




Article

# Claim Prediction and Premium Pricing for Telematics Auto Insurance Data Using Poisson Regression with Lasso Regularisation

Farha Usman <sup>1,\*</sup>, Jennifer S. K. Chan <sup>1,\*</sup>, Udi E. Makov <sup>2</sup>, Yang Wang <sup>1</sup> and Alice X. D. Dong <sup>3</sup>

<sup>1</sup> School of Mathematics and Statistics, The University of Sydney, Camperdown, NSW 2050, Australia; ywan7081@uni.sydney.edu.au

<sup>2</sup> Department of Statistics, University of Haifa, Haifa 3103301, Israel; makov@stat.haifa.ac.il

<sup>3</sup> Transdisciplinary School, University of Technology Sydney, Ultimo, NSW 2007, Australia; xiaodan.dong@uts.edu.au

\* Correspondence: fusm0507@uni.sydney.edu.au (F.U.); jennifer.chan@sydney.edu.au (J.S.K.C.)

† These authors contributed equally to this work.

**Abstract:** We leverage telematics data on driving behavior variables to assess driver risk and predict future insurance claims in a case study utilising a representative telematics sample. In the study, we aim to categorise drivers according to their driving habits and establish premiums that accurately reflect their driving risk. To accomplish our goal, we employ the two-stage Poisson model, the Poisson mixture model, and the Zero-Inflated Poisson model to analyse the telematics data. These models are further enhanced by incorporating regularisation techniques such as lasso, adaptive lasso, elastic net, and adaptive elastic net. Our empirical findings demonstrate that the Poisson mixture model with the adaptive lasso regularisation outperforms other models. Based on predicted claim frequencies and drivers' risk groups, we introduce a novel usage-based experience rating premium pricing method. This method enables more frequent premium updates based on recent driving behaviour, providing instant rewards and incentivising responsible driving practices. Consequently, it helps to alleviate cross-subsidization among risky drivers and improves the accuracy of loss reserving for auto insurance companies.

**Keywords:** usage-based insurance pricing; lasso regression; Poisson mixture model; ROC curve; experience rating auto insurance premium



**Citation:** Usman, Farha, Jennifer S. K. Chan, Udi E. Makov, Yang Wang, and Alice X. D. Dong. 2024. Claim

Prediction and Premium Pricing for Telematics Auto Insurance Data Using Poisson Regression with Lasso

Regularisation. *Risks* 12: 137.

<https://doi.org/10.3390/risks12090137>

Received: 23 July 2024

Revised: 15 August 2024

Accepted: 22 August 2024

Published: 28 August 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Traditional auto insurance premiums have been based on *driver-related risk (demographic) factors* such as age, gender, marital status, claim history, credit risk and living district, and *vehicle-related risk factors* such as vehicle year/make/model, which represent the residual value of an insured vehicle. Although these *traditional* variables or factors indicate claim frequency and size, they do not reflect true driving risk and often lead to cross-subsidising higher-risk drivers by lower-risk drivers to balance the claim cost. These premiums have been criticised for being inefficient and socially unfair because they do not punish aggressive driving nor encourage prudent driving. Chassagnon and Chiappori (1997) reported that the accident risk depends not only on demographic variables but also on driver behaviour that reflects how drivers drive cautiously to reduce accident risk.

Usage-based insurance (UBI) relies on telematic data, often augmented by global positioning systems (GPSs), to gather vehicle information. UBI encompasses two primary models: Pay As You Drive (PAYD) and Pay How You Drive (PHYD). PAYD operates on a drive-less-pay-less principle, taking into account *driving habits* and travel details such as route choices, travel time, and mileage. This model represents a significant advancement over traditional auto insurance approaches. For instance, Ayuso et al. (2019) utilised a Poisson regression model to analyse a combination of seven traditional and six travel-related variables. However, Kantor and Stárek (2014) highlighted limitations in PAYD policies, notably their sole focus on kilometres driven, neglecting crucial driver behaviour aspects.

By integrating a telematics device into the vehicle, PHYD extends the principles of PAYD to encompass the monitoring of driving behaviour profiles over a specified policy period. *Driving behaviour*, encompassing operational choices such as speeding, harsh braking, hard acceleration, or sharp cornering in varying road types, traffic conditions, and weather, serves as a defining aspect of drivers' styles (Tselentis et al. 2016; Winlaw et al. 2019). These collected driving data offer valuable insights into assessing true driving risks, enabling the calculation of the subsequent UBI experience rating premium. This advancement over traditional premiums incorporates both historical claim experiences and current driving risks. The UBI premium can undergo regular updates to provide feedback to drivers, incentivising improvements in driving skills through premium reductions. Research by Soleymanian et al. (2019) indicated that individuals drive less and safer when incentivised by UBI premiums. Moreover, Bolderdijk et al. (2011) demonstrated that monitoring driving behaviours can effectively reduce speeding and accidents by promoting drivers' awareness and behavioural changes. Wouters and Bos (2000) showed that monitoring of driving resulted in a 20% reduction in accidents. The monitoring system enables early intervention for risky drivers, potentially saving lives (Hurley et al. 2015). Finally, Ellison et al. (2015) concluded that personalised feedback coupled with financial incentives yields the most significant changes in driving behaviour, emphasising the importance of a multifaceted approach to risk reduction.

The popularity of PHYD policies has surged in recent years, driven by the promise of lower premiums for safe driving behaviour. QBE Australia (Q for Queensland Insurance, B for Bankers' and Traders' and E for The Equitable Probate and General Insurance Company), renowned for its innovative approaches, introduced a product called the *Insurance Box*, a PHYD policy featuring in-vehicle telematics. This product not only offers lower premiums to good drivers but also delivers risk scores as actionable feedback on driving performance. These risk scores directly influence the calculation of insurance premiums. In essence, PHYD policies epitomise personalised insurance (Barry and Charpentier 2020), nurturing a culture of traffic safety while concurrently reducing congestion and environmental impact by curbing oil demand and pollutant emissions.

To assess UBI premiums, extensive driving data are initially gathered via telematics technology. Subsequently, a comprehensive set of driving behaviour variables, termed driving variables (DVs), is generated. These variables encompass four main categories: driver-related, vehicle-related, driving habits, and driving behaviours. These DVs are then analysed through regression against insurance claims data to unveil correlations between driving habits and associated risks, which is a process commonly referred to as knowledge discovery (Murphy 2012). Stipancic et al. (2018) determined drivers' risk by analysing the correlations of accident frequency and accident severity with specific driving behaviours such as hard braking and acceleration.

To forecast and model accident frequencies, Guillen et al. (2021) proposed utilising Poisson regression models applied to both traditional variables (related to drivers, vehicles, and driving habits) and critical incidents, which encompass risky driving behaviours. Through this approach, the study delineates insurance premiums into a baseline component and supplemental charges tailored to *near-miss events*—defined as critical incidents like abrupt braking, acceleration, and smartphone usage while driving—which have the potential to precipitate accidents. Building on this, Guillen et al. (2020) employed negative binomial (NB) regression, regressing seven traditional variables, five travel-related factors, and three DVs to the frequency of near-miss events attributable to acceleration, braking, and cornering. Notably, the study suggests that these supplementary charges stemming from near-miss events could be dynamically updated on a weekly basis.

To comprehensively assess the potential nonlinear impacts of DVs on claim frequencies, Verbelen et al. (2018) utilised Poisson and negative binomial regression models within generalised additive models (GAMs). They focused on traditional variables, as well as telematic risk exposure DVs, such as total distance driven, yearly distance, number of trips, distance per trip, distance segmented by road type (urban, other, motorways, and abroad), time slot, and weekday/weekend. While these exposure-centric DVs can serve as offsets in regression models, they fail to capture the subtle details of actual driving behaviour. To attain a deeper understanding of driving behaviour, it becomes essential to

extract a broader array of DVs that can discern between safe and risky driving practices while also delineating claim risk. For instance, rather than merely registering a braking event, a more comprehensive approach involves constructing a detailed 'braking story' that accounts for various factors such as road characteristics (location, lanes, angles, etc.), braking style (abrupt, continuous, repeated, intensity, etc.), braking time (time of day, day of the week, etc.), road type (speed limit, normative speed, etc.), weather conditions, preceding actions (turning, lane changing, etc.), and more. Furthermore, the inclusion of additional environmental and traffic variables obtained through GPS enhances the richness of available information, facilitating a more thorough analysis of driving behaviour and associated risk factors.

As the number of variables describing driving behaviour increases, the data can become voluminous, volatile, and noisy. Managing this influx of variables is crucial to mitigate computational costs and address the issue of multicollinearity among them. Multicollinearity arises due to significant overlap in the predictive power of certain variables. For instance, a driver residing in an area with numerous traffic lights might engage in more forceful braking, or an elderly driver might tend to drive more frequently during midday rather than during typical rush hours or late nights. Consequently, it is possible for confusion to arise between factors such as location and aggressive braking or age and preferred driving times. Multicollinearity can lead to overfitting and instability in predictive models, diminishing their effectiveness. Thus, streamlining the variables to a manageable number is not only essential for computational efficiency but also critical for addressing multicollinearity and enhancing the reliability of predictive models.

Machine learning, employing statistical algorithms, holds remarkable potential in mitigating overfitting and bolstering the stability and predictability of various predictive models. These algorithms are typically categorised as supervised or unsupervised. In the realm of unsupervised learning, [Osafune et al. \(2017\)](#) conducted an analysis wherein they developed classifiers to distinguish between safe and risky drivers based on acceleration, deceleration, and left-acceleration frequencies gleaned from smartphone-equipped sensor data from over 800 drivers. By labelling drivers with at least 20 years of driving experience and no accident records as safe and those with more than two accident records as risky, they achieved a validation accuracy of 70%. [Wüthrich \(2017\)](#) introduced pattern recognition techniques utilising two-dimensional velocity and acceleration (VA) heat maps via K-means clustering. However, it is worth noting that neither study offers predictions related to insurance claims.

With claim risk information such as claim making, claim frequency, and claim size, supervised machine learning models embedded within generalised linear models (GLMs) can be constructed to unfold the hidden patterns in big data and predict future claims for premium pricing. Various machine learning techniques are widely utilised in predictive modelling, including clustering, decision trees, random forests, gradient boosting, and neural networks. [Gao et al. \(2019\)](#) investigated the effectiveness of Poisson GAMs, integrating two-dimensional speed–acceleration heat maps alongside traditional risk factors for predicting claim frequencies. They employed feature extraction methods outlined in their previous work ([Gao and Wüthrich 2018](#)), such as K-medoids clustering to group drivers with similar heatmaps and principal component analysis (PCA) to reduce the dimensionality of the design matrix, thereby enhancing computational efficiency. Furthermore, [Gao et al. \(2019\)](#) conducted an extensive analysis focusing on the predictive power of additional driving style and habit covariates using Poisson GAMs. In a similar vein, [Makov and Weiss \(2016\)](#) integrated decision trees into Poisson predictive models, expanding the repertoire of predictive algorithms in insurance claim forecasting.

In assessing various machine learning techniques, [Paefgen et al. \(2013\)](#) discovered that neural networks outperformed logistic regression and decision tree classifiers when analysing claim events using 15 travel-related variables. [Ma et al. \(2018\)](#) employed logistic regression to explore accident probabilities based on four traditional variables and 13 DVs, linking these probabilities to insurance premium ratings. [Weerasinghe and Wijegunasekara \(2016\)](#) categorised claim frequencies as low, fair, and high and compared neural networks, decision trees, and multinomial logistic regression models. Their findings indicated that neural networks achieved the best predictive performance, although logistic regression was

recommended for its interpretability. Additionally, [Huang and Meng \(2019\)](#) utilised logistic regression for claim probability and Poisson regression for claim frequency, incorporating support vector machine, random forest, advanced gradient boosting, and neural networks. They examined seven traditional variables and 30 DVs grouped by travel habits, driving behaviour, and critical incidents, employing stepwise feature selection and providing an overview of UBI pricing models integrated with machine learning techniques.

However, machine learning techniques often encounter challenges with overfitting. One strategy to address both multicollinearity and overfitting is to regularise the loss function by penalising the likelihood based on the number of predictors. While ridge regression primarily offers shrinkage properties, it does not inherently select an optimal set of predictors to capture the best driving behaviours. [Tibshirani \(1996\)](#) introduced the Least Absolute Shrinkage and Selection Operator (lasso) regression, incorporating an L1 penalty for the predictors. Subsequently, the lasso regression framework underwent enhancements to improve model fitting and variable selection processes. For instance, [Zou and Hastie \(2005\)](#) proposed the elastic net, which combines the L1 and L2 penalties of lasso and ridge methods linearly. [Zou \(2006\)](#) introduced the adaptive lasso, employing adaptive weights to penalise different predictor coefficients in the L1 penalty. Moreover, [Park and Casella \(2008\)](#) presented the Bayesian implementation of lasso regression, wherein lasso estimates can be interpreted as Bayesian posterior mode estimates under the assumption of independent double-exponential (Laplace) distributions as priors on the regression parameters. This approach allows for the derivation of Bayesian credible intervals of parameters to guide variable selection. [Jeong and Valdez \(2018\)](#) expanded upon the Bayesian lasso framework proposed by [Park and Casella \(2008\)](#) by introducing conjugate hyperprior distributional assumptions. This extension led to the development of a new penalty function known as log-adjusted absolute deviation, enabling variable selection while ensuring the consistency of the estimator. While the Bayesian approach is applicable, the running of MCMC is often time-consuming.

When modelling claim frequencies, a common issue arises from an abundance of zero claims, which Poisson or negative binomial regression models may not effectively capture. These zero claims do not necessarily signify an absence of accidents during policy terms but rather indicate that some policyholders, particularly those pursuing no-claim discounts, may refrain from reporting accidents. To identify factors influencing zero and nonzero claims, [Winlaw et al. \(2019\)](#) employed logistic regression with lasso regularisation on a case-control study, assessing the impact of 24 DVs on acceleration, braking, speeding, and cornering. Their findings highlighted speeding as the most significant driver behaviour linked to accident risk. In a different approach, [Guillen et al. \(2019\)](#) and [Deng et al. \(2024\)](#) utilised zero-inflated Poisson (ZIP) regression models to model claim frequencies directly and [Tang et al. \(2014\)](#) further integrated the model with the EM algorithm and adaptive lasso penalty. However, [Tang et al. \(2014\)](#) observed suboptimal variable selection results for the zero-inflation component, suggesting a lower signal-to-noise ratio compared to the Poisson component. [Banerjee et al. \(2018\)](#) proposed a multicollinearity-adjusted adaptive lasso approach employing ZIP regression. They explored two data-adaptive weighting schemes: inverse of maximum likelihood estimates and inverse estimates divided by their standard errors. For a comprehensive overview of various modelling approaches in UBI, refer to Table A1 in [Eling and Kraft \(2020\)](#).

Numerous studies in UBI have employed a limited number of DVs to characterise a broad spectrum of driver behaviours. For instance, [Jeong \(2022\)](#) analysed synthetic telematic data sourced from [So et al. \(2021\)](#), encompassing 10 traditional variables and 39 DVs, including metrics like sudden acceleration and abrupt braking. While [Jeong \(2022\)](#) utilised PCA to reduce dimensionality and enhance predictive model stability, the interpretability of the principal components derived from PCA remains constrained. Regularisation provides a promising alternative for dimension reduction while addressing the challenge of overfitting. The literature on UBI predictive models employing GLMs with machine learning techniques, such as lasso regularisation to mitigate overfitting, is still relatively sparse, particularly concerning forecasting claim frequencies and addressing challenges like excessive zero claims and overdispersion arising from heterogeneous driving behaviours.

Our main objective is to propose predictive models to capture the impact of driving behaviours (safe or risky) on claim frequencies, aiming to *enhance prediction accuracy, identify relevant DVs, and classify drivers based on their driving behaviours*. This segmentation will enable the application of differential UBI premiums for safe and risky drivers. More importantly, we advocate for the regular updating of these UBI premiums to provide ongoing feedback to drivers through the relevant DVs and encourage safer driving habits.

We demonstrate the applicability of the proposed predictive models through a case study using a representative telematics dataset comprising 65 DVs. The proposed predictive models includes two-stage threshold Poisson (TP), Poisson mixture (PM), and ZIP regression models with lasso regularisation. We extend the regularisation technique to include adaptive lasso and elastic net, facilitating the identification of distinct sets of DVs that differentiate safe and risky behaviours. In the initial stage of regularised TP models, drivers are classified into risky (safe) group if their annual predicted claim frequencies, estimated by a single-component Poisson model, exceed (not exceed) predefined thresholds. Subsequently, in stage two, regularised Poisson regression models are refitted to each driver subgroup (exceeding thresholds or not) using different sets of selected DVs in each group. Alternatively, PM models simultaneously estimate parameters and classify drivers. Our findings reveal that PM models offer greater efficiency compared to TP models, providing added flexibility and capturing overdispersion akin to NB distributions.

In ZIP models, we observe that the structural zero component may not necessarily indicate safe drivers, as safe drivers may claim less frequently but not necessarily abstain from claims altogether, while risky drivers may avoid claims due to luck or incentives like bonus rewards. [So et al. \(2021\)](#) investigated the cost-sensitive multiclass adaptive boosting method, defining classes based on zero claims, one claim, and two or more claims, differing from our proposed safe and risky driver classes. We argue that the level of accident risk may not solely correlate with the number of claims but rather with driving behaviours. Hence, the regularised PM model proves more efficient in tracking the impact of DVs on claim frequencies, allowing for divergent effects between safe and risky drivers. This proposed PM model constitutes the primary contribution of this paper, addressing a critical research gap in telematics data analysis.

Our second contribution is to bolster the robustness of our approach and mitigate overfitting by incorporating resampling and cross-validation (CV) apart from lasso regularisation. These techniques help us attain more stable and reliable results. Additionally, we utilise the area under curve (AUC) from the receiver operating characteristic (ROC) curve as one of the performance metrics, which evaluates classification accuracy highlighting the contribution of predictive models in classifying drivers.

Our third contribution involves introducing an innovative UBI experience rating premium method. This method extends the traditional experience rating premium method by integrating classified claim groups and predicted claim frequencies derived from the best-trained model. This dynamic pricing approach also enables more frequent update of premiums to incentivise safer and reduced driving. Moreover, averaged and individual driving scores from the identified relevant DVs for each driver can inform their driving behaviour possibly with warnings and encourage skill improvement. By leveraging these advanced premium pricing models, we can improve loss reserving practices, and we can even evaluate the legitimacy of reported accidents based on driving behaviours.

Lastly, we highlight a recent paper ([Duval et al. 2023](#)) with similar aims to this paper. They applied logistic regression with elastic net regularisation to predict the probability of claims, but this paper considers two-group PM regression instead of logistic regression to predict claim frequency and allow different DVs to capture the distinct safe and risky driving behaviours. For predictive variables, they used driving habits information (when, where, and how much the insured drives) from telematics, as well as traditional risk factors such as gender and vehicle age, whereas this paper focuses on driving behaviour/style (how the insured drives). To avoid handcrafting of telematics information, they proposed measures using the Mahalanobis method, Local Outlier Factor, and Isolation Forest to summarise trip information into local/routine anomaly scores by trips and global/peculiar anomaly scores by drivers, which were used as features. This is an *innovative* idea in the literature. On the other hand, this paper uses common handcraft practices to summarise

driving behaviour by drivers, using both driving habits (where and when) and driving styles (how) information by defining driving events such as braking and turning considering time, location, and severity of events. Duval et al. (2023) demonstrated that the improvement in classification using lower global/peculiar Mahalanobis anomaly scores enables a more precise pure premium (product of claim probability from logistic regression to insured amount) calculation. As stated above, this paper provides differential contributions by classifying drivers into safe and risky groups, predicting claims for drivers in their groups using regularised PM models (among regularised TP and ZIP models), which is pioneering in the UBI literature, and calculating premiums using the proposed innovative UBI experience rating premium based on drivers’ classifications (safe/risky) and predicted annual claims.

