British Actuarial Journal (2026), Vol. **31**, e5, pp. 1–36
doi:10.1017/S1357321725100317

Institute and Faculty of Actuaries

CONTRIBUTED PAPER

Dual evaluation of performance and fairness from machine learning models for non-life insurance pricing

Tarun Israni[1], Linda Daly[1], John Condon[1] and Eric Wolsztynski[1,2]

[1]School of Mathematical Sciences, University College Cork, Cork, Ireland and [2]Insight Research Ireland Centre for Data Analytics, University College Cork, Cork, Ireland
Corresponding author: Tarun Israni; Email: tisrani@gmail.com

Abstract
An increasing number of reports highlight the potential of machine learning (ML) methodologies over the conventional generalised linear model (GLM) for non-life insurance pricing. In parallel, national and international regulatory institutions are accentuating their focus on pricing fairness to quantify and mitigate algorithmic differences and discrimination. However, comprehensive studies that assess both pricing accuracy and fairness remain scarce. We propose a benchmark of the GLM against mainstream regularised linear models and tree-based ensemble models under two popular distribution modelling strategies (Poisson-gamma and Tweedie), with respect to key criteria including estimation bias, deviance, risk differentiation, competitiveness, loss ratios, discrimination and fairness. Pricing performance and fairness were assessed simultaneously on the same samples of premium estimates for GLM and ML models. The models were compared on two open-access motor insurance datasets, each with a different type of cover (fully comprehensive and third-party liability). While no single ML model outperformed across both pricing and discrimination metrics, the GLM significantly underperformed for most. The results indicate that ML may be considered a realistic and reasonable alternative to current practices. We advocate that benchmarking exercises for risk prediction models should be carried out to assess both pricing accuracy and fairness for any given portfolio.

Keywords: Actuarial pricing; Fairness; General insurance; Machine learning; Non-life insurance; Pricing bias; Pricing structure; Protected variables; Rate making

1. Introduction

This section gives an introduction to the study by providing a summary of modelling strategies and fairness measures. Furthermore, this section details the purpose of this study, which is primarily to design a benchmarking framework for assessing pricing fairness and accuracy simultaneously.

1.1 Non-Life Insurance Premium Pricing

Insurance is the business of transferring risk for a premium, and as such, it requires suitable methodologies to evaluate an appropriate value for transferring the risk. The standard approach for evaluating this appropriate value remains the generalised linear model (GLM) (De Jong & Heller, 2008; Frees, 2015; Haberman & Renshaw, 1996; Henckaerts et al., 2021; Nelder & Wedderburn, 1972). The GLM combines information collected about the policy, policyholder and vehicle in a linear representation to describe claims frequencies, claims severities or other quantities of interest that characterise the risk associated with a given policyholder. Under appropriate assumptions, the GLM yields statistically unbiased estimates (McCullagh & Nelder, 1989) and allows for straightforward interpretation, a critical requirement for both the underwriter and policyholder.

The most conventional dual GLM approach consists of modelling claims frequencies and severities separately before recombining the respective estimates into a single premium estimate. It is based on two separate models, typically using a discrete distribution for frequency modelling and a continuous distribution for severity modelling (Campo & Antonio, 2022; Henckaerts et al., 2021; Noll et al., 2020; Ohlsson & Johansson, 2010). The important theoretic property of unbiasedness at the portfolio level (McCullagh & Nelder, 1989; Wüthrich, 2020), along with the transparency and ease of interpretation of linear regression models, contributes to the predominance of GLM approach in non-life insurance pricing. This is despite some well-known limitations, such as the underlying assumption of a simplistic (linear) functional structure of the response variable, which may be unrealistic, making it difficult to incorporate complex interactions and non-linear associations (Campo & Antonio, 2022; Henckaerts & Antonio, 2022; König & Loser, 2020; Spedicato et al., 2018). Another limitation is the propensity of the GLM model to overfit with an increase in the number of variables used (McCullagh & Nelder, 1989). Given the growing amount of data available, the propensity of the GLM model to overfit should be considered where pricing is based on large tabulated variables, a common practice in the field (Campo & Antonio, 2022).

An alternative GLM approach based on a Tweedie distribution is also, but less commonly, considered (Campo & Antonio, 2022; Marin-Galiano & Christmann, 2004) where premiums are modelled directly by compounding frequency and severity information into a single (Tweedie) random variable. This approach can also be used to capture zero-inflated claims distributions, which, unlike in the conventional dual GLM approach, allows the use of the entire dataset to characterise the risk associated with policyholder profiles with no claims in the portfolio. The joint frequency-severity distribution is controlled by a power parameter, which implicitly assumes a level of correlation between the two quantities and imposes a particular shape on that distribution, which can limit the flexibility of the fitting procedure (Kurz, 2017). There is increasing interest in exploring alternative modelling strategies which can better exploit policyholder information so as to leverage non-linear associations between variables.

1.2 Alternative Modelling Strategies

ML modelling strategies can provide competitive alternatives to GLM, and improve risk stratification in non-life insurance pricing (Campo & Antonio, 2022; Dal Pozzolo et al., 2011; Guelman, 2012; Henckaerts et al., 2021;Wuthrich & Buser, 2023). Early ML models for pricing included tree-based approaches, as they alleviate the requirement of stringent model assumptions, albeit to the detriment of model interpretability. Regression tree and boosted tree models were considered for claims frequency prediction (Henckaerts et al., 2021; Noll et al., 2020). Noll et al. (2020) demonstrated that feature interactions were captured more effectively using tree-based techniques than with GLM. Henckaerts et al. (2021) compared GLM against various decision trees, random forests (RFs) and boosted trees for pricing, showcasing visualisation tools to obtain insights from the resulting models and their economic value. The loss of model interpretability of ML techniques is usually somewhat mitigated by some form of sensitivity analysis based, for example, on variable importance or partial dependency plots (Henckaerts et al., 2021; Kuo & Lupton, 2020). Alternative ML approaches include technical adaptations of GLM, for instance, adding a LASSO or elastic net regularisation penalty to mitigate overfitting (Devriendt et al., 2021). Such regularisation procedures, however, still imply using a specific parametric distribution model.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

Some studies have explored the potential of neural networks (NN) for pricing. Such studies include the study of use of a shallow feed-forward NN (Noll et al., 2020), an exploration of the value of autoencoders for pricing and premium bias correction (Blier-Wong et al., 2022; Grari et al., 2022; Wüthrich & Merz, 2022) and a framework based on generative adversarial networks to synthesise insurance datasets (Kuo, 2019). NN models can approximate non-linear functions very effectively but may appear overly elaborate to analyse portfolios with a small number of variables (Noll et al., 2020; Wüthrich, 2020). Wüthrich (2020) argued that NNs may provide pricing accuracy at the individual policy level but that state-of-the-art use of NNs does not assess unbiasedness (pricing fairness) at the portfolio level, which may be controlled with an early stopping rule in the gradient descent.

1.3 Pricing Fairness

Pricing fairness has always been a concern for insurance companies and is mainly understood in terms of premium bias and the nature of the policyholder information used (European Insurance and Occupational Pensions Authority, 2023b; Financial Conduct Authority, 2021). Attention towards pricing fairness is growing further with the development of ML-based pricing models that aim to improve risk differentiation for pricing (Henckaerts et al., 2021; Kuo & Lupton, 2020) without paying appropriate attention towards pricing fairness. ML models may not systematically result in fairer premiums because of less interpretability as compared to GLM. With the advent of ML models, more efforts are needed to control unlawful discrimination against specific customer groups (Baumann & Loi, 2023; Lindholm et al., 2022b). The European Insurance and Occupational Pensions Authority (EIOPA) focused in particular on the use of protected characteristics such as tenure and gender, advocating for increasing sophistication of risk management and governance processes (European Insurance and Occupational Pensions Authority, 2023a). For the purpose of pricing fairness, the Central Bank of Ireland (CBI)'s consumer protection framework was revised to mitigate the risks emerging from the use of innovative models (Central Bank of Ireland, 2024).

Some examples of generic guidelines for the promotion of fairness in pricing include: using a statistical model that is free of proxy discrimination, such that the dependent variable Y (claims frequency, severity or policy premium) is sufficiently well described by non-protected information X on the policyholder and product alone (Lindholm et al., 2022b); the use of an empirical microeconomic framework to evaluate the impact of fairness and accountability in the entire insurance pricing process (Shimao & Huang, 2022); the use of sets of fairness criteria (e.g. fairness through unawareness and through awareness, counterfactual fairness, independence from protected characteristics, etc.) to evaluate discriminatory effects on specific features or groups of policyholder profiles (Lindholm et al., 2022a; Mosley & Wenman, 2022; Xin & Huang, 2024); and the use of methods to remove various types of biases, through either pre-, in- or post-processing (Becker et al., 2022; Denuit et al., 2021; Mosley & Wenman, 2022; Wüthrich, 2020).

There are multiple measures of fairness, but not all measures of fairness and accuracy can be satisfied simultaneously (Baumann & Loi, 2023; Frees & Huang, 2023; Iturria, 2023; Lindholm et al., 2024b). In practice, fairness issues are often framed in terms of market behaviour, such as targeting specific customer groups (such as young drivers) or through "price walking." The Central Bank of Ireland (2022) "price walking" defines "price walking" as a practice where a customer's premium is gradually increased in longer-term policies. Formally, one could think of a fair premium as one that is as close to the true underlying risk of that individual as possible, adopting a risk-based pricing (Xin & Huang, 2024). The literature on individual fairness is scarce, and is primarily focused on group fairness, with a range of complementary definitions that predominantly include demographic parity (whereby premium distributions should be equal across protected groups overall), actuarial group fairness (whereby premium distributions should be equal across protected groups for any given risk profile) and calibration (in the sense that for a given predicted risk level, the actual risk should be the same across protected groups) (Barocas et al., 2023; Dolman & Semenovich, 2018; Dwork et al., 2011; Kleinberg et al., 2016; Lindholm et al., 2023; Xin & Huang, 2024). Attaining demographic parity does not eliminate unfairness at the individual level, as individuals from lower-risk groups may end up paying more than they should, while those from higher-risk groups might pay less than their risk would warrant (Dwork et al., 2011). Similarly, achieving actuarial group fairness ensures similar outcomes across protected groups within bands of homogeneous risk, but may sometimes make the overall model less accurate (Dolman & Semenovich, 2018). Another fairness metric called "disparate impact" consists of evaluating whether outcomes affect one group more than others (Barocas et al., 2023). Optimising for "disparate impact" may reduce accuracy by forcing similar outcomes across groups with different risk profiles (Xin & Huang, 2024).

Improving model fairness may thus impact other aspects of model performance depending on the model and the criteria being considered. Although model performance has been considered in GLM and a number of other modelling strategies (Colella & Jones, 2023; Fauzan & Murfi, 2018; Ferrario & Hämmerli, 2019; Henckaerts et al., 2021; Kuo & Lupton, 2020), towards, an overall picture of model performance combining predictive potential and fairness is rarely evaluated using a comparative benchmark. At the time of writing, we are only aware of one other paper reporting on such a combined assessment of fairness and pricing accuracy (Xin & Huang, 2024). Furthermore, the majority of published research concentrates on classification settings, evaluating fairness in binary decisions (see Dolman & Semenovich, 2018; Lindholm et al., 2022a and references within). Fairness in insurance pricing, treated as a regression problem, remains scarcely explored. This paper aims to narrow this gap in the literature. The subsection below details the goal and contribution of this paper.

1.4 Goals and Contribution

The aim of this work is to assess the potential benefits of ML-based pricing solutions against the current industry standard GLM-based approach in a regression setting. Recent papers report on studies of fairness that capture some aspects of the analyses proposed in this paper (for example, these studies may consider performance of the product model, or focus on fairness criteria) (Lindholm et al., 2022a, 2024b; Moriah et al., 2024). The proposed study brings together multiple aspects into a comprehensive comparative analysis of popular ML models, evaluating both model performance and pricing fairness simultaneously. This framework is described in Section 2. The performance benchmark used for this study is not unlike that of Henckaerts et al. (2021) However, Henckaerts et al. (2021) in their work used only a Poisson-gamma modelling strategy, whereas we extend the evaluation to include a Tweedie modelling strategy. Our assessment of pricing fairness combines the approach carried out by the Central Bank of Ireland (Central Bank of Ireland, 2022) with three common fairness metrics for demographic parity, group fairness and calibration (Dolman & Semenovich, 2018). We define these metrics with respect to conditional distributions of raw actual premiums in Section 2.5.9, and propose a practical quantitation for these three axioms. The models are compared on two open-source datasets in Section 3 to evaluate methods on two distinct common types of cover: third-party liability and fully comprehensive liability. Section 4 presents the details of a fairness analysis carried out on the second dataset. The findings of this study are discussed further in Section 5 and demonstrate that ML strategies provide consistent improvements over GLM, but also that there is not necessarily one optimal ML model for non-life premium pricing. Furthermore, Section 5 also illustrates that ML strategies can yield better performance, which can be achieved together with improvement in various fairness criteria.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

2. Premium Modelling and Assessment

This section describes the premium modelling methodology and defines the key performance indicators and measures of pricing fairness considered for this study.

2.1 Quantities of Interest

We consider a portfolio of n independent policies with data $(\mathbf{D}_i, \mathbf{Z}_i, N_i, L_i, e_i)$, $\forall$ $i = 1, \ldots, n$, where:

- $\mathbf{D}$ denotes a vector of unprotected policyholder information (e.g. vehicle age, area, etc.), which can be used to rate on,
- $\mathbf{Z}$ denotes protected policyholder variables (e.g. gender, ethnicity, etc.), which may not be used for pricing,
- N and L, respectively, denote the number of claims and the aggregated claims amount for each policy over its exposure e.

In what follows, we will in turn use either the unprotected policyholder data matrix $\mathbf{X} = \mathbf{D}$, or the combined $n \times p$ data matrix $\mathbf{X} = [\mathbf{D}, \mathbf{Z}]$ comprising of both $\mathbf{D}$ and $\mathbf{Z}$. The nature of protected variables varies by jurisdiction and laws applicable in the country. In this paper, we consider fairness with respect to gender, a typical variable that is protected in Ireland. Factors such as gender, age and marital status are considered protected characteristics under EU anti-discrimination law. However, gender has a unique status among these protected characteristics; regardless of its commercial or underwriting relevance to insurance risks, it cannot be used in setting insurance prices (European Parliamentary Research Service, 2017).

Using this information, we are interested in evaluating the intrinsic policyholder risk

$$P_T = \mathbb{E}(L|\mathbf{D}, \mathbf{Z})$$

and a best feasible estimate of the expected loss for a policy

$$P_A = \mathbb{E}(L|\mathbf{D})$$

Quantities P_T and P_A can be seen, respectively, as the “best estimate price” and “unaware price” (Lindholm et al., 2022a), to which the company would apply a number of commercial margins and adjustments to account for expenses, levies, etc. Although P_T and P_A as defined above do not include any such additional margins and adjustments, they provide a basis for intrinsic evaluation of risk/loss. For the sake of simplicity, and as described in other works (Lindholm et al., 2024b) in the rest of the paper, we refer to P_T and P_A, respectively, as technical premium (TP) and actual premium (AP). In this work, we consider two distinct modelling strategies to estimate P_T and P_A, namely, the Poisson-gamma (product) model, and the Tweedie model.

