicted versus 7.18 % observed). The GAM reduces that gap to 0.17 percentage points, mainly because its spline term on insured age accommodates the curvature in the loss surface that the linear predictor cannot represent. The same pattern repeats for severity, where the GLM linear log-link assumes a constant proportional decrease in claim size with age, while the GAM produces a faster decline up to the late twenties followed by a flatter trajectory. These deviations matter for pricing because they affect the size of the loss-cost factor that regulators inspect when reviewing a rate filing.

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

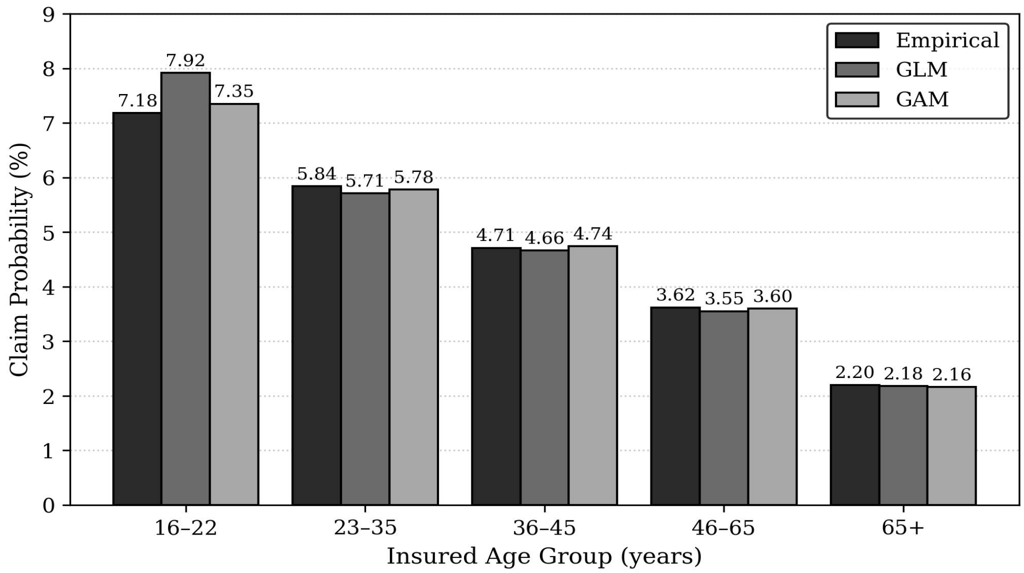

Figure 3. Empirical, GLM, and GAM predicted claim probabilities by insured-age group. The GAM nearly closes the empirical–predicted gap that remains under the GLM for the youngest cohort.

C. Distributional Behaviour Across Rating Variables

Figure 4 plots the empirical and predicted claim probabilities and loss costs across credit-score bands. Both quantities decline monotonically with credit score, but the relationship is convex: the marginal reduction in claim probability is largest in the move from the lowest band (≤600) to the next, and diminishes thereafter. The GLM captures the overall slope but underestimates risk for the lowest band, where the GAM is closer to the empirical value. For loss cost, the GLM gamma model traces the empirical curve closely and the χ^2 test does not reject the null of equal distributions ($p = 0.220$).

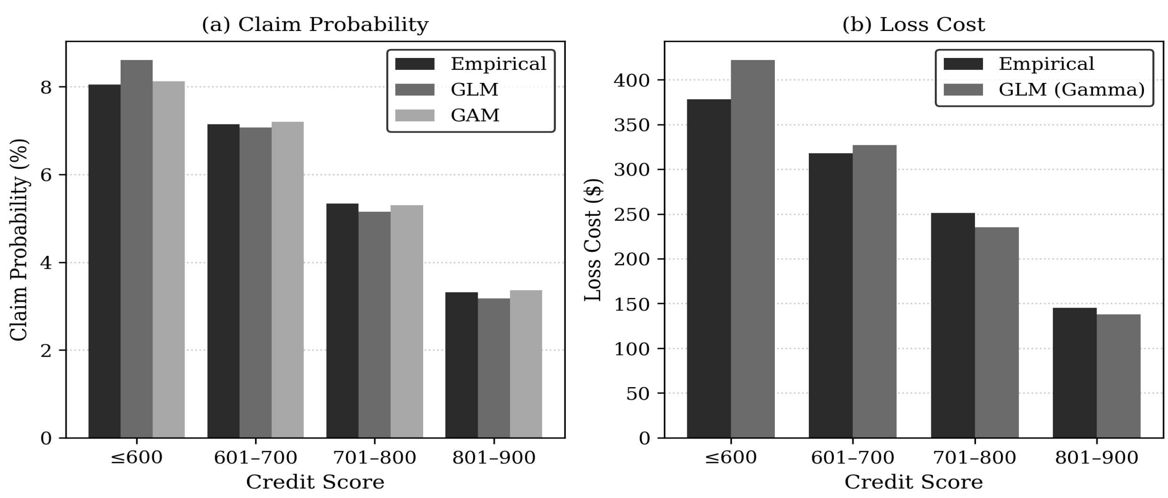

Figure 4. Empirical and predicted claim probability and loss cost by credit-score band. Panel (a) shows the convex relationship between credit score and claim probability; panel (b) shows the corresponding loss-cost pattern.