The paper is structured as follows: Section 2 outlines the GLMs, including Poisson, PM, and ZIP regression models, alongside lasso regularisation and its extensions. Section 3 presents the telematics data and conducts an extensive empirical analysis of the two-stage TP, PM, and ZIP models. Section 4 introduces the proposed UBI experience rating premium method. Lastly, Section 5 offers concluding remarks and implementation guidelines and explores potential avenues for future research.

## 2. Methodologies

We derive predictive models for the claim rate of any driver using GLMs with lasso regularisation when the true model is assumed to have a sparse representation of 65 DVs. This section also considers some model performance measures including AUC.

### 2.1. Regression Models

#### 2.1.1. Poisson Regression Model

The Poisson regression model is commonly applied to count data, like claims. It is defined as

$$Y_i \sim \text{Poisson}(\mu_i(\boldsymbol{\beta})), \quad \mu_i(\boldsymbol{\beta}) = n_i a_i = n_i \exp(\mathbf{x}_{i\bullet} \boldsymbol{\beta}) = \exp(\mathbf{x}_{i\bullet} \boldsymbol{\beta} + \log(n_i)) \quad (1)$$

where  $\mathbf{Y} = Y_{1:N}$ ,  $\boldsymbol{\beta} = \beta_{0:J}$ , and  $J$  are the number of selected DVs in the model;  $a_i = \exp(\mathbf{x}_{i\bullet} \boldsymbol{\beta})$  estimates the number of claim per year for driver  $i$ ; and  $\log(n_i)$  is the offset parameter in a regression model. Poisson regression assumes equidispersion and is applied to the 2-stage TP model in Section 3.4. For overdispersed data, negative binomial (NB) distribution, as in the Poisson–Gamma mixture, provides extra dispersion. With NB regression, the term “Poisson( $\mu_i$ )” in (1) is replaced with NB distribution  $\text{NB}(v, q_i) = \text{NB}(v, \mu_i / (v - \mu_i))$ —where  $v$  is the shape parameter, and  $q_i$  is the success probability of each trial. NB distribution converges to Poisson distribution if  $v$  tends to infinity.

#### 2.1.2. Poisson Mixture Model

The finite mixture Poisson model is another popular model for modelling unobserved heterogeneity. The model assumes  $G$  unobserved groups each with probability  $\pi_g$ ,  $0 < \pi_g < 1, g = 1, \dots, G$ , and  $\sum_{g=1}^G \pi_g = 1$ . Focusing on classifying safe and risky drivers, we model  $G = 2$  groups. Assume that the claim frequency  $Y_i$  for driver  $i$  in group  $g$  follows a Poisson distribution with mean  $\mu_{ig}$ , that is,  $Y_i \sim \text{Poisson}(\mu_{ig})$  at probability  $\pi_g, g = 1, \dots, G$ . The probability mass function (pmf) of the Poisson mixture model is

$$f(y_i | \pi, \mu_{i1}(\boldsymbol{\beta}_1), \mu_{i2}(\boldsymbol{\beta}_2)) = \pi f_1(y_i | \mu_{i1}(\boldsymbol{\beta}_1)) + (1 - \pi) f_2(y_i | \mu_{i2}(\boldsymbol{\beta}_2)) \quad (2)$$

where  $\pi_1 = \pi$  and  $\pi_2 = 1 - \pi$ ,  $\boldsymbol{\beta}_g = (\beta_{0:J,g})^\top$ ,  $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_G^\top, \pi_1, \dots, \pi_{G-1})^\top$  is the model parameter vector, and  $f_g(y_i | \mu_{ig}(\boldsymbol{\theta}))$  is the Poisson pmf with mean function  $\mu_{ig}(\boldsymbol{\beta}_g) = \exp(\mathbf{x}_{i\bullet} \boldsymbol{\beta}_g + \log(n_i))$ .

The expectation–maximisation (EM) algorithm is often used to estimate parameters  $\theta$ . In the E step, the posterior group membership for driver  $i$  is estimated by

$$z_{ig} = \frac{\pi_g f_g(y_i | \mu_{ig}(\beta_g))}{\sum_{g'=1}^G \pi_{g'} f_{g'}(y_i | \mu_{ig'}(\beta_{g'}))} \tag{3}$$

The marginal predicted claim is

$$\hat{y}_i = \hat{z}_{i1} \mu_{i1}(\beta_1) + (1 - \hat{z}_{i1}) \mu_{i2}(\beta_2). \tag{4}$$

If there is a high proportion of zero claims, the ZIP model (Lambert 1992) may be suitable to capture the excessive zeros. The model is a special case of a two-group mixture model that combines a zero point mass in group 1 with a Poisson distribution in group 2. The zeros may come from the point mass (structural zero) or the zero count (natural zero) in a Poisson distribution. The model is given by

$$\Pr(Y_i = 0) = \pi_i + (1 - \pi_i) \exp(-\mu_i), \text{ and } \Pr(Y_i = y_i) = (1 - \pi_i) \frac{\mu_i^{y_i} \exp(-\mu_i)}{y_i!}, y_i \geq 1 \tag{5}$$

where the two regression models (called zero and count) for the probability  $\pi_i$  of structural zero and the expected counts (including nonstructural zero)  $\mu_i$ , respectively, are given by

$$\pi_i(\theta) = \frac{\exp(x_{i\bullet} \psi + \log(n_i))}{1 + \exp(x_{i\bullet} \psi + \log(n_i))}, \text{ and } \mu_i(\beta) = \exp(x_{i\bullet} \beta + \log(n_i)), \tag{6}$$

and the logistic regression parameters  $\psi = (\psi_0, \dots, \psi_{J_\psi})^\top$  define a vector of  $J_\psi$  selected DVs to estimate the probability of extra zero; the vector of model parameters is  $\theta = (\psi^\top, \beta^\top, \pi)^\top$ .

### 2.2. Regularisation Techniques

The stepwise procedure to search for a good subset of DVs often suffers from high variability, a local optimum, and ignorance of uncertainty in the searching procedures (Fan and Li 2001). Lasso (L) regularisation offers an alternative approach to select variables for parsimonious models. It is further extended to adaptive lasso (A), elastic net (E), and adaptive elastic net (N). These regularisations with L1 penalty provide a simple way to enforce sparsity in variable selection by shrinking some coefficients  $\beta_j$  to zero. This aligns with our aim to select important DVs, that is, those with coefficients not shrunk to zero.

To implement these regularisation techniques, we consider the penalised log likelihood (PLL) (Banerjee et al. 2018; Bhattacharya and McNicholas 2014). For the case of the most general adaptive elastic net regularisation, coefficients  $\beta_{\lambda, w, \alpha, N}$  of Poisson regression in (1) estimated by minimising the penalised log likelihood are given by

$$\text{LOSS}_{\lambda, \alpha, w}(\beta) = - \sum_{i=1}^N \log f(y_i; \mu_i(\beta)) + \lambda \left[ \frac{1 - \alpha}{2} \sum_{j=1}^J w_j \beta_j^2 + \alpha \sum_{j=1}^J w_j |\beta_j| \right] \tag{7}$$

where the first term is the negative log likelihood (NLL), the second term is the penalty,  $f(y_i; \mu_i(\beta))$  is the pmf of Poisson model, and  $w_j$  are the data-driven adaptive weights. Equation (7) includes special cases:  $\alpha = 1, w_j = 1$  for lasso,  $\alpha = 1$  for adaptive lasso, and  $w_j = 1$  for elastic net.

The development of (7) starts with the basic lasso regularisation with  $\alpha = 1$  and  $w_j = 1$ . The parameter estimates condition on  $\lambda$  are given by

$$\hat{\beta}_{j, \lambda} = \beta_{j, \text{NLL}} \max \left( 0, 1 - \frac{N\lambda}{|\beta_{j, \text{NLL}}|} \right) \tag{8}$$

where  $\beta_{\text{NLL}} = (\beta_{1, \text{NLL}}, \dots, \beta_{J, \text{NLL}})$  minimise the NLL when  $\lambda = 0$ . As  $\lambda$  increases, the term  $1 - \frac{N\lambda}{|\beta_{j, \text{NLL}}|}$  becomes negative, and so  $\hat{\beta}_{j, \lambda}$  will shrink to zero. Then, the penalty term

$\lambda \sum_{j=1}^J |\hat{\beta}_{j,\lambda}|$  will drop as more  $\hat{\beta}_{j,\lambda} = 0$ , but NLL increases as  $\beta_\lambda = (\hat{\beta}_{1,\lambda}, \dots, \hat{\beta}_{J,\lambda})$  get further away from  $\beta_{\text{NLL}}$  so that one can choose a  $\lambda_{\text{min}}$  that minimises the PLL to obtain  $\beta_{\lambda_{\text{L}}}$ . Alternatively, one can perform a  $K$ -fold CV and choose  $\lambda_{\text{min}}$  that provides the best overall model fit for all  $K$  validated samples. Different criterion may suggest different optimal  $\lambda_{\text{min}}$  and hence the estimates  $\beta_{\lambda_{\text{min}}}$ . Details are provided points 1–2 in Appendix A.

However, [Meinshausen and Bühlmann \(2006\)](#) showed the conflict of optimal prediction and consistent variable selection in lasso regression. Moreover, whether lasso regression has an oracle procedure is debatable. An estimating procedure is an oracle if it can identify the right subset of variables and has an optimal estimation rate so that estimates are unbiased and asymptotically normal. [Städler et al. \(2010\)](#) also proclaimed these issues and addressed some bias problems of the (one-stage) lasso, which may shrink important variables too strongly. [Zou \(2006\)](#) introduced the two-stage *adaptive lasso* as a modification of lasso in which each coefficient  $\beta_j$  is given its own weight  $w_j$  to control the rate as each coefficient is shrunk towards 0.

Adaptive lasso deals with three issues, namely, inconsistent selection of coefficients, lack of oracle property, and unstable parameter estimation when working with high dimensional data. As smaller coefficients  $\beta_{j,\text{NLL}}$  in (8) will leave the model faster than larger coefficients, [Zou \(2006\)](#) suggested the weights  $w_j = |\hat{\beta}_{j,R}|^{-\gamma}$  in (7), where the tuning parameter  $\gamma > 0$  is to ensure that the adaptive lasso has oracle properties and that  $\hat{\beta}_{j,R}$  is an initial estimate from ridge regression. The weights are rescaled so that their sum equals to the number of DVs. [Städler et al. \(2010\)](#) suggested the tuning parameter  $\gamma = 1$  for low-claim threshold and  $\gamma = 2$  for high-claim threshold. We adopted  $\gamma = 2$  as the best tuning parameter to estimate weights  $w_j$  in the subsequent adaptive lasso models.

[Zou and Zhang \(2009\)](#) argued that L1 penalty can perform poorly when there are multicollinearity problems, which is common in high-dimensional data. This severely degrades the performance of lasso. They proposed *elastic net*, which takes a weighted average of two penalties: ridge ( $\alpha = 0$ ) and lasso ( $\alpha = 1$ ). The mixing parameter  $\alpha \in (0, 1)$  in (7) balances the two penalties with  $\alpha > 1/3$ , indicating a heavier lasso penalty.

When regularisation is applied to ZIP and PM models, the penalised log likelihood in (7) is extended to

$$\begin{aligned} \text{LOSS}_{\lambda,\alpha,w}(\beta_1, \beta_2) = & - \sum_{i=1}^N \log f(y_i; \pi, \mu_{i1}(\beta_1), \mu_{i2}(\beta_2)) + \lambda_1 \left[ \frac{1-\alpha}{2} \sum_{j=1}^J \beta_{j1}^2 + \alpha \sum_{j=1}^J w_{j1} |\beta_{j1}| \right] \\ & + \lambda_2 \left[ \frac{1-\alpha}{2} \sum_{j=1}^J \beta_{j2}^2 + \alpha \sum_{j=1}^J w_{j2} |\beta_{j2}| \right] \end{aligned} \tag{9}$$

where  $f(y_i; \pi, \mu_{i1}(\beta_1), \mu_{i2}(\beta_2))$  is given by (2).

The optimal  $\alpha$  needs to be searched over to identify the best  $\alpha$  with the lowest mean square error (MSE), root MSE (RMSE), or R-squared. We searched for  $\alpha$  in (7) and (9) for different models summarised in Table 1. For example, model TPAL-2 refers to a stage 2 TP model when adaptive lasso regularisation is applied in stage 1 and lasso is applied in stage 2. Different stage 1 TP, stage 2 (under TPL-1 and TPA-1) TP with threshold  $\tau$  (to split predicted annual claim  $a_i = y_i/n_i$  into low and high groups), PM, and ZIP models were considered. We ran each model over five  $\alpha$  values (0.100, 0.325, 0.550, 0.775, 1.000) and identified the best  $\alpha$ , which gives the lowest RMSE with  $K = 10$  folds CV. To ensure the search is robust, results were repeated  $R = 100$  times for each model based on  $R = 100$  70% subsamples  $S_{1:R}$ . Results show that low  $\alpha = 0.1$  should be adopted for stage 1 TP models, the low group of most stage 2 TP models, the PME model, and the PMN model, whereas higher  $\alpha = 0.775, 1$  should be adopted for the higher group of stage 2 TP models. See point 1 in Appendix B for the implementation of all lasso regularisation procedures under Poisson regression and point 2 in Appendix B for the implementation details using caret package in R.



**Table 1.** Model names for TP, PM, and ZIP models with different lasso regularisation.

Stage 1 Threshold Poisson		Stage 2 Threshold Poisson			Poisson Mixture		Zero-Inflated	
TPL-1	Lasso	TPLL-2	TPAL-2	Lasso	PML	Lasso	ZIPL	Lasso
TPE-1	Elastic net	TPLE-2	TPAE-2	Elastic net	PME	Elastic net		
TPA-1	Adaptive lasso	TPLA-2	TPAA-2	Adaptive lasso	PMA	Adaptive lasso	ZIPL	Adaptive lasso
TPN-1	Adaptive elastic net	TPLN-2	TPAN-2	Adaptive elastic net	PMN	Adaptive elastic net		

### 2.3. Model Performance Measures

Model performance can be evaluated from different aspects depending on the aims and model assumptions. The goodness of model fit, prediction accuracy, and classification of drivers are the main types of criteria that are linked to different metrics.

Firstly, the Bayesian information criterion (BIC) is a popular model fit measure that contains a deviance and a parameter penalty term using the log of sample size as the model complexity penalty weight. The Akaike information criterion (AIC) can also be used when the parameter penalty term uses 2 as the weight. Deviance (without parameter penalty) is also used by some packages to select models.

Secondly, for prediction accuracy, we adopted the popular mean square error  $MSE = \sum_{i=1}^N (y_i - \mu_i)^2 / N$  and mean absolute error  $MAE = \sum_{i=1}^N |y_i - \mu_i| / N$ . The third measure we considered is the correlation  $\rho$  between observed and predicted annual claim frequencies (instead of claim frequencies in MSE). A higher correlation shows better performance.

Lastly, to quantify classification performance, the difference between observed group membership and predicted group membership should be quantified. In machine learning, AUC for ROC curve (Fawcett 2006) is a measure of model classification power. It constructs a confusion matrices condition on the classifier (e.g.,  $a_i$  in (1) for TP-2 and  $z_{i1}$  in (3) for PM) cutoff, calculates the true positive rate (TPR) (sensitivity) and the false positive rate (FPR) (1-specificity), and plots the TPR against the FPR as the discrimination cutoff for the classifier varies to obtain the ROC curve. AUC is the probability that a randomly chosen member of the positive class has a lower estimated probability of belonging to the negative class than a randomly chosen member of the negative class. See point 3 in Appendix B for implementation. For the claim data, we let the binary classifier be the low-claim (safe driver) and high-claim (risky driver) groups. However, the group membership of each driver is not observed, so it is approximated using K-means clustering, which minimises the total within-cluster variation using the selected DVs for each model. These four types of measures, namely BIC, MSE,  $\rho$ , and AUC, assessing different performance perspectives, were applied to assess the performance of a set of models  $\mathbb{M}$ .

Although these four measures are popular in statistical and machine learning models, they are not particularly built for count models. Czado et al. (2009) and Verbelen et al. (2018) proposed six scores for claim count models based on the idea of probability integral transform (PIT) or, equivalently, the predictive CDF. The six scores are defined as

$$\begin{aligned}
 \text{Logarithmic:} \quad & \text{Log}(F, y) &= -\log(f_y) \\
 \text{Quadratic (Quad):} \quad & \text{Quad}(F, y) &= -2f_y + \|f\| \\
 \text{Spherical:} \quad & \text{Spher}(F, y) &= -f_y / \|f\| \\
 \text{Ranked Probability:} \quad & \text{RankProb}(F, y) &= \sum_{k=1}^{\infty} [F_y - \mathbf{1}(y \leq k)]^2 \\
 \text{Dawid-Sebastiani:} \quad & \text{Dawid}(F, y) &= \left(\frac{y - \mu_F}{\sigma_F}\right)^2 + 2\log(\sigma_F) \\
 \text{Squared error:} \quad & \text{SqErr}(F, y) &= (y - \mu_F)^2
 \end{aligned}$$

where  $Y \sim \text{Poisson}$ ,  $f_y = \Pr(Y = y)$ ,  $F_y = \Pr(Y \leq y)$ ,  $\mu_F = E(Y)$ ,  $\sigma_F = \text{Var}(Y)$ , and  $\|f\| = \sum_{k=0}^{\infty} f_k^2$ . For PM models,  $f_y = \pi_1 f_{y1} + (1 - \pi_1) f_{y2}$ ,  $F_y$ , and  $\mu_F$  are similarly defined, and  $\sigma_F^2 = \sum_{k=0}^{\infty} (k - \mu_F)^2 f_k$ . To accommodate the effect of driver classification, the prior

probability  $\pi_g$  is replaced by posterior probabilities  $z_{ig}$  in (3). These scores are averaged over drivers, and lower scores indicate better predictions. The Logarithmic score is the common NLL, which is a model-fit measure. Quadratic and Spherical scores are similar to Logarithmic scores for assessing model fit using different functional forms. Dawid–Sebastiani and Squared error (MSE) scores measure prediction accuracy. For the Dawid–Sebastiani score, the term  $2 \log(\sigma_F)$  adjusts for the fact that the first term decays to zero as  $\sigma_F$  tends to infinity. The Ranked probability score calculates the sum of squares value to summarise the PP plot, plotting the fitted cumulative probability  $F_y$  against the observed proportion  $\mathbf{1}(y \leq k)/N$  when averaged. These six measures are used to select the final model to enrich the versatility of our model selection criteria.

To facilitate model selection, we ranked each model  $\mathcal{M}$  in the model class  $\mathbb{M}$  in descending order of preference for each performance measure  $m_{\mathcal{M},l;4}$  for (BIC, MSE,  $\rho$ , AUC) and  $m_{\mathcal{M},l;6}$  for (Log, Quad, Spher, RankProb, Dawid, SqErr) to obtain ranks  $\mathfrak{R}_{\mathcal{M},l} = \text{rank}_{\mathcal{M} \in \mathbb{M}}(m_{\mathcal{M},l})$  and sum of ranks

$$\mathfrak{R}_{\mathcal{M}} = \sum_{l=1}^l \mathfrak{R}_{\mathcal{M},l}, \quad l = 4, 6 \tag{10}$$

to reflect the performance of each model  $\mathcal{M}$ .

### 3. Empirical Studies

#### 3.1. Data Description

The dataset is originated from cars driven in the US, where special UBI sensors were installed. The University of Haifa Actuarial Research Center provided the data, where UBI modelling was analysed (Chan et al. 2022). It contains two column vectors of claim frequencies  $\mathbf{y} = y_{1:N}$  and policy duration or exposure  $\mathbf{n} = n_{1:N}$  in a year. Ninety-two percent of  $\mathbf{y}$  are zero. Figure 1 displays three histograms for  $\mathbf{y}$ ,  $\mathbf{n}$ , and annual claim frequencies  $\mathbf{a}$  ( $a_i = y_i/n_i$ ), respectively. The dataset also contains  $J_0 = 65$  numerical DVs constructed based on the information collected from telematics and GPS for  $N = 14157$  drivers. Figure A2 in Appendix D.1 visualises the DVs by plotting  $x_{ij}$  across driver  $i$ , with colours indicating the number of claims  $y_i = 0, 1, 2^+$ . We remark that the DVs are labelled up to 77 with some skips of numbers. For example, DV 6, 11, 12, etc. do not exist in Figure A2. Each DV describing a specific event (details in the next section) has been aggregated over time to obtain certain incidence rates (per km or hour of driving) and scaled to normalise their ranges for better interpretability of their coefficients in the predictive models. These procedures transformed the multidimensional longitudinal DVs into a single row for each driver, which is the unit of analysis. All DVs are presented as column vectors  $\mathbf{x}_{\bullet j}, j = 1, \dots, J_0$ .

