Transportation Research Part A 107 (2018) 20–34

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Good drivers pay less: A study of usage-based vehicle insurance models

Yiyang Bian[a,c], Chen Yang[b,*], J. Leon Zhao[c], Liang Liang[a]

[a] School of Management, University of Science and Technology of China, Hefei, Anhui, China
[b] College of Management, Shenzhen University, Shenzhen, Guangdong, China
[c] Department of Information Systems, College of Business, City University of Hong Kong, China

ARTICLE INFO

Keywords:
Usage-based insurance
Driving behavior
Driver risk-level classification
Behavior-centric vehicle insurance pricing

ABSTRACT

Usage-based insurance (UBI) has been attracting more and more attention; however, two open research questions are how behavioral data of drivers affects driving risk and how driver behavior should affect UBI pricing schemas. This paper proposes a driver risk classification model to evaluate the risk level of drivers based on in-car sensor data. A Behavior-centric Vehicle Insurance Pricing model (BVIP) and a vehicle premium calculation prototype are developed in this paper. Based on empirical data, our research results show that BVIP achieves better accuracy in terms of risk-level classification and the prototype achieves good performance in terms of effectiveness and usability.

1. Introduction

In recent years, a noticeable trend of data-driven business service in vehicle insurance and transportation industry is Usage-Based Insurance (UBI). Numerous business opportunities and service modes are created because companies could get access to individual behavior data (Miah et al., 2017). Vehicle insurance pricing (premium) is known as the amount of money that an insurant must pay for an insurance policy of vehicle travel. Basing premiums on "how much you drive", UBI premium would be transferred from annual costs to variable charges such as miles and other driving behavior variables. It allows an insurance company to accurately target discounts at careful drivers and charge more aggressive customers an appropriately higher amount based directly on how much the vehicles are driven during the lifetime of an insurance policy (policy term). Information Technology (IT) provides facilities for collecting instantaneous driving data and calculating various driving indicators through on-board diagnostics (OBD) and online vehicle network platform (Baecke and Bocca, 2017; Baek and Jang, 2015). The real-time vehicle related parameters and driving information data can be uploaded by OBD to the vehicle network platform. Specifically, IT extends the individual vehicle insurance pricing indicators from traditional factors (e.g., age, gender and auto purchase price) to new driving factors like mileage-per-trip and driver habits. Such an approach changes the existing business model of vehicle insurance that lower-risk drivers pay less and higher-risk drivers pay more for their auto insurance (Litman, 2005).

UBI is now changing the incumbent business model of vehicle insurance. Traditional insurance and actuarial science has estimated individuals' driving risk based on driver-related personal information. Studies showed that demographic variables (such as age, gender) and personalities have significant impacts on driving risk and insurance pricing (Guo and Fang, 2013; Litman, 2005; Miyajima et al., 2007). However, driving behaviors are also powerful predictors for assessing individual driving risk. Automobile insurance companies have been trying for years to convince customers to pay premiums based on their driving behavior; however

* Corresponding author at: No. 3688 Nanhai Rd, College of Management, Shenzhen University, Shenzhen, Guangdong Province, China.
E-mail address: yangc@szu.edu.cn (C. Yang).

https://doi.org/10.1016/j.tra.2017.10.018
Received 28 May 2017; Received in revised form 14 October 2017; Accepted 30 October 2017
Available online 15 November 2017
0965-8564/

such programs are still not widely used. According to Boulton's report, people have questions like "Does the pricing model of UBI lower premiums logically?" "Will premiums increase if insurance companies know too much about them?" (Boulton, 2013). Therefore, insurers and researchers are still trying to find an appropriate path for UBI. Currently, basic usage-based premiums are calculated by dividing existing premiums by *pay as you drive* and *pay how you drive*, offered by Metromile and Progressive companies, respectively. These two insurance programs have emerged in the US, and a couple of case studies have examined these models for premiums. For instance, Desyllas and his colleagues discuss an example of how firms can profit from business model innovation using the prominent case of PAYD auto insurance (Desyllas and Sako, 2013). Unfortunately, few statistics are provided in these studies that identify factors associated with individual driving risk and help predict high-risk drivers. In prior papers, most of the chosen driving features (e.g., demographic data, mileage and time) were too general to capture driving risk. In fact, companies have been trying to collect and take advantage of big data in the changing and competitive business environment (Jukić et al., 2015). For UBI research, much of the available influential vehicle sensor data and indicators are still unemployed.

The **key research question** in this research is *how to utilize massive behavior data to offer assistance for making personalized UBI pricing strategy*. It is noted that factors of personalized driving behavior are strongly correlated with the driver's traffic accident rate. Hence, a novel behavior-centric insurance model should be developed and applied to complete UBI studies as Paefgen et al. (2013) indicated in his research.

Nevertheless, how to develop a personalized vehicle premium model according to the driving behavior data is one of the open research issues for UBI research. To fill this research gap, this research proposes a behavior-centric pricing mechanism for vehicle insurance and strives to develop novel features for UBI pricing. One important advancement of this study is to utilize supervised machine-learning approach to train the risk-level classification model with relevant sensor features and extend the existing research scope of usage-based insurance by designing a differential behavior-centric pricing mechanism based on in-car sensor data. The study also adds practical insights on UBI business via a down-to-earth demonstration. The premium calculation prototype shows potential practical value for organizations and companies to exploit UBI related business.

This study is organized as follows. In Section 2, we review the literature on user behavior variables, risk-level classification models and insuring pricing methods. A behavior-centric vehicle insurance pricing model is proposed in Section 3. Section 4 describes a prototype for vehicle premium calculation. The validity of the risk-level classification model and the prototype are evaluated in Section 5. Section 6 summarizes contributions of this study and outlines future work.

2. Literature review

2.1. Development of UBI premium strategies

Conventional vehicle insurance premium is established through an actuarial rating. Insurance companies use actuarial science to quantify the risks based on the policyholder's basic information such as type of car owned, age and gender (Azzopardi and Cortis, 2013). The conventional vehicle insurance is inefficient and inequitable because it ignores the differential driving behavior of drivers. Drivers who are similar in age, gender and automobile price may pay nearly the same premiums no matter how and how often they drive.

Usage-based insurance premium as a business pricing strategy was first introduced in 1994 at Progressive. The built-in telematics devices with GPS enabled tracking vehicle routing and emergency response. Progressive modeled premiums by combining factors of speed, location, mileage and time when driving occurred (Desyllas and Sako, 2013). The new usage-based model offered various benefits to both customers and insurers. Past studies proposed variant forms of usage-based premium options. Paefgen et al. (2013) treated mileage as a most important rating factor into the insurance premium model. It is the simplest option to implement but is constrained by the weight that can be placed on self-reported mileage estimates. Around that time, another method, *Pay-at-the-Pump* (Sugarman, 1994), funded basic insurance coverage through a surcharge (about 50 cents per gallon) on fuel sales (Litman, 2007). Efficient management of fuel sales is a critical issue in vehicle related industry (Suzuki, 2009). However, this model is less popular nowadays because the payments of this model are only based on vehicle fuel consumption and do not incorporate risk factors into the existing premium option. A premium model called *Per-Mile Premiums (PMP)* (Butler, 1993; Ferreira and Minikel, 2012) changes the unit of exposure from the vehicle year to the vehicle miles or kilometers. Drivers should pay their insurance premium based on the distance they drive. The mileage-based *PMP model* significantly improves actuarial accuracy since odometer audits provide more accurate mileage data than the self-reported methods. Prior studies showed that *PMP* has a different financial impact on different customer segments. *PMP* provides significant consumer savings, particularly to younger drivers and lower income households. *Per-Minute Premiums* is a similar approach. It uses a small electronic device to calculate the minutes of vehicle operation as the unit of exposure (Litman, 2007). The Per-Minute model allows insurance premium rates to vary by time of day. For instance, if drivers avoid peak-period travel, they can reduce their insurance premiums. So drivers can adjust their driving patterns in order to receive an extra incentive. Another method is called *GPS-Based Pricing* (Bomberg et al., 2009). It calculates insurance premiums based on when and where driving occurs. An in-car sensor installed on the vehicle could track the related rating factors. In this case, users can monitor their driving pattern using GPS data or vehicle routing systems (Santos et al., 2011), and insurers can improve the accuracy of their insurance premiums with such data.

