Peiris, Hashan, Himchan Jeong, and Bin Zou. 2025. "Development of Telematics Safety Scores in Accordance with Regulatory Compliance." *Variance* 18 (October). https://doi.org/10.66573/001c.145072.

Financial and Statistical Methods

Development of Telematics Safety Scores in Accordance with Regulatory Compliance

Hashan Peiris[1a], Himchan Jeong[1b], Bin Zou[2c]

[1] Department of Statistics and Actuarial Science, Simon Fraser University, Canada, [2] Department of Mathematics, University of Connecticut, USA

Keywords: Usage-based insurance (UBI), Machine learning, Safety score, Generalized linear model (GLM), Telematics Insurance, Insurance regulations

https://doi.org/10.66573/001c.145072

Variance

Vol. 18, 2025

The paper proposes a ratemaking framework for claim frequency that uses informative telematics data and complies with a "discount-only" regulatory requirement of the sort proposed in the 2023–2024 session of the New York State Assembly. The proposed framework uses a feedforward neural network to extract a one-dimensional safety score from multidimensional telematics features and integrates that score with traditional features in generalized linear models (GLMs). To meet the discount-only requirement, we impose constraints on the safety score and its regression parameter. The results show that the proposed models, with a suitable safety score function, can outperform a standard GLM in both in-sample goodness of fit and out-of-sample prediction performance. Furthermore, the analysis reveals that while the discount-only constraint may drive insurers to raise base premiums to offset revenue losses from the relativity cap, the regulation could achieve its intended goal in scenarios with strong favorable selection.

This work was supported by a 2024 Individual Research Grant from the Casualty Actuarial Society.

Address for Correspondence: himchan_jeong@sfu.ca

1. INTRODUCTION

Telematics automobile insurance, also known as usage-based insurance, leverages technology to monitor driving behavior and adjust policy premiums accordingly. Such coverage involves the use of telematics devices to collect real-time driving data, which distinguishes it from traditional auto insurance, which relies only on demographic features and historical claims data. The detailed information on driving behavior allows insurers to assess risk more accurately and incentivizes safe driving by offering discounts to policyholders showing low-risk driving behavior. Telematics insurance's appeal is echoed by its significant growth in North America, Europe, and other regions. According to Research and Markets, the market size of telematics insurance was USD 4.77 billion in 2024 and will grow at an annual rate of 18.92%, reaching USD 13.77 billion by 2030. However, integrating telematics information into insurance pricing introduces new challenges in privacy, data usage policies, fairness and discrimination issues, and ethical concerns (Handel et al. 2014). These challenges, in turn, compel regulators to establish new requirements on the collection and use of telematics data by insurers.

Telematics data encompasses a wide range of driving behaviors, such as speed, acceleration and breaking patterns, turning radius, and operating hours, among other factors.

a Hashan Peiris is a PhD candidate in statistics at the Department of Statistics and Actuarial Science, Simon Fraser University. He earned his MSc in actuarial science from the same institution in 2023. His research focuses on data integration in insurance, predictive analytics, and, more recently, the application of sports analytics from an actuarial perspective.

b Himchan Jeong holds a PhD in mathematics with a concentration in actuarial science from the University of Connecticut and is also a Fellow of the Society of Actuaries (SOA). Professionally, Himchan has authored over 30 peer-reviewed publications, appearing in the well-known actuarial science and statistics journals such as *Insurance: Mathematics and Economics* and *Journal of Royal Statistical Society: Series A*. He has also been awarded grants from the Canadian Institute of Actuaries, the Casualty Actuarial Society, and the NSERC of Canada. His current research interest is predictive modeling for ratemaking and reserving of property and casualty insurance.

c Bin Zou is an associate professor of actuarial science at the Department of Mathematics, University of Connecticut. Previously, he worked as a postdoc at the University of Washington and the Technical University of Munich, Germany. He obtained his PhD in mathematical finance from the University of Alberta in 2015. His main research interests include optimal decision making in insurance and finance, predictive analytics, and, more recently, cryptocurrencies and DeFi.

Recent studies have demonstrated the usefulness of telematics data in risk assessment and ratemaking—see, for instance, Verbelen, Antonio, and Claeskens (2018), Ayuso, Guillen, and Nielsen (2019), Arumugam and Bhargavi (2019), Denuit, Guillen, and Trufin (2019), Guillen et al. (2019), Huang and Meng (2019), Pesantez-Narvaez, Guillen, and Alcañiz (2019), So, Boucher, and Valdez (2021a), Che, Liebenberg, and Xu (2022), and Henckaerts and Antonio (2022). Specifically, G. Gao, Meng, and Wüthrich (2022) and Y. Gao, Huang, and Meng (2023) show that using telematics data leads to a more precise prediction of the key features that are crucial for identifying high-risk drivers. Given such dominant evidence supporting the use of telematics data, we will utilize telematics features, along with traditional features, to adjust premiums in the proposed framework. However, we face two immediate questions in this endeavor:

1. How do we build a ratemaking model that uses telematics features?
2. How can we ensure the compliance of the model with telematics regulations?

The current literature on telematics insurance mostly focuses on addressing the first question (Q1) but, to the best of our knowledge, completely ignores question two (Q2). This motivates us to propose an insurance ratemaking framework that makes use of new telematics features and, at the same time, complies with the related regulatory requirements. As it turns out, the compliance requirement has a direct impact on how our proposed framework uses telematics information.

Before we introduce our framework, we briefly review several mainstream approaches (that is, answers to Q1) that process telematics data in insurance. One straightforward, perhaps less sophisticated, approach is to integrate telematics features directly with traditional features and use them together in a standard model—in that respect, see Ayuso, Guillen, and Nielsen (2019) and Peiris et al. (2024). Because of the often high-dimensional nature of telematics datasets, applying dimension-reduction techniques when processing the data can be beneficial. Indeed, Jeong (2022) and Chan et al. (2024) show that doing so improves model performance and interpretability. Given the complexity of telematics data, it is not surprising that various machine learning methods, such as neural networks, find great application in telematics insurance (G. Gao, Wang, and Wüthrich 2022; Dong and Quan 2025). In our proposed framework, we use a feedforward neural network (FNN), a special type of neural network, to help process telematics information.

Given the general nature of Q2, a universal answer is unlikely, and a case-by-case approach is more suitable to address that question. As such, we take as our motivation a bill addressing telematics automobile insurance introduced in the 2023–2024 session of the New York State legislature, Assembly Bill 2023-A7614. Clause (D) of that bill prohibits "a premium increase for a driver or vehicle due to a telematics program that measures driving behavior during the current policy period." This particular provision motivates us to impose a *discount-only* constraint on telematics policies.[1] As hinted earlier, such a constraint leads to two significant challenges or consequences. First, because of the complexity and high dimensionality of telematics data, it is nontrivial to propose a model that satisfies the discount-only constraint. Second, the constraint would likely prompt insurers to increase the base premium and would lead to favorable selection bias—that is, drivers with low-risk driving behavior would be more likely to enroll in a telematics policy (Cather 2020).

Having discussed the background and motivation, we now introduce our ratemaking framework for claim frequency. To utilize both traditional and telematics features, we turn to a standard (Poisson) generalized linear model (GLM) with risk embedding via an FNN. To be precise, we use an FNN to build a map from the available telematics features y to a safety score R, that is, $R = f(y)$, in which f, referred to as the *risk-embedding function*, is learned by the FNN. Next, we propose a modified GLM that takes the safety score R, along with the recorded traditional features x, as input, which reads as $\log \mu = \alpha + \beta R + \gamma x$, with $\mu = \mathbb{E}[N]$ denoting the expected claim frequency. With the dimension reduction (y is a high-dimensional vector, but $R = f(y)$ is a one-dimensional scalar) and the above GLM, the discount-only constraint is satisfied if we impose $f(\cdot) \geq 0$ and $\beta < 0$ (which together generate a *nonpositive* component βR in the GLM). As is obvious, the choice of a risk-embedding function f plays a key role in our framework, and we propose two methods, both based on an FNN, to compute the safety score $R = R^{(i)}$, $i = 1, 2$ (see equations (6) and (7) in Section 2 for details). Note that the modified GLM with each $R^{(i)}$ yields a model under the proposed framework, and we consider two such models in this paper.