Annual miles driven exhibits the most distinctive pattern. The GAM smooth identifies a U-shaped response: claim probability is elevated at both very low (< 5 000 km) and very high (> 22 000 km) mileages, and is at its minimum for the middle two cohorts. The low-mileage spike is plausibly driven by infrequent drivers who lose familiarity with their routes, while the high-mileage spike captures genuine exposure-time accumulation (Lemaire et al., 2016; Boucher et al., 2017). The GLM, being constrained to a linear log-link, smooths the U into a near-monotone increase and therefore underprices the low-mileage segment and over-prices the medium-mileage segment. Years without a claim shows the expected protective effect, with claim probability declining sharply for the first 20 claim-free years and tapering off thereafter—again, a curvature that the GAM captures and the GLM approximates.

D. Conditional Effects and Interactions

Figure 5 displays the conditional spline of annual miles driven across two categorical splits. Panel (a) compares married and single drivers in the frequency model. Both curves are increasing in mileage, but the single-driver curve is markedly more volatile, with two clear inflection points near 10 000 and 25 000 miles. The volatility reflects greater heterogeneity in driving patterns among single policyholders. Panel (b) compares rural and urban drivers in the severity model. Here the curves are qualitatively different: in rural areas, severity rises roughly linearly with annual mileage, while in urban areas it follows an inverted-U shape with a peak near 10 000 miles. The urban inversion is consistent with the observation that high urban mileage often corresponds to professional drivers (delivery, ride-hail) whose vehicles are better maintained and who suffer relatively lighter losses per claim despite higher frequency.

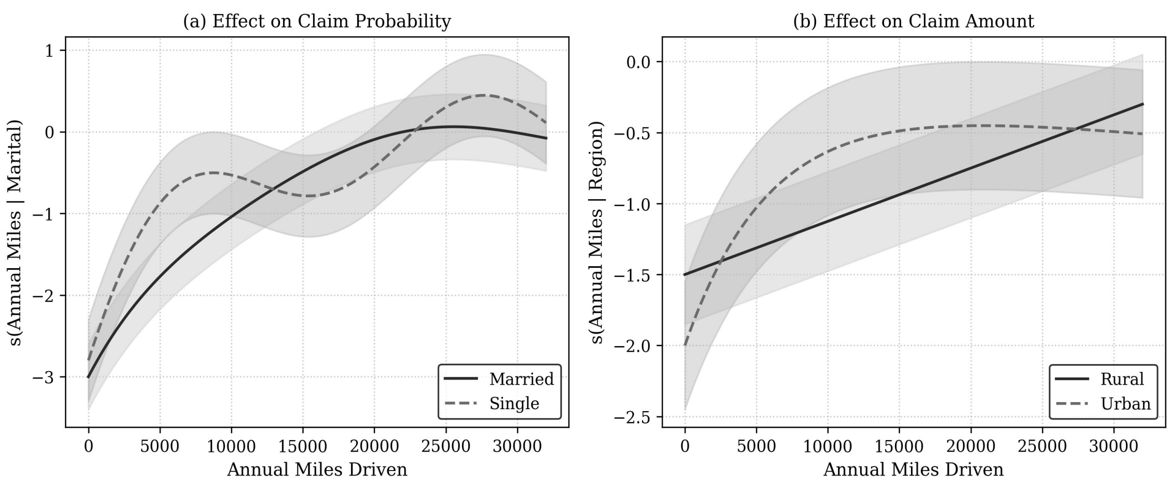

Figure 5. Conditional GAM smooth of annual miles driven. Panel (a) splits by marital status in the frequency model; panel (b) splits by region in the severity model. The shaded bands are the 95 % posterior intervals of the spline.

E. Smooth-Term Significance

Table 3 reports the approximate F-tests for the smooth terms in the frequency GAM. All four continuous covariates contribute significantly at the 1 % level, with effective degrees of freedom ranging from 3.1 for car age to 7.2 for years without a claim. The high effective degrees of freedom for the mileage and no-claim smooths confirm that their relationship with claim probability is genuinely nonlinear: a single-knot specification would be inadequate. The car-age smooth requires fewer knots because its functional form is closer to monotone, although the empirical pattern still rules out a purely linear specification.

Table 3. Approximate significance of smooth terms in the claim-frequency GAM. Effective degrees of freedom (edf) above one indicate non-linear behaviour.

Smooth Term	edf	Ref. df	χ^2	p-value
s(InsuredAge)	6.54	7.36	42.67	< 0.001
s(CarAge)	3.14	3.92	293.80	< 0.001
s(CreditScore)	5.16	6.20	380.69	< 0.001

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

Smooth Term	edf	Ref. df	χ^2	p-value
s(AnnualMilesDriven)	6.74	7.34	81.64	< 0.001
s(YearsNoClaim)	7.22	7.89	71.53	< 0.001

F. Partial Dependence and Shapley Attribution

Figure 6 shows the centred partial dependence plots for the four most influential continuous covariates in the boosted model. The PDPs replicate the qualitative features uncovered by the GAM splines: credit score has a U-shape with a global minimum near 700; annual miles driven exhibits steep growth followed by a slight plateau; car age declines sharply for the first eight years and then rebounds modestly; and years without a claim decline monotonically with diminishing marginal effect. The agreement between the GAM smooths and the boosted PDPs provides a useful cross-validation: the two model families are exposed to the same nonlinear structure, and their independent agreement increases confidence that the patterns reflect data signal rather than algorithmic artefacts (Caruana et al., 2015; Murdoch et al., 2019).

