IMPROVING AUTOMOBILE INSURANCE CLAIMS FREQUENCY PREDICTION WITH TELEMATICS CAR DRIVING DATA

BY

SHENGWANG MENG, HE WANG, YANLIN SHI AND GUANGYUAN GAO

ABSTRACT

Novel navigation applications provide a driving behavior score for each finished trip to promote safe driving, which is mainly based on experts' domain knowledge. In this paper, with automobile insurance claims data and associated telematics car driving data, we propose a supervised driving risk scoring neural network model. This one-dimensional convolutional neural network takes time series of individual car driving trips as input and returns a risk score in the unit range of (0,1). By incorporating credibility average risk score of each driver, the classical Poisson generalized linear model for automobile insurance claims frequency prediction can be improved significantly. Hence, compared with non-telematics-based insurers, telematics-based insurers can discover more heterogeneity in their portfolio and attract safer drivers with premiums discounts.

KEYWORDS

Telematics car driving data, automobile insurance pricing, one-dimensional convolutional neural network, limited fluctuation credibility model, generalized linear model, claims frequency.

JEL codes: G22, C45

1. INTRODUCTION

Traditionally, insurers use classical actuarial risk factors to classify policyholders and determine premiums for different homogeneous risk groups. With the advancements of telematics technology, insurers can collect detailed car driving information. Telematics car driving data are personalized, which

Astin Bulletin 52(2), 363–391. doi:10.1017/asb.2021.35 © The Author(s), 2021. Published by Cambridge University Press on behalf of The International Actuarial Association.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

encode differences in driving behavior not included in classical social-demographic risk factors, such as regions and ages. For example, some young drivers may have a cautious driving style, while some matured drivers may still demonstrate a wild and aggressive driving behavior. This work is among the first to explore how such information can be learned from telematics car driving data.

The most commonly used methods in automobile insurance ratemaking include (1) two sub-models for claims frequency and severity, respectively and (2) Tweedie's compound Poisson models for pure premium (Yang *et al*. 2016). Insurance industry typically prefers the first approach, and Delong *et al*. (2021) provide both theoretical and empirical support for such a preference. In this paper, our goal is to improve automobile insurance claims frequency prediction by utilizing telematics car driving data.

Admittedly, there are concerns on using personalized data and algorithmic prediction (Mahapatra 2019; Cevolini and Esposito 2020). Despite this, we prefer the statement by Cather (2020) that incorporating telematics data into automobile insurance risk classification systems would "minimize insurance discrimination and increase cream skimming adverse selection." Specifically, asymmetric information favoring telematics-based insurers supports premium discounts that attract safer drivers, prompting an underpopulation of low-risk drivers among non-telematics insurers.

There are three streams of existing literature that are related to our work: actuarial science, transportation, and engineering (i.e., IEEE related). Listed below is a summary of literature review. Where necessary, we briefly discuss both the merits and limitations of existing method. For the actuarial science stream, among others, the following papers are related to our work:

- Weidner *et al*. (2016); Weidner *et al*. (2017) extract covariates from time series of telematics data using the discrete Fourier transforms. This approach transform time series of varying length into constituent frequency components of the same length, which could facilitate the subsequent statistical modeling. The interpretation, however, is not straightforward.
- Huang and Meng (2019), Paefgen *et al*. (2014), and Verbelen *et al*. (2018) calculate and adopt descriptive statistics of telematics data. This approach is easy to implement with interpretable results. The limitation is that it depends on researchers' domain knowledge (ad hoc) and some important statistics might be missed.
- Gao *et al*. (2021) apply the convolutional neural networks (CNNs) to learn patterns in speed-acceleration heatmaps, with explanations of this network. To apply CNNs, one needs to rely on sophisticated machine learning platform such as TensorFlow, which may not be familiar to actuaries.
- Other scholars also explore the *risk exposure* of driving distances or driving duration; see Ayuso *et al*. (2016a); Ayuso *et al*. (2016b), Boucher *et al*.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

(2017), and Lemaire *et al*. (2016) for examples. The exposure of distance is known to be a significant risk factor for pricing and has been already used in the usage-based insurance.

- Guillen *et al*. (2020) study the association of telematics car driving data with near-miss events such as cornering, braking, and accelerating. This approach adopts near-miss events as response variable, which provides a more efficient "supervision" on model fitting than the number of claims. One concern is how to define a near-miss event quantitatively and objectively.
- Richman (2020) discusses and compares potential approaches to analyze telematics car driving data. Geyer *et al*. (2020) explore the effect of driving behavior on risk and insurance selection from economic perspective.
- Since telematics data are posterior experiences, Denuit *et al*. (2019) propose a credibility model to incorporate the posterior information of driving behavior via multivariate mixed models.

In transportation field, among others, existing literature includes

- Joubert *et al*. (2016), Ma *et al*. (2018), and Hu *et al*. (2019) investigate driving behaviors and find that certain driving behaviors are more relevant to at-fault accidents. Those works consider both telematics data and contextual data such as road conditions, traffic flow, speed limits. In insurance pricing, however, those contextual data cannot be employed as risk factors directly. Instead, regions are used to integrate those information implicitly.
- Ho *et al*. (2014), Hung *et al*. (2007) and Kamble *et al*. (2009) adopt telematics data and driving cycles to understand vehicular emissions, energy consumption, and impacts on traffic in different cities around the world. The speed-acceleration heatmaps are examined in those papers, which are also studied in actuarial research including Gao *et al*. (2021).

As for the IEEE Transactions journals and conference papers, there are many works on smartphone-based telematics. Listed below are some typical works potentially related to our research and other actuarial literature.

- Wahlström *et al*. (2015) detect the dangerous cornering using GPS data, which is relevant to near-miss events defined in Guillen *et al*. (2020). In their work, Wahlström *et al*. (2015) argue that the developed framework can be applied to smartphone-based insurance telematics applications.
- Wahlström *et al*. (2017) review influential research on smartphone-based vehicle telematics, such as vehicle navigation, driver classification and road condition monitoring. Although our data are collected from devices installed in cars, smartphone-based telematics data record much similar information. Such smartphone-based data are attractive to insurance

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

companies for its low costs and contextual information such as weather and road conditions.

- Wahlström *et al*. (2018) investigate the fusion of on-board-diagnostics (OBD) systems and global navigation satellite system (GNSS) measurements of speed from an electrical engineering perspective. This is related to our data cleaning, since in our case, multiple measurements may exist for the same variable. The fusion or selection of those measurements makes the very first (preliminary) step of our analysis.
- Savelonas *et al*. (2020), Girma *et al*. (2019) and Carvalho *et al*. (2017) employ recurrent neural networks to identify drivers or to learn different driving behaviors such as normal, moderate, and aggressive. The merit of those recurrent neural networks is their automated feature engineering, whereas the difficulties associated are interpretation and how to label trips therein.

As briefly outlined above, the data employed in this paper are collected by OBD systems rather than smartphones. Also, our telematics data are privately owned and thus not publicly available. In terms of methodology, we focus on convolutional neural network (CNN), one of the most critical deep learning algorithms (LeCun *et al*. n.d.). It is capable of learning representation and conducting shift-invariant classification. From 1980s to 1990s, LeCun invented the notion of convolutional network in his epoch-making LeNet-5, which used convolution to extract similar spatial features at multiple locations with few parameters. After 20 years development, CNNs demonstrate outstanding performance in computer vision. In automobile insurance field, Singh (2016) implemented two CNNs to classify pictures of drivers. From a similar perspective, Masood *et al*. (2020) propose a machine learning algorithm based on CNNs to not only detect distracted drivers but also identify the cause of distraction. The input images are obtained from a camera module installed in automobiles.

We employ a one-dimensional convolutional neuron network (1D CNN) to classify trips into either *potentially risky or potentially safe*. 1D CNN can learn representations along time series of car driving trips. We demonstrate that those representations learned from such a binary classification task are related automobile insurance claims frequency, which is an essential component in automobile insurance pricing. More specifically, archetypal drivers are selected and their individual trips are labeled binarily, on which a 1D CNN is calibrated. The output neuron of 1D CNN is defined as risk score of individual trip. The credibility average risk score is derived for each driver in a limited fluctuation credibility model, where the variation in individual trip risk scores and number of trips are considered. By incorporating credibility average risk score, the classical Poisson generalized linear model (GLM) for automobile claims frequency prediction is significantly improved, which leads to a more fair automobile insurance classification pricing. Therefore, telematics-based insurers can discover more heterogeneity in their portfolio and attract safer

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

drivers with premiums discounts, while non-telematics-based insures may be exposed to antiselection risk. In summary, our contributions are threefold: (1) the proposed model is among the first to evaluate the risk of individual car driving trips; (2) we develop an effective method to select archetypal drivers and label their trips; and (3) the limited fluctuation credibility model is employed to calculate credibility average risk score.

This paper is structured as follows. Section 2 describes telematics variables and preprocesses telematics data into time series. Section 3 describes automobile insurance data and establishes a Poisson GLM for claims frequency prediction with classical actuarial risk factors. Section 4 proposes a risk scoring 1D CNN for individual trips and develops a limited fluctuation credibility model for average risk score. Section 5 improves the classic claims frequency model proposed in Section 4 by incorporating credibility average risk score and compares with two competing methods. Section 6 concludes the paper with important findings.

2. Preprocessing telematics car driving data

In this section, we first describe telematics variables collected. Then data cleaning is conducted. Finally, we extract time series of individual car driving trips. For a detailed process of the telematics data cleaning, please refer to the supplementary material. In this work, we have spent at least 80% of total time to preprocess telematics data.

2.1. Telematics variables

A trip is defined as the period from engine start to engine switch off. During a trip, we receive records from three independent sensors: GPS, instrumental panel, and three-axis accelerometer.