To test the performance of the two proposed models, we compare them with two benchmark models: Model 1 is a Poisson GLM that uses only the traditional features ($\log \mu = \alpha + \gamma x$); Model 2 is a Poisson GLM that uses both the traditional and telematics features in a parallel manner ($\log \mu = \alpha + \gamma x + \eta y$). We show that the proposed models, with a suitable safety score R, outperform the benchmark models in both in-sample goodness of fit and out-of-sample prediction. We also find that the discount-only constraint on telematics insurance could lead to a potential increase of the base premium for certain groups (high-risk drivers). In addition, when there is a relatively high degree of favorable selection (that is, good drivers are more likely than bad drivers to choose telematics insurance), the discount-only

1 Under such a discount-only constraint, a telematics policy should receive a discount in premium compared with a traditional policy with identical traditional features (covariates).

requirement works as desired by rewarding low-risk drivers, but not at the cost of raising premiums for high-risk drivers.

The paper contributes to a growing body of literature on telematics insurance from a unique perspective—regulatory compliance. To the best of our knowledge, it is the first paper to take into account proposed regulatory requirements on insurers' use of telematics data. Motivated by the discount-only constraint drawn from the introduced New York Assembly Bill A7614, we propose a general ratemaking framework that incorporates the predictive advantage of telematics information and complies with the discount-only regulatory provision. We remark that the proposed framework can be easily modified to accommodate a quantitative, not just a directional (no increase), requirement on the premium of telematics policies. Indeed, we can scale the safety score R to a closed interval (say $[0, 1]$) and impose a threshold on β to limit the impact of the telematics data to achieve any desired bound set by regulators.

The remainder of the article is organized as follows. We introduce our ratemaking framework in Section 2 and test the performance, both in sample and out of sample, of two specific models under the proposed framework in Section 3. We dedicate Section 4 to the study of selection biases on the implied relativities for the telematics policies. We conclude in Section 5.

2. FRAMEWORK

This section introduces a ratemaking framework for claim frequency that utilizes telematics features in the dataset and complies with a discount-only regulatory requirement—such as was proposed in New York Assembly Bill 2023-A7614—that prohibits premium surcharges in the current policy period based on telematics data.

We consider an automobile insurance portfolio with traditional features (such as gender, age, vehicle information, etc.) and telematics features (such as sudden acceleration and braking, turning speed, time and day of driving, etc.); there are $I \in \mathbb{N}$ policyholders in the insurance portfolio. For every policyholder i, $i = 1, \ldots, I$, denote their data entry in the portfolio by $(x_i, y_i; N_i)$, where $x_i \in \mathbb{R}^m$ and $y_i \in \mathbb{R}^n$ record their traditional and telematics features, respectively, and N_i is the claim frequency. We often suppress the subscript i when we consider a generic policyholder. Since the collection of telematics data is often optional and requires the policyholder's consent, it is expected that telematics information is available only on a subset of the entire insurance portfolio.

We start by reviewing two existing approaches that use telematics features in ratemaking. The first, and less sophisticated, approach is to treat both traditional and telematics features in the same way and adopt the following GLM:

$$\log \mu_i = \alpha + \eta \cdot y_i + \gamma \cdot x_i, \tag{1}$$

in which $\alpha \in \mathbb{R}$, $\eta \in \mathbb{R}^n$, $\gamma \in \mathbb{R}^m$ are regression parameters, and $\mu_i = \mathbb{E}[N_i]$ denotes the expected claim frequency of policyholder i and is connected to the linear predictor $\alpha + \eta \cdot y_i + \gamma \cdot x_i$ via the canonical link function (which is set to be the log function here). The second, and more modern, approach is to use a machine learning technique that relaxes the linear structure in the GLM (1). Along this direction, one may directly combine the traditional and telematics features to fine-tune an FNN for the ratemaking purpose. We illustrate one example of this approach in the left panel of Figure 1. However, due to the "black-box" nature of neural networks, this approach will likely fail the regulatory requirements regarding the use of telematics data in insurance ratemaking.

To address the noncompliance drawback of the existing approaches reviewed above, we propose a two-step modeling approach, described as follows. The first step is to make use of the available telematics features in the dataset. To that end, we modify the standard GLM and propose the following model:

$$\log \mu_i = \alpha + \beta R_i + \gamma \cdot x_i, \quad \text{with } R_i = f(y_i), \tag{2}$$

in which $f(\cdot) : \mathbb{R}^n \mapsto \mathbb{R}$ is a *risk-embedding function*. Comparing the modified GLM above in (2) with the standard GLM in (1), the essential difference is that we do *not* directly regress the telematics features y but instead apply a risk-embedding function f to reduce the *multidimensional* features y into a *one-dimensional* scalar $R = f(y)$, which we call the *safety score*, and then use the transformed safety score R, along with the traditional features x, in the regression. We will discuss how we compute the safety score later.

In the second step, we "translate" the regulatory requirements into mathematical conditions and impose them on the risk-embedding function f and/or the regression parameters in (2). In this way, the modified GLM in (2), together with appropriate regulatory constraints, will fully comply with the regulations on telematics data in insurance ratemaking. In this work, we are particularly interested in the *discount-only* regulatory requirement on telematics policies. For that purpose, we impose the following constraints on (2):

$$f(\cdot) \geq 0 \quad \text{and} \quad \beta < 0. \tag{3}$$

We easily see that, all things being equal, the combined constraints in (3) produce a *nonpositive* component βR_i in (2), which can be seen as a "discount" in premium to telematics policies. Note that with the constraints (3) in force, the smaller the safety score R_i, the higher the risk associated with the policyholder i.

Remark 2.1. *We use this remark to make several technical comments on the proposed two-step modeling approach. First, the modification of $\eta \cdot y_i$ in the standard GLM (1) to $\beta R_i = \beta f(y_i)$ in the proposed model (2) allows us to impose regulatory constraints effectively. Recall that the dimension of the telematics features, n, is often large; if one were to impose constraints directly on its regression coefficient η, it would result in a large number of constraints and thus reduce the efficiency of estimation and prediction. Second, the introduction of the risk-embedding function f leads to a highly flexible model in (2). In fact, (2) is a family of models that can take parametric or nonparametric forms depending on the choice of the risk-embedding function f. We will discuss different methods for constructing f later. Third, imposing constraints on the model (2) can affect the overall rate level (base premium). For example, the constraints in (3) will place a cap on the relativity from the telematics features by 1; consequently, insur-*

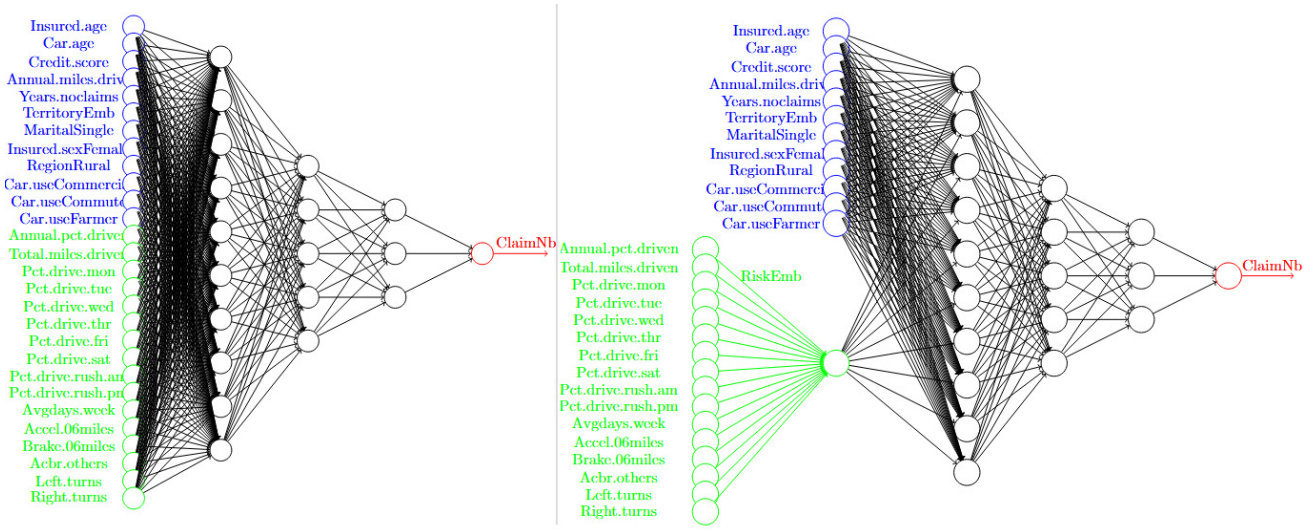

Figure 1. Examples of FNN architectures with both traditional and telematics features.