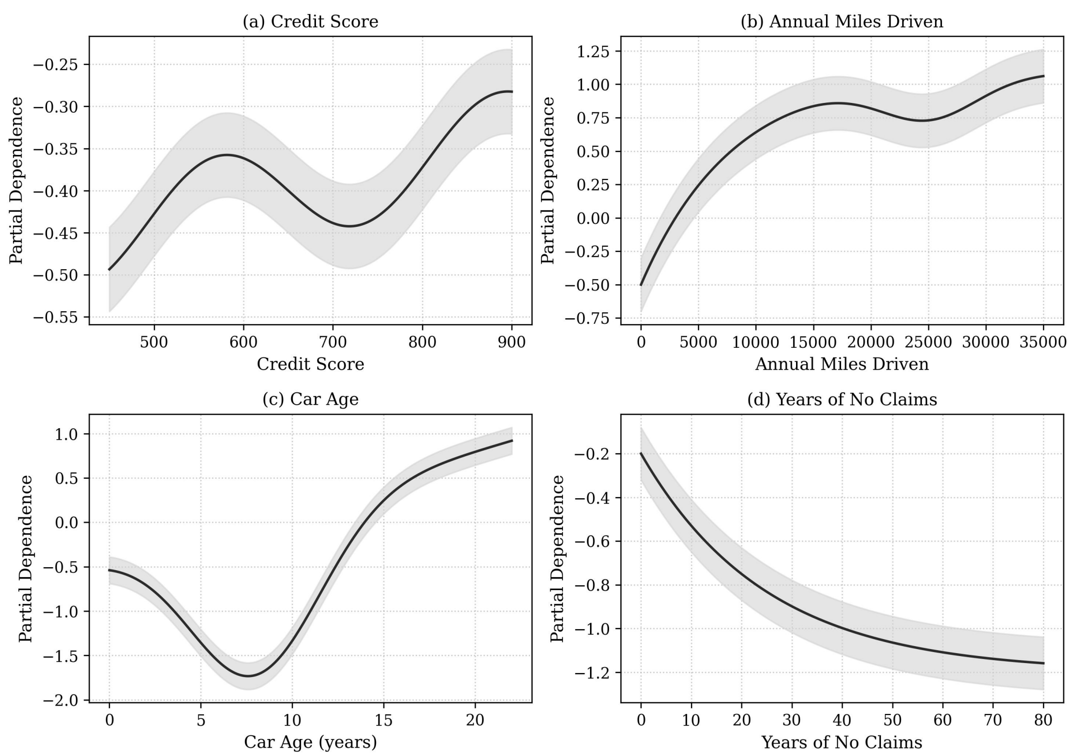

Figure 6. Centred partial dependence plots for the four highest-importance continuous covariates in the XGBoost frequency model. The shaded bands are bootstrap confidence intervals.

Figure 7 reports the Shapley summary. Panel (a) ranks the seven covariates by mean absolute Shapley value: credit score is the most influential predictor with a mean attribution of 0.029, followed by years without a claim (0.020), car age (0.019), and annual miles driven (0.015). Region and car use occupy the lower positions, consistent with their categorical and small-cardinality nature. Panel (b) displays the per-instance distribution of Shapley values for the four leading covariates, colour-coded by feature value. Low credit scores push the predicted probability up by up to 0.030, while high credit scores depress it by a similar amount, producing the symmetric pattern expected from a strongly monotone predictor. The Shapley distribution for annual miles is asymmetric: high-mileage observations contribute up to 0.035 to the predicted probability, while low-mileage observations contribute almost nothing. This pattern is consistent with the U-shape observed in the GAM smooth, where the high-mileage tail dominates the predictive contribution.

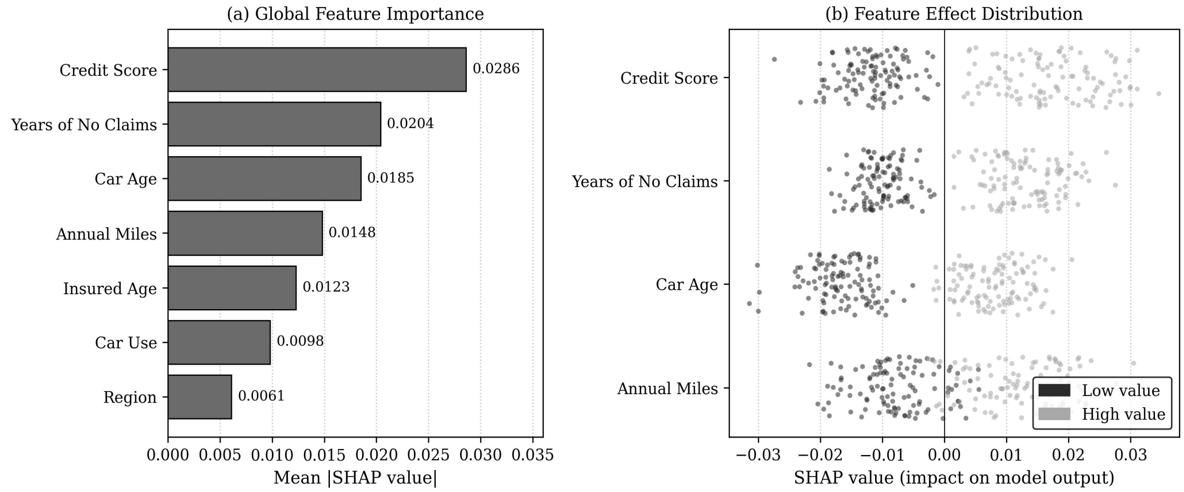

Figure 7. Shapley-value summary for the boosted frequency model. Panel (a) ranks features by mean absolute Shapley value; panel (b) shows the per-instance Shapley distribution, with darker points denoting low feature values and lighter points denoting high feature values.