2.1.1. GPS variables

The following five variables are recorded from GPS by seconds:

- `GPS_Latitude`. This variable is the GPS latitude in decimal degree.
- `GPS_Longitude`. This variable is the GPS longitude in decimal degree.
- `GPS_Heading`. This variable is the approaching direction of the vehicle $\psi^{(gps)}$ in decimal degree. It ranges from 0 to 360.
- `GPS_Speed`. This variable is the vehicle speed $v^{(gps)}$ in km/h.
- `Positional_Quality`. This binary variable indicates the GPS signal quality. When it is zero, the vehicle cannot be located by the GPS satellite, and the four GPS variables listed above are invalid.

The spherical coordinates (`GPS_Latitude`, `GPS_Longitude`) need be converted to the plane coordinates (x,y), which are used to calculate distance,

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

speed $v^{(xy)}$, angle $\psi^{(xy)}$, acceleration $a^{(xy)}$, and angle change $\Delta^{(xy)}$. When GPS signal drift problems occur, GPS coordinates will be affected, so will variables dependent on them.

2.1.2. Instrument panel variables

The following two variables are recorded from the instrumental panel by seconds:

- `VSS_Speed`. This variable is vehicle speed $v^{(vss)}$ shown in the instrument panel. We are informed by the insurer that when valid, it is more reliable than `GPS_Speed` $v^{(gps)}$.
- `Engine_RPM`. This variable is the rotation of engine per minute shown in the instrument panel, which is not used in the study.

2.1.3. Accelerometer variables

The following three variables are measured by a three-axis accelerometer.

- `Accel_Lateral`. The lateral acceleration $a'^{(acc)}$ is perpendicular to the car approaching direction, which measures the change of car approaching direction:
- `Accel_Longitudinal`. The longitudinal acceleration $a^{(acc)}$ is along the car approaching direction, which measures the change of speed values.
- `Accel_Vertical`. The vertical acceleration is always at around the gravity $9.8m/s^2$.

There is calibration bias with accelerometer. Please refer to the supplementary material for more details.

2.2. Telematics data cleaning

Telematics car driving data have a large size and are deemed big data. Data cleaning for such big data is very challenging since the same procedure needs to be applied to all the trips of all cars. Thus, the data cleaning should be adequately flexible for this aim. An investigation using several visual tools implies that we need to consider two data cleaning issues. First, there are missing values in GPS coordinates (x,y), GPS speed $v^{(gps)}$, GPS heading $\psi^{(gps)}$, and instrument panel speed $v^{(vss)}$. These missing values need to be interpolated, respectively. Second, there are frequent calibration issues with the accelerometer variables $a^{(acc)}$, $a'^{(acc)}$. It is difficult to remove the calibration bias since the timing and severity of such a bias are rather random. Consequently, it is better to consider other variables to describe the acceleration in two directions.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

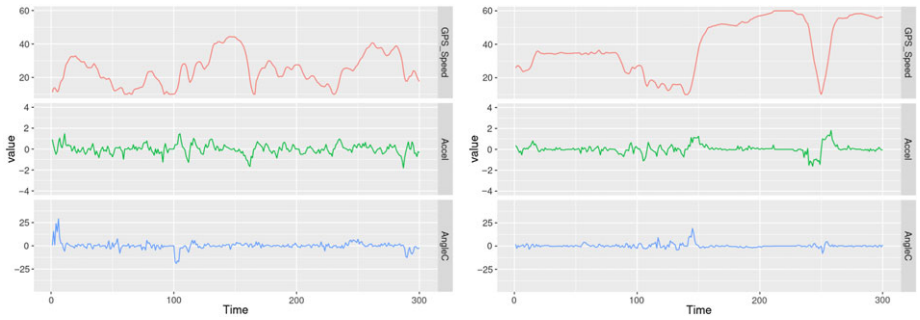


FIGURE 1: Two sample trips.

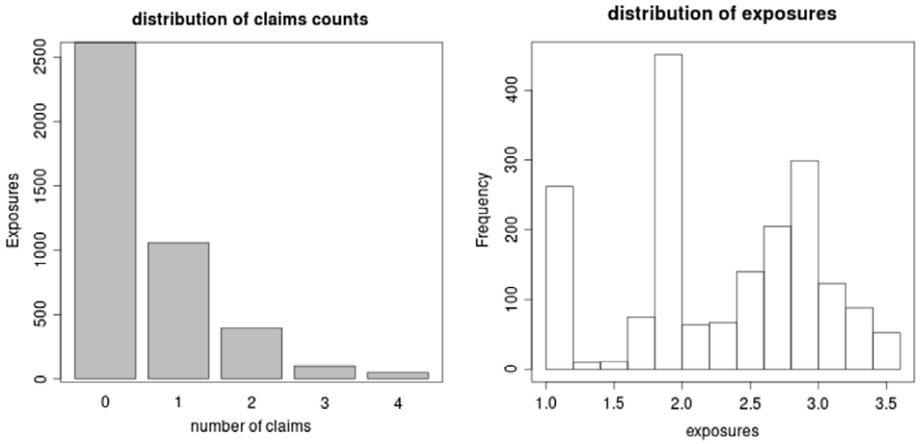


FIGURE 2: The distributions of claims number and exposures.

2.2.1. Data imputation

Data imputation is essential if time series of telematics variables are analyzed. On the contrary, if only summary statistics of telematics variables are employed (e.g., mean of speed or standard deviation of acceleration), the interpolation of missing values is not necessary according to the law of large numbers.

In this paper, missing values of GPS coordinates (x,y), GPS speed $v^{(gps)}$, GPS heading $\psi^{(gps)}$, and instrument panel speed $v^{(vss)}$ are linearly interpolated, respectively. Note that missing values at boundaries of trips cannot be interpolated, and such missing values are therefore truncated. The imputation of GPS heading $\psi^{(gps)}$ around 0 or 360 degree is challenging and may leads to biased GPS heading imputation. However, such biased imputation should have little effects on angle change due to the useness of sine and arc-sine functions; see Equation (2). Nevertheless, the proportion of missing values in our data set is very small. Thus, we reply on our data quality with high confidence. Please see Figures 1–4 in the supplementary material for details.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

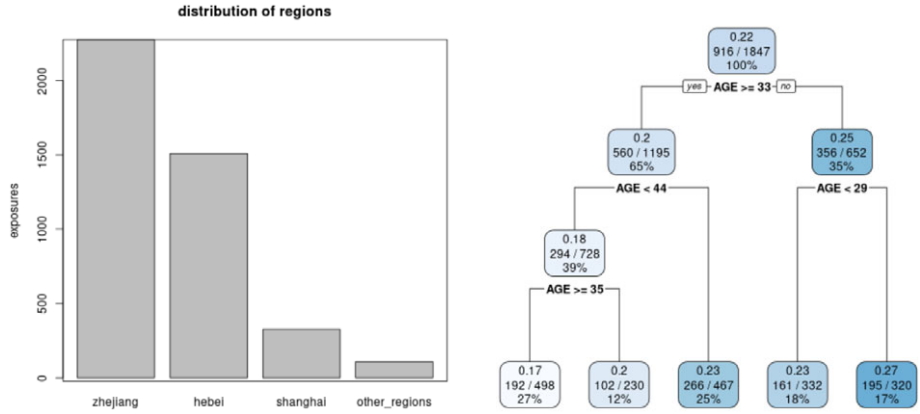


FIGURE 3: The distribution of exposures across four regions. The marginal Poisson tree with driver's age covariate.

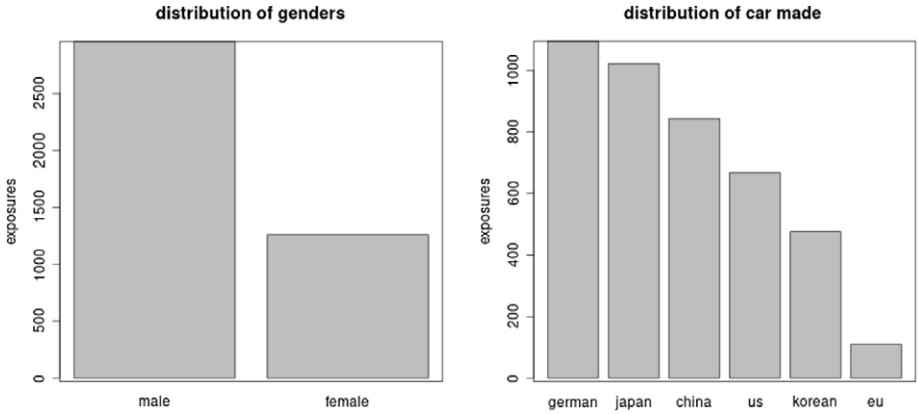


FIGURE 4: The distribution of exposures across genders. The distribution of exposure across car made countries.