The most recent approach is *Pay-How-You-Drive* premium. It has already been recognized as the most promising differentiated commercial vehicle insurance premium strategy in UBI (Nai et al., 2016). Because the premium is calculated based on driving patterns, the Pay-How-You-Drive vehicle premiums become more personalized. Therefore, driving behavior indicators have drawn many researchers' attention, such as speed violations (Lahrmann et al., 2012), experience (Aseervatham et al., 2016) and driving time

(Paefgen et al., 2013). These driving indicators could help the insurance industry to target driving risk classes more effectively (Hultkrantz et al., 2012). All these mechanisms show that differentiated UBI premium models with detailed behaviors represent a new vehicle insurance premium approach and influence the business.

2.2. Behavior-based variables for UBI

Since in-car sensors produce large amounts of auto-related data every trip, one problem is how to deal with the massive sensor data. Prior researchers did some works on exposing the variables that hide in sensor data as a substitute for established rate factors in insurance. For instance, researchers have proven that mileage is one of the most relevant factors for predicting accident risk (Chipman et al., 1993). Some studies indicated that increased car use results in traffic intensities that may increase accident risk (Dickerson et al., 1998). Jun et al. (2011) suggested that driving factors such as velocity and acceleration might have relationship with car accidents. Speed is another significant related variable for predicting road accidents as Dickerson et al. (1998) mentioned. Paefgen et al. (2013) combined different driving times and other variables with speed as impact factors for car insurance. Location data such as road type and driving environment may also have great effect on predicting driving risk (Aarts and Van Schagen, 2006). All these driving related variables provided us new insights for understanding driving risk and UBI.

2.3. The driving risk classification models

Driving behavior features can be leveraged to predict the driver's risk level for auto insurance application (Paefgen et al., 2013). The risk-level classification model could be trained when the data of applicant's historical behavior data and driving risk level (a label which reflects the user's accident claim frequency and claim amount) is obtained. Several classic classification models have been utilized in previous auto insurance premium studies, such as logistic regression, neural network and decision tree classifiers (Guelman, 2012).

Siordia et al. (2010) developed an automatic driving risk classification mechanism based on expert knowledge. A system proposed in this research and the system considers the three traffic safety basic elements: driver, road and vehicle. Guelman (2012) employed the Gradient Boosting classification method to predict auto accident loss cost with a real dataset obtained from a Canadian insurance company. The proposed method can train the model parameters with little data, and the experimental result has an advantage over the Generalized Linear Model approach. Guo and Fang (2013) classified drivers into three risk groups based on crash and near-crash rate using a K-mean cluster method (Guo and Fang, 2013). Fifteen location-based driving features are applied to three kinds of classification models for risk-level prediction in Paefgen's study (984 accident-free vehicles and 583 accident-involved vehicles in this case). The Supervised Neural Network method achieved the best performance for insurance cost estimation while logistic regression classification has better fitness from an actuarial view (Paefgen et al., 2013). Baecke and Bocca (2017) studied the added value of driving behavior variables in addition to traditional accident risk predictors. Specifically, they chose Logistic Regression, Random Forests, Artificial Neural Networks as main data mining techniques to classify the drivers' vehicle telematics data. The results demonstrate that the standard telematics variables significantly improve the risk assessment of drivers.

2.4. Usage based insurance pricing models

Paefgen et al. (2013) applied supervised machine-learning approaches to select meaningful prediction variables from initial sensor data and develop an auto insurance premium model (Paefgen et al., 2013). The experimental results reflect that vehicle sensor data has great potential to predict driver's insurance payment cost. Supervised models such as Logistic Regression and Neural Network have achieved good performance in cost estimation. Husnjak et al. (2015) and some other researchers proposed the data model that was used in billing for usage-based auto insurance. They presented a typical sample set of extrapolated environmental and behavioral factors. Such information could provide some indirect instructions for insurance companies. A vehicle insurance loss cost model is presented in Guelman's paper (2012), where the theory of Gradient Boosting is leveraged to estimate the loss cost as an additive model.

Several premium pricing models have been reviewed by David (2015). As illustrated in the paper, the Generalized Linear Models (GLMs) usually consist of two parts-the estimation model of claim frequency and the estimation model of claim cost. The calculation model for insurance premiums can be represented by the arithmetic product of the two mentioned components. Such insurance models inspire us to decompose the insurance pricing model into two parts-the estimation of average vehicle insurance cost per mile and the driver's mileage.

3. Behavior-centric vehicle insurance pricing model

This research presents a differential pricing mechanism for commercial vehicle insurance through examining factors affecting individual driving risk and estimating individual driving risk models based on detailed vehicle sensor data. The pricing model of usage-based vehicle insurance can leverage the key idea of GLMs, because the data mining approach can help to analyze and estimate the insurant's driving risk level and potential insurance cost, while the total premium may be positively correlated with distance or duration. A low-risk insurant who drives a long distance during the insurance period could incur high expenses.

Traditional vehicle insurance pricing model does not consider the impact of various potential driver behavior features extracted from the vehicle telematics data (Guelman, 2012; Husnjak et al., 2015). Hsu et al. found that drivers' road traffic accidents rates is

Table 1
Driving-related variables for UBI.

Category	Variables	Reference	Category	Variables	Reference
Basic information	Age	Rhodes and Pivik (2011) and Yannis et al. (2005)	Sensor-related variables	Mileage	Bailey and Simon (1960), Jun et al. (2007) and Litman (2005)
	Gender	Bailey and Simon (1960), Lonczak et al. (2007) and Rhodes and Pivik (2011)		Time	Jun et al. (2007) and Paefgen et al. (2013)
	Salary	Shinar et al. (2001) and Yannis et al. (2005)		Average speed	
	Education level	Lourens et al. (1999) and Yannis et al. (2005)		Over speed	
	Driving years (experience)	Lajunen and Summala (1995) and Yannis et al. (2005)		Rapid Acceleration	af Wåhlberg (2004) and Jun et al. (2011)
	Endorsement (violation record)	Jun et al. (2011)		Maximum deceleration	Malta et al. (2009) and Zheng et al. (2014)
	Family status	Litman (2005) and Yannis et al. (2005)		Sudden turn	Quddus et al. (2002)
	Physical status	Rhodes and Pivik (2011)			
Vehicle status	Time length of vehicle use	Beirão and Cabral (2007)	Driver-related variables	Risk preference	Guo and Fang (2013), Lajunen et al. (1998), Lajunen and Summala (1995) and Miyajima et al. (2007)
				Aggressiveness	
	Initial purchase price			Accident prediction ability	
	Intended use (private or commercial)			Assistive technology using ability	
Geospatial	Road type	Jun et al. (2007) and Paefgen et al. (2013)		Carefulness	
	Near crash location	Jun et al. (2007)		Safety awareness	
	Crash Object Type	Zheng et al. (2014)		Driving enjoyment	
	Potential crash type			Dangerous behavior while driving (Texting, dialing, intoxication, etc.)	Donovan and Marlatt (1982), Gupta et al. (2016), Nelson et al. (2009) and Nemme and White (2010)
	Triggering factors				

positive related with their the purchase of vehicle insurance (Hsu et al., 2015). It is hard to directly deduce the insurance cost from the driver behavior data, because the extracted features have different degrees of importance for the cost of vehicle insurance. Thus, to efficiently leverage the behavior data, we employ an ensemble learning-based approach to obtain the user's risk-level classification model to calculate the potential compensation payouts.

In this section, firstly, we extract various related factors as shown in Table 1. Secondly, we employ an ensemble learning-based classification approach for user's driving risk prediction. Thirdly, a pricing model is established by combining the driver's estimated risk level and behavior features.

3.1. Variables for driving risk-level classification

The key for formulating a preconceived UBI premium for drivers is to determine their driving risk level. Prior researchers proved that driving behavior activity patterns differ between crash-drivers and non-crash drivers. So the following work intend to extend our understanding of UBI by finding the most influential behavior features.

UBI considers a much broader variety and more objective variables than self-reported data. We summarized the most influential driving-related variables for predicting driving risk level in Table 1.