2.2 Poisson-gamma (Product) Model

The first modelling strategy considered is the most prevalent, Poisson-gamma (product) model. This strategy consists of estimating frequency and severity as two separate steps. Claims frequencies are typically modelled by a discrete count distribution. The Poisson distribution being a common choice (Henckaerts et al., 2021; Noll et al., 2020) and the distribution used hereafter, such that for policyholder $i \in \{1, \ldots, n\}$,

$$N_i \sim Poi(\mu_i e_i), \quad \mu_i > 0, \ \ e_i > 0$$

for some rate μ_i. Then the claims frequency F over a unit period of exposure (say, one-year exposure), i.e. the number of claims N proportional to exposure e, is such that

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

$$\mathbb{E}(F) = \mathbb{E}\left(\frac{N}{e} \mid e > 0\right)$$

or, for a given policy, evaluated as

$$F_i = \frac{N_i}{e_i}$$

such that $\mathbb{E}(F_i) = \mu_i$ and $Var(F_i) = \frac{\mu_i}{e_i}$, i.e. F_i is pseudo-Poisson random variable (either over or under-dispersed depending on the sign of $e_i - 1$; usually $e_i \leq 1$). In practice, we can estimate μ_i as the maximum likelihood estimator (MLE)

$$\hat{\mu}_i = F_i$$

A conventional approach consists of fitting a GLM to the policyholder information $(\mathbf{D}, \mathbf{Z}, N, e)$ to explain claims frequencies for individual profiles as

$$\log\mathbb{E}(F|\mathbf{X}) = g_F(\mathbf{X}) = \mathbf{X}\beta_F \tag{1}$$

with real-valued regression coefficients $\beta_F \in \mathbb{R}^p$. Model (1) may be fit to the data by minimising the Poisson deviance between claims frequency data $\{y_i\}_{i=1}^n$ (e.g. where $Y = F$) and a set of model predictions $\{g_F(\mathbf{x}_i)\}_{i=1}^n$ obtained from policyholder data $\mathbf{X}$, defined as in (Henckaerts et al., 2021; Venables & Ripley, 2013) by

$$\mathcal{D}_F\big(y, g_F(\mathbf{X})\big) = 2\sum_{i=1}^{n}\left(y_i log\left(\frac{y_i}{g_F(\mathbf{x}_i)}\right) - \big(y_i - g_F(\mathbf{x}_i)\big)\right) \tag{2}$$

Let us then define claim severity S as the aggregated loss (or claim amount) L divided by the number of claims from a subset of polices that have at least one or more claims reported and paid out, such that

$$\mathbb{E}(S) = \mathbb{E}\left(\frac{L}{N} \mid N > 0\right)$$

or, for a given policy i, evaluated as

$$S_i = \frac{L_i}{N_i}$$

Hereafter, following a conventional approach, S is modelled by a gamma distribution to allow for skewness arising naturally in such data. Similar to equations (1) and (2), claim severities may be modelled using a distinct GLM, that we may denote as $g_S(X) = \mathbf{X}\beta_S$ with $\beta_S \in \mathbb{R}^p$. This model may be achieved by minimising the gamma deviance between the severity observations $\{y_i\}_{i=1}^n$ (e.g. where $Y = S$) and a set of model predictions $\{g_S(\mathbf{x}_i)\}_{i=1}^n$ defined as in (Henckaerts et al., 2021) by

$$\mathcal{D}_S\big\{y, g_S(\mathbf{X})\big\} = 2\sum_{i=1}^{n} w_i\left(\frac{y_i - g_S(\mathbf{x}_i)}{g_S(\mathbf{x}_i)} - \log\left(\frac{y_i}{g_S(\mathbf{x}_i)}\right)\right) \tag{3}$$

where w_i is the weight for each observation. Hereafter we use $w_i = N_i$.

The overall policy premium P may then be defined as the expected loss L per unit of exposure, such that

$$P = \mathbb{E}\left(\frac{L}{e}\right)$$

under the assumptions of a non-zero chance that a claim be made, and that claim frequencies and severities are independent of each other. A policy premium estimate $\hat{P}$ over a unit exposure may

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

then be obtained as the product of estimates for frequency $\hat{F}$ and severity $\hat{S}$ (Denuit et al., 2007; Frees et al., 2014; Parodi, 2014), i.e.

$$\hat{P} = \hat{F} \times \hat{S}$$

such that (Henckaerts et al., 2021)

$$\mathbb{E}(\hat{P}) = P = \mathbb{E}(F \times S) = \mathbb{E}\left(\frac{N}{e}\right) \times \mathbb{E}\left(\frac{L}{N} \,|\, N > 0\right).$$

2.3 Tweedie Model

A relatively common alternative to the Poisson-gamma product model consists of modelling the quantity $\mathbb{E}(\hat{P})$ directly as a compound variable under the Tweedie distribution (Quijano Xacur & Garrido, 2015; Shi, 2016). This member of the exponential dispersion family is characterised by a variance function of the form $\mathrm{Var}(Y) = \phi\gamma^{\alpha}$, where ϕ is the dispersion parameter, γ is the mean, and α is the power parameter such that $\alpha \neq 0$ and $\alpha \neq 1$. Given a set of n independent observations $\{y_1, \ldots, y_n\}$ of a Tweedie random variable Y, an estimate of the premium P is derived by minimising the Tweedie deviance between the observations $\{y_i\}_{i=1}^{n}$ and a set of model predictions $\{g_T(\mathbf{x}_i)\}_{i=1}^{n}$, defined as

$$\mathcal{D}\big(y, g_T(\mathbf{X})\big) = \sum_{i=1}^{n} \frac{2}{\phi}\left(\frac{y_i^{2-\alpha}}{(1-\alpha)(2-\alpha)} - \frac{y_i g_T(\mathbf{x}_i)^{1-\alpha}}{1-\alpha} + \frac{g_T(\mathbf{x}_i)^{2-\alpha}}{2-\alpha}\right) \tag{4}$$

with respect to (ϕ, γ, α) and the model parameters for g_T.

2.4 Premium Modelling

A number of models $g(\mathbf{X})$ using the Poisson, gamma or Tweedie deviance $\mathcal{D}$ defined, respectively, by equations (2), (3) or (4), as cost function were compared. They include the conventional GLM, with a structure similar to equation (1), allowing for interaction terms (McCullagh & Nelder, 1989; Ohlsson & Johansson, 2010) in a forward stepwise selection of the final model ESL; an elastic net and a LASSO (Tibshirani, 1996; Zou & Hastie, 2005), using $\delta = 0.5$ and $\delta = 1$, respectively, and a regularisation parameter $\lambda > 0$ in

$$\mathcal{D}_{\lambda}\big(\hat{\beta}\big) = \sum_{i=1}^{n} \mathcal{D}(y_i, \mathbf{x}_i\hat{\beta}_j)^2 + \lambda \sum_{j=1}^{p}\left(\frac{1}{2}(1-\delta)\beta_j^2 + \delta\big|\beta_j\big|\right)$$

Furthermore, they also include a pruned regression tree (Breiman et al., 1984; Friedman, 2001) partitioning the dataset into J leafs $R = \{R_1, \ldots, R_J\}$ and with complexity penalisation $c > 0$ as described in (Henckaerts et al., 2021) by optimising

$$\sum_{j=1}^{J} \sum_{i:\mathbf{x}_i \in R_j} \mathcal{D}(y_i, \hat{y}_{R_j}) + cJ \sum_{i:\mathbf{x}_i \in R} \mathcal{D}(y_i, \hat{y}_R)$$

where $\hat{y}_R$ denotes the average value of Y within a leaf or set of leaves R as defined above; a RF (Breiman, 2001) using an ensemble of M trees; a gradient boosting model (GBM) (Friedman, 2001) using a penalised loss gradient at each step $m = 2, \ldots, M$, the latter being defined by

$$-g_m(\mathbf{x}) = -\left[\frac{\partial\mathcal{D}\big(y, g(\mathbf{x})\big)}{\partial g(\mathbf{x})}\right]_{g(\mathbf{x}) = g_{m-1}(\mathbf{x})}$$

as well as extreme gradient boosting (XGB) (Chen & Guestrin, 2016), penalising (a second-order linearisation of) the deviance by

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

$$\sum_{i=1}^{n} \mathcal{D}\big(y_i, g(\mathbf{x}_i)\big) + \xi J + \frac{\tau}{2}||\omega||$$

where J is the number of terminal nodes or leaves in a tree, $\xi > 0$ a user defined penalty encouraging pruning, and for some regularisation $\tau > 0$ of the final leaf weights ω determined for the ensemble model by stochastic gradient descent. XGB in particular has gained a lot of popularity in the actuarial field in recent years, some research reporting it to outperform other models for insurance claims and frequency prediction (Colella & Jones, 2023; Fauzan & Murfi, 2018; Ferrario & Hämmerli, 2019; Henckaerts et al., 2021; Kuo & Lupton, 2020).

2.5 Performance Evaluation

2.5.1 Resampling Framework

Prior to the analyses, each of the two datasets considered in Section 3 were split randomly into a build sample and a test sample, comprising, respectively, 83% and 17% of the complete dataset. This random splitting was done to ensure that each of the cross-validation folds contained roughly the same amount of data as the test set. The build set was used for model training using nested 5-fold cross-validation (CV) (Hastie et al., 2009). The model hyperparameters were set from the CV training folds, via a grid search, to optimise deviance $\mathcal{D}$ as defined by equations (2), (3) or (4) as appropriate. The performance metrics defined hereafter were then evaluated on each of the five validation folds and averaged to obtain a first measure of performance. The resulting model fits were then applied to the test sample for external model performance evaluation, using .632-bootstrapping ESL to compute standard errors and non-parametric confidence intervals for the performance indicators points from 250 bootstrap resamples.

2.5.2 Post-Modelling Calibration

All model predictions (premium estimates) were calibrated to adjust for model bias before they were used to carry out analysis. The calibration factor C was calculated based on the model fits obtained on the training dataset and is evaluated as

$$C = \frac{\sum_{i=1}^{n} L_i}{\sum_{i=1}^{n} \hat{P}_i}$$

2.5.3 Premium Estimation Error and Bias

All models were first compared on the test set with respect to bootstrapped Poisson, gamma and Tweedie deviances as defined in equations (2), (3) or (4), to evaluate overall model error. Model bias was also evaluated on the test data as the difference between the total sum of observed losses and the premium estimates obtained from each model, defined as

$$\sum_{i=1}^{n} \left(L_i - \hat{P}_{T,i}\right)$$

2.5.4 Risk Stratification

Lorenz curves were used to compare premiums by analysing the distribution of premiums versus losses, where both quantities were ordered by increasing relativities

$$r_i = \frac{\hat{Q}_i}{\hat{P}_i},\ i = 1, \ldots, n$$

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

i.e. by the ratio of the premium prediction obtained from the competing model ($\hat{Q}$) to that of the benchmark model ($\hat{P}$) (Henckaerts et al., 2021), whose empirical c.d.f. is denoted hereafter by F_n. Sorting policies in terms of their increasing relativity, the ordered Lorenz curve then has coordinates $\left(\hat{F}_L, \hat{F}_{\hat{P}}\right)$ defined by

$$\hat{F}_{\hat{P}}(s) = \frac{\sum_{i=1}^{n} \hat{P}(\mathbf{x}_i) I(F_n(r_i) \leq s)}{\sum_{i=1}^{n} \hat{P}(\mathbf{x}_i)}, \quad s \in [0, 1]$$

and

$$\hat{F}_L(s) = \frac{\sum_{i=1}^{n} y_i I(F_n(r_i) \leq s)}{\sum_{i=1}^{n} y_i}, \quad s \in [0, 1]$$

where $I(.)$ is the indicator function. The curve $\left(\hat{F}_L, \hat{F}_{\hat{P}}\right)$ follows the identity line $\left(y = x\right)$ in the case of benchmark estimation. The Gini index is calculated as twice the area between the ordered Lorenz curve and the identity line (Frees et al., 2014; Henckaerts et al., 2021). Larger index indicates greater risk differentiation can be obtained by using the competing premium $\hat{Q}$ compared to the reference premium $\hat{P}$, and hence the competing pricing model is more profitable. This case corresponds to a concave Lorenz curve. The profitability point is as per the theory (Frees et al., 2014) that the profitability is the area under the Lorenz curve. However, if Company B has a better estimate of elasticity and conversion/retention rates then they could operate more profitably than Company A, even if A has the more profitable pricing model. In this analysis, we assume that the price elasticity modelling capabilities between companies are similar and they do not play a role.

2.5.5 Loss Ratios and Conversion Rates in Open-market Competition

A two-player game was implemented to simulate an open-market competition between the industry-standard GLM methodology and each of the competing alternatives for premium prediction. The models are evaluated in terms of their loss ratio (LR) defined as the earned premium $\hat{P}_A$ over paid claims L ($LR = \frac{L_i}{\hat{P}_{A,i}}$). In this game, for a particular policyholder, the model yielding the lowest premium wins the business and hence secures the premium predicted for the policy as revenue. If $\hat{P}^{GLM}_{A,i} < \hat{P}^{ML}_{A,i}$, GLM secures the business, thereby obtaining both the associated revenue and the cost of claims from that policy.

This simulation thus shows the range of premiums a model is expected to earn whilst competing with the GLM reference, as well as the amount of claims it is expected to pay out. Loss ratios for each model are then calculated by taking the ratio of premium and claims amount, to measure profitability. The model that produces a lower ratio indicates a more profitable book of business.

Note that a measure of the expenses was not available in the datasets analysed here, and hence, a combined ratio (LR plus expense ratio) could not be measured. An expense ratio is generally significantly smaller than the corresponding LR. To carry out the study, we assumed that expense ratios are consistent between providers.

2.5.6 Sensitivity Analysis

2.5.6.1 Variable Importance. Variable importance was assessed to understand which variables drive premium prediction in each of the models considered. Any discrepancies in the variable importance values may indicate specific model sensitivity to the information collected on policyholders. Variable importance in regression models (GLM and its regularised alternatives) was measured directly from the fitted coefficients by normalising their magnitude (in absolute value) by the sum of all model coefficient magnitudes (Fryda, 2023). Variable importance in a tree was derived by taking the sum of improvements in the loss function at each split (Breiman et al., 1984)

$$\mathcal{J}\Big(X_{(j)}\Big) = \sum_{\ell=1}^{\eta-1} I\Big(v_\ell = X_{(j)}\Big)(\Delta\mathcal{D})_\ell,\ j = 1, \ldots, p$$

where η denotes the number of internal nodes (or splits) of the tree, $v_\ell \in \mathbf{X}$ is the variable used to perform the ℓ^{th} split, and $(\Delta\mathcal{D})_\ell$ measures the magnitude of the difference in deviance $\mathcal{D}$ (defined by equations (2) or (3) or (4)) before and after split ℓ. For the other tree-based models, namely RF, GBM and XGB, variable importance was calculated as the average of importances measured for the variable over all trees in the ensemble (Breiman et al., 1984; Hastie et al., 2009). Importance analysis was also used to assess the impact of protected variables on actual premiums derived from the various models.

2.5.6.2 Model Agnostic Evaluation. The impact of individual features on model predictions was further evaluated in model-independent ways, using SHapley Additive exPlanations (SHAP) analyses and partial dependence plots. For a prediction function g, the SHAP value ϕ_j for feature $X_{(j)}$ is a feature importance value computed as the average marginal contribution of feature $X_{(j)}$ across all possible subsets of features (Lundberg & Lee, 2017) as

$$\phi_j = \sum_{S \subseteq \mathbf{X} \setminus \{X_{(j)}\}} \frac{|S|!(|\mathbf{X}| - |S| - 1)!}{|\mathbf{X}|!} \Big(g(S \cup \{X_{(j)}\}) - g(S)\Big),$$

where $\mathbf{X}$ is the whole feature set, S is a subset of features not containing $X_{(j)}$. $g(S \cup \{X_{(j)}\})$ and $g(S)$ denote the premium prediction function evaluations obtained when only the features in the respective subset $S \cup \{X_{(j)}\}$ and subset S are known. In practice, $g(S)$ can be evaluated by integrating over all possible values of $X_{(j)}$, i.e. $g(S) \approx E[g(\mathbf{X})|S]$, typically by sampling the missing features conditionally on the known ones (S) (Molnar, 2022).

A partial dependence plot (PDP) is a popular model-agnostic visualisation tool for model interpretability. It depicts the marginal effect of one or more input features on the predicted outcome, averaging over the joint distribution of the remaining features (Friedman, 2001; Molnar, 2022). Formally, for a premium prediction function $g(x)$ and a feature $X_{(j)} \in \mathbf{X}$, the partial dependence function is defined as

$$\hat{g}(x_{(j)}) = \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{X}_i | X_{(j)} = x_{(j)}).$$

This marginalisation estimates the expected prediction when $X_{(j)} = x_{(j)}$ is fixed and all other features vary according to the empirical data distribution. For continuous features, we discretise the feature range (using deciles) into a grid of values and compute the partial dependence at each grid point.

2.5.7 Defining Actual and Technical Premiums

The ratio of actual premium to technical premium (APTP)

$$\rho = \frac{P_A}{P_T} \tag{5}$$

may be used to assess how the removal of protected information from the model may impact the unaware price P_A relative to the best estimate risk P_T. An APTP ratio greater than 1 indicates a higher premium than required to cover the expected costs of the policy for that policyholder profile. On the other hand, an APTP ratio lower than 1 indicates the customer paid less than the expected cost of insuring the risk. Univariate analyses of APTP ratios were carried out to illustrate

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

how some typical protected variables affected the pricing methodologies listed in Section 2.4. Since our quantitations of P_A and P_T are raw model output (see Section 2.1), ρ may be seen as a proxy for an APTP ratio.

2.5.8 Statistical Distance Between Populations

In order to compare premium distributions across fairness groups, we used the Wasserstein distance (Pichler, 2014), with order 1. Given two premium estimators P and P', with probability distributions F_P and $F_{P'}$, this distance is defined as

$$W(P, P') = \int_{-\infty}^{\infty} |\mathrm{F_P(u)} - \mathrm{F_{P'}(u)}|\mathrm{du} \tag{6}$$

In practice, this quantity can be evaluated by using the empirical distribution functions for sets of premium estimates $\{P_i\}_{i=1}^n$ and $\{P'_i\}_{i=1}^n$, i.e. $F_Q(u) = \frac{1}{n}\sum_{i=1}^n I(Q_i \leq u)$ for either $Q = P$ or $Q = P'$, in a numerical approximation of the integral term. Wasserstein distance as defined in equation (6) is an actual metric that measures the area between the marginal distributions F_P and $F_{P'}$ (Chhachhi & Teng, 2023; De Angelis & Gray, 2021). The Wasserstein distance yields a positive value that increases with the distance between the two distributions that is bounded upwards. In this analysis, we use the first order of the Wasserstein distance to mitigate sensitivity to the natural skewness in the distributions of premiums (Chhachhi & Teng, 2023).