2.2.2. Derived acceleration and angle change

Since the quality of accelerometer variables is relatively questionable, alternative measures are derived to measure the acceleration and direction changes. We derive longitudinal acceleration rates $a^{(gps)}, a^{(vss)}, a^{(xy)}$ from GPS speed $v^{(gps)}$, instrument panel speed $v^{(vss)}$, and GPS coordinates (x,y), respectively,

$$\begin{aligned} a_t^{(gps)} &= v_t^{(gps)} - v_{t-1}^{(gps)} \\ a_t^{(vss)} &= v_t^{(vss)} - v_{t-1}^{(vss)} \\ a_t^{(xy)} &= v_t^{(xy)} - v_{t-1}^{(xy)} \end{aligned}, \tag{1}$$

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

where

$$v_t^{(xy)} = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2}.$$

We replace lateral acceleration $a'^{(acc)}$ by angle changes of heading directions $\Delta^{(gps)}$, $\Delta^{(xy)}$, which are derived from GPS heading $\psi^{(gps)}$ and GPS coordinates (x,y), respectively,

$$\begin{aligned} \Delta_t^{(gps)} &= \arcsin\left(\sin\left(\psi_t^{(gps)} - \psi_{t-1}^{(gps)}\right)\right) \\ \Delta_t^{(xy)} &= \arcsin\left(\sin\left(\psi_t^{(xy)} - \psi_{t-1}^{(xy)}\right)\right), \end{aligned} \tag{2}$$

where

$$\psi_t^{(xy)} = \mathbb{1}_{(-\infty,0)}\left(\arctan 2\left(\frac{x_t - x_{t-1}}{y_t - y_{t-1}}\right)\right) \times 2\pi + \arctan 2\left(\frac{x_t - x_{t-1}}{y_t - y_{t-1}}\right) \in [0, 2\pi].$$

Note that a positive value of Δ indicates a right turn whereas a negative one suggests a left turn. The jump between 0 and 360 is not an issue since we take sine values.

2.2.3. Selection of telematics variables

After data imputation and derivation of the corresponding telematics variables, we discuss variable selection among speed $v^{(gps)}$, $v^{(vss)}$, $v^{(xy)}$, acceleration $a^{(gps)}$, $a^{(vss)}$, $a^{(xy)}$, angle $\psi^{(gps)}$, $\psi^{(xy)}$, and angle change $\Delta^{(gps)}$, $\Delta^{(xy)}$. Due to the GPS coordinates drift, we should avoid using variables derived from plane coordinate (x,y). Instead, the telematics variables $v^{(vss)}$, $v^{(gps)}$, $a^{(vss)}$, $a^{(gps)}$, $\psi^{(gps)}$, and $\Delta^{(gps)}$ are more reliable. The three variables v, a, Δ are related to driving behavior, while heading direction $\psi^{(gps)}$ is irrelevant. Also, variables from the same sensor should be adopted for a certain study, that is, we should use either $(v^{(vss)}, a^{(vss)})$ or $(v^{(gps)}, a^{(gps)}, \Delta^{(gps)})$ only.

2.3. Time series of individual trips

A trip may contain several standing still phases $\{t{:}v_t = 0, a_t = 0, \Delta_t = 0\}$. Those phases should be removed since they are irrelevant to driving behavior. Also, it is worth noting that GPS variables tend to be more unreliable with a slow driving speed (e.g., when parking). Thus, by focusing on a meaningful and usual speed interval [10,60] km/h, we can make more sensible comparison among drivers. It is worth pointing out that one may consider several speed intervals, for example, [10,60], (60,90], (90,120] km/h, and study partial trips on those speed intervals, respectively. We refrain from attempting this approach since

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

our portfolio is small, and many drivers in downtown have very few trips with speed above 60 km/h in our data set.

We first extract partial trips in [8, 62] km/h and set the acceleration and angle change in the *shifting parts* $[8, 10] \cup [60, 62]$ km/h to 0. Further, we truncate speed in the shifting parts $[8, 10] \cup [60, 62]$ km/h to 10 and 60 km/h, respectively. The first $\tau = 5 \times 60$ s of such concatenated trips are then saved for each trip. If a trip did not spend more than 5 min in speed interval [8,62] km/h, it would be discarded in the analysis.

We extract at most 500 preprocessed trips for each driver. Therefore, for each driver $i = 1, \ldots, n$, we have $c_i \in (0, 500]$ individual trips. For driver i, we stack its trips into a $c_i \times 300 \times 3$ array $\boldsymbol{z}_i = (z_{i,1}, \ldots, z_{i,c_i})'$, where each trip $z_{i,j} \in [10, 60]^{300} \times [-4, 4]^{300} \times [-45°, 45°]^{300}$. Note that the unbolded notation z_j refers to a trip j without considering which driver it belongs to. Two sample trips are plotted in Figure 1, which shows that speeding, braking, and direction changes are always associated with low speed. Note that in Section 4.3, it is demonstrated that stable results can be obtained based on those 500 trips.

Remarks. One may choose the middle stage of a trip to improve data reliability. However, it is difficult to define such a stage. On one hand, for instance, if the middle 5 min were selected, it would be unfair for trips with varying durations. On the other hand, if the first several minutes (e.g., 10 min) were discarded and the next 5 min were selected, it would be not appropriate for short trips (e.g., shorter than 15 min). Therefore, by extracting and focusing on the first 5 min, we fairly compare driving behavior in the early stage of a trip. In addition, we have specified particular speed interval and truncate acceleration and angle change. A particular speed interval makes the comparison among drivers more sensible, and the truncation eliminates leverage effects of outliers. We only focus on the first 5 min of a trip rather than a longer period or the whole trip for two reasons. First, a 1D CNN will be employed to fit the trips which only allows for inputs of the same length. Second, by choosing this short period of trips, we can largely reduce the computational time with limited resource. Admittedly, by using cloud computing and other learning algorithms, one may directly work on the complete trips to further improve the accuracy. This remains for future works.

3. CLAIMS FREQUENCY MODELING WITH CLASSICAL ACTUARIAL RISK FACTORS

We consider the claims history of $i = 1, \ldots, n = 1847$ cars whose exposure e_i is ranging from 1 to 3.48 years. Those cars have over 100 and up to 500 preprocessed trips of 5 min as discussed in Section 2.3. Note that there are 1598 car drivers with 500 such pre-processed telematics trips. We show the distribution of claims number N_i and exposures e_i in Figure 2. We observe that most drivers do not make a claim. The total claims number is $\sum_{i=1}^{n} N_i = 933$,

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 1

DATA PARTITIONS.

Data	Number of drivers	Exposure	Number of claims	Empirical claims frequency
Train	1107	2535	557	0.2197
Validation	372	843	192	0.2279
Test	368	837	184	0.2198

and the total exposures is $\sum_{i=1}^{n} e_i = 4215$. The empirical claims frequency is $\sum_{i=1}^{n} N_i / \sum_{i=1}^{n} e_i = 22.14\%$ per driver per year.

Remarks. Our preliminary data cleaning ensures that the main driver of a car does not change over the entire observation period, and we concatenate policy renewals of the same driver over this observation period. Thus, we can follow the same driver for at most 3 years and 5 months from 01/01/2014 to 31/05/2017. Insurance policies over multiple years are monitored, but only for the most recent periods telematics data become available. For this reason, we typically have a longer observation period of claims history on insurance policies than that of the corresponding telematics data. An implicit assumption is that driving behaviors in the most recent periods are good representations for the entire observation period of insurance exposure.

3.1. General setting of claims frequency modeling

We assume that claims number N_i follows a Poisson distribution with mean of $e_i\lambda(\boldsymbol{x}_i)$

$$N_i \overset{ind}{\sim} \text{Poi}(e_i\lambda(\boldsymbol{x}_i)),$$

where e_i is the exposure, and $\lambda(\boldsymbol{x}_i)$ is the claims frequency per driver per year given the risk factors $\boldsymbol{x}_i \in \mathcal{X}$. The function λ is a mapping from risk factors to claims frequency:

$$\lambda: \mathcal{X} \to \mathbb{R}_+, \boldsymbol{x} \mapsto \lambda(\boldsymbol{x}).$$

We first establish the base line generalized linear model (GLM) with λ as a log-linear function in this section. Then in Section 4, we improve it by introducing telematics covariates to expand the covariate space $\mathcal{X}$.

We will compare models in terms of out-of-sample prediction performances. Stratified split (w.r.t. number of claims) is used to partition the data into train $\mathcal{D}_{\text{train}}$, validation $\mathcal{D}_{\text{validation}}$, and test data $\mathcal{D}_{\text{test}}$ with a proportion of (0.6:0.2:0.2). We denote the corresponding index sets by $\mathcal{I}_{\text{train}}, \mathcal{I}_{\text{validation}}$, and $\mathcal{I}_{\text{test}}$, respectively. The exposures, claims counts, and claims frequencies of each data set are listed in Table 1. The index of learning data $\mathcal{D}_{\text{learn}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{validation}}$ is denoted by $\mathcal{I}_{\text{learn}} = \mathcal{I}_{\text{train}} \cup \mathcal{I}_{\text{validation}}$.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

It is natural to use the Poisson deviance loss to compare different models. The out-of-sample Poisson deviance loss on the data $\mathcal{D}_{\text{test}}$ is defined as

$$\mathcal{L}(\hat{\lambda}, \mathcal{D}_{\text{test}}) = \frac{2}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \left(e_i \hat{\lambda}(\boldsymbol{x}_i) - N_i - N_i \ln e_i \hat{\lambda}(\boldsymbol{x}_i) + N_i \ln N_i \right), \quad (3)$$

where the estimated mapping $\hat{\lambda}$ contains estimated parameters using the learning data $\mathcal{D}_{\text{learn}}$.

3.2. Preprocessing classical actuarial risk factors

The available actuarial risk factors are `regions`, `driver's age`, `driver's gender`, `car brand`, `car's age`, `seat count`, `car's price`, and `average daily distance`. For categorical variables, we merge levels containing few exposures. For continuous variables, we investigate their marginal effect on claims frequency and discretize them when necessary. Note that the last covariate `average daily distance` is obtained from the original telematics data rather than the preprocessed c_i trips defined in Section 2.3. Strictly speaking, this variable is not a classical actuarial risk factor. We include this variable together with other classical actuarial risk factors, since the distance exposure has been studied intensively in actuarial literature and widely used in the usage-based insurance.

Regions. There are three main regions, Hebei Province, Zhejiang Province, and Shanghai City, which have accounted for 97.67% of total cars. Hence, we merge the remaining regions as one, denoted by "other regions". The distribution of exposures across those four regions is shown in Figure 3. Note that one may create a continuous variable of population density in each region. This variable should be related to claims frequency, since it is a proxy of traffic density.