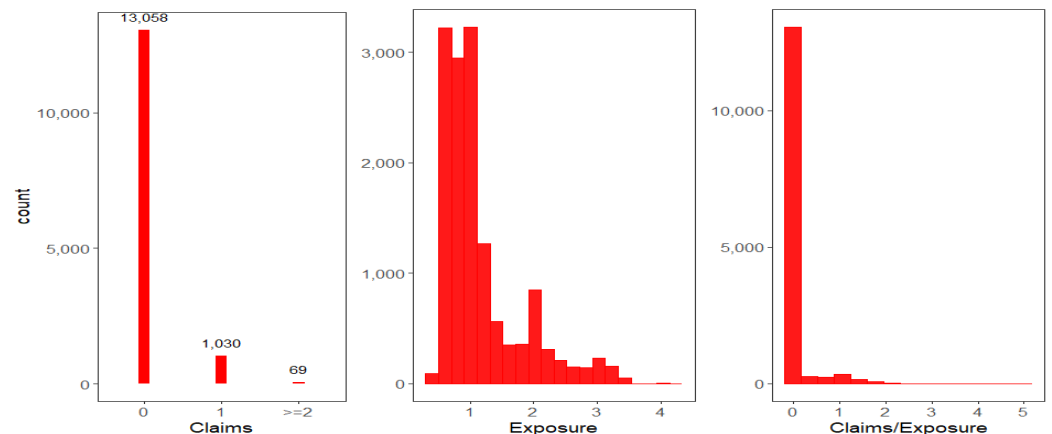


Figure 1. Histogram of claims (left), exposure (mid), and claims per exposure (right).

### 3.2. Data Cleaning and DVs Setting

Telematics sensors are installed by car manufacturers to provide much cleaner signals. Therefore, standard data cleaning techniques, including the removal of outliers, were applied. External environmental information from GPS was utilised to minimise false signals, recognising that driving behaviours are often responsive to varying conditions. Then, the DVs can be defined to indicate specific driving events, which can associate with certain driving risks. However, while rapid acceleration is typically undesirable, it may be necessary when merging onto a busy highway. To accurately process and analyse telematics and GIS data, roads were categorised into specific types such as highways, junctions, roundabouts, and others. This segmentation enables a more precise assessment of driving behaviours across different contexts, improving safety measures and performance evaluations. Given the complexity of telematics data, including metrics like acceleration and braking, the definition of events like rapid acceleration or hard braking was adapted to account for varying road conditions depending also on time. Hence, the DVs were defined for a range of driving events as combinations of event types (e.g., accelerating, braking, left/right turning), environmental condition (e.g., interchange, junction), and time (e.g., the morning rush from 6 am to 9 am). Then, rates of the events (over standardised period or mileage) were evaluated and normalised. Appendix C provides the labels and interpretation of these DVs.

### 3.3. Exploratory Data Analyses

To summarise the variables, their averages are presented:  $\bar{y} = 0.083$ ,  $\bar{n} = 1.146$ , and  $\bar{a} = 0.075$ . We split the drivers into three classes:  $C_b$ ,  $b = 0, 1, 2^+$ , with 0, 1, and at least 2 claims and class sizes  $N_b$ . Their proportions— $p_b = N_b/N$ —came out to (0.92, 0.07, 0.005), averaged exposures— $\bar{n}_b = \sum_{i \in C_b} n_i / N_b$ —came out to (1.13, 1.38, 1.64), and averaged annual claim frequencies— $\bar{a}_b = \sum_{i \in C_b} a_i / N_b$ —came out to (0, 0.92, 1.71). The average claim frequency for  $C_{2^+}$  was 2.11. Regressing the claim frequencies  $y_i$  on the exposure  $n_i$ , the  $R^2$  was only 0.014, showing that the linear effect of exposure on claim frequency is weak and insignificant. Hence, it is possible that other effects, such as driving behaviour as measured by the DVs  $x_{\bullet j}$ , may impact the claim frequency  $y_i$ . Sections 3.4–3.6 will analyse such effects of DVs on claim frequencies.

Section 2.1.1 introduced the Poisson and NB regression for equidispersed and overdispersed data, respectively. To assess the level of dispersion, we used sample variance  $\text{Var}(\mathbf{y}) = 0.089$ , which shows equidispersion possibly due to the large proportion of zeros. We also tested the equidispersion assumption with the null hypothesis— $H_0: \text{Var}(Y_i) = \mu_i$ —and alternative hypothesis— $H_1: \text{Var}(Y_i) = \mu_i + \psi g(\mu_i)$ —where  $g(\cdot) > 0$  is a transformation function (Cameron and Trivedi 1990), and  $\psi > 0$  ( $\psi \leq 0$ ) indicates overdispersion (underdispersion). See point 4 in Appendix B for the implementation. For model TPL-1,  $\psi = 0.0369$  ( $p = 0.0482$ ), and for model TPA-1,  $\psi = 0.0369$  ( $p = 0.0477$ ), which are marginally significant outcomes. Moreover, the TP, PM, and ZIP models can capture some overdispersion by splitting according to threshold and mixture components. Hence, we focused on Poisson regression for all subsequent analyses.

Moreover, noninformative DVs can lead to unstable models. In Figure A2, seven DVs (1, 2, 7, 8, 10, 14, and 28) are shown to be sparse, with at most 13 nonzeros. Hence, we explored the information content of each DV. Firstly, nonsparsity  $S_j$ , defined as the proportion of nonzero data for each DV, is reported. A refined measure is Shannon's entropy (Shannon 2001)  $H_j$ , which measures the degree of disorder/information of each DV. While  $H_j$  provides no information of the relationship with  $\mathbf{y}$ , the information gain  $IG_j$  evaluates the additional information that the  $j$ th DV provides about the claims  $\mathbf{y}$  with respect to the three classes  $C_b, b = 0, 1, 2^+$ .

Apart from the information content of the DVs, it became clear that the multicollinearity between DVs also affects the stability of a regression model. Figure A3a in Appendix D.2 plots the correlation matrix of  $\text{Corr}(\mathbf{x}_{\bullet 1:j}, \mathbf{y}, \mathbf{a})$ , and the correlations of the DVs with  $\mathbf{y}$  are denoted by  $\rho_j = \text{Corr}(\mathbf{x}_{\bullet j}, \mathbf{y})$ . The correlation matrix shows that the DVs up to 16 (except 9) are nearly uncorrelated with each other, the next up to DV 39 are mildly correlated, and the rest are moderately correlated, reflecting some pattern of these DVs. However, they are

only weakly correlated with  $y$  and  $a$ , showing low signal content of each DV in predicting  $y$  and  $a$ .

Table 2 reports  $\rho_j$ ,  $H_j$ ,  $S_j$ , and  $IG_j$  to quantify the information content of the 65 DVs, flags a DV as “X” when  $H_j < 1$  and  $S_j < 1\%$  (“✓” otherwise), and highlights  $IG_j$  in boldface when  $IG_j > 0$  indicates information gain. Asterisks are added to the DVs’ ID to indicate the two levels of information content in “✓” and boldface, respectively. Twenty DVs with “X” were classified as having low information content. Including them in the more complicated PM and ZIP models led to unstable results. Thus, we dropped them and considered  $J_1 = 65 - 20 = 45$  DVs in the PM and ZIP models, but we considered all  $J_0 = 65$  DVs in the TP models. All the DVs were normalised before analyses to ensure efficient modelling.

Figure A3 in Appendix D.2 plots the Euclidean distance  $d(j, j') = \sqrt{\sum_{i=1}^N (x_{ij} - x_{ij'})^2}$  between the  $(j, j')$  pair of DVs and demonstrates the hierarchical clustering based on  $d(j, j')$ . The results show one major cluster of size 54 and two more smaller clusters (49\*\*, 36\*, 43\*\*, 72\*, 56\*, 63\*) and (55\*, 59\*, 71\*, 27\*\*, 58\*), with increasing pairwise distance from the major cluster. All spare DVs labelled as noninformative are in the major cluster. These DV features guided our interpretation of the selected DVs in subsequent analyses. Refer to Table A1 and Appendix C for the interpretation of these DVs.

**Table 2.** Identification of informative DVs. DVs with one asterisk have  $H_j \geq 1$  and  $S_j \geq 1\%$ . DVs with two asterisks have  $IG_j > 0$ , indicating information gain.

DVs	$\rho$	$H_j$	$S_j(\%)$	$IG_j$	Flag	DVs	$\rho$	$H_j$	$S_j(\%)$	$IG_j$	Flag	DVs	$\rho$	$H_j$	$S_j(\%)$	$IG_j$	Flag	DVs	$\rho$	$H_j$	$S_j(\%)$	$IG_j$	Flag	
1	-0.002	0.08	0.012	0	X	22*	0.003	89.45	12.991	0	✓	39	0.008	0.42	0.076	0	X	59*	0.002	45.89	8.312	0	✓	
2	0.012	0.02	0.002	0	X	23*	-0.001	87.75	12.871	0	✓	43**	-0.053	<b>99.99</b>	<b>13.789</b>	<b>0.002</b>	✓	60**	-0.041	<b>99.93</b>	<b>13.787</b>	<b>0.001</b>	✓	
3*	0.018	7.04	1.231	0	✓	24	0.017	1.02	0.192	0	X	44*	-0.004	19.69	3.795	0	✓	61*	0.018	35.71	6.472	0	✓	
4	-0.011	1.91	0.310	0	X	25	-0.002	0.30	0.046	0	X	45*	0.002	18.61	3.678	0	✓	63*	0.006	32.91	6.462	0	✓	
5	0.003	0.79	0.119	0	X	26*	0.005	17.50	3.990	0	✓	46*	-0.003	90.51	13.008	0	✓	64*	-0.021	61.78	10.149	0	✓	
7	-0.002	0.01	0.001	0	X	27**	<b>-0.060</b>	<b>99.69</b>	<b>13.773</b>	<b>0.002</b>	✓	47*	-0.035	92.68	13.246	0	✓	65	0.0005	1.22	0.295	0	X	
8	0.004	0.10	0.014	0	X	28	-0.004	0.03	0.003	0	X	49**	<b>-0.061</b>	<b>99.98</b>	<b>13.789</b>	<b>0.002</b>	✓	66*	0.008	4.41	1.339	0	✓	
9*	0.010	28.69	5.666	0	✓	29*	0.023	4.41	1.288	0	✓	50*	0.012	6.65	1.247	0	✓	67**	<b>-0.060</b>	<b>99.54</b>	<b>13.766</b>	<b>0.002</b>	✓	
10	-0.002	0.01	0.001	0	X	31*	0.014	15.93	3.698	0	✓	51*	-0.025	67.41	10.718	0	✓	68*	-0.019	76.17	11.953	0	✓	
13	-0.003	0.45	0.069	0	X	32	0.247	4.41	1.229	0	X	52*	-0.039	94.18	13.357	0	✓	69*	-0.007	7.83	1.895	0	✓	
14	-0.006	0.06	0.009	0	X	33*	0.011	39.01	7.957	0	✓	53	0.015	3.00	0.645	0	X	71*	0.006	32.11	6.585	0	✓	
15	-0.0001	0.50	0.076	0	X	34*	-0.001	21.44	5.114	0	✓	54*	0.023	5.03	1.161	0	✓	72*	0.007	41.24	7.861	0	✓	
16	0.006	0.24	0.036	0	X	35*	0.010	35.54	7.257	0	✓	55*	-0.002	21.21	4.424	0	✓	73*	0.023	11.09	2.775	0	✓	
18**	<b>-0.010</b>	<b>99.90</b>	<b>13.785</b>	<b>0.001</b>	✓	36*	0.009	54.80	9.701	0	✓	56*	0.001	34.25	6.654	0	✓	74	0.013	1.03	0.222	0	X	
19*	0.003	77.88	12.129	0	✓	37*	0.024	2.40	0.856	0	✓	57	-0.012	1.23	0.229	0	X	75*	0.029	10.85	2.669	0	✓	
20*	0.006	67.10	10.980	0	✓	38*	0.022	3.84	1.355	0	✓	58*	-0.008	61.74	10.378	0	✓	76*	-0.021	35.50	7.043	0	✓	
																			77*	0.011	13.61	3.354	0	✓

### 3.4. Two-Stage Threshold Poisson Model

The TP model fit Poisson regression in Section 2.1.1 twice at stages 1 and 2 with the aim of classifying drivers into safe and risky groups at stage 1 and determining predictive DVs for each group at stage 2 using a single-component Poisson model. The DVs for TP models were selected from  $J = J_0 = 65$  DVs.

At stage 1, lasso-regularised Poisson regression models were trained through resampling and applied to predict claim frequencies for all drivers. To ensure model robustness and reduce overfitting, we repeated regularised Poisson regression  $R = 100$  times with 70% simulated subsamples of size  $N_r = 9910$  each and selected DVs for each repeat. For each repeat  $r$ , the optimal  $\lambda_{\min,r}$  was selected with  $K = 10$  (default setting) folds CV. Then, the DVs most frequently selected were identified using a weighted count  $I_j$  in (A3) based on the RMSE. There were  $J^{T1} = 52$  selected DVs for TPL-1 model and  $J^{T1} = 39$  DVs for TPA-1 model. The details are provided in Appendix A. Then, Poisson regression models with the selected DVs were refitted to all drivers. Parameter estimates  $\beta_j^{T1}$  are reported in Table A2 in Appendix E under  $\beta_j^{T1}$ . To visualise these coefficients, Figure 2a plots the heat map of  $\beta_j^{T1}$  for all PT-1 models. Let  $J_S^{T1}$  be the number of significant  $\beta_j^{T1}$  with a  $p$  value  $< 0.05$ . For the TPA-1 model, Table A2 shows that  $J_S^{T1} = 13$  out of  $J^{T1} = 44$  selected DVs are significant. Table 3a shows that TPL-1 and TPE-1 provided a similar selected number  $J^{T1}$  of DVs. The same applied to models TPA-1 and TPN-1 with adaptive weights. As feature selection is important, we selected the best model in each group, and they are highlighted in Table 3a.

**Table 3.** Model performance measures for stage 1 TP and ZIP models, stage 2 TP and PM models, and final selection.

Models	$J^{T1}$	AIC	BIC	MSE	Model	$J^{T1}$	AIC	BIC	MSE
<b>Stage 1 Threshold Poisson models</b>									
TPL-1	52	8020	8421	0.0866	TPA-1	39	7998	8300	0.0866
TPE-1	57	8029	8468	0.0866	TPN-1	39	8000	8301	0.0866
<b>Zero-inflated Poisson models</b>									
ZIPL	$J_0^Z=0$ $J_c^Z=44$	8028	8368	0.0868	ZIPA	$J_0^Z=4$ $J_c^Z=45$	8155	8258	0.0873

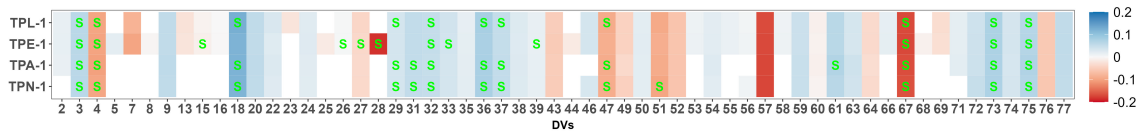
**a** Model performance measures for the stage 1 TP and ZIP models.

		$J_{hh}^{T2}$		$J_{Sh}^{T2}$		Performance measures				Ranks				Sum
		Low	High	Low	High	BIC	MSE	$\rho$	AUC	BIC	MSE	$\rho$	AUC	$\mathfrak{R}_M$
$\tau_i$	Models	<b>Stage 2 threshold Poisson model</b>												
$\tau_{0,08}$	TPLL-1	0	2	0	2	-	-	-	-	-	-	-	-	-
	TPLE-1	0	1	0	1	-	-	-	-	-	-	-	-	-
	TPLA-1	16	22	0	3	8286	<b>0.0863</b>	<b>0.1336</b>	0.5990	4.0	4.0	5.0	4.0	<b>17.0</b>
	TPLN-1	23	19	1	1	8326	0.0864	0.1261	<b>0.6119</b>	3.0	2.0	2.0	5.0	12.0
	TPAL-1	5	0	2	0	-	-	-	-	-	-	-	-	-
	TPAE-1	17	2	4	1	<b>8132</b>	0.0867	0.1176	0.5958	5.0	1.0	1.0	3.0	10.0
	TPAA-1	25	20	4	4	8347	<b>0.0863</b>	0.1324	0.5757	2.0	4.0	4.0	1.0	11.0
	TPAN-1	27	19	3	4	8356	<b>0.0863</b>	0.1323	0.5781	1.0	4.0	3.0	2.0	10.0
$\tau_{0,09}$	TPLL-1	26	3	4	2	<b>8204</b>	0.0864	0.1237	0.6186	8.0	1.5	1.0	7.0	17.5
	TPLE-1	39	4	7	2	8321	0.0863	0.1282	0.6111	5.0	3.5	4.0	5.0	17.5
	TPLA-1	38	14	6	1	8398	<b>0.0862</b>	0.1332	0.6129	3.0	6.5	7.0	6.0	<b>22.5</b>
	TPLN-1	41	11	8	1	8400	<b>0.0862</b>	0.1291	<b>0.6276</b>	2.0	6.5	5.0	8.0	21.5
	TPAL-1	35	3	10	3	8281	0.0863	0.1263	0.5972	7.0	3.5	2.5	4.0	17.0
	TPAE-1	38	3	10	2	8311	0.0864	0.1263	0.5936	6.0	1.5	2.5	3.0	13.0
	TPAA-1	30	15	9	0	8341	<b>0.0862</b>	0.1292	0.5921	4.0	6.5	6.0	2.0	18.5
	TPAN-1	34	18	9	1	8404	<b>0.0862</b>	<b>0.1338</b>	0.5890	1.0	6.5	8.0	1.0	16.5
$\tau_{0,10}$	TPLL-1	49	3	12	0	8400	0.0862	0.1281	<b>0.6284</b>	4.0	2.5	2.0	8.0	16.5
	TPLE-1	51	5	13	0	8434	0.0861	0.1319	0.6276	2.0	5.0	4.0	7.0	18.0
	TPLA-1	38	18	12	0	8427	0.0860	0.1329	0.6218	3.0	7.0	5.0	6.0	21.0
	TPLN-1	42	23	11	0	8507	<b>0.0859</b>	<b>0.1398</b>	0.6170	1.0	8.0	8.0	5.0	<b>23.0</b>
	TPAL-1	38	6	11	0	8327	0.0862	0.1313	0.5944	7.0	2.5	3.0	2.0	14.5
	TPAE-1	38	5	11	0	<b>8321</b>	0.0863	0.1271	0.5933	8.0	1.0	1.0	1.0	11.0
	TPAA-1	35	17	9	0	8396	0.0861	0.1353	0.6005	6.0	5.0	7.0	3.0	21.0
	TPAN-1	35	17	9	0	8397	0.0861	0.1341	0.6012	5.0	5.0	6.0	4.0	20.0
$\tau_{0,11}$	TPLL-1	50	6	13	1	8431	0.0861	0.1310	<b>0.6214</b>	4.0	2.0	1.0	8.0	15.0
	TPLE-1	51	12	13	0	8492	0.0860	0.1391	0.6200	3.0	4.5	5.0	7.0	19.5
	TPLA-1	41	24	13	0	8503	0.0859	0.1429	0.6183	2.0	6.5	7.0	6.0	21.5
	TPLN-1	43	28	12	0	8554	<b>0.0857</b>	<b>0.1457</b>	0.6083	1.0	8.0	8.0	5.0	<b>22.0</b>
	TPAL-1	38	8	15	1	<b>8344</b>	0.0861	0.1347	0.6013	8.0	2.0	2.0	2.0	14.0
	TPAE-1	38	9	15	0	8353	0.0861	0.1357	0.6008	7.0	2.0	3.0	1.0	13.0
	TPAA-1	34	18	15	1	8396	0.0859	0.1366	0.6031	6.0	6.5	4.0	4.0	20.5
	TPAN-1	35	20	15	1	8422	0.0860	0.1403	0.6029	5.0	4.5	6.0	3.0	18.0
<b>Poisson mixture</b>														
	PML	45	18	4	3	<b>8551</b>	0.0832	0.2345	<b>0.6237</b>	4.0	1.0	1.0	4.0	10.0
	PME	45	35	2	4	8704	0.0825	0.2581	0.6216	2.0	2.0	2.0	3.0	9.0
	PMA	39	40	6	5	8692	<b>0.0805</b>	0.2972	0.6003	3.0	4.0	3.0	1.0	<b>11.0</b>
	PMN	45	44	1	7	8781	0.0811	<b>0.3002</b>	0.6073	1.0	3.0	4.0	2.0	10.0