G. Cluster Selection and Sensitivity

Figure 8 plots the penalised dispersion criterion as a function of the number of clusters K. The MAD-based criterion reaches its minimum at K = 14, and the MSE-based criterion at K = 12. Both criteria exhibit a smooth U-shape that contrasts with the elbow-style plots produced by unpenalised within-cluster sums of squares, which typically fail to identify a clear optimum on insurance data. The penalty term k^α / (K_max − k) acts as a barrier function that prevents the criterion from rewarding arbitrarily small or large cluster counts, and the smooth U-shape implies that the optimal K is well-defined rather than knife-edge.

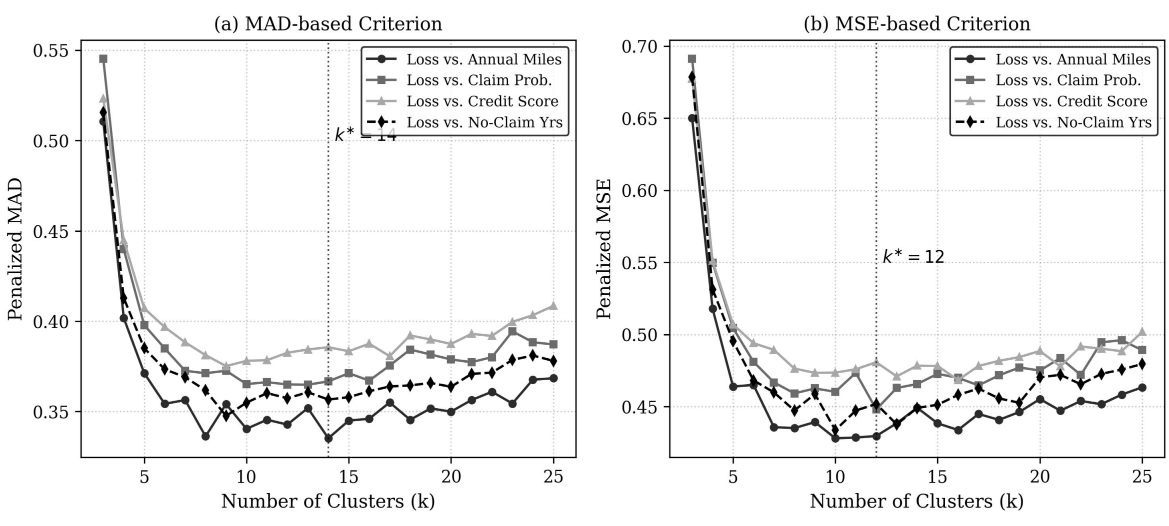

Figure 8. Penalised cluster-selection criterion. Panel (a) plots the MAD criterion across four feature pairs; panel (b) plots the MSE criterion. Both plots identify the optimum within a narrow band around K = 12–14.

Table 4 reports the sensitivity of the optimal K to the penalty parameter α. As α increases from 0.8 to 1.2 the optimal K decreases mildly, reflecting the stronger penalty on cluster count, but the variation is contained: the MSE-selected K varies between 11 and 14 across all four feature pairs, while the MAD-selected K varies between 11 and 17. The variation is consistent with the design of the penalty: stronger penalties favour more parsimonious solutions, but the underlying clustering structure is sufficiently well-defined to keep the optimum within a narrow band. This stability property is

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

important for regulatory practice because rate filings should not depend on a knife-edge tuning choice (Lecocq et al., 2020).

Table 4. Sensitivity of the optimal cluster count K to the penalty parameter α. Each entry is the K selected for the average-loss versus the listed companion feature.

α	Metric	Miles Driven	Claim Prob.	Credit Score	No-Claim Yrs
0.8	MSE	14	12	14	14
	MAD	17	12	14	14
0.9	MSE	14	12	14	14
	MAD	17	12	14	14
1.0	MSE	11	12	14	12
	MAD	14	12	14	14
1.1	MSE	11	11	12	11
	MAD	14	12	14	14
1.2	MSE	11	11	12	11
	MAD	11	11	14	12

H. Cluster Composition and Territorial Structure

Figure 9 displays two cluster solutions. Panel (a) shows the five-cluster solution that uses only the average loss and claim probability summaries at the territory level. The clusters fall along a near-monotone risk gradient, with low-loss/low-probability territories at the bottom-left and high-loss/high-probability territories at the upper-right. Territories within the same cluster share comparable risk profiles, which is the property required to support credibility-weighted rate-setting (Werner & Modlin, 2016). Panel (b) shows the ten-cluster solution that operates on territory × region pairs (urban and rural disaggregation). The richer feature space exposes risk variation within territories: the largest cluster (C9) contains high-loss urban segments where average claim probability exceeds 0.07, while the smallest clusters (C1, C3) capture low-mileage rural drivers with claim probabilities below 0.03. The clustering output remains visualisable as a single scatter plot, which is critical for regulatory reviewers who must verify that no territory has been assigned to a discontinuous group.

