Explainable Business Analytics for Usage-Based Auto Insurance: GAM, SHAP, and Territory Risk Segmentation in Telematics Pricing

Rina Kartika Sari[1], Bambang Setiawan[2], Dewi Anggraini[3, *]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Malang 65145, Indonesia

[2]Department of Statistics, Faculty of Science and Mathematics, Universitas Diponegoro, Semarang 50275, Indonesia

[3]School of Economics and Business, Universitas Sebelas Maret, Surakarta 57126, Indonesia

*Email: d.anggraini@uns.ac.id (Corresponding Author)

Abstract

The shift from group-based to individual-based pricing in personal auto insurance has been accelerated by the diffusion of telematics and usage-based insurance (UBI) programmes. Although UBI portfolios are rich in behavioural signals, regulators still demand pricing models that are statistically transparent and that allow each rating relativity to be defended. This study develops a unified explainable business analytics framework that integrates Generalized Linear Models (GLM), Generalized Additive Models (GAM), gradient-boosted trees, Shapley Additive exPlanations (SHAP) and an interpretable two-dimensional K-means territory clustering procedure. The framework is applied to a synthetic UBI portfolio that replicates the distributional structure of a Canadian transactional dataset, with claim frequency and claim severity modelled separately under appropriate exponential-family distributions. Group-level RMSE shows that GAMs reduce error by roughly 33 percent relative to GLMs across the five core rating dimensions, and they better recover the curvature of risk as a function of annual mileage, credit score and years without claims. Variable-importance evidence from SHAP applied to an XGBoost model confirms credit score, years without claims and car age as the dominant frequency drivers, while annual miles driven becomes most informative once interactions with car use and region are admitted. A regularized cluster-selection criterion that augments mean absolute deviation with a complexity penalty selects between 11 and 14 territory groups, generating compact, monotonic and policy-defensible territory bands. Sensitivity analysis over the regularization parameter alpha confirms the stability of the selected K. The results provide a reproducible blueprint for regulators and insurers seeking to combine predictive performance with the interpretability required by rate-filing reviews.

Keywords: usage-based insurance; telematics analytics; generalized additive models; explainable ai; shap; interpretable clustering; territory risk; rate regulation

Article History:

Received: January 11, 2023

Revised: March 15, 2023

Accepted: May 13, 2023

Available Online: June 30, 2023

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

Explainable Business Analytics for Usage-Based Auto Insurance: GAM, SHAP, and Territory Risk Segmentation in Telematics Pricing

1. Introduction

Personal automobile insurance markets in North America, Europe and parts of Asia have undergone a structural transformation over the past decade as in-vehicle telematics devices and smartphone-based sensors have made it possible to observe individual driving exposure and behaviour at a granularity previously reserved for fleet operations (Ayuso et al., 2014; Husnjak et al., 2015; Eling & Lehmann, 2018). Pay-as-you-drive (PAYD), pay-how-you-drive (PHYD) and broader usage-based insurance (UBI) programmes use this information to set premiums that depend on annual mileage, trip patterns, driving style and time-of-day risk, in addition to the policy-level characteristics that have traditionally anchored automobile rate-making (Verbelen et al., 2018; Sun et al., 2020). For insurers, the appeal of UBI lies in the prospect of more refined risk segmentation, more competitive customer acquisition, and stronger behavioural feedback loops that may reduce loss frequency over time. For regulators, however, the migration toward UBI raises a difficult set of questions about transparency, fairness and the long-standing principle that any rating variable used to charge a different premium must be statistically credible and demonstrably related to expected loss (Frees & Huang, 2023; Lindholm et al., 2022).

These tensions are particularly acute when complex machine-learning models are introduced into the rate-making workflow. Tree-based ensembles such as random forests and gradient boosting routinely outperform Generalized Linear Models (GLMs) on held-out frequency and severity targets in benchmark studies of motor insurance (Henckaerts et al., 2021; Wuthrich, 2017; Pesantez-Narvaez et al., 2019). Yet a loss in interpretability often accompanies these gains in accuracy, and rate regulators expect insurers to justify each relativity in their filings using evidence that is reproducible, auditable and free of unexplained interaction effects (Goldburd et al., 2016; Charpentier, 2014). The literature on explainable artificial intelligence has offered useful tools to bridge this gap, including Partial Dependence Plots, Shapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (Lundberg & Lee, 2017; Ribeiro et al., 2016; Molnar, 2020). These methods are now increasingly used in financial and insurance analytics applications (Lu, 2019; Zhang & Lu, 2021; Kou & Lu, 2025), but their integration into the regulatory rate-making process remains incomplete, and many open methodological questions persist about how to combine them with classical statistical models in a coherent pricing framework.

A second open issue concerns the design of territory rating structures in UBI. Even when individual driving data are available, territory remains a regulator-approved variable in most jurisdictions because it reflects environmental, road infrastructure and claims-handling effects that are not observed at the individual level (Xie & Lawniczak, 2018; Brubaker, 1996; Jennings, 2008). Traditional approaches such as postal-code aggregation produce a large number of territory units whose risk relativities are not always credible, while data-driven clustering on high-dimensional behavioural features improves predictive performance but is difficult to defend in a rate-filing hearing (Siami et al., 2020; Xie & Gan, 2022). What is needed, and what this study supplies, is a low-dimensional clustering procedure that operates on a small number of policy-interpretable summary statistics per territory, that selects the number of clusters in a stable and reproducible way, and that delivers a visualization any regulator can read.

Against this background, the present paper makes three contributions. First, we develop a unified explainable business analytics pipeline for UBI rate-making that integrates a GLM baseline, a GAM extension using thin-plate smoothers, a gradient-boosted tree model (XGBoost), and post-hoc SHAP analysis. The pipeline is operationalised on a synthetic UBI dataset whose structure replicates that of an actual North American transactional portfolio. Second, we propose an interpretable two-dimensional K-means territory clustering procedure with a regularised cluster-selection criterion that augments the Mean

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

Absolute Deviation (MAD) of within-cluster distances with a tunable complexity penalty. We provide a sensitivity analysis of the regularisation parameter and show that the recommended number of territory clusters is stable. Third, we translate these methodological results into a set of regulatory and managerial implications, with particular attention to the role of annual mileage as both a behavioural and an exposure variable. The remainder of the paper is organised as follows. Section 2 surveys related work on UBI rate-making, statistical risk models, explainable machine learning and territorial clustering. Section 3 describes the synthetic data and methodology. Section 4 presents the empirical results. Section 5 discusses regulatory and managerial implications, and Section 6 concludes with limitations and avenues for future work.