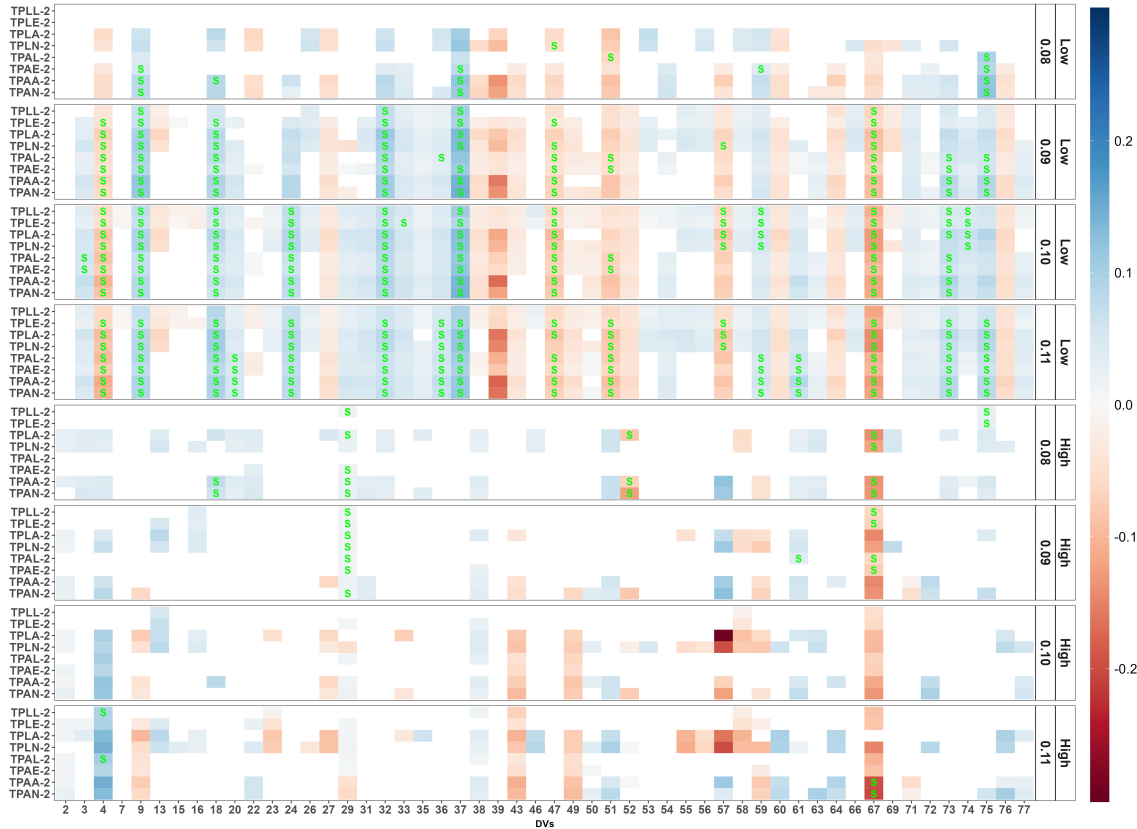
**b** Number of significant selected DVs and performance measures for stage 2 TP and PM models.

Models	Performance measures for count models						Ranks						Sum
	Log	Quad	Spher	RankProb	Dawid	SqErr	Log	Quad	Spher	RankProb	Dawid	SqErr	
$\tau_{0,08}$ : TPLA	0.1360	-0.8550	-1.0015	0.0778	0.0956	0.0906	2.0	4.0	4.0	3.0	4.0	3.0	20.0
$\tau_{0,09}$ : TPLA	0.1355	-0.8549	-1.0009	0.0777	0.7710	<b>0.0900</b>	3.5	3.0	3.0	4.0	1.0	5.0	19.5
$\tau_{0,10}$ : TPLN	0.1355	-0.8545	-1.0005	0.0780	0.5285	0.0905	3.5	2.0	2.0	2.0	2.0	4.0	15.5
$\tau_{0,11}$ : TPLN	0.1363	-0.8541	-1.0002	0.0783	0.4674	0.0912	1.0	1.0	1.0	1.0	3.0	2.0	9.0
PMA	<b>0.1190</b>	<b>-0.8613</b>	<b>-1.0235</b>	<b>0.0760</b>	<b>-0.2932</b>	0.0929	5.0	5.0	5.0	5.0	5.0	1.0	<b>26.0</b>

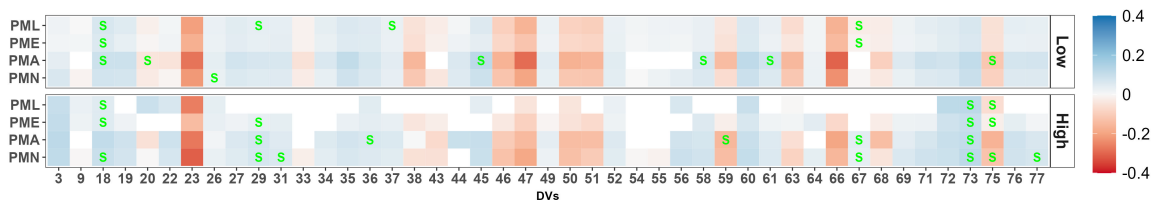
**c** Model performance measures for final selection of stage 2 TP and PM models.



(a) Coefficients  $\beta^{T1}$  for the selected DVs of stage 1 TP regression models.



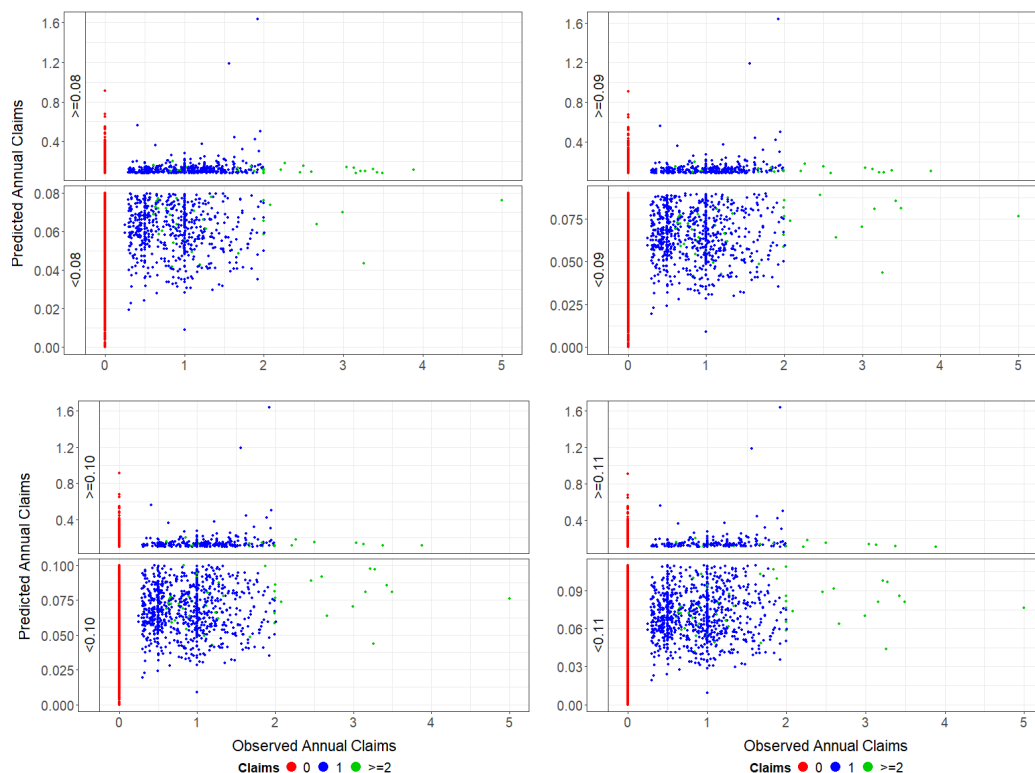
(b) Coefficients  $\beta_{Li}^{T2}, \beta_{Hi}^{T2}, h = 1, \dots, 4$  for the stage 2 TP regression models.



(c) Coefficients  $\beta_{Lj}^M$  and  $\beta_{Hj}^M$  for PM regression models according to low- and high-claim groups.

**Figure 2.** Heat map of coefficients, with significant values denoted by “S”.

At stage 2, predicted claims  $\hat{y}_i = \mu_i$  were calculated using the fitted means in (1) and  $\beta^{T1}$ . Then, the predicted annual claim frequencies  $\hat{a}_i = \hat{y}_i/n_i$  were calculated, and drivers were classified into low- and high-claim groups according to  $\hat{a}_i < \tau_h$  and  $\hat{a}_i \geq \tau_h$ , respectively. We considered four thresholds  $\tau = \{\tau_h\} = (0.08, 0.09, 0.10, 0.11)$ , and the proportion  $\mathcal{P}_h$  of drivers classified into the low claim group out of all drivers was (0.70, 0.79, 0.85, 0.90), respectively. Figure 3 shows how the drivers were classified to low- and high-claim groups according to the four thresholds  $\tau$  using  $\hat{a}_i$  from TPA-1 and visualises the relationship between the observed  $a_i$  and predicted  $\hat{a}_i$ . We attribute the nonlinear pattern in these scatter plots partially to the impact of driving behaviour revealed by the DVs.



**Figure 3.** Scatter plots of observed annual claim frequencies  $a_i$  against predicted annual claim frequencies  $\hat{a}_i$  using TPA-1 model cross-classified with low- and high-claim groups by the four thresholds, with colour indicating claims  $y_i = 0, 1, \geq 2$ .

To improve model robustness and reduce overfitting, subsamples  $S_{1:R}$  of size  $N_r = 0.7N$  were drawn again, and each  $S_r$  was further split into two groups

$$\mathcal{G}_{L,rh}^{T2} = \{y_i \in S_r : \hat{a}_i < \tau_h\} \quad \text{and} \quad \mathcal{G}_{H,rh}^{T2} = \{y_i \in S_r : \hat{a}_i \geq \tau_h\}, \quad h = 1, \dots, 4 \quad (11)$$

where T2 indicates stage-2 of the TP model. Then, regularised Poisson regression was applied to each  $\mathcal{G}_{L,rh}^{T2}$  and  $\mathcal{G}_{H,rh}^{T2}$ . Let the index sets  $\mathcal{I}_{Lh}^\beta$  and  $\mathcal{I}_{Hh}^\beta$  for nonzero coefficients (that is, selected at least once from  $S_r$ ) be defined similar to  $\mathcal{I}^\beta$  in (A2) for  $h = 1, \dots, 4$ . Then,  $\beta_{Lh} = (\beta_{Lh,j \in \mathcal{I}_{Lh}^\beta})$  and  $\beta_{Hh} = (\beta_{Hh,j \in \mathcal{I}_{Hh}^\beta})$  are averaged parameter estimates for the low- and high-claim groups, respectively, obtained in a similar manner to  $\beta$  defined in (A1), and  $I_{Lh}^{T2}$  and  $I_{Hh}^{T2}$  are importance measures based on the RMSE  $r_{Lh}$  and RMSE  $r_{Hh}$  defined similarly to  $I$  in (A3);  $J_{Lh}^{T2}$  and  $J_{Hh}^{T2}$  are the number of frequently selected DVs out of  $\beta_{Lh}$  and  $\beta_{Hh}$ , with  $I_{Lhj} > 43$  ( $62 \times 0.70$ ;  $\mathcal{R}_1 = 0.70$  for  $\tau_{0.08}$  and  $\max(I_j, j \in \mathcal{I}^\beta) = 62$  is the lower threshold of  $I_j$ ), 49 ( $\tau_{0.09}$ ), 53 ( $\tau_{0.10}$ ), 56 ( $\tau_{0.11}$ ),  $I_{Hhj} > 19$  ( $62 \times 0.30$ ), 13, 9, and 6, respectively, (and dropped otherwise). Poisson regression models with various selected DVs were refitted to the low- and high-claim groups for each  $h$ . Table A3 in Appendix E reports the parameter estimates  $\beta_{Lh}^{T2}, \beta_{Hh}^{T2}$  of the best model (TPLA-2 for  $\tau = 0.08, 0.09$  and TPLN-2 for  $\tau = 0.10, 0.11$ ) when the stage 1 model was TPL-1 (from  $J^{T1} = 52$  selected DVs) or TPA-1 ( $J^{T1} = 39$ ). Table 3b reports the number  $J_{Lh}^{T2}, J_{Hh}^{T2}$  of selected  $\beta_{Lhj}^{T2}$  and  $\beta_{Hhj}^{T2}$  and the number  $J_{LS_h}^{T2}, J_{HS_h}^{T2}$  of significant  $\beta_{Lhj}^{T2}$  and  $\beta_{Hhj}^{T2}$  with  $p$  values  $< 0.05$ .

To visualise these coefficients, Figure 2a presents the heat map for models in the low- and high-claim groups. It shows that the DVs, which were mostly selected and significant in the low-claim groups, are 4, 9\*, 18\*\*, 24, 32, 37\*, 47\*, 57, 67\*\*, 73\*, and 74, while the least are 2, 7, 15, 16, 23\*, 46\*, and 53. For the high-claim group, DVs 4, 29\*, 52\*, and 67\*\* were mostly selected. The information content of each DV is indicated by asterisks. See Table 2 for details and Table A1 for the interpretation of these DVs. We observed that there were more selected DVs for the low claim group with thresholds  $\tau_{2:4} = 0.09, 0.10, 0.11$ , and the

selected DVs are relatively more informative. Two DVs, 4 and 67, were selected in both the low- and high-level claim groups but with differential effects: negative for the low-claim group and positive for the high-claim group for DV 4, whereas DV 67 had a consistent negative effect for both groups.

For model selection, Table 3b summarises the model performance measures, BIC, MSE,  $\rho$ , and AUC (see Section 2.3) using all data for 32 models, with 8 models under each threshold. The two criteria, BIC and AUC, in Table 3b were averaged over the two groups using the ratio  $\mathcal{R}_h$  in Table A3. For each threshold, top ranked measures and the sum of rank  $\mathfrak{R}_{\mathcal{M}}$  in (10) are boldfaced and yellow highlighted. We first dropped those models with  $J_h^{T2}, J_{Sh}^{T2} = 0$  for either group and chose the best model  $\mathcal{M}$  with the top  $\mathfrak{R}_{\mathcal{M}}$ . The best stage two model is TPLA-2 for  $\tau = 0.08, 0.09$  and TPLN-2 for  $\tau = 0.10, 0.11$ . The results will be compared for the PM and ZIP models in Section 3.7.

### 3.5. Poisson Mixture Model

To facilitate driver classification, we considered lasso-regularised PM models. To robustify our results, we again performed 70% resampling to obtain  $R = 100$  subsamples of size  $N_r = 9910$ . The parameters were selected from the  $J = J_1 = 45$  more informative DVs to provide stable results (see Section 3.3). In each subsample  $\mathcal{S}_r$ , the regularised PM model was estimated using  $K = 10$  folds CV. Then, the drivers were classified into  $\mathcal{G}_{L,r}^M$  and  $\mathcal{G}_{H,r}^M$  according to  $\hat{z}_{ig} \geq 0.5$  or  $< 0.5$ , respectively, where  $\hat{z}_{ig}$  was defined in (3). Let the index sets  $\mathcal{I}_L^\beta$  and  $\mathcal{I}_H^\beta$  for a nonzero coefficient (that is, selected at least once from  $\mathcal{S}_r$ ) be defined similar to  $\mathcal{I}^\beta$  in (A2). Then,  $\beta_L = (\beta_{L,j \in \mathcal{I}_L^\beta})$  and  $\beta_H = (\beta_{H,j \in \mathcal{I}_H^\beta})$  are averaged parameter estimates for the low- and high-claim groups, respectively, obtained in a similar manner to  $\beta$  defined in (A1);  $I_L^M$  and  $I_H^M$  are importance measures based on the RMSE  $_{r,L}^M$  and RMSE  $_{r,H}^M$  defined similar to  $I$  in (A3);  $\mathcal{R}^M$  is the average ratio of the low group size over  $R = 100$  subsamples; and  $J_L^M$  and  $J_H^M$  are the number of selected DVs out of  $\beta_L$  and  $\beta_H$ , with  $I_{Lj}^M > 43$  ( $62 \times 0.69$  and  $\mathcal{R}^M = 0.69$  for PML),  $45$  ( $62 \times 0.73$  for PMA),  $I_{Hj}^M > 19$  ( $62 \times 0.31$  for PML), and  $17$  ( $62 \times 0.27$  for PMA) similar to  $J^{T1}$ . We note that some subsamples had too low of sample size ( $< 200$ ) for the high-claim group or too low differences ( $< 0.005$ ) of observed annual claim frequencies between the two groups or both. Both criteria indicate ineffective grouping and should be eliminated. Consequently, 172 and 113 subsamples were drawn for the PML and PMA models, respectively, in order to collect 100 effective subsamples.

To obtain the overall parameter estimates  $\beta_L^M$  and  $\beta_H^M$ , the selected DVs were refitted to the PM model again. Table A4 in Appendix E reports  $\beta_L^M$  and  $\beta_H^M$  of the PML and PMA models, together with  $I_L^M$  and  $I_H^M$ . Table 3b reports the number of DVs selected,  $J_L^M, J_H^M$  ( $\beta_{Lj}^M, \beta_{Hj}^M \neq 0$  and  $I_{Lj}^M, I_{Hj}^M > 62$ ), and the number of significant selected DVs,  $J_{LS}^M, J_{HS}^M$ , for each PM model. We note that the PM models had more selected and significant DVs for the high-claim group in general than the TP models. Figure 2c plots the heat map of the parameter estimates of the two groups for the four PM models. Across all four models, the two sets of mostly selected and significant variables are (**18\*\***, **20\***, **26\***, **29\***, **37\***, **45\***, **58\***, **61\***, **67\*\***, **75**) and (**18\*\***, **29\***, **31\***, **36\***, **59\***, **67\*\***, **73\***, **75\***, **77\***) for the low- and high-claim groups, respectively. These two sets of significant DVs are quite different from those of TP models, as only two DVs from each group (in boldface) were also selected by TP models. Again, DV 67 was selected by both groups, which was the same as the case of the TP models. To select the best PM model, Table 3b reports the performance measures BIC, MSE,  $\rho$ , and AUC. According to the sum of ranks  $\mathcal{R}_{\mathcal{M}}$  in (10), the PMA model was selected. For the selected PMA model, Table A4 shows that there were  $J_L^M = 39$  DVs selected for the low-claim group and  $J_H^M = 40$  DVs selected for the high-claim group, of which six (18, 20, 45, 58, 61, 75) and five (29, 36, 59, 67, 73) DVs are significant. See point 5 in Appendix B for the implementation of the PM models, Appendix C for the interpretation of the significant selected DVs, and Section 3.7 for the implication of these DVs on risky driving.