Note: In the left panel, the telematics features are used directly, in the same way as traditional features, to train the FNN. In the right panel, the telematics features are first transformed into a safety score via a hidden layer; next, the safety score and the traditional features are used to train the FNN.

ers may increase the base premium to compensate for the loss of premium due to the relativity cap (see Werner et al. [2016, 279–80] for an example).

Remark 2.2. *Note that almost all policyholders can receive some level of discount under GLM (2) when the constraints in (3) are imposed (implying $\beta R_i \leq 0$). As discussed earlier, the nonpositivity of βR_i is the key to compliance with the discount-only regulatory constraint. However, as a referee pointed out, this may not align perfectly with intuition, since insurers may want to penalize bad drivers (with small safety scores R_i). To further accommodate this feature, one solution is to group policies by their safety scores and introduce a reference group in ratemaking. We outline the key idea of this solution below and refer the reader to Section 4 (see equation (10)) for a detailed implementation. Suppose that the insurer separates all telematics policies into G groups based on their safety scores, and group 1 (respectively, group G) is the least (respectively, most) preferred to the insurer. We replace the term βR_i in (2) by*

$$\sum_{g=1}^{G} \beta_g \cdot \mathbb{1}_{\{R_i \in [\kappa_{g-1}, \kappa_g)\}},$$

in which policy i is in group g if $R_i \in [\kappa_{g-1}, \kappa_g)$. We choose group 1 as the reference group by setting $\beta_1 = 0$ (so that policyholders in group 1 do not *receive any discount) and impose $\beta_G < \beta_{G-1} < \cdots < \beta_2 < 0$ (so that policyholders in a more preferred group receive more discounts). Last, "discount" is a* relative *concept and, thus, relies on the choice of a benchmark (or base premium). With the discount-only regulatory constraint in place, insurers cannot increase the premium only for telematics policies, but they can adjust the base premium for* all *policies upward to offset the discounts offered to telematics policies. In addition, favorable selection bias could occur, in the sense that good drivers are more likely to sign up for telematics insurance than bad drivers. We conduct a detailed study of this effect on the base premium in Section 4.*

As is clear from the proposed GLM (2), a key component of the model is the safety score $R = f(y)$. In the rest of this section, we introduce two methods for constructing the safety score, both of which employ the powerful FNN but in different ways. Note that the way to construct the safety score is certainly not unique, and we mention a variant of Method 1 in Remark 3.

Method 1. We first train a fully connected FNN for risk classification, as shown in the left panel of Figure 1, and denote the prediction for policyholder i by $\mu_i^{FNN} := \mu_i^{FNN}(x_i, y_i)$. Note that μ_i^{FNN} utilizes both traditional features x_i and telematics features y_i. In the meantime, we consider a standard Poisson GLM that uses only traditional features; assume that N_i follows a Poisson distribution with intensity (mean) μ_i^{Trad}, $N_i \sim \mathcal{P}(\mu_i^{Trad})$, and that

$$\log \mu_i^{Trad} = \alpha + \gamma \cdot x_i, \tag{4}$$

in which $\alpha \in \mathbb{R}$, $\gamma \in \mathbb{R}^m$ are parameters. Now, with both μ_i^{FNN} and μ_i^{Trad} in hand, we propose the first method for obtaining the safety score by

$$R_i^{(1)} := \log \mu_i^{FNN} - \log \mu_i^{Trad}. \tag{5}$$

We remark that the safety score in (5) captures the explanatory power of the telematics features as a one-dimensional variable. We apply a simple affine transformation so that the final safety score is between 0 and 1. To be precise, we apply the following transformation:

$$R_i^{(1)} \to \frac{R_i^{(1)} - R_m^{(1)}}{R_M^{(1)} - R_m^{(1)}}, \tag{6}$$

in which $R_m^{(1)} := \min_i R_i^{(1)}$ and $R_M^{(1)} := \max_i R_i^{(1)}$ denote the minimum and maximum values of the safety scores among all policyholders computed from (5). With a little abuse of notation, we still denote the right-hand side of (6) by $R_i^{(1)}$.

Method 2. Recall that μ_i^{FNN} in Method 1 is obtained from an FNN that is trained with both the traditional and telematics features as direct inputs. In Method 2, we adopt an FNN with an extra hidden layer at the initial stage, $f^*(\cdot) : \mathbb{R}^n \to \mathbb{R}$; please refer to the right-hand panel of Figure 1 for illustration. The purpose of such a hidden layer

is to map the high-dimensional telematics features $y \in \mathbb{R}^n$ to a one-dimensional variable $f^*(y) \in \mathbb{R}$. Note that the embedding map f^* is extracted from the fine-tuned FNN, and as such, the construction of f^* takes into account not only the relationship between the claim frequency N and the telematics features y, but also the possible interaction between the traditional features x and the telematics features y. Once the fully connected FNN is calibrated and f^* is extracted, we define the safety score under Method 2 by

$$R_i^{(2)} = f^*(y_i) \to \frac{R_i^{(2)} - R_m^{(2)}}{R_M^{(2)} - R_m^{(2)}}, \tag{7}$$

which is then normalized so that $R_i^{(2)}$ takes values between 0 and 1. $R_m^{(2)}$ and $R_M^{(2)}$ in (7) are defined in a similar fashion as their counterparts in (6). After the linear transformation, $R_i^{(2)}$ is always between 0 and 1.

Remark 2.3. *When constructing $R^{(1)}$ under Method 1, we compare the prediction difference between a simple GLM using only the traditional features x in equation (4) and an FNN using both the traditional and telematics features x and y in the left panel of Figure 1. An anonymous referee points out an alternative method to us, and the suggestion is to replace the GLM prediction μ_i^{Trad} by the FNN prediction only using x, $\mu_i^{FNN(x)}$. Under this alternative method, we construct a new safety score by*

$$R_i^{(3)} := \log \mu_i^{FNN} - \log \mu_i^{FNN(x)}$$

which then goes through the same transformation as in (6) so that $R_i^{(3)} \in [0, 1]$. Recall that μ_i^{FNN} is the FNN prediction using both x and y. However, through a detailed study (available upon request), we find that the model with $R^{(1)}$ outperforms the model with $R^{(3)}$ in all metrics considered under both in-sample and out-of-sample tests. For this reason, we do not include the model with $R^{(3)}$ in the subsequent analysis.

We close this section with some technical details on the FNNs shown in Figure 1. The general neural networks here include two types of input: one for traditional features, which we call "TradInput" (colored blue in Figure 1), and the other for telematics features, which we call "TeleInput" (colored green in Figure 1). Similar to Schelldorfer and Wüthrich (2019), we consider three hidden layers (hidden1, hidden2, hidden3) and choose the hyperbolic tangent activation function. A regression layer using a linear activation function is set as the output layer. An additional input (LogVolGLM) is used for the nontrainable exposure, which is concatenated with the output layer of the FNN to return the expected number of claims. The model is then compiled using the Poisson loss function and the Nadam optimizer to adaptively adjust the learning rate. To train the model, we set the number of epochs to 500 (that is, the entire training dataset would be passed through the model 500 times). The batch size is set to 200, and the model updates its weights after each batch of 200 samples. To overcome the potential overfitting issue, we reserve 20% of the training data for validation when we train the FNN. Using a grid search, we find the number of nodes in the three hidden layers as (10,5,3), exactly as shown in Figure 1, which yields the minimum loss under the validation dataset. Finally, we train the model as specified in Method 1 and Method 2, respectively.