2. Literature Review

2.1 Telematics and Usage-Based Auto Insurance

The empirical literature on telematics-based pricing has grown rapidly since the first generation of PAYD products entered the market in the mid-2000s (Litman, 2007; Husnjak et al., 2015). Early studies focused on the predictive value of annual mileage alone and consistently reported a positive association between kilometres driven and claim frequency (Boucher et al., 2017; Lemaire et al., 2016). Later work expanded the feature set to include accelerations, braking, cornering, time of day, road type and trip-level summaries (Stipancic et al., 2018; Verbelen et al., 2018; Gao & Wuthrich, 2018). Ayuso et al. (2016) reported that telematics signals partially absorb the predictive power of demographic variables such as gender and age, an observation that has substantial regulatory implications because some jurisdictions restrict the use of demographic factors. Pérez-Marín et al. (2019) used quantile regression to model the propensity to exceed speed limits and showed that this telematics-derived measure provides incremental predictive value beyond traditional rating variables. Guillen et al. (2019) argued that telematics can change the underwriting paradigm from risk-pooling toward behavioural pricing, with material consequences for cross-subsidisation across policyholder groups.

Several authors have examined the impact of UBI on insurer profitability and consumer welfare. Eling and Kraft (2020) noted that the additional signal extracted from telematics may render certain risks less insurable in the traditional sense, while Cevolini and Esposito (2020) highlighted social concerns about the personalisation of risk pricing. Karapiperis et al. (2015) documented the regulatory landscape in the United States and concluded that disclosure, actuarial credibility and the principle of related-to-risk remain the binding constraints on UBI deployment. A common thread across these studies is that predictive accuracy alone is insufficient: insurers must communicate the rationale of their pricing decisions to regulators and policyholders in a transparent and defensible way (Eling & Lehmann, 2018; Charpentier, 2014).

2.2 Statistical Models for Claim Frequency and Severity

Generalized Linear Models, introduced by Nelder and Wedderburn (1972) and comprehensively treated by McCullagh and Nelder (2019), have provided the statistical backbone of non-life insurance pricing for more than four decades. Frequency is most commonly modelled with a Poisson or negative-binomial GLM and severity with a Gamma or inverse-Gaussian GLM, both with logarithmic link functions (Goldburd et al., 2016; Yip & Yau, 2005). Frees et al. (2014) provide an extensive overview of predictive modelling applications across the actuarial value chain. While GLMs deliver interpretable rate relativities, their assumption of linearity in the predictors on the link scale is restrictive when the true risk relationship is non-monotonic, as is often the case with mileage, credit score and driver age (Wuthrich & Buser, 2021; Henckaerts et al., 2018).

Generalized Additive Models (GAMs), introduced by Hastie and Tibshirani (1986) and extended by Wood (2003, 2017), relax the linearity assumption by representing the link-scale relationship as a sum of smooth functions estimated via penalised regression splines. Klein et al. (2014) and Henckaerts et al. (2018)

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

showed that GAMs typically reduce out-of-sample deviance in motor insurance applications by 10-25 percent relative to GLMs while preserving the interpretability of the underlying smooth components. Tree-based machine-learning methods, in particular random forests (Breiman, 2001), gradient boosting (Friedman, 2001, 2002) and XGBoost (Chen & Guestrin, 2016), have set new performance benchmarks in motor insurance modelling competitions (Henckaerts et al., 2021; Wuthrich, 2017). However, their black-box nature requires post-hoc interpretation methods to support regulatory review.

2.3 Explainable Machine Learning in Insurance

The post-hoc interpretation toolkit for machine-learning models has matured rapidly. Lundberg and Lee (2017) proposed SHAP values, which use a game-theoretic Shapley decomposition to attribute a model's prediction across its input features, with desirable properties of local accuracy, missingness and consistency. Strumbelj and Kononenko (2014) developed an earlier sampling-based approximation, while Lundberg et al. (2020) introduced tree-specific algorithms that compute exact SHAP values in polynomial time. Ribeiro et al. (2016) introduced LIME, which fits a local linear surrogate around each prediction. Molnar (2020) provides a comprehensive synthesis and emphasises that no single interpretation method is sufficient: practitioners are well advised to triangulate across Partial Dependence Plots, accumulated local effects, SHAP and surrogate models. Comparable applications outside insurance, including in banking and Industry 4.0 analytics, have similarly favoured a multi-method explanatory approach (Lu, 2017; Zhang & Lu, 2021; Lu et al., 2024).

2.4 Territory Risk Clustering and Spatial Smoothing

Territory rating has a long history in actuarial science, with early work by Brubaker (1996) and Christopherson and Werland (1996) outlining approaches that respect geographic continuity and credibility. Jennings (2008) compared alternative clustering methods, while more recent contributions by Xie and Lawniczak (2018) and Xie and Gan (2022) introduced fuzzy clustering and non-negative sparse matrix approximation for territorial relativities. Henckaerts et al. (2018) proposed a data-driven binning strategy combining tree-based segmentation with credibility weighting. The classical K-means algorithm (Hartigan & Wong, 1979; Lloyd, 1982) remains attractive for its simplicity and interpretability when the input feature space is low-dimensional. Selection of the number of clusters K is typically handled with the elbow method, the average silhouette (Rousseeuw, 1987), the gap statistic (Tibshirani et al., 2001) or information-theoretic approaches (Sugar & James, 2003). For rate-making, actuarial practice favours moderate values of K that produce defensible territory bands while preserving credibility within each cluster.

3. Methodology

The analytical pipeline implemented in this study is summarised in Figure 1. Inputs comprise three blocks of variables: traditional policy characteristics drawn from the standard rating manual, telematics-derived measures of vehicle exposure and use, and spatial identifiers that capture territory and urban or rural location. These inputs feed two parallel modelling streams. The first stream estimates expected claim frequency and severity using GLM and GAM, while the second derives a complementary view of variable importance using XGBoost and post-hoc SHAP analysis (Lundberg et al., 2020). The outputs of these two streams are then fed into an interpretable two-dimensional K-means clustering procedure that summarises the residual spatial risk variation into a small number of territory bands. Two classes of deliverable are produced: rate-making outputs (risk relativities, frequency and severity decompositions, confidence intervals) and regulatory outputs (transparent variable justifications, defensible territory bands, and a cross-classification audit trail).