### 3.6. Zero-Inflated Poisson Model

Since 92% of the claims are zero, we applied the ZIP model in (5) and (6) (Lambert 1992; Zeileis et al. 2008) to capture the structural zero portion of the claims and test if the structural zero claim group should be included in modelling claims. As with the TP and PM models, we drew subsamples  $\mathcal{S}_{1:100}$ , each with  $N_r = 9910$  drivers, and ZIP lasso-regularised regression was applied to each  $\mathcal{S}_r$  to robustify the selection of DVs. The procedures were similar to the cases of the TP and PM models. As with the PM model, the DVs are selected from  $J = J_1 = 45$  more informative DVs to provide stable results.

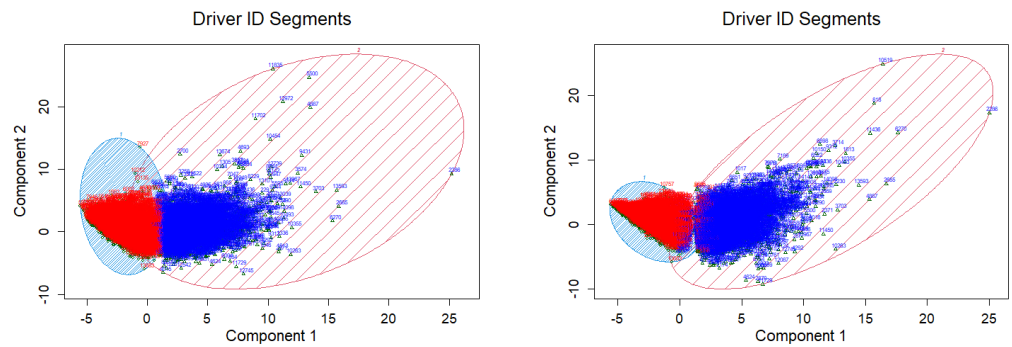
Let  $\mathcal{I}_0^\beta$  and  $\mathcal{I}_c^\beta$  be the index sets of nonzero parameter estimates (that is, selected at least once from  $\mathcal{S}_r$ ) in the zero and count models, respectively, defined in a similar manner to  $\mathcal{I}^\beta$  in (A2). Then, the averaged parameter estimates  $\beta_0 = (\beta_{0,j \in \mathcal{I}_0^\beta})$  for the zero model and  $\beta_c = (\beta_{c,j \in \mathcal{I}_c^\beta})$  for the count model are obtained in a similar manner to  $\beta$  in (A1);  $I_0^Z$  and  $I_c^Z$  are importance measures based on the RMSE  $Z_{0,r}$  and RMSE  $Z_{c,r}$ , respectively, defined similarly to  $I$  in (A3), and  $J_0^Z$  and  $J_c^Z$  are the number of selected DVs out of  $\beta_0$  and  $\beta_c$ , with  $I_{0,j}^Z > 62$  and  $I_{c,j}^Z > 62$  similar to  $J^{T1}$ . Next, the  $J_0^Z$  and  $J_c^Z$  selected DVs were refitted to the ZIP model for all data to obtain the overall parameter estimates  $\beta_0^Z$  and  $\beta_c^Z$ . Parameters  $\beta_0, \beta_c$  were averaged before refit, parameters  $\beta_0^Z, \beta_c^Z$  were averaged after refit, and the importance measures  $I_0^Z, I_c^Z$  are reported in Table A4. Table 3a reports the number of selected DVs  $J_0^Z$  and  $J_c^Z$ , and performance measures AIC, BIC, and MSE for the ZIPL and ZIPA models following the regularisation choices from the two chosen TPL-1 and TPA-1 models. Between the two ZIP models, the ZIPA model was chosen, because it had nonzero DVs selected for the zero component and a lower BIC. However, the MSE was the highest among all the models shown in Table 3b, indicating low predictive power. More importantly, the zero model estimates the probability of structural zero among all zero, but it does not guide the classification of safe and risky drivers, because safe drivers can claim less but not necessarily none, and risky drivers can claim none by luck or for a no-claim bonus. Hence, drivers classified to the structural zero claim group are not necessarily safe drivers. As a result, the ZIPA model was excluded from model comparison and selection. See Appendix B.6 for the implementation of ZIP models.

### 3.7. Model Comparison and Selection

We compared the performance of the TP and PM models in terms of claim prediction, predictive DVs selection, and driver classification. Table 3b displays the performance of all 32 TP models and four PM models. The best TP model for each threshold and the best PM model were selected according to  $\mathfrak{R}_M$  in (10) using four measures. Among the selected TPLA-2 ( $\tau = 0.08, 0.09$ ), TPLN-2 ( $\tau = 0.10, 0.11$ ), and PMA models, Table 3c shows that the PMA model succeeded through the final selection using six count model scores, confirming the superiority of PM model in many aspects and its preferability over the Poisson model. For an interpretation of the significant selected DVs (29, 36, 59, 67, 73; all with positive coefficients except 59) using the PMA model, *risky driving is associated with more frequent severe brake to slow-down at weekday and weekend nights, as well as more frequent severe right turns at the junction at weekday nights and Friday rush time.*

Apart from the numerical measures, we also visualised their performance using ROC curves. Section 2.3 introduces the AUC according to three classes of drivers with  $y_i = 0, 1, \geq 2$  or an overall class of all drivers. For each of these classes, the AUC was drawn for the binary classifier of low- and high-claim groups comparing the predicted group, with the *proxy observed* group of each driver estimated by K-means clustering using  $J^{T1} = 52$  selected DVs (from TPL-1 model) for all stage 2 TP models and  $J_1 = 45$  informative DVs for the PM model. Figure 4 plots the two clusters using the first two principal components (PCs) that explain 22.11% and 17.44% of variance, respectively, using selected DVs for the two cases. The results show that the first PC could separate the two clusters well for both cases. For the stage 2 TP (PM) model,  $N_1 = 9450$  (9372) claims were assigned to the low claim cluster accounting for 67% (66%) of drivers using K-means clustering. These two proportions of the low claim group using K-means clustering are not far away from the

estimated proportions  $\mathcal{P}_h = 0.7$  to  $0.89$  in Tables A3 using the TPL-1 model, as well as  $\mathcal{P}^M = 0.73$  using the PMA model.



(a) Using  $J_1^T = 52$  selected DVs from TPL-1. (b) Using  $J^M = 45$  informative DVs for PM.

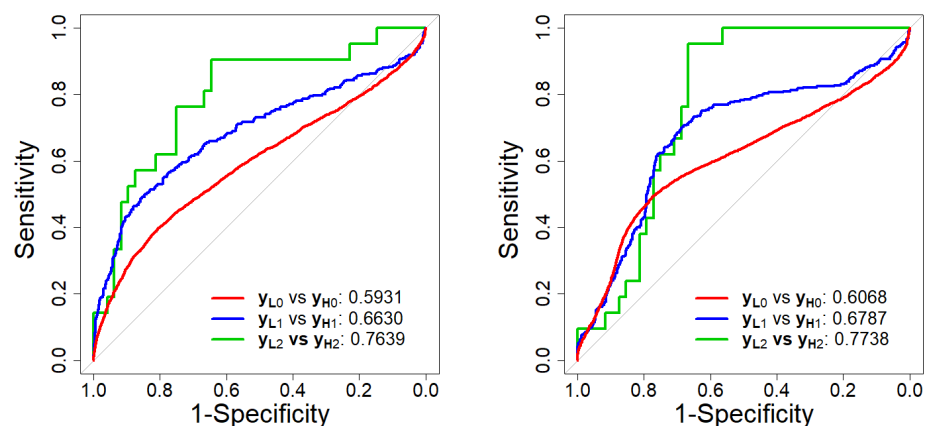
**Figure 4.** K-means clustering analysis to segment drivers into low-claim cluster in red, with blue shade ellipse and high-claim cluster in blue, as well as red shade ellipse for TP and PM models.

Figure 5a–e plots the ROC curves with AUC values using classifier  $\hat{y}_i/n_i$  for the best stage 2 TP models under each threshold and classifier  $z_{i1}$  in (3) for the best PMA model. The AUCs have been calculated into subgroups  $y_i = 0$  (red),  $1$  (blue),  $\geq 2$  (green) in Figure 5a–e, and all the data are shown in Figure 5f. The zigzag patterns of the green lines indicate small sample sizes for drivers with  $y_i \geq 2$ . Table 3b reports the overall AUC values, which are the weighted averages of the three AUC values in each of Figure 5a–e and show that model TPLN-2 when  $\tau = 0.10$  is the best classifier of low- and high-claim groups, while PMA displays the third classifying power. However, the accuracy of the results depends on whether K-means clustering can estimate the true latent groups well.

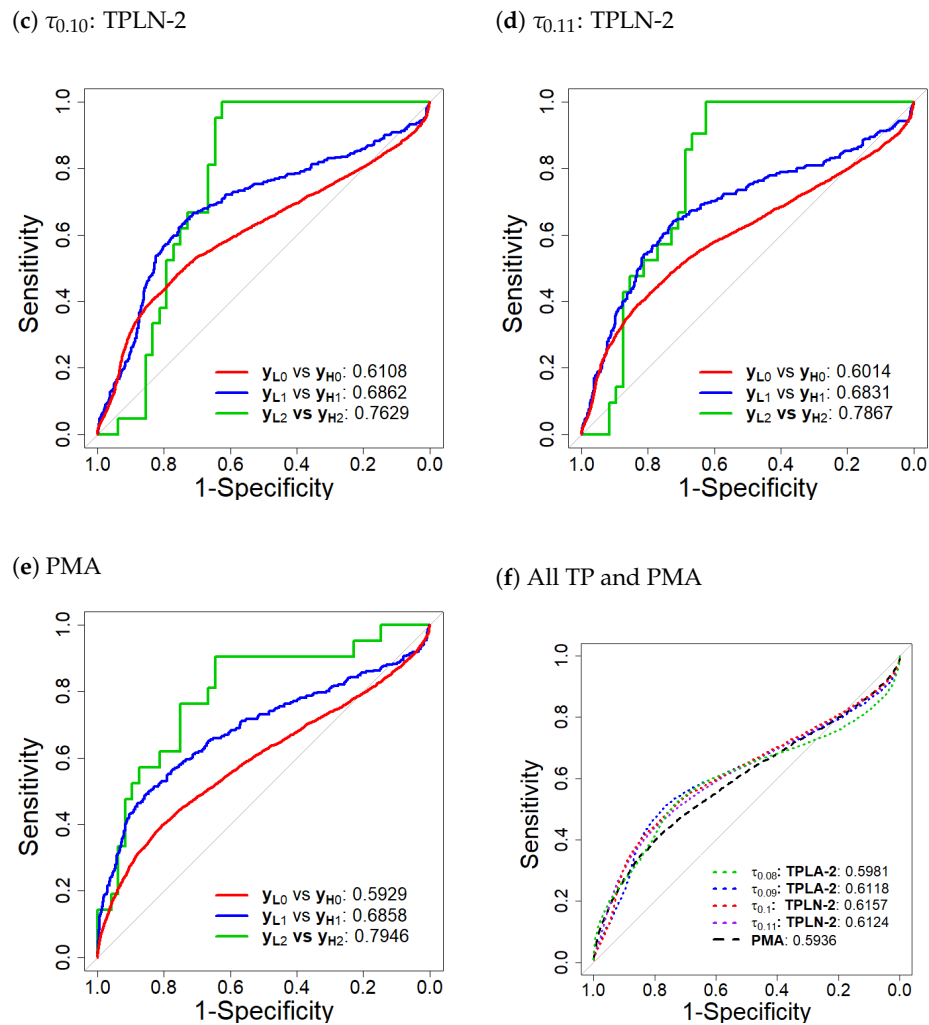
Since zero-claim drivers are not necessarily safe drivers after considering the DVs, we expect some zero-claim drivers to be risky and non-zero-claim drivers to be safe. This disagreement reflects the impact of DVs on assessing driving risk apart from the claim information. Zero disagreement (ZD) is defined as the proportion, out of all drivers, of those zero-claim drivers classified as risky and non-zero claim drivers classified as safe. This ZD is 3.2% for the best PMA model. This small ZD is due to the low MSE showing agreement between the predicted and observed claims.

(a)  $\tau_{0.08}$ : TPLA-2

(b)  $\tau_{0.09}$ : TPLA-2



**Figure 5.** Cont.



**Figure 5.** ROC curve and AUC values for (a–d) the four best stage 2 TP; (e) PMA model; and (f) four best stage 2 TP models and one PM model.

#### 4. UBI Experience Rating Premium

In the context of general insurance, a common approach for assessing risk in the typical short-tail portfolio involves multiplying predicted claims frequency by claims severity to determine the risk premium. This derived risk premium is subsequently factored into the profit margin, alongside operating expenses, to determine the final premium charged to customers. This paper centers on claims frequency, and in the premium calculation discussed herein, we assume that claim severity remains constant. Consequently, the premium calculation relies on predicting claims frequency.

The traditional experience rating method prices premiums using historical claims and offers the same rate for drivers within the same risk group (low/high or safe/risky). If individual claim history is available, premiums can be calculated using individual claims relative to overall claims—both from historical records. However, although this extended historical experience rating method can capture the individual differences of risk within a group, it still fails to reflect drivers’ recent driving risk. The integration of telematic data enables us to tailor pricing to *current* individual risks. This enhanced method is called the *UBI experience rating* method. We leverage premium pricing as a strategic approach to refine our pricing methodology.

Suppose that a new driver  $i$  was classified to claim group  $g$  with index set  $C_g$  of all drivers in this group and  $i \in C_g$ . Let  $P_{it}$  be his premium for year  $t$ ,  $L_{i,t-1}$  be the historical claim/loss in year  $t - 1$ ,  $L_{t-1}^g = \sum_{i' \in C_g} L_{i',t-1}$  be the total claim/loss from the claim group  $g$  that driver  $i$  was classified to, and  $P_{t-1}^g = \sum_{i' \in C_g} P_{i',t-1}$  be the total premium from the

claim group  $g$ . Moreover, suppose that the best PMA model was trained using the sample of drivers. Let  $x_{i\bullet,t}$  be the observed DVs for driver  $i$  at time  $t$ ,  $g = 1$  safe group ( $g = 2$  risky group) be the classified group if the group indicator  $\hat{z}_{i1} > 0.5$  (otherwise),  $\hat{y}_{i,t}$  in (4) be the predicted claim frequency given  $x_{i\bullet,t}$ ,  $\hat{y}_t^g = (\sum_{i' \in C_g} \hat{y}_{i',t}) / N_g$  be the average predicted claim frequencies from the claim group  $g$  that driver  $i$  was classified to, and  $N_g$  be the size of group  $g$ .

Using the proposed *UBI experience rating* method, the premium  $P_{it}$  for driver  $i$  in year  $t$  is given by

$$P_{it,\varkappa} = (1 + R_{i,t}^\Delta) \times \bar{P}_t^g \times E_{it} \times F + \bar{P}_t^* \times E_{it} \times (1 - F) \tag{12}$$

where  $\bar{P}_t^g$  is the group average annual premium in period  $t$  from the group data,  $\bar{P}_t^*$  is the average annual premium from all data or some other data source,  $F$  is the credibility factor (Dean 1997),  $E_{it}$  is the exposure of driver  $i$ , and  $R_{i,t-1}^\Delta$  is the individual adjustment factor to the overall group loss ratio given by

$$R_{i,t}^\Delta = R_{i,t-1}^{\Delta,H} + \varkappa R_{i,t}^{\Delta,UB}, \tag{13}$$

which is the sum of the *historical* loss rate change adjustment  $R_{i,t-1}^{\Delta,H}$  and weighted *UBI predicted* loss rate change adjustment  $R_{i,t-1}^{\Delta,UB}$ ;  $\varkappa \in [0, 1]$  is the *UBI policy* parameter to determine how much UBI adjustment is applied to  $R_{i,t-1}^{\Delta,UB}$  in  $R_{i,t}^\Delta$  when updating the premium to account for current driving behaviour. The *historical* loss rate change  $R_{i,t-1}^{\Delta,y}$ , historical individual loss ratio  $R_{i,t-1}$ , and historical group loss ratio  $R_{t-1}^g$  are, respectively,

$$R_{i,t-1}^{\Delta,H} = \frac{R_{i,t-1} - R_{t-1}^g}{R_{t-1}^g}, \quad R_{i,t-1} = \frac{L_{i,t-1}}{P_{i,t-1}}, \quad \text{and} \quad R_{t-1}^g = \frac{L_{t-1}^g}{P_{t-1}^g}. \tag{14}$$

The *UBI predicted* loss rate change  $R_{i,t}^{\Delta,UB}$ , *UBI predicted* individual loss ratio  $R_{i,t}^y$ , and *UBI predicted* group loss ratio  $R_t^{y,g}$  are, respectively,

$$R_{i,t}^{\Delta,UB} = \frac{R_{i,t}^y - R_t^{y,g}}{R_t^{y,g}}, \quad R_{i,t}^y = \frac{\hat{y}_{i,t}}{P_{i,t}}, \quad \text{and} \quad R_t^{y,g} = \frac{\hat{y}_t^g}{P_t^g}.$$

The credibility factor  $F$  is the weight of the best linear combination between the premium estimate  $(1 + R_{i,t}^\Delta) \times \bar{P}_t^g$  using the sample data to the premium estimate  $\bar{P}_t^*$  using all data or data from another source to improve the reliability of the premium estimate  $P_{it}$ . The credibility factor increases with the business size and, hence, the number of drivers in the sample. Dean (1997) provided some methods to estimate  $F$  and suggested full credibility  $F = 1$  when the sample size  $N$  is large enough, such as above 10,000 in an example. As this requirement is fulfilled for the telematic data with size  $N = 14,157$ , and all data are used to estimate the chosen PMA model, a full credibility of  $F = 1$  was applied. In cases where insured vehicles are less in number in the sample, the credibility factor  $F$  may vary, and external data sources may be used to improve the reliability of the premium estimate. Moreover, as the selected PMA model can classify drivers, the premium calculation can focus on the classified driver group to provide a more precise premium calculation.

We give an example to demonstrate the experience rating method and its extension to UBI. Suppose that driver  $i$  is classified as a safe driver ( $g = 1$ ) in a driving test and wants to buy auto insurance for the next period ( $E_{it} = 1$ ). As summarised in Table 4, the annual premium for the safe group is  $\bar{P}_t^1 = \$300$  and for the risky group is  $\bar{P}_t^2 = \$500$ . Driver  $i$  has recorded  $L_{i,t-1} = 0.2$  in annual claim frequency and paid an annual premium of  $P_{i,t-1} = \$500$  before. The safe group has recorded an average of  $L_{t-1}^1 = 0.1$  in annual claim frequency and paid an annual premium if  $P_{t-1}^1 = \$310$  per driver before. The risky group has recorded an average of  $L_{t-1}^2 = 0.3$  claims/loss and paid  $P_{t-1}^2 = \$510$  in annual premium per driver before. Driver  $i$  has more claims than the average of a safe group

before. According to these historical claim frequencies, driver  $i$  is expected to be relatively more risky than the average of the safe group, so he should pay more.

To illustrate the UBI experience rating method, additional assumptions about the predicted annual claim frequencies for driver  $i$  have been added to the last row of Table 4. Assume that driver  $i$  has a predicted annual claim frequency  $\hat{y}_{i,t-1} = 0.15$  before; then, that of the safe group is  $\hat{y}_{t-1}^1 = 0.105$ , and that of the risk group is  $\hat{y}_{t-1}^2 = 0.305$ . This suggests that driver  $i$  operates his vehicle more safely than his historical claims indicate. This information is summarised in Table 4.