Incorporating appropriate features and variables into classification model is vital (Zhang et al., 2014). In Table 1, there are 5 main categories of driving-related variables. The listed variables are having a direct or indirect relationship with driving accidents. The variables of driving information and vehicle status that are essential for risk prediction have been added to vehicle insurance pricing model for several decades. Geographical information and driver-related variables are also widely discussed in predict driving risk. In this study, the behavioral variables that could be extracted from on-board diagnostics (OBD) devices or a driver's mobile phone (Sensor-related driving variables) are selected for further analysis. The behavioral variables could reflect driver's operational choices at real time in handling vehicle. These choices are directly linked to the probability of getting involved in a traffic accident (Tselentis et al., 2017). However, either of the two kinds of devices will generate a large amount of behavior data per day, per trip or even per hour. Based on prior research, we choose to extract the numeric data of 7 behavior-related variables from in-car sensors. The variables of mileage, time and average speed are three conventional but valuable indicators for UBI. Specifically, this research divides variable time into two indicators: time of day and day of week. However, more exposed driving-related data could reduce the complexity and improve the accuracy of usage-based insurance pricing in practice (Paefgen et al., 2013). In fact, these four detailed behavior variables play a more important role in predicting drivers' risk level. (Jun et al., 2011) observed that drivers who had crash experiences tended to drive at higher speed and exhibited higher tendencies of non-compliance with the posted speed limit. More frequent rapid acceleration, deceleration (braking) rate and sharp turns may increase driving risk and damage levels (Quddus et al., 2002). Thus, in this study, we try to integrate these more detailed behavior variables into the pricing model and risk-level classification model to explore their impact on UBI pricing.

3.2. The behavior-centric risk-level classification model

Driving risk can be described as the potential probability of vehicle crashes or losing something of value caused by the driver. Driving risk level is proposed to represent the harmful degree of different types of behaviors. In this study, the drivers are classified into five risk levels according to insurants' traffic accident records. Based on the insurants' driving risk level and the corresponding behavior features as described in the previous section, it is possible to explore the potential relationship between driving-related characteristics and risk results. This study employs the ensemble learning-based classification approach to train and learn the model for risk-level prediction. Based on the features, a classifier is generated for training objects into different categories.

Ensemble learning methods utilize a series of classifiers (usually belonging to one type) to solve the classification problem (Kiang, 2003; Wang et al., 2013). Bagging is one of the most useful ensemble learning techniques (Han and Kamber, 2006). The main contribution of the ensemble learning technique is combining multiple classifiers and making predictions based on the classification result votes of each classifier, which could reduce the behavior data variance in a single classifier and make use of the diversity in each of the classifiers. The idea of bagging (also known as Bootstrap Aggregating) explains how to deal with the training data using bootstrap sampling (Han and Kamber, 2006). For a given dataset that consists of n samples, a sample is randomly selected from the set and then put back into the set. After n times selection process, a new subset is built. Repeating this K times and getting K subsets for classification model building. The base classifiers train the model parameters separately, and the classification results for the same

Table 2
Bagging algorithm procedure.

Input: Training Set $D = (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$;
N is the number of samples in the original dataset and sampling set, K is the number of sampling set.
Procedure:
for $k = 1, 2, 3, \ldots, K$ **do**
generate the sampling set D_k with N samples randomly selected from the original dataset;
train a base classifier C_k on the dataset D_k
end for
Output: $C^*(x) = \underset{y}{argmax} \sum_k I(C_k(x) = y)$

sample are aggregated by a voting mechanism. The basic procedure of the bagging algorithm is described in the following table (see Table 2).

It is noted that the classifier's performance are context sensitive and the chosen methods in our study showed good performance for the given data set. This study finds an effective approach that achieved better performance than the existing approaches such as logistic regression, neural networks, and decision trees.

After employing Naive Bayes (John and Langley, 2013), Logistic Regression (Cessie and Houwelingen, 1992), SMO (Keerthi et al., 2006), and Locally Weighted Learning (LWL) (Frank et al., 2002) methods in our context, this study shows that the best base classifier for the bagging-based ensemble learning strategy is the Naive Bayes classifier (Han and Kamber, 2006; John and Langley, 2013; Yang and Davis, 2002). This method has shown its strength in training the classifier for the samples in our experiments comparing with other candidates. The Naive Bayes model originated from classical mathematical theory has stable classification efficiency and outstanding performance for small-scale data set. It can be utilized to deal with multi classification tasks such as text mining tasks (Han and Kamber, 2006) and be employed to makes perdition for the given vector set based on the value of the posterior probability estimation. We estimate the possibility that X belongs to Category C in the following equation:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{1}$$

where C_i denotes the estimated category for vector X. If the input value is continuous, we need further works to assume the continuous attribute obey Gauss distribution and conduct calculation. μ_{C_i} and σ_{C_i} denote the average value and standard deviation of category C_i respectively. The two variables can be used to predict $P(x_k|C_i)$.

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

$$P(x_k|C_i) = g(x_k,\mu_{C_i},\sigma_{C_i}) \tag{3}$$

After feature extraction and normalization, several lists of numerical values can be generated according to the several main streams of behavior-related measures in our study. Here, using the sign $r_1, r_2, \ldots, r_n$ to denote the score of the normalized value lists. So the input data for the weight learning of driving risk is Input $= < r_1, r_2, \ldots, r_n, label >$. The label is an assigned risk level that is determined by the quotient of the insurance payouts and the mileage in the records, which indicates the average insurance cost of a driver for one mile. A higher risk level means the driver would have a higher insurance cost in the future. The supervised bagging-based classification approach could perform iterations with the training data and obtain the final convergent classifier parameters.

In our study, the bagging-based classification approach serves as the risk-level prediction approach. Next, a fraction of data will be selected from the dataset as the training set to be used for classifier parameter learning and turning on the explanatory variables.

3.3. The behavior-centric vehicle insurance pricing model (BVIP Model)

To alleviate the described research issue, this section proposes a personalized *Behavior-centric Vehicle Insurance Pricing Model (BVIP model)* for UBI. The *BVIP Model* takes advantage of known vehicle insurance pricing models and charts a new path for calculating differential vehicle insurance pricing. It incorporates the predicted behavior-based risk level provided by the proposed ensemble learning driven classification model as described in Section 3.2. The insurants can be classified into different risk levels so corresponding pricing strategies are applied. Thus, when new customers intend to choose UBI as the form of their vehicle insurance, insurance companies can acquire the driver information such as driving risk level. However, the driving behavior data is usually available as the record is stored in OBD. In this way, insurance companies could classify insurants' driving behavior via proposed bagging-based classification approach in order to calculate the actual driving risk level.

Fig. 1 illustrates the development of vehicle insurance pricing models, which are more personalized and more elaborate than existing models. From conventional vehicle insurance to usage-based insurance, different factors have played a decisive role in pricing models (see Fig. 1). In UBI, a pay-as-you-drive model is said to have many shortcomings, because it is focused only on the number of driven kilometers and not on driving behavior (Kantor and Stárek, 2014). The pay-how-you-drive model is more promising because it models the driving behavior patterns efficiently and integrates the behavior variables in model making (Tselentis et al., 2017). Specifically, The *BVIP Model* combines the essential indicators in prior premium actuarial models with the detailed behavior features in order to establish a differential vehicle insurance model for insurants.

A fixed cost is generally considered in structuring an insurance premium model since once a policy is purchased. In UBI premium models, mileage is the most essential indicator. This means that usage-based vehicle insurance is sold by the vehicle-mile (or kilometer) rather than the vehicle-year (Litman, 1997). The driver's risk level is a key determiner for calculating vehicle insurance premium as it is positively correlated with the accident rate and insurance cost during the insured period (Vukina and Nestić, 2015). According to Tselentis et al. (2017)'s research, each driver could be assigned a risk level (a probability of crash involvement) based on his/her driving behavior in UBI premium model. Thus, the unit cost of each driver based on historical behavior data can be estimated. In the *BVIP* Model, the estimated average insurance cost of the driver per mile is multiplied by the mileage during the insured period to obtain the final vehicle premium. It is noted that the behavior-centric pricing model would not only use the mileage (also known as the pay-as-you-drive model) but also leverages the personalized behavior indicators to make a precise prediction of a driver's future vehicle insurance cost. The arithmetic product of the predicted risk level, an adjusted weight and the driving mileage is used to reflect the estimation of the insurance cost per year, which is shown in the following equation.

