# **Integration of Traditional and Telematics data for E** cient Insurance Claims **Prediction**

 $\overleftarrow{\textbf{w}}$ 

**Hashan Peiris**

B.Sc., University of Sri Jayawardenepura, 2018



**© Hashan Peiris 2023 SIMON FRASER UNIVERSITY Spring 2023**

Copyright in this work is held by the author. Please ensure that any reproduction or real is done in a condition in a condition in a condition in a copyright legislation. The relevant legislation is  $\mathcal{L}(\mathcal{A})$ 

# <span id="page-1-0"></span>**Declaration of Committee**



### <span id="page-2-0"></span>**Abstract**

While driver telematics has gained attention for risk classifcation in auto insurance, the scarcity of observations with telematics features  $\mathbb{R}^n$  and could be observations  $\mathbb{R}^n$ to the either provides selection concerns or adverse selection compared to the data points with features. To handle this issue, we propose a data integration technique based on calibration weights. It is shown that the proposed technique can effect the so-called technique can effect the so-called traditional d and the and also contribution is and also coper with possible adverse selection is usually in the availability on the availability of  $\mathcal{A}$ of the telematics data. Our findings are supported by a simulation study and empirical and empi on a synthetic telematics dataset. **Keywords: Adverse selection; Automobile insurance; Automobile integration; Driver telematics; Driver telematics;**  $W \leq 1$ 

# **Dedication**

<span id="page-3-0"></span>For those who paved the path to achieving my dreams, especially my parents and family.

<span id="page-4-0"></span>**Acknowledgements**

# <span id="page-5-0"></span>**Table of Contents**



**[18.995 \(elematics\)-3322997 \(in\)-3P14 \(.](#page-16-0))-496 (Data)Descrip0.909-661TD** 85 (.)-500.004 (.)-



### <span id="page-7-0"></span>**List of Tables**

[Table 4.1 Summary of models](#page-29-1) . 21 [Table 4.2 Estimation performance with the simulated data](#page-30-0) . . . . . . . . . . . 22 [Table 4.3 Out-of-sample validation performance with the simulated data](#page-31-0) . . . . 23 [Table 5.1 Variable names and descriptions of the pre-processed dataset](#page-33-1) . . . . . 25  $T$ nd-sample validation performance with bootstrapping from the ac[tual data](#page-37-0) . 29

## <span id="page-8-0"></span>**List of Figures**



### <span id="page-9-0"></span>**Chapter 1**

### **Introduction**

 $I$ insurers have been modeling data to set prices for a utomobile insurance prices for automobile insurance prices for any  $\alpha$ decades. Since the insurance industry consists of competitors with various products with minimal diferences related to automobile insurance, fair and attractive pricing of products is benefcial for both the companies and the customers. Assessing the risk of new customers would here in market share while meeting the professional meeting the professional meeting the professional market margins as it enables in  $\mathbb{R}^n$  $\mathcal{R}=\mathcal{R}+\mathcal{R}+\mathcal{R}$  and policies. Over time the insurance industry has been time the insurance industry  $\mathcal{R}=\mathcal{R}+\mathcal{R}$  $u_{\mathcal{A}}$ upgraded with different products which use a vehicle about drivers and vehicles.  $\mathbb{C}$  data collection is based not only on the data collection is bills and bi  $\mathbf{v}$  also on devices that generate data about different measurements in  $\mathbf{v}$ telematics devices on vehicles. Thus, these diferent datasets are available for the operational  $\mathbf{y} = \mathbf{x}^T \mathbf$  $\mathbf{A}$ s analyze tend tend tend tend tenderated datasets with large datasets with large dimensions with large during their operations to set the prices of their products. Hence two types of datasets are recognized, traditional and telematics, based on the generation process, for the ratemaking.  $\mathbb{A}^{\mathcal{A}}$ nd also analyzing the risk of operations; effects of operations; effects of operations; effects of operations; effects of  $\mathbb{A}$  $c$ cases, evaluations, and make accurate accurate accuracy, and make accurate cost predictions. Therefore such predictions  $\mathbb{R}^n$ methods and insurance counts and the insurance products and the insurance products and insurance products and  $\mathcal{R}_\mathbf{X}$ data are discussed in this chapter.

<span id="page-9-1"></span>**1.1 Modeling Claim Counts**

Interesting the number of claims as assessing the risk in motor in motor insurance in motor insurance in motor insurance in motor i beneficial in premium calculation. There are different approaches for insurance rate  $\alpha$ and risk assessment in the actuarial literature. Among the actual literature. Among the actual literature. Amon  $\mathbb{C}_{\mathrm{cyl}}$  in the successor of classical linear models [\(Haberman and Ren](#page-45-0)[shaw](#page-45-0) [\(1996\)](#page-45-0)). One can interpret the efects of each variable using these models without being limited to provide the progression of studies shows the applicability of  $\mathbb{K}$  applicability of different shows the applicability of different shows the applicability of different shows the applicability of diffe linear modeling in insurance ratemaking as it is explained by a single as it is explained by a single brief of



 $\mathbf{A} \subset \mathbb{R}$  compact but computed in statistical learning approaches that are are as a computations of  $\mathbb{R}$ used in the actuarial literature is given in [Vermet](#page-46-0) [\(2018\)](#page-46-0). [Ferrario et al.](#page-44-1) [\(2020\)](#page-44-1) provides a complete guide in using neural network models to evaluate claim counts. Thus, it is observed that methods to model counts and counts are presented with a broad spectrum.

#### <span id="page-11-0"></span>**1.2 Telematics in Insurance**

 $\mathbb{R}^n$  now insurers have gained access to data that  $\mathbb{R}^n$  and driving habits and drivers  $\mathbb{R}^n$ haviours on roads. Telematics is a technology that can provide data that contain such addition since it generates data related to many variables characterized for the many variables characterized f each driver including but not limited to total miles driven, number of breaks or accelerations, and at what time the whole technological and the wheel. With the technological advancements in the tech  $\mathcal{R}$ the automobile industry with driver telematics service providers of  $\mathcal{R}$ mation to users for  $\mathbb{Z}$  and location, the location based information, customers  $\mathbb{Z}$ who use the attracted to telematics-based insurance products-based insurance products insurance products insur effective to the insurers can add new features to the insurers to the databases in addition to the databases i traditional features that can be used in classifications in classifications in a uniform prediction  $\mathcal{A}$ frame.

 $\mathbb{R}^n$  in the insurance (UBI) is an innovative product in the innovative product in the insurance in  $\mathbb{R}^n$ dustry based on the risk profile of a driver. The risk profile of a driver profile of a driver. The risk profile of a driver. The r [et al.](#page-46-1) [\(2017\)](#page-46-1) provides a comprehensive review of frequently used UBI methodologies, Payas-you-drive (PAYD) and Pay-how-you-drive (PHYD), show-you-drive (PHYD), showing the varieties of a p of the telematics in the rewarded based on the rewarded based on  $\mathcal{S}$ driving as it leads to lower premiums when the region of the respective respectively. The respectively in the respectivel insurance contract. Husnia et al. [\(2015\)](#page-45-1) states that UBI is a beneficial option for insurance companies to deal with revenue loss due to inconsistent pricing with individual risk with  $\alpha$ and competitive pricing conditions in the market as well as well as well as well as  $\mathcal{R} = \mathcal{R} = \mathcal$ Thus, the closely observe driving behaviours insurance driving behaviours and set insurance driving behaviours  $p$ remium rates. According to [Bolderdijk et al.](#page-43-2) [\(2011\)](#page-43-2), when a discount on premium is a discount on premium is appearance on premium is appearance of  $\mathcal{R}$ plied, the number of the definition of the definition of the definition of the contractes the contrac has been reduced by considering a distance-based insurance-based insurance-based insurance product similar to  $\mathcal{H}$  $\mathcal{X}_1$  and a lower premium when lower premium when  $\mathcal{X}_2$  is achieved. Moreover, Denvis et al.  $\lambda_0$  inducts as a mechanism as a mechanism that leads to effect risk evaluation and risk evaluation and risk evaluation and recognize  $\lambda_0$ acknowledges the though the safer three problems. The constant the constant relationship are problems are part ular UBI products the products can obliquely improve road safety. When  $\mathcal{U} = \mathcal{U}$ it is expected, UBI products called and expected,  $\mathcal{U}$  and environmental benefits. The new results of  $\mathcal{U}$ 

