Personalized Vehicle Insurance: Driving Behavior Matters

Yiyang Bian
Department of Information Systems
City University of Hong Kong
Byiyang2-c@my.cityu.edu.hk

Chen Yang
College of Management
Shenzhen University
yangc0201@gmail.com

J Leon Zhao
Department of Information Systems
City University of Hong Kong
jlzhao@cityu.edu.hk

Ben Niu
College of Management
Shenzhen University
drniuben@gmail.com

Abstract

In-car sensor techniques enable insurance corporations to capture the insurant's driving behavior and correspondingly propose the personalized vehicle insurance, namely the usage-based insurance (UBI). The schema of personalized vehicle insurance needs to be designed by analyzing what factors and how the factors influence the driving risk. Though the existing literature about UBI focuses on the association of demographic and driving distance with driving risk, how individual driving behavior affects driving risk has not been investigated. As a result of this research gap, existing auto insurance pricing models are somewhat irrational because a good driver and a bad driver would pay the same premium. To alleviate this research issue, our study extends the existing research scope by collecting driving behavior data with in-car sensors and then designs a pricing model for personalized vehicle insurance. One important implication of our research is to disrupt the insurance industry via personalized auto insurance based on Internet-of-Cars and big data analytics.

1. Introduction

Basing premiums on how and how much you drive—a concept known as usage-based insurance (UBI) or "pay-as-you-drive" insurance—will allow insurance company to accurately target discounts at careful drivers, and charge more spirited customers an appropriately higher amount. (White 2012). With the development of information technology and communication technology, UBI has been used gradually (Dijksterhuis et al. 2015). UBI is based directly on how much it is driven during the policy term (Litman 2001). This insurance is done by changing the unit of exposure from traditional factors (e.g., purchase price) to new driving factors like mileage per trip and driver's habits. UBI can enable lower-risk drivers pay less and higher-risk drivers pay more for their auto insurance (Litman 2005).

The insurance and actuarial science has estimated individuals' driving risk based on various demographic variables. Studies showed that some demographic variables have significant impacts on insurance pricing. For instance, experience, age (Vlakveld 2005), gender (Lonczak et al. 2007) and family status (Litman 2005) of a vehicle driver. Beside demographic variables, driver personality is also known to play a

significant role in individual driving risk (Guo et al. 2013; Miyajima et al. 2007). Prior studies have shown the association between personality characteristics and risky driving behaviors (Hamdar et al. 2008). A recent study shows that driving behavior is a powerful predictor of individual driver risk but is hard to measure in real driving situations. According to Boulton's report, automobile insurance companies have been trying for years to convince customers to pay premiums based on their driving behavior, however the programs have still not been widely used today for some reasons. People still have questions like "Is the pricing model of UBI can lower premiums logically?" "Are the premiums will increase if insurance companies know too much about them?" (Boulton 2013) Insurers and researchers are still not found an appropriate future path for UBI. For now, basic PAYD premiums are calculated by dividing existing premiums by *pay as you drive* and *pay how you drive*, known as *Metromile* and *Progressive*[1]. These two insurance programs are widely used in US currently and a few of studies talked about these models based on some cases. For instance, Desyllas and his colleagues gave us an example of how firms can profit from business model innovation using the prominent case of PAYD auto insurance (Desyllas et al. 2013). Prior studies focused on identifying factors that associated with individual driver risk and predicting high-risk drivers using demographic and a few driving characteristic data. Seldom statistics are provided in prior papers. Most of the driving features for insurance they chose were too simple to capture the driving risk. There are still lots of valuable vehicle sensor data and data of driver personality that were not used. It is noted that the factors of personalized driving behavior are strongly correlated with the driver's traffic accident risk. The key research question is how to effectively handle a wide range of driver's behavior data (include the sensor data and driver's personality data), in order to offer assistance for the personalized premium pricing. Hence a novel behavior-based insurance model should be put forward and be applied to complete the UBI study (Paefgen et al. 2013). How to develop a personalized pricing model according to the in-car sensor data is one of the key issues for the auto insurance premium model. In this research, we will try to propose a novel pricing model for vehicle insurance and dig deeper into these driving features because these features are essential for making pricing strategy for UBI. The proposed behavior-based pricing model can extend our understanding of traditional UBI models, and provide pricing references for insurers in practice. Further more, our model can also be extended to predict potential driving risks, which is beneficial to regulate driver's behavior and reduce accident risk.

2. Theoretical Background

2.1 Development of UBI Pricing Strategies

Conventional insurance pricing is established through an actuarial rating. Insurance firms use actuarial science to quantify the risks based on the policyholder's basic information such as a type of car owned, age and gender (Azzopardi et al. 2013). A new idea of PAY-AS-YOU-GO insurance pricing model was first introduced in 1994