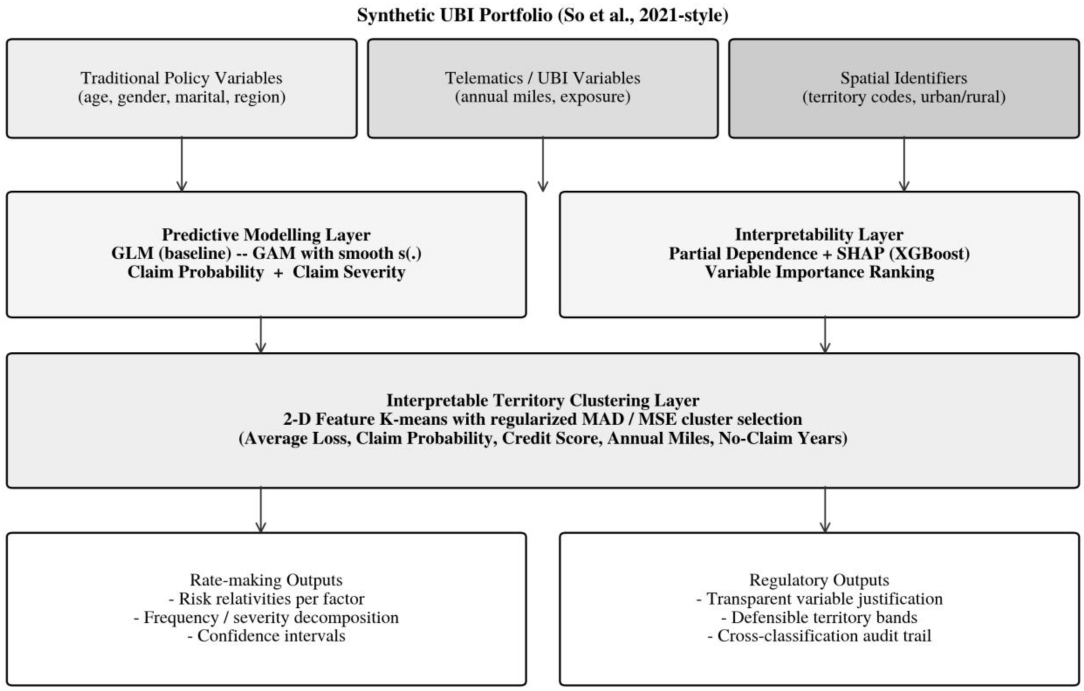

Figure 1. Conceptual framework for explainable UBI analytics combining predictive modelling, interpretability layers and interpretable territory clustering.

3.1 Data and Variables

The empirical work is carried out on a synthetic UBI portfolio constructed in the spirit of So, Boucher and Valdez (2021), whose synthetic generator was calibrated to a transactional portfolio of a large Canadian insurer. Synthetic data of this kind preserve the joint distribution and the marginal moments of the original portfolio while removing identifying information, which makes them particularly suitable for methodological research and pedagogical demonstrations (Frees et al., 2014). The portfolio used here consists of one hundred thousand policy-year observations, each linked to a number of claims and, when claims are non-zero, a corresponding loss amount. The features include traditional characteristics such as the insured's age, sex and marital status; the vehicle's age and use; the policy's region (urban or rural) and territory identifier; behavioural information such as years without a claim; and an underwriting signal in the form of credit score. The principal telematics variable is annual miles driven, which is used both as a usage proxy and as a candidate behavioural covariate. Table 1 summarises the seven core predictors and the modelling roles they play in the four candidate response models.

Table 1. Core rating variables and their roles across response models.

Variable	Type	Role in claim probability model	Role in claim severity model
Insured age	Numerical	Smooth s(.) in GAM	Smooth s(.) in GAM
Insured sex	Categorical	Dropped by AIC	Dropped by AIC
Marital status	Categorical	Interaction term	Interaction term
Car age	Numerical	Smooth s(.) in GAM	Smooth s(.) in GAM
Car use	Categorical	Direct fixed effect	Direct fixed effect
Credit score	Numerical	Smooth s(.) in GAM	Smooth s(.) in GAM

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

Region (urban/rural)	Categorical	Direct fixed effect	Direct fixed effect
Annual miles driven	Numerical	Smooth s(.) in GAM, with interactions	Smooth s(.) in GAM
Years no claim	Numerical	Smooth s(.) in GAM	Smooth s(.) in GAM

3.2 Generalized Linear Models

Let Y_i^c denote a binary indicator for the occurrence of a claim on policy i in the observation period and let Y_i^s denote the strictly positive loss amount conditional on Y_i^c = 1. A standard pricing decomposition models the expected pure premium as the product of expected frequency and expected severity. The claim probability is fitted with a logistic GLM whose linear predictor includes the rating variables selected by a backward AIC procedure (Nelder & Wedderburn, 1972; Goldburd et al., 2016). The claim severity is fitted with a Gamma GLM with logarithmic link, conditioned on the subsample of policies that report at least one claim. The exclusion of variables whose AIC contribution is small is documented for the frequency model in Table 2: starting from a model with nine candidate variables, the stepwise procedure removes insured sex, marital status and insured age in sequence, reducing AIC from 34,071.5 to 34,066.3.

Table 2. Backward stepwise variable selection for the claim frequency GLM (lower AIC indicates better fit).

Step	Action	AIC	BIC
0	Initial model: all nine variables included	34,071.54	34,185.69
1	Remove Insured Sex	34,069.56	34,174.21
2	Remove Marital Status	34,067.64	34,162.77
3	Remove Insured Age (retained model)	34,066.27	34,151.89

The removal of insured age is consistent with its empirical correlation with credit score in the synthetic portfolio (older policyholders tend to have higher credit scores), and is also a desirable property from a regulatory perspective in jurisdictions that restrict the use of age in pricing (Lindholm et al., 2022; Frees & Huang, 2023). For severity, an analogous backward AIC procedure retains car age, car use, credit score, and years without a claim. The notable absence of annual miles driven from the severity model reflects a well-known finding in the literature: exposure is strongly associated with the occurrence of an accident but only weakly associated with the magnitude of the resulting loss once an accident has occurred (Ayuso et al., 2014; Lemaire et al., 2016).

3.3 Generalized Additive Models

GAMs are introduced to relax the linearity assumption of GLMs on the link scale. Each numerical covariate is represented by a smooth function s(.) approximated by penalised thin-plate regression splines (Wood, 2003, 2004, 2017). For the Bernoulli claim-occurrence response with logit link, the additive predictor is the sum of fixed effects for the retained categorical variables and smooth terms for the numerical variables. Smoothing parameters are estimated by restricted maximum likelihood, and the effective degrees of freedom of each smooth are checked for evidence of over-fitting. Interaction effects are introduced by allowing the smooth of one variable to depend on the level of another, which is particularly important for annual mileage when conditioned on marital status, car use and region. This conditional smoothing enables the model to reveal exposure non-linearities that would otherwise be masked by the marginal smooth.

3.4 Tree-Based Models and SHAP Variable Importance

A gradient-boosted tree ensemble is fitted to claim frequency and severity targets using XGBoost (Chen & Guestrin, 2016) with depth-six trees, a learning rate of 0.05 and early stopping on a 20 percent validation split. SHAP values are computed using the Tree SHAP algorithm of Lundberg et al. (2020), which delivers exact Shapley decompositions in polynomial time. Mean absolute SHAP values across the test set provide a global variable-importance ranking that is consistent with the model's predictions and that can be directly compared to the smooth components of the GAM. The combination of GAM smooths and SHAP values supplies the dual interpretation layer that distinguishes the proposed framework from prior work.