#### <span id="page-12-0"></span>**1.2.1 Uses of Telematics Data**

#### <span id="page-13-0"></span>**1.2.2 Challenges**

 $\mathbf{A}$ s it is mentioned to policyholders insurance companies  $\mathbf{A}$  insurance companies  $\mathbf{A}$ that contain that contains age, including but not limited but not limited to the driver's age, general age, ge vehicle characteristics. On the other hand, a telematics dataset can have a smaller number of data points with respect to the traditional dataset when the number of  $\mathcal{A}$ owns a police relation of the telematics is low. [Meyers and Hoyweghen](#page-45-3) [\(2020\)](#page-45-3) describes a study of the study of about a telematics insurance product that had failed to achieve the target number of par- $\mathbb{C}$  and discount of participation of premium for participation. The proposed method method method method of [Castignani et al.](#page-43-3) [\(2015\)](#page-43-3) tries to improve the availability of telematics data by using smartphone solutions as a remedy to the reluctance of drivers to install telematics devices which results in an initial cost. And Max et al. [\(2018\)](#page-45-4) mentions that the lack of the lack of telematics that data availability is a challenge in identifying the factors of policyholder behaviour with  $\mathcal{A}$ the actual pricing methods incorporated with  $\mathcal{X}_\mathbf{u} = \mathcal{X}_\mathbf{u}$  $\mathbb{A}^{\mathcal{A}}$  et al. [\(2015\)](#page-45-1) states that a method to maintain the  $\mathcal{A}$ can be questionable as the collection process can be interrupted due to technical or environmental in particles in particles and  $G$ uillen et al. [\(2021\)](#page-45-5) uses a modeling approach for insurance approach for  $\mathcal{R}=\mathcal{R}$  and the telematics data but is limited to a small number of  $\mathcal{R}$ features as available data is limited. In terms of risk classifcation, [Tuna and Cengiz](#page-46-2) [\(2021\)](#page-46-2) recommends using telematics data that are observed for a predetermined short period. Furthermore, they have mentioned that collecting less telematics information helps to meet the recommendations of insurance regulators.

Indeed and consideration and consideration of the collection of the collection of the relatively recent and the are still ongoing concerns about privacy is about privacy is a  $\mathcal{R}^{\mathcal{A}}$  many policyholders relations rel agree on the provision of the insurers. [Duri](#page-44-4)ng the insurers data to the insurers data to the insurers data to [et al.](#page-44-4) [\(2004\)](#page-44-4) state that the success of the success of the success of the awareness of  $s$ service providers that the data is received accurately which accurately which are provided users. The privacy of end users,  $\frac{1}{2}$  $\mathbf{A}^\mathcal{A}$  also also [Milanović et al.](#page-45-6) [\(2020\)](#page-45-6) states and idea in terms of telematics acceptance of telematics acceptan technology which is shown to be dependent on privacy concerns. Similarly, [Buxbaum](#page-43-4) [\(2006\)](#page-43-4)  $\epsilon$  $S_{\rm t}$ states the police telematics will the provide telematics data tend to have less concern about privacy is subsequently in the four perspective,  $\mathcal{A}_{\mathbf{u}} = \mathcal{A}_{\mathbf{u}}$ telematics data while recognizing the privacy and suggesting and security is and suggestion  $\mathcal{X}^{\mathcal{X}}$  $\mathcal{R} \subset \mathcal{R}$  and the proposed at the proposed a framework to propose that is based as framework to protect data that is based as  $\mathcal{R}$  $\mathbf{c}_p = \mathbf{p}$  and security technology. But the study of [Eling and Kraft](#page-44-5) [\(2020\)](#page-44-5) which provides a compact description of the insurability of risk using telematics data, also highlights some recommendations from the literature that can improve the number of telematics-based  $p_a = 4k$ 

 $\mathbf{A}^{\text{max}} = \mathbf{A}^{\text{max}} = \mathbf{A}^{\text{max}}$ telematics data may depend more on the type of  $\mathcal{X}$  of  $\mathcal{X$ 

 $\mathcal{S} = \mathcal{S} \mathcal{S} = \mathcal{S} \mathcal{S} = \mathcal{S} \mathcal{S} = \mathcal{S} \mathcal{S} = \mathcal{S} \mathcal{S}$ [et al.](#page-44-6) [\(2022\)](#page-44-6) mentions that the attraction of safer of the insurer of safer drivers is beneficial for the insurer as insure  $\mathbb{C}^d$  in conclusion, all the fact that conclusion, all those studies have raised the fact that conclusion, and  $\mathbb{C}^d$ for the action of the above to the area being monitored. Although  $\mathcal{A}$ it is an existent for the insured, and society, the insure  $\mathcal{A}$  insured, and society, the insured, and in missing some insights about more risky drivers in terms of an analytical point of view. But using the explanation in the study of [Cohen and Siegelman](#page-44-7) [\(2010\)](#page-44-7) about adverse selection in the selection in the selection  $\mathcal{L}_\text{c}$ in insurance, one can think of a situation where a buyer knows that he can have a lower premium by achieving a lower claim count as a case weekly related to the adverse selection.  $T$ Thus we can consider the aforements as challenges to having a telematics as challenges to having a telematics a telematics and the second situations are challenges to having a telematics and the second situations are c dataset as informative as the fnite population of drivers.

#### <span id="page-14-0"></span>**1.3 Motivation**



- estimator. Thus, here we focus on using the propensity score estimation method proposed by Wang and  $\mathcal{M}$  which can be used as a uniform can be used as a uniform combining information  $\mathcal{M}$ from multiple datasets with the consideration of possible non-ignorable sampling bias. This method is also similar to the data integration where in [Wang et al.](#page-46-4) [\(2022\)](#page-46-4) where the data in Wang et al. (2022) data integration is defined by regression in  $\mathcal{A}_{\text{max}}$ and  $\delta$  incorporate information from different data sources. It suggests a model of  $\delta$ calibration technique to get partial information from external data sources (traditional data sources (traditional d dataset in our case) and integrate the data sources at once.
- <span id="page-15-0"></span>**1.4 Summary**
- In the following objectives of the following objectives of the study are recognized, which will be described as appropriately in later chapters. In later chapters,  $\mathcal{R}_1$ 
	- $\bullet$  Beautiful that a new few feed for a few feed for a few feed for deal with multiple sources of data few feed for  $\bullet$ in insurance ratemaking due to the scarcity of telematics data compared to the tradi-

<span id="page-16-0"></span>**Chapter 2**

# <span id="page-16-1"></span>**Data structure and problem description**

- **x***i*<sup>1</sup> be the traditional features of the *i th* policyholder. Those are available in both <sup>0</sup>  $\mathbf{a}_{1}$
- $\bullet$   $\frac{1}{2}$  be the box  $\frac{1}{2}$  be the *i*<sup>2</sup> be  $\frac{1}{2}$  be
- $\mathbf{X} = \mathbf{X} \times \mathbf{M}$

<span id="page-17-0"></span> $\mathbf{A}^\bullet$  summary of description of description of data is given in Figure [2.1](#page-17-0) which depicts the asymmetric department of  $\mathbf{A}^\bullet$  $\mathcal{S}$ structure of the full dataset in reality. The full data set in reality of the full data set in reality.