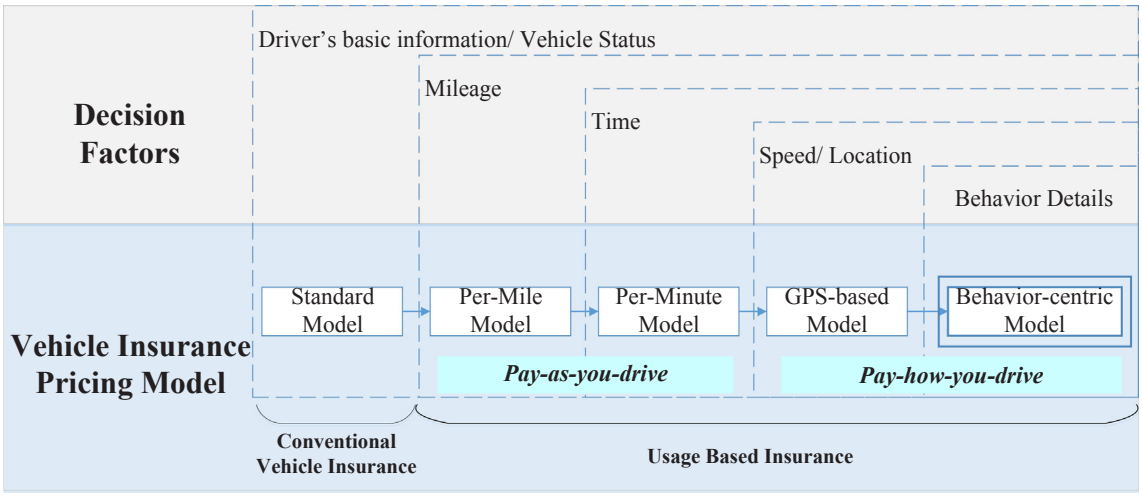


Fig. 1. Driving behavior-centric differential insurance pricing model.

$$InsurancePrice = C_0 + Risk_p * w_p * M \tag{4}$$

C_0 denotes the basic cost of a vehicle per policy term. It is given by the managers in an insurance company according to their experience. M represents the mileage of the vehicle. $Risk_p$ indicates the risk level estimated by the behavior-centric classification model. w_p is a given weight that reflects the average unit insurance cost for the corresponding risk level. With this function, driving behaviors are captured, and the personalized pricing for different drivers is calculated. The effectiveness of the premium function is verified in latter sections.

4. Prototype for behavior-centric vehicle premium calculation

In this section, a *BVIP* prototype and its design logic are presented. In order to assess the efficacy of the proposed *BVIP model*, a prototype for vehicle premium calculation is implemented. The prototype is designed to resolve a pragmatic problem (March and Smith, 1995) – personalized vehicle insurance pricing, using a design process to add and test the premium decision-making prototype system (von Alan et al., 2004).

The *Vehicle Premium Calculation Prototype* supports insurance agents' determination of personal vehicle premium by identifying insurant's driving risk level and behavior score. The prototype is built with *JavaScript* and also uses various techniques such as *requires* and *bootstrap*. Based on the *BVIP Model*, this prototype is designed to calculate behavior-centric vehicle premiums. Hence, an insurant's behavior, as one of the key factors, is included in the analysis processing. The *Vehicle Premium Calculation Prototype* interface is comprised of three components: (1) *access selector*, (2) *input boxes* and (3) *exhibit board* of calculation results.

The *access selector* helps users choose a function module for individual/agency usage. For individual users, the prototype analyzes and categorizes their behaviors, then generates a driving score and a feasible insurance fee. Users input their basic information, condition of vehicle and sum insured amount into the *input box*. Then the prototype will calculate their driving scores based on their matching behavior data extracted from their OBD device. The *exhibit board* will present the driving score and generate a feasible insurance premium for a particular user. The driving score is displayed in a radar chart located at the bottom left of the interface, reflecting the driver's behavior. A feasible premium (color coded as dark blue) is displayed at the bottom right of the *BVIP Model* as shown in Fig. 3b. For an insurance agency, the prototype creates personalized vehicle premiums for its clients and helps make vehicle insurance pricing by analyzing customer trends and insurants' driving risk-level categories.

The architecture of prototype is shown in Fig. 2. The computational formulas are based on the BVIP model. To deal with the non-personalized vehicle premium issue, we integrate driver's behavior data into the analysis process.

The calculation process includes the following steps:

(1) Inputting basic information. The input parameters include insurant ID, OBD device number, driver information and vehicle status (see Fig. 3a). Specifically, OBD device number is the serial number of the On-Board Diagnostics device.
(2) Analyzing driving behavior. If applicants have an OBD device record, the system will calculate driving score by analyzing the data stored in the device. The behavior variables include mileage, average speed, daytime/nighttime driving hours, weekday/weekend driving hours, over speed times, rapid acceleration times, maximum deceleration and sudden turn times, as shown in Fig. 3a. If the insurant does not have a valid OBD device record, this step will be skipped.
(3) Evaluating driving risk. The system will evaluate applicants based on their objective behavior records, background information and geographic data. Driver location and the geographic data are retrieved using GPS.
(4) Generating insurance pricing and driving score. Based on the *BVIP Model*, the prototype will generate a driving score and feasible insurance pricing for the insurance applicant.

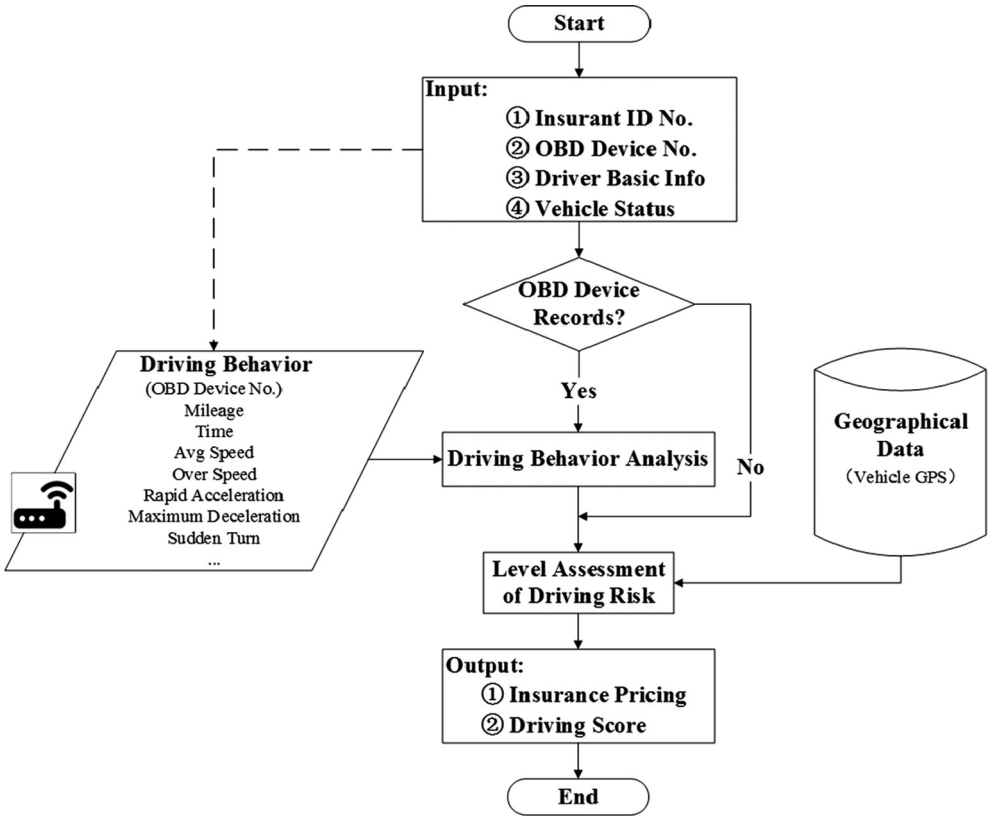


Fig. 2. Process view of vehicle premium calculator.

To better understand how the prototype works, screenshots of exemplary calculation results are exhibited in Fig. 3b. The radar chart shows the details of one insurant's behavior, whose total driving score is 69.37. The sum insured is RMB 180,000 and the final feasible insurance premium is RMB 1,928.75.

The 3D map in Fig. 4 shows the overall distribution of the observations. It presents the data storage and visualization functions of the prototype. The X-axis represents total insured amount, the Y-axis represents premium and the Z-axis shows the driving score of the insurants. With these functions, the insurance agent can see the whole picture of their clients more intuitively. The overall situation and customer trends can help insurance agents to understand their insurants' distribution and make further business strategy.