3. MODEL ANALYSIS

3.1. MODEL SPECIFICATIONS

Based on the framework proposed in Section 2, we consider the following four different models for ratemaking:

1. "Trad Model" (Model 1) is a standard Poisson GLM specified in (4) and using only the traditional features x in the dataset.
2. "TRaw Model" (Model 2) is also a Poisson GLM, specified in (1), but in contrast to Trad Model, it utilizes both the traditional features x and the telematics features y in the dataset.
3. "TScore1 Model" (Model 3) is a modified Poisson GLM specified in (2) along with the regulatory constraint (3), with the safety score in (2) given by $R_i^{(1)}$ in (6).
4. "TScore2 Model" (Model 4) is the same as TScore1 Model, except that the safety score is given by $R_i^{(2)}$ in (7).

Note that among the four models, Model 1 is the only one that does *not* use the telematics features in estimation and prediction. The remaining three models, Models 2, 3, and 4, do use the telematics features. From the discussion in Section 2, we know that Model 2 may fail to comply with the discount-only regulatory requirement, but Models 3 and 4 are fully compliant with that requirement, due to the constraint (3) imposed on the model parameters (2).

We consider a synthetic telematics dataset used in Jeong (2022), which is a processed version of the synthetic dataset from So, Boucher, and Valdez (2021b). This dataset contains 100,000 observations and records 11 traditional features and 10 telematics features. We provide details on the dataset in Appendix A. In the study, we randomly split the full dataset consisting of 100,000 observations into two parts: the first part, with 90,000 observations, is used solely for training the models; the second part, containing the remaining 10,000 observations, is reserved for out-of-sample model validation.

3.2. MODEL ESTIMATIONS

In this section, we use the training dataset (which consists of 90% of the full dataset) to estimate the four models introduced in Section 3.1. We present the detailed estimation results on all model parameters in Table 1. In the paragraphs that follow, we summarize the key observations from Table 1.

First, the signs of the γ parameters (for the traditional features) are mostly consistent across the four models. This result shows the "consensus" among the models on how a traditional feature contributes to claim frequency. Taking the γ parameter for the feature `Credit.score` as an example, we observe that it is always negative with a negligible p-value (high statistical significance), an observation that implies that policyholders with higher credit scores are less likely to get involved in a car accident than those with lower credit scores. Among the 11 traditional features, `Car.age`, `Credit.score`, `Annual.miles.drive`, `Years.noclaims`, and `Ter-`

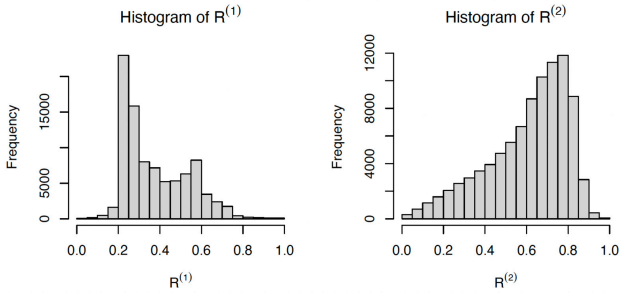

Figure 2. Histograms of the safety scores $R^{(1)}$ and $R^{(2)}$.

ritoryEmb are statistically significant at the 1% level for all four models.

Second, we observe that the estimated signs of η for Pct.drive.day—a telematics feature that records the percentages of driving on a day (from Monday to Saturday)—are always positive when the reference level is Sunday. This result suggests that it is much safer to drive on Sundays than other days, possibly because of reduced traffic on that day. We also find that rush hour driving in the afternoons (mostly from the workplace to home) is much riskier than that in the mornings. All the telematics features, with the exception of Acbr.others (number of sudden accelerations and brakes), are statistically significant at the 1% level, demonstrating a good potential of predictive power from telematics data.

Regarding the α parameter (intercept in a model), we observe that it is negative in Trad Model and TRaw Model but positive in TScore1 Model and TScore2 Model, and the difference in value is rather noticeable across models. Aligned with the fact that the estimated β is negative in both TScore1 Model and TScore2 Model, it implies that the base premium for these models is greater than that for Trad Model and TRaw Model.

Last, we report in Figure 2 the distribution of the safety scores, $R^{(1)}$ in TScore1 Model (see its definition in (6)) and $R^{(2)}$ in TScore2 Model (see its definition in (7)), obtained from the training dataset. We observe a significant difference in their distributions, with $R^{(1)}$ more concentrated on small values and $R^{(2)}$ on large values. However, it is worth noting that the safety scores obtained from different methods are *not* directly comparable, and the takeaway message from Figure 2 is that the construction methods for the safety score could lead to major differences in their values.

3.3. MODEL EVALUATIONS

First, we conduct an in-sample goodness-of-fit test on all four models introduced in Section 3.1 and compare how those models fit the training data. In the test, we consider three metrics—the log-likelihood (logLik), Akaike information criterion (AIC), and Bayesian information criterion (BIC) values. For the log-likelihood metric, a model with higher values is preferred, but for both the AIC and BIC, a model with small values is preferred. We present the results in Table 2.

In all three metrics, Trad Model (Model 1) has the worst fit to the training data, and it is noticeably worse than the other models. Recall that Trad Model is the traditional GLM and the only model considered that does not use the telematics features. As such, we conclude that the information contained in the telematics features is valuable and helps improve the model goodness of fit. We observe, across all three metrics, the ranking among the three models that utilize the telematics features as follows:

$$\begin{aligned}\text{TScore1 (Model 3)} &\succ \text{TRaw (Model 2)} \\ &\succ \text{TScore2 (Model 4)}.\end{aligned} \tag{8}$$

That is, the model with the safety score $R = R^{(1)}$ in (6) is the best, and the one with the safety score $R = R^{(2)}$ in (7) is the worst, with TRaw always in between. Recall that the modified GLM (2) in Models 3 and 4 takes the *one-dimensional* safety score R, not the multidimensional telematics features y, as a single input, but TRaw (Model 2) directly uses y in its GLM (1). Therefore, it is pleasing to see from (8) that such a dimension reduction does *not* necessarily lead to a poorer fit of the data. Instead, the method (mapping) that reduces y to R itself plays an important role in the final goodness-of-fit results. For the two methods we consider, the first in (6) outperforms the second in (7) in terms of model fitness. Note that it is also possible to understand the performance differences between TScore1 (with $R^{(1)}$) and TScore2 (with $R^{(2)}$) by their difference in capturing the interaction effect between traditional features x and telematics features y. To be precise, $R^{(1)}$ captures all interactions between x and y simultaneously; in comparison, $R^{(2)}$ captures only the interactions between x and a one-dimensional risk-embedding $f^*(y)$, which is reduced from "raw" telematics features y. Therefore, TScore1's superiority over TScore2 could be attributed to the fact that $R^{(1)}$ better captures potential hidden interaction effects between x and y. For example, if a vehicle is frequently used for delivery or ride-sharing service, it is expected that the annual mileage and night-driving time of the vehicle are much higher than that of a vehicle primarily used for personal purposes, and the interactions between those two features can be captured by $R^{(1)}$.