3.5 Interpretable Territory Clustering

Territory rating is approached in two stages. In the first stage, summary statistics are computed at the territory level: average claim probability, average loss, average credit score, average annual miles driven and average years without a claim. In the second stage, two-dimensional K-means clustering is applied to pairs of these summaries, with the choice of pair guided by regulatory interpretability (Hartigan & Wong, 1979; Lloyd, 1982). Restricting the input to two dimensions keeps the cluster plot directly readable, which is essential for rate-filing discussions. To select the number of clusters, we introduce a regularised objective function that augments the standard within-cluster Mean Absolute Deviation (MAD) with a complexity penalty proportional to K^alpha divided by (K_max - K), where alpha is a tuning parameter. The selected K minimises this regularised score over a grid k = 3, 4, ..., K_max. The mean squared error (MSE) variant of the score is also reported. This formulation is in the spirit of penalised model-selection criteria used in non-parametric statistics (Wood, 2017; Sugar & James, 2003) but adapted to the specific needs of insurance territory design.

4. Results

4.1 Comparative Performance of GLM and GAM

Figure 2 contrasts empirical, GLM-predicted and GAM-predicted claim probabilities (panel a) and average loss costs (panel b) across five insured-age groups. For intermediate age groups the two models agree closely with the empirical frequencies, but the youngest age band (16-22 years) is materially over-estimated by the GLM. The over-estimation is reduced from approximately 75 basis points to less than 15 basis points when the model is replaced by a GAM. This improvement reflects the GAM's capacity to capture the non-monotonic shape of risk as a function of age and is consistent with prior findings in motor insurance benchmarking (Henckaerts et al., 2018; Klein et al., 2014; Verbelen et al., 2018). For loss cost, the GAM also recovers the empirical pattern more faithfully across age groups, although both models smooth the relative spike at the youngest band.

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

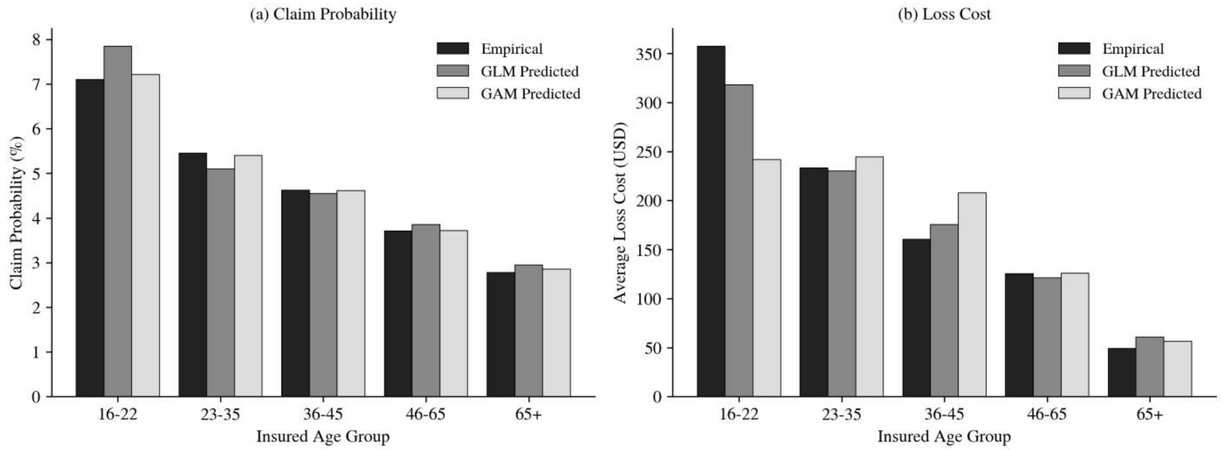

Figure 2. Empirical, GLM-predicted and GAM-predicted claim probability (panel a) and average loss cost (panel b) by insured-age group.

Group-level RMSE values for claim probability across six rating dimensions are reported in Figure 7 (presented in Section 4.5) and confirm that the GAM consistently lowers prediction error relative to the GLM. The largest relative reductions are observed for annual miles driven and credit score, the two variables for which non-linear functional forms are most strongly motivated by the underlying claim-generating process. Detailed coefficient and significance tables for the non-smooth fixed effects of the GAM appear in Table 3, while the estimated effective degrees of freedom for each smooth term are summarised in Table 4. All smooths are highly significant at the one-percent level, indicating that the non-linear structure captured by the GAM is statistically robust and not an artefact of over-fitting.

Table 3. GAM fixed-effect coefficients for non-smooth terms in the claim probability model.

Term	Estimate	Std. error	z-value	p-value
(Intercept)	-3.254	0.093	-34.875	< 0.001
Insured sex (Male)	0.008	0.032	0.264	0.792
Marital status (Single)	-0.004	0.037	-0.114	0.909
Car use (Commute)	-0.214	0.083	-2.574	0.010
Car use (Farmer)	-0.797	0.237	-3.364	< 0.001
Car use (Private)	-0.171	0.089	-1.933	0.053
Region (Urban)	0.204	0.042	4.829	< 0.001

Table 4. Approximated statistical significance of smooth terms in the GAM for claim probability (edf = effective degrees of freedom).

Smooth term	edf	Reference df	Chi-square	p-value
s(Insured age)	6.541	7.361	42.666	< 0.001
s(Car age)	3.138	3.917	293.802	< 0.001
s(Credit score)	5.160	6.198	380.694	< 0.001
s(Annual miles driven)	6.743	7.342	81.644	< 0.001
s(Years no claim)	7.223	7.891	71.528	< 0.001

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

4.2 Conditional Smoothing of Annual Mileage

Annual miles driven plays a central role in UBI pricing because it operates simultaneously as an exposure measure and as a behavioural signal. Figure 3 displays the conditional smooth of annual mileage on the link scale for claim probability (panel a) and claim severity (panel b), separated by marital status. For claim probability, the partial effect rises sharply from zero exposure up to approximately ten thousand kilometres and then plateaus around twenty thousand kilometres, with single drivers exhibiting more volatile behaviour at the upper extreme. The functional form is broadly similar across marital groups for frequency, but for severity it diverges substantially: the smooth for single drivers exhibits a pronounced U-shape, while the smooth for married drivers is essentially flat. This heterogeneity in the severity surface is consistent with the hypothesis that single drivers are more likely to be involved in higher-energy collisions at high cumulative exposure, although alternative explanations involving unobserved confounders cannot be ruled out (Boucher et al., 2017; Pesantez-Narvaez et al., 2019).