2.5.9 Characterisation of Fairness

Fairness is commonly assessed by analysing actual premium distributions with respect to protected variables. The nature of the departure from a fair actual premium may be investigated in more detail, especially to assess whether it affects a specific protected group. Building on prior studies (Dolman & Semenovich, 2018; Lindholm et al., 2022a; Mosley & Wenman, 2022; Xin & Huang, 2024), a framework for the evaluation model fairness can be extended to assess compliance of the actual premiums P_A with the following three axiom.

Axiom 1 (Demographic Parity). *The actual premium P_A must not depend on the protected features. In particular, where the feature is a categorical variable with L levels, $Z \in \{l_1, \ldots, l_L\}$ (e.g. gender), then it should be the case that*

$$\mathbb{P}(P_A|Z = l_i) = \mathbb{P}(P_A|Z = l_j)\ \forall i \neq j$$

Axiom 2 (Actuarial Group Fairness). *The distribution of actual premium P_A conditional on technical premium P_T must not depend on the protected information Z, i.e.*

$$\mathbb{P}(P_A|P_T, Z = l_i) = \mathbb{P}(P_A|P_T, Z = l_j)\ \forall i \neq j$$

Axiom 3 (Calibration). *The distribution of actual loss Y conditional on actual premium P_A must be conditionally independent of any protected feature Z, i.e.*

$$\mathbb{P}(Y|P_A, Z = l_i) = \mathbb{P}(Y|P_A, Z = l_j)\ \forall i \neq j$$

Without loss of generality, Z could be a numerical variable and the conditional distributions may be evaluated at discretised Z-value bins.

Alignment with Axiom 1 may be assessed quantitatively by measuring the magnitude of disparities between probability distributions $\mathbb{P}(\hat{P}_A|Z = l_i)$ and $\mathbb{P}(\hat{P}_A|Z = l_j)$ for any two groups l_i and l_j of a given protected variable Z. In our framework, we propose using the Wasserstein distance between the distributions of two subgroups of policyholders. As an example, for a binary categorical variable $Z \in \{l_1, l_2\}$ over the n_{test} data points:

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

$$d_Z = W\left(\hat{P}_A \middle| Z = l_1, \hat{P}_A \middle| Z = l_2\right) \tag{7}$$

Similarly, adherence to Axiom 2 may be measured on the basis of the Wasserstein distance in the test dataset for a given category $\{Z = l\}$, and particular TP percentiles of interest. As an example, subsetting the dataset into bands defined by the TP quintiles $\{\tilde{P}_{T,0}, \tilde{P}_{T,1}, \ldots, \tilde{P}_{T,5}\}$ such that

$$\mathcal{P}_T \in \left\{\left(\tilde{P}_{T_{k-1}}, \tilde{P}_{T_k}\right], k \in \{1,2,3,4,5\}\right\}, \tag{8}$$

to measure adherence for each band we calculate the Wasserstein distance as:

$$d_{l,\mathcal{P}_T} = W\left(\hat{P}_A \middle| \left(\hat{P}_T \in \mathcal{P}_T, Z = l_1\right), \hat{P}_A \middle| \left(\hat{P}_T \in \mathcal{P}_T, Z = l_2\right)\right) \tag{9}$$

Adherence to Axiom 3 may be measured on the basis of the Wasserstein distance of observed losses in the test dataset between protected groups in a given actual premium quintile band, as follows:

$$d_{L,Z} = W\left(L \middle| \hat{P}_A, Z = l_1, L \middle| \hat{P}_A, Z = l_2\right) \tag{10}$$

In each case, a lower Wasserstein distance from equations (7), (9) or (10) indicates better conformity to the axioms. In addition, two-sided Mann–Whitney tests were performed at the 5% significance level to assess whether any of the conditional distributions compared in equations (7), (9) and (10) were statistically different. These quantitations are reported on in Section 4.1.

Other metrics may be considered to complement indicators from equations (7), (9) and (10). Quantitatively, in what follows we also evaluate the distance between distribution medians, to provide additional context around the value of the Wasserstein distance. Complementary assessment may also be performed qualitatively. For the qualitative assessment, we consider evaluating the statistical significance of a nonparametric Wilcoxon test applied to the samples of premium estimates across protected groups.

2.6 Implementation

All analyses were carried out using R version 4.2.0 (R Core Team, 2021) and its H2O interface (Fryda, 2023). Some of our implementation builds upon source code developed by (Henckaerts et al., 2021), available at https://github.com/henckr/distRforest. The RF and regression tree models rely on the distRforest (Henckaerts et al., 2021) and rpart (Therneau & Atkinson, 2019) R packages. Dedicated packages were also used for model glmnet (Friedman et al., 2010; Zou & Hastie, 2005), GBM (Ridgeway, 2014; Southworth, 2015) SHAP analyses and PDPs were implemented using the iml package in R (Molnar, 2022), which supports both individual and grouped feature effects. To calculate the Wasserstein distance, the R package Transport (Dominic et al., 2023) was used. All source code for our analyses is available in open access on Github at https://github.com/tisrani/Pricing.

3. Case Studies

In this section, we present the outputs obtained by benchmarking the modelling methodologies discussed in Section 2 on two distinct datasets of French motor insurance claims, with respect to the key performance indicators defined in Section 2.5.

3.1 Datasets

For this study we used two open-access datasets which comprised of both frequency and severity. These datasets yielded the opportunity to create a benchmark on two distinct types of cover namely: third-party liability and fully comprehensive cover. Furthermore, they allowed exploring

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

the impact of different sets of covariates such as driver age. Additionally, Dataset 2 contained information on protected variables, such as gender, that allowed for a realistic benchmark on fairness.

Dataset 1 comprised of two parts, namely freMTPL2freq and freMTPL2sev from the R package CASdatasets (Charpentier, 2014; Dutang & Charpentier, 2015), containing data from a French motor third-party liability book of business over one calendar year. The merged dataset contained the numbers of claims and aggregated amounts incurred by each of 678,013 policies, of which 24,944 had non-zero claims. Dataset 1 also included information on the policy (bonus-malus), main driver's age and location (area, density and region), and vehicle (age, brand, power, fuel type). We refer the reader to (Noll et al., 2020) for a more detailed description of Dataset 1. Dataset 1 has also been used in a number of other works (Ciatto et al., 2023; Delcaillau et al., 2022; Denuit et al., 2021; Ferrario et al., 2020; Krasniqi et al., 2022; Lorentzen & Mayer, 2020; Schelldorfer & Wuthrich, 2019; Su & Bai, 2020).

Dataset 2 also had two parts, namely, pg17trainpol and pg17trainclaim (Dutang & Charpentier, 2015) from the R package CASdatasets. Dataset 2 contains observations of 64,515 fully comprehensive cover policies with one full year of exposure. Out of the 64,515 policies, there were claims in 8,648 of them. Since there was no information on the type of claims, we could not identify how much of the losses were related to third-party claims. Hence, an additional 35,485 third-party-only liability policies were removed from this second dataset. This yielded a more statistically balanced dataset, and also allowed for a more effective benchmark design, because now Datasets 1 and 2 thus contained a single and distinct type of policy cover. Dataset 2 also included variables on the policy (bonus-malus, coverage, payment frequency, pay-as-you-drive), main driver (INSEE code[1], age, gender, number of years with full licence), and vehicle (type, make, model, age, number of cylinders, DIN[2], fuel type, maximum speed vehicle can achieve, weight and value). Information on a second driver and other secondary variables were ignored to simplify the analysis. Furthermore, as the vehicle age was already known, the vehicle sales dates were ignored. Dataset 2 has also been used in other works (Bove et al., 2022; Brouste et al., 2024; Havrylenko & Heger, 2022; Simjanoska, 2022).

Claims with extremely low and high amounts outside the range of (€50, €10,000) were removed in both datasets. Following this, Datasets 1 and 2 contained $n = 24,123$ and $n = 8,500$ policies with non-zero claims, respectively. The base level for each factor (e.g. vehicle gas) was set at the level with the highest number of policies.

Following a conventional actuarial representation of policyholder information, categorical variables (for Dataset 1, region, vehicle brand and area; for Dataset 2, vehicle make and vehicle model) were recategorised into broader, well-defined groups, to select key differentiating features and use interpretable modelling approaches to enhance the significance and understanding of predictions for each group. Here, a tree-based binning strategy (Henckaerts et al., 2021) was applied to create categorical risk factors with levels being separated based on the observed claims values at each level. Tree-based binning strategy is typically used to improve the generalisability of estimated premiums and to simplify the model in line with the parsimony principle. The distributions of these recategorised variables are provided in Appendix A. Following this step, information on vehicle type in Dataset 2 was removed due to its high correlation with other vehicle features based on Spearman's or Cramer's correlation, as appropriate.

[1]Variable INSEE corresponds to an official 5-digit alphanumeric code of the French city/municipality where the policyholder lives, attributed by the French National Institute for Statistics and Economic Studies. There are about 36,000 "communes" in France, but not every one of them is present in the dataset (which contains only 18,000 of them). The first 2 digits of the INSEE code identify the department (they are 96, not including overseas departments). The INSEE code or department code may be used to merge external data to the datasets: population density, OSM data, etc.

[2]Variable DIN represents motor power.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

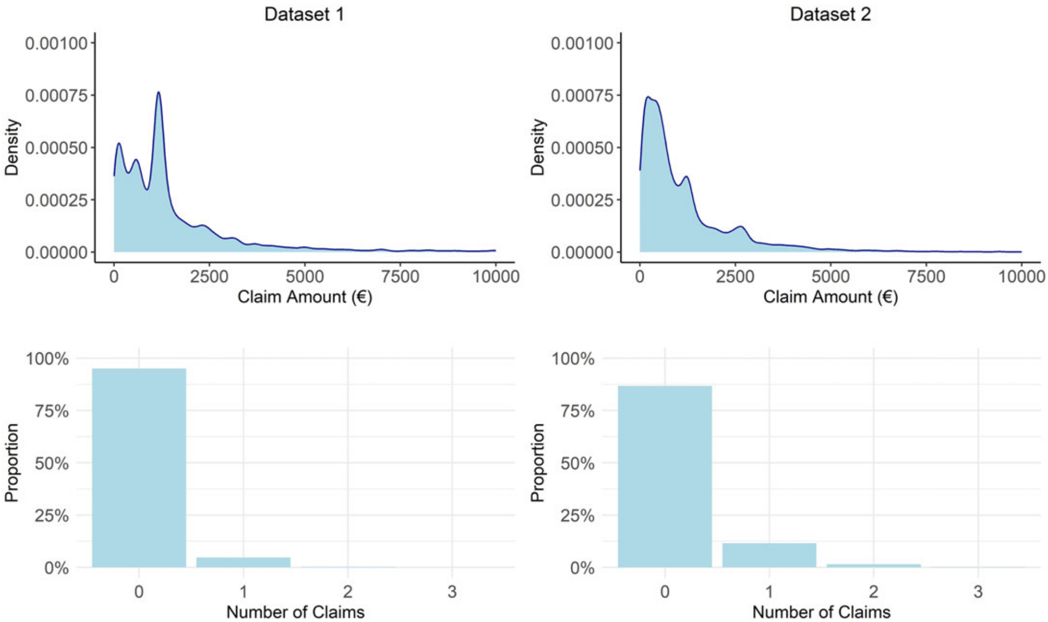

Figure 1. Distributions of claims amounts (top) and number of claims (bottom) for Dataset 1 (left) and Dataset 2 (right).

The overall distributions of claims numbers (N) and amounts (L) in the datasets after data cleaning and pre-processing are shown in Figure 1 for both datasets. These were found to be globally similar in shape, although the distribution of claims amount in Dataset 1 has a relatively fatter tail.

3.2 Which Factors Drive the Premiums?

3.2.1 Variable Importance

Analyses of variable importance showed both similarities and differences between models, influenced by the nature of modelling strategy: the contrast between linear regression and tree-based modelling strategy. These similarities and discrepancies can be observed in Table 1. The figure also indicates that only a small number of variables effectively drove any of the models, particularly for the estimation of frequency in both datasets, and of severity in Dataset 2.

Considering the product modelling approach in Dataset 1, overall the variable importance values and patterns were similar across models for frequency estimation, with only minor discrepancies across model types. This was also largely the case for severity estimation, although the XGB model tended to leverage more variables than the other models. A similar conclusion could also be drawn for the Tweedie modelling approach in Dataset 1. Under the Tweedie approach, variable importance values from XGB tended to be more spread out across the top 6 variables in the table, and regression models in this framework used more variables. For the regression model, increased importance was observed for variables area and vehicle gas.

In Dataset 2, different patterns of variable importance were observed between regression-based and tree-based techniques. However, similar patterns of variable importance were observed within each modelling family. In the product modelling approach, tree-based models placed more emphasis on vehicle value, age and weight for frequency estimation, while vehicle make was more important in regression models. For severity estimation, tree-based techniques placed higher importance to vehicle weight and engine cylinder, whilst bonus-malus was a large contributor in regression settings. Under the Tweedie modelling approach, vehicle value had a larger

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

Table 1. Variable importance metrics for frequency (left), severity (centre) and Tweedie (right) models trained on Dataset 1 (top) and Dataset 2 (bottom), as defined in Section 2.5.6. The greyed-out cells indicate variables that were removed during the stepwise selection in the GLM model building process. A value of 0% in the table corresponds to an importance below 0.5%

Frequency Dataset 1	GLM	Lasso	Elastic Net	RF	GBM	XGB
Bonus Malus	45%	25%	25%	32%	39%	37%
Region	0%	0%	0%	2%	3%	1%
Driver Age	10%	7%	7%	7%	6%	6%
Vehicle Age	26%	59%	59%	36%	32%	34%
Vehicle Brand	1%	0%	0%	14%	10%	9%
Density	3%	0%	0%	1%	2%	1%
Vehicle Power	4%	9%	9%	1%	2%	4%
Area	7%	0%	0%	1%	0%	0%
Vehicle Gas	3%	0%	0%	5%	7%	6%

Severity Dataset 1	GLM	Lasso	Elastic Net	RF	GBM	XGB
Bonus Malus	20%	12%	12%	11%	20%	18%
Region	17%	39%	39%	26%	35%	14%
Driver Age	9%	5%	5%	10%	4%	16%
Vehicle Age	21%	10%	10%	14%	24%	22%
Vehicle Brand	12%	17%	17%	14%	14%	10%
Density	7%	2%	2%	10%	1%	15%
Vehicle Power	7%	4%	4%	3%	1%	5%
Area	6%	10%	10%	9%	1%	0%
Vehicle Gas		1%	1%	1%	0%	1%

Tweedie Dataset 1	GLM	Lasso	Elastic Net	RF	GBM	XGB
Bonus Malus	42%	62%	61%	64%	73%	51%
Region	6%	4%	4%	6%	4%	6%
Driver Age	9%	6%	7%	13%	8%	16%
Vehicle Age	8%	4%	4%	2%	3%	6%
Vehicle Brand	1%	8%	8%	5%	3%	9%
Density	4%	3%	3%	5%	4%	11%
Vehicle Power	7%	4%	4%	1%	1%	2%
Area	13%	0%	0%	3%	1%	0%
Vehicle Gas	11%	9%	9%	1%	2%	0%

Frequency Dataset 2	GLM	Lasso	Elastic Net	RF	GBM	XGB
Vehicle Make	10%	21%	21%	0%	4%	2%
Vehicle Age	19%	2%	2%	29%	11%	28%
Vehicle Speed	3%	0%	0%	2%	7%	2%
Vehicle Value		0%	0%	31%	20%	24%
Vehicle Weight		0%	0%	14%	10%	5%
Driver Age	4%	0%	0%	1%	6%	4%
Density	10%	0%	0%	4%	11%	10%
Payment Frequency	12%	14%	15%	0%	3%	2%
Bonus Malus	10%	34%	33%	4%	10%	7%
Vehicle Fuel	11%	10%	10%	8%	3%	8%
Policy Usage	7%	8%	8%	0%	2%	0%
Policy Duration		0%	0%	1%	4%	3%
Pay as You Drive	4%	7%	8%	0%	0%	1%
Sex	3%	2%	2%	0%	1%	0%
Engine Cylinder	7%	0%	0%	4%	6%	4%

Severity Dataset 2	GLM	Lasso	Elastic Net	RF	GBM	XGB
Vehicle Make		4%	4%	1%	1%	0%
Vehicle Age	16%	12%	12%	6%	9%	13%
Vehicle Speed	7%	8%	8%	12%	14%	13%
Vehicle Value	16%	15%	15%	17%	15%	26%
Vehicle Weight		0%	0%	15%	17%	1%
Driver Age	20%	20%	20%	10%	12%	19%
Density	11%	8%	7%	8%	12%	12%
Payment Frequency	0%	2%	2%	3%	0%	0%
Bonus Malus	12%	10%	10%	3%	5%	6%
Vehicle Fuel	12%	9%	9%	3%	1%	5%
Policy Usage		8%	8%	3%	1%	0%
Policy Duration	6%	4%	4%	5%	3%	1%
Pay as You Drive		0%	0%	0%	1%	0%
Sex		1%	1%	1%	0%	0%
Engine Cylinder		0%	0%	13%	9%	4%

Tweedie Dataset 2	GLM	Lasso	Elastic Net	RF	GBM	XGB
Vehicle Make	10%	22%	12%	1%	3%	1%
Vehicle Age	25%	21%	25%	9%	33%	36%
Vehicle Speed	2%	2%	3%	6%	2%	3%
Vehicle Value	3%	3%	2%	8%	29%	28%
Vehicle Weight		0%	0%	6%	2%	1%
Driver Age	9%	4%	5%	27%	4%	4%
Density		2%	3%	3%	12%	11%
Payment Frequency	13%	8%	9%	10%	1%	1%
Bonus Malus	12%	2%	3%	1%	9%	8%
Vehicle Fuel	5%	3%	5%	1%	0%	0%
Policy Usage	8%	3%	4%	4%	2%	0%
Policy Duration		2%	3%	16%	1%	2%
Pay as You Drive	5%	15%	12%	0%	1%	1%
Sex	3%	2%	2%	6%	0%	0%
Engine Cylinder	5%	10%	11%	3%	3%	3%

contribution in tree-based models; on the other hand, variables vehicle make, pay-as-you-drive and vehicle cylinder variables were more important in the regression models.