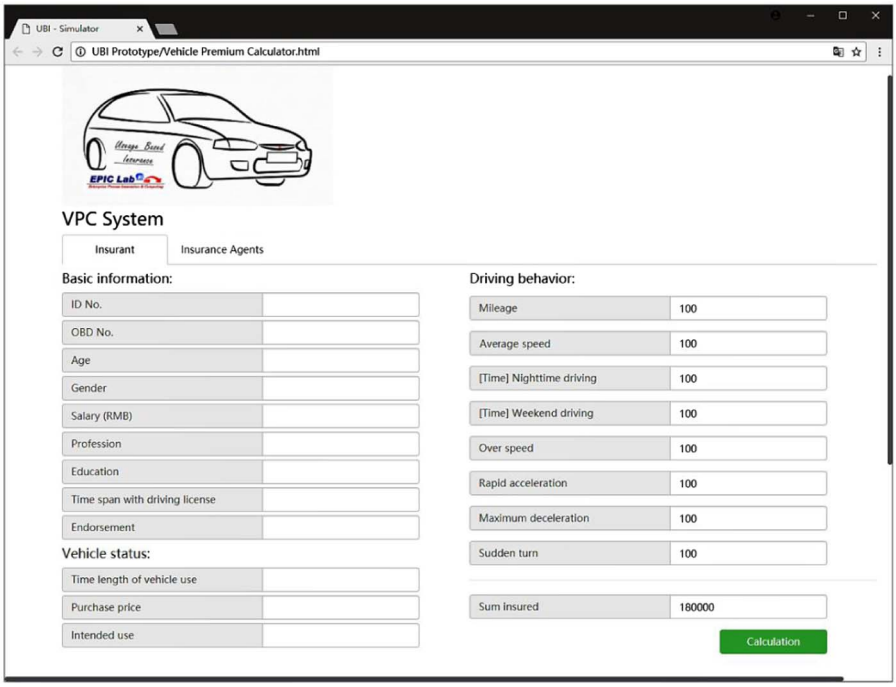
5. Validation

This section contains two parts: the behavior-centric risk-level classification model validation and prototype evaluation. Intended to detect the effectiveness of the proposed classification model, the validation procedure is conducted basing on real world data.

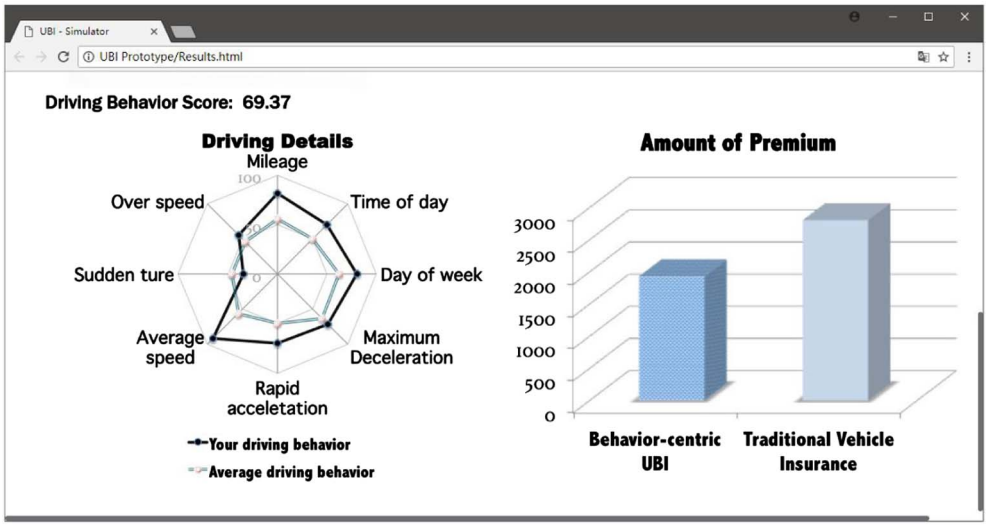
5.1. Sample selection and data processing

The dataset in our validation process contains two parts: insurants' driving accident records and matched behavior data. The behavior data were collected from an online platform (see Fig. 5) maintained by a Chinese data service company. This company has been a provider of On-Board Diagnostics (OBD) devices and has cooperated with some insurance companies for years. The OBD device is a computer-based system device that designed to monitor the performance of a vehicle engine's major components and to access the GPS information, accelerometer information, etc. During vehicle operation, behavior data (consisting of instantaneous velocity speed, ignition status of the vehicle, engine speed, acceleration, etc.) and geographic location (latitude and longitude) updates every 1 s.

The description of the total mileage (unit: kilometer) per month (MIL), nighttime driving hours per month (NDH), workday driving hours per month (WDH) and monthly average speed (AS) together with times of over speed (OST), acceleration (RA), maximum deceleration (MD) and sharp turn (ST) from the OBD data is shown in the following Table 3. Specifically, we define the driving behavior as over speed when the vehicle speed is higher than road speed limits. The over speed times is calculated by matching the engine speed data with the road information (GPS signal). The standards of the limited speed in urban, countryside and highway are different that ranging from 30 km/h to 120 km/h depends on the road type.



(a)



(b)

Fig. 3. Details of vehicle premium calculator interface.

The insurants' accident records are acquired from a Chinese insurance company located in Southern China. 206 accident records of insurants whose vehicles had an OBD device for more than 6 months are obtained. This study processes the data as follows: (1) Match the documented accident records with the particular insurant by their OBD device's serial number; eventually obtain 206 accident records. (2) Remove the records with insurance policy period of less than 3 months. (3) Match the insurants' accident data with their driving behavior data. (4) Process the missing data and errors in data recording. Finally, 198 individual observations (there are 73 accident-free drivers and 125 accident-involved drivers in the data set) with 215,736 trip records are obtained in total.

The observational data is classified into five risk levels (Level I–Level V) according to insurants' traffic accident involve times. Thus, the driving risk level function could be presented as $R(n)$, n denotes for the accident involve times of the insurant see formula (5).

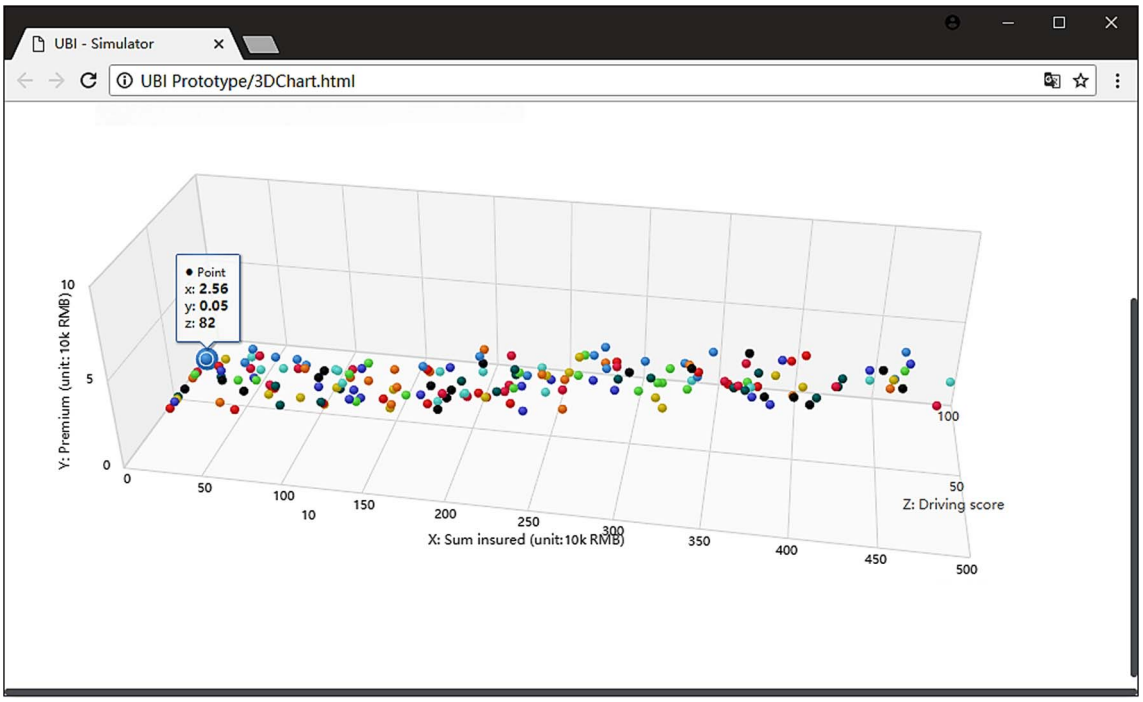


Fig. 4. Overall distributions of insurants.

Fig. 5. Real-time vehicle monitoring platform.

$$R(n) = \begin{cases} \text{I}, & n \in [0,2) \\ \text{II}, & n \in [2,4) \\ \text{III}, & n \in [4,6) \\ \text{IV}, & n \in [6,8) \\ \text{V}, & n \in [8,+\infty) \end{cases} \tag{5}$$

Table 3
Statistics description of driving related variables.