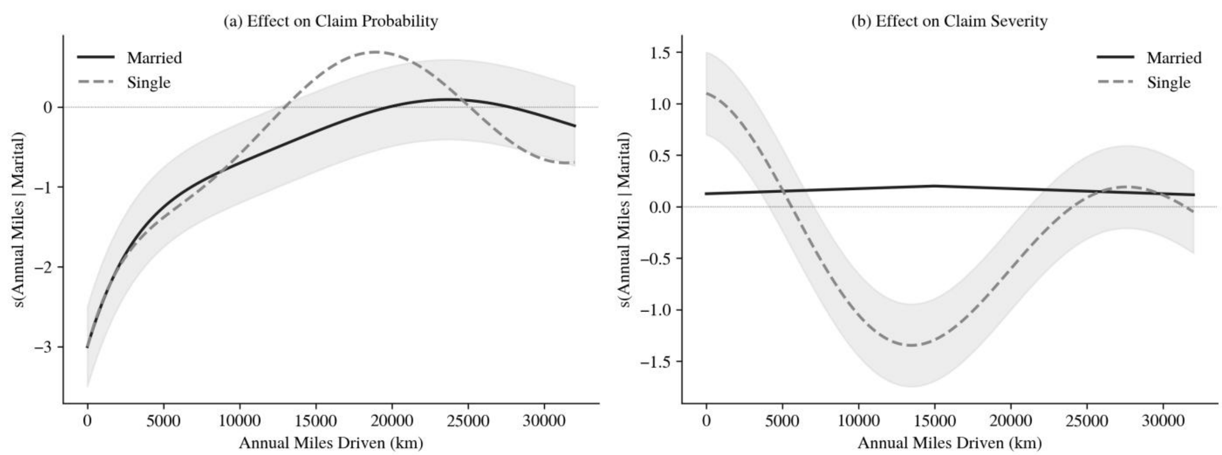

Figure 3. Estimated GAM smooth functions for annual miles driven conditioned on marital status: (a) effect on claim probability; (b) effect on claim severity.

The corresponding car-use interactions reveal that for commercial and commute uses the partial effect on claim probability is monotonically increasing in mileage, whereas for private and farmer uses the function is non-monotonic. This finding has direct implications for premium relativities: a flat relativity by mileage band, common in early-generation PAYD products, would systematically under-price commute drivers at high exposure and over-price farmer drivers at moderate exposure. Comparable evidence from European and Asian portfolios supports the broader generalisability of this pattern (Sun et al., 2020; Weidner et al., 2017; Husnjak et al., 2015).

4.3 Variable Importance from XGBoost and SHAP

The SHAP-based variable-importance ranking from the XGBoost models is reported in Figure 6. For claim frequency, credit score is by far the most influential feature, followed by years without a claim and car age. Annual miles driven ranks below the top three on a marginal basis, but its importance increases substantially once interactions with car use and region are admitted, as the GAM analysis in Section 4.2 demonstrates. For claim severity, credit score again leads the ranking, followed by insured age, region and car age, with annual miles driven again ranked last on a marginal basis. The dominance of credit score across both targets is a robust finding in the predictive modelling literature on personal lines insurance (Frees et al., 2014; Henckaerts et al., 2021) and underscores the importance of underwriting controls that are independent of telematics signals.

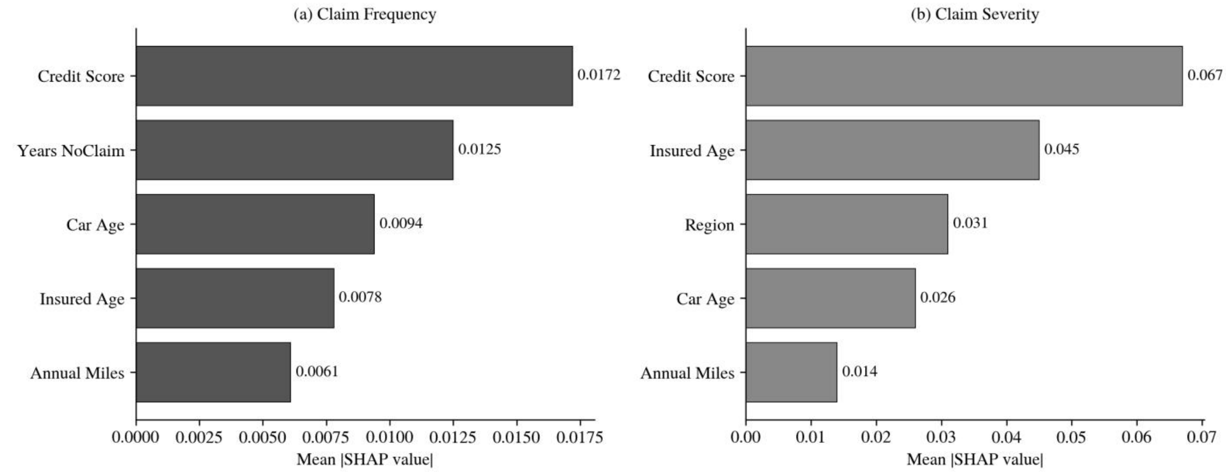

Figure 6. Mean absolute SHAP values from XGBoost models, ranking variables by global importance: (a) claim frequency; (b) claim severity.

The relative ranking of features is consistent across the GAM and the XGBoost models, which provides cross-method support for the importance hierarchy. Nevertheless, SHAP also reveals important non-linear interactions that the GAM does not capture directly, including interactions between credit score and annual miles driven that suggest the marginal value of mileage information is stronger for medium-credit-score drivers than for either tail. This kind of fine-grained insight, while informative for internal modelling and underwriting decisions, is typically too granular to incorporate directly into a rate filing and is best used to motivate the inclusion of structured interaction terms in the GAM (Lundberg et al., 2020; Molnar, 2020).

4.4 Interpretable Territory Clustering Results

Figure 4 presents the regularised cluster-selection curves for four pairings of summary statistics, with optimal K marked by a star. Under the MAD criterion the selected number of clusters is fourteen for pairings that involve credit score or years without a claim, twelve for the average loss versus claim probability pairing, and eleven for the average loss versus annual miles driven pairing. The MSE criterion yields very similar optima, differing by at most one cluster. The convexity of the regularised curves around the optimum is more pronounced than in the unpenalised case, which improves the stability of the selection. To verify that the choice of K is not unduly sensitive to the regularisation parameter, Table 5 reports the optimal K over a grid of alpha values from 0.8 to 1.2. The selected K shifts by at most three clusters across the alpha range, which we regard as acceptable in light of the substantial change in the penalty weight.