**2.2 Problem Description**



unless the provision of the provision of provision of provision outweight the possible concerns. The possible concerns of the provision  $\mathcal{S}$ one can think about the following conjectures:

- $\bullet$  Those who are  $\mathbb{R}$  are  $\mathbb{R}$  and they are less  $\mathcal{R}$ reluctant to the sampling probability of an observation in  $\mathcal{R}$  $\alpha_0$  is inversely proportion is inverse in the driver  $\alpha_0$
- $\bullet$  Those who are less risky tend to a great tend to provide tend to provide tend the tensor tend to  $\sim$ accessibility of **x***i*<sup>2</sup> is prone to adverse selection, which implies the sampling probability of an observation in  $\mathcal{A}$  from the population is in  $\mathcal{A}$  from the number  $p$  $\bullet$  *ni*

While our main task is not to detect possible selection biases in the availability of the availability of the a<br>The availability of the availa telematics features and prove such conjectures, we conjecture the situations where such conjectures where such conjectures, we consider the such conjectures, we conjecture such conjectures, where such consider the situati jectures do hold and discuss the benefts of the proposed method compared to pre-existing benchmark methods in various situations.

<span id="page-19-0"></span>**Chapter 3**

**Methodology**



In general, we are interested in estimating in the regression model ( *<sup>i</sup>* | **x**) = *m*(**x** ) where *m*(·) is a known function and is an unknown parameter while *<sup>i</sup>* is the observed number of claims for policyholder *i* The census estimating equation for can be written as *M i*=1 *n<sup>i</sup>* − *m*(**x***<sup>i</sup>* )}*h*(**x***<sup>i</sup>* ) = 0

<span id="page-20-0"></span>for some *h*(·)



**3.2 Proposed Method**

Keeping the GLM framework with the assumption on the distribution of *<sup>i</sup>* , we now adopt the single study of  $\mathcal{A}_{\text{tr,2}}$  and  $\mathcal{A}_{\text{tr,3}}$  in the data structure structure structure structure structure  $i$  in  $\mathcal{L}$ 

**3.2.1 Estimation of Parameters**

<span id="page-21-0"></span>
$$
x_{k+1} = (x_{k+1} + x_{k+2}) x_{k+2} x_{k+3} x_{k+4} x_{k+5} x_{k+6} x_{k+7} x_{k+8} x_{k+8} x_{k+1} x_{k+1} x_{k+2} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+5} x_{k+6} x_{k+7} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+5} x_{k+6} x_{k+7} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+5} x_{k+6} x_{k+7} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+2} x_{k+3} x_{k+4} x_{k+5} x_{k+6} x_{k+7} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+6} x_{k+7} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+5} x_{k+6} x_{k+7} x_{k+8} x_{k+1} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+6} x_{k+7} x_{k+8} x_{k+1} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+6} x_{k+7} x_{k+8} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+6} x_{k+7} x_{k+8} x_{k+1} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+6} x_{k+1} x_{k+1} x_{k+2} x_{k+3} x_{k+4} x_{k+6} x_{k+1} x_{k+2} x_{k+3} x_{k+2} x_{k+1} x_{k+2} x_{k+3} x_{k+2} x_{k+2} x_{k+3} x_{k+2} x_{k+4} x_{k+3} x_{k+2} x_{k+3} x_{k+2} x_{k+3} x_{k+2} x_{k+4} x_{k+2} x_{k+3} x_{k+2} x_{k+3} x_{k+2} x_{k+4} x_{k+2} x_{k+3} x_{k+2} x_{k+4} x_{k+2} x_{k+3} x_{k+2}
$$

 $N_{\rm eff}$  and  $N_{\rm eff}$  functions in [\(3.4\)](#page-21-0), we impose

<span id="page-22-0"></span>
$$
i S_0
$$
  

$$
i I \rightarrow 1 i \cdots I
$$
  

$$
i J_0
$$
  

$$
i J_1
$$
  

$$
i J_2
$$
  

$$
i J_1
$$
  

$$
i J_2
$$

as a constraint for  $\mu \ll \sigma$   $\mu \ll \gamma$  *ii*  $\mu \Rightarrow \mu(\mu_1 \land \mu_2)$  is the different point  $s_n = s_n$   $s_n$   $[x_{1i}, \ldots, x_{Li}] = [1, x_{i1}, \ldots, x_{i1}]$ ,  $s_n$   $s_n$   $s_n$   $s_n$   $s_n$   $s_{n+1}$  where is the number of features in *xi*1. Constraint [\(3.5\)](#page-22-0) is often called the covariate-balancing  $\mathbb{R}^n$ uich, 2014) or calibration property ( $\mathbb{R}^n$  and  $\mathbb{R}^n$  and  $\mathbb{R}^n$  $\mathbb{R}^n$  is as as  $\mathbb{R}^n$  is satisfactory is satisfactory in the canonical structure  $\mathbb{R}^n$ 

$$
i S_0
$$
\n
$$
i ( ; , i * i) = \begin{array}{c} M \\ j & j ( ; , i * j) + (1 - j i) \end{array} \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text
$$

$$
\mathbf{w} = (\mathbf{a}_0 \mathbf{a}_1 \cdots \mathbf{a}_L) \mathbf{P} \qquad \mathbf{w} = (\mathbf{a}_1 \mathbf{a}_1 \cdots \mathbf{a}_L) \mathbf{P}
$$
\n
$$
\mathbf{w} = \mathbf{a}_1 \mathbf{a}_2 \cdots \mathbf{a}_L \mathbf{a}_L \mathbf{P}
$$
\n
$$
\mathbf{w} = (\mathbf{a}_1 \mathbf{a}_1 \cdots \mathbf{a}_L) \mathbf{P} \qquad \mathbf{w} = \mathbf{a}_2 \mathbf{a}_2 \mathbf{a}_2 \mathbf{a}_3 \mathbf{P}
$$

where  $\mathcal{C}$ 

$$
i \quad ( \quad ; \quad i \cdot s_i) = \frac{M}{i-1} \qquad i \quad ( \quad ; \quad i \cdot s_i) + (1)
$$

<span id="page-23-0"></span>Once *ϕ*<sup>0</sup> · · · *, ϕ<sup>L</sup>* are estimated by [\(3.5\)](#page-22-0), we can use  $\hat{i}$  i = 1 +  $\frac{1}{0}$ exp *ϕ* ^0 + *ø* ^{1} 1*i* + · · · + *ø* ^{1} L*i*  $\mathbf{x} = \mathbf{x} \sqrt{\mathbf{x}} \mathbf{z} \quad \mathbf{x} = \mathbf{z} \sqrt{\mathbf{x}} \mathbf{z}$ *i*∈S *i*  $\hat{i}(x) = (x^2, y^2, y^2) = 0$  $\mathcal{R}^{\mathcal{R}} = \pm \theta \mathcal{H}$  is the propension of propensity the propensity  $\mathcal{R}^{\mathcal{R}}$  the propensity the propensity of  $\mathcal{R}^{\mathcal{R}}$ calibration equation in [\(3.5\)](#page-22-0), this equation satisfes the self-efciency without estimating  $\mathcal{R} \setminus \mathcal{R}$  in the regression  $\mathcal{R}$  in the regression model  $\mathcal{R}$  $\left( \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right)$   $\left( \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right)$   $\left( \begin{array}{ccc} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right)$ *L k*=1 *αkbl*(**x***i*1*, ni*)*.*

#### **3.2.2 Standard Errors of Estimates**

We are also interested in the standard errors of the standard errors of the errors of the estimates. The calcu  $\mathcal{R}$  and standard extra is done according to  $\mathcal{R}$  and  $\mathcal{R}$ described by [Wang and Kim](#page-46-3) [\(2021a\)](#page-46-3), the Richard and Kim (2021a), the PS of the PS  $\mathbf{A}$ model (with parameter is the other is the regression of the regression of the regression of the regression of