	Accident-free drivers N = 73				Accident-involved drivers N = 125			
	Mean	1st Q	2nd Q Median	3rd Q	Mean	1st Q	2nd Q Median	3rd Q
MIL	912.5	625.7	858.5	1184.0	1361.3	945.5	1194.5	1723.2
NDH	2.5	1.1	2.0	3.6	4.7	2.6	3.9	5.7
WDH	24.6	17.2	23.7	31.5	29.8	22.3	28.9	33.0
AS	40.7	27.0	40.2	54.4	52.6	41.6	47.8	57.5
OST	3.4	0.7	2.7	4.9	7.5	2.5	5.2	13.6
RA	2.8	1.1	2.0	4.0	8.6	4.4	7.2	13.1
MD	13.5	5.7	12.0	18.0	29.5	18.5	25.1	42.0
ST	9.2	3.0	6.0	11.0	21.9	9.5	21.0	27.5

5.2. Risk-level classification validation

Our evaluation of the risk-level classification consists of two stages. The first stage demonstrates the effectiveness of our bagging-based approach compared with several benchmark classification models. The second stage uses the bagging-based approach to evaluate the proposed behavior-centric classification model together with previous usage-based classification methods for UBI.

5.2.1. Evaluation of the effectiveness of the bagging-based approach

First, the comparison results of proposed bagging-based ensemble learning method and several common classification models such as Naive Bayes (John and Langley, 2013), Logistic Regression (Cessie and Houwelingen, 1992), SMO (Keerthi et al., 2006) and Locally Weighted Learning (LWL) (Frank et al., 2002) are presented. Four common classification result analysis metrics are used to evaluate the experimental results, including Correctly classified instances measure, Kappa statistic, Mean absolute error (MAE), and Root mean squared error (RMSE). Correctly classified instances percentage is an intuitive measure that represents the rate of the correctly classified instances and the total number of instances.

$$Correctly\ classified\ instances\ percentage = \frac{correctly\ classified\ instances}{total\ number\ of\ instances} \tag{6}$$

Kappa statistic is a normalized measure and is used to represent the difference degree between the classification results of the test classifier and random classification. It can be calculated by the following equation:

$$Kappastatistic = \frac{P(A)-P(E)}{1-P(E)} \tag{7}$$

where $P(A)$ indicates the percentage agreement and $P(E)$ denotes the chance agreement. The larger Kappa measure indicates the classifier prediction and the ground truth has a high consistency. Mean Absolute Error (MAE) computes the deviation between predicted ratings and actual ratings. It accumulates the absolute value of the difference between the predicted risk level and the real one, which is defined in the following equation, where p_i denotes the prediction and r_i denotes the actual rating.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |p_i - r_i| \tag{8}$$

Root Mean Square Error (RMSE) is similar to MAE, but places more emphasis on larger deviation by calculating the square root of sum of squares, which can be represented in the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (p_i - r_i)^2} \tag{9}$$

The dataset consists of multiple behavior features that were collected from the in-car sensors of 198 private and commercial vehicles in several Chinese cities. The data used in this experiment contains 8 behavior-related features and 1 risk-level label. A preliminary insurance pricing model can be established based on the predicted insurance cost (or accident risk level). Tenfold cross-validation is employed to get the robust results. Nine out of ten instances are used to train the model and the remaining one-tenth serves as the gold standard. The first stage experiment results are shown in the following table.

It is obvious that the proposed bagging-boosted classification approach achieved the best performance compared with the four commonly used benchmarks for classification analysis. Specifically, for the correctly classified instances measure and Kappa statistic, the higher value represents the better performance. And the bagging-boosted classification approach achieved 0.66 and 0.4939. For the Mean absolute error and Root mean squared error, the lower rating suggests the better performance. The scores for the bagging approach are 0.1434 and 0.3156.

The bagging-based ensemble-learning algorithm achieved the best result across all four kinds of evaluation metrics. In this study, the base classifier for the bagging strategy is the Naive Bayes classifier. A significant improvement can be obtained using the ensemble learning process with the multiple base classifiers. The second best classification model is the SMO classifier that achieves a

Table 4
Comparison of proposed model and benchmark methods.

	Correctly classified instances percentage	Kappa statistic	Mean absolute error	Root mean squared error
Locally Weighted Learning	0.55	0.3118	0.2236	0.346
Logistic	0.57	0.3765	0.179	0.3591
Naïve Bayes	0.6	0.4028	0.1516	0.353
SMO	0.63	0.4556	0.26	0.3464
Bagging_NB	0.66	0.4939	0.1434	0.3156
Improvements of SMO (second best)	3.13%	8.4%	44.85%	8.89%
P value in T-test	0.0412*[a]			

[a] For MAE and RMSE, lower means better. To facilitate statistical significance test, this study subtract the original value by 1 for the T-test. Thus the SMO values are 0.74 and 0.6536. For Bagging, the values are 0.8566 and 0.6844. In this paper '*' represents the level of significance ($p < 0.05$), '**' represents the level of significance ($p < 0.01$), and '***' represents the level of significance ($p < 0.001$).

decent score among the four baselines. To show statistical significance of proposed model and benchmarks, this process check whether there is a significant statistical difference between the results of proposed model and the second best method. As can be seen from the table, the bagging classifier improves the SMO classifier. The T-test result indicates the improvement is statistically significant at the 0.05 level.

5.2.2. Evaluation of behavior-centric classification model for UBI

At this stage, five benchmark approaches are chosen to compare with our proposed *Behavior-centric classification model* (BC model). These baselines are established separately from the mileage (MC model), average speed (SC model) and time use (TC model) aspects, following the illustration in prior section. Multifactorial driver classification (MC model) is another notable UBI model (Paefgen et al., 2013). Several classic classification models have been utilized in Paefgen's studies, which employ four usage-based features as predictors (mileage, time of day, day of the week and average velocity). This study incorporates convictive behaviors-related features (times of over speed, sudden turn, emergency brake and rapid acceleration) into the classification model. The evaluation approaches are performed with SMO and bagging methods, as these two approaches had higher performance in stage 1 (see Table 5).

The results in Table 4 show that the proposed bagging-based ensemble learning approach that leverages comprehensive behavior features achieved the best performance. The best benchmark is the one using four basic car usage-related features, as described in a previous study with the bagging model (Paefgen et al., 2013). For the two baselines with four basic features, bagging-based classification performed better than the SMO classifier once again. The results indicate that our proposed behavior-boosted risk-level approach is valid. Similarly, to ensure there is a significant improvement between the previous method and our classification model, statistics tests are performed and the results are shown in Table 5.

The behavior-centric classification model with full features improves the bagging classifier over the prior four features classification method. T-tests are performed to further show the statistical significance of our approach and the state-of-the-art approaches. The T-test statistics results also indicate the improvement is significant at the 0.05 level.

5.3. The effectiveness of prototype for insurance business

System adaptability has been noted as key characteristic of success in insurance companies (Busquets et al., 2009). A well-designed system should fit the potential users and their requirements (Vitharana et al., 2012). Thus a questionnaire was designed to evaluate the effectiveness and usability of the prototype in insurance companies. Our survey respondents consist of two groups:

Table 5
Comparison of behavior-centric model with conventional modes.

	Evaluation approach	Correctly classified instances percentage	Kappa statistic	Mean absolute error	Root mean squared error
MC model	Bagging	0.59	0.3793	0.2054	0.3335
TC model	Bagging	0.53	0.2905	0.2112	0.3499
SC model	Bagging	0.55	0.2401	0.2507	0.3597
MC model	SMO	0.56	0.3145	0.2652	0.3538
	Bagging	0.6	0.3957	0.1853	0.342
BC model	SMO	0.63	0.4556	0.26	0.3464
	Bagging	0.66	0.4939	0.1434	0.3156
Improvements of MC model (second best)	Bagging	11%	24.82%	22.61%	7.72%
P value in T-test		0.0176*[b]			

[b] To facilitate statistical significance test, this study subtract the original value by 1 for MAE and RMSE for the T-test. Thus the values for Conventional UBI driver classification are 0.8147 and 0.658. For behavior-centric classification model, the values are 0.8566 and 0.6844.

Table 6
Results of prototype evaluation.