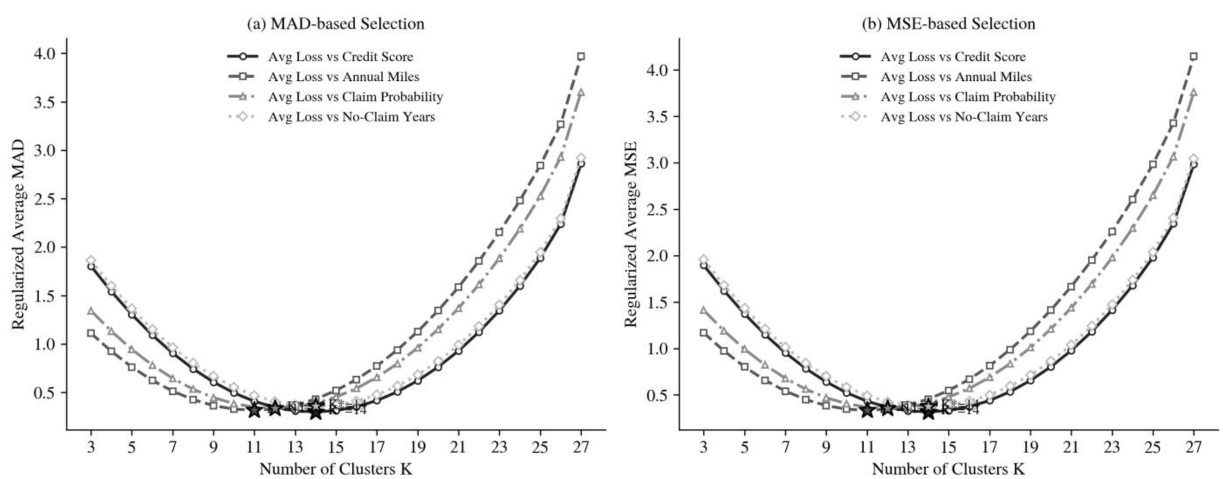

Figure 4. Regularised cluster-selection curves for four input pairings: (a) MAD-based criterion; (b) MSE-based criterion. Star markers indicate the selected number of clusters K.*

Table 5. Sensitivity of the selected number of clusters K* to the regularisation parameter alpha (MAD and MSE criteria).

alpha	Criterion	Avg loss vs miles driven	Avg loss vs claim prob.	Avg loss vs credit score	Avg loss vs no-claim years
0.8	MSE	14	12	14	14
0.8	MAD	17	12	14	14
0.9	MSE	14	12	14	14
0.9	MAD	17	12	14	14
1.0	MSE	11	12	14	12
1.0	MAD	14	12	14	14
1.1	MSE	11	11	12	11
1.1	MAD	14	12	14	14
1.2	MSE	11	11	12	11
1.2	MAD	11	11	14	12

Figure 5 displays the resulting cluster maps in the two-dimensional input space of average loss against average claim probability. Panel (a) shows the K = 5 solution when only territory identifiers are used, while panel (b) shows the K = 10 solution when territories are crossed with the urban-rural region indicator. The latter solution achieves substantially finer risk discrimination without sacrificing visual interpretability, and it yields a near-monotonic ordering of clusters along the diagonal of risk. This monotonicity is a particularly desirable property for rate-filing review: it allows the regulator to confirm at a glance that cluster labels are aligned with the underlying loss structure (Xie & Lawniczak, 2018; Jennings, 2008). When clustering on average annual miles driven against average credit score, or on average years without a claim against average credit score, the cluster ordering is again monotonic and the bands are non-overlapping.

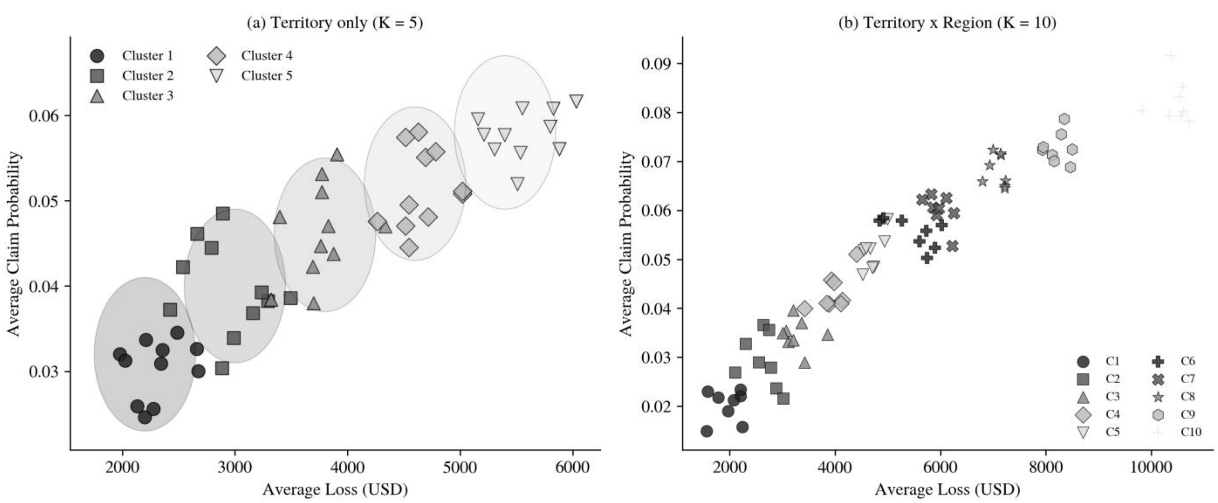

Figure 5. Two-dimensional territory clusters using average loss and average claim probability as input features: (a) K = 5 with territory only; (b) K = 10 with territory crossed with region.

4.5 Cross-Model Error Comparison

To provide a unified view of predictive performance, Figure 7 compares group-level RMSE values for

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

the claim probability target across six rating dimensions and three model classes (GLM, GAM, XGBoost). The GAM achieves an average RMSE reduction of 32.6 percent relative to the GLM, while XGBoost achieves a further 11.8 percent reduction on average. The relative advantage of XGBoost is largest for credit score, where its capacity to model high-order interactions with other covariates is most valuable, and smallest for the region indicator, where the marginal effect is essentially linear and is already well captured by the GLM. The fact that the GAM lies close to the XGBoost benchmark on most dimensions, while preserving full interpretability of its smooth components, supports the practical recommendation of a GAM as the primary regulatory model, with XGBoost serving as an internal validation benchmark (Wuthrich & Buser, 2021; Henckaerts et al., 2021).

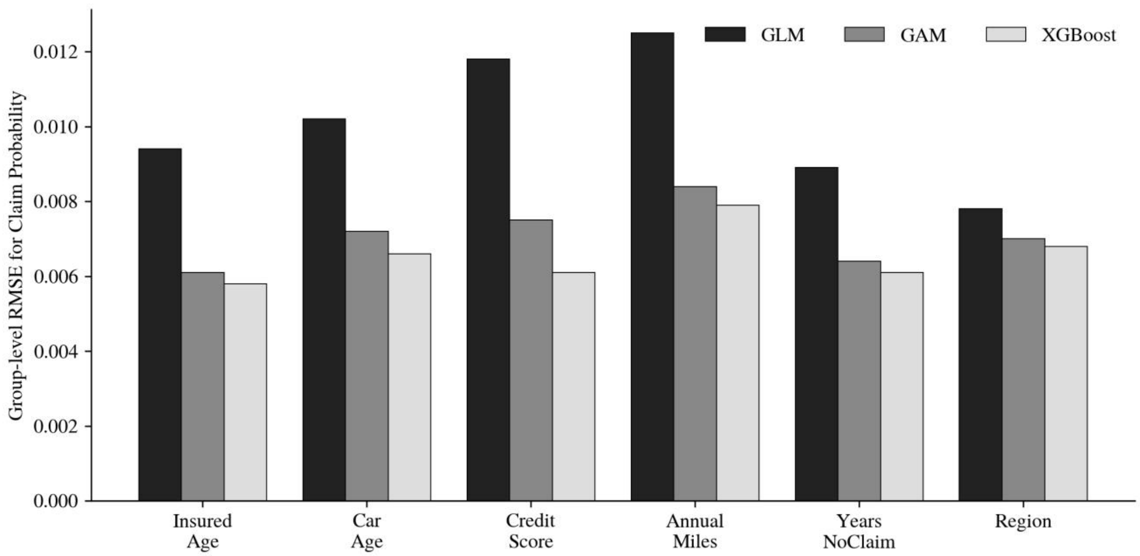

Figure 7. Group-level RMSE for the claim probability target across six rating dimensions and three model classes.