Driver's age. It is widely acknowledged that driver's age has a nonlinear relationship with claims frequency. Thus, we discretize driver's age into five groups according to a marginal Poisson regression tree as shown in Figure 3. The cut-off values of age are 29, 33, 35, and 44. We list driver's age group and their exposures in Table 2.

Gender. Male drivers are almost twice the size of female drivers as shown in Figure 4.

Car brands. There are 66 different car brands. Due to the small sample sizes of individual car brands, we create a new categorical variable of car made country with 6 levels: Germany, Japan, China, US, Korea, and Europe (excluding Germany). Figure 4 displays the distribution of exposures across car made countries.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 2

DRIVER'S AGE GROUPS.

Age interval	Age group	Exposure
[18, 29)	young	708
[29, 33)	middle1	724
[33, 35)	middle2	395
[35, 44)	mature1	1295
[45, 100)	mature2	1092

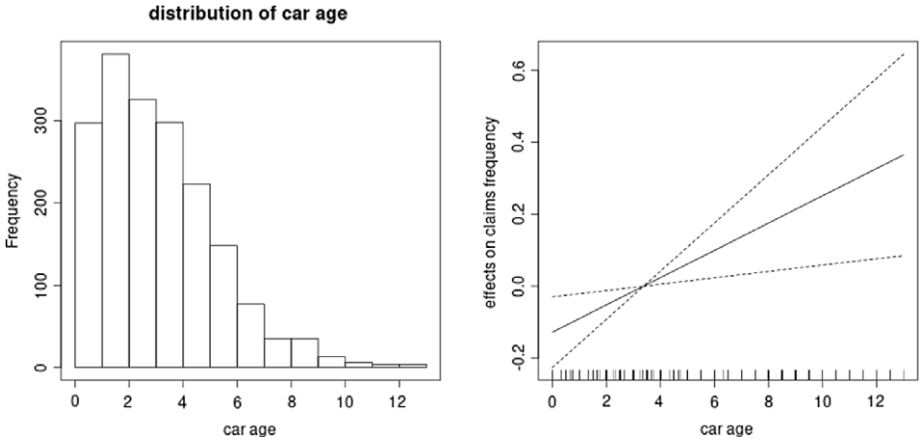


FIGURE 5: Histogram of car's age. The effect of car's age on claims frequency (from a marginal GAM).

Car's age. The distribution of car's age is plotted in Figure 5. A marginal generalized additive model (GAM) for claims frequency indicates that car's age is related to claims frequency in a log-linear fashion (due to the logarithm link function) as shown in Figure 5. Thus, discretization is not needed for this variable.

Seat count. Around 95.67% cars have 5 seats as shown in Figure 6. Hence, this variable is not quite useful for claims frequency prediction.

Car's price. The distribution of car's price in logarithm is shown in Figure 6. A preliminary analysis (using marginal tree, GAM, and GLM) demonstrates that this variable does not have a close relationship with claims frequency. This may be due to the relative small size $n = 1847$ of our portfolio.

Average daily distance. We fit a marginal GAM to investigate nonlinear effect of average daily distance on claims frequency. The left plot in Figure 7 suggests that the logged daily distance are mainly distributed between 2.5 and 4.5. Thus, we truncate the logged daily distance at 2.5 and 4.5, and re-fit a marginal GAM with truncated values. The right plot in Figure 7 indicates that the truncated logged daily distance is related to claims frequency in a log-linear fashion (due

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

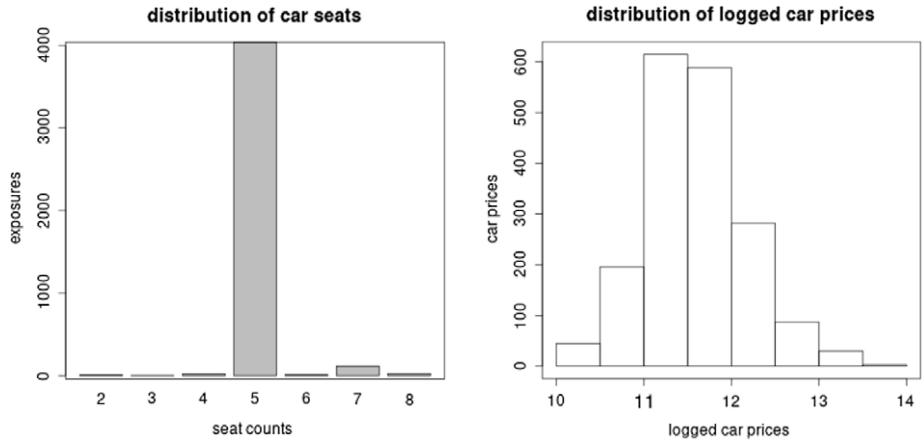


FIGURE 6: The distribution of exposure across seat count. Histogram of logged car's price.

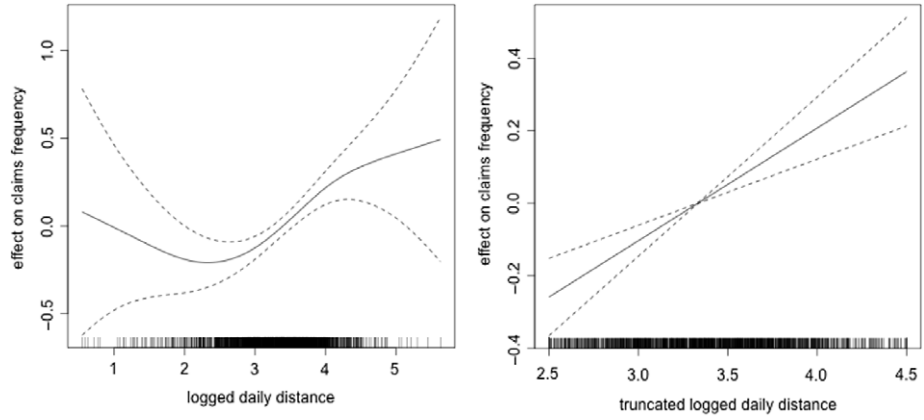


FIGURE 7: The effects of logged daily distance on claims frequency (from two marginal GAMs).

to the logarithm link function). We will then use the truncated and logged average daily distance in GLM. Figure 8 compares the distributions of original and truncated variables.

In summary, we have 4 categorical covariates: regions, driver's age groups, gender, and car made countries and 2 continuous covariates: car's age and (truncated and logged) average daily distance. Seat count, and car's price will not be used in the following analysis.

3.3. Poisson generalized linear models for claims frequency

We begin with the classical generalized linear model (GLM), which assumes the following log-linear mapping:

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

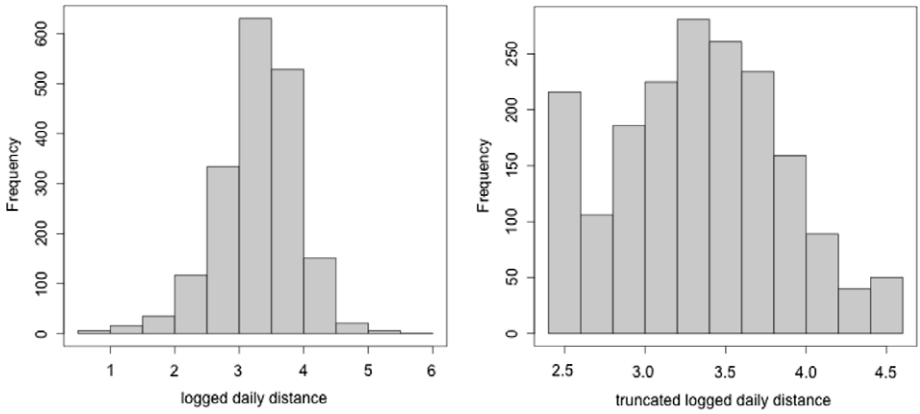


FIGURE 8: The distributions of logged daily distance without and with truncation at 2.5 and 4.5.

$$\begin{aligned}\ln \lambda^{(\mathrm{GLM})}(\boldsymbol{x}) = &\beta_0 + \alpha_{\mathrm{region}} + \gamma_{\mathrm{age_group}} + \zeta_{\mathrm{female}} + \delta_{\mathrm{car_made}} \\ &+ \beta_1 \mathrm{car_age} + \beta_2 \mathrm{trun_logged_distance},\end{aligned} \tag{4}$$

where the Zhejiang Province, mature age 1, female, and Germany made car set as reference levels. The coefficients are estimated using the learning data $\mathcal{D}_{\mathrm{learn}}$. We then perform a step-wise variable selection according to the Akaike information criterion (AIC). The final model is selected as follows:

$$\begin{aligned}\ln \lambda^{(\mathrm{GLM})}(\boldsymbol{x}) = &\beta_0 + \alpha_{\mathrm{hebei}} + \gamma_{\mathrm{young}} + \gamma_{\mathrm{middle1}} + \gamma_{\mathrm{mature2}} + \delta_{\mathrm{china}} + \delta_{\mathrm{eu}} \\ &+ \beta_1 \mathrm{car_age} + \beta_2 \mathrm{trun_logged_distance}.\end{aligned} \tag{5}$$

Hence, we have merged Shanghai Province and other regions with Zhejiang Province, middle age 2 with mature age 1, and car made in Japan, Korea, and US with Germany. Also, we have removed gender from the full model (4). The estimated coefficients are listed in Table 3. Therefore, drivers in Hebei Province have a lower claims frequency than those in Zhejiang Province, and drivers aged 18 to 32 and older than 45 have a higher claims frequency than those aged 33 to 44. Claims frequencies are not significantly different among different car made countries. Furthermore, claims frequency is increasing log-linearly with both car's age and average daily driving distance.