3.2.2 SHAP Analysis

The SHAP analysis presented in Table 2 offers additional insights into feature importance, and allows for a direct comparison of the product and Tweedie modelling frameworks.

For the product framework (Table 2, left), distinct patterns emerge. For Dataset 1 (top-left panel), the two regularised models have identical SHAP profiles. In contrast, GLM displays a different pattern, with more evenly distributed SHAP values. This differentiation was not seen in the model-based variable importance analysis of Section 3.2.1, and suggests that regularisation may enhance strong predictors for this data, in a way that is not impacted by collinearity (since the patterns for LASSO and Elastic Net are almost identical). RF and GBM yield comparable SHAP profiles, suggesting a shared interpretation of the feature space, but different from that of XGB, which again differs from a common trend seen in Section 3.2.1 across all tree-based models. Bonus-malus is relevant in the regularised models (16%) and in XGB (28%), along with at least one other predictor (density, vehicle power and vehicle gas for the linear models; vehicle and driver ages for XGB). However, for Dataset 2 (bottom-left panel), the regularised linear models yield different SHAP profiles, as do RF and GBM. In most models, the distributions of SHAP values are more spread out for Dataset 2. Density and vehicle speed are influential predictors for a number of models, but there are no clear strong predictors across all models. Noticeably, gender contributes around 10% of the risk evaluation in elastic net and GBM, which would potentially raise ethical and fairness implications.

Table 2. SHAP value metrics for Product (left) and Tweedie (right) models trained on Dataset 1 (top) and Dataset 2 (bottom), as defined in Section 2.5.6. The greyed-out cells indicate variables that were removed during the stepwise selection in the GLM model-building process. A value of 0% in the table corresponds to an SHAP value below 0.5%

Product Premium						
Dataset 1	**GLM**	**Lasso**	**Elastic Net**	**RF**	**GBM**	**XGB**
Bonus Malus	10%	16%	16%	7%	2%	28%
Region	16%	4%	4%	1%	2%	5%
Driver Age	18%	1%	1%	17%	17%	14%
Vehicle Age	22%	4%	4%	9%	5%	26%
Vehicle Brand	7%	9%	9%	17%	26%	6%
Density	6%	25%	25%	12%	8%	12%
Vehicle Power	15%	14%	14%	13%	12%	7%
Area	7%	7%	7%	14%	16%	0%
Vehicle Gas		21%	21%	10%	13%	1%

Tweedie Premium						
Dataset 1	**GLM**	**Lasso**	**Elastic Net**	**RF**	**GBM**	**XGB**
Bonus Malus	13%	44%	44%	38%	24%	37%
Region	12%	3%	3%	0%	8%	3%
Driver Age	15%	9%	9%	21%	6%	15%
Vehicle Age	12%	9%	9%	13%	5%	10%
Vehicle Brand	14%	4%	4%	0%	9%	12%
Density	7%	2%	2%	18%	11%	9%
Vehicle Power	11%	5%	5%	10%	19%	5%
Area	9%	14%	14%	0%	8%	4%
Vehicle Gas	8%	10%	10%	0%	11%	5%

Product Premium						
Dataset 2	**GLM**	**Lasso**	**Elastic Net**	**RF**	**GBM**	**XGB**
Vehicle Make		4%	10%	0%	11%	0%
Vehicle Age	4%	7%	6%	5%	11%	11%
Vehicle Speed	15%	16%	2%	14%	4%	15%
Vehicle Value		7%	6%	6%	2%	14%
Vehicle Weight		5%	10%	4%	12%	1%
Driver Age	9%	3%	7%	7%	6%	18%
Density	18%	17%	2%	10%	4%	14%
Payment Frequency	14%	10%	5%	11%	5%	0%
Bonus Malus	1%	8%	6%	17%	1%	9%
Vehicle Fuel	16%	13%	3%	5%	8%	11%
Policy Usage		7%	11%	6%	10%	0%
Tenure		0%	8%	3%	7%	3%
Pay as You Drive		3%	7%	3%	6%	0%
Gender		1%	9%	1%	12%	0%
Engine Cylinder		1%	8%	8%	1%	4%

Tweedie Premium						
Dataset 2	**GLM**	**Lasso**	**Elastic Net**	**RF**	**GBM**	**XGB**
Vehicle Make	8%	3%	3%	0%	7%	4%
Vehicle Age	7%	24%	25%	21%	12%	25%
Vehicle Speed	9%	2%	2%	5%	4%	4%
Vehicle Value	8%	7%	7%	27%	7%	24%
Vehicle Weight		0%	0%	7%	4%	1%
Driver Age	4%	13%	13%	7%	5%	8%
Density		7%	7%	13%	6%	13%
Payment Frequency	8%	6%	6%	0%	3%	3%
Bonus Malus	9%	8%	7%	12%	1%	10%
Vehicle Fuel	8%	7%	7%	0%	8%	0%
Policy Usage	11%	5%	5%	0%	8%	0%
Tenure		4%	4%	3%	9%	2%
Pay as You Drive	12%	2%	2%	0%	8%	1%
Gender	7%	3%	3%	0%	7%	0%
Engine Cylinder	8%	10%	10%	5%	7%	5%

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

For the Tweedie modelling framework (Table 2, right), we observe different SHAP-based variable rankings, and more consistency across all ML models (for both datasets), whereas SHAP values for GLM are spread more uniformly across the range of features. The fact that SHAP profiles differ so much between the product and Tweedie frameworks emphasises how the choice of loss function alters the learning dynamics, leading to a redistribution of predictive importance across features. For Dataset 1 (top-right panel), bonus-malus in particular yields high importance across all ML models. For Dataset 2 (bottom-right panel), vehicle age and vehicle value arise as important predictors. Density and bonus-malus appear with moderate importance across most models. As driver age and gender appear meaningful in a number of models, this may raise issues around pricing fairness.

3.2.3 Partial Dependence Analysis

Figure 2 presents a selection of partial dependence patterns that emerge when comparing the product and Tweedie framework predictions across models for Dataset 1. Both product and Tweedie models demonstrate a sharp increase in average prediction with increasing bonus-malus (top left), with slightly stronger gradients for Tweedie XGB and RF at higher premiums. This greater sensitivity is consistent with the SHAP-based findings. For both pricing frameworks, all models yielded comparable variations in average premiums with respect to a number of variables, such as region (bottom left). However, more variability was observed across models and/or frameworks for certain variables, such as vehicle age (top right) and density (bottom right). For instance, for vehicle age, trends differed across models, but those differences were globally consistent across product and Tweedie frameworks. For density, however, variations in average prediction appeared consistent across all models in the product framework, but different in the Tweedie framework. As these variables appeared important (see Section 3.2.1), the subset of PDP

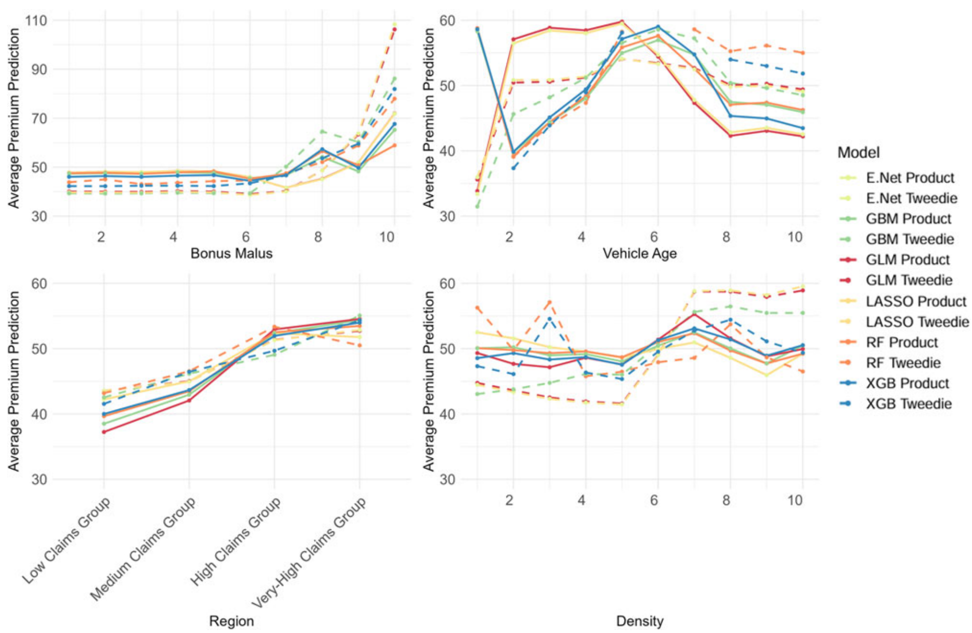

Figure 2. Partial dependence plot by feature level for categorical and discretised (by deciles) numerical variables for models trained on Dataset 1. Each subplot shows the average predicted premium across levels of a single feature, under the product (solid line) and Tweedie (dashed line) modelling strategies.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

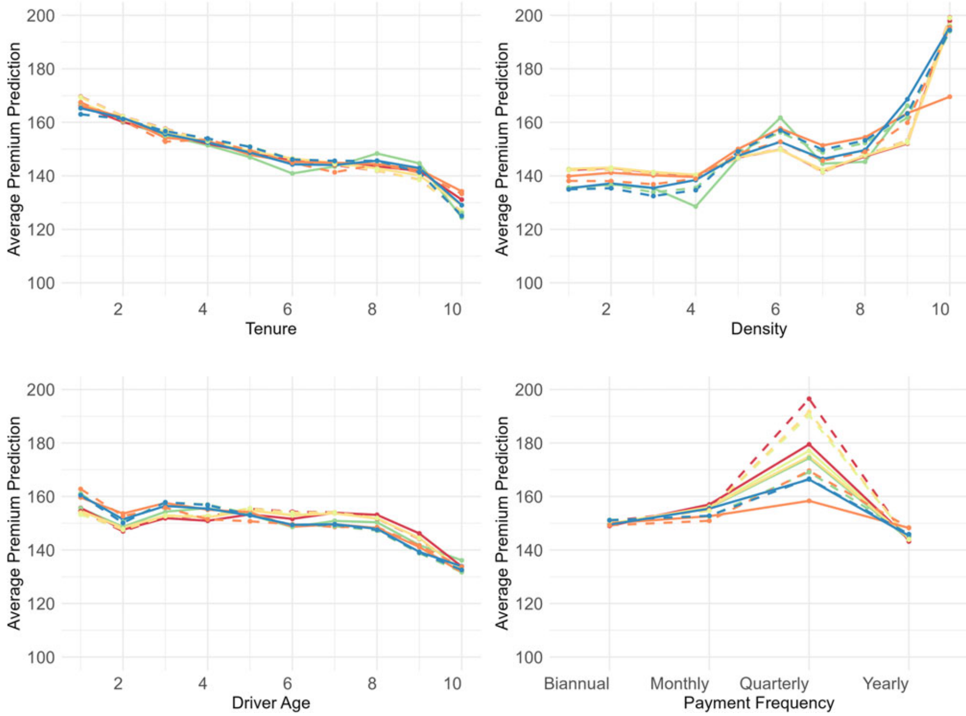

Figure 3. Partial dependence plot by feature level for categorical and discretised (by deciles) numerical variables for models trained on Dataset 2. Each subplot shows the average predicted premium across levels of a single feature, under the product (solid line) and Tweedie (dashed line) modelling strategies.

graphs shown in Figure 2 is sufficient to highlight the importance of an independent sensitivity analysis of any given model with respect to the information available for pricing. In summary, the interpretation of several variables in terms of the pricing outcome is inconsistent across models.

For Dataset 2 (Figure 3), trends were comparable across most variables, and in particular, for all numerical variables (left and top-right panels). The magnitude of the change in average premiums with respect to changes in levels for a few categorical variables showed more significant differences, although trends always had the same direction. This was the case, for example, for payment frequency (bottom-right panel).

3.3 Deviance and Bias

Figure 4 shows the bootstrap distributions of Poisson, gamma and Tweedie deviances, obtained from the test subsets for Datasets 1 and 2. Two-sample one-sided Mann–Whitney tests were carried out at the 5% significance level on the bootstrapped test deviances to assess statistical significance of relative differences in performance between any two models.

For estimation of claims frequency, XGB was found to yield the lowest Poisson deviance overall, outperforming all other models for Dataset 1 ($p < 10^{-4}$). For Dataset 2, both regularised models and the RF model performed at par with GLM (all $p > 0.05$). However, the boosting models performed significantly better than the GLM (GBM: $p < 10^{-4}$, XGB: $p = 0.0356$).

For estimation of claims severity, GLM performed at par with all other models ($p > 0.05$) for Dataset 1. For Dataset 2, regularised models performed comparably with GLM (LASSO: $p = 0.2871$, E.Net: $p = 0.2829$). However, the boosting models and RF performed significantly better than GLM (all $p < 10^{-4}$).

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

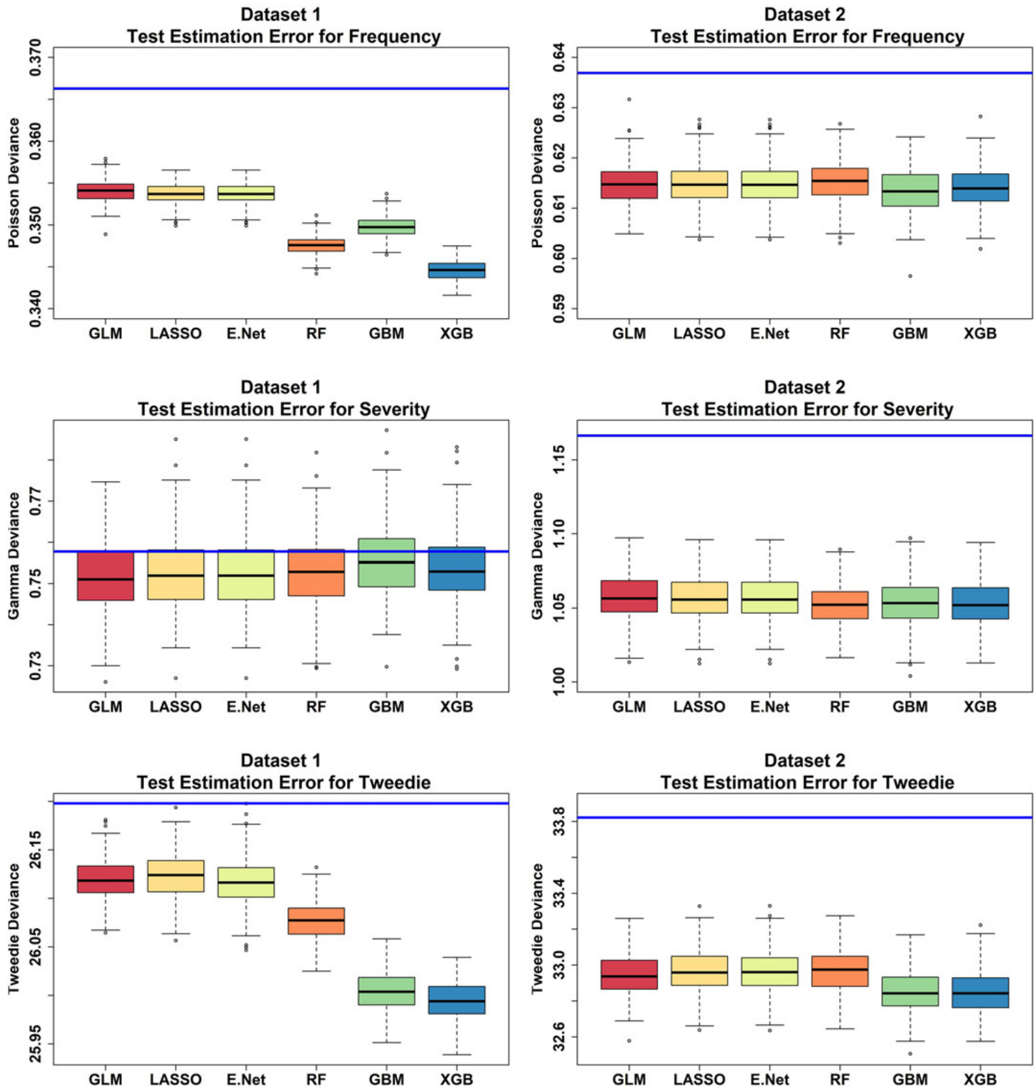

Figure 4. Bootstrap distributions of test Poisson (top), gamma (middle) and Tweedie (bottom) deviances, for Datasets 1 (left) and 2 (right). The solid blue line indicates the deviance measure from a null model (average value observed from training data).