 $\mathbb{W}_{\mathbf{u}}$  constructions functions functions for extensions for extensions for  $\mathbb{W}_{\mathbf{u}}$  and  $\mathbb{W}_{\mathbf{u}}$  $\mathbf{I} = \mathbb{I}(\ell \in \mathbf{0})$ 

$$
\hat{i}_1(\cdot) = \frac{i}{s} \sum_{j=1}^{s} \hat{j}_j(\cdot) - 1 \sum_{j=1}^{s} \hat{j}_j
$$
\n
$$
\hat{i}_2(\cdot) = \frac{i}{s} \sum_{j=1}^{s} \hat{j}_j(\cdot) (\cdot; i \cdot \hat{i}_j)
$$

 $\mathbf{W} = (1 \, 1 \, 1 \, 1 \, 1 \, 1 \, 1 \, 1)$  and  $i_j$ ( ) = 1 +  $\frac{1}{0} \exp(\phi)$  0  $\#$   $j_1$   $j_1$  +  $\cdots$   $\#$   $j_l$   $j_l$   $k_l$  $T$  is the solution to the joint estimation to the  $\frac{1}{2}$  $\hat{a}_1(\ )=0$  **d**  $\hat{a}_2(\ )=0.$ 

 $\mathbf{W}_{\mathbf{W}^*}$   $\mathbf{X} = ( )$  define

$$
f(x) = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix}
$$
  
\n
$$
f(x) = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix}
$$
  
\n
$$
f(x) = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{-1} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1} = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1}
$$
  
\n
$$
f(x) = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1} = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1} = \frac{1}{2} \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-1}
$$

$$
\tilde{\gamma} = \frac{1}{\gamma - 1} \left( \gamma \right) \qquad \mathbf{d} \quad \tilde{\gamma} \left( \tilde{\gamma} \right) = \frac{1}{\gamma - 1} \left( \tilde{\gamma} \right) \left( \tilde{\gamma} \right) - \frac{1}{\gamma - 1} \mathbf{d}
$$

$$
R = \frac{1}{2} \sum_{i=1}^{N} \frac{1}{i} \sum_{i=1}^{N} \frac{1}{i} \left( \frac{1}{i} \sum_{i=1}^{N} R_i \cdot R_i \right) \cdot R_i R_i \quad \text{and} \quad R_i = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{i} \sum_{i=1}^{N} \frac{1}{i
$$

$$
\tilde{a}_j = \begin{cases} i & \text{if } i \neq j \text{ and } j \neq j \text{ and }
$$

### **3.3 Estimation scheme**

Now, the estimation scheme for the study is listed orderly according to the requirement of the requirement of  $\mathcal{R}$  $\mathcal{R}^{\mathcal{R}}_{\mathcal{R}}=\mathcal{R}^{\mathcal{R}}_{\mathcal{R}}=\mathcal{R}^{\mathcal{R}}_{\mathcal{R}}=\mathcal{R}^{\mathcal{R}}_{\mathcal{R}}$ 

$$
\mathbb{R} \mathcal{H} = \mathcal{H} \mathcal{H} = \mathcal{H} \mathcal{H} \mathcal{H} \mathcal{H} \mathcal{H} \mathcal{H} \mathcal{H} \mathcal{H}
$$

<span id="page-25-1"></span>**Chapter 4**

**Simulation study**

<span id="page-25-0"></span>

 $N = N$  in  $N = N$  for all individuals to  $N = N$  for all individuals to  $N = N$  for all individuals for a simplification for a simplification for all individuals for  $N = N$  for all individuals for all individuals for a simplifica  $\mathbb{Z}^2$  which results in a model with  $\mathbb{Z}^2$ Furthermore, we recognize the variables as below.

• *xAi* refers to a traditional continuous variable with quadratic efect (e.g., driver's age)

• *xGi* refers to a traditional binary variable (e.g., gender)

 $\bullet$   $x$ *T* i<sup>n</sup>  $\infty$  in the risk profile of significant impact on the risk profile  $\infty$ 

<span id="page-26-0"></span> $N$  Note that the driver as an example of the  $\alpha$  $\mathcal{R}_\text{c}$  that  $\mathcal{R}_\text{c}$  denote the risky due to the risky d aftected by other biological facts similar to  $\mathcal{A}$  for  $\mathcal{A}$  facts similar to  $G$ uillen et al. [\(2019\)](#page-45-9). Here, each feature contains 100,000 data points. The generated for the generation for generation for generatio population in accordance with the notation used in Section [2.1.](#page-16-1)

- **Step 3: The Remaining areas of databaset as a large data points are used as a large data points are used as a l** features *<sup>i</sup>* **x***i*1}, which is equivalent to <sup>1</sup> in Section [2.1.](#page-16-1) These data are all different models. Here, we consider the following models. Here, we consider the following  $\alpha$ models to estimate  $1$  and  $2$  where  $N$  and  $2$  where  $N$  appropriate partitions of the generation,  $2$  where  $N$ that are used in  $\mathcal{A}$  are indicated in  $\mathcal{A}$  . The indicated in Figure • Naive model  $\mathbb{F}^{\mathbf{b} \mathbf{b}}$  **L**<br>**Fig. 1** • **Traditional model** Fit a use a use a using only the traditional model Fit a set of the traditional features and the traditional model of  $\mathbb{R}^n$  $\mathbf{x} \times \mathbf{x} = \mathbf{x} - \mathbf{i} \times \mathbf{n}$ *V*  $\rightarrow$   $\theta$  +  $\mathbf{x}$  as  $\mathbf{x}$  as  $\mathbf{x}$  to  $\mathbf{x}$ the telematics information at all in the risk classifcation. • Full model: It it's all the data points in − <u>P</u>oints in Section coefficients in Section coefficients in Section co cients of a usual point it is expected to provide the best estimation of  $\mathcal{R} \times \mathcal{R}$  estimation best estimation of  $\mathcal{R} \times \mathcal{R}$ 
	- performance. Note that might not be available in practice.
	- **Boosting model**  $\mathbf{I} \cdot \mathbf{I}$  **is used in the same of**  $\mathbf{I} \cdot \mathbf{I}$  **is a set of**  $\mathbf{I} \cdot \mathbf{I}$  **and**  $\mathbf{I} \cdot \mathbf{I}$  **and \mathbf{I} \**  $\therefore$  **i**  $\mathbf{N}^*$ *i*  $\eta \hat{i} = \exp(i \hat{i} \hat{i})$  ii.  $\mathbf{N}$  is  $\mathbf{N}$  in  $\mathbf{N}$  in  $\mathbf{N}$  in  $\mathbf{N}$ GLM is ftted with <sup>0</sup> where the telematics information, **x***i*2, is the only regressor and  $\log^{N}$ **ind is me**ntioned as  $\log^{N}$  is  $\log^{N}$  is mentioned as the boosting  $\log$  $\mathbf{a}$

• Proposed model I this the estimation proposed model I  $\mathcal{R}=\mathcal{R}$ errors descri

After all models are function are function are for the regression estimates from the models were used to find the following to find the models were used to find the following to find the models were used to find the follo the predictive value ˆ *<sup>i</sup>* for *i th* policyholder in the out-of-sample validation set T . Since these models use the generated data different lying and prediction  $\mathcal{A}$  summary of each prediction  $\mathcal{A}$ model is given in Table [4.1.](#page-29-1)

#### <span id="page-27-0"></span>**4.3 Evaluation Procedure**

 $N=3$ nite that generation of the formulation of the formulation ( ), and prediction, and pred are repeated that different random seeds. This contribution is to evaluate the random seeds us to evaluate us to evaluate the random seeds. This expedition is to evaluate the random seeds. This expedition is to evaluate t and compare the methods computation  $\mathcal{A}$  $u \geq 0$ ias, root mean-squared error (RMSE) and 90% confidence interval coverage (CI) of  $\mathbb{R}$ *<sup>j</sup>* . Let these statistics be defned as follows:  $\mathbf{a}$  =  $\frac{1}{2}$   $\mathbf{b}$ *r*=1  $(j - \hat{i}^{(r)}_i)$ *j* )  $\mathbf{H} = \frac{1}{2}$ *r*=1  $(y - \hat{f}^{(r)})$ *j* ) 2