Next, we study the out-of-sample prediction performance of all four models using the 10,000 observations in the validation dataset. We consider three popular metrics in model evaluation: the root-mean-square error (RMSE), the mean absolute error (MAE), and the Poisson deviance (DEV). Both RMSE and MAE are widely used evaluation metrics, and their names indicate how they are computed. However, as DEV is slightly less well known, we give its definition as follows:

$$\text{DEV} = \frac{2}{|\mathcal{T}|}\sum_{i\in\mathcal{T}}\left[N_i\log(N_i/\hat{\mu}_i) + (N_i - \hat{\mu}_i)\right],$$

in which $|\mathcal{T}|$ is the size of the test dataset $\mathcal{T}$, and $\hat{\mu}_i$ is the predicted value of N_i under a given model. Note that there is no consensus on which metric is a better choice for model evaluation; for all three metrics, however, the smaller the value, the better the prediction performance. We compute the RMSE, MAE, and DEV values for all four models and present them in Table 3. Similar to the in-sample goodness-of-fit results in Table 2, Trad (Model 1) delivers the worst performance among the four models, and this finding shows that telematics information is also valuable for pre-

Table 1. Detailed estimation results for all four models introduced in Section 3.1

		Trad	TRaw	TScore1	TScore2
α	(Intercept)	−0.4331 (0.0058)	−5.2711 (0.0000)	2.6114 (0.0000)	1.0082 (0.0000)
γ	Insured.age	0.0025 (0.1858)	0.0035 (0.0751)	0.0028 (0.1485)	0.0087 (0.0000)
	Insured.sexFemale	0.0265 (0.4059)	−0.0387 (0.2290)	0.0470 (0.1444)	0.0081 (0.7999)
	Car.age	−0.0697 (0.0000)	−0.0623 (0.0000)	−0.0689 (0.0000)	−0.0544 (0.0000)
	MaritalSingle	0.0332 (0.3433)	0.0389 (0.2665)	−0.0238 (0.5114)	0.0525 (0.1328)
	Car.useCommercial	0.1775 (0.0403)	−0.0014 (0.9875)	0.2816 (0.0010)	0.0509 (0.5552)
	Car.useCommute	−0.0220 (0.5713)	−0.0560 (0.1699)	0.0188 (0.6353)	−0.0732 (0.0618)
	Car.useFarmer	−0.4616 (0.0465)	−0.3728 (0.1080)	−0.5331 (0.0223)	−0.3949 (0.0885)
	Credit.score	−0.0030 (0.0000)	−0.0029 (0.0000)	−0.0032 (0.0000)	−0.0026 (0.0000)
	RegionRural	−0.1177 (0.0059)	−0.0801 (0.0726)	−0.1129 (0.0087)	−0.2495 (0.0000)
	Annual.miles.drive	0.0000 (0.0000)	0.0000 (0.0001)	0.0000 (0.0000)	0.0000 (0.1973)
	Years.noclaims	−0.0091 (0.0000)	−0.0048 (0.0084)	−0.0107 (0.0000)	−0.0078 (0.0000)
	TerritoryEmb	0.5013 (0.0000)	0.4042 (0.0000)	0.4872 (0.0000)	0.4491 (0.0000)
η	Annual.pct.driven		2.0186 (0.0000)		
	Total.miles.driven		0.0000 (0.0000)		
	Pct.drive.mon		3.9104 (0.0000)		
	Pct.drive.tue		2.9269 (0.0000)		
	Pct.drive.wed		2.3149 (0.0007)		
	Pct.drive.thr		4.3848 (0.0000)		
	Pct.drive.fri		2.5579 (0.0001)		
	Pct.drive.sat		2.0609 (0.0099)		
	Pct.drive.rush.am		−2.0283 (0.0000)		
	Pct.drive.rush.pm		0.9061 (0.0007)		
	Avgdays.week		0.0512 (0.0059)		
	Accel.06miles		0.0016 (0.0000)		
	Brake.06miles		0.0021 (0.0000)		
	Acbr.others		−0.0002 (0.2064)		
	Left.turns		0.0000 (0.0069)		
	Right.turns		0.0000 (0.0022)		
β	Tscore			−9.4039 (0.0000)	−3.5006 (0.0000)

Table 2. In-sample goodness-of-fit results

	Trad	TRaw	TScore1	TScore2
logLik	−15882.95	−14625.14	−14008.52	−14769.31
AIC	31791.90	29308.28	28045.04	29566.61
BIC	31914.20	29581.10	28176.75	29698.32

diction. Together, our results from Tables 2 and 3 favor the collection and use of telematics data in insurance ratemaking.

As in the case of goodness-of-fit ranking, TScore1 (Model 3) achieves the best performance in all out-of-sample validation criteria, while the performance of TScore2 (Model 4) is somewhat comparable to that of TRaw (Model 2). To quickly summarize, the proposed framework not only leads to ratemaking models that are fully compliant with the discount-only regulatory constraint, but also has the potential to outperform TRaw (Model 2), which does not satisfy the regulatory requirement. Indeed, TScore1 (Model 3) is strictly preferred to TRaw (Model 2) under all three metrics.

Lastly, we use the validation dataset to compare the premiums under Trad (Model 1) and TScore$j$ (Model 3 with $j = 1$ and Model 4 with $j = 2$). Recall that Trad (Model 1) does not use the telematics features. For that purpose, we define the relative difference of premiums between TScore$j$ and Trad by

Table 3. Out-of-sample prediction results

	Trad	TRaw	TScore1	TScore2
RMSE	0.04787	0.04549	0.04538	0.04547
MAE	0.08621	0.08213	0.08172	0.08221
DEV	0.27418	0.23835	0.23722	0.24069

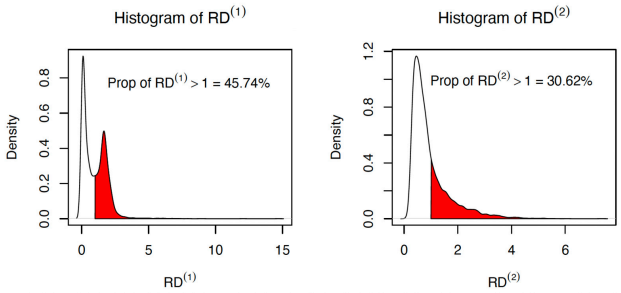

Figure 3. Distributions of the relative premium differences $RD^{(1)}$ and $RD^{(2)}$.

$$RD_i^{(j)} = \frac{\mu_i^{\text{TScore}j}}{\mu_i^{\text{Trad}}}, \quad j = 1, 2.$$

Figure 3 displays the distributions of $RD^{(1)}$ and $RD^{(2)}$ under the validation dataset. We observe that about 46% (respectively, 31%) of policyholders will pay a higher premium if the insurer replaces Trad with TScore1 (respectively, TScore2) for ratemaking.

We conclude this section by remarking on the practical applicability of Models 3 and 4. Consider a new driver who wants to purchase a telematics insurance policy but has no "safety score" due to the lack of observed driving behavior. In this case, an insurer can accept her into its telematics portfolio by offering her an up-front discount (or treating her as in a safe group in terms of driving behavior) to attract new customers. The insurer then observes her actual driving via the telematics device or app over a certain period of time before it can determine her risk score (class). Upon renewal, the insurer uses the collected telematics information to decide whether the driver should continue to receive the premium discount. Such a practice is indeed followed by insurers in real markets, and one example is Progressive (see their telematics insurance website, https://www.progressive.com/auto/discounts/snapshot/, for full details).

4. IMPACT OF SELECTION BIAS ON BASE PREMIUM

We show in Section 3 that both the in-sample and out-of-sample performance of the embedding models is promising. However, as already hinted in Remark 1, this might come at a cost—(favorable) selection bias; see, for instance, Cather (2020) for a recent treatment of this topic. In this section, we study the impact of favorable selection on the base premium.

To start, we partition the full dataset $\mathcal{S}$ into two exclusive sub-datasets $\mathcal{S}_0$ and $\mathcal{S}_1$ as follows: for policyholders in $\mathcal{S}_0$, both traditional and telematics features are observed, but for policyholders in $\mathcal{S}_1$, only traditional features are recorded. We further assume that N_i, the claim frequency of the policyholder i, follows a Poisson distribution, $N_i \sim \mathcal{P}(\mu_i)$, with $\mu_i = \mathbb{E}[N_i]$. We first consider the traditional GLM in (4) with log-link function, $\log \mu_i = \alpha + \gamma \cdot x_i$, for all policyholders $i \in \mathcal{S}_0 \cup \mathcal{S}_1$. Denote $(\hat{\alpha}, \hat{\gamma})$ the estimates of the traditional GLM; that is, $(\hat{\alpha}, \hat{\gamma})$ solves the following optimization problem:

$$(\hat{\alpha}, \hat{\gamma}) = \underset{(\alpha,\gamma)\in\mathbb{R}\times\mathbb{R}^m}{\operatorname{argmin}} \sum_{i\in\mathcal{S}_0\cup\mathcal{S}_1} \left(-\mu_i + \log N_i \cdot \log \mu_i\right). \quad (9)$$

Given $\hat{\alpha}$ from (9), we define *base premium* M by $M = \exp(\hat{\alpha})$.