The out-of-sample Poisson deviance loss (3) for model (5) is 1.0306, compared with 1.1003 for a homogeneous model

$$\ln \lambda^{(\mathrm{Homo})}(\boldsymbol{x}) = \beta_0. \tag{6}$$

Hence, with classical actuarial risk factors insurers can better understand risk heterogeneity of the portfolio and divide it into several homogeneous risk groups.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 3

THE ESTIMATED COEFFICIENTS IN THE GLM (5).

coefficients	estimate	std. error	z-value	p-value
β_0	–2.7864	0.2741	–10.1648	0.0000
α_{hebei}	–0.5320	0.0864	–6.1591	0.0000
γ_{young}	0.2764	0.1111	2.4874	0.0129
γ_{middle1}	0.3719	0.1023	3.6371	0.0003
γ_{mature2}	0.2656	0.0939	2.8277	0.0047
δ_{china}	0.1411	0.0905	1.5589	0.1190
δ_{eu}	0.3885	0.2053	1.8920	0.0585
β_1	0.0295	0.0170	1.7344	0.0829
β_2	0.3283	0.0734	4.4721	0.0000

Remarks. We have also fitted a generalized boosted regression model and a XGBoost. It turns out that the out-of-sample Poisson deviance loss is not improved by those two boosting methods. Please refer to Sections 3.3.1 and 3.3.2 of Gao *et al*. (2021) for more details. Thus, the GLM (5) has sufficiently captured the prediction power of those risk factors. It may be attributed to the appropriateness of our preprocessing of covariates and the nonexistence of significant interaction effects. On the other hand, our data only contain $n = 1847$ cars with 4215 years exposure, so we may not discover potential nonlinear effects and interaction effects based on such a small portfolio. In addition, one may consider a zero-inflated Poisson (ZIP) model to address the excess of zero claims. This has been outlined in Section 6 as directions for future works.

4. A SUPERVISED DRIVING RISK SCORING NEURAL NETWORK

In this section, we employ one-dimensional convolutional neural networks (1D CNNs) to evaluate driving risk associated with individual trips. The application of 1D CNNs is appropriate in our case, for their capability of learning patterns from time series data. Our proposed method includes three steps: First, in Section 4.1, we select **archetypal drivers** and label their trips as either **potentially risky or potentially safe**. Second, in Section 4.2, we calibrate a 1D CNN to classify those labeled trips of archetypal drivers. The sigmoid probability of output neuron is defined as **risk score** of a trip. Third, in Section 4.3, we employ the calibrated 1D CNN to evaluate risk score of all trips and calculate **credibility average risk score** for each driver.

4.1. Selection of archetypal drivers

Ideally, reasonable risk scores of individual trips should improve our claims frequency predictions. More specifically, reasonable risk scores should explain some variations in residuals of the GLM (5). This motivates how we select archetypal drivers and label their trips. We calculate the deviance residuals of the Poisson GLM (5) as follows:

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 4

10 ARCHETYPAL RISKY DRIVERS.

e_i	N_i	Region_i	age_i	r_i	$\hat{\lambda}(\boldsymbol{x}_i)^{(\text{GLM})}$
3.1836	4	Other_regions	33.5	7.7461	0.2070
2.9096	4	Hebei	28.0	8.8773	0.1917
2.0000	4	Zhejiang	45.0	9.0390	0.2724
2.7918	3	Hebei	56.0	6.7977	0.1458
2.2055	4	Zhejiang	29.0	7.7946	0.2966
2.5288	3	Hebei	50.0	8.3549	0.1200
3.0000	4	Zhejiang	34.0	10.0270	0.1576
1.0000	2	Shanghai	40.0	5.2395	0.2218
2.9507	3	Hebei	29.0	5.5946	0.1748
2.0000	3	Shanghai	36.0	6.4458	0.2179

$$r_i = \text{sign}\left(N_i - e_i\hat{\lambda}^{(\text{GLM})}(\boldsymbol{x}_i)\right) \times \sqrt{2e_i\hat{\lambda}^{(\text{GLM})}(\boldsymbol{x}_i) - 2N_i - 2N_i \ln\left(e_i\hat{\lambda}^{(\text{GLM})}\right) + 2N_i \ln N_i}.$$

We select 10 drivers with the largest deviance residuals in the learning data $\mathcal{D}_{\text{learn}}$ as **archetypal risky drivers** and label their 5000 trips as potentially risky trips (coded as $R_{i,j} = 1$). For those archetypal risky drivers, their exposure, number of claims, region, age, deviance residuals, and estimated claims frequency in GLM (5) are listed in Table 4. They have made at least 2 claims in total, but the average estimated claims frequency is only 20.06% from GLM (5). We anticipate that those drivers may have a risky driving behavior not coded in age, region, or other classical risk factors.

We then select 10 drivers with no claim and the largest exposure in the learning data $\mathcal{D}_{\text{learn}}$ as **archetypal safe drivers** and label their 5000 trips as potentially safe trips (coded as $R_{i,j} = 0$). For those archetypal safe drivers, their exposure, number of claims, region, age, deviance residuals, and estimated claims frequency in GLM (5) are listed in Table 5. The average estimated claims frequency is 18.37%. We denote the index set of the archetypal drivers by $\mathcal{I}_{\text{sel}}$.

Remarks. We do not select drivers with the smallest residuals. The reason is that the corresponding drivers to this criterion will be those with high estimated claims frequency but no claims. Such drivers can be potentially risky drivers, since even the *true* claims frequency is high, the chance of making no claim is also high. Our attempts show that if we chose those drivers with the smallest residuals, we would not calibrate a useful 1D CNN for the goal of improving the GLM.

4.2. A one-dimensional convolutional neural network

We add two more telematics variables, squared acceleration rates a^2 and squared angle changes Δ^2, to time series of individual trips. Those second

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 5

10 ARCHETYPAL SAFE DRIVERS.

e_i	N_i	Region_i	Age_i	r_i	$\hat{\lambda}(\boldsymbol{x}_i)^{(\text{GLM})}$
3.4438	0	Hebei	32.5	–1.4907	0.2164
3.4795	0	Hebei	39.5	–0.8265	0.1188
3.4658	0	Zhejiang	29.5	–2.1375	0.3084
3.4466	0	Zhejiang	62.0	–1.6836	0.2442
3.4466	0	Hebei	33.5	–0.7848	0.1139
3.4329	0	Shanghai	44.5	–1.5932	0.2321
3.4438	0	Hebei	30.5	–1.2430	0.1805
3.4795	0	Hebei	35.5	–0.7124	0.1024
3.4274	0	Hebei	43.5	–1.2207	0.1781
3.4603	0	Hebei	23.5	–0.9866	0.1426

moments measure the variation in acceleration and angle change, which are related to one's driving style. From the perspective of the universality approximation theorem (Hornik *et al*. 1989), although including those variables may not be necessary, doing so may improve the model fitting on relatively small samples. We normalize telematics variables v, a, Δ, a^2 and Δ^2 using the min-max normalization, that is, $z_{i,j} \in [-1, 1]^{300\times 5}$ for $i \in \mathcal{I}_{\text{sel}}, j = 1, \ldots, 500$. Note that all archetypal drivers have 500 trips. We have labeled the trips of 10 archetypal risky drivers as $R_{i,j} = 1$ and those of archetypal safe drivers as $R_{i,j} = 0$. We split all trips $(R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}, j=1:500}$ into training data $(R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}}$, $j = 1{:}300$, validation data $(R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}, j=301:400}$, and test data $(R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}}$, $j = 401{:}500$.

With R interface to `keras` and `tensorflow`, we construct a 1D CNN to classify the trips of selected archetypal drivers:

$$\phi{:}[-1, 1]^{300\times 5} \to (0, 1), z \mapsto \phi(z). \tag{7}$$

Using the `keras` code shown in Listing 1 of the supplementary material with `q1=32, q2=16, q3=8, s1=7, s2=5, s3=7, rate1=0.5, rate2=0.5`, we specify the feed-forward structure for a 1D CNN as shown in Table 6.

A 1D convolution layer `cov1` extracts `q1=32` features from a sliding window of `s1=7` s. A average pooling layer `ave1` calculates average of each feature for every `s2=5` s without overlapping. Another 1D convolution layer `cov2` further extracts `q2=16` features from a sliding window of `s3=7` s. A global average pooling layer `ave2` calculates average of each feature along the entire time series. Two dropout layers `drop1` and `drop2` are inserted before two densely connected layers `dense1` and `dense2`, respectively, to reduce overfitting. The first densely connected layer `dense1` explores interactions among `q2=16` features. The second densely connected (output) layer `dense2` explores interactions among `q3=8` features, which returns an output neuron in (0,1) via a sigmoid activation function, defined as **risk score** of trip z. If an output neuron $\phi(z)$ is closer to 1, then the corresponding trip z receives a higher risk score.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 6

THE SPECIFICATION OF 1D CNN.

Layer name	Type	Activation	Kernal size	# Kernals	Output dim.	# Params
`trips`	Input	–	–	–	$[-1, 1]^{300\times 5}$	0
`conv1`	Conv.	tanh	`s1=7`	`q1=32`	$(-1, 1)^{294\times 32}$	1152
`ave1`	Pooling	(Average)	`s2=5`	–	$(-1, 1)^{58\times 32}$	0
`cov2`	Conv.	tanh	`s3=7`	`q2=16`	$(-1, 1)^{52\times 16}$	3600
`ave2`	Global pooling	(Average)	(52)	–	$(-1, 1)^{16}$	0
`drop1`	Dropout	(`rate1=0.5`)	–	–	$(-1, 1)^{16}$	0
`dense1`	Dense	tanh	–	(`q3=8`)	$(-1, 1)^{8}$	136
`drop2`	Dropout	(`rate2=0.5`)	–	–	$(-1, 1)^{8}$	0
`dense2`	Dense	Sigmoid	–	–	(0,1)	9

parameters; binary cross-entropy loss; Adam optimizer.