<span id="page-28-0"></span>

 $F: \mathbb{R} \longrightarrow \mathbb{R}$  at data structure considered in each the model in each the

 $\mathbf{l}_j = \frac{1}{n}$ *r*=1  $\frac{1}{2}$  {/ *j* − ^(r) /<1*.* SE( ^)(r) }  $\mathcal{R}$   $\left(\begin{matrix}r\\l\end{matrix}\right)$  $j$  is the  $j$  for  $k$  for  $f$  and  $j$   $j$ *j* is the estimated standard  $\begin{matrix} \wedge(r) \\ \vdots \end{matrix}$ *j* After the entity of each model is assessed, we use the out-of-sample  $\mathcal{A} \longrightarrow \mathcal{A}$  *r*  $\mathcal{A} \longrightarrow \mathcal{A}$  *= 1* 

<span id="page-29-1"></span>

<span id="page-29-0"></span>**4.4 Results**

As a summary, Table [4.2](#page-30-0) shows estimation results of the regression coefcients under different model specifcations and sampling schemes. Here **N T B F**, and **P** refer to Naive,  $T$ aditional, Boosting, Proposed models, respectively. Functions,  $\mathbb{E}[T]$ It is computation in the sampling mechanism of  $\mathcal{X}$  is purely random, then the  $\mathcal{X}$ use of the naive model is in the national intervals in the full state  $\mathcal{U}$ model shows the best performance in the estimation performance followed by the proposed model, the boosting model (and correspondingly the traditional model) sufers from the  $\infty$   $\infty$   $\infty$   $\infty$  $T$  **T**  $\mathcal{B}$  . But the naive model is less than  $\mathcal{B}$  is the full and proposed to the full and proposed is  $T$ models, by having a larger RMSE of the estimated regression coefcients.  $A$ when the sample mechanism is a generator is age selection, it is shown that the nation, it is shown that the national is in the nation of the national is in the nation of the nation of the nation of the nation of the n still unbiased and less efcient compared to the full and proposed models. Beyond the changes of bias of the naive model, one can hardly observe a noticeable difference in the noticeable difference performance of models within both random and age selection results. On the other hand, if the sample mechanism of  $\mathcal{N}$  and  $\mathcal{N}$  is the difference in the dif estimation performance are more dramatic. Unlike the random sampling case, naive model  $s$ severely sufficient and biases in the estimates, since  $\mathcal{S}_\mathbf{r}$  in the estimates, since  $\mathcal{S}_\mathbf{r}$ sample of the fnite population anymore.



<span id="page-30-0"></span>

<span id="page-31-0"></span>

and Right.turns were were selected to represent the four for four clusters out of for four features for four f each cluster) while all the variables in the remaining cluster were used without representa- $\sim$ After the data pre-processing, the following retained the following variables which are used to in the contract of the described in Table [5.1.](#page-33-1) Hence the traction of the contract of the dataset with reduced d For more details about the data pre-processing, see [Jeong](#page-45-10) [\(2022\)](#page-45-10).

<span id="page-33-1"></span>

#### <span id="page-33-0"></span>**5.2 Estimation and Evaluation**



traditional and telematics features, but a relatively small number of observations, which incorporates the information in the traditional dataset.

<span id="page-34-0"></span>**5.2.1 Estimation**

 $\mathbb{E}_{\mathbf{z}} \times \mathbb{X}$  treat the processes as the formulation and use bootstrap samples bootstr for the analysis to ensure each observation has the same empirical distribution as the fnite population. Thus, we take a bootstrap sample, T , of size 100,000 for out-of-sample validation at  $\mathbf{a}$  is taken bootstrap sample of size 100,000 is taken as  $\mathbf{a}$ probabilities that are listed in Chapter [4.](#page-25-1) After that, a bootstrap sample of size 800,000 is taken as  $B \times S$  subjects to the complex probabilities of  $B$   $\mathcal{L}$  $\frac{1}{1}$  is the boundary elimination by eliminating telematics features from  $B$  $\mathbb{F}^N$  and convenience of convenience of calculations, we use sampling schemes with a slight change as  $\mathbb{F}^N$ 

• **Random selection**: The data points as a data points as a data point at random.

• Age selection:  $\mathbb{R}^d$  and  $\mathbb{R}^d$  assigned to  $\mathbb{R}^d$  assigned to  $\mathbb{R}^d$  $1/(1 + \exp(0.031\,\textnormal{nsured.\,age}_i))$ 

• Adverse selection: Each data point assigned to 2 is chosen with the sampling assigned to 0 is chosen with the sampling of th  $p_{\mathbf{v}}$  and  $1/(1 + \exp(NB_C \text{C} \text{I} \text{a} \text{I} \text{m}_i))$ 

 $\mathcal{L}_\mathcal{A}$  and the process of ftetting and testing and testing the process of  $\mathcal{L}_\mathcal{A}$  for  $\mathcal{L$  $= 500$  to compare the estimation and predictive performance performance performance performance performance performance  $\alpha$ method. Note that all data points in the coefficients in the full second to estimate the coefficients in the full second to estimate the model.

<span id="page-34-1"></span>**5.2.2 Evaluation**

To assess the in-sample estimation performance, we compare the estimated regression coefcients from each method and sampling scheme with the estimated regression coefcients obthat the finite population as in  $\mathcal{A}=\mathcal{A}$  as in  $\mathcal{A}=\mathcal{A}$  , root mean-squared me error (RMSE), and 90% confidence interval confidence interval confidence interval confidence interval confidence in defned as follows:  $\mathbf{a}$  =  $\frac{1}{2}$   $\mathbf{b}$ *r*=1 *N* The Sole of The Contract of Times and the Contract of Times a  *(cmeTf* 

Implications of the in-sample estimation results, under each sampling mechanism, with  $\omega$  $x \nvert x$  are  $x$ 

- $\bullet$  in the case of random selection, only the biases of  $\mathcal{R}$  model sufficient mo the regression coefficients and the estimation performance in the estim mance between the national proposed models. It is in a long as  $\mathbb{R}^n$  in as long as the sampling as the sample of  $\mathbb{R}^n$ mechanism of  $\mathcal{A}$  and  $\mathcal{A}$  and the traditional and telematics features  $\mathcal{A}$ the finite population is purely random, one can ignore  $\mathcal{U}$  (a large dataset only with  $\mathcal{U}$  $\mathbb{R}^d$  different purposes of  $\mathbb{R}^d$  of  $\mathbb{R}^d$  of  $\mathbb{R}^d$  by  $\mathbb{R}^d$  ,  $\mathbb{R}^d$
- $\bullet$   $\mathbb{I}$  in the case of age selection, the naive model is model in the estimation of  $\mathbb{I}$ the the traditional covariates (especially the intercept term), compared to the proposed term  $\mathcal{R}$ model. It is in the observable that in the observable telematics features on the telematics features depends o traditional features, the proposed approach might be helpful to be helpful to be helpful to be helpful to be h  $\mathcal{R}^\bullet$ the underlying impacts of the covariates on the covariate on the co
- $\bullet$  Lastly, in the case of adverse selection, the proposed model is no model is no model is no more unbiased but  $\mathcal{A}$ the naive model is still more biased in the estimation of the regression coefcients.  $T$ hens, if the accessibility of the telematics features is affected by a feature selection, and a verse selection,  $\alpha$ then it is recorded to the missingle transition of the missingle transition  $\mathcal{X} = \mathcal{X}$ telematics features.

 $\mathbf{N}$ ud differences are also visualized in Figures [5.1,](#page-38-0) 3.1, 3.2, and [5.3.](#page-40-0) It is consistent lying to  $\mathbf{N}$ that in the case of either age or adverse selection, the proposed model is the proposed model is the second-best,  $\alpha$ following the full model that is unattained in practice. Furthermore, it is visible that is v the proposed model is performed than the boosting model despite than the boosting model despite the sampling model of the sample of the s mechanism. The clear winds and  $\mathbb{R}^n$  values of the computation of the comparison of the co  $r_{\rm B}$ ranges of estimates, the boosting model tends to have model tends to have  $r_{\rm B}$ other models.