To better assess the impact of selection bias resulting from the embedding safety score, we group policyholders by their safety scores and consider a version of the discretized telematics score embedding model, under which the range $[0, 1]$ of the safety scores $R^{(1)}$ and $R^{(2)}$ are divided into four groups as follows: Group 1 with $R_i^{(j)} \in [0, q_{[1]}^{(j)})$, Group 2 with $R_i^{(j)} \in [q_{[1]}^{(j)}, q_{[2]}^{(j)})$, Group 3 with $R_i^{(j)} \in [q_{[2]}^{(j)}, q_{[3]}^{(j)})$, and Group 4 with $R_i^{(j)} \in [q_{[3]}^{(j)}, 1]$, in which $q_{[1]}^{(j)}$, $q_{[2]}^{(j)}$, and $q_{[3]}^{(j)}$ denote the first, second, and third quartiles, respectively, of the empirical safety scores $R^{(j)}$ calculated using the training dataset for $j = 1, 2$. (The empirical distributions of $R^{(1)}$ and $R^{(2)}$ are quite different, as seen from Figure 2, so their quartiles $q_{[k]}^{(j)}$ are also different.) With the above grouping, we consider the discretized embedding safety score models

$$\begin{aligned} \log \mu_i = \hat{\alpha} + \hat{\gamma} \cdot x_i &+ \beta_1 \cdot \mathbb{1}_{\{R_i^{(j)} \in [0, q_{[1]}^{(j)})\}} \\ &+ \beta_2 \cdot \mathbb{1}_{\{R_i^{(j)} \in [q_{[1]}^{(j)}, q_{[2]}^{(j)})\}} \\ &+ \alpha^* + \beta_3 \cdot \mathbb{1}_{\{R_i^{(j)} \in [q_{[2]}^{(j)}, q_{[3]}^{(j)})\}} \\ &+ \beta_4 \cdot \mathbb{1}_{\{R_i^{(j)} \in [q_{[3]}^{(j)}, 1]\}}, \end{aligned} \quad (10)$$

for $i \in \mathcal{S}_0$ and $j = 1, 2$, in which $\hat{\alpha}$ and $\hat{\gamma}$ are obtained from (9), and the parameters satisfy

$$\beta_4 \le \beta_3 \le \beta_2 \le \beta_1 = 0.$$

The above constraint implies that if a policyholder belongs to a safer safety score band, she should expect a bigger discount on the premium. Because the larger the safety score, the lower the riskiness of a policy, Group 4 is the most preferred and Group 1 the least preferred to the insurer. As Group 1 is the least preferred, there shall be no discount applied to that group as implied in $\beta_1 = 0$. Lastly, α^* accounts for the different mean risk level between $\mathcal{S}_0$ and the full dataset $\mathcal{S}$, which is the so-called market segmentation effect. If we can observe the telematics features from all policyholders so that $\mathcal{S}_0 = \mathcal{S}$, then by definition $\alpha^* = 0$.

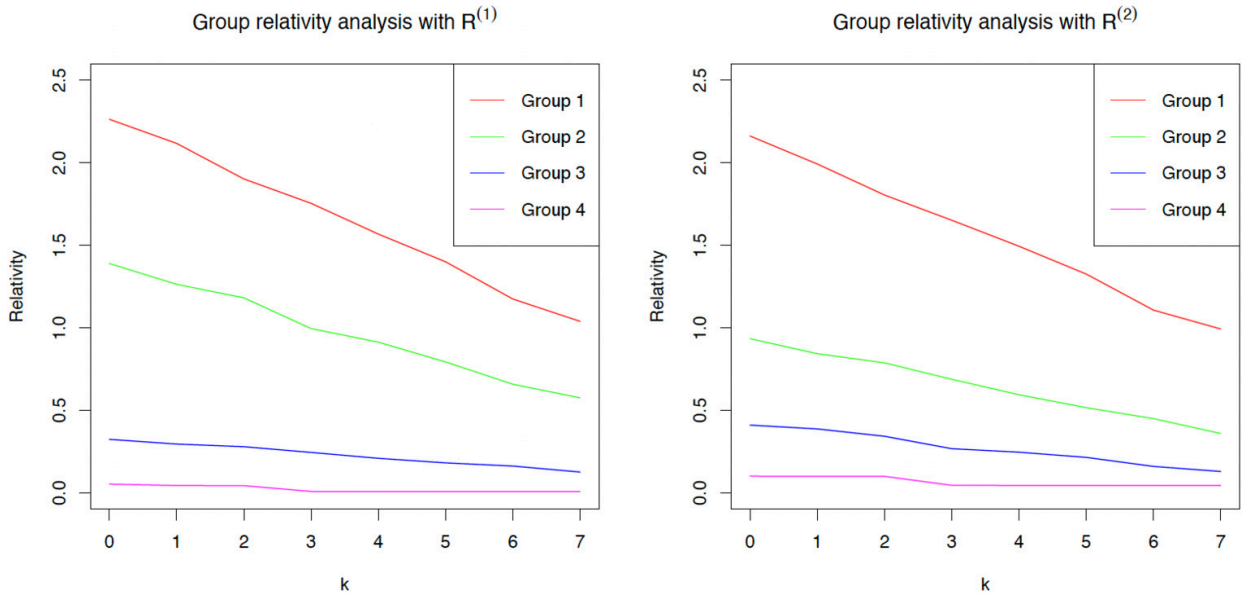

Figure 4. Relativities RL in (11) for the four safety score groups under different k.

The models proposed in (10) allow us to easily analyze the *realized relativities* due to the (observed) safety score computed from the telematics features of a policyholder in $\mathcal{S}_0$. Indeed, given the βs from (10), we define the relativity of policyholders in Group m, compared to the base premium due to their safety score, by

$$RL_m = \exp\left(\beta_m + \alpha^*\right), \quad m = 1, 2, 3, 4. \tag{11}$$

Since the degree of favorable selection can be modeled by the split of $\mathcal{S}_0$ and $\mathcal{S}_1$, we consider diverse selection schemes for traditional and telematics policies and study the corresponding changes in the implied relativities RL_1, RL_2, RL_3, and RL_4 due to the telematics safety score.

Assume that the sampling probability of a policy with claim frequency N_i in $\mathcal{S}_0$ (the dataset with both the traditional and telematics features) is given by

$$p_i = \frac{1}{1 + \exp(k \cdot N_i/6)},$$

in which k is a parameter controlling the degree of favorable selection.[2] The larger the value of k, the stronger the favorable selection. To see this, consider $N_i = 0, k = 7$ and $N_i = 3, k = 7$; we compute $p_i(N_i = 0, k = 7) = 0.5$ and $p_i(N_i = 3, k = 7) = \frac{1}{1+\exp(3.5)} \simeq 0.029$, so policies with three claims are much less likely to appear in the telematics dataset $\mathcal{S}_0$ than the policies with no claim. We consider various levels of k, ranging from 0 to 7. Note that $k = 0$ corresponds to the random selection scheme, since $p_i = 0.5$ for all i, while $k = 7$ represents an extreme scenario of favorable selection. We compute the relativities, defined by (11), for the four groups based on the telematics safety scores, and present the results in Figure 4. The key findings are summarized below.

- When there is no selection bias ($k = 0$), and thus $\mathcal{S}_0$ is a random sample of the population (that is, $\mathcal{S}_0$ has the same distribution as the population of all policyholders), the two proposed models, TScore1 (Model 3) and TScore2 (Model 4), can lead to higher relative premiums for some groups. For example, if the safety score is calculated by $R^{(1)}$, both Groups 1 and 2 have relativities greater than 1, so policyholders in those groups experience premium surcharges. This result is not surprising because there is an *implicit* increase in the base premium for the telematics policies, implied by the difference in the estimated values of α parameter between Trad (–0.4331) and the TScore1/TScore2 models (2.6114 and 1.0082, respectively) in Table 1.
- As the degree of favorable selection increases (k increases), the base premium for the telematics policies in Groups 1 and 2 decreases significantly but remains relatively stable for Groups 3 and 4; this observation applies to both $R^{(1)}$ and $R^{(2)}$. In the extreme case of $k = 7$, the telematics policies in Groups 2 through 4 when $R = R^{(1)}$ (respectively, all groups when $R = R^{(2)}$) in the telematics dataset $\mathcal{S}_0$ receive a premium discount compared with those with only the traditional features. This result shows that the discount-only regulatory requirement could work as intended if there is a strong favorable selection (policyholders with good claim history are more attracted to the telematics policies).