The 1D CNN in Table 6 contains 4897 trainable parameters (network weights) θ. The loss function is selected as binary cross-entropy loss. The Adam optimizer is used to perform gradient decent, which leads to a local minimum of the binary cross-entropy loss w.r.t. network weights θ on training data $\boldsymbol{T}_{\text{train}} = (R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}, j=1:300}$:

$$\mathcal{L}_{\text{CNN}}(\phi, \boldsymbol{T}_{\text{train}}) = -\frac{1}{300 \times |\mathcal{I}_{\text{sel}}|} \sum_{i\in\mathcal{I}_{\text{sel}}} \sum_{j=1}^{300} R_{i,j} \log \phi(z_{i,j};\theta) + (1 - R_{i,j}) \log (1 - \phi(z_{i,j};\theta)),$$

which is equivalent to the log-likelihood related to Bernoulli distribution.

Using the `keras` code shown in Listing 2 of the supplementary material, we calibrate network weights θ for at most 1000 epoches. In each epoch, the gradient descent algorithm updates network weights for around 6000/512 times. And in each update, 512 randomly selected training samples are used to calculate the gradient. This procedure is called stochastic gradient decent method, and the training loss on all training samples does not necessarily reduce after an epoch; see the non-monotonous trend in Figure 9. We stop calibration if the validation loss on $\boldsymbol{T}_{\text{validation}} = (R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}, j=301:400}$ does not improve in a sequence of 20 consecutive epoches. The network weights $\hat{\theta}$ resulting in the lowest validation loss are then saved for the rest of modeling. The calibration shown in Figure 9 indicates that overfitting starts after 300 epoches, and the final validation accuracy is around 70%. Note that the hyperparameters of `q1, q2, q3, s1, s2, s3` and dropout rates have been tuned by monitoring validation loss.

The calibrated 1D CNN $\hat{\phi} = \phi(\cdot;\hat{\theta})$ is employed to test trips $(R_{i,j}, z_{i,j})_{i\in\mathcal{I}_{\text{sel}}, j=401:500}$. The test accuracy is around 70%, which is much better than a random guess. We compare the distribution of test risk scores $\hat{\phi}(z_{i,j})$ for potentially risky trips with $R_{i,j} = 1$ with that for potentially safe trips

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

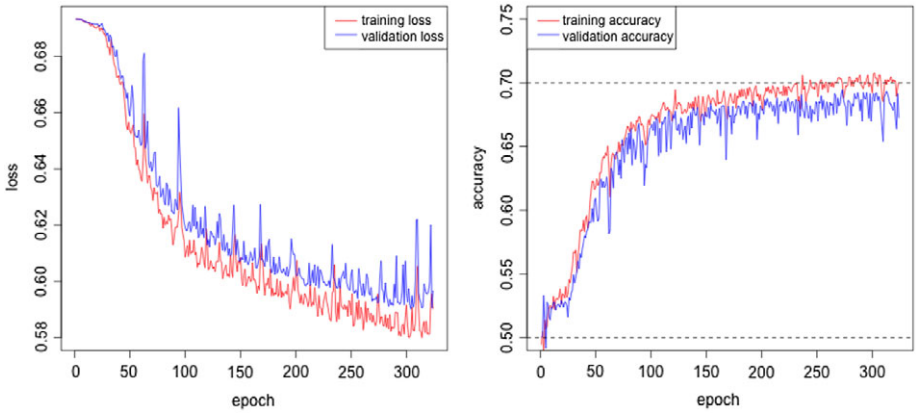


FIGURE 9: Calibration of the 1D CNN in Table 6.

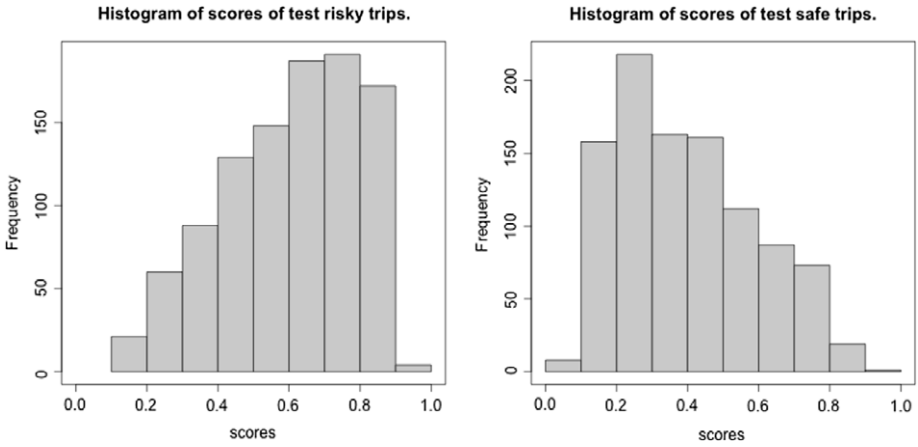


FIGURE 10: Distributions of risk scores for test trips $(z_{i,j})_{i\in\mathcal{I}_{\text{sel}},j=401:500}$.

with $R_{i,k} = 0$ in Figure 10. The test potentially risky trips tend to have higher risk scores than those of safe ones.

Remarks. From our preliminary analyses, we found that the hyperbolic tangent activation is better than the relu activation function, and the average pooling is better than the max pooling for this particular classification task. Including squared acceleration rates and squared angle changes can improve prediction accuracy of 1D CNN. Moreover, one may choose the entire 500 trips of 6/2/2 archetypal risky drivers and 6/2/2 archetypal safe drivers as training/validation/test data, that is, splitting w.r.t. drivers rather than trips. However, according to our attempts, the neural network calibration is very sensitive to the partition, that is, the validation error cannot be reduced for

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

some partitions. The reason may be that some archetypal safe drivers could still have many risky trips and vice versa.

4.3. Credibility average risk scores

The calibrated 1D CNN $\hat{\phi}$ is now employed to evaluate risk scores of c_i individual trips for each driver i. We calculate average risk scores for each driver i:

$$\bar{\phi}_i = \frac{1}{c_i} \sum_{j=1}^{c_i} \hat{\phi}(z_{i,j}). \tag{8}$$

We consider robustness of average risk scores. Given a driver i, suppose that individual trips scores $\hat{\phi}(z_{i,j})$ identically and independently follow a distribution with expectation μ_i and standard deviation σ_i. To fully credit an average risk score, it needs to be close to μ_i sufficiently with a high probability. In other words, the minimal number of trips is determined such that the following inequality holds:

$$\Pr\left(\mu_i - r\mu_i \leq \bar{\phi}_i \leq \mu_i + r\mu_i\right) \geq 1 - \alpha. \tag{9}$$

This implies that average risk score has at least $(1-\alpha)$ chance lying in a short interval $[\mu_i - r\mu_i, \mu_i + r\mu_i]$ around μ_i. As detailed in the supplementary material, the minimal number of required trips is derived as

$$c_i^{(f)}(\alpha, r) = \left(\frac{\Phi^{-1}(1-\alpha/2)}{r}\right)^2 \left(\frac{\sigma_i}{\mu_i}\right)^2, \tag{10}$$

where Φ^{-1} is the inverse of standard normal distribution function. Equation (10) is then employed as the standard of full credibility for an average risk score $\bar{\phi}_i$. If the number of individual trips does not meet this standard, we can use the **credibility average risk score** instead

$$\tilde{\phi}_i = Z_i \bar{\phi}_i + (1 - Z_i)\mu,$$

where

$$Z_i = \min\left(1, \sqrt{\frac{c_i}{c_i^{(f)}}}\right),$$

is partial credibility, and μ is the overall expectation of individual trip risk score. We select $\alpha = 10\%$, $r = 10\%$ and estimate parameters by the corresponding sample statistics:

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

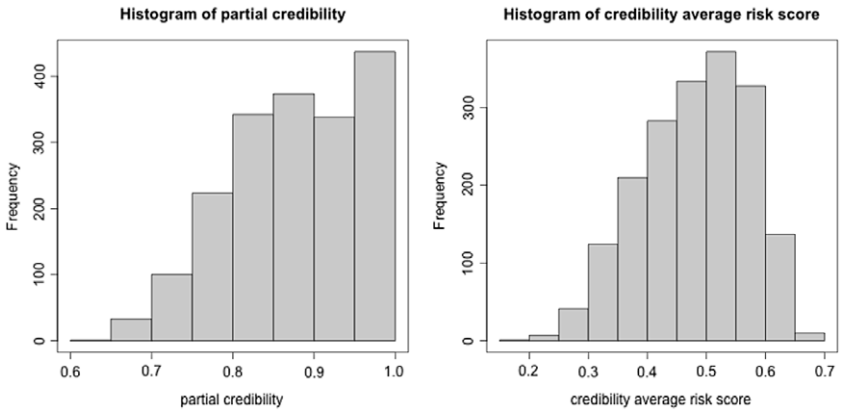


FIGURE 11: Histograms of partial credibilities and credibility average risk scores.

$$\begin{aligned}\hat{\mu}_i &= \bar{\phi}_i \\ \hat{\sigma}_i &= \frac{1}{c_i - 1}\sqrt{\sum_{j=1}^{c_i}\left(\hat{\phi}(z_{i,j}) - \bar{\phi}_i\right)^2} \\ \hat{\mu} &= \frac{\sum_{i=1}^{n} c_i \bar{\phi}_i}{\sum_{i=1}^{n} c_i}.\end{aligned} \tag{11}$$

The histograms of partial credibilities and credibility average risk scores are displayed in Figure 11. Around 200 drivers meet the standard of full credibility for average risk score, and most drivers have partial credibility larger than 0.8. Thus, for most drivers, the credibility average risk scores are mainly determined by their own driving trips.