	Mean		Std. deviation		*t*-test #	
	Users (n = 50)	Experts (n = 10)	Users	Experts	Users	Experts
Effectiveness						
1. Assist in assessing driving risks associated with insurants.	5.36	5.60	1.241	1.075	7.746***	4.707***
2. Provide an effective way to form personalized UBI pricing.	5.12	5.40	1.319	1.265	6.003***	3.500**
3. Assist in operating UBI related business effectively.	5.18	5.50	1.424	1.354	5.859***	3.503**
Usability						
4. Learning to operate the prototype would be easy for me.	5.30	5.80	1.249	1.619	7.357***	3.515**
5. My interaction with the prototype would be clear and understandable.	4.56	5.60	1.327	1.265	2.983**	4.001**
6. I find the prototype to be flexible to interact with.	4.80	5.40	1.229	1.174	4.603***	3.772**
7. The prototype's commands are self-explained and easy to understand.	4.48	5.90	1.418	1.197	2.394*	5.019***
8. I find the prototype easy to use.	4.72	6.00	1.278	1.247	3.982***	5.071***
9. The system is user friendly.	4.92	5.80	1.085	1.135	5.996***	5.014***
10. I'd like to recommend this prototype to other users.	4.90	5.50	1.129	1.354	5.635***	3.503**

* Represents the level of significance ($p < 0.05$).
** Represents the level of significance ($p < 0.01$).
*** Represents the level of significance ($p < 0.001$).

system users and domain experts. According to the evaluation procedure in (Ngai and Wat, 2005), 50 system users (20 insurants and 30 agents salesmen) and 10 domain experts (6 insurance managers and 4 university professors) are requested to fill questionnaires or take interviews after the completion of prototype implementation. Ten measurement items indicated by the seven-point Likert scale (1 = strongly disagree, 4 = undecided, 7 = strongly agree) were used to assess the perception of prototype effectiveness and usability. The evaluation results are shown in Table 6.

The mean scores of the 10 items rated by users are higher than 4 (undecided) that ranging from 4.48 to 5.36, whereas the experts rated each item from 5.40 to 6.00. A one-sample *t*-test using the neutral value "4" was conducted for the each items. Results ensured that the values of most median responses had a statistically significant difference from the neutral value at the 0.05, 0.01 and 0.001 level. Overall, the statistical results validated that the prototype could assist in assessing their driving risks and helping usage based insurance services in general.

6. Discussions

6.1. Theoretical and practical implications

The mechanics of UBI is a worthwhile research topic for transportation researchers (Paefgen et al., 2014). The technological advances of in-car sensors and information technology innovation enable insurance applicants to share more detailed driving information with an insurance company, which would inevitably affect the future development of vehicle insurance premium models. This research proposed a novel BVIP Model for personalized vehicle insurance that employs data analysis approaches to utilize the driver's behavior data. We developed a system prototype to help insurance company to seek the most appropriate business premium mode for different insurants. Specifically, this study used a bagging-based classification technique to obtain driver's risk level using features from the behavior data, which are obtained from a large real-world sample of insurants and their vehicle records. The results show the behavior-centric classification approach performs well and improve the accuracy and reliability of prior methods.

This study makes several research contributions. It is the first study that extends the existing research scope of usage-based insurance by designing a differential *BVIP Model* for UBI. First, the model employed more detailed behavior features. The effect of these features was verified to be essential for driver's risk level justification in the experiments. To obtain the premium discount, drivers should adjust their damaging aggressive driving behavior such as reducing their instances of rapid acceleration and over speeding. Second, most prior studies used a single classification method for driver risk assessment in UBI (Baecke and Bocca, 2017; Guelman, 2012). As far as we know, the proposed *BVIP Model* was the first UBI model that employed the ensemble learning method to model driving behavior of drivers. The results indicated that the ensemble-learning method is valid and achieves good performance. Third, to answer the call of Tselentis et al. (2017), the *BVIP Model* was established by combining the driver's estimated risk level and behavior features, which provides a way for estimating risk for new insurants and for regulating insurants' driving behavior by charging them more premiums.

Moreover, this study adds new insights for understanding UBI business models via a down-to-earth demonstration that has practical implications for several audiences. The aim of studying UBI is to develop a premium calculation system based on driving behavioral characteristics (Tselentis et al., 2017). However, previous UBI papers rarely talk about this practical issue. That is, this research goes one step further by not only studying the UBI premium models but also implementing a prototype of UBI premium calculation system. By means of the prototype of premium calculation system, insurants and insurance agents could acquire a more well-rounded view when developing related UBI strategies.

6.2. Research limitations and future works

This study is also subject to several limitations. First, because of the privacy issues, we couldn't get access to respondents' personal information such as their risk perception and family status. However, their personal characteristics and risk perception level may play a role in behavioral decision-making (Rhodes and Pivik, 2011). Second, this study used data of insurants located in China. In spite of this, given the imperfect insurance system in China, one would expect that these behavioral indicators might play a different role in usage-based vehicle insurance pricing calculation in other countries.

It is worth mentioning that because of the emergence of the differential vehicle premium models, many UBI-related research questions are of interest. One interesting problem is to explore the economic effects of the personalized insurance pricing models on both the insurance market and society. In the future, we'll study the adverse UBI-selection activities of drivers with negative driving records via a big data analysis approach. Specifically, when drivers' behavior records worsen, what should insurance agents do to keep their insurants paying behavior-centric premiums instead of traditional premiums? It is also interesting to include observations from other countries to contrast with the China case. Moreover, we encourage researchers to apply economic methods in the investigation of social welfare issues and business impacts of UBI on benefit sharing between insurance agencies and insurants.

Acknowledgements

This work is supported by grants from the Humanity and Social Science Youth Foundation of Ministry of Education of China [16YJC630153], National Natural Science Foundation of China [Nos.: 71701134, 71681360327 and 71471157], and Natural Science Foundation of Guangdong Province of China [2017A030310427].