2 The particular form of the probability p_i is also used in Peiris et al. (2024) to describe the favorable selection scheme.

5. CONCLUSION

The rapid growth of telematics insurance in recent years poses new challenges and concerns in the areas of policyholder privacy, use of telematics data, and fairness/discrimination in pricing, among others. Such concerns will necessarily compel regulators to implement new laws or policies regulating the insurance industry's collection and use of telematics data. As a prime example, legislators in the state of New York introduced Assembly Bill 2023-A7614 in the 2023–2024 session, which contained a provision to prohibit the increase of premiums based on telematics data in the current policy year, which we term a "discount-only" regulatory requirement. The essential goal of this paper is to propose a ratemaking framework for claim frequency that, on the one hand, utilizes the available telematics features and, on the other hand, complies with such a discount-only regulatory requirement. To that end, we combine the powerful FFN with a standard GLM to build modified GLMs via a two-step approach. We first apply an FNN to transform the telematics features into a one-dimensional safety score, referred to as risk embedding, and next use it along with the traditional features in GLMs. In a second step, we impose appropriate constraints on the safety score and its regression coefficient in the GLMs to satisfy the discount-only regulatory requirement. We show that for a suitable choice of risk embedding, the proposed models can outperform standard GLMs (with or without the use of telematics features) in both in-sample goodness of fit and out-of-sample prediction. We also study the impact of selection bias (i.e., good drivers are more likely to enroll in telematics policies than bad drivers) on the base premium. Our results confirm that the discount-only constraint may force insurers to increase the base premium to compensate for the loss of revenue due to the relativity cap. However, this unwanted effect is largely alleviated in the presence of a sufficient level of favorable selection. Lastly, the safety scores we construct also have the potential to help reduce the industry's reliance on some of the controversial traditional covariates (such as age and gender) in ratemaking practices, which aligns with recent findings in Ayuso, Guillen, and Pérez-Marín (2016) and Boucher and Pigeon (2024).

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their valuable comments on an earlier version of the paper. This project was funded by a Casualty Actuarial Society 2024 individual research grant. The first two authors were also partially supported by a Natural Sciences and Engineering Research Council of Canada grant (R832535).

Submitted: December 09, 2024 EDT. Accepted: August 24, 2025 EDT. Published: October 28, 2025 EDT.

REFERENCES

Arumugam, S., and R. Bhargavi. 2019. "A Survey on Driving Behavior Analysis in Usage Based Insurance Using Big Data." *Journal of Big Data* 6:1–21. https://doi.org/10.1186/s40537-019-0249-5.

Ayuso, M., M. Guillen, and J. P. Nielsen. 2019. "Improving Automobile Insurance Ratemaking Using Telematics: Incorporating Mileage and Driver Behaviour Data." *Transportation* 46:735–52. https://doi.org/10.1007/s11116-018-9890-7.

Ayuso, M., M. Guillen, and A. M. Pérez-Marín. 2016. "Telematics and Gender Discrimination: Some Usage-Based Evidence on Whether Men's Risk of Accidents Differs from Women's." *Risks* 4 (2): 10. https://doi.org/10.3390/risks4020010.

Boucher, J.-P., and M. Pigeon. 2024. "Balancing Risk Assessment and Social Fairness: An Auto Telematics Case Study." CAS Research Paper: Series on Race and Insurance Pricing. Casualty Actuarial Society.

Cather, D. A. 2020. "Reconsidering Insurance Discrimination and Adverse Selection in an Era of Data Analytics." *The Geneva Papers on Risk and Insurance—Issues and Practice* 45:426–56. https://doi.org/10.1057/s41288-020-00166-7.

Chan, I. W., S. C. Tseung, A. L. Badescu, and X. S. Lin. 2024. "Data Mining of Telematics Data: Unveiling the Hidden Patterns in Driving Behaviour." *arXiv*, July. https://doi.org/10.48550/arXiv.2304.10591.

Che, X., A. Liebenberg, and J. Xu. 2022. "Usage-Based Insurance—Impact on Insurers and Potential Implications for InsurTech." *North American Actuarial Journal* 26 (3): 428–55. https://doi.org/10.1080/10920277.2021.1953536.

Denuit, M., M. Guillen, and J. Trufin. 2019. "Multivariate Credibility Modelling for Usage-Based Motor Insurance Pricing with Behavioural Data." *Annals of Actuarial Science* 13 (2): 378–99. https://doi.org/10.1017/S1748499518000349.

Dong, P., and Z. Quan. 2025. "Automated Machine Learning in Insurance." *Insurance: Mathematics and Economics* 120:17–41. https://doi.org/10.1016/j.insmatheco.2024.10.002.

Gao, G., S. Meng, and M. V. Wüthrich. 2022. "What Can We Learn from Telematics Car Driving Data: A Survey." *Insurance: Mathematics and Economics* 104:185–99. https://doi.org/10.1016/j.insmatheco.2022.02.004.

Gao, G., H. Wang, and M. V. Wüthrich. 2022. "Boosting Poisson Regression Models with Telematics Car Driving Data." *Machine Learning* 111 (1): 243–72. https://doi.org/10.1007/s10994-021-05957-0.

Gao, Y., Y. Huang, and S. Meng. 2023. "Evaluation and Interpretation of Driving Risks: Automobile Claim Frequency Modeling with Telematics Data." *Statistical Analysis and Data Mining* 16 (2): 97–119. https://doi.org/10.1002/sam.11599.

Guillen, M., J. P. Nielsen, M. Ayuso, and A. M. Pérez-Marín. 2019. "The Use of Telematics Devices to Improve Automobile Insurance Rates." *Risk Analysis* 39 (3): 662–72. https://doi.org/10.1111/risa.13172.

Handel, P., I. Skog, J. Wahlstrom, F. Bonawiede, et al. 2014. "Insurance Telematics: Opportunities and Challenges with the Smartphone Solution." *IEEE Intelligent Transportation Systems Magazine* 6 (4): 57–70. https://doi.org/10.1109/MITS.2014.2343262.

Henckaerts, R., and K. Antonio. 2022. "The Added Value of Dynamically Updating Motor Insurance Prices with Telematics Collected Driving Behavior Data." *Insurance: Mathematics and Economics* 105:79–95. https://doi.org/10.1016/j.insmatheco.2022.03.011.

Huang, Y., and S. Meng. 2019. "Automobile Insurance Classification Ratemaking Based on Telematics Driving Data." *Decision Support Systems* 127:113156. https://doi.org/10.1016/j.dss.2019.113156.

Jeong, H. 2022. "Dimension Reduction Techniques for Summarized Telematics Data." *Journal of Risk Management* 33 (2022): 1–24. https://doi.org/10.2139/ssrn.4281943.

Peiris, H., H. Jeong, J.-K. Kim, and H. Lee. 2024. "Integration of Traditional and Telematics Data for Efficient Insurance Claims Prediction." *ASTIN Bulletin* 54 (2): 263–79. https://doi.org/10.1017/asb.2024.6.

Pesantez-Narvaez, J., M. Guillen, and M. Alcañiz. 2019. "Predicting Motor Insurance Claims Using Telematics Data—XGBoost Versus Logistic Regression." *Risks* 7 (2): 70. https://doi.org/10.3390/risks7020070.

Schelldorfer, J., and M. V. Wüthrich. 2019. "Nesting Classical Actuarial Models into Neural Networks." Preprint, SSRN. https://doi.org/10.2139/ssrn.3320525.