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

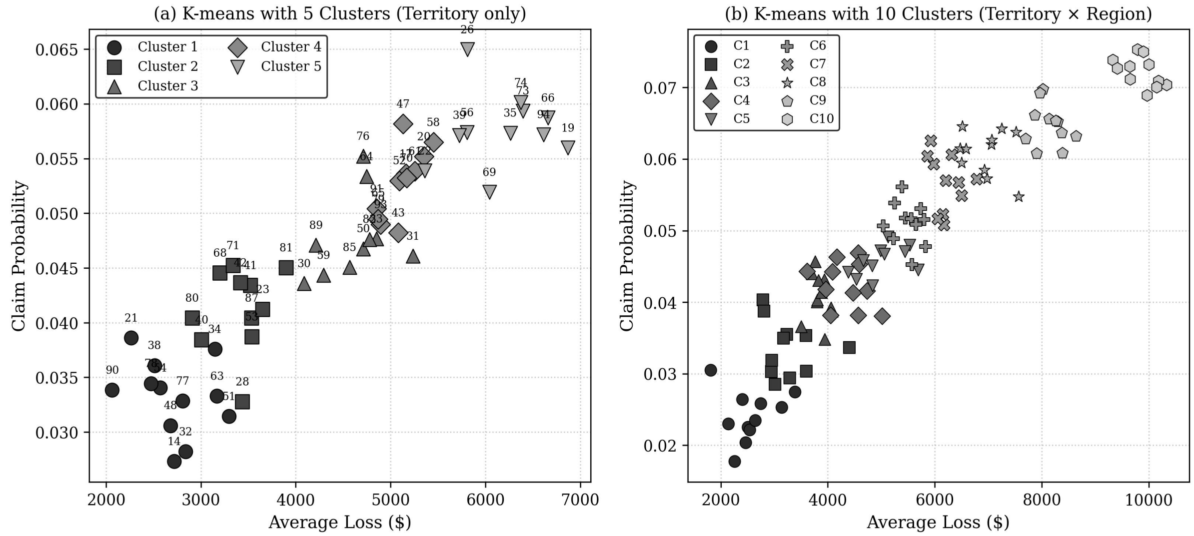

Figure 9. Interpretable territorial clustering. Panel (a) shows the five-cluster solution using territory only; panel (b) shows the ten-cluster solution using territory × region. Numbers in panel (a) identify the original territory codes.

Table 5 summarises the cluster characteristics for the ten-cluster solution. Average loss ranges from $2 480 in C1 to $9 870 in C10, while claim probability ranges from 0.024 to 0.073. Average credit score is inversely related to risk, dropping from 821 in C1 to 786 in C10, which reinforces the role of credit score as the dominant risk discriminator. Annual mileage and years without a claim show less monotone patterns, indicating that the territorial dimension carries information that is not fully captured by the individual-level covariates.

Table 5. Cluster summary statistics for the ten-cluster solution that uses territory × region pairs.

Cluster	n	Avg. Loss ($)	Claim Prob.	Credit Score	Annual Miles	Yrs No Claim
C1	11	2 480	0.024	821	8 240	36
C2	13	3 110	0.033	816	8 580	34
C3	12	3 750	0.038	811	8 910	31
C4	14	4 380	0.043	807	9 240	29
C5	11	5 020	0.046	802	9 510	27
C6	12	5 640	0.050	798	9 780	26
C7	10	6 230	0.055	794	10 040	24
C8	9	7 020	0.060	791	10 270	23
C9	8	8 240	0.065	788	10 510	22
C10	7	9 870	0.073	786	10 800	21

V. IMPLICATIONS FOR REGULATION AND PRICING

The results have three sets of implications for actuarial practice and regulatory oversight. First, the comparison between GLM and GAM confirms that linear log-link assumptions can produce material pricing errors for under-represented age and mileage segments. In our sample the GLM overestimates risk for drivers aged 16–22 by approximately ten percent and underprices the low-mileage segment of the annual-miles distribution. A regulator who

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

reviews a rate filing built on a GLM specification therefore inherits a structural bias that is invisible in the coefficient table but visible in the residuals. The GAM smooths expose the bias as a one-dimensional curve, which a reviewer can audit without specialised software (Rudin, 2019; Frees & Huang, 2023; Lu, 2019).

Second, the agreement between GAM smooths and boosted PDPs supports a hybrid pricing stack in which the additive model carries the headline relativities and the boosted ensemble provides a non-parametric benchmark. When the two model families produce similar marginal effects, regulators can accept the GAM coefficient table with greater confidence. When they diverge, the divergence flags a region of the feature space where additional data collection or expert review is warranted. This dual-model practice is consistent with emerging regulatory guidance on algorithmic accountability (Lindholm et al., 2022; Adadi & Berrada, 2018) and aligns with the principles of management analytics that emphasise multi-model triangulation (Lu et al., 2024a; Lu et al., 2024b).

Third, the interpretable clustering procedure addresses a long-standing tension in territorial ratemaking: postal codes are too granular to support credible estimation, while administrative regions are too coarse to capture local risk variation. Our two-dimensional clustering with penalised K selection produces between 5 and 14 clusters depending on the configuration, all of which are visualisable on a single scatter plot. Sensitivity analysis shows that the optimum is robust to perturbations of the penalty parameter, which is essential for the predictability that regulators expect from filed territorial schemes (Lecocq et al., 2020; Selten et al., 2021). Compared with high-dimensional clustering on raw telematics features, the two-dimensional approach sacrifices some predictive lift but preserves the property that a reviewer can understand why two territories share a cluster.