References

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: a review. Accid. Anal. Prev. 38, 215–224.
af Wåhlberg, A.E., 2004. The stability of driver acceleration behavior, and a replication of its relation to bus accidents. Accid. Anal. Prev. 36, 83–92.
Aseervatham, V., Lex, C., Spindler, M., 2016. How do unisex rating regulations affect gender differences in insurance premiums? Geneva Pap. Risk Insurance Issues Pract. 41, 128–160.
Azzopardi, M., Cortis, D., 2013. Implementing automotive telematics for insurance covers of fleets. J. Technol. Manage. Innovation 8, 59–67.
Baecke, P., Bocca, L., 2017. The value of vehicle telematics data in insurance risk selection processes. Decis. Support Syst. 98, 69–79.
Baek, S.-H., Jang, J.-W., 2015. Implementation of integrated OBD-II connector with external network. Inf. Syst. 50, 69–75.
Bailey, R.A., Simon, L.J., 1960. Two studies in automobile insurance ratemaking. ASTIN Bull. 1, 192–217.
Beirão, G., Cabral, J.S., 2007. Understanding attitudes towards public transport and private car: a qualitative study. Transp. Policy 14, 478–489.
Bomberg, M., Baker, R.T., Goodin, G.D., 2009. Mileage-based user fees – a path toward implementation: phase 2: an assessment of technology issues.
Boulton, C., 2013. Auto insurers bank on big data to drive new business. Wall Street J.
Busquets, J., Rodon, J., Wareham, J., 2009. Adaptability in smart business networks: an exploratory case in the insurance industry. Decis. Support Syst. 47, 287–296.
Butler, P., 1993. Cost-based pricing of individual automobile risk transfer: car-mile exposure unit analysis. J. Actuarial Pract. 1, 51–84.
Cessie, S.L., Houwelingen, J.C.V., 1992. Ridge estimators in logistic regression. Appl. Stat. 41, 191–201.
Chipman, M.L., Macgregor, C.G., Smiley, A.M., Lee-Gosselin, M., 1993. The role of exposure in comparisons of crash risk among different drivers and driving environments. Accid. Anal. Prev. 25, 207–211.
David, M., 2015. Auto insurance premium calculation using generalized linear models. Proc. Econ. Finance 20, 147–156.
Desyllas, P., Sako, M., 2013. Profiting from business model innovation: evidence from Pay-As-You-Drive auto insurance. Res. Policy 42, 101–116.
Dickerson, A., Peirson, J., Vickerman, R., 1998. Road accidents and traffic flows: an econometric investigation. Economica 67, 101–121.
Donovan, D.M., Marlatt, G.A., 1982. Personality subtypes among driving-while-intoxicated offenders: relationship to drinking behavior and driving risk. J. Consult. Clin. Psychol. 50, 241.
Ferreira Jr., J., Minikel, E., 2012. Measuring per mile risk for pay-as-you-drive automobile insurance. Transp. Res. Rec.: J. Transp. Res. Board 97–103.
Frank, E., Hall, M., Pfahringer, B., 2002. Locally weighted naive bayes. In: Proceedings of Nineteenth Conference on Uncertainty in Artificial Intelligence, pp. 249–256.
Guelman, L., 2012. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst. Appl. 39, 3659–3667.
Guo, F., Fang, Y., 2013. Individual driver risk assessment using naturalistic driving data. Accid. Anal. Prev. 61, 3–9.
Gupta, P.B., Burns, D.J., Boyd, H., 2016. Texting while driving: an empirical investigation of students' attitudes and behaviors. Inf. Syst. Manage. 33, 88–101.
Han, J., Kamber, M., 2006. Data mining concept and techniques.
Hsu, Y.C., Shiu, Y.M., Chou, P.L., Chen, Y.M.J., 2015. Vehicle insurance and the risk of road traffic accidents. Transp. Res. Part A Policy Pract. 74, 201–209.
Hultkrantz, L., Nilsson, J.-E., Arvidsson, S., 2012. Voluntary internalization of speeding externalities with vehicle insurance. Transp. Res. Part A: Policy Pract. 46, 926–937.
Husnjak, S., Peraković, D., Forenbacher, I., Mumdziev, M., 2015. Telematics system in usage based motor insurance. Proc. Eng. 100, 816–825.
John, G.H., Langley, P., 2013. Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, pp. 338–345.
Jukić, N., Sharma, A., Nestorov, S., Jukić, B., 2015. Augmenting data warehouses with Big Data. Inf. Syst. Manage. 32, 200–209.
Jun, J., Guensler, R., Ogle, J., 2011. Differences in observed speed patterns between crash-involved and crash-not-involved drivers: application of in-vehicle monitoring technology. Transp. Res. Part C: Emerg. Technol. 19, 569–578.
Jun, J., Ogle, J., Guensler, R., 2007. Relationships between crash involvement and temporal-spatial driving behavior activity patterns: use of data for vehicles with global positioning systems. Transp. Res. Rec.: J. Transp. Res. Board 246–255.
Kantor, S., Stárek, T., 2014. Design of algorithms for payment telematics systems evaluating driver's driving style. Trans. Transp. Sci. 7, 9.
Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K., 2006. Improvements to platt's SMO algorithm for SVM classifier design. Neural Comput. 13, 637–649.
Kiang, M.Y., 2003. A comparative assessment of classification methods. Decis. Support Syst. 35, 441–454.
Lahrmann, H., Agerholm, N., Tradisauskas, N., Berthelsen, K.K., Harms, L., 2012. Pay as You Speed, ISA with incentives for not speeding: results and interpretation of speed data. Accid. Anal. Prev. 48, 17–28.
Lajunen, T., Corry, A., Summala, H., Hartley, L., 1998. Cross-cultural differences in drivers' self-assessments of their perceptual-motor and safety skills: Australians and Finns. Personality Individ. Differ. 24, 539–550.
Lajunen, T., Summala, H., 1995. Driving experience, personality, and skill and safety-motive dimensions in drivers' self-assessments. Personality Individ. Differ. 19, 307–318.
Litman, T., 1997. Distance-based vehicle insurance as a TDM strategy. Transp. Q. 51, 119–137.
Litman, T., 2005. Pay-as-you-drive pricing and insurance regulatory objectives. J. Insurance Regul. 23.

Litman, T., 2007. Distance-Based Vehicle Insurance Feasibility, Costs and Benefits. Victoria Transport Policy Institute, Victoria, BC, Canada.
Lonczak, H.S., Neighbors, C., Donovan, D.M., 2007. Predicting risky and angry driving as a function of gender. Accid. Anal. Prev. 39, 536–545.
Lourens, P.F., Vissers, J.A., Jessurun, M., 1999. Annual mileage, driving violations, and accident involvement in relation to drivers' sex, age, and level of education. Accid. Anal. Prev. 31, 593–597.
Malta, L., Miyajima, C., Takeda, K., 2009. A study of driver behavior under potential threats in vehicle traffic. IEEE Trans. Intell. Transp. Syst. 10, 201–210.
March, S.T., Smith, G.F., 1995. Design and natural science research on information technology. Decis. Support Syst. 15, 251–266.
Miah, S.J., Vu, H.Q., Gammack, J., Mcgrath, M., 2017. A big data analytics method for tourist behaviour analysis. Inf. Manage. 54, 771–785.
Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., 2007. Driver modeling based on driving behavior and its evaluation in driver identification. Proc. IEEE 95, 427–437.
Nai, W., Chen, Y., Yu, Y., Zhang, F., Dong, D., Zheng, W., 2016. Fuzzy risk mode and effect analysis based on raw driving data for pay-how-you-drive vehicle insurance. In: Proceedings of 2016 IEEE International Conference on Big Data Analysis (ICBDA), pp. 1–5.
Nelson, E., Atchley, P., Little, T.D., 2009. The effects of perception of risk and importance of answering and initiating a cellular phone call while driving. Accid. Anal. Prev. 41, 438–444.
Nemme, H.E., White, K.M., 2010. Texting while driving: psychosocial influences on young people's texting intentions and behaviour. Accid. Anal. Prev. 42, 1257–1265.
Ngai, E.W., Wat, F., 2005. Fuzzy decision support system for risk analysis in e-commerce development. Decis. Support Syst. 40, 235–255.
Paefgen, J., Staake, T., Fleisch, E., 2014. Multivariate exposure modeling of accident risk: insights from Pay-as-you-drive insurance data. Transp. Res. Part A: Policy Pract. 61, 27–40.
Paefgen, J., Staake, T., Thiesse, F., 2013. Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach. Decis. Support Syst. 56, 192–201.
Quddus, M.A., Noland, R.B., Chin, H.C., 2002. An analysis of motorcycle injury and vehicle damage severity using ordered probit models. J. Saf. Res. 33, 445–462.
Rhodes, N., Pivik, K., 2011. Age and gender differences in risky driving: the roles of positive affect and risk perception. Accid. Anal. Prev. 43, 923–931.
Santos, L., Coutinho-Rodrigues, J., Antunes, C.H., 2011. A web spatial decision support system for vehicle routing using Google Maps. Decis. Support Syst. 51, 1–9.
Shinar, D., Schechtman, E., Compton, R., 2001. Self-reports of safe driving behaviors in relationship to sex, age, education and income in the US adult driving population. Accid. Anal. Prev. 33, 111–116.
Siordia, O.S., de Diego, I.M., Conde, C., Reyes, G., Cabello, E., 2010. Driving risk classification based on experts evaluation. In: Proceedings of Intelligent Vehicles Symposium (IV), 2010 IEEE, pp. 1098–1103.
Sugarman, S.D., 1994. "Pay at the Pump" auto insurance: the vehicle injury plan (VIP) for better compensation, fairer funding, and greater safety. J. Policy Anal. Manage. 13, 363–368.
Suzuki, Y., 2009. A decision support system of dynamic vehicle refueling. Decis. Support Syst. 46, 522–531.
Tselentis, D.I., Yannis, G., Vlahogianni, E.I., 2017. Innovative motor insurance schemes: a review of current practices and emerging challenges. Accid. Anal. Prev. 98, 139–148.
Vitharana, P., Jain, H., Zahedi, F., 2012. A knowledge based component/service repository to enhance analysts' domain knowledge for requirements analysis. Inf. Manage. 49, 24–35.
von Alan, R.H., March, S.T., Park, J., Ram, S., 2004. Design science in information systems research. MIS Q. 28, 75–105.
Vukina, T., Nestić, D., 2015. Do people drive safer when accidents are more expensive: testing for moral hazard in experience rating schemes. Transp. Res. Part A Policy Pract. 71, 46–58.
Wang, G., Sun, J., Ma, J., Xu, K., Gu, J., 2013. Sentiment classification: the contribution of ensemble learning. Decis. Support Syst. 57, 77–93.
Yang, S., Davis, G.A., 2002. Bayesian estimation of classified mean daily traffic. Transp. Res. Part A Policy Pract. 36, 365–382.
Yannis, G., Kanellopoulou, A., Aggeloussi, K., Tsamboulas, D., 2005. Modelling driver choices towards accident risk reduction. Saf. Sci. 43, 173–186.
Zhang, D., Yan, Z., Jiang, H., Kim, T., 2014. A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites. Inf. Manage. 51, 845–853.
Zheng, Y., Wang, J., Li, X., Yu, C., Kodaka, K., Li, K., 2014. Driving risk assessment using cluster analysis based on naturalistic driving data. In: Proceedings of 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 2584–2589.