In terms of Tweedie deviance, GLM and regularised models performed at the same level (LASSO: $p = 0.9817$ and E.Net: $p = 0.1699$). However, the tree-based models yielded a lower Tweedie deviance than GLM ($p < 10^{-6}$). Additionally, the boosting models outperformed the RF ($p < 10^{-6}$). For Dataset 2, GLM, RF and regularised models perform at par (all $p > 0.05$). However, the boosting models performed significantly better than GLM ($p < 10^{-6}$).

In all comparisons, Poisson, gamma and Tweedie deviances from GLM estimates were either significantly larger than, or comparable to those of other models. Globally, in any of these deviance comparisons, at least two tree-based methods out of RF, GBM and XGB tended to outperform models based on the linear regression principle. The only exception to the tree-based methods domination was under the product modelling strategy for Dataset 2 and for severity

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

prediction in Dataset 1. Another key finding was that with respect to all deviance measures in both datasets, none of the ML models stood out as a single optimal technique.

Based on the training dataset, we calculated a calibration factor, so that the sum of the overall predicted premium equals the actual amount of claims paid for the portfolio. The derived value of the factor was then used to calibrate the test dataset. Based on 250 bootstrapped data runs of the test samples, we calculated a measure of the portfolio-level bias by comparing the sum of predicted premiums to the actual claims amount. The confidence intervals for bootstrapped bias calculated for each model and dataset are given in Table 3. All intervals include zero, indicating an absence of significant bias, which was expected given the models were calibrated. The variation in bias values is higher in Dataset 2 than in Dataset 1 due to the relatively smaller size of Dataset 2.

Table 3. Bootstrapped confidence intervals for model bias on the predicted premium values

	Product model		Tweedie model	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
GLM	[−0.85%, 2.52%]	[−1.18%, 4.06%]	[−1.08%, 2.53%]	[−2.04%, 4.44%]
LASSO	[−0.53%, 2.69%]	[−1.21%, 4.05%]	[−0.98%, 2.62%]	[−1.59%, 4.54%]
E.Net	[−0.53%, 2.69%]	[−1.21%, 4.05%]	[−0.71%, 2.84%]	[−1.58%, 4.54%]
RF	[−1.08%, 2.52%]	[−2.03%, 3.88%]	[−1.55%, 2.22%]	[−2.00%, 4.26%]
GBM	[−0.96%, 2.47%]	[−2.45%, 6.54%]	[−1.01%, 2.65%]	[−1.59%, 4.28%]
XGB	[−0.74%, 2.60%]	[−1.67%, 3.76%]	[−0.69%, 2.70%]	[−1.47%, 4.16%]

3.4 Risk Stratification

Table 4 below provides the bootstrapped mean Gini indices (defined in Section 2.5.4) and corresponding standard errors obtained from the comparison of the machine learning (ML) methodologies with the reference GLM approach, on the test subset. Under the product modelling strategy, this comparison demonstrated that all tree-based models (RF, GBM and XGB) significantly improved policyholder risk profiling into high- and low-risk policies when compared to the GLM model ($p < 10^{-6}$).

For the product model, improvement appeared strongest from RF and XGB for Dataset 1 and from GBM for Dataset 2. This observation, similar to the results from the previous section, shows that these models performed relatively differently for the two datasets. We note that the

Table 4. Bootstrapped Gini indices (and corresponding standard errors) for comparison of risk stratification from the ML pipelines against that of the reference GLM

	Product model		Tweedie model	
	GLM (Dataset 1)	GLM (Dataset 2)	GLM (Dataset 1)	GLM (Dataset 2)
LASSO	2.5 (1.3)	1.6 (2.1)	3.3 (1.2)	1.7 (2.2)
E.Net	2.5 (1.3)	1.9 (2.1)	3.3 (1.2)	1.7 (2.2)
RF	14.4 (1.3)	10.7 (2.1)	14.2 (1.2)	5.1 (2.3)
GBM	12.0 (1.2)	12.9 (2.1)	17.4 (1.2)	6.7 (2.2)
XGB	17.5 (1.2)	9.6 (2.1)	12.4 (1.2)	6.6 (2.2)

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

regularised linear regression models (elastic net and LASSO) also yielded improvement over GLM for Dataset 1 ($p < 10^{-6}$), but not for Dataset 2 ($p = 0.2725$).

For the Tweedie model, all tree-based models (RF, GBM and XGB) significantly improved policyholder risk profiling compared to the GLM ($p < 10^{-6}$). Improvement appeared strongest from RF and GBM for Dataset 1 and from GBM for Dataset 2, which, compared to the results from the previous section, shows that these models performed relatively similarly for the two datasets. We note that the regularised linear regression models (elastic net and LASSO) also yielded improvement over GLM for Dataset 1 ($p < 10^{-6}$), but not for Dataset 2 ($p = 0.3164$). Globally, these results demonstrate that the ML methodologies led to a positive economic benefit (i.e. higher profit, assuming modelling of price elasticity capabilities are similar and ignoring customer loyalty), compared to the GLM pricing approach.

3.5 Loss Ratios and Conversion Rates in Open-Market Competition

Market competition is a key consideration in the pricing process. We build upon the work of Li et al. (2021) to develop a dynamic pricing game. The game compares pricing strategies based on LRs. Multiple insurers (models) compete to sell insurance contracts based on their predicted premiums. In this setup, we construct a zero-sum game entirely based on premiums. Figure 5 shows the bootstrap distributions of total (aggregated) actual premiums, total claims and LRs for each dataset in a direct two-player competition between the GLM, XGB and RF methodologies.

The LR charts in Figure 5 for GLM versus XGB (third row) and GLM versus RF (bottom) show a similar result when GLM is compared to these tree-based models. However, the difference in LRs varies in intensity when comparing GLM to XGB versus GLM to RF. This demonstrates that the RF and XGB models compete with GLM in different ways. The top and second row charts show that the comparisons between models yield different findings between Datasets 1 and 2.

For Dataset 1, under the product model strategy, when comparing GLM and XGB, the XGB tended to yield lower premiums for a larger book of business. The lower premium prediction resulted in XGB securing a higher amount of total premium, with the XGB median total premium sitting higher at around € 2,700k compared to the GLM median at around € 2,100k. However, the distributions of total claims payable were comparable for both models. This led to a lower ratio and more profitable business when using an XGB pricing model compared to a GLM model. This aligns with the results of the risk stratification analysis, where the XGB model was found to improve risk profiling over GLM. To make the comparison fair, models were calibrated to quote (on average) the same premium for the considered group of policies.

For Dataset 1 again, under the Tweedie model strategy, the trend was different, with GLM predicting lower individual premiums (not shown in the figure) compared to XGB, which resulted in the XGB securing a lower amount of business, i.e. a lower total book premium, as shown in Figure 5, top left. However, the business secured with XGB tended to attain a lower LR because of comparably lower overall claims payable. This result suggests that XGB yields a lower LR (risk-averse) pricing, which leads to a comparatively lower overall book premium while still retaining higher profits.

In Dataset 2, the distributions of LRs between GLM, XGB and RF are comparable. Hence, there is no clear indication that ML models outperform GLM in terms of LR. This can be driven by the relative difference in size and type of policies between Datasets 1 and 2. Here, the results from the product and Tweedie model strategies are consistent. Comparing GLM and XGB models in an open-market scenario also indicated improvement in the distribution of XGB LRs. However, there were differences in the type and amount of business the two models earned. For Dataset 2 (under product model), XGB earned a lower amount of total premium (with medians of € 830k and € 760k for GLM and XGB, respectively), but with significantly lower amounts of claims payable (with medians of € 900k and € 790k for GLM and XGB,

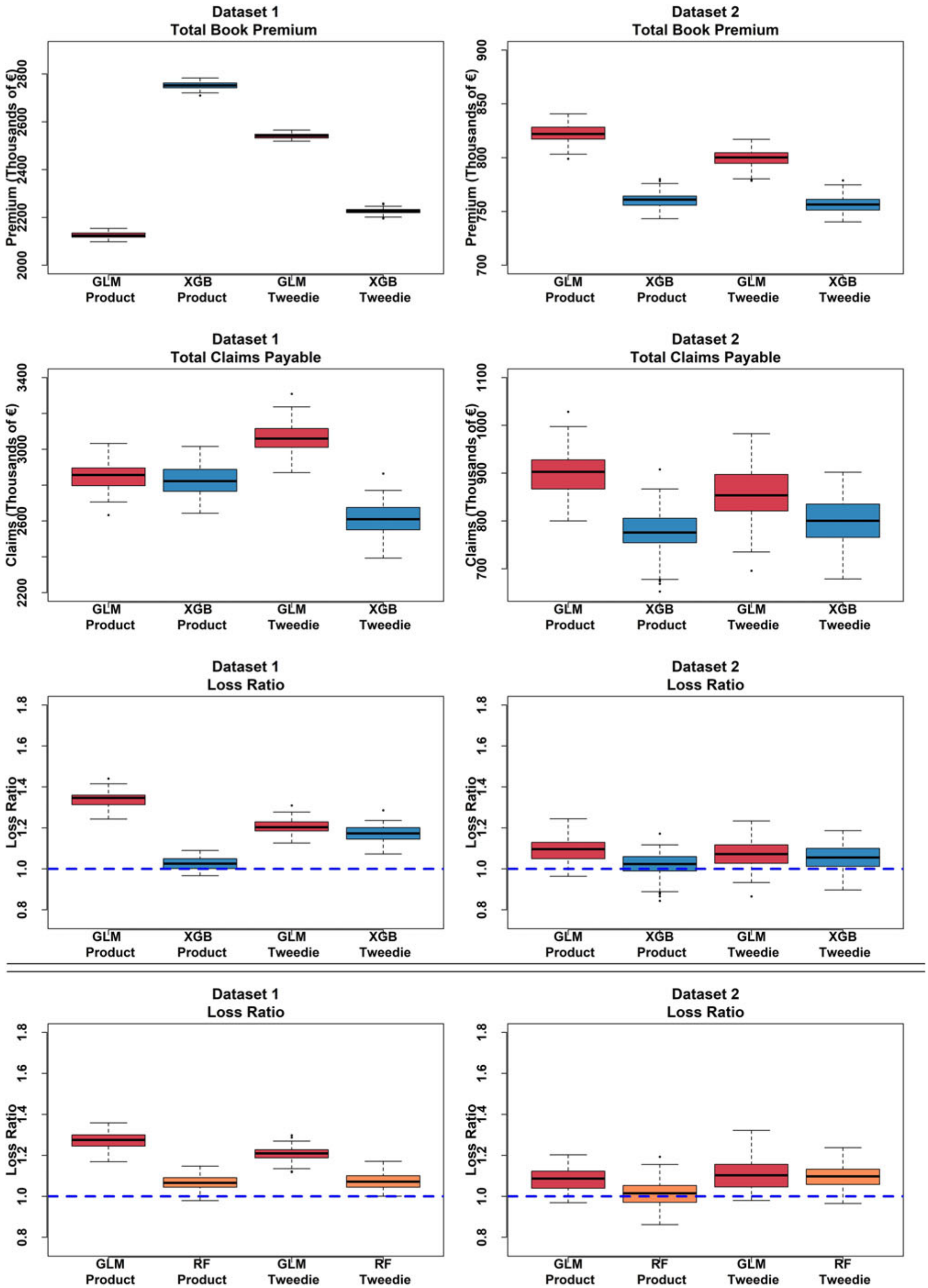

Figure 5. Bootstrap distributions of total book premiums (top), total claims (2^{nd} row) and loss ratios (3^{rd} row) comparison between two models, using, respectively, a GLM and an XGB methodology for premium pricing, on Dataset 1 (left) and Dataset 2 (right). The dashed line indicates a loss ratio of 1. The same analysis was carried out to compare RF with GLM, with loss ratio results (bottom) consistent with the comparison between XGB and GLM.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

respectively). Although XGB led to writing a lower number of policies compared to RF, these policies had a lower LR.

3.6 Overall Comparison

Figure 6 captures the essential metrics (namely bias, Poisson deviance, gamma deviance, Tweedie deviance, Gini index and conversion rate) in the form of radar charts to summarise the previous analyses and allow for overall model comparison on Datasets 1 and 2. For Dataset 2, additional metrics (calibration, actuarial group fairness and demographic parity) are shown on the radar chart that help measure model fairness. These metrics are described in more detail in Section 4 on group fairness.

In this analysis, the scales were calibrated independently for each dataset and key performance indicator. This calibration ensured that in each chart, the minimum and maximum values achieved among all models for a given metric were represented as the minimum and maximum achieved for that metric. A maximum value in any metric indicates comparably superior, i.e. preferable, model performance. Since the elastic net and LASSO yielded comparable performance, only the LASSO is represented in these summaries.

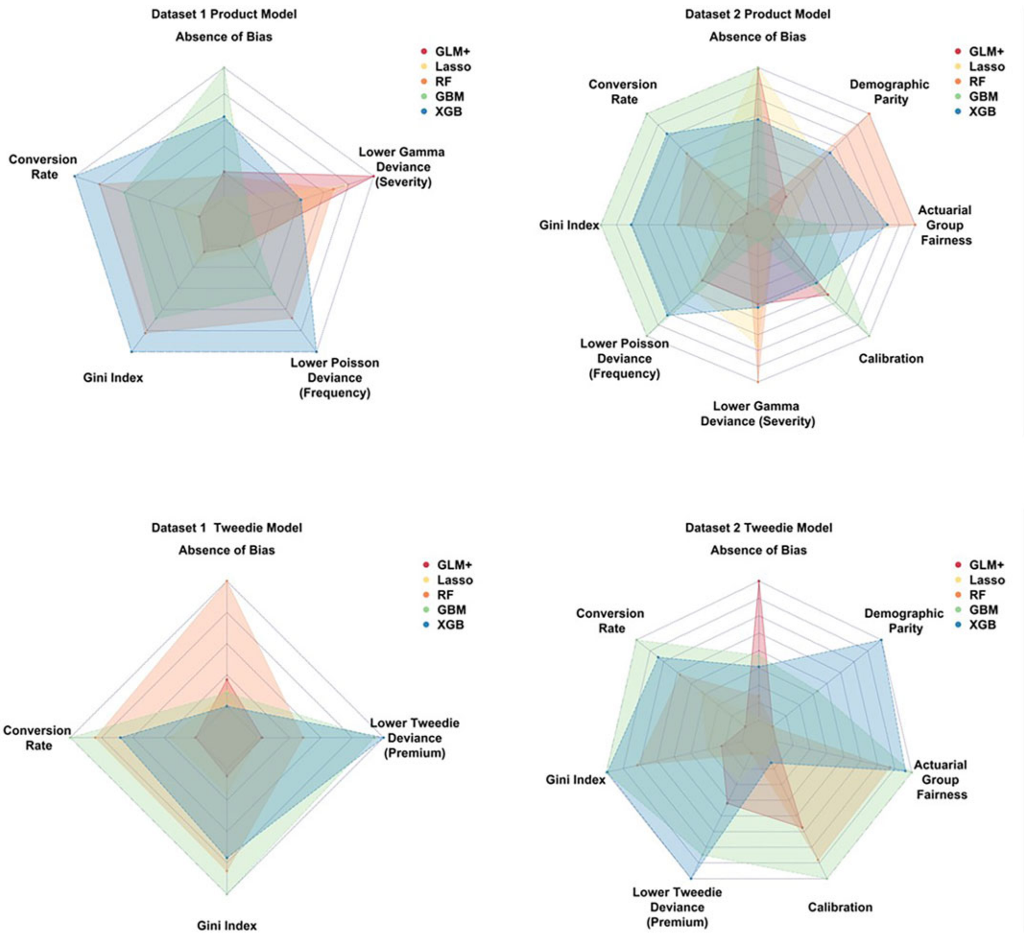

Figure 6. Radar charts comparing the reference GLM with LASSO and the tree-based models on all key performance indicators.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

For Dataset 1, the overall footprints of the RF and XGB models appear larger than those of other models. The tree-based models led to the most competitive frequency, severity and Tweedie premium estimation performance for both datasets. Under the product model, for Dataset 1, XGB also tended to be the optimal model in terms of the Gini index and conversion rate. In contrast, GBM yielded the lowest bias on the overall portfolio level but appeared inferior for risk stratification (measured by Gini index). Under the Tweedie model, for Dataset 1, GBM was the optimal model in terms of the Gini index and conversion rate. In contrast, XGB yielded the lowest deviance on the overall portfolio-level prediction.