 $T^{\mathcal{A}}$  discuss the out-of-sample value  $T$  and  $\mathcal{A}$  which shows that shows the sh the proposed model is the only comparable model is the only comparable model in terms of proposed in terms of p and DEV on average, especially when the observability of telematics features is prone to adverse selection. It is also observed that the national, and boosting models of the national, and boosting mod not outperform the proposed model in most of the bootstrap samples as shown in the values of Prop\_pRMSE and Prop\_DEV, regardless of the selection scheme. Therefore, the proposed approach is a reasonable alternative in the absence of a fnite population with both traditional and telematics features or the sample with both types of features is not representative. And also it depicts that the out-of-sample performance of the traditional model does not change with sampling sampling schemes. It is a sign of the information of the information  $\mathcal{L}$ that telematics data care about the risk of a driver.

<span id="page-37-0"></span>

Furthermore, Figure [5.4](#page-41-0) showcases the distributions of proportional improvements of pRMSE and DEV. It is shown that while the proportional improvements of pRMSE are symmetric and centered at 0 with the random or age selection, the proportional improve- $\mathcal{R}=\mathcal{B}=\mathcal{R}_{\text{tot}}/\mathcal{R}_{\text{tot}}$  and all  $\mathcal{R}_{\text{tot}}$  centered at  $\mathcal{B}=\mathcal{B}$ the adverse selection which also supports the usefulness of the proposed method upon the propose existence of adverse selection in the provision of telematics features.

<span id="page-38-0"></span>

<span id="page-39-0"></span>

<span id="page-40-0"></span>





 $\mathfrak{z}$ 



<span id="page-41-0"></span>Drancead ve Naivo din norportaged from Dandom Cologicae ... Drancead ve Naivo din norportoral from Dandom Cologican

## <span id="page-43-0"></span>**Bibliography**

<span id="page-43-5"></span>**A. Agreement Categorical data analysis**  $\mathcal{A}_{\mathcal{N}}$  $M_{\rm eff}$  and  $M_{\rm eff}$  and  $M_{\rm eff}$  and  $M_{\rm eff}$  and  $M_{\rm eff}$  and distance to fraction  $T$ driving patterns of young drivers with pay-as-you-drivers and *allysis and Prevention*  $\mathbf{F}$   $\mathbf{J}$   $\mathbf{A}$  $M_{\rm eff}$  M. A. M.  $M_{\rm eff}$  in  $M_{\rm eff}$  and  $M_{\rm eff}$  and  $M_{\rm eff}$ usage-based evidence on whether men's risks of accedents differs accedents accedents accedents and women's. Als (2):10, 2016.  $M_{\rm eff}$  and  $M_{\rm eff}$  and  $M_{\rm eff}$  insurance ratemaking and  $M_{\rm eff}$ telematics: incorporation: incorporation driver behaviour data. Transportation driver behaviour data. Transportation  $\mathbf{r} \in \mathbf{A}_1$ P. Baecke and L. Bocca. The value of vehicle telematics data in insurance risk selection procession Support Systems  $\mathbf{a}^{\prime}$ <sup>2</sup>+  $\mathbf{a}$ <sub>1</sub> $\mathbf{a}$  $\mathcal{R}$  and  $\mathcal{R}$  insurance  $\mathcal{R}$  in  $\mathcal{R}$  in  $\mathcal{R}$  insurance rate  $\mathcal{R}$ *Insurance: Mathematics and Economics*,  $\rightarrow$ <sup>3</sup><sup>5</sup> *IT*(2)  $\rightarrow$  $J_{\mathbf{w}}$  A. A. Bolder, A. Steg, and E. Steg, and E. T. Verhoef. E. Steg, and E. T. Verhoef. Efects of pay-as-you-drive vehicle insurance on young drivers' speed choice: Results of a dutch feld experiment. *Accident Analysis & Prevention*, 2346, 2011.  $\mathbf{A}$ -P. Boucher and M. Denvis random effects in Poisson random effects in Poisson and  $\mathbf{A}$ claim countries in the study with motor insurance. **ASTIN Bulletin**: The Journal of the *IAA*, *1*, *1, 2*6(1):2006.

<span id="page-43-4"></span><span id="page-43-3"></span><span id="page-43-2"></span><span id="page-43-1"></span> $\mathcal{P}$ 

<span id="page-45-10"></span><span id="page-45-9"></span><span id="page-45-8"></span><span id="page-45-7"></span><span id="page-45-5"></span><span id="page-45-1"></span><span id="page-45-0"></span> $\mathcal{L} \cap \mathcal{L} \subset \mathcal{L}$  and Associated Marin. The use of telematics devices of telematics devices device devices device devices of the use of telematics devices device devices of the use of telematics devices of the use to improve automobile insurance rates. *Risk analysis*, 39(3):662–672, 2019. M. Guillen, J. P. Nielsen, and A. M. Pérez-Marín. Near-miss telematics in motor insurance. *Journal of RTsk and Insurance*  $\mathbf{B}$   $\mathbf{B}$   $\mathbf{F}$  of  $\mathbf{B}$   $\mathbf{F}$  and  $\mathbf{F}$ S. Haberman and A. E. Renshaw. Generalized linear models and actuarial science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(4):407–436, 1996.  $\mathbf{A}$ . I. Heras, I. Moreno, A. Moreno, A. L. L. J. L. Vilares,  $\mathbf{A}$ to insurance *insurance ratemaking. Scandinavian Actuarial Journal*  $\lambda_1$  and  $\pi$  and  $\lambda_2$  and  $S$ . Husniak, D. Peraković, I. Forenbacher, and M. Mumdziev. Telematics system in usage in usag **based motor insurance. Insurance motor insurance insurance. In the contract of the contract of the contract of** K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*  $\mathbf{M}$ *<sup>2</sup>:*  $\rightarrow$  *71.*  $\lambda$  $A \times A \times A \times B$ <br>intelligent transport systems to proceed (its): Privacy implications of vehicle infotainment and telematics systems. In *Proceedings of the 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications*,  $\mathcal{L}$  25–31, 2016. H. Jeong. Dimension reduction techniques for summarized telematics data. *Journal of Risk Management*, 337 H. Jeong and E. A. Valdez. Ratemaking application of Bayesian LASSO with conjugate hyperprior. *Available at SSRN 3251623*, 2018.  $\mathbb{L}^{\times}$  decay and  $\mathbb{R}$  and  $\mathbb{R}^{\times}$  and  $\mathbb{R}^{\times}$  and  $\mathbb{R}^{\times}$  mixed models for  $\mathbb{R}^{\times}$ dependent compound risk models. *Variance*, 14(1), 2021.  $N$  with  $\mathcal{A}$  denote reflect the  $N$  -  $\mathcal{A}$  management  $\mathcal{A}$ with Bayesian generalized and shape. Insurance: **Insurance:** and shape. Insurance: *Mathematics and Economics*,  $\overrightarrow{H}$ ,  $\overrightarrow{A}$  $\mathbb{F}$ . Klinker. Klinker. Generalized mixed mixed models for  $\mathbb{R}$ ibility into a generalized linear model setting. In a general setting. In *Let in a* generalized linear model society E-Forum, *Winter 2011 Volume 2*  $\lambda$ <sub>101</sub> T. Littl**an** Bectronic Library **A**  $\delta$ ,  $\delta$ ,  $\delta$ ,  $\delta$  /  $\delta$  /  $\delta$  /  $\delta$  /  $\delta$  / [https://policycommons.net/artifacts/1189025/](https://policycommons.net/artifacts/1189025/distance-based-vehicle-insurance/1742147/) [distance-based-vehicle-insurance/1742147/](https://policycommons.net/artifacts/1189025/distance-based-vehicle-insurance/1742147/) Y.-L. Ma, X. Zhu, X. Hu, and Y.-C. Chiu. The use of context-sensitive insurance telematics data in autor insurance *insurance making. The making making. Transportation Research Part A: Policy and Practice*  $1 \neq 2$ <sup>1</sup> $\mathbf{3}$ G. Meyers and I. V. Hoyweghen. 'Happy failures': Experimentation with behaviour-based personalisation in car insurance. *Big Data & Society*, 7(1):2053951720914650, 2020. N. Milanović, M. Milosavljević, S. Benković, D. Starčević, and Ž. Spasenić. An acceptance **b** a production in the novel technologies in care in care in care in care in the sustainability, 11(24):10331, 2020. In the main care i