So, B., J.-P. Boucher, and E. A. Valdez. 2021a. "Cost-Sensitive Multi-Class AdaBoost for Understanding Driving Behavior Based on Telematics." *ASTIN Bulletin* 51 (3): 719–51. https://doi.org/10.1017/asb.2021.22.

———. 2021b. "Synthetic Dataset Generation of Driver Telematics." *Risks* 9 (4): 58. https://doi.org/10.3390/risks9040058.

Verbelen, R., K. Antonio, and G. Claeskens. 2018. "Unravelling the Predictive Power of Telematics Data in Car Insurance Pricing." *Journal of the Royal Statistical Society Series C: Applied Statistics* 67 (5): 1275–1304. https://doi.org/10.1111/rssc.12283.

APPENDIX A

SYNTHETIC TELEMATICS DATASET

In this appendix, we provide further details on the synthetic telematics dataset used in the empirical study. The original telematics dataset from So, Boucher, and Valdez (2021a, 2021b) is acquired from a Canadian-owned firm on its auto insurance policies and claims in Ontario from 2013 to 2016. To protect privacy and to comply with the use and sharing of proprietary data, So, Boucher, and Valdez (2021b) apply an algorithm, called "Extended SMOTE," to create a synthetic dataset with both traditional and telematics features and claims information based on the original data. Note that such a synthetic dataset preserves the key characteristics of the original dataset and is accessible to the public for research purposes. However, a noticeable drawback of the synthetic dataset in So, Boucher, and Valdez (2021b) is that it contains a total of 52 features (variables), and some categorical features have a large menu of categories (for instance, `Territory` has 55 categories). As such, this drawback may lead to an ineffective fit of GLMs to the dataset. To address this issue, Jeong (2022) proposes a dimension-reduction technique to process the dataset in So, Boucher, and Valdez (2021b) so that it can be trained to fit GLMs in a more robust way. We use the same processed dataset from Jeong (2022) in this paper.

The dataset consists of 100,000 observations and contains 11 traditional features and 10 telematics features; the response variable we consider is the number of observed claims. We present a detailed summary of all features included in the dataset in Table 4. As a remark, we mention that `TerritoryEmb` in Table 4 is an embedded variable from the `Territory` variable in So, Boucher, and Valdez (2021b), which is a categorical variable with 55 categories. Also, careful readers may notice that `Annual.miles.drive` is listed as a traditional variable in Table 4—that is because it is *not* reported by telematics devices or apps, but rather declared by policyholders. However, the `Total.miles.driven` variable is the actual mileage recorded by the telematics devices or apps.

Tables 5 and 6 show the empirical distributions of the response variable (`NB_Claim`) and the corresponding summary statistics for some of the covariates. One can clearly see that the dataset is quite unbalanced—most of the observed claim frequency is 0, and only a few policies have more than one claim per year, which is a prevalent situation in a personal auto insurance portfolio. We also observe that the response variable (`NB_Claim`) and the values of `Annual.pct.driven` and `Total.miles.driven` are positively associated as shown in their means grouped by `NB_Claim`.

Table 4. Description of the traditional and telematics features in the dataset

Type	Variable	Description
Traditional	`Duration`	Duration of the insurance coverage of a given policy, in days
	`Insured.age`	Age of insured driver, in years
	`Insured.sex`	Sex of insured driver (male/female)
	`Car.age`	Age of vehicle, in years
	`Marital`	Marital status (single/married)
	`Car.use`	Use of vehicle: private, commute, farmer, commercial
	`Credit.score`	Credit score of insured driver
	`Region`	Type of region where driver lives: rural, urban
	`Annual.miles.drive`	Annual miles expected to be driven declared by driver
	`Years.noclaims`	Number of years without any claims
	`TerritoryEmb`	Embedded value from the territorial location of vehicle
Telematics	`Annual.pct.driven`	Annualized percentage of time on the road
	`Pct.drive.xxx`	Percent of driving day xxx of the week: mon/tue/.../sun
	`Pct.drive.rush.am`	Percent of driving during a.m. rush hours
	`Pct.drive.rush.pm`	Percent of driving during p.m. rush hours
	`Avgdays.week`	Mean number of days used per week
	`Accel.06miles`	Number of sudden acceleration 6 mph/s per 1,000 miles
	`Brake.06miles`	Number of sudden brakes 6 mph/s per 1,000 miles
	`Acbr.others`	Total number of sudden acceleration and brakes 8/9/.../14 mph/s per 1,000 miles
	`Left.turns`	Number of left turns per 1,000 miles with intensity ≥ 8
	`Right.turns`	Number of right turns per 1,000 miles with intensity ≥ 8
Response	`NB_Claim`	Number of observed claims

Table 5. Empirical distribution of the response variable (NB Claim)

`NB_Claim`	0	1	2	3	Total
# of observations	95,728	4,061	200	11	100,000

Table 6. Summary statistics for selected covariates grouped by `NB Claim`

Variable Name	NB_Claim	Summary Statistics						
		Mean	SD	Min	Q1	Median	Q3	Max
Annual.pct.driven	0	0.49	0.30	0.00	0.24	0.48	0.72	1.00
	1	0.75	0.25	0.04	0.53	0.84	0.97	1.00
	2	0.76	0.25	0.18	0.51	0.87	0.99	1.00
	3	0.86	0.06	0.76	0.83	0.84	0.92	0.93
Total.miles.driven	0	4,659.30	4,422.08	0.10	1,472.60	3,322.24	6,502.19	47,282.60
	1	8,656.91	5,383.89	67.22	4,764.64	7,717.27	11,245.79	41,019.58
	2	10,047.12	6,015.31	1,192.16	5,152.69	9,601.15	13,107.07	30,415.65
	3	15,193.33	9,182.25	7,501.71	8,457.83	9,613.06	25,700.82	28,363.82
Avgdays.week	0	5.52	1.26	0.20	4.88	5.88	6.49	7.00
	1	5.82	0.89	0.96	5.40	6.02	6.48	7.00
	2	6.06	0.78	2.92	5.70	6.40	6.59	6.90
	3	5.88	0.82	4.92	5.06	6.43	6.65	6.70
Pct.drive.rush.am	0	0.10	0.08	0.00	0.04	0.08	0.14	0.99
	1	0.10	0.07	0.00	0.05	0.08	0.14	0.41
	2	0.10	0.06	0.00	0.04	0.09	0.14	0.28
	3	0.13	0.06	0.07	0.08	0.15	0.16	0.24
Pct.drive.rush.pm	0	0.14	0.07	0.00	0.09	0.13	0.17	0.99
	1	0.15	0.06	0.00	0.10	0.14	0.18	0.42
	2	0.17	0.08	0.02	0.13	0.15	0.19	0.44
	3	0.11	0.04	0.05	0.06	0.12	0.13	0.16
Insured.age	0	51.58	15.48	16.00	39.00	52.00	64.00	103.00
	1	46.82	14.46	18.00	35.00	47.00	57.00	90.00
	2	46.43	15.76	19.00	32.00	45.50	55.25	81.00
	3	56.09	13.09	38.00	47.00	51.00	68.50	71.00
Credit.score	0	802.28	82.75	422.00	769.00	826.00	856.00	900.00
	1	771.64	90.89	428.00	719.00	794.00	840.00	900.00
	2	732.35	91.43	511.00	662.75	751.00	806.00	900.00
	3	761.82	22.64	740.00	743.50	759.00	762.50	804.00

Variable Name	NB_Claim	Summary Statistics						
TerritoryEmb	0	-0.03	0.35	−8.82	−0.25	0.00	0.19	0.66
	1	0.01	0.28	-0.63	-0.24	0.03	0.19	0.66
	2	0.10	0.28	-0.63	-0.08	0.13	0.22	0.66
	3	0.04	0.25	-0.35	-0.20	0.19	0.22	0.22
Years.noclaims	0	29.07	16.13	0.00	15.00	29.00	42.00	79.00
	1	23.91	15.01	0.00	10.00	23.00	35.00	74.00
	2	20.75	14.52	0.00	7.75	19.00	33.00	64.00
	3	29.45	21.14	3.00	10.00	22.00	50.00	53.00