For Dataset 2, the GBM and XGB appeared as the two most competitive models for most metrics except for gamma deviance (product) and bias (Tweedie). Every time, GBM yielded the most effective risk stratification and conversion rate and, hence, was the most profitable model. (Recall that these relative differences were found to be statistically significant in previous sections.) Overall, GLM overall appeared weaker than all other methods for most criteria.

From the fairness aspect, GBM was the optimal model when calibration was used as the fairness measure. For demographic parity and actuarial group fairness RF (under the product model) and XGB/GBM (under the Tweedie model) were optimal. These measures are discussed in more detail in the next section.

4. Analysis of Fairness

We now assess model fairness, specifically in Section 4.1 in terms of the three axioms defined in Section 2.5.9. As there were no protected variables in Dataset 1, we focus on Dataset 2 in this section. For the sake of presentation clarity, we focus on a comparison between XGB and GLM models, the former being one of the better performing model across our earlier benchmarks and of particular interest to the industry at the moment (Colella & Jones, 2023; Dolman & Semenovich, 2018; Ferrario & Hämmerli, 2019; Henckaerts et al., 2021; Kuo & Lupton, 2020). The analysis aims to assess quantitatively if any one cohort of policyholders is subsidising for any other cohort. Furthermore, for any segment of the public, the analysis aims to assess if the AP charged to policyholders is in line with their expected cost of claim.

4.1 Group Fairness, Demographic Parity and Calibration

Figures in each section below present outputs specific to the evaluation of the three axioms defined in Section 2.5.9.

4.1.1 Demographic Parity (Axiom 1)

A comparison of test set distributions of actual premium $\hat{P}_A$ with respect to protected variable gender Z_{gender} showed that the distributions from the XGB model ($\mathbb{P}(\hat{P}_{A_{XGB}}|Z_{gender} = male)$ and $\mathbb{P}(\hat{P}_{A_{XGB}}|Z_{gender} = female)$) were comparable for both product (Figure 7, left) and Tweedie (Figure 7, right) modelling strategies, unlike the corresponding GLM model. In addition, the Wasserstein distance d_{gender} defined similar to equation (7) by

$$d_{gender} = W\left(\hat{P}_A \middle| Z_{gender} = male, \hat{P}_A \middle| Z_{gender} = female\right)$$

calculated for the test set distribution were 20.0 and 19.1, respectively, for GLM and XGB under the product modelling, and 20.2 and 15.5 under the Tweedie modelling strategies. Table 5 gives the Wasserstein distances, the differences in distribution medians, and the p-values of two-sided Mann–Whitney (Wilcoxon) tests for all the models used in the study. Based on the test outcomes reported in Table 5, we can conclude that the actual premium distributions differed significantly between males and females.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

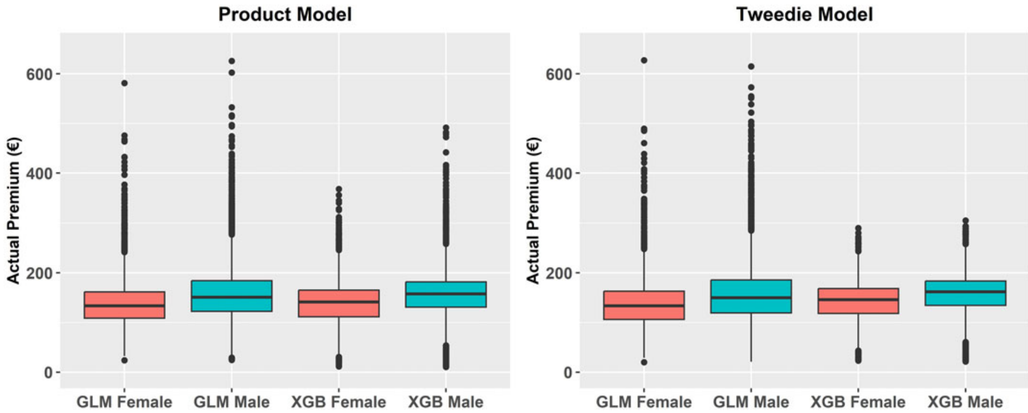

Figure 7. Distributions of the GLM and XGB model actual premium, by gender, under the Product (left) and Tweedie (right) model strategies.

Table 5. Measure of the distance (d_{gender}), difference in distribution medians and Wilcoxon test p-values of the actual premium models under the product model (left) and Tweedie model (right) strategies

Model	Product			Tweedie		
Measure	Distance	Median difference	p-value	Distance	Median difference	p-value
GLM	20.0	17.3	<0.0001	20.2	16.4	<0.0001
LASSO	19.3	17.0	<0.0001	20.2	16.2	<0.0001
E.Net	19.3	17.1	<0.0001	20.2	16.4	<0.0001
RF	18.3	17.2	<0.0001	20.2	15.3	<0.0001
GBM	20.3	17.0	<0.0001	18.3	18.5	<0.0001
XGB	19.1	16.6	<0.0001	15.5	15.9	<0.0001

4.1.2 Actuarial Fairness (Axiom 2)

A comparison of test data distributions of actual premiums ($\mathbb{P}(\hat{P}_A|\hat{P}_T, Z_{gender})$) conditional on predicted TP bands (defined by TP quintiles as in equation (8)) and on gender is provided in Figure 8. The results showed greater alignment between distributions of XGB-derived ratios for male, female and overall drivers, i.e. greater alignment with Axiom 2, compared to those from GLM. Discrepancies $d_{l,\mathcal{Y}}$ defined by equation (9) were calculated for the five TP bands successively, as

$$d_{l,P_T} = W\left(\hat{P}_A \middle| \hat{P}_T, Z_{gender} = male, \hat{P}_A \middle| \hat{P}_T, Z_{gender} = female\right)$$

The Wasserstein distances for the GLM and XGB models for the product and Tweedie modelling strategies are given in Tables 6 and 7, respectively. This quantitation also indicates that the distance between the XGB distributions of both groups were closer compared to the same distributions from the GLM model.

4.1.3 Calibration (Axiom 3)

A comparison of actual claim distributions conditional on actual premium bands ($\mathbb{P}(L|\hat{P}_A, Z_{gender})$) with respect to the protected variable gender is provided in Figure 9. Figure 9 does not demonstrate overall large discrepancies across gender groups from GLM-based pricing,

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

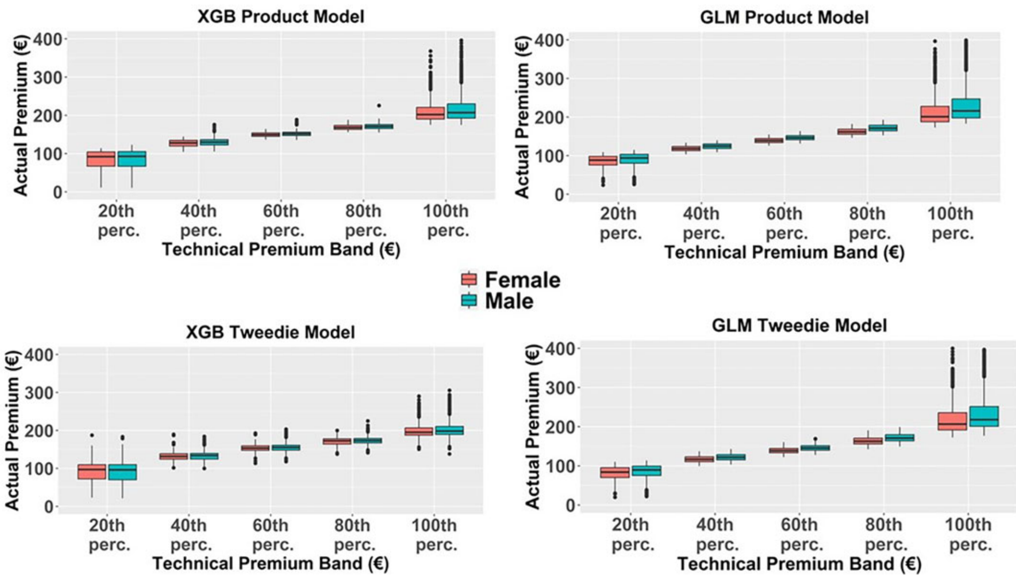

Figure 8. Distributions of actual premium across technical premium quintile bands for the XGB model (left) and the GLM model (right) under the product (top) and Tweedie (bottom) modelling strategies.

for the majority of the data. Quantitatively, $d_{L,Z}$ defined by equation (10) were evaluated for each actual premium band defined as in equation (8), but using the AP, as follows:

$$d_{L,P_A} = W\left(L\middle|\hat{P}_A, Z_{gender} = male, L\middle|\hat{P}_A, Z_{gender} = female\right)$$

These discrepancies between male and female groups for XGB and GLM actual claims by actual premium bands for the product and Tweedie modelling strategies are also given in Tables 6 and 7, respectively.

4.1.4 APTP Analysis

APTP ratios were compared between male and female groups by actual premium bands for the product and Tweedie modelling strategies. These are shown in Tables 6 (product model) and 7 (Tweedie model). The difference between the male and female APTP ratios (ρ) is greater for the GLM and regularised models compared to the tree-based models. This may partially be due to tree-models being less reliant on the gender factor as compared to the GLM models (Table 1). Furthermore, all models were found to overcharge males ($\rho > 1$) compared to females ($\rho < 1$). Axioms 1 and 2 measure the distance between the levels of the protected characteristics, however they do not provide a directional sense if this difference can be attributed to under- or over-pricing. The analysis of APTP ratios provide this complementary insight, and can be used to detect and quantify subsidisation across protected groups. In this case study, overall, males tend to cross-subsidise females.

5. Discussion

5.1 Summary of Findings

While the use of ML models in non-life insurance pricing has been extensively researched, their use remains hindered because of their untransparent nature. The aim of this research was to compare the traditional GLM with a number of common ML models under a comprehensive range of key criteria, including bias, accuracy, risk differentiation, competitiveness, LRs,

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

Table 6. For the Product model, Wasserstein distances, median difference and Wilcoxon test *p*-values comparing actual premium distributions by technical premium band for actuarial group fairness (left) and calibration (right), as well as mean APTP ratios and associated standard deviation by actual premium band (centre). Significant *p*-values are italicised

		Actuarial Group Fairness			APTP Values (SD)		Calibration		
Percentile	**Model**	**Distance**	**Median Difference**	**p-value**	**Male**	**Female**	**Distance**	**Median Difference**	**p-value**
20th Percentile	GLM Product	5.3	5.6	*<0.0001*	1.03 (0.02)	0.98 (0.02)	16.5	-10.6	0.5100
40th Percentile	GLM Product	6.2	6.7	*<0.0001*	1.02 (0.02)	0.98 (0.02)	14.4	15.0	0.4900
60th Percentile	GLM Product	7.5	7.2	*<0.0001*	1.02 (0.02)	0.97 (0.02)	24.1	-8.3	0.4200
80th Percentile	GLM Product	9.3	9.4	*<0.0001*	1.02 (0.02)	0.97 (0.01)	38.5	-18.5	0.0700
100th Percentile	GLM Product	19.4	15.7	*<0.0001*	1.01 (0.02)	0.97 (0.02)	25.4	17.1	0.7800
Average	**GLM Product**	**9.5**	**8.9**		**1.02 (0.02)**	**0.97 (0.02)**	**23.8**	**-1.0**	
20th Percentile	Lasso Product	4.7	5.1	*<0.0001*	1.03 (0.03)	0.98 (0.02)	17.7	-8.1	0.4200
40th Percentile	Lasso Product	5.3	5.7	*<0.0001*	1.02 (0.03)	0.98 (0.02)	15.3	15.6	0.3400
60th Percentile	Lasso Product	6.5	6.4	*<0.0001*	1.02 (0.02)	0.98 (0.02)	30.3	-9.0	0.3200
80th Percentile	Lasso Product	8.1	8.1	*<0.0001*	1.02 (0.02)	0.97 (0.02)	38.1	-24.1	*0.0300*
100th Percentile	Lasso Product	18.4	13.5	*<0.0001*	1.01 (0.02)	0.97 (0.02)	31.4	19.5	0.4700
Average	**Lasso Product**	**8.6**	**7.8**		**1.02 (0.02)**	**0.98 (0.02)**	**26.6**	**-1.2**	
20th Percentile	E.Net Product	4.8	5.0	*<0.0001*	1.03 (0.03)	0.98 (0.02)	17.1	-7.8	0.5600
40th Percentile	E.Net Product	5.3	5.7	*<0.0001*	1.02 (0.03)	0.98 (0.02)	16.0	13.6	0.4300
60th Percentile	E.Net Product	6.4	6.3	*<0.0001*	1.02 (0.03)	0.98 (0.02)	33.1	-3.2	0.3700
80th Percentile	E.Net Product	8.0	8.1	*<0.0001*	1.02 (0.03)	0.97 (0.02)	39.0	-30.5	*0.0400*
100th Percentile	E.Net Product	18.1	13.8	*<0.0001*	1.01 (0.02)	0.97 (0.02)	27.7	21.7	0.7600
Average	**E.Net Product**	**8.5**	**7.8**		**1.02 (0.03)**	**0.98 (0.02)**	**26.6**	**-1.2**	
20th Percentile	RF Product	4.1	4.7	*0.0001*	0.98 (0.08)	0.97 (0.08)	12.1	-4.8	0.5500
40th Percentile	RF Product	1.3	1.1	*0.0053*	1.00 (0.04)	1.00 (0.05)	23.3	4.2	0.2900
60th Percentile	RF Product	1.0	1.0	*0.0072*	1.01 (0.04)	1.00 (0.04)	11.8	-17.1	0.7400
80th Percentile	RF Product	0.4	0.4	0.9904	1.01 (0.04)	1.01 (0.04)	64.8	-31.8	0.2000
100th Percentile	RF Product	3.0	1.8	*0.0407*	1.02 (0.04)	1.03 (0.04)	20.1	-4.3	0.8800
Average	**RF Product**	**2.0**	**1.8**		**1.00 (0.05)**	**1.00 (0.05)**	**26.4**	**-10.8**	
20th Percentile	GBM Product	3.9	3.8	*0.0004*	1.02 (0.07)	0.99 (0.05)	17.5	-7.6	0.7400
40th Percentile	GBM Product	5.2	4.2	*<0.0001*	1.03 (0.06)	0.99 (0.03)	13.1	-1.9	0.5400
60th Percentile	GBM Product	4.2	3.7	*<0.0001*	1.01 (0.05)	0.98 (0.03)	33.7	-37.1	0.4600
80th Percentile	GBM Product	5.6	5.1	*<0.0001*	1.01 (0.04)	0.98 (0.03)	15.6	3.5	0.9400
100th Percentile	GBM Product	15.0	9.4	*<0.0001*	1.01 (0.04)	0.98 (0.04)	29.1	-16.4	0.2500
Average	**GBM Product**	**6.8**	**5.2**		**1.02 (0.05)**	**0.98 (0.04)**	**21.8**	**-11.9**	
20th Percentile	XGB Product	1.7	0.8	0.2700	1.00 (0.05)	0.99 (0.03)	17.3	-9.5	0.1900
40th Percentile	XGB Product	2.3	2.0	*<0.0001*	1.01 (0.03)	0.99 (0.01)	15.9	0.4	0.9800
60th Percentile	XGB Product	2.1	2.4	*<0.0001*	1.00 (0.02)	0.99 (0.01)	22.8	-42.8	0.7900
80th Percentile	XGB Product	2.3	2.7	*<0.0001*	1.00 (0.01)	0.99 (0.01)	24.9	8.4	0.8200
100th Percentile	XGB Product	8.9	5.5	*<0.0001*	1.00 (0.01)	1.00 (0.01)	40.7	-13.7	0.1200
Average	**XGB Product**	**3.5**	**2.7**		**1.00 (0.02)**	**0.99 (0.01)**	**24.3**	**-11.5**	

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

Table 7. For the Tweedie model, Wasserstein distances, median difference and Wilcoxon test *p*-values comparing actual premium distributions by technical premium band for actuarial group fairness (left) and calibration (right), as well as mean APTP ratios and corresponding standard deviation by actual premium band (centre). Significant *p*-values are italicised