**Table 4.** Assumptions summary in a case study in thousand dollars.

	Driver $i$ (Safe)	Safe Group	Risky Group
Average annual premium $\bar{P}_t^g$	-	0.3	0.5
Historical annual premium $P_{i,t-1}, P_{t-1}^g$	0.5	0.31	0.51
Historical annual claims $L_{i,t-1}, L_{t-1}^g$	0.2	0.1	0.3
Predicted annual claim frequencies $\hat{y}_{i,t-1}, \hat{y}_{t-1}^g$	0.15	0.105	0.305

Taking the policy parameter  $\varkappa = 1$ , the UBI experience rating premium is given by

$$P_{it,1} = (1 + R_{i,t}^\Delta) \times \bar{P}_{ci,t} \times E_i \times F = (1 + 0.1260) \times 300 \times 1 \times 1 = \$337.80$$

where the *historical* loss rate change  $R_{i,t-1}^\Delta$ , the historical loss ratio  $R_{i,t-1}$  for driver  $i$ , the historical loss ratio for safe group  $R_{t-1}^1$ , the *UBI predicted* loss rate change  $R_{i,t-1}^{\Delta,UB}$ , the UBI predicted loss ratio  $R_{i,t}^y$  for driver  $i$ , and the UBI predicted loss ratio  $R_t^{y,1}$  for the safe group are, respectively,

$$R_{i,t}^\Delta = R_{i,t-1}^{\Delta,H} + 1 \times R_{i,t}^{\Delta,UB} = 0.2403 - 0.1143 = 0.1260, \tag{15}$$

$$R_{i,t-1}^{\Delta,H} = \frac{R_{i,t-1} - R_{t-1}^1}{R_{t-1}^1} = \frac{0.4 - 0.3225}{0.3225} = 0.2403, R_{i,t-1} = \frac{L_{i,t-1}}{P_{i,t-1}} = \frac{0.2}{0.5} = 0.4, R_{t-1}^1 = \frac{L_{t-1}^1}{P_{t-1}^1} = \frac{0.1}{0.31} = 0.3225, \tag{16}$$

$$R_{i,t}^{\Delta,UB} = \frac{R_{i,t}^y - R_t^{y,1}}{R_t^{y,1}} = \frac{0.3 - 0.3387}{0.3387} = -0.1143, R_{i,t}^y = \frac{\hat{y}_{i,t}}{P_{i,t-1}} = \frac{0.15}{0.5} = 0.3, R_t^{y,1} = \frac{\hat{y}_t^1}{P_t^1} = \frac{0.105}{0.31} = 0.3387$$

using (12). So, the premium for driver  $i$  using the UBI experience rating method is \$337.80. This premium is higher than the premium  $\bar{P}_t^1 = \$300$  for the safe group because the loss ratio for driver  $i$  is higher relative to the overall ratio in the safe group using historical claims. However, his current loss ratio due to current safe driving reduces the adverse effect due to the higher historical claims.

Nevertheless, we recognise that not all insured vehicles are equipped with telematic devices, introducing potential data gaps in the telematics insights. In response to this challenge, the *UBI policy* parameter  $\varkappa$  in (13) can be set to 0. This adaptation to the UBI pricing model in (12) also allows for the application to newly insured drivers with only historical records (traditional demographic variables). This premium called *historical experience rating* premium for driver  $i$  during period  $t$  is

$$P_{it,0} = (1 + R_{i,t-1}^{\Delta,H}) \times \bar{P}_t^1 \times E_i \times F = (1 + 0.24031) \times 300 \times 1 \times 1 = \$372.09$$

where the *historical* loss rate change  $R_{i,t-1}^\Delta$  is given by (16). This loss rate change can capture individual differences within a claim group using historical claims but fails to reflect the recent driving risk. Hence, this premium is higher than the UBI experience rating premium calculated using both historical and current driving experience. Thus, the historical experience rating method is unable to provide immediate compensation/reward for safe driving.

Moreover, the UBI premium can track driving behaviour more frequently and closely using regularly updated claim class and annual claim frequency prediction  $\hat{y}_{i,t}$ . The updating period can be reduced to monthly or even weekly to provide more instant feedback using the live telematic data. In summary, the proposed UBI experience rating

premium provides a correction of the loss rate change  $R_{i,t}^{\Delta}$  of the experience rating only premium using the sum of both the *historical* loss rate change  $R_{i,t-1}^{\Delta,H}$  and the *UBI predicted* loss rate change  $R_{i,t}^{\Delta,UB}$ . Here, the proposed PMA model can predict more instantly the annual claim frequencies  $\hat{y}_{i,t}$  using live telematic data. Hence, the UBI premium can be updated more frequently to provide incentives for safe driving. The proposed UBI experience rating premium provides an incremental innovation to the business processes allowing the company to gradually transit to the new regime of UBI by adjusting the UBI policy factor  $\varkappa$  in (13) such that  $\varkappa$  can gradually increase from 0 to 1 if driver  $i$  wants his premium to gradually account for his current driving.

We remark that our analyses made a few assumptions. Firstly, we assumed that the annual premium  $\bar{P}_i^{\$}$  covers the total cost with possibly some profit, and the expectations of loss ratios  $R_{i,t-1}^{\Delta,H}$  and  $R_{i,t}^{\Delta,UB}$  across drivers  $i$  in group  $g$  are around zero. To assess the validity of the assumptions on expectations, one can obtain the distributions of  $R_{i,t-1}^{\Delta,H}$ ,  $R_{i,t}^{\Delta,UB}$  based on the most recent data. If their means  $m_g^{\Delta,H}$ ,  $m_g^{\Delta,UB}$  are not zero, the overall loss ratio  $R_{i,t}^{\Delta}$  in (13) can be adjusted as

$$R_{i,t}^{\Delta} = m_g^{\Delta,H} R_{i,t-1}^{\Delta,H} + m_g^{\Delta,UB} \varkappa R_{i,t}^{\Delta,UB} \quad (17)$$

for group  $g$ . For conservative purposes, the means  $m_g^{\Delta,H}$ ,  $m_g^{\Delta,UB}$  can be replaced by say 75% quantiles  $q_{g,0.75}^{\Delta,H}$ ,  $q_{g,0.75}^{\Delta,UB}$  of the distributions. Secondly, it also implicitly assumes perfect or near-perfect monitoring. However, the advent of monitoring technologies reduces the extent of asymmetric information between insureds and insurers and reduces moral hazard costs.

## 5. Conclusions

In summary, our study, based on claim data from 14,157 drivers exhibiting equidispersion and a substantial 92% of zero claims, introduces a novel approach using two-stage TP, PM, and ZIP regressions. Employing regularisation techniques such as lasso, elastic net, adaptive lasso, and adaptive elastic regularisation, we aimed to predict annual claim frequencies, identify significant DVs, and categorised drivers into low-claim (safe driver) and high-claim (risky driver) groups. To ensure the robustness of our findings, we performed 100 resampling iterations, each comprising 70% of the drivers for all TP, PM, and ZIP models. Our empirical results show that PMA model with adaptive lasso regularisation displayed the best performance in this study. This finding provides relevant guidelines for practitioners and researchers, as the analysis is based on a sound representative telematics sample. Moreover, the PMA model is highly favoured in Table 3c, and its implementation is more straightforward than the TP models.

Furthermore, we proposed to utilise the best-performing PMA model for implementing a UBI experience rating method, aiming to enhance the efficiency of premium pricing strategies. This approach shifts the focus from traditional claim history to recent driving behaviour, offering a nuanced assessment of drivers' risk profiles. Notably, our proposed UBI premium pricing method departs from the annual premium revision characteristic of traditional methods and instead allows for more frequent updates based on recent driving performance to provide instant rewards for safe driving practices and feedback against risky driving using scores of the selected significant DVs for the high-claim group. This dynamic pricing approach not only incentivises responsible and less-frequent driving but also minimises the cross-subsidisation of risky drivers. By enabling a more accurate and timely reflection of driver risk, the UBI contributes to improved loss reserving practices for the auto insurance industry. In essence, our findings support the adoption of UBI experience rating methods as a progressive and effective means of enhancing both driver behaviour and the overall operational efficiency of auto insurance companies.

To implement the PMA models for premium pricing, Section 3.5 provides the modelling details and Appendix B.5 the technical application. If it is challenging, some data analytic companies are experienced to support the handling of telematics data, running of PMA models, predicting drivers' claims, and revising the UBI experience rating premiums.

Updating the frequency for models, claim predictions, and premiums depends on the resources and type of policies. As a suggestion, the PMA models can be updated annually to reflect the change in road conditions, transport policies, etc., and the drivers' predicted annual claims can be updated fortnightly or monthly depending on drivers' mileage. When their predicted annual claims are updated, the premium can also be updated to provide an incentive for good driving. Averaged and individual driving scores for the selected significant DVs (e.g., 29, 36, 59, 67, 73 for the high-claim group of PMA model) can be sent possibly with warnings to inform driving behaviour and encourage skill improvement. These selected significant DVs are associated with more frequent severe brake to slow down at weekday and weekend nights, as well as more frequent severe right turns at the junction at weekday nights and Friday rush time.

In the context of future research within this domain, expanding the classification of driver groups to three or more holds the potential to encompass a wider range of driving styles, ultimately leading to more accurate predictions of claim liability. Introducing an intermediary driver group, distinct from the existing safe and risky classifications, offers an avenue to capture unique driving behaviours and potentially enhances the predictive power of our models. This extension not only enables a closer examination of different driving behaviours but also poses challenges in terms of identifying and interpreting these additional groups. While the application of similar mixture models and regularisation techniques for modelling multiple components remains viable, unravelling the intricacies of distinct groups within the expanded framework introduces interpretative complexities. Determining whether the third group is a composite of the existing two or represents a genuinely distinct category presents additional challenges. Moreover, handling the label switching problem becomes more intricate when dealing with mixture models featuring multiple groups.

A parallel trajectory for future exploration centers around the integration of neural networks as an alternative modelling approach. In contrast to the selection of key driving variables, neural networks employ hidden layers to capture intricate dynamics, incorporating diverse weights and interaction terms. This modelling paradigm allows for the application of network models to trip data without temporal aggregation, as exemplified by [Ma et al. \(2018\)](#), facilitating a more detailed analysis of driving behaviours in conjunction with real-time information on surrounding traffic conditions.

**Author Contributions:** Conceptualisation, J.S.K.C. and U.E.M.; methodology, J.S.K.C.; software, F.U. and A.X.D.D.; validation, F.U., J.S.K.C. and A.X.D.D.; formal analysis, F.U. and Y.W.; investigation, F.U. and Y.W.; resources, U.E.M.; data curation, F.U.; writing—original draft preparation, F.U., J.S.K.C. and Y.W.; writing—review and editing, J.S.K.C.; visualisation, F.U.; supervision, J.S.K.C.; project administration, J.S.K.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors upon request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

A	Adaptive lasso
AIC	Akaike information criterion
BIC	Bayesian information criterion
DVs	Driver behaviour variables
E	Elastic net
GLM	Generalized linear model
GPS	Global positioning system
IG	Information gain
L	Lasso

MSE	Mean squared error
N	Adaptive elastic net
NB	Negative binomial
PAYD	Pay As You Drive
PHYD	Pay How You Drive
PM	Poisson mixture
RMSE	Root mean squared error
ROC	Receiver operating characteristic curve
TP	Two-stage threshold Poisson
UBI	Usage-based auto insurance
ZIP	Zero-inflated Poisson

**Appendix A. Details of Stage 1 TP Model Procedures**

1. Draw subsamples  $\mathcal{S}_r = \{(x_{i\bullet}, n_i, y_i), i \in \mathcal{I}_r\}$ ,  $r = 1, \dots, R$ , with each containing  $N_r = 9910$  drivers, where the index set  $\mathcal{I}_r$  contains all  $i$  being sampled. The  $K$ -fold CV ( $K = 10$ ) further splits  $\mathcal{S}_r$  into 10 nonoverlapping and equal-sized ( $N_k = 991$ ) CV sets

$$\mathcal{S}_{rk} = \{(x_{i\bullet}, n_i, y_i), i \in \mathcal{I}_{rk}\}, \quad k = 1, \dots, K \text{ with index set } \mathcal{I}_{rk}$$

and the training sets are  $\mathcal{S}_{rk}^T = \mathcal{S}_r \setminus \mathcal{S}_{rk}$ , with index set  $\mathcal{I}_{rk}^T = \mathcal{I}_r \setminus \mathcal{I}_{rk}$ . Set  $\lambda = (\lambda_1, \dots, \lambda_M)$  for some  $M$  to be the list of potential  $\lambda$ .

2. Estimate  $\beta_{\lambda_m, rk} = \underset{\beta}{\operatorname{argmin}} \operatorname{LOSS}_{\lambda, \alpha, w}(\beta)$  in (7) for each  $\lambda_m \in \lambda$  and training set  $\mathcal{S}_{rk}^T$  at repeat  $r$  and CV  $k$ . Find optimal  $\lambda_m$  that minimises some regularised CV test statistic such as MSE, MAE, or Deviance (Dev). Taking Dev as an example,

$$\lambda_{r, \min} = \underset{\lambda_m \in \lambda}{\operatorname{argmin}} \operatorname{Dev}_r(\lambda_m) = \underset{\lambda_m \in \lambda}{\operatorname{argmin}} \frac{1}{N_r} \sum_{k=1}^K \sum_{i \in \mathcal{I}_{rk}} -2 \log f(y_{rki}; \mu_{rki}(\beta_{\lambda_m, rk}))$$

where the mean  $\mu_{rki, \lambda_m} = \exp(x_{i\bullet} \beta_{\lambda_m, rk} + \log(n_i))$ . Among MSE, MAE, and Dev statistics, optimal  $\lambda_{r, \min}$  using Dev is selected according to the RMSE of predicted claims for all subsamples. Using  $\lambda_{r, \min}$ ,  $\beta_r = (\beta_{r1}, \dots, \beta_{rj})$  is re-estimated based on the subsample  $\mathcal{S}_r$ . Figure A1a plots Poisson deviance with SE against  $\log(\lambda_m)$ , showing how it drops to  $\lambda_{r, \min}$  for the first subsample ( $r = 1$ ). Figure A1b shows how  $\beta_{rj}$  shrinks to zero as  $\lambda$  increases.

3. Average those nonzero coefficients (selected at least once) over repeats as below:

$$\beta_j = \frac{\sum_{r=1}^R \beta_{rj} \mathbb{I}(\beta_{rj} \neq 0)}{\sum_{r=1}^R \mathbb{I}(\beta_{rj} \neq 0)}, \quad j \in \mathcal{I}^\beta \tag{A1}$$

where  $\mathbb{I}(A)$  is the indicator function of event  $A$ , and the index set

$$\mathcal{I}^\beta = \{j : \exists r = 1, \dots, R, \beta_{rj} \neq 0\} \tag{A2}$$

contains those DVs selected at least once over  $R$  subsamples in stage 1. The averaged coefficients  $\beta_{j \in \mathcal{I}^\beta}$  (based on Dev) are reported in Table A2 for the TPL-1 and TPA-1 models using the optimal  $\lambda_{\min}$ . For example, DV 10 is not even selected once for the TPL-1 model.

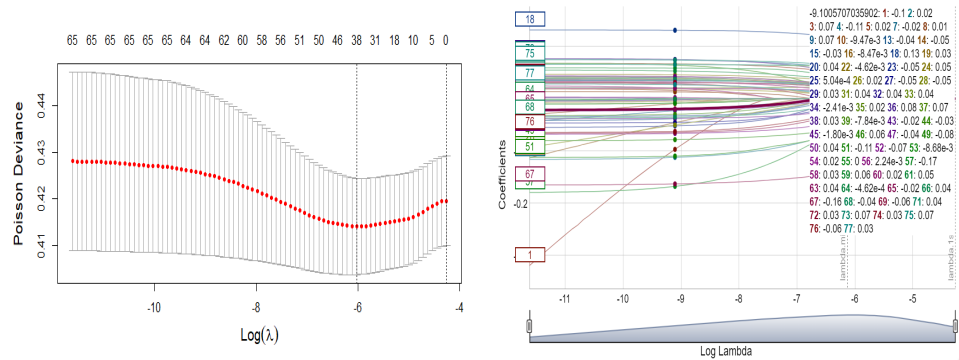
4. Further select DVs that are frequently (not rarely) selected according to a weighted selection frequency measure given by

$$I_j = \sum_{r=1}^R \frac{1}{\operatorname{RMSE}_r} \mathbb{I}(\beta_{rj} \neq 0), \quad j \in \mathcal{I}^\beta \tag{A3}$$

weighting inversely to  $\operatorname{RMSE}_r$ . Superscripts T1, T2, M, and Z are added to  $I_j$  when applied to the stage 1 TP, stage 2 TP, PM, and ZIP models, respectively. This weighted selection counts  $I = (I_{j \in \mathcal{I}^\beta})$  using MSE, MAE, and deviance, which are also reported



in Table A2 for the TPL-1 and TPA-1 models. Table 3a shows that the TPL-1 and TPA-1 models have been selected according to the model performance measures AIC, BIC, and MSE. The results of the TPL-1 model in Table A2 show that 12 DVs in deep grey highlight with  $I_j < 0.2 \max(I_j) = 62$  have been dropped, as they are rarely selected, resulting in  $J^{T1} = 65 - 1 - 12 = 52$  DVs. These  $J^{T1}$  DVs can be interpreted as frequently selected DVs or simply selected DVs.



(a) Poisson deviance for CV

(b) Coefficients path

**Figure A1.** (a) The Poisson deviance CV criteria across  $\log \lambda_m$  to find  $\lambda_{\min}$ . (b) Coefficient  $\beta_j$  across  $\log \lambda$  for stage 1 TP model using lasso regularisation based on the first subsample ( $r = 1$ ).

**Appendix B. Some Technical Details of Model Implementation**

- This study utilises R commands `glm` to fit Poisson regression and `glmnet` to fit Poisson regression with lasso regularisation (Zeileis et al. 2008). The latter command begins with adopting the R function `sparse.model.matrix` as `data_feature <- sparse.model.matrix(~., dt_feature)`. We use the argument `penalty.factor` in `cv.glmnet` for adaptive lasso. We remark that the `glmnet` package does not provide a  $p$  value. We extract the  $p$  value for the selected DVs by refitting the model using `glm` procedure.
- We use the 100 simulated dataset in stages 1 and 2 of the TP and PM models to explore optimal  $\alpha$  values in the elastic net. We first set up our 10-fold CV strategy. Using `caret` package in R, we use `train()` with `method = "glmnet"` to fit the elastic net.

```

XX = model.matrix(Claims ~ . -EXP-1,data=stage1)
YY = stage1$Claims
OFF = log(stage1$EXP)
Fit_stage1 <- caret::train(
  x = cbind(XX,OFF),
  y = YY,
  method = "glmnet",
  family = "poisson",
  tuneLength = 10,
  trControl = trainControl(method="cv", number = 10, repeats = 100)
)

```
- We use `roc()` in the `pROC` package to calculate the AUC. The `latex2exp` package also provides an ROC plot.
- We implement the `AER` package in R using the built-in command `dispersiontest()` that assesses the alternative hypothesis  $H_1 : \text{Var}(Y_i) = \mu_i + \Psi \times \text{trafo}(\mu_i)$ , where the transformation function `trafo`( $\mu_i$ ) =  $\mu_i$  (by default, `trafo` = `NULL`) corresponds to the Poisson model with  $\text{Var}(Y_i) = (1 + \Psi)\mu_i$ . If the dispersion  $1 + \Psi$  is greater than 1, it indicates overdispersion.
- The PM regression model is estimated using `FLXMRglmnet(formula = ~., family = c("gaussian","binomial","poisson"), adaptive = TRUE, select = TRUE, offset = NULL, ...)` in the R package `flexmix` (Leisch 2004) to fit mixtures of GLMs with lasso regularisation. Setting `adaptive = TRUE` for the adaptive lasso triggers a two-step process. Initially, an unpenalised model is fitted to obtain the preliminary coefficient estimates

- $\hat{\beta}_j$  for the penalty weights  $w_j = 1/|\hat{\beta}_j|$ . Then,  $w_j$  values are applied to each coefficient in the subsequent model fitting. With the selected DVs for the low- and high-claim groups, `FLXMRglmfix()` refits the model, provides the significance of the coefficients, predicts claims, supports CV values and evaluates various goodness-of-fit measures.
- The ZIP regression model is estimated using the `zipath()` function for lasso and elastic net regularisation and the `ALasso()` function for adaptive lasso regularisation from the `mpath` and `AMAZonn` packages. The optimal lambda minimum is searched via 10-fold cross-validation with `cv.zipath()` and applied to both fitted models, `ZIPL` and `ZIPA`, for  $R = 100$  subsamples, each with 70% data. Full data are refitted to the PM model based on the selected DVs using `Poisson zeroinf`.