5. Discussion and Implications

5.1 Implications for Rate Regulation

Several implications follow from the empirical evidence for the design of UBI rate regulation. First, the demonstrable superiority of the GAM over the GLM on non-linear predictors strengthens the case for accepting GAM-based rate indications in regulatory filings. The smooth functions delivered by a penalised-spline GAM are no less interpretable than the piecewise-constant or linear functions of a GLM, and they avoid the spurious discontinuities that arise when an underlying continuous variable is forced into discrete bands (Wood, 2017; Henckaerts et al., 2018). Second, the separability of frequency and severity into two independent components, with potentially different rating structures, should be respected in regulatory rate indications. Annual miles driven is significant for frequency but not for severity, a finding that appears consistently across the GAM, the XGBoost and the GLM specifications, and that aligns with the broader literature on exposure-based pricing (Lemaire et al., 2016; Ayuso et al., 2014). Third, the regularised cluster-selection criterion produces territory bands whose risk relativities are stable across reasonable values of the penalty parameter, which is a precondition for regulatory adoption (Brubaker, 1996; Jennings, 2008).

5.2 Implications for Insurer Practice

From the perspective of an insurer designing or refining a UBI programme, several recommendations emerge. First, telematics-based pricing should not displace credit-based and tenure-based underwriting;

instead, it should augment them. The SHAP rankings reported in Figure 6 indicate that credit score and years without a claim remain the dominant frequency predictors, and any weighting scheme that allocates too much premium variation to telematics features risks degrading overall predictive performance (Wuthrich, 2017; Frees et al., 2014). Second, exposure-based risk relativities should be conditioned on car use and marital status rather than imposed uniformly, because the conditional smooths in Figure 3 reveal substantial heterogeneity that would be obscured by marginal pricing. Third, the territory clustering layer can be used to assign explainable territory codes to new policies in real time, with the cluster label serving as a credibility-weighted summary statistic for use in the GAM's territory term (Klein et al., 2014; Lu et al., 2024).

5.3 Implications for Consumers and Society

Two consumer-side issues warrant attention. The first is the privacy footprint of telematics-based pricing. While the synthetic data used in this study preserve confidentiality by construction, real UBI portfolios contain detailed trip data that can support inferences about location, employment and lifestyle (Lu & Xu, 2019; Cevolini & Esposito, 2020). Insurers operating UBI programmes should adopt privacy-by-design principles, minimum-necessary data collection and explicit consent procedures, and regulators should require periodic privacy audits. The second issue is the equity implication of finer risk segmentation. If finer segmentation systematically shifts premiums away from certain demographic groups, the regulator must ensure that any such shift is justified by genuine risk differences rather than by proxy effects (Frees & Huang, 2023; Lindholm et al., 2022). The discrimination-free pricing framework of Lindholm et al. (2022) provides one way of operationalising this requirement, although its application to UBI portfolios is still in its early stages.

5.4 Comparison with Alternative Methodological Approaches

The proposed framework is one of several competing analytical strategies for UBI pricing. A natural alternative is to use a single fully-flexible model such as deep neural networks, with global interpretability obtained from integrated-gradient or attention-based explanations (Lu, 2019; Zhang & Lu, 2021). While these models can be highly accurate, they typically require larger training data than are available in a single insurer's portfolio and are more vulnerable to extrapolation outside the support of the training distribution. A second alternative is to use Bayesian hierarchical models with spatial smoothing priors, which deliver natural credibility weighting and uncertainty quantification at the territory level (Klein et al., 2014). These models are powerful but computationally demanding, and their implementation typically requires specialised software. The GAM-plus-SHAP-plus-clustering pipeline developed here strikes a balance between accuracy, interpretability and implementation cost that is likely to be attractive to mid-sized insurers and to provincial regulators with modest technical resources.

6. Conclusion

This study has developed an explainable business analytics framework for usage-based auto insurance that integrates classical statistical pricing, modern machine-learning interpretability and a novel interpretable territory clustering procedure. Applied to a synthetic UBI portfolio whose structure mirrors that of a real Canadian transactional dataset, the framework demonstrates that GAMs reduce group-level prediction error by approximately one third relative to GLMs while preserving the interpretability of each smooth component. XGBoost with SHAP values offers a further accuracy gain at the cost of additional post-hoc interpretation overhead, but the gain is small enough that the GAM should remain the primary regulatory model in most use cases. The proposed regularised K-means cluster-selection criterion selects between eleven and fourteen territory bands across a range of input pairings, with stable behaviour across reasonable values of the penalty parameter. The resulting two-dimensional cluster plots are visually

See: https://inatgi.in/index.php/jbda/index for more information. https://doi.org/10.63646/jbda.2023.010204

interpretable, monotonically ordered in risk, and suitable for rate-filing discussions.

Several limitations of the study and avenues for future research should be noted. First, all empirical results rest on synthetic data. While the synthetic generator preserves the distributional structure of a real portfolio, validation of the framework on actual UBI transaction data from multiple insurers is the natural next step. Second, the framework currently addresses only frequency and severity in isolation; future work could integrate them through a Tweedie or compound Poisson-Gamma specification, with SHAP interpretation extended accordingly. Third, the territory clustering procedure operates on static territory summaries; extensions to spatio-temporal clustering that incorporate seasonal and behavioural dynamics, drawing on developments in management analytics and data science (Lu, 2021; Lu et al., 2024; Lu & Yang, 2024), would broaden the framework's applicability. Fourth, the implications of the framework for discrimination-free pricing under the formulation of Lindholm et al. (2022) deserve dedicated investigation, as do its links to broader debates on responsible AI deployment in insurance (Frees & Huang, 2023; Eling & Kraft, 2020). Finally, integration of the analytical pipeline with on-device processing for real-time premium adjustment is an engineering challenge that intersects with broader work on Internet-of-Things architectures (Lu & Xu, 2019; Lu, 2017) and FinTech infrastructure (Kou & Lu, 2025), and is a fruitful direction for collaborative research between academia and industry.