		Actuarial Group Fairness			APTP Values (SD)		Calibration		
Percentile	**Model**	**Distance**	**Median Difference**	**p-value**	**Male**	**Female**	**Distance**	**Mean Difference**	**p-value**
20th Percentile	GLM Tweedie	4.8	5.6	*<0.0001*	1.03 (0.03)	0.99 (0.03)	12.6	-10.6	0.8700
40th Percentile	GLM Tweedie	5.3	5.4	*<0.0001*	1.02 (0.03)	0.98 (0.03)	19.3	15.0	0.9100
60th Percentile	GLM Tweedie	6.4	6.6	*<0.0001*	1.02 (0.03)	0.98 (0.03)	14.5	-8.3	0.7200
80th Percentile	GLM Tweedie	8.1	8.1	*<0.0001*	1.02 (0.03)	0.97 (0.03)	38.8	-18.5	0.0500
100th Percentile	GLM Tweedie	19.0	12.5	*<0.0001*	1.01 (0.03)	0.97 (0.03)	30.8	17.1	0.6900
Average	**GLM Tweedie**	**8.7**	**7.6**		**1.02 (0.03)**	**0.98 (0.03)**	**23.2**	**-1.0**	
20th Percentile	Lasso Tweedie	4.7	5.5	*<0.0001*	1.03 (0.03)	0.99 (0.03)	12.9	-8.1	0.8200
40th Percentile	Lasso Tweedie	4.9	5.3	*<0.0001*	1.02 (0.03)	0.98 (0.03)	20.8	15.6	0.9300
60th Percentile	Lasso Tweedie	6.2	6.5	*<0.0001*	1.02 (0.03)	0.98 (0.03)	16.8	-9.0	0.7200
80th Percentile	Lasso Tweedie	8.3	8.2	*<0.0001*	1.02 (0.03)	0.97 (0.03)	36.3	-24.1	0.0600
100th Percentile	Lasso Tweedie	19.4	13.8	*<0.0001*	1.01 (0.03)	0.97 (0.03)	39.4	19.5	0.5100
Average	**Lasso Tweedie**	**8.7**	**7.9**		**1.02 (0.03)**	**0.98 (0.03)**	**25.2**	**-1.2**	
20th Percentile	E.Net Tweedie	4.7	5.4	*<0.0001*	1.03 (0.03)	0.99 (0.03)	12.1	-7.8	0.8500
40th Percentile	E.Net Tweedie	5.1	5.1	*<0.0001*	1.02 (0.03)	0.98 (0.03)	22.6	13.6	0.9000
60th Percentile	E.Net Tweedie	6.2	6.3	*<0.0001*	1.02 (0.03)	0.98 (0.03)	16.0	-3.2	0.7200
80th Percentile	E.Net Tweedie	8.1	8.1	*<0.0001*	1.02 (0.03)	0.97 (0.03)	33.2	-30.5	0.0800
100th Percentile	E.Net Tweedie	19.3	13.9	*<0.0001*	1.01 (0.03)	0.97 (0.03)	28.5	21.7	0.7400
Average	**E.Net Tweedie**	**8.7**	**7.7**		**1.02 (0.03)**	**0.98 (0.03)**	**22.5**	**-1.2**	
20th Percentile	RF Tweedie	2.9	0.1	0.3854	0.99 (0.02)	0.98 (0.02)	8.7	-4.8	0.9300
40th Percentile	RF Tweedie	1.1	1.4	*<0.0001*	1.00 (0.02)	1.00 (0.02)	27.9	4.2	0.6200
60th Percentile	RF Tweedie	1.2	1.5	*<0.0001*	1.00 (0.02)	0.99 (0.02)	19.1	-17.1	0.4900
80th Percentile	RF Tweedie	0.5	0.5	0.3929	1.00 (0.03)	1.00 (0.03)	20.0	-31.8	0.8000
100th Percentile	RF Tweedie	8.5	6.2	*0.0093*	1.01 (0.07)	1.01 (0.07)	35.7	-4.3	0.1400
Average	**RF Tweedie**	**2.9**	**1.9**		**1.00 (0.03)**	**1.00 (0.03)**	**22.3**	**-10.8**	
20th Percentile	GBM Tweedie	1.2	-2.0	0.6793	1.00 (0.04)	1.00 (0.03)	9.6	-7.6	0.5100
40th Percentile	GBM Tweedie	1.2	1.6	*0.0037*	1.00 (0.03)	1.00 (0.02)	10.7	-1.9	0.6100
60th Percentile	GBM Tweedie	0.7	0.1	0.1470	1.00 (0.02)	1.00 (0.01)	37.6	-37.1	0.1900
80th Percentile	GBM Tweedie	0.8	1.8	*0.0241*	1.00 (0.02)	1.00 (0.01)	19.2	3.5	0.8800
100th Percentile	GBM Tweedie	4.7	2.5	*0.0205*	1.00 (0.01)	1.00 (0.01)	31.8	-16.4	0.2500
Average	**GBM Tweedie**	**1.7**	**0.8**		**1.00 (0.02)**	**1.00 (0.02)**	**21.8**	**-11.9**	
20th Percentile	XGB Tweedie	1.8	-0.9	0.5716	1.13 (0.21)	1.12 (0.15)	13.5	-9.5	0.3800
40th Percentile	XGB Tweedie	1.4	2.2	*0.0294*	1.04 (0.06)	1.04 (0.05)	9.7	0.4	0.4600
60th Percentile	XGB Tweedie	1.7	2.0	*<0.0001*	1.02 (0.05)	1.02 (0.05)	29.9	-42.8	0.3900
80th Percentile	XGB Tweedie	1.8	0.9	*<0.0001*	1.00 (0.04)	0.99 (0.04)	29.7	8.4	0.7700
100th Percentile	XGB Tweedie	3.6	3.1	*0.0001*	0.94 (0.06)	0.94 (0.05)	42.0	-13.7	0.1900
Average	**XGB Tweedie**	**2.1**	**1.5**		**1.03 (0.08)**	**1.02 (0.07)**	**25.0**	**-11.5**	

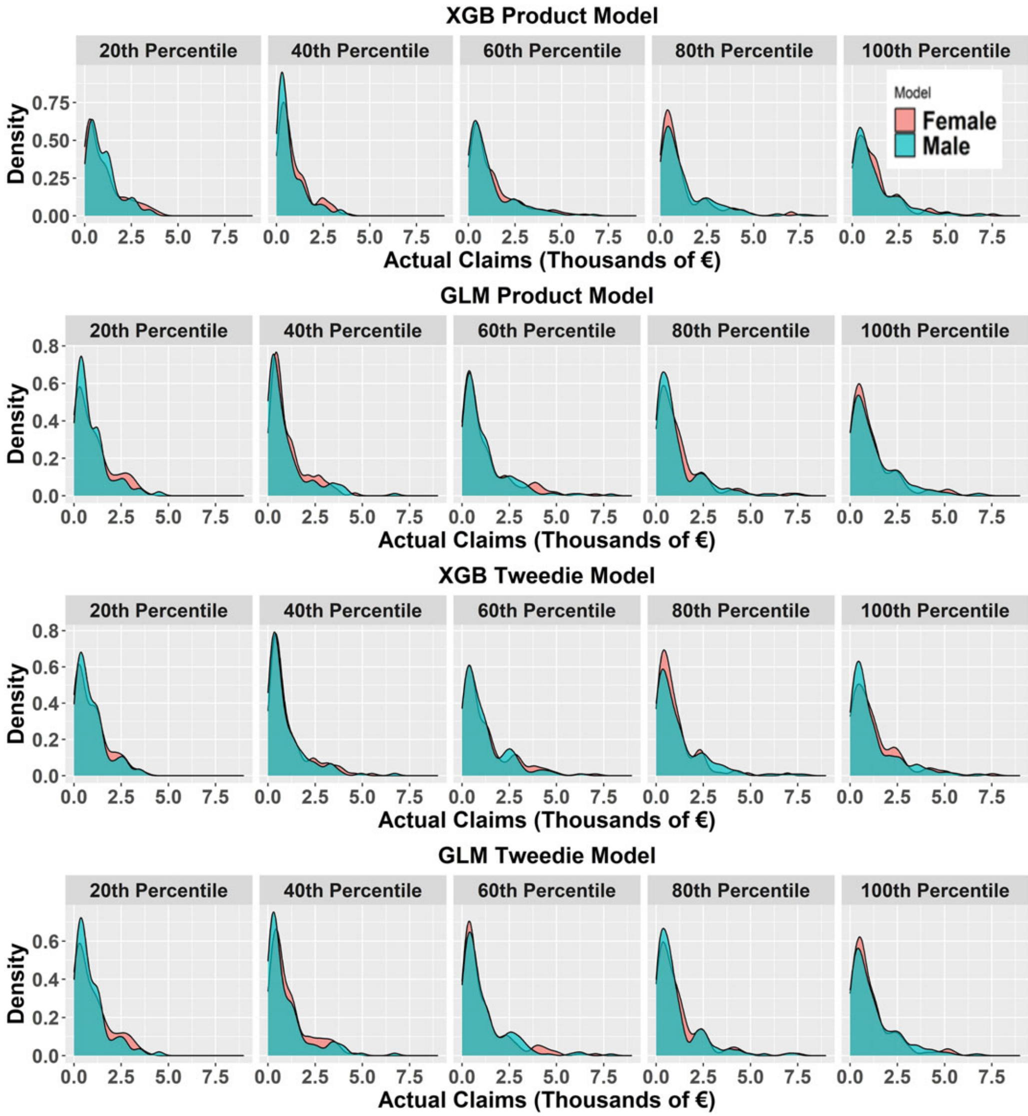

Figure 9. Distributions of actual claims by actual premium quintile band for the XGB and GLM models under the Product and Tweedie modelling strategies.

discrimination and fairness. Pricing performance and fairness were assessed on the same samples of premium estimates obtained from each model, so we were able to evaluate whether any of these pricing models could be accurate and fair simultaneously.

The results highlighted that ML methodologies should be considered as a realistic and reasonable alternative to the traditional GLM approach. While no single ML model outperformed the others across all metrics, the GLM underperformed for most. For instance, GLM was one of the worst-performing models in predicting both frequency and severity of claims and was significantly outperformed by all tree-based models. Additionally, the tree-based models significantly outperformed the GLM in terms of risk differentiation.

Interestingly, for Dataset 2, a conversion rate analysis identified that the XGB model resulted in lower overall premium income (compared to GLM), yet the smaller claim payouts associated with

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

these policies resulted in a lower overall LR (or higher overall profit). The same was not observed in Dataset 1 (under the product model), underscoring the requirement of a comprehensive, methodical analysis to determine whether any ML model may be deemed a reasonable alternative. We demonstrated that ML models may include a systematic pricing bias for one cohort of policyholders compared to another. This tendency is not unique to ML models, and similar characteristics are common in traditional models such as GLM (Kuo & Lupton, 2020). Variable importance analysis highlighted that the six models exploited information differently from both datasets. Broadly, tree-based models (RF, GBM, XGB) extracted the information in a similar manner when compared to regression-based models (GLM, LASSO, elastic net) particularly in terms of claims frequency. However, no two models used the data in the exact same way.

With the increasing availability of pricing data and the fact that non-life insurance companies are progressively using ML-based models to calculate (Blier-Wong et al., 2020), issues of fairness and discrimination might arise (Kleinberg et al., 2016). We devised a robust quantitative framework to evaluate fairness in regression settings on the basis of three complementary criteria: demographic parity, group fairness and calibration. We proposed a set of specific metrics to evaluate each of these criteria, enabling a numerical assessment and direct comparison of fairness between ML models and the GLM.

The results of the proposed benchmark underscore the value of analysing actual and TPs to detect instances of price discrimination irregularities in any potential pricing framework. XGB-based premiums proved to be fairer than GLM-based premiums in terms of demographic parity with respect to gender. The equitable separation of risk by TP bands and protected groups was also evaluated with respect to gender. The XGB model was found to outperform the GLM. Lastly, the assessment of premium calibration against associated losses, with respect to (gender) group membership, was challenging for all models.

5.2 Limitations and Future Work

In the prevalent non-life insurance market practices, discrimination-free pricing can be difficult to achieve due to practical considerations such as data limitations, regulatory constraints around data gathering on sensitive factors, and the trade-off between different aspects of fairness and predictability. The findings of this study are limited by the characteristics of the datasets used, and cannot be generalised for any of the ML models. The performance of the ML models remains dependent of the data. To mitigate this limitation a systematic approach was taken in this study. The availability of open-access datasets with protected variables remains scarce. In practice, modellers commonly use claims data that is broken down at a peril level, however the datasets used in this study have claims available only at an aggregate level. There are practical differences in how each country implements EU regulations or regulates the use of particular variables. We used two popular French datasets here, and the definition of what is a protected variable differs across countries (e.g. bonus-malus in France *vs* no claims discount in Ireland).

The datasets we used did not include the original actual (or technical) premiums charged to the policyholders. Hence, we could not compare our premium estimates against the actual premiums charged. It is pertinent to note that original actual premium charged would be biased towards a particular pricing methodology. Most likely, the charge premiums would have been obtained from GLM-based pricing and may thus not have been a suitable benchmark. We carried out the analysis over one complete year of data. Further analysis may be performed over multiple years to see how the premium for the same policyholder can vary over time with changes in policyholder risk characteristics.

This study does not make any assumptions on price elasticity across insurers to simplify the market simulation and highlight core competitive dynamics. We modelled a new business environment where price-sensitive customers choose the lowest premium, aligning with rational consumer behaviour. Future work could explore this direction further by incorporating elasticity modelling.

A number of questions around data pre-processing arose during the development of this study. For example, if further analyses of the impact of exposure threshold on model performance should be conducted with Dataset 1, as the exposure threshold should align with modelling decisions made in practical settings. The way missing or outlying values should be handled was debated as this decision could have impacted model performance in terms of overall or group-specific bias. Different choices of model calibration strategies may impact the benchmark (in particular it could affect the bias, deviances and APTP ratios). For instance, the Tweedie distribution is also known to under-evaluate overdispersion, leading to biased representation of the underlying data (Hilbe, 2011; Smyth, 1996). Autocalibration and approximation strategies have been considered recently to mitigate this model bias (Becker et al., 2022; Denuit et al., 2021).

It is recommended that in the future a similar study may be carried out to assess the impact of the choice of pricing model on retention modelling and on claims reserving processes. Concern for fairness is not an issue for reserving; however, future works should assess changes in pricing performance that may have a significant impact on a particular reserving strategy, depending on the models used.

While the fairness metrics used in this study are commonly used in pricing fairness, they can be mathematically incompatible under realistic conditions (Chouldechova, 2017; Kleinberg et al., 2016). In particular, a model cannot simultaneously satisfy both calibration and demographic parity unless risk characteristics are equal across groups, which is rarely the case in insurance contexts. This implies that firms must make decisions about which fairness criterion to prioritise, depending on their regulatory obligations, market positioning and ethical commitments. For example, a regulator may favour demographic parity to ensure equal approval or pricing rates across protected groups, while an insurer may prioritise calibration to maintain pricing accuracy and solvency. One could aim to define a fairness-accuracy frontier, in order to allow firms to quantify the impact of choosing one fairness criterion over another on business objectives such as LRs or customer retention. This could be achieved with a suitable optimisation framework, to balance these conflicting fairness goals within a pricing process. A potential technical development could be the introduction of a penalty to the regression cost function used to fit the premium models, in order to achieve a suitable trade-off between desired fairness metrics and other important pricing performance metrics.

Future work could explore the applicability of pricing models in different jurisdictions to understand how regional variations in data and market conditions affect model performance. Additionally, the use of synthetic datasets could be investigated to supplement real-world data, thereby addressing limitations in data availability and enhancing the robustness of the models. Expanding the analysis beyond the current two datasets would provide a more comprehensive evaluation and contribute to the development of more reliable and versatile pricing models.

The scope of this study was limited to the use of regularised regression and tree-based models, excluding NNs. NN models are less commonly considered than tree-based techniques due to their opacity, and may also be overly calibrated if only a few variables are available for pricing (Rudin, 2019). However, progress has recently been made on their explainability, and we will explore NN in future work, across a range of suitable network architectures. Recent developments in model-agnostic interpretability techniques have been applied successfully to deep learning models, making them increasingly viable for applications in regulated jurisdictions. Furthermore, the availability of larger datasets and advancements in training efficiency may help address concerns related to overfitting and computational cost. As fairness and accountability remain central to actuarial modelling, future work can assess whether NNs, enhanced by explainability tools, can meet the dual requirements of performance and transparency in insurance pricing.

On the topic of interpretability, to support regulatory transparency in the context of annual pricing review, we propose to communicate variable importance to regulators in an interpretable

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

manner. Furthermore, we propose that rather than emphasising technical model metrics, the focus should be on clear, stakeholder-friendly reporting that highlights which features most significantly influence pricing outcomes. One effective approach is to provide ranked variable importance summaries accompanied by concise explanations of the impact of each variable. Additionally, firms can assess and justify the continued use of variables that may raise fairness concerns, such as region, occupation, or payment method. Incorporating a dashboard-type analysis into model governance documentation and submitting an interpretability report alongside the pricing review can enhance compliance and support supervisory dialogue.

While this study focuses on gender as a protected attribute, we acknowledge that fairness in insurance pricing is a multidimensional issue that extends beyond single-variable assessments. In practice, individuals may face compounded disadvantages based on intersections of gender, race, socio-economic status and geographic location. Although our dataset did not contain direct indicators for attributes such as race or income, future research could benefit from incorporating multi-attribute fairness frameworks, such as intersectional fairness or subgroup fairness, to assess the compounded effects of algorithmic decision-making. This is especially pertinent given increasing regulatory attention in various jurisdictions. For example, the United States has introduced proposals addressing algorithmic bias in mortgage lending and insurance that emphasise the role of race and ethnicity (U.S. Department of the Treasury, 2024), while the European Union's AI Act encourages auditing for discrimination across multiple protected characteristics (European Commission, 2023).