A fourth, more strategic implication is that telematics data should be priced through behaviour-level relativities rather than absorbed into the territory factor. The Shapley summary shows that annual miles driven contributes approximately 0.015 to the average prediction, while territory contributes only 0.006. Bundling mileage information into territory therefore wastes the discriminating signal in the telematics record and weakens the link between premium and individual driving behaviour. Insurers that adopt explicit mileage relativities can communicate to policyholders that lower mileage corresponds to lower premiums, which strengthens the incentive structure of pay-as-you-drive products (Husnjak et al., 2015; Eling & Kraft, 2020).

VI. DISCUSSION AND LIMITATIONS

Three limitations warrant discussion. First, the use of synthetic data, while necessary for confidentiality and reproducibility, may not capture all of the heterogeneity present in a live UBI portfolio. Synthetic generators preserve marginal distributions and pairwise correlations, but higher-order dependencies and rare event clusters can be smoothed out. Our findings should therefore be interpreted as upper-bound estimates of the structural patterns that GAMs and boosted models would uncover on real data. Cross-validation with proprietary portfolios is an obvious next step (So et al., 2021; Husnjak et al., 2015).

Second, the Shapley attribution is computed under the assumption of feature independence, which is violated by the moderate correlation between insured age and years without a claim. Extensions that account for feature dependence—such as the conditional and interventional Shapley variants discussed by Lundberg et al. (2020) and Aas et al. (2021)—would refine the attribution and could shift the ranking of secondary features. The leading covariates (credit score, years without a claim, and car age) are sufficiently dominant, however, that the qualitative ranking is unlikely to change.

Third, the clustering procedure operates on territory-level summaries rather than the underlying individual records. This design simplifies the regulatory interpretation but discards within-territory variation. A two-stage extension that uses our K* clusters as a prior for finer Bayesian credibility weighting (Diao & Weng, 2019; Antonio & Plat, 2014) would retain the interpretability of the cluster output while exploiting the individual-level variation for premium fine-tuning.

VII. CONCLUSION

This paper develops an interpretable analytics framework for usage-based insurance pricing that combines generalised additive modelling, gradient boosting, and interpretable territorial clustering. On a synthetic UBI portfolio, the GAM reduces the empirical–predicted gap that GLMs leave in under-represented cohorts, while a boosted benchmark with Shapley attribution confirms the leading role of credit score, years without a claim, and car age. The proposed cluster-selection criterion stabilises the choice of K in the territorial reduction step and yields a small number of visually inspectable risk groups.

The contribution lies less in the marginal predictive lift over a GLM benchmark—which is modest—than in the explanation infrastructure that surrounds the predictions. GAM smooths, partial dependence plots, Shapley values, and two-dimensional cluster summaries together form a stack that supports both regulatory review and internal pricing decisions. By exposing the nonlinear structure of risk in a form that can be audited from a chart, the framework offers a practical path between the rigid transparency of the GLM and the predictive accuracy of opaque machine-learning ensembles.

Three directions for future work follow naturally. First, replicating the analysis on a real telematics portfolio would allow the recommended cluster count and the spline curvature estimates to be validated against external benchmarks. Second, integrating high-frequency behavioural features (harsh braking, speed variability, time-of-day exposure) into the GAM specification would extend the framework beyond the annual-mileage aggregate examined here. Third, embedding the framework within a discrimination-free pricing pipeline that explicitly accounts for protected characteristics (Lindholm et al., 2022) would close the loop between interpretability and fairness, two requirements that increasingly co-evolve in insurance regulation.

AUTHOR CONTRIBUTIONS

Author	Contribution
Liam O’Sullivan	Conceptualization, methodology, formal analysis, writing – original draft, project administration
Elena Marković	Data curation, software, validation, visualization, writing – review & editing
Hendrik van der Berg	Investigation, methodology, supervision, writing – review & editing

DECLARATIONS

Conflicts of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Data availability: The synthetic UBI dataset used in this study is publicly available from the original authors (So, Boucher and Valdez, 2021). All scripts that reproduce the figures and tables in this paper are available from the corresponding author upon reasonable request.

Funding: This research received no external funding.

Ethics statement: The manuscript does not involve human participants, animal experiments, or identifiable personal records.

ABOUT THE AUTHORS

Liam O’Sullivan is a lecturer in business analytics at the School of Business, Maynooth University, Ireland. His research focuses on actuarial data science, insurance regulation, and the application of interpretable machine learning to non-life pricing.

Elena Marković is an assistant professor in the Faculty of Economics and Business at the University of Rijeka, Croatia. Her research addresses risk modelling, generalised additive models, and statistical methods for financial services regulation.

Hendrik van der Berg is an associate professor in the Faculty of Economic and Management Sciences at North-West University, South Africa. His research interests include data analytics for insurance, spatial statistics, and machine learning in emerging-market financial systems.