5. IMPROVING GLM (5) WITH CREDIBILITY AVERAGE RISK SCORES

In this section, we investigate the (additional) predictive power of credibility average risk score for claims frequency prediction. By incorporating credibility average risk score, we expect to discover more heterogeneity in the portfolio and charge more fair premiums. Furthermore, we compare the proposed method with two alternative methods and conduct a sensitivity analysis.

5.1. Claims frequency modeling with credibility average risk scores

We improve the GLM (5) with credibility average risk score in another GLM as follows:

$$\ln \lambda(\boldsymbol{x}) = \ln \hat{\lambda}^{(\mathrm{GLM})}(\boldsymbol{x}) + \ln \lambda^{(\mathrm{TEL})}(\boldsymbol{x}), \tag{12}$$

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

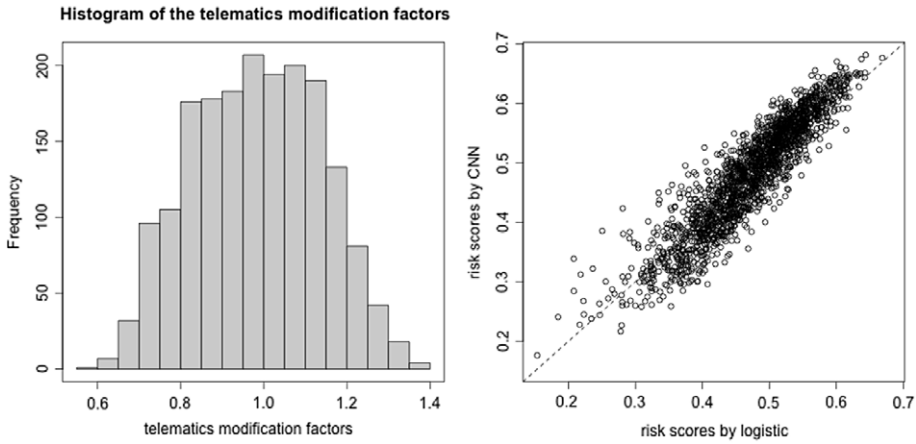


FIGURE 12: Histogram of telematics modification factors for all drivers. Comparison of credibility average risk scores for all drivers obtained from 1D CNN and logistic regression.

where $\hat{\lambda}^{(\mathrm{GLM})}(\boldsymbol{x})$ is the estimated claims frequency in GLM (5) and being treated as an offset term, and the (logged) **telematics modification factor** is

$$\ln \lambda^{(\mathrm{TEL})}(\boldsymbol{x}) = \beta_3 + \beta_4 \tilde{\phi}.$$

Note that we have expanded the covariate space $\mathcal{X}$ to include the credibility average risk score. In actuarial practices, $\hat{\lambda}^{(\mathrm{GLM})}$ is corresponding to a pure premium charged at the inception of a policy, and $\lambda^{(\mathrm{TEL})}$ is a posterior telematics modification factor based on driving experience during the policy. Similar to design of bonus-malus system, premiums of policy renewals can be based on the telematics modification factor.

It turns out that the out-of-sample Poisson deviance loss for GLM (12) is 1.0286, improved over 1.0306 for the GLM (5). The estimated telematics modification factor is

$$\hat{\lambda}^{(\mathrm{TEL})} = \exp(\hat{\beta}_3 + \hat{\beta}_4 \tilde{\phi}) = \exp(-0.8859 + 1.7795 \times \tilde{\phi}).$$

Credibility average risk scores for all drivers are in $(0.1763, 0.6810)$, so the telematics modification factors fall in $(0.5643, 1.3855)$, which is quite reasonable. We plot the histogram of telematics modification factors for all drivers in Figure 12, which implies that most drivers have a telematics modification factor between 0.8 and 1.2.

Remarks. One may consider to calibrate a Poisson 1D CNN on all drivers' trips with claims counts as response variable. However, such a calibration fails using either individual trips or jointed trips. One may consider to add credibility average risk score to GLM (5) and estimate all coefficients simultaneously, which actually leads to a similar result as that of GLM (12).

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

5.2. Competing methods

In this subsection, we first discuss one alternative to 1D CNN, then compare the proposed method with two alternative methods. All those alternative models need the input of manually engineered features of telematics variables.

5.2.1. Archetypal drivers' trips classification by a logistic model

We replace the proposed 1D CNN by a logistic model. The benefit is that the specification and calibration of logistic model are much simpler than those of 1D CNN. However, a thorough feature engineering needs to be performed. The 35 manually engineered features (covariates) include mean, median, minimum, maximum, 10% and 90% quantiles and standard deviation of speed v, acceleration a, angle change Δ, squared acceleration a^2, and squared angle change Δ^2. For each trip $\boldsymbol{z}_{i,j}$, the manually engineered features are denoted by $\boldsymbol{s}_{i,j} \in \mathbb{R}^{35}$. All of them are used in a logistic regression for trips classification of archetypal drivers $\mathcal{I}_{\text{sel}}$. The test accuracy of logistic regression is 66.7%, which is not far from that of 1D CNN 70.5%.

It may be concluded that the logistic regression also captures the major difference between potentially risky trips and potentially safe trips by utilizing adequate summary statistics. This might be explained by the fact that the chronological property in individual trips does not play a critical role in trips classification. Note that this logistic regression can be represented as a shallow dense neural network with one hidden layer (Wüthrich 2021). For this shallow dense neural network, we still need to perform a thorough feature engineering as in the logistic regression since it cannot learn complicated representations of input variables by using only one hidden layer. Employing such a neural network and contrasting it against our 1D CNN is worth consideration in future works.

5.2.2. Two alternative models to (12)

We consider two alternative models to (12), both of which use the manually engineered features $\boldsymbol{s}_{i,j}$ as discussed in Section 5.2.1.

Alternative 1: Averaged telematics summary statistics. We calculate the average of summary statistics $\boldsymbol{s}_{i,j}$ over all trips j for each driver i:

$$\boldsymbol{u}_i = \frac{1}{c_i} \sum_{j=1}^{c_i} \boldsymbol{s}_{i,j}.$$

Those averaged telematics summary statistics $\boldsymbol{u}_i \in \mathbb{R}^{35}$ are then used as covariates in the Poisson GLM for claims frequency prediction:

$$\ln \lambda(\boldsymbol{x}) = \ln \hat{\lambda}^{(\text{GLM})}(\boldsymbol{x}) + \ln \lambda^{(\text{ALT1})}(\boldsymbol{u}) = \ln \hat{\lambda}^{(\text{GLM})}(\boldsymbol{x}) + \langle \beta^{(u)}, \boldsymbol{u} \rangle, \tag{13}$$

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

TABLE 7

TEST POISSON DEVIANCE LOSS FOR DIFFERENT DATA PARTITIONS.

Test index	homo.(6)	glm (5)	Proposed (12)	alt_1 (13)	alt_2 (14)
1	1.1095	1.0981	**1.0918**	1.1071	1.0952
2	1.1003	1.0306	**1.0286**	1.0379	1.0287
3	1.0949	1.0641	1.0439	**1.0190**	1.0373
4	1.0952	1.0721	1.0654	1.0889	**1.0617**
5	1.0996	1.0318	1.0268	1.0975	**1.0262**

where $\beta^{(u)} \in \mathbb{R}^{36}$ includes an intercept. We then conduct a step-wise variable selection based on the AIC. The test Poisson deviance loss is 1.0379 which is even worse than that of the GLM (5) with classical actuarial risk factors.

Alternative 2: Risk scoring with a logistic regression. This alternative method follows the same procedure as our proposed method, but replaces the 1D CNN by a logistic regression as discussed in Section 5.2.1. We have the following alternative model:

$$\ln \lambda(\boldsymbol{x}) = \ln \hat{\lambda}^{(\mathrm{GLM})}(\boldsymbol{x}) + \ln \lambda^{(\mathrm{ALT2})}(\boldsymbol{s}) = \ln \hat{\lambda}^{(\mathrm{GLM})}(\boldsymbol{x}) + \beta_3 + \beta_4 \tilde{\phi}^{(\mathrm{ALT2})}(\boldsymbol{s}), \tag{14}$$

where $\tilde{\phi}^{(\mathrm{ALT2})}(\boldsymbol{s})$ is the credibility average risk score by the logistic regression $\hat{\phi}^{(\mathrm{ALT2})}(\boldsymbol{s})$ based on individual trips summary statistics $\boldsymbol{s}$. This alternative method also uses 35 manually engineered features $\boldsymbol{s}_{i,j}$, whereas our proposed method uses automatically engineered features by the neural network. The test Poisson deviance loss is 1.0287 for this alternative model, which is close to that of our proposed method (1.0286). Figure 12 compares credibility average risk scores for all drivers from the logistic regression and the 1D CNN with a correlation of 0.9153, which implies that both models return similar risk scores, that is, $\hat{\phi}^{(\mathrm{ALT2})}(\boldsymbol{s}_{i,j}) \approx \hat{\phi}(z_{i,j})$.

5.3. Sensitivity analysis

Since our portfolio size is relatively small, we conduct a sensitivity analysis to test the robustness of our results. Similar to cross-validation, we repeat the above analysis 5 times and evaluate test Poisson deviance loss on 5 mutually exclusive test data for the homogeneous model (6), the GLM with classical actuarial risk factors (5), our proposed model (12), and two alternative models (13) and (14), respectively. The results are shown in Table 7. Note that the above analysis corresponds to test index 2.