5.3 Conclusion

This paper evaluated GLM and ML models for non-life insurance pricing and prepared a benchmarking framework to simultaneously assess premium estimation accuracy and fairness. Furthermore, we proposed a quantitative framework for the evaluation of pricing models with respect to protected policyholder information (such as gender). The results of our analysis showed that GLM was outperformed by ML on most metrics, and thus ML provides reasonable alternatives to GLM. This study also illustrated that different ML techniques can be more appropriate for frequency and severity modelling. Mixing any two of these modelling techniques can achieve improved levels of performance across a given policyholder portfolio. For the evaluation of fairness, the novel quantitative framework we designed may serve as a valuable resource for insurance companies to better monitor pricing discrimination.

Data Availability Statement. The datasets used in this study are openly available from the R package CASdatasets (Charpentier, 2014; Dutang & Charpentier, 2015). Additionally, all source code supporting the findings of this study is openly accessible on GitHub at https://github.com/tisrani/Pricing.

Financial Support. This publication has emanated from research conducted with the financial support of Research Ireland under Grant number 12/RC/2289-P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Competing Interests Statement. None.

References

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.
Baumann, J., & Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. *Philosophy & Technology*, **36**(3), 45.
Becker, S., Jentzen, A., Müller, M. S., & von Wurstemberger, P. (2022). Learning the random variables in Monte Carlo simulations with stochastic gradient descent: Machine learning for parametric PDEs and financial derivative pricing. *ArXiv Preprint ArXiv:2202.02717*.
Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2020). Machine learning in P&C insurance: A review for pricing and reserving. *Risks*, **9**(1), 4.

Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2022). Geographic ratemaking with spatial embeddings. *ASTIN Bulletin: The Journal of the IAA*, **52**(1), 1–31.

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2022). Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. *27th International Conference on Intelligent User Interfaces*, 807–819.

Breiman, L. (2001). Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Brouste, A., Dutang, C., & Rohmer, T. (2024). A closed-form alternative estimator for GLM with categorical explanatory variables. *Communications in Statistics-Simulation and Computation*, **53**(5), 2444–2460.

Campo, B. D., & Antonio, K. (2022). Insurance pricing with hierarchically structured data: An illustration with a workers' compensation insurance portfolio. *ArXiv Preprint ArXiv:2206.15244*.

Central Bank of Ireland. (2022). Review of differential pricing in the private car and home insurance markets. https://www.centralbank.ie/docs/default-source/publications/consultation-papers/cp143/differential-pricingreview—final-report-and-public-consultation.pdf?sfvrsn = 5

Central Bank of Ireland. (2024). Consultation paper on the consumer protection code. https://www.centralbank.ie/docs/default-source/publications/consultation-papers/cp158/cp158-consultation-paper-consumer-protectioncode.pdf?sfvrsn = 45d631a_4

Charpentier, A. (2014). *Computational actuarial science with R*. CRC Press.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chhachhi, S., & Teng, F. (2023). On the 1-Wasserstein distance between location-scale distributions and the effect of differential privacy. *ArXiv Preprint ArXiv:2304.14869*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, **5**(2), 153–163.

Ciatto, N., Verelst, H., Trufin, J., & Denuit, M. (2023). Does autocalibration improve goodness of lift? *European Actuarial Journal*, **13**(1), 479–486.

Colella, S., & Jones, H. (2023). Machine learning and ratemaking: Assessing performance of four popular algorithms for modeling auto insurance pure premium. *Casualty Actuarial Society (CAS) E-Forum*.

European Commission (2023). Proposal for a regulation of the european parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act). https://artificialintelligenceact.eu

Dal Pozzolo, A., Moro, G., Bontempi, G., & Le Borgne, D. Y. A. (2011). Comparison of data mining techniques for insurance claim prediction (Master's thesis). *Universita Degli Studi Di Bologna*.

De Angelis, M., & Gray, A. (2021). Why the 1-Wasserstein distance is the area between the two marginal CDFs. *ArXiv preprint ArXiv:2111.03570*.

De Jong, P., & Heller, G. Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

Delcaillau, D., Ly, A., Papp, A., & Vermet, F. (2022). Model transparency and interpretability: Survey and application to the insurance industry. *European Actuarial Journal*, **12**(2), 443–484.

Denuit, M., Charpentier, A., & Trufin, J. (2021). Autocalibration and Tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, **101**, 485–497.

Denuit, M., Maréchal, X., Pitrebois, S., &Walhin, J.-F. (2007). *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons.

Devriendt, S., Antonio, K., Reynkens, T., & Verbelen, R. (2021). Sparse regression with multi-type regularized feature modeling. *Insurance: Mathematics and Economics*, **96**, 248–261.

Dolman, C., & Semenovich, D. (2018). Algorithmic fairness: Contemporary ideas in the insurance context. In *IFoA GIRO Conference*, 23–26.

Dominic, S., Björn, B., & Nicolas, B. (2023). The transport package. https://cran.r-project.org/web/packages/transport/index.html

Dutang, C., & Charpentier, A. (2015). Casdatasets manual. https://cas.uqam.ca/pub/web/CASdatasets-manual.pdf

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2011). Fairness through awareness. *ArXiv Preprint ArXiv:1104.3913*.

European Insurance and Occupational Pensions Authority. (2023a). Artificial intelligence governance principles towards ethical and trustworthy artificial intelligence in the european insurance sector. https://www.eiopa.europa.eu/document/download/30f4502b-3fe9-4fad-b2a3-aa66ea41e863_en?filename = Artificial/20intelligence/20governance/20principles.pdf

European Insurance and Occupational Pensions Authority. (2023b). Supervisory statement on differential pricing practices in non-life insurance lines of business. https://www.eiopa.europa.eu/system/files/2023-03/EIOPA-BoS-23-076-Supervisory-Statement-on-differential-pricing-practices_0.pdf

European Parliamentary Research Service. (2017). The EU Directive. https://www.europarl.europa.eu/RegData/etudes/STUD/2017/593787/EPRS_STU(2017)593787_EN.pdf

Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. *International Journal of Advances in Soft Computing and its Applications*, **10**(2), 159–171.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

Ferrario, A., & Hämmerli, R. (2019). On boosting: Theory and applications. Available at SSRN 3402687.

Ferrario, A., Noll, A., &Wuthrich, M. V. (2020). Insights from inside neural networks. Available at SSRN 3226852.

Financial Conduct Authority. (2021). General insurance pricing practices – Amendments. https://www.fca.org.uk/publication/policy/ps21-11.pdf

Frees, E. W. (2015). Analytics of insurance markets. *Annual Review of Financial Economics*, **7**, 253–277.

Frees, E. W., & Huang, F. (2023). The discriminating (pricing) actuary. *North American Actuarial Journal*, **27**(1), 2–24.

Frees, E. W., Meyers, G., & Cummings, A. D. (2014). Insurance ratemaking and a Gini index. *Journal of Risk and Insurance*, **81**(2), 335–366.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. https://www.jstatsoft.org/v33/i01/

Fryda, T. (2023). R interface for 'h2o'. https://cran.r-project.org/web/packages/h2o/index.html

Grari, V., Charpentier, A., & Detyniecki, M. (2022). A fair pricing model via adversarial learning. *ArXiv Preprint ArXiv:2202.12008*.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, **39**(3), 3659–3667.

Haberman, S., & Renshaw, A. E. (1996). Generalized linear models and actuarial science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **45**(4), 407–436.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. **2**). Springer.

Havrylenko, Y., & Heger, J. (2022). Detection of interacting variables for generalized linear models via neural networks. *ArXiv Preprint ArXiv:2209.08030*.

Henckaerts, R., & Antonio, K. (2022). The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance: Mathematics and Economics*, **105**, 79–95.

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.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Iturria, C. A. A. (2023). Discrimination in insurance pricing (Doctoral dissertation). University of Waterloo.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *ArXiv Preprint ArXiv:1609.05807*.

König, D., & Loser, F. (2020). GLM, neural network and gradient boosting for insurance pricing, part 1: Claim frequency. https://www.kaggle.com/code/floser/glm-neural-nets-and-xgboost-for-insurance-pricing

Krasniqi, D., Bardet, J.-M., & Rynkiewicz, J. (2022). Parametric and XGBoost hurdle model for estimating accident frequency. https://hal.science/hal-03739838v2/document

Kuo, K. (2019). Generative synthesis of insurance datasets. *ArXiv Preprint ArXiv:1912.02423*.

Kuo, K., & Lupton, D. (2020). Towards explainability of machine learning models in insurance pricing. *ArXiv Preprint ArXiv:2003.10674*.

Kurz, C. F. (2017). Tweedie distributions for fitting semicontinuous health care utilization cost data. *BMC Medical Research Methodology*, **17**, 1–8.

Li, D., Li, B., & Shen, Y. (2021). A dynamic pricing game for general insurance market. *Journal of Computational and Applied Mathematics*, **389**, 113349.

Lindholm, M., Lindskog, F., & Palmquist, J. (2023). Local bias adjustment, duration-weighted probabilities, and automatic construction of tariff cells. *Scandinavian Actuarial Journal*, **2023**(10), 946–973.

Lindholm, M., Richman, R., Tsanakas, A., & Wuthrich, M. (2022a). Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, **52**(1), 55–89.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022b). A discussion of discrimination and fairness in insurance pricing. *ArXiv preprint ArXiv:2209.00858*.

Lindholm, M., Richman, R., Tsanakas, A., & Wuthrich, M. (2024a). Sensitivity-based measures of discrimination in insurance pricing. https://papers.ssrn.com/sol3/papers.cfm?abstract_id = 4897265

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024b). What is fair? Proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*, **2024**(9), 935–970.

Lorentzen, C., & Mayer, M. (2020). Peeking into the black box: An actuarial case study for interpretable machine learning. Available at SSRN 3595944.

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

Marin-Galiano, M., & Christmann, A. (2004). Insurance: An R-program to model insurance data. (No. 2004, 49). Technical Report.

McCullagh, P., & Nelder, J. (1989). Binary data. In *Generalized Linear Models* (pp. 98–148). Springer.

Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable. *Proceedings of the IEEE*, 109(3), 247–278.
Moriah, M., Vermet, F., & Charpentier, A. (2024). Measuring and mitigating biases in motor insurance pricing. *European Actuarial Journal*, **14**(3), 833–869.
Mosley, R., & Wenman, R. (2022). Methods for quantifying discriminatory effects on protected classes in insurance. Casualty Actuarial Society (CAS) Research Paper Series on Race and Insurance Pricing.
Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384.
Noll, A., Salzmann, R., &Wuthrich, M. V. (2020). Case study: French motor third-party liability claims. Available at SSRN 3164764.
Ohlsson, E., & Johansson, B. (2010). The basics of pricing with GLMs. In *Non-Life Insurance Pricing with Generalized Linear Models* (pp. 15–38). Springer.
Parodi, P. (2014). *Pricing in general insurance*. CRC Press.
Pichler, A. (2014). Insurance pricing under ambiguity. *European Actuarial Journal*, **4**, 335–364.
Quijano Xacur, O. A., & Garrido, J. (2015). Generalised linear models for aggregate claims: To Tweedie or not? *European Actuarial Journal*, **5**(1), 181–202.
R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
Ridgeway, G. (2014). GBM: Generalized boosted regression models. https://cran.r-project.org/web/packages/gbm/gbm.pdf
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.
Schelldorfer, J., &Wuthrich, M. V. (2019). Nesting classical actuarial models into neural networks. Available at SSRN 3320525.
Shi, P. (2016). Insurance ratemaking using a copula-based multivariate tweedie model. *Scandinavian Actuarial Journal*, **2016**(3), 198–215.
Shimao, H., & Huang, F. (2022). Welfare cost of fair prediction and pricing in insurance market. SSRN Manuscript ID, 4225159.
Simjanoska, T. (2022). D-vine regression in insurance (Master's thesis). https://mediatum.ub.tum.de/doc/1658790/1658790.pdf
Smyth, G. K. (1996). Regression analysis of quantity data with exact zeros. In *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management, Gold Coast, Australia*, 17–19.
Southworth, H. (2015). GBM: Generalized boosted regression models. https://github.com/harrysouthworth/gbm
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.
Su, X., & Bai, M. (2020). Stochastic gradient boosting frequency-severity model of insurance claims. *PLOS ONE*, **15**(8), e0238000.
Therneau, T., & Atkinson, E. (2019). An introduction to recursive partitioning using the rpart routines (vignette). https://cran.r-project.org/web/packages/rpart/index.html
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.
U.S. Department of the Treasury. (2024). Assessing the impact of artificial intelligence on financial services. https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf
Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.
Wuthrich, M. V., & Buser, C. (2023). Data analytics for non-life insurance pricing. Swiss Finance Institute Research Paper No. 16-68.
Wüthrich, M. V. (2020). Bias regularization in neural network models for general insurance pricing. *European Actuarial Journal*, **10**(1), 179–202.
Wüthrich, M. V., & Merz, M. (2022). Deep learning. In *Statistical Foundations of Actuarial Learning and its Applications* (pp. 267–379). Springer.
Xin, X., & Huang, F. (2024). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, **28**(2), 285–319.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

https://doi.org/10.1017/S1357321725100317 Published online by Cambridge University Press

Appendix A. Grouping for Categorical Variables

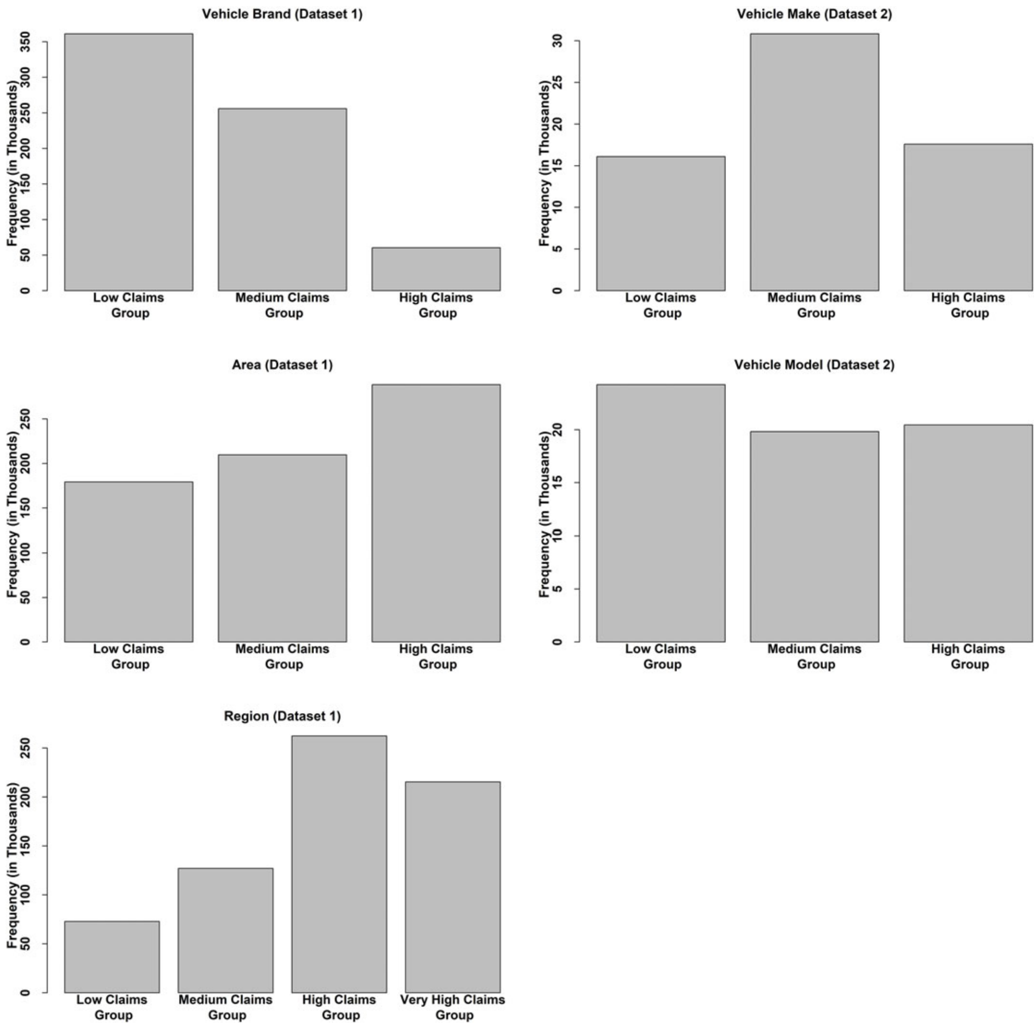

Figure A1. Distributions of recategorised variables. Based on the average observed claims, variables were recategorised into broader, well-defined groups to select key differentiating features and use interpretable modelling approaches to enhance the significance and understanding of predictions for each group.

Cite this article: Israni T, Daly L, Condon J, and Wolsztynski E. (2026). Dual evaluation of performance and fairness from machine learning models for non-life insurance pricing. *British Actuarial Journal*. https://doi.org/10.1017/S1357321725100317