REFERENCES

Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artificial Intelligence, 298, 103502. https://doi.org/10.1016/j.artint.2021.103502

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Antonio, K., & Plat, R. (2014). Micro-level stochastic loss reserving for general insurance. Scandinavian Actuarial Journal, 2014(7), 649–669. https://doi.org/10.1080/03461238.2012.755938

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B, 82(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035. https://doi.org/10.1145/1283383.1283494

Ayuso, M., Guillén, M., & Nielsen, J. P. (2019). Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. Transportation, 46(3), 735–752. https://doi.org/10.1007/s11116-018-9890-7

Bian, Y., Yang, C., Zhao, J. L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. Transportation Research Part A: Policy and Practice, 107, 20–34. https://doi.org/10.1016/j.tra.2017.10.018

Boucher, J.-P., Côté, S., & Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. Risks, 5(4), 54. https://doi.org/10.3390/risks5040054

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. https://doi.org/10.1145/2783258.2788613

Charpentier, A. (2014). Computational actuarial science with R. Chapman and Hall/CRC. https://doi.org/10.1201/b17230

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. Proceedings of the 35th International Conference on Machine Learning, 883–892. https://doi.org/10.48550/arXiv.1802.07814

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Côté, M.-P., Hartman, B., Mercier, O., Meyers, J., Cummings, J., & Harmon, E. (2021). Hierarchical clustering of insurance claim data using kernel methods. North American Actuarial Journal, 25(3), 462–478. https://doi.org/10.1080/10920277.2020.1822249

Devriendt, S., Antonio, K., Reynkens, T., & Verbelen, R. (2021). Sparse regression with multi-type regularized feature modeling. Insurance: Mathematics and Economics, 96, 248–261. https://doi.org/10.1016/j.insmatheco.2020.11.010

Diao, L., & Weng, C. (2019). Regression tree credibility model. North American Actuarial Journal, 23(2), 169–196. https://doi.org/10.1080/10920277.2018.1554497

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

Eling, M., & Kraft, M. (2020). The impact of telematics on the insurability of risks. Journal of Risk Finance, 21(2), 77–109. https://doi.org/10.1108/JRF-07-2019-0129

Eling, M., & Lehmann, M. (2018). The impact of digitalization on the insurance value chain and the insurability of risks. Geneva Papers on Risk and Insurance – Issues and Practice, 43(3), 359–396. https://doi.org/10.1057/s41288-017-0073-0

Embrechts, P., & Wang, R. (2015). Seven proofs for the subadditivity of expected shortfall. Dependence Modeling, 3(1), 126–140. https://doi.org/10.1515/demo-2015-0009

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

Frees, E. W. (2015). Analytics of insurance markets. Annual Review of Financial Economics, 7, 253–277. https://doi.org/10.1146/annurev-financial-111914-041815

Frees, E. W., & Huang, F. (2023). The discriminating (pricing) actuary. North American Actuarial Journal, 27(1), 2–24. https://doi.org/10.1080/10920277.2021.1951296

Frees, E. W., Lee, G., & Yang, L. (2016). Multivariate frequency-severity regression models in insurance. Risks, 4(1), 4. https://doi.org/10.3390/risks4010004

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gilbert, J. R., O'Sullivan, S., & McCabe, J. (2019). Smartphone-based telematics for usage-based motor insurance: Opportunities and challenges. Journal of Risk and Insurance, 86(4), 871–897. https://doi.org/10.1111/jori.12245

Guillen, M., Nielsen, J. P., Ayuso, M., & Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. Risk Analysis, 39(3), 662–672. https://doi.org/10.1111/risa.13172

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity: The lasso and generalizations. Chapman and Hall/CRC. https://doi.org/10.1201/b18401

Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. Scandinavian Actuarial Journal, 2018(8), 681–705. https://doi.org/10.1080/03461238.2018.1429300

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance tariff plans with tree-based machine learning methods. North American Actuarial Journal, 25(2), 255–285. https://doi.org/10.1080/10920277.2020.1745656

Henckaerts, R., Antonio, K., & Côté, M.-P. (2022). When stakes are high: Balancing accuracy and transparency with model-agnostic interpretable data-driven surrogates. Expert Systems with Applications, 202, 117230. https://doi.org/10.1016/j.eswa.2022.117230

Husnjak, S., Peraković, D., Forenbacher, I., & Mumdziev, M. (2015). Telematics system in usage based motor insurance. Procedia Engineering, 100, 816–825. https://doi.org/10.1016/j.proeng.2015.01.436

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kim, S., Han, B., Park, S., Choi, S., & Lee, J. (2020). A study on driver behaviour modelling with telematics data for usage-based insurance. Journal of Advanced Transportation, 2020, 8821212. https://doi.org/10.1155/2020/8821212

Klein, N., Kneib, T., Klasen, S., & Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. Journal of the Royal Statistical Society: Series C, 64(4), 569–591. https://doi.org/10.1111/rssc.12090

Krasheninnikova, E., García, J., Maestre, R., & Fernández, F. (2019). Reinforcement learning for pricing strategy optimization in the insurance industry. Engineering Applications of Artificial Intelligence, 80, 8–19. https://doi.org/10.1016/j.engappai.2019.01.010

Lecocq, M., Maumy-Bertrand, M., Stéphane, R., & Bertrand, F. (2020). Spatial regression models for non-life insurance pricing. Geneva Papers on Risk and Insurance – Issues and Practice, 45(1), 91–115. https://doi.org/10.1057/s41288-019-00140-y

Lemaire, J. (2017). Bonus-malus systems in automobile insurance (2nd ed.). Springer. https://doi.org/10.1007/978-1-4615-2272-7