We conclude that the second alternative model (14) performs as well as our proposed method, and the first alternative model (13) is the least robust. Despite its simple specification and calibration, the second alternative model (14) relies on manually engineered features, which require thorough

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

considerations. In contrast, our proposed method learns the representations automatically from time series of individual trips, and thus needs less manual intervention.

6. Conclusions

In this paper, we develop a driving risk scoring 1D CNN, which is calibrated in a binary classification of individual trips for selected archetypal risky/safe drivers. The output neuron defines a risk score of individual trip in the unit range of (0,1). To address the randomness associated with average risk score for each driver, we propose to use its credibility estimate in a limited fluctuation credibility model. By incorporating credibility average risk score into the GLM for claims frequency, the out-of-sample prediction performance is significantly improved, comparing with that using only classical actuarial risk factors. This experience rating mechanism benefits both drivers (policyholders) and insurers. On one hand, it promotes safe driving. On the other hand, it reveals more heterogeneity in insurers' portfolio.

Further comparison between 1D CNN and logistic regression indicates that the chronological property in individual trips does not play a critical role in archetypal drivers' trips classification. Same as 1D CNN, the logistic regression can also capture essential features from individual trips related to claims frequency. Nevertheless, with 1D CNN, we do not need to perform a thorough manually feature engineering as in the alternative methods, which is an advantage of the proposed method.

Some future directions to extend this research may be considered. First, due to the unobserved heterogeneity, there are systematic effects in the claim counts which lead to overdispersion. We may consider to use zero-modified Poisson models such as zero-truncated Poisson model or zero-inflated Poisson model. Second, note that the logistic regression model in Section 5.2.1 can be represented as a shallow dense neural network with one hidden layer in machine learning jargon. It is worth contrasting the proposed 1D CNN against a shallow neural network. Those possible extensions remain for future works.

Acknowledgments

Guangyuan Gao gratefully acknowledges financial support from the National Natural Science Foundation of China (71901207). Shengwang Meng gratefully acknowledges financial support from the National Social Science Fund of China (16ZDA052). We particularly thank the Editor and two anonymous referees for providing valuable and insightful comments on earlier drafts. This research was also supported by Public Computing Cloud, Renmin University of China.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/asb.2021.35.

REFERENCES

AYUSO, M., GUILLEN, M. and PÉREZ-MARÍN, A.M. (2016a) Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, **4**(2), 10.

AYUSO, M., GUILLEN, M. and PÉREZ-MARÍN, A.M. (2016b) Using GPS data to analyse the distance traveled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, **68**, 160–167.

BOUCHER, J.-P., CÔTÉ, S. and GUILLEN, M. (2017) Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, **5**(4): 54.

CARVALHO, E., FERREIRA, B.V., FERREIRA, J., DE SOUZA, C., CARVALHO, H.V., SUHARA, Y., PENTLAND, A.S. and PESSIN, G. (2017) Exploiting the use of recurrent neural networks for driver behavior profiling. In 2017 International Joint Conference on Neural Networks (IJCNN), pp. 3016–3021.

CATHER, D. A. (2020) Reconsidering insurance discrimination and adverse selection in an era of data analytics. *Geneva Papers on Risk and Insurance - Issues and Practice*, **45**, 426–456.

CEVOLINI, A. and ESPOSITO, E. (2020) From pool to profile: Social consequences of algorithmic prediction in insurance. *Big Data & Society*, **7**(2).

DELONG, L., LINDHOLM, M. and WÜTHRICH, M.V. (2021) Making Tweedie's compound Poisson model more accessible. *European Actuarial Journal*, **11**, 185–226.

DENUIT, M., GUILLEN, M. and TRUFIN, J. (2019) Multivariate credibility modelling for usage-based motor insurance pricing with behavioural data. *Annals of Actuarial Science*, **13**(2), 378–399.

GAO, G., SHI, Y. and WANG, H. (2021) Telematics car driving data analytics. *The Society of Actuaries Research Institute*, pp. 1–50.

GAO, G., WANG, H. and WÜTHRICH, M.V. (2021) Boosting poisson regression models with telematics car driving data. Machine Learning.

GEYER, A., KREMSLEHNER, D. and MUERMANN, A. (2020) Asymmetric information in automobile insurance: Evidence from driving behavior. *Journal of Risk and Insurance*, **87**(4), 969–995.

GIRMA, A., YAN, X. and HOMAIFAR, A. (2019) Driver identification based on vehicle telematics data using LSTM recurrent neural network. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 894–902.

GUILLEN, M., NIELSEN, J.P., PÉREZ-MARÍN, A.M. and ELPIDOROU, V. (2020) Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal*, **24**(1), 141–152.

HO, S.-H., WONG, Y.D. and CHANG, V.W.C. (2014) Developing Singapore driving cycle for passenger cars to estimate fuel consumption and vehicular emissions. *Atmospheric Environment*, **97**, 353–362.

HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989) Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**(5), 359–366.

HU, X., ZHU, X., MA, Y.-L., CHIU, Y.-C. and TANG, Q. (2019) Advancing usage-based insurance – A contextual driving risk modelling and analysis approach. *IET Intelligent Transport Systems*, **13**(3), 453–460.

HUANG, Y. and MENG, S. (2019). Automobile insurance classification ratemaking based on telematics driving data, *Decision Support Systems*, 127.

HUNG, W. T., TONG, H. Y., LEE, C. P., HA, K. and PAO, L. Y. (2007). Development of practical driving cycle construction methodology: A case study in Hong Kong. *Transportation Research Part D: Transport and Environment*, **12**(2), 115–128.

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

JOUBERT, J.W., DE BEER, D. and DE KOKER, N. (2016) Combining accelerometer data and contextual variables to evaluate the risk of driver behaviour. *Transportation Research Part F: Traffic Psychology and Behaviour*, **4**, 80–96.

KAMBLE, S.H., MATHEW, T.V. and SHARMA, G.K. (2009) Development of real-world driving cycle: Case study of Pune, India. *Transportation Research Part D: Transport and Environment*, **14**(2), 132–140.

LECUN, Y., BENGIO, Y. and HINTON, G. (n.d.) Deep learning. *Nature*, **521**, 436–444.

LEMAIRE, J., PARK, S. C. and WANG, K. (2016) The use of annual mileage as a rating variable. *ASTIN Bulletin*, **46**(1), 39–69.

MA, Y.-L., ZHU, X., HU, X. and CHIU, Y.-C. (2018) The use of context-sensitive insurance telematics data in auto insurance rate making. *Transportation Research Part A: Policy and Practice*, **113**, 243–258.

Mahapatra, A. (2019) Privacy: A growing risk in the insurance industry. *Creative Components*, 410.

MASOOD, S., RAI, A., AGGARWAL, A., DOJA, M.N. and AHMAD, M. (2020) Detecting distraction of drivers using convolutional neural network. *Pattern Recognition Letters*, **139**, 79–85.

PAEFGEN, J., STAAKE, T. and FLEISCH, E. (2014) Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, **61**, 27–40.

RICHMAN, R. (2020) AI in actuarial science - A review of recent advances. Annals of Actuarial Science.

SAVELONAS, M., MANTZEKIS, D., LABIRIS, N., TSAKIRI, A., KARKANIS, S. and SPYROU, E. (2020) Hybrid time-series representation for the classification of driving behaviour. In 2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA), pp. 1–6.

SINGH, D. (2016) Using convolutional neural networks to perform classification on state farm insurance driver images, *Technical report, Stanford University*.

VERBELEN, R., ANTONIO, K. and CLAESKENS, G. (2018) Unraveling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **67**(5), 1275–1304.

WAHLSTRÖM, J., SKOG, I. and HÄNDEL, P. (2015) Detection of dangerous cornering in GNSS-data-driven insurance telematics. *IEEE Transactions on Intelligent Transportation Systems*, **16**(6), 3073–3083.

WAHLSTRÖM, J., SKOG, I. and HÄNDEL, P. (2017) Smartphone-based vehicle telematics: A ten-year anniversary. *IEEE Transactions on Intelligent Transportation Systems,* **18**(10), 2802–2825.

WAHLSTRÖM, J., SKOG, I., NORDSTRÖM, R.L. and HÄNDEL, P. (2018) Fusion of OBD and GNSS measurements of speed. *IEEE Transactions on Instrumentation and Measurement* **67**(7), 1659–1667.

WEIDNER, W., TRANSCHEL, F.W.G. and WEIDNER, R. (2016) Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal*, **6**(1), 3–24.

WEIDNER, W., TRANSCHEL, F.W.G. and WEIDNER, R. (2017) Telematic driving profile classification in car insurance pricing. *Annals of Actuarial Science*, **11**(2), 213–236.

WÜTHRICH, M. V. (2021) The balance property in neural network modelling. *Statistical Theory and Related Fields*, 1–9.

YANG, Y., QIAN, W. and ZOU, H. (2016) Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business and Economic Statistics*, **36**(3), 456–470.

SHENGWANG MENG
Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China, 100872
E-Mail: mengshw@ruc.edu.cn

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press

HE WANG
Department of Finance, Shenzhen Humanities & Social Sciences Key Research Bases and Ying Shang Nan Ke Actuarial Science Center, Southern University of Science and Techology, Shenzhen, China, 518055
E-Mail: wangh9@sustc.edu.cn

YANLIN SHI
Department of Actuarial Studies and Business Analytics, Macquarie University, North Ryde 2109, Australia
E-Mail: yanlin.shi@mq.edu.au

GUANGYUAN GAO (Corresponding author)
Center for Applied Statistics and School of Statistics, Renmin University of China, Beijing, China, 100872
E-Mail: guangyuan.gao@ruc.edu.cn

https://doi.org/10.1017/asb.2021.35 Published online by Cambridge University Press