### Appendix C. Driving Variable Description

#### Event type

- ACC** Acceleration Event—Accelerating/From full stop
  - C1 Smooth acceleration (acceleration to 30 MPH in more than 12 s)
  - C2 Moderate acceleration (acceleration to 30 MPH in 5–11 s)
- BRK** Braking Event—Full Stop/Slow down
  - C1 Smooth, even slowing down (up to about 7 mph/s)
  - C2 Mild to sharp brakes with adequate visibility and road grip (7–10 mph/s)
- LFT** Left turning Event—None (Interchange, curved road, overtaking)/At Junction
  - C1 Smooth, even cornering within the posted speed and according to the road and visibility conditions
  - C2 Moderate cornering slightly above the posted speed (cornering with light disturbance to passengers)
- RHT** Right turning Event—None (Interchange, curved road, overtaking)/At Junction
  - C1 and C2 are the same as LFT

#### Time type

- T1 Weekday late evening, night, midnight, early morning
- T2 Weekday morning rusk, noon, afternoon rush
- T3 Weekday morning, afternoon, no rush
- T4 Friday rush
- T5 Weekend night
- T6 Weekend day

**Table A1.** Driving variable labels.

DV1	ACC_ACCELERATING_T3_C1	DV19	BRK_FULLSTOP_T1_C1	DV39	LFT_NONE_T1_C1	DV57	RHT_NONE_T1_C1
DV2	ACC_ACCELERATING_T3_C2	DV20	BRK_FULLSTOP_T1_C2	DV43	LFT_NONE_T6_C1	DV58	RHT_NONE_T1_C2
DV3	ACC_ACCELERATING_T4_C1	DV22	BRK_FULLSTOP_T2_C2	DV44	LFT_NONE_T6_C2	DV59	RHT_NONE_T4_C1
DV4	ACC_ACCELERATING_T4_C2	DV23	BRK_FULLSTOP_T3_C1	DV45	LFT_ATJUNCTION_T1_C1	DV60	RHT_NONE_T4_C2
DV5	ACC_ACCELERATING_T5_C1	DV24	BRK_FULLSTOP_T3_C2	DV46	LFT_ATJUNCTION_T1_C2	DV61	RHT_NONE_T5_C1
DV7	ACC_ACCELERATING_T5_C2	DV25	BRK_FULLSTOP_T4_C1	DV47	LFT_ATJUNCTION_T2_C1	DV63	RHT_NONE_T5_C2
DV8	ACC_FROMFULLSTOP_T1_C1	DV26	BRK_FULLSTOP_T4_C2	DV49	LFT_ATJUNCTION_T3_C1	DV64	RHT_NONE_T6_C1
DV9	ACC_FROMFULLSTOP_T1_C2	DV27	BRK_FULLSTOP_T6_C1	DV50	LFT_ATJUNCTION_T3_C2	DV65	RHT_NONE_T6_C2
DV10	ACC_FROMFULLSTOP_T2_C1	DV28	BRK_FULLSTOP_T6_C2	DV51	LFT_ATJUNCTION_T4_C1	DV66	RHT_ATJUNCTION_T1_C1
DV13	ACC_FROMFULLSTOP_T3_C2	DV29	BRK_SLOWDOWN_T1_C1	DV52	LFT_ATJUNCTION_T4_C2	DV67	RHT_ATJUNCTION_T1_C2
DV14	ACC_FROMFULLSTOP_T4_C1	DV31	BRK_SLOWDOWN_T2_C1	DV53	LFT_ATJUNCTION_T5_C1	DV68	RHT_ATJUNCTION_T2_C1
DV15	ACC_FROMFULLSTOP_T4_C2	DV32	BRK_SLOWDOWN_T2_C2	DV54	LFT_ATJUNCTION_T5_C2	DV69	RHT_ATJUNCTION_T2_C2
DV16	ACC_FROMFULLSTOP_T5_C1	DV33	BRK_SLOWDOWN_T4_C1	DV55	LFT_ATJUNCTION_T6_C1	DV71	RHT_ATJUNCTION_T3_C2
DV18	ACC_FROMFULLSTOP_T5_C2	DV34	BRK_SLOWDOWN_T4_C2	DV56	LFT_ATJUNCTION_T6_C2	DV72	RHT_ATJUNCTION_T4_C1
		DV35	BRK_SLOWDOWN_T5_C1			DV73	RHT_ATJUNCTION_T4_C2
		DV36	BRK_SLOWDOWN_T5_C2			DV74	RHT_ATJUNCTION_T5_C1
		DV37	BRK_SLOWDOWN_T6_C1			DV75	RHT_ATJUNCTION_T5_C2
		DV38	BRK_SLOWDOWN_T6_C2			DV76	RHT_ATJUNCTION_T6_C1
						DV77	RHT_ATJUNCTION_T6_C2

Appendix D. Visualisation of Driver Variables  
 Appendix D.1. Driving Variables by Claim Frequency

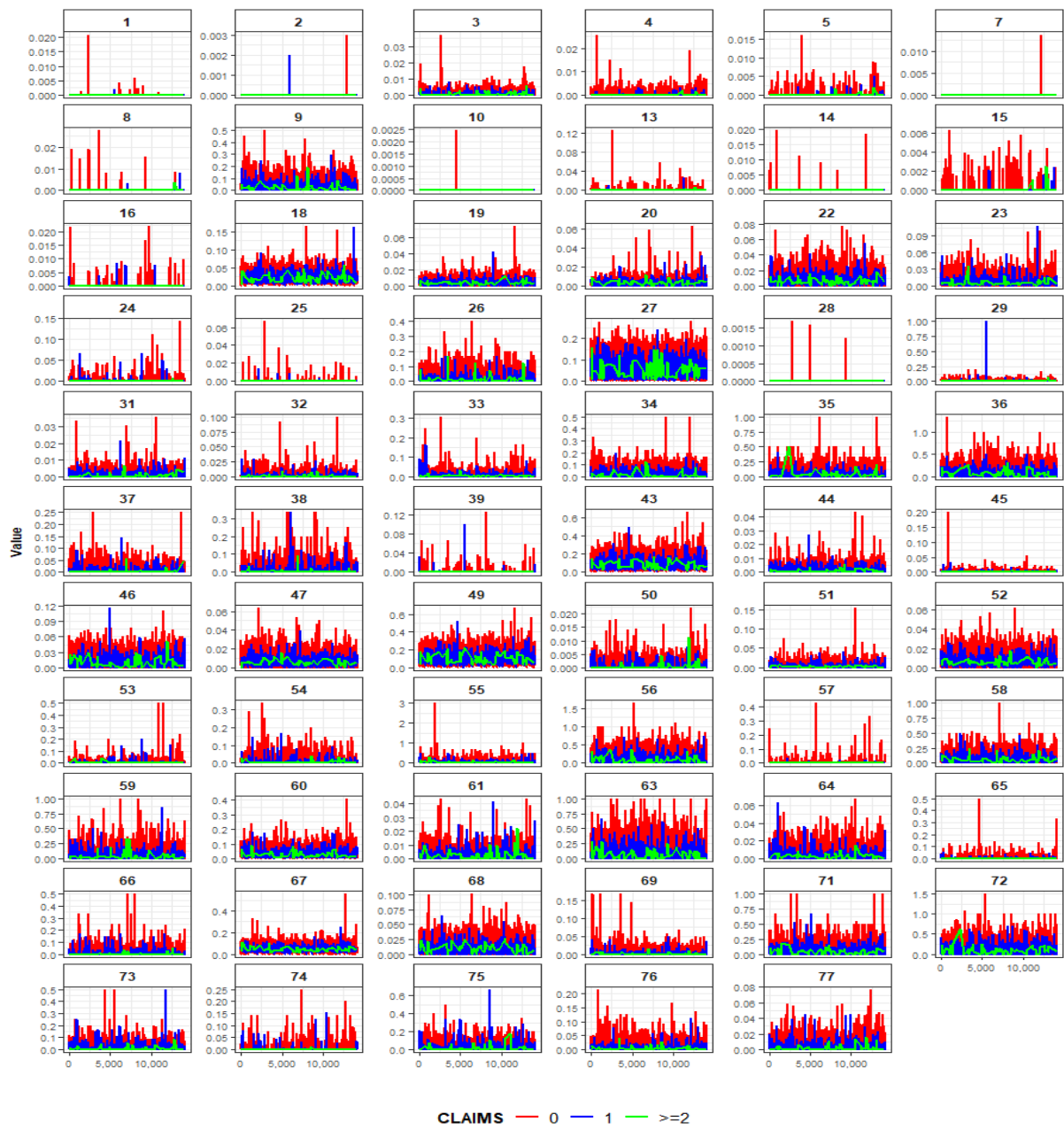
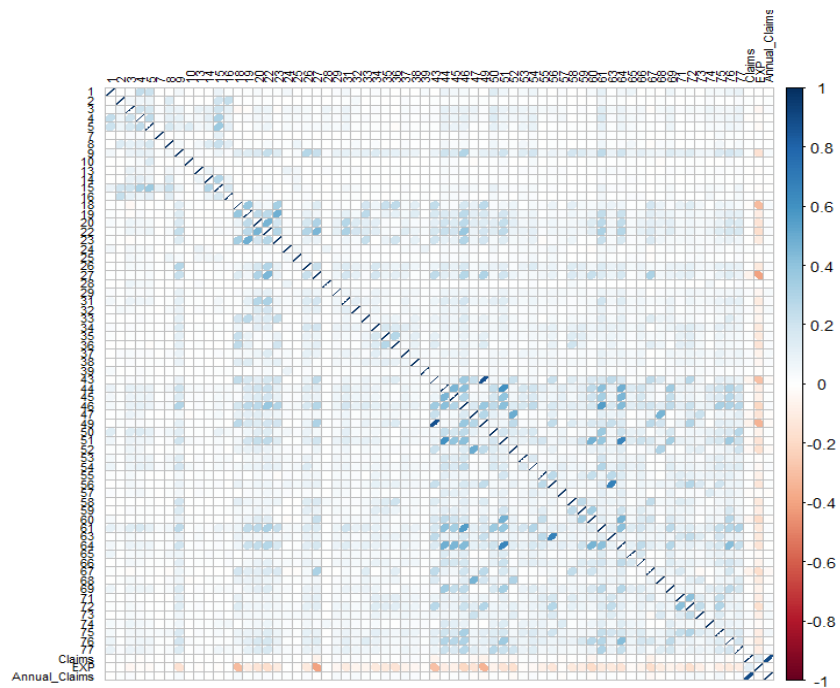
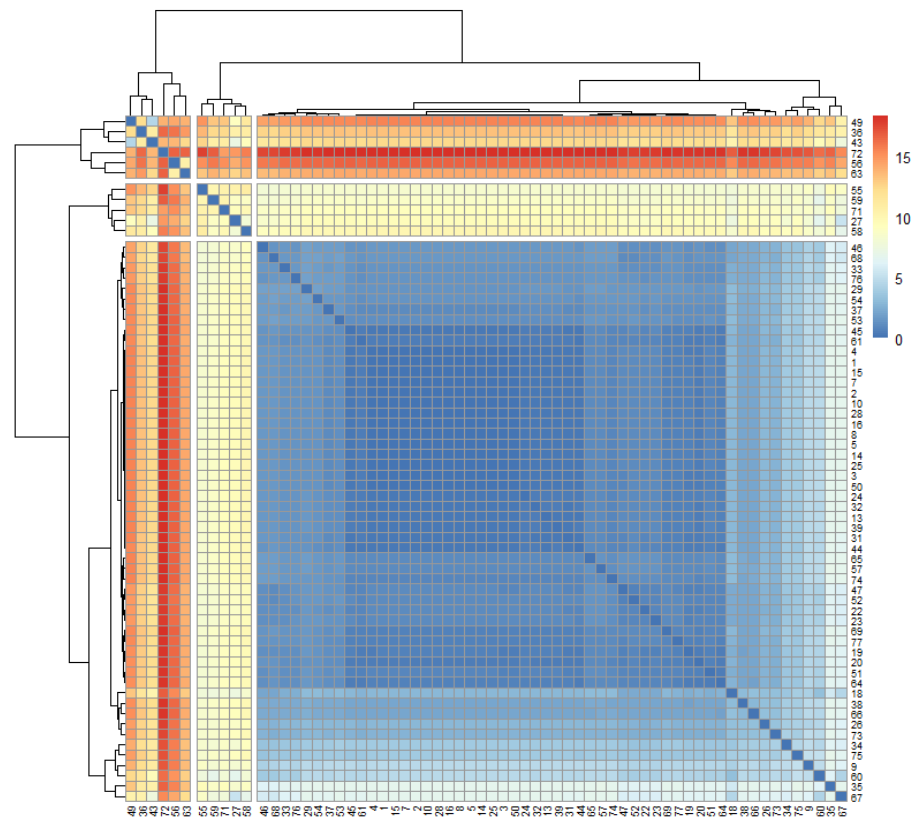


Figure A2. Value against driver ID with colours showing claim frequency  $y_i = 0, 1, \geq 2$  for 65 DVs.

Appendix D.2. Correlation Matrix and Hierarchical Clustering of Driving Variables



(a) Correlation matrix between 65 DVs, claims, exposure, and annual claims.



(b) Distance score and hierarchical clustering between 65 DVs.

Figure A3. Relationship between variables using correlation matrix and hierarchical clustering.

Appendix E. Parameter Estimates of all Models

**Table A2.** Parameter estimates  $\beta$  in (A1) for stage 1 TP models before refit with  $R = 100$  subsamples of 70% data using Poisson glmnet,  $\beta^{T1}$  after refitting to full data using Poisson glm on the selected DVs with  $I_j^{T2} > 62$  (otherwise dropped as indicated in grey highlight), and selection criteria  $I^{T1}$  in (A3). There are  $J^{T1} = 52$  DVs selected for TPL-1 and  $J^{T1} = 39$  DVs selected for TPA-1 under columns  $\beta_j^{T1}$ . The bold with yellow highlighted under  $\beta_j^{T1}$  are significant.

TPL-1																	
Measures	glmnet with 100 Repeats				glm	Measures	glmnet with 100 Repeats				glm	Measures	glmnet with 100 Repeats				glm
	MSE	MAE	Deviance	Poisson	$\beta_j^{T1}$		MSE	MAE	Deviance	Poisson	$\beta_j^{T1}$		MSE	MAE	Deviance	Poisson	$\beta_j^{T1}$
DVs	$I_j^{T1}$			$\beta_j$	$\beta_j^{T1}$	DVs	$I_j^{T1}$			$\beta_j$	$\beta_j^{T1}$	DVs	$I_j^{T1}$			$\beta_j$	$\beta_j^{T1}$
1	-	34	7	-0.0029	-	28	3	126	61	-0.0031	-	55	-	119	68	0.0082	0.0134
2	89	232	228	0.0176	0.0159	29	227	276	279	0.0279	<b>0.0360</b>	56	37	136	123	0.0149	0.0061
3	180	317	337	0.0402	<b>0.0619</b>	31	251	324	337	0.0409	0.0535	57	139	310	334	-0.0446	-0.1696
4	140	307	334	-0.0409	<b>-0.0987</b>	32	302	327	341	0.0474	<b>0.0513</b>	58	24	147	109	0.0115	0.0095
5	3	109	61	0.0085	-	33	133	273	266	0.0256	0.0393	59	68	252	229	0.0282	0.0448
7	-	136	116	-0.0021	-0.0830	34	-	85	20	-0.0073	-	60	95	198	191	-0.0243	-0.0091
8	7	95	37	-0.0011	-	35	146	245	242	0.0222	0.0168	61	255	317	320	0.0426	0.0626
9	272	310	320	0.0417	0.0546	36	262	320	334	0.0576	<b>0.0797</b>	63	98	242	235	0.0219	0.0346
10	-	17	-	-	-	37	292	327	334	0.0424	<b>0.0518</b>	64	30	194	160	-0.0264	-0.0348
13	10	140	72	-0.0154	-0.0253	38	184	290	289	0.0238	0.0263	65	-	99	41	-0.0084	-
14	-	105	14	-0.0031	-	39	71	232	228	0.0122	0.0160	66	41	164	133	0.0164	0.0173
15	14	119	68	-0.0190	-0.0065	43	78	204	204	-0.0381	-0.0505	67	329	330	341	-0.1400	<b>-0.1706</b>
16	3	113	65	0.0001	-0.0004	44	3	112	41	-0.0113	-	68	17	102	61	-0.0135	-
18	316	327	341	0.0969	<b>0.1254</b>	45	3	78	48	0.0172	-	69	17	188	140	-0.0166	-0.0320
19	-	85	20	0.0021	-	46	31	194	164	0.0210	0.0361	71	78	231	224	0.0172	0.0185
20	173	310	323	0.0363	0.0563	47	177	314	330	-0.0611	<b>-0.0918</b>	72	187	303	297	0.0319	0.0418
22	41	205	177	0.0133	0.0309	49	95	245	252	-0.0341	-0.0448	73	302	327	341	0.0587	<b>0.0743</b>
23	-	133	75	-0.0051	-0.0235	50	72	228	218	0.0157	0.0205	74	102	242	242	0.0216	0.0350
24	136	272	262	0.0213	0.0236	51	116	289	306	-0.0517	-0.0944	75	336	330	341	0.0621	<b>0.0659</b>
25	-	119	58	-0.0087	-	52	150	307	324	-0.0397	-0.0623	76	48	239	235	-0.0236	-0.0565
26	58	160	139	0.0129	0.0024	53	65	174	157	0.0156	0.0107	77	157	307	324	0.0324	0.0549
27	17	160	129	-0.0205	-0.0402	54	163	262	272	0.0209	0.0208						
TPA-1																	
Measures	glmnet with 100 Repeats				glm	Measures	glmnet with 100 Repeats				glm	Measures	glmnet with 100 Repeats				glm
	MSE	MAE	Deviance	Poisson	$\beta_j^{T1}$		MSE	MAE	Deviance	Poisson	$\beta_j^{T1}$		MSE	MAE	Deviance	Poisson	$\beta_j^{T1}$
DVs	$I_j^{T1}$			$\beta_j$	$\beta_j^{T1}$	DVs	$I_j^{T1}$			$\beta_j$	$\beta_j^{T1}$	DVs	$I_j^{T1}$			$\beta_j$	$\beta_j^{T1}$
1	3	41	24	-0.0712	-	28	-	79	-	-	-	55	-	41	20	0.0030	-
2	61	95	68	0.0360	0.0149	29	228	279	276	0.0311	<b>0.0357</b>	56	14	99	61	0.0311	-
3	160	317	310	0.0512	<b>0.0608</b>	31	217	316	313	0.0514	<b>0.0536</b>	57	78	306	327	-0.0797	-0.1726
4	89	296	310	-0.0630	<b>-0.1035</b>	32	319	337	340	0.0552	<b>0.0503</b>	58	-	38	3	0.0297	-
5	-	61	10	0.0482	-	33	78	248	214	0.0352	0.0363	59	31	204	139	0.0374	0.0478
7	-	24	-	-	-	34	-	38	17	-0.0201	-	60	58	143	126	-0.0288	-0.0093
8	-	34	13	-0.0067	-	35	102	190	177	0.0258	0.0199	61	163	283	282	0.0595	<b>0.0702</b>
9	187	300	289	0.0517	0.0579	36	248	334	330	0.0703	<b>0.0775</b>	63	41	194	150	0.0358	0.0422
10	-	-	-	-	-	37	285	334	330	0.0513	<b>0.0530</b>	64	27	143	89	-0.0551	-0.0317
13	-	48	20	-0.0144	-	38	160	231	218	0.0298	0.0271	65	-	17	10	-0.0100	-
14	-	62	-	-	-	39	7	116	71	0.0132	0.0163	66	17	109	55	0.0221	-
15	3	86	31	-0.0200	-	43	48	235	204	-0.0577	-0.0510	67	336	340	340	-0.1752	<b>-0.1686</b>
16	-	14	3	-0.0088	-	44	-	44	7	-0.0294	-	68	-	85	55	-0.0201	-
18	333	340	340	0.1212	<b>0.1205</b>	45	3	72	44	0.0293	-	69	7	99	34	-0.0334	-
19	10	65	17	0.0146	-	46	10	130	51	0.0410	-	71	37	129	102	0.0291	0.0204
20	112	286	272	0.0426	0.0567	47	170	327	327	-0.0773	<b>-0.0913</b>	72	156	269	248	0.0479	0.0443
22	20	143	85	0.0300	0.0282	49	58	194	187	-0.0470	-0.0367	73	289	340	337	0.0710	<b>0.0733</b>
23	-	55	14	-0.0171	-	50	27	129	92	0.0259	0.0206	74	51	228	167	0.0301	0.0341
24	58	188	147	0.0237	0.0230	51	51	282	262	-0.0718	-0.0918	75	316	340	333	0.0748	<b>0.0709</b>
25	-	34	3	-0.0843	-	52	136	306	303	-0.0493	-0.0618	76	41	225	184	-0.0446	-0.0565
26	21	68	38	0.0201	-	53	10	71	54	0.0182	-	77	116	290	273	0.0457	0.0554
27	10	153	105	-0.0349	-0.0359	54	109	176	183	0.0259	0.0256						

Table A3. Parameter estimates beta\_Lhj^T2, beta\_Hhj^T2 for the stage 2 TP models with R = 100 subsamples of 70% data. Parameters are based on J^T1 = 52 DVs in stage 1, and J^T2 refers to the number of frequently selected DVs with I\_Lhj > 43 (tau\_0.08), 49 (tau\_0.09), 53 (tau\_0.10), 56 (tau\_0.11), and I\_Hhj > 19, 13, 9, and 6, respectively, which differ across threshold. Significant beta\_Lhj^T2, beta\_Hhj^T2 are in boldface with yellow highlighted.