<span id="page-45-6"></span><span id="page-45-4"></span><span id="page-45-3"></span><span id="page-45-2"></span> $\mathbf{r}$ 

<span id="page-46-5"></span><span id="page-46-4"></span><span id="page-46-3"></span><span id="page-46-2"></span><span id="page-46-1"></span><span id="page-46-0"></span> $\mathcal{A}\times\mathcal{A}$  and  $\mathcal{A}\times\mathcal{A}$  and  $\mathcal{A}\times\mathcal{A}$ telematics data—XGBoost versus logistic regression. *Risks*, 7(2):70, 2019.  $\mathbf{J}$  **by** Inference and mission metrika,  $\mathbf{J}$  starts and  $\mathbf{J}$  $K$  sakthivel and  $K$  and  $K$  comparative study of  $K$  comparative study of  $K$  and  $K$ artificial network international Journal of Statistics and model in comparation of *International Journal of Statistics and*  $S$ *vstems*,  $111/2$ r/2<sub>0</sub>  $\mathbf{A}$ . A. Smith,  $\mathbf{A}$ ,  $\mathbf{A}$ , claim patterns using data mining: A case study. *Journal of the operational research society*,  $\vec{J}$   $\vec{J}$   $\vec{J}$ B. So, J.-P. B. A. A. Valdez. Synthesis and E. A. Valdez. Synthetic dataset generation of driver telematics. S  $R$ *isks*,  $\bullet$ <sup>\*\*\*</sup> $\mathbf{r}'$  $D\cup \{X\}$ i. G. J. E. I. V. V. E. I. Vlahogianni. Innovative motor insurance motor insurance motor insurance schemes: A **review うちゃく こうしょう こうしゃ こうしょう こうしょう しゅうしょう しゅうしょう しゅうしょう しゅうしょう しゅうしゃ しゅうしゃ しゅうしゃ しゅうしゃ こうしゃかい こうしゃかい こうしゃかい こうしゃ**  $8^{44} - 8^{4}9$ <sub>1</sub>  $\mathbb{F}$  and K. Century. The mobile internet: current situation and  $\mathbb{F}$ *Principles and Applications of Narrowband Internet of Things (NBIoT)*,  $\ast$  **6** – **1**  $\boldsymbol{M}$ R. Verbelen, K. Antonio, and G. Claeskens. Unravelling the predictive power of telematics data in care *insural of the Royal Statistical Society: Series C (Applied*).  $\bullet$ *Statistics)*, **14** *Tu* = 0<sub>1</sub> *A*<sub>1</sub> *a*<sup>4</sup> F. Vermet. Statistical learning methods. Statistical learning methods. **1:** 1:43–82, 2:43–82, 2:43–82, 2018. **1:43–82, 2018.** 2:43–82, 2018. **1:43–82, 2:43–82, 2:43–82, 2:43–82, 2:43–82, 2:43–82, 2:43–82, 2:43–82, 2:43–82,**  $\mathcal{A}_{\mathbf{w}}$  and  $\mathcal{A}_{\mathbf{w}}$  and  $\mathcal{A}_{\mathbf{w}}$  approach to projection approach to propensity score estimation projection  $\mathcal{A}_{\mathbf{w}}$ for handling selection bias under missing at random. *arXiv e-prints*, pages arXiv–2104,