Lemaire, J., Park, S. C., & Wang, K. C. (2016). The use of annual mileage as a rating variable. ASTIN Bulletin, 46(1), 39–69. https://doi.org/10.1017/asb.2015.25

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. ASTIN Bulletin, 52(1), 55–89. https://doi.org/10.1017/asb.2021.23

Liu, X., Jiang, M., Yu, Z., & Chen, C. (2022). Fairness in insurance underwriting: A review and research agenda. Insurance: Mathematics and Economics, 107, 25–43. https://doi.org/10.1016/j.insmatheco.2022.07.001

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y., Ivanov, L. A., Wang, F., Pisarenko, Z. V., & Ye, C. (2024a). Management analytics: A bibliometric analysis. Nanotechnologies in Construction, 16(3), 257–266. https://doi.org/10.15828/2075-8545-2024-16-3-257-266

Lu, Y., Pisarenko, Z. V., Yang, L., & Ye, C. (2024b). Advancing decision-making: The role of management analytics in modern business practices. Nanotechnologies in Construction, 16(5), 431–440. https://doi.org/10.15828/2075-8545-2024-16-5-431-440

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable. Journal of Open Source Software, 5(54), 2467. https://doi.org/10.21105/joss.02467

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, 116(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Noll, A., Salzmann, R., & Wüthrich, M. V. (2020). Case study: French motor third-party liability claims. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3164764

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data – XGBoost versus logistic regression. Risks, 7(2), 70. https://doi.org/10.3390/risks7020070

Quan, Z., & Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. Dependence Modeling, 6(1), 377–407. https://doi.org/10.1515/demo-2018-0022

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schelldorfer, J., & Wüthrich, M. V. (2019). Nesting classical actuarial models into neural networks. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3320525

Selten, F., Antonio, K., & Henckaerts, R. (2021). Spatial smoothing of insurance loss data using credibility theory. European Actuarial Journal, 11(2), 419–445. https://doi.org/10.1007/s13385-021-00270-5

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 180–186. https://doi.org/10.1145/3375627.3375830

So, B., Boucher, J.-P., & Valdez, E. A. (2021). Synthetic dataset generation of driver telematics. Risks, 9(4), 58. https://doi.org/10.3390/risks9040058

So, B., & Valdez, E. A. (2024). SAMME.C2 algorithm for imbalanced multi-class classification. Quantitative Finance and Economics, 8(3), 583–613. https://doi.org/10.3934/QFE.2024023

Spedicato, G. A., Dutang, C., & Petrini, L. (2018). Machine learning methods to perform pricing optimization. A comparison with standard GLMs. Variance, 12(1), 69–89. https://doi.org/10.2139/ssrn.3127236

Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency–severity model of insurance claims. PLOS ONE, 15(8), e0238000. https://doi.org/10.1371/journal.pone.0238000

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 3319–3328. https://doi.org/10.48550/arXiv.1703.01365

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B, 63(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis and Prevention, 98, 139–148. https://doi.org/10.1016/j.aap.2016.10.006

Verbelen, R., Antonio, K., & Claeskens, G. (2018). Unravelling the predictive power of telematics data in car insurance pricing. Journal of the Royal Statistical Society: Series C, 67(5), 1275–1304. https://doi.org/10.1111/rssc.12283

Werner, G., & Modlin, C. (2016). Basic ratemaking (5th ed.). Casualty Actuarial Society.

RESEARCH ARTICLE
Vol. 1, No. 1 (2023), pp. 57–75
DOI: 10.63646/jaiaa.2023.010104

Wood, S. N. (2017). Generalized additive models: An introduction with R (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Wüthrich, M. V. (2017). Covariate selection from telematics car driving data. European Actuarial Journal, 7(1), 89–108. https://doi.org/10.1007/s13385-017-0149-z

Wuthrich, M. V., & Buser, C. (2021). Data analytics for non-life insurance pricing. Swiss Finance Institute Research Paper, No. 16-68. https://doi.org/10.2139/ssrn.2870308

Xin, X., & Huang, F. (2024). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. North American Actuarial Journal, 28(2), 285–319. https://doi.org/10.1080/10920277.2023.2168506

Yang, Y., Qian, W., & Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. Journal of Business and Economic Statistics, 36(3), 456–470. https://doi.org/10.1080/07350015.2016.1200981

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, J., Wu, R., & Wang, Y. (2020). Insurance fraud detection with XGBoost combined with focal loss. International Journal of Computational Intelligence Systems, 13(1), 1145–1156. https://doi.org/10.2991/ijcis.d.200818.001

Kou, G., & Lu, Y. (2025). FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1–34. https://doi.org/10.1186/s40854-024-00668-6

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. Pattern Recognition, 36(2), 451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Soleymanian, M., Weinberg, C. B., & Zhu, T. (2019). Sensor data and behavioral tracking: Does usage-based auto insurance benefit drivers? Marketing Science, 38(1), 21–43. https://doi.org/10.1287/mksc.2018.1126

Gao, G., Wang, H., & Wüthrich, M. V. (2022). Boosting Poisson regression models with telematics car driving data. Machine Learning, 111(1), 243–272. https://doi.org/10.1007/s10994-021-05957-0

Reig Mullor, J., Garrido, J., & Lemus, F. (2020). Premium calculation and the calibration of statistical models in motor insurance. Risk Management, 22(2), 125–146. https://doi.org/10.1057/s41283-020-00060-5