Table with 17 columns: Groups, DVs, and parameter estimates across four thresholds (tau\_0.08: TPLA-2, tau\_0.09: TPLA-2, tau\_0.10: TPLN-2, tau\_0.11: TPLN-2). The table lists DVs from 2 to 77 with their respective beta values and I values. Significant values are highlighted in bold and yellow.

**Table A4.** Parameter estimates  $\beta_0, \beta_c$  for ZIP model before refit,  $\beta_L^M, \beta_H^M$  for PM, and  $\beta_0^Z, \beta_c^Z$  for ZIP models after refitted to all data based on selected DVs and selection criteria  $I_L^M, I_H^M, I_0^Z, I_c^Z$ , with  $R = 100$  subsamples of 70% data. For PM model,  $J_L^M, J_H^M$  refer to the number of frequently selected DVs with  $I_{L_j}^M > 43$  (PML), 45 (PMA), and  $I_{H_j}^M > 19$  (PML), 17 (PMA); otherwise, they are dropped, as in grey highlight. For ZIP model,  $J_0^Z, J_c^Z$  refer to the number of frequently selected DVs with  $I_{0_j}^Z > 62$  and  $I_{c_j}^Z > 62$ ; otherwise,  $\beta_{0j}$  and  $\beta_{cj}$  are excluded, as in grey highlight. Significant parameters  $\beta_{L_j}^M, \beta_{H_j}^M, \beta_{0_j}^Z, \beta_{c_j}^Z$  are boldfaced and yellow highlighted.

DV <sub>s</sub>	PML				PMA				ZIPA							
	43		19		45		17		62			62				
	45		18		39		40		4			45				
DV <sub>s</sub>	Low		High		DV <sub>s</sub>	Low		High		DV <sub>s</sub>	Zero			Count		
	$I_{L_j}^M$	$\beta_{L_j}^M$	$I_{H_j}^M$	$\beta_{H_j}^M$		$I_{L_j}^M$	$\beta_{L_j}^M$	$I_{H_j}^M$	$\beta_{H_j}^M$		$I_{0_j}^Z$	$\beta_{0_j}^Z$	$\beta_{0_j}^Z$	$I_{c_j}^Z$	$\beta_{c_j}^Z$	$\beta_{c_j}^Z$
3	118	0.0182	47	0.0983	3	71	0.0380	126	0.1182	3	44	-0.0022	-	291	0.0404	<b>0.0514</b>
9	88	-0.0003	20	0.0275	9	27	-0.0226	16	0.0106	9	20	-0.0047	-	172	0.0170	0.0450
18	324	<b>0.0636</b>	27	<b>0.0477</b>	18	206	<b>0.0877</b>	149	0.1078	18	-	-	-	88	0.0044	<b>0.1352</b>
19	311	0.0491	3	0.1058	19	200	0.0877	82	0.0821	19	10	-0.0003	-	136	0.0050	-0.0263
20	78	-0.0197	37	0.0919	20	71	<b>-0.0443</b>	27	-0.0532	20	34	-0.0053	-	291	0.0552	<b>0.0717</b>
22	162	0.0098	54	0.0633	22	91	-0.0484	101	0.0839	22	47	-0.0104	-	217	0.0333	0.0441
23	335	-0.2076	24	-0.2634	23	219	-0.2801	159	-0.2732	23	-	-	-	84	-0.0025	-0.0247
26	250	0.0259	98	0.0370	26	212	0.0375	81	0.0119	26	3	$1.91 \times 10^{-5}$	-	125	0.0069	0.0014
27	338	0.0450	3	0.0068	27	251	0.0755	60	0.0576	27	10	0.0004	-	339	-0.1900	-0.0495
29	324	<b>0.0355</b>	17	0.0633	29	172	0.0663	79	<b>0.0722</b>	29	24	-0.0005	-	267	0.0182	<b>0.0363</b>
31	294	0.0352	14	0.0390	31	165	0.0637	87	0.0496	31	20	-0.0024	-	234	0.0265	0.0515
33	138	-0.0150	-	-	33	50	-0.0516	13	-0.0265	33	55	-0.0123	-	213	0.0243	0.0426
34	287	0.0369	7	0.1254	34	127	0.0586	57	0.0533	34	3	-0.0001	-	88	-0.0022	-0.0234
35	331	0.0578	7	0.0130	35	299	0.1089	43	0.0735	35	20	-0.0029	-	132	0.0099	0.0232
36	335	0.0501	31	0.0450	36	214	0.0690	120	<b>0.0642</b>	36	30	-0.0103	-	281	0.0464	<b>0.0752</b>
37	304	<b>0.0284</b>	13	0.0522	37	183	0.0416	51	0.0567	37	10	-0.0009	-	298	0.0429	<b>0.0524</b>
38	230	-0.0516	7	0.0002	38	121	-0.1553	40	0.0017	38	98	-0.0265	-39.0618	121	0.0076	-0.0065
43	88	-0.0267	3	-0.0238	43	36	-0.0296	44	-0.0703	43	17	0.0008	-	173	-0.0397	-0.0471
44	57	0.0078	-	-	44	54	0.0575	37	0.0932	44	-	-	-	155	-0.0099	-0.0287
45	274	0.0541	24	0.0395	45	249	<b>0.1177</b>	79	0.0982	45	10	-0.0005	-	153	0.0090	0.0188
46	338	-0.0957	14	-0.0442	46	259	-0.1665	78	-0.1081	46	30	-0.0042	-	264	0.0569	0.0362
47	338	-0.1644	20	-0.0469	47	292	-0.2914	77	-0.1493	47	3	0.0000	-	322	-0.0843	<b>-0.0940</b>
49	249	0.0241	10	0.0235	49	91	0.0430	23	0.0378	49	165	0.0366	-0.0193	251	-0.1070	-0.0597
50	331	-0.0898	24	-0.0447	50	197	-0.1691	126	-0.1419	50	-	-	-	122	0.0077	0.0145
51	335	-0.0981	14	-0.0381	51	225	-0.1602	119	-0.1378	51	17	0.0004	-	301	-0.0865	-0.0868
52	314	0.0265	37	0.0265	52	93	0.0447	81	0.0538	52	10	0.0023	-	311	-0.0738	-0.0660
54	84	0.0134	-	-	54	40	0.0111	7	0.0496	54	17	-0.0013	-	213	0.0197	0.0182
55	112	0.0183	7	0.0196	55	33	0.0027	14	-0.0053	55	7	0.0001	-	88	0.0016	0.0109
56	142	0.0193	44	0.0636	56	33	0.0478	114	0.0780	56	3	-0.0003	-	75	0.0012	0.0002
58	240	0.0298	17	0.0632	58	154	<b>0.0718</b>	62	0.0692	58	3	-0.0001	-	128	0.0087	0.0157
59	294	-0.0433	17	-0.0817	59	115	-0.1431	73	<b>-0.1543</b>	59	24	-0.0035	-	200	0.0189	0.0409
60	318	0.0574	24	0.0934	60	183	0.1142	162	0.0945	60	13	0.0024	-	231	-0.0259	-0.0053
61	223	0.0297	3	0.0193	61	167	<b>0.0773</b>	37	0.0514	61	20	-0.0032	-	244	0.0403	0.0560
63	189	-0.0652	20	-0.0031	63	162	-0.1524	125	-0.0583	63	37	-0.0057	-	98	0.0088	0.0362
64	162	0.0132	7	0.0346	64	71	0.0231	7	0.0641	64	-	-	-	139	-0.0220	-0.0409
66	338	-0.1689	7	-0.0538	66	297	-0.3054	64	-0.1943	66	7	-0.0001	-	81	0.0050	0.0175
67	88	<b>-0.0263</b>	-	-	67	37	-0.0730	40	<b>0.0416</b>	67	20	0.0044	-	339	-0.1312	<b>-0.1510</b>
68	210	-0.0265	7	-0.0102	68	90	-0.0940	52	-0.1397	68	7	0.0001	-	67	-0.0013	-0.0041
69	199	0.0210	17	0.0611	69	95	0.0537	26	0.0420	69	7	0.0005	-	180	-0.0190	-0.0286
71	270	0.0388	14	0.0725	71	141	0.0869	76	0.0686	71	34	-0.0040	-	145	0.0098	0.0114
72	318	0.0473	183	0.1172	72	217	0.0774	194	0.0950	72	30	-0.0040	-	268	0.0338	0.0391
73	335	0.0631	88	<b>0.1193</b>	73	242	0.1072	109	<b>0.1064</b>	73	75	-0.0129	-0.1323	288	0.0391	<b>0.0617</b>
75	321	-0.0545	27	<b>-0.0622</b>	75	142	<b>-0.1177</b>	136	-0.1645	75	20	-0.0022	-	301	0.0511	<b>0.0634</b>
76	257	0.0361	10	0.0327	76	161	0.0722	83	0.0827	76	7	0.0009	-	244	-0.0380	-0.0656
77	284	0.0324	17	0.0370	77	170	0.0740	74	0.0615	77	98	-0.0189	-34.9726	149	0.0132	-0.0313

References

Ayuso, Mercedes, Montserrat Guillen, and Jens Perch Nielsen. 2019. Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation* 46: 735–52. [CrossRef]

Banerjee, Prithish, Broti Garai, Himel Mallick, Shrabanti Chowdhury, and Saptarshi Chatterjee. 2018. A note on the adaptive lasso for zero-inflated Poisson regression. *Journal of Probability and Statistics* 2018: 2834183. [CrossRef]

Barry, Laurence, and Arthur Charpentier. 2020. Personalization as a promise: Can big data change the practice of insurance? *Big Data & Society* 7: 2053951720935143.

Bhattacharya, Sakyajit, and Paul D. McNicholas. 2014. An adaptive lasso-penalized BIC. *arXiv arXiv:1406.1332*.

Bolderdijk, Jan Willem, Jasper Knockaert, E. M. Steg, and Erik T. Verhoef. 2011. Effects of Pay-As-You-Drive vehicle insurance on young drivers' speed choice: Results of a dutch field experiment. *Accident Analysis & Prevention* 43: 1181–86.

- Cameron, A. Colin and Pravin K. Trivedi. 1990. Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics* 46: 347–64. [CrossRef]
- Chan, Jennifer S. K., S. T. Boris Choy, Udi Makov, Ariel Shamir, and Vered Shapovalov. 2022. Variable selection algorithm for a mixture of poisson regression for handling overdispersion in claims frequency modeling using telematics car driving data. *Risks* 10: 83. [CrossRef]
- Chassagnon, Arnold and Pierre-André Chiappori. 1997. Insurance under moral hazard and adverse selection: The case of pure competition. Delta-CREST Document Available online: <https://econpapers.repec.org/paper/fthlavale/28.htm> (accessed on 1 August 2024).
- Czado, Claudia, Tilmann Gneiting, and Leonhard Held. 2009. Predictive model assessment for count data. *Biometrics* 65: 1254–61. [CrossRef]
- Dean, Curtis Gary. 1997. An introduction to credibility. In *Casualty Actuary Forum: Arlington: Casualty Actuarial Society*. pp. 55–66. Available online: [https://www.casact.org/sites/default/files/database/forum\\_97wforum\\_97wf055.pdf](https://www.casact.org/sites/default/files/database/forum_97wforum_97wf055.pdf) (accessed on 1 August 2024).
- Deng, Min, Mostafa S. Aminzadeh, and Banghee So. 2024. Inference for the parameters of a zero-inflated poisson predictive model. *Risks* 12: 104. [CrossRef]
- Duval, Francis, Jean-Philippe Boucher, and Mathieu Pigeon. 2023. Enhancing claim classification with feature extraction from anomaly-detection-derived routine and peculiarity profiles. *Journal of Risk and Insurance* 90: 421–58. [CrossRef]
- Eling, Martin, and Mirko Kraft. 2020. The impact of telematics on the insurability of risks. *The Journal of Risk Finance* 21: 77–109. [CrossRef]
- Ellison, Adrian B., Michiel C. J. Bliemer, and Stephen P. Greaves. 2015. Evaluating changes in driver behaviour: A risk profiling approach. *Accident Analysis & Prevention* 75: 298–309.
- Fan, Jianqing, and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348–60. [CrossRef]
- Fawcett, Tom. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861–74.
- Gao, Guangyuan, and Mario V. Wüthrich. 2018. Feature extraction from telematics car driving heatmaps. *European Actuarial Journal* 8: 383–406. [CrossRef]
- Gao, Guangyuan, Mario V. Wüthrich, and Hanfang Yang. 2019. Evaluation of driving risk at different speeds. *Insurance: Mathematics and Economics* 88: 108–19. [CrossRef]
- Gao, Guangyuan, Shengwang Meng, and Mario V. Wüthrich. 2019. Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal* 2019: 143–62. [CrossRef]
- Guillen, Montserrat, Jens Perch Nielsen, Ana M. Pérez-Marín, and Valandis Elpidorou. 2020. Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal* 24: 141–52. [CrossRef]
- Guillen, Montserrat, Jens Perch Nielsen, and Ana M. Pérez-Marín. 2021. Near-miss telematics in motor insurance. *Journal of Risk and Insurance* 88: 569–89. [CrossRef]
- Guillen, Montserrat, Jens Perch Nielsen, Mercedes Ayuso, and Ana M. Pérez-Marín. 2019. The use of telematics devices to improve automobile insurance rates. *Risk Analysis* 39: 662–72. [CrossRef]
- Huang, Yifan, and Shengwang Meng. 2019. Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems* 127: 113156. [CrossRef]
- Hurley, Rich, Peter Evans, and Arun Menon. 2015. *Insurance Disrupted: General Insurance in a Connected World*. London: The Creative Studio, Deloitte.
- Jeong, Himchan. 2022. Dimension reduction techniques for summarized telematics data. *The Journal of Risk Management* 33: 1–24. [CrossRef]
- Jeong, Himchan, and Emiliano A. Valdez. 2018. Ratemaking Application of Bayesian LASSO with Conjugate Hyperprior. Available online: <https://ssrn.com/abstract=3251623> (accessed on 1 December 2018).
- Kantor, S., and Tomas Stárek. 2014. Design of algorithms for payment telematics systems evaluating driver's driving style. *Transactions on Transport Sciences* 7: 9. [CrossRef]
- Lambert, Diane. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34: 1–14. [CrossRef]
- Leisch, Friedrich. 2004. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* 11: 1–18. [CrossRef]
- Ma, Yu-Luen, Xiaoyu Zhu, Xianbiao Hu, and Yi-Chang Chiu. 2018. The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice* 113: 243–58. [CrossRef]
- Makov, Udi, and Jim Weiss. 2016. Predictive modeling for usage-based auto insurance. *Predictive Modeling Applications in Actuarial Science* 2: 290.
- Meinshausen, Nicolai, and Peter Bühlmann. 2006. Variable selection and high-dimensional graphs with the lasso. *Annals of Statistics* 34: 1436–62. [CrossRef]
- Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.
- Osafune, Tatsuaki, Toshimitsu Takahashi, Noboru Kiyama, Tsuneo Sobue, Hirozumi Yamaguchi, and Teruo Higashino. 2017. Analysis of accident risks from driving behaviors. *International Journal of Intelligent Transportation Systems Research* 5: 192–202. [CrossRef]
- Paefgen, Johannes, Thorsten Staake, and Frédéric Thiesse. 2013. Evaluation and aggregation of Pay-As-You-Drive insurance rate factors: A classification analysis approach. *Decision Support Systems* 56: 192–201. [CrossRef]
- Park, Trevor, and George Casella. 2008. The Bayesian lasso. *Journal of the American Statistical Association* 103: 681–86. [CrossRef]



- Shannon, Claude Elwood. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5: 3–55. [\[CrossRef\]](#)
- So, Banghee, Jean-Philippe Boucher, and Emiliano A Valdez. 2021. Cost-sensitive multi-class adaboost for understanding driving behavior based on telematics. *ASTIN Bulletin: The Journal of the IAA* 51: 719–51. [\[CrossRef\]](#)
- Soleymanian, Miremad, Charles B. Weinberg, and Ting Zhu. 2019. Sensor data and behavioral tracking: Does usage-based auto insurance benefit drivers? *Marketing Science* 38: 21–43. [\[CrossRef\]](#)
- Städler, Nicolas, Peter Bühlmann, and Sara Van De Geer. 2010. L1-penalization for mixture regression models. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* 19: 209–56. [\[CrossRef\]](#)
- Stipancic, Joshua, Luis Miranda-Moreno, and Nicolas Saunier. 2018. Vehicle manoeuvres as surrogate safety measures: Extracting data from the GPS-enabled smartphones of regular drivers. *Accident Analysis & Prevention* 115: 160–69.
- Tang, Yanlin, Liya Xiang, and Zhongyi Zhu. 2014. Risk factor selection in rate making: EM adaptive lasso for zero-inflated Poisson regression models. *Risk Analysis* 34: 1112–27. [\[CrossRef\]](#)
- Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58: 267–88. [\[CrossRef\]](#)
- Tselentis, Dimitrios I., George Yannis, and Eleni I. Vlahogianni. 2016. Innovative insurance schemes: Pay As/How You Drive. *Transportation Research Procedia* 14: 362–71. [\[CrossRef\]](#)
- Verbelen, Roel, Katrien Antonio, and Gerda Claeskens. 2018. Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67: 1275–304. [\[CrossRef\]](#)
- Weerasinghe, K. P. M. L., and M. C. Wijegunasekara. 2016. A comparative study of data mining algorithms in the prediction of auto insurance claims. *European International Journal of Science and Technology* 5: 47–54.
- Winlaw, Manda, Stefan H. Steiner, R. Jock MacKay, and Allaa R. Hilal. 2019. Using telematics data to find risky driver behaviour. *Accident Analysis & Prevention* 131: 131–36.
- Wouters, Peter I. J. and John M. J. Bos. 2000. Traffic accident reduction by monitoring driver behaviour with in-car data recorders. *Accident Analysis & Prevention* 32: 643–50.
- Wüthrich, Mario V. 2017. Covariate selection from telematics car driving data. *European Actuarial Journal* 7: 89–108. [\[CrossRef\]](#)
- Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. Regression models for count data in R. *Journal of Statistical Software* 27: 1–25. [\[CrossRef\]](#)
- Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101: 1418–29. [\[CrossRef\]](#)
- Zou, Hui, and Hao Helen Zhang. 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* 37: 1733. [\[CrossRef\]](#)
- Zou, Hui, and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: 301–20. [\[CrossRef\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.