<span id="page-47-0"></span>**Appendix A**

**Results**





<span id="page-50-0"></span>**Appendix B**

# **Code**

 $a_k$  for  $\mathcal{A}_k$ 

```
x4 \leftarrow \text{rnorm}(1)lambda <- exp(-1.3-4*x1 + 3.4*x2 + 0.1*x3 + 0.5*x4)NB_Claim <- rpois(I, lambda)
Duration \leq- rep(1, 1)
fdata \leq as. data. frame(cbind(x1, x2, x3, x4, Duration, NB_Claim))
#for testing
set.seed(j+1000)
test_ind <- sample(1:nrow(fdata), 10000)
forr.data<-fdata[ test_ind, ]
trtt.data<-fdata[-test_ind,]
#sampling- when using a specific sampling method, comment other two sampling sections.
##################################################RS
set.seed(j + 2000)tele_ind <- sample(1:90000, nrow(fdata)*0.1)
ntr <- length(tele_ind)
###################################################NIS(advsel)
#set.seed(j+2000)
#dz <- 1/(1+exp(2*trtt.data$NB_Claim))
#dz <- dz/mean(dz)/9
#dzz <- rbinom(90000, 1, dz)
#tele_ind <- (1:90000)*(dzz==1)
#rm(dz, dzz)
#tele_ind <- tele_ind[tele_ind!=0]
#ntr <- length(tele_ind)
###################################################MAR(agesel)
#set.seed(j+2000)
#dz <- 1/(1+exp(3*trtt.data$x1))
#dz <- dz/mean(dz)/9
#dzz <- rbinom(90000, 1, dz)
#tele_ind <- (1:90000)*(dzz==1)
#rm(dz, dzz)
#tele_ind <- tele_ind[tele_ind!=0]
#ntr <- length(tele_ind)
####################################################
\begin{bmatrix} SO <- trtt.data[ tele_ind, \end{bmatrix}# A small dataset that contains both telematics and traditional features
S1 \le trtt.data[-tele ind,-tele ind,-tele ind,)
```

```
<code>T*Td(<-)Tj0g14.1220Td[(trtt.data[)-525.004(tele_ind,)]TJ150220Td[(trt.12214_in9D4]TJ5atetelematics)-</code>
```

```
b_S0 \leq a s. matrix(cbind(S0[, c(5, 1:3)], S0[, 6]*S0[, c(5, 1:3)]))
b_S <- as.matrix(cbind(S[, c(5, 1:3)], S[, 6]*S[, c(5, 1:3)]))
#function for optimize using nleqslv()
cal eqn \leq function(parm) {
  result <- colSums(as.vector(1+nrow(S1)/nrow(S0)*exp(parm%*%t(b_S0)))*b_S0)-colSums(b_S)
 return(result) }
#find for parameters of basis functions
fit2 <- nleqslv(rep(0,8), cal_eqn)
#calculate weights from information projection
w3 \leq 1+nrow(S1)/nrow(S0)*exp(b S0 %*% fit2$x)
#############################
#combine weights to S0
SS6<-\text{cbi} nd (SO, w3)
#fitted the model with ws
glm. freq. S3 \leq -q/m(NB) Claim \sim. -Duration-w3, offset=log(Duration),
                   weights= w3, data=SS6, family=poisson())
x_S0 <- model.matrix(glm.freq.S3)
#coef of proposed model
prop2_coef[j,] <- summary(glm.freq.S3)$coefficients[,1]
# sandwich formula for variance estimation
Ui <- cbind(c(as.vector(w3)-1, rep(-1, nrow(S1)))* b_S,
 c(w3*(SS6$NB_Clain-fitted(qlm.freq.S3)), rep(0, nrow(S1)))*as.matrix(S[,c(5,1:4)]))
U_i \leq U_i - \text{rep}(col Means(U_i), each = nrow(S))V_U \le - (t(U_i))^2 % Ui
tau \le rbind(cbind(t(b S0) %*% (as.vector(w3-1)*b S0),
        matrix(0, ncol=ncol(x_S0), nrow=ncol(b_S0))),
 cbind(t(x_S0) %*% (as.vector(w3-1)*(SS6$NB_Claim-fitted(glm.freq.S3))*b_S0),
        -t(x, S0) %*% (as.vector(w3*fitted(glm.freq.S3))*x S0) ))
invtau <- solve(tau)
prop2_stde[j,] <- sqrt(diag(invtau \% \vee U \wedge \% invtau))[-(1:ncol(b_S0))]
##########################boosting model
glm. freq. boost \leq qlm(NB_Claim \sim x4-1, data=S0,
            offset=log(Duration)+predict(glm.freq.trad, S0), family=poisson())
#coefficients and SE
boost_coef[j,] <- c(summary(glm.freq.trad)$coefficients[,1],
                    summary(glm.freq.boost)$coefficients[,1])
boost_stde[j,] <- c(summary(glm.freq.trad )$coefficients[,2],
                    summary(glm.freq.boost)$coefficients[,2])
################################try forecast
pred.naive \leq predict(glm.freq.naive, newdata = forr.data, type="response")
pred.full <- predict(glm.freq.full , newdata = forr.data, type="response")
pred. S3 \leq exp(as. matrix(forr. data[, c(5, 1:4)]) %*% prop2_coef[j,])
pred.trad <- predict(glm.freq.trad , newdata = forr.data, type="response")
pred.boost <- pred.trad * exp(coeff(g/m, freq. boost)*forr. data $x4)#remove datasets for this split
rm(tele_ind, test_ind)
#RMSE
RMSEs[j, -4] <- sqrt(c(
 mean((forr.data$NB_Claim-pred.naive)^2),
```

```
mean((forr.data$NB Claim-pred.trad )^{2}),
      mean((forr.data$NB_claim-pred.ful 1)^2),
      mean((forr.data$NB_C1aim-pred.S3)<sup>\wedge2),</sup>
      mean((forr.data$NB Claim-pred.boost)^2)))
    #MAE
    MAES[i, -4] \leq Cmean(abs(forr.data$NB_Claim-pred.naive)),
      mean(abs(forr.data$NB_Claim-pred.trad)),
      mean(abs(forr.data$NB_Claim-pred.full )),
      mean(abs(forr.data$NB_Claim-pred.S3 )),
      mean(abs(forr.data$NB Claim-pred.boost)))
    #DEV
    DEVs[i, -4] <- c(
      Poisson. Deviance(pred. naive, forr. data$NB_Claim),
      Poisson. Deviance(pred.trad, forr.data$NB_Claim),
      Poisson. Deviance(pred. full, forr. data$NB_Claim),
      Poisson. Deviance(pred. S3, forr. data$NB_Claim),
      Poisson.Deviance(pred.boost, forr.data$NB_Claim))
 })
#summarizing the outputs
col Means (RMSEs)
colMeans(MAEs)
col Means (DEVs)
#true coeffients used for data generation
true coef \leq C(-1.3, -4, 3.4, 0.1, 0.5)#bias of each estimator
bias_naive <- true_coef - colMeans(naive_coef)
bias_trad <- true_coef - colMeans(trad_coef)
bias_prop2 <- true_coef - colMeans(prop2_coef)#proposed
bias_boost <- true_coef - colMeans(boost_coef)
bias_full <- true_coef - colMeans(full_coef)
#RMSE of estimates
rmse_naive <- sqrt(col \text{Means}((nai \text{ ve } coef-rep(true\_coef, each=J))^2))rmse trad <- sqrt(colMeans((trad_coef -rep(true_coef, each=J))^2))
rmse_prop2 <- sqrt(colMeans((prop2_coef-rep(true_coef, each=J))^2))#proposed
rmse_boost <- sqrt(colMeans((boost_coef-rep(true_coef, each=J))^2))
rmse_full <- sqrt(col \text{Means}((ful \text{close} - rep (true \text{code}, each=J))^2))#CI of estimator
naive_90CI <- colMeans((naive_coef-1.645*naive_stde<rep(true_coef, each=J))*
                          (naive_coef+1.645*naive_stde>rep(true_coef, each=J))*1)
trad_90Cl <- colMeans((trad_coef -1.645*trad_stde<rep(true_coef, each=J))*
                         (train\_coef + 1.645*trad\_stde\geq rep(true\_coef, each=J))*1)prop2_90CI <- colMeans((prop2_coef-1.645*prop2_stde<rep(true_coef, each=J))*
                          (prop2\_coef+1.645*prop2\_stde\ge rep(true\_coef, each=J))^*1,na.rm=TRUE)#proposed
boost_90CI <- colMeans((boost_coef-1.645*boost_stde<rep(true_coef, each=J))*
                          (boost_coef+1.645*boost_stde>rep(true_coef, each=J))*1)
full_90Cl <- colMeans((full_coef -1.645*full_stde<rep(true_coef, each=J))*
                         (ful Lcoef +1.645*ful Lstde>rep(true.coef, each=J))*1)
```
*#### Preliminary analysis #### #naive model* glm.freq.naive <- glm(NB\_Claim ~ .-Duration, offset=log(Duration), data=S0, family=poisson()) *#full model* glm.freq.full <- glm(NB\_Claim ~ .-Duration, offset=log(Duration), data=S, family=poisson()) *#traditional model* glm.freq.trad <- glm(NB\_Claim ~ .-Duration, offset=log(Duration), data=S[,  $c(1:13, 30)$ ], family=poisson())

```
data=S0, offset=log(Duration)+predict(glm.freq.trad, S0)
                        , family=poisson())
#coefficients and SE
boost_coef[j,] <- c(summary(glm.freq.trad)$coefficients[,1],
                    summary(glm.freq.boost)$coefficients[,1])
boost stde[j,] <- c(summary(glm.freq.trad )$coefficients[,2],
                    summary(glm.freq.boost)$coefficients[,2])
################################try forecast
pred.naive \leq predict(glm.freq.naive, newdata = forr.data, type="response")
pred.full <- predict(glm.freq.full , newdata = forr.data, type="response")
pred. S3 \leq exp(as. matrix(forr. data[, c(2:29)]) \frac{9}{2} \frac{1}{2}propd_coef[j, 2: 29]+propd_coef[j, 1]+ log(forr.data[, 1]))
pred.trad <- predict(glm.freq.trad, newdata = forr.data, type="response")
pred.boost <- pred.trad * exp(as.matrix(forr.data[14:29])%*%coef(glm.freq.boost))
#remove datasets for this split
rm(tele_ind, test_ind)
#RMSE
RMSEs[i, 1 \leftarrow \text{sqrt}(c)mean((forr.data$NB.Claim-pred.naive)^2),
  mean((forr.data$NB_Claim-pred.trad)^2),
 mean((forr.data$NB_Claim-pred.full )^2),
  mean((forr.data$NB Claim-pred.S3 )^2),
  mean((forr.data$NB_Claim-pred.boost)^2)))
#DEV
DEVs[j, ] \leftarrow c(Poisson. Deviance(pred. naive, forr. data$NB_Claim),
  Poisson. Deviance(pred.trad, forr.data$NB_Claim),
  Poisson. Deviance(pred. full, forr.data$NB_Claim),
  Poisson. Deviance(pred. S3, forr. data$NB_Claim),
  Poisson.Deviance(pred.boost, forr.data$NB_Claim))
```
<span id="page-57-0"></span>**Appendix C**

### **Basic Setup of Proposed Method**



<span id="page-58-2"></span><span id="page-58-1"></span><span id="page-58-0"></span>
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx = \int_{0}^{2} f(x) dx
$$
\n
$$
\int_{0}^{2} f(x) dx
$$

<span id="page-58-3"></span>E

The results of  $\mathbb{R}$  is the results of  $\mathbb{R}$  in the results of  $\mathbb{R}$