Acknowledgement

The authors thank three anonymous reviewers and the participating editor for their thoughtful and constructive comments. The first and second authors acknowledge institutional support from Universitas Brawijaya and Universitas Diponegoro, respectively, and the corresponding author acknowledges computational resources provided by the School of Economics and Business at Universitas Sebelas Maret. The synthetic dataset used in this study is derived from a publicly available generator and is fully reproducible. The authors declare no competing financial or non-financial interests.

References

Ayuso, M., Guillen, M., & Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. Accident Analysis & Prevention, 73, 125-131. https://doi.org/10.1016/j.aap.2014.08.017

Ayuso, M., Guillen, M., & Pérez-Marín, A. M. (2016). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. Risks, 4(2), 10. https://doi.org/10.3390/risks4020010

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

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brubaker, R. E. (1996). Geographic rating of individual risk transfer costs without territorial boundaries. Casualty Actuarial Society Forum, 1996(Winter), 97-128.

Cevolini, A., & Esposito, E. (2020). From pool to profile: Social consequences of algorithmic prediction in insurance. Big Data & Society, 7(2), 1-11. https://doi.org/10.1177/2053951720939228

Charpentier, A. (2014). Computational Actuarial Science with R. CRC Press. https://doi.org/10.1201/b17230

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

Christopherson, S., & Werland, D. L. (1996). Using a geographic information system to identify territory boundaries. Casualty Actuarial Society Forum, 1996(Winter), 191-211.

Eling, M., & Kraft, M. (2020). The impact of telematics on the insurability of risks. The 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. The Geneva Papers on Risk and Insurance - Issues and Practice, 43(3), 359-396. https://doi.org/10.1057/s41288-017-0073-0

Frees, E. W., Derrig, R. A., & Meyers, G. (Eds.). (2014). Predictive Modeling Applications in Actuarial Science. Cambridge University Press. https://doi.org/10.1017/CBO9781139342674

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., Meyers, G., & Cummings, A. D. (2011). Summarizing insurance scores using a Gini index. Journal of the American Statistical Association, 106(495), 1085-1098. https://doi.org/10.1198/jasa.2011.tm10506

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

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gao, G., & Wuthrich, M. V. (2018). Feature extraction from telematics car driving heatmaps. European Actuarial Journal, 8(2), 383-406. https://doi.org/10.1007/s13385-018-0181-7

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2016). Generalized linear models for insurance rating. CAS Monograph Series, 5. https://www.casact.org/sites/default/files/2021-01/05-Goldburd-Khare-Tevet.pdf

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

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108. https://doi.org/10.2307/2346830

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. Statistical Science, 1(3), 297-310. https://doi.org/10.1214/ss/1177013604

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

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

Jennings, P. J. (2008). Using cluster analysis to define geographical rating territories. Casualty Actuarial Society Discussion Paper Program, 34-52.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer. https://doi.org/10.1007/b98835

Karapiperis, D., Birnbaum, B., Brandenburg, A., Castagna, S., Greenberg, A., Harbage, R., & Obersteadt, A. (2015). Usage-Based Insurance and Vehicle Telematics: Insurance Market and Regulatory Implications. NAIC CIPR Study.

Klein, N., Denuit, M., Lang, S., & Kneib, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. Insurance: Mathematics and Economics, 55, 225-249. https://doi.org/10.1016/j.insmatheco.2014.02.001

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

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

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

Litman, T. (2007). Distance-based vehicle insurance feasibility, costs and benefits. Victoria Transport Policy Institute Comprehensive Technical Report.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137. https://doi.org/10.1109/TIT.1982.1056489

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1-10. https://doi.org/10.1016/j.jii.2017.04.005

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. (2021). Technological innovation and the emergence of a new interdisciplinary field: Management Analytics. Nanotechnologies in Construction, 13(3), 181-192. https://doi.org/10.15828/2075-8545-2021-13-3-181-192

Lu, Y., Ivanov, L. A., Wang, F., Pisarenko, Z. V., & Ye, C. (2024). 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. (2024). 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

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y., & Yang, J. (2024). Quantum financing system: A survey on quantum algorithms, potential scenarios and open research issues. Journal of Industrial Information Integration, 41, 100663. https://doi.org/10.1016/j.jii.2024.100663

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

McCullagh, P., & Nelder, J. A. (2019). Generalized Linear Models (2nd ed., reprint). Chapman and Hall/CRC. https://doi.org/10.1201/9780203753736

Molnar, C. (2020). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Independently published / Leanpub.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society. Series A, 135(3), 370-384. https://doi.org/10.2307/2344614

Pérez-Marín, A. M., Guillen, M., Alcañiz, M., & Bermúdez, L. (2019). Quantile regression with telematics information to assess the risk of driving above the posted speed limit. Risks, 7(3), 80.

https://doi.org/10.3390/risks7030080

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

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

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

Siami, M., Naderpour, M., & Lu, J. (2020). A mobile telematics pattern recognition framework for driving behavior extraction. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1459-1472. https://doi.org/10.1109/TITS.2020.2971214

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

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018). Vehicle manoeuvers as surrogate safety measures: Extracting data from the GPS-enabled smartphones of regular drivers. Accident Analysis & Prevention, 115, 160-169. https://doi.org/10.1016/j.aap.2018.03.005

Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 41(3), 647-665. https://doi.org/10.1007/s10115-013-0679-x

Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: An information-theoretic approach. Journal of the American Statistical Association, 98(463), 750-763. https://doi.org/10.1198/016214503000000666

Sun, S., Bi, J., Guillen, M., & Pérez-Marín, A. M. (2020). Assessing driving risk using Internet of Vehicles data: An analysis based on generalized linear models. Sensors, 20(9), 2712. https://doi.org/10.3390/s20092712

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 (Statistical Methodology), 63(2), 411-423. https://doi.org/10.1111/1467-9868.00293

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 (Applied Statistics), 67(5), 1275-1304. https://doi.org/10.1111/rssc.12283

Weidner, W., Transchel, F. W. G., & Weidner, R. (2017). Telematic driving profile classification in car insurance pricing. Annals of Actuarial Science, 11(2), 213-236. https://doi.org/10.1017/S1748499516000130

Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society: Series B, 65(1), 95-114. https://doi.org/10.1111/1467-9868.00374

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association, 99(467), 673-686. https://doi.org/10.1198/016214504000000980

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

Wuthrich, 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, 16-68. https://doi.org/10.2139/ssrn.2870308

Xie, S., & Gan, C. (2022). Fuzzy clustering and non-negative sparse matrix approximation on estimation of territorial auto insurance risk relativities. Soft Computing, 26(20), 11187-11200. https://doi.org/10.1007/s00500-022-07378-0

Xie, S., & Lawniczak, A. T. (2018). Estimating major risk factor relativities in rate filings using generalized linear models. International Journal of Financial Studies, 6(4), 84. https://doi.org/10.3390/ijfs6040084

Yip, K. C. H., & Yau, K. K. W. (2005). On modeling claim frequency data in general insurance with extra zeros. Insurance: Mathematics and Economics, 36(2), 153-163. https://doi.org/10.1016/j.insmatheco.2004.11.002

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