[1]https://en.wikipedia.org/wiki/Usage-based_insurance

at Progressive. The built-in telematics device with GPS system enabled tracking vehicle routing and emergency response. In 1998, Progressive modeled premium by combining data on speed, location, mileage, and the time of day where driving occurred (Desyllas et al. 2013). The new usage-based model offers various benefits to both customers and insurers. Past studies proposed variant forms of usage-based pricing options. Hunstad et al. (1994) mentioned that auto insurance could incorporate estimates of customers' annual mileage as rating factor into the existing price structure (Mileage Rate Factor). It is the simplest option to implement but is constrained by the weight that can be placed on self-reported mileage estimates. Vehicle drivers have a difficulty of predicting their annual mileage and tend to underestimate the mileage as well. Another method named *Pay-at-the-Pump (PATP)* (Wenzel 1995) funds basic insurance coverage through a surcharge -- about 50 cents per gallon – on fuel sales (Litman 2008). It is not actuarially accurate because payments are based on vehicle fuel consumption and does not incorporate risk factors into the existing price option. In particular, this method is one of the most attractive options in jurisdictions with high uninsured driving rate since it force all the vehicles to be insured. The third*, Per-Mile Premiums (PMP)* (Ferreira Jr et al. 2010) changes the unit of exposure from the vehicle year to the vehicle mile or vehicle kilometer. Drivers should pay their insurance premium based on the miles or kilometers they drive. Mileage-based PMP method significantly improves actuarial accuracy since odometer audits provide more accurate mileage data than the self-reported methods does. Prior studies showed that there is a different financial impact of PMP on different customer segments. PMP provides significant consumer savings, particularly to younger drivers and lower income households. The fourth method is *Per-Minute Premiums*, it uses an electronic meter to keep track of engine-operating time. Per-Minute Premiums allows insurance premium rates to vary by time of day. Vehicles drivers can adjust their driving pattern in order to receive an extra incentive. If drivers avoid their peak-period travel, they can reduce their insurance premium. *The last one is called GPS-Based Pricing* (Bomberg et al. *2009*).This method calculates insurance premium based on when and where driving occurs. A GPS (in-car sensor) installed on the vehicle could track any rating factors related to driver, vehicle, time and location of car travel. In this case, not only users can monitor their driving pattern using data from GPS but also insurers can improve an accuracy of their price with such rich dataset. More and more GPS-based pricing model emerged in UBI studies gradually.

2.2 Behavior-based Variables for UBI

Since in-car sensors produce large amounts of data data every second, one problem is *how to deal with massive data?* It is essential to identify the influential variables that hide in the massive amounts of sensor data. UBI considers a much broader variety and more objective of variables than self-reported data. Prior researchers did some meaningful works on exposing these variables as a substitute for established rate factors in insurance. According to the literature, variable of mileage should be one of the most relevant factors for predicting accident risk and be proven by researchers (Chipman et al. 1993). The Progressive Insurance Report presented a regression

analysis of relationship between mileage and insurance claims (Insurance 2005). Some studies indicated that increased car use results in heavier traffic intensities that may increase the accident risk (Dickerson et al. 2000). Jun et al. suggested that some driving factors such as velocity and acceleration might have some relationships with car accident (Jun et al. 2011). Speed is another significant related variable for predict road accidents as Dickerson et al. (2000) mentioned and it could be retrieved through in-car sensor. In Paefgen's research, they put different driving time and other variables together with speed as impact factors of car insurance (Paefgen et al. 2013). As Aarts and Schoon mentioned, some location data such as road type and driving environment may also have great effect on predicting the driving risk (Aarts et al. 2006; SCHOON et al. 2006). Moreover, some information of the automobile may also affect UBI like vehicle type and service time as illustrated in Boucher et al. (2013).

2.3 Vehicle insurance pricing model

Some researchers have applied supervised machine-learning approaches to select meaningful prediction variables from initial sensor data and develop auto insurance premium model (Paefgen et al. 2013). The experimental results reflect that vehicle sensor data has a great application potential to predict driver's insurance payment cost and the 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 the billing process of usage based auto insurance. They presented a typical sample set of extrapolated environmental and driver's behavior based factors, such as average speed, total duration and total distance. It could provide some indirect instructions for the insurance company. A vehicle insurance loss cost model is presented in Guelman (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 that 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 of insurance premium can be represented by the arithmetic product of the two mentioned components. This type of insurance models gives us a good inspiration to decompose the insurance pricing model into two parts – the estimation of average vehicle insurance cost per mile and the driver's mileage.

In this research, we try to present a pricing model for commercial vehicle insurance through examining factors affecting individual driving risk and estimating individual driving risk models based on these deeper and more detailed vehicle sensor data. In particular, we first extract various related factors such as the driver's demographic features, vehicle features, driving behavior, geographical features and driver personality as showed in Table 1. Secondly, we adopt a supervised learning model to obtain the weighting degree of these features. Thirdly, a pricing model is established based on the driver's estimated average vehicle insurance cost per mile, the driver's mileage and driving behavior features.

3. The Proposed Pricing Model for Vehicle Insurance

The pricing model of usage-based vehicle insurance can leverage the key idea of the GLMs, because the data mining approach can help to analysis and infer the insurant's driving risk and potential insurance cost, while the total premium is positively correlated with the distance or duration. Even an insurant with low risk drive a long distance in the period of insurance could also incur high expenses. Hence we employ the similar strategy in David's paper (David, 2015).

In this study, we have collected the related vehicle sensor data from a high-tech corporation named Beierjia in China. This company provides multiple driving behavior features which were collected through the in-car sensors of 150 private and commercial vehicles in several Chinese cities during 12 months. The inferred novel and original features may provide inspirations for the usage-based insurance pricing Thus we also want to employ supervised machine-learning models to train the parameters that represent the importance of the extracted sensor features and other driving behavior related features. A preliminary insurance pricing model can be established based on the predicted insurance cost (or accident risk).

The traditional definition of 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). 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 driving behavior data, we employ a supervised learning approach to obtain the weight of the various features for potential compensation payouts.

3.1 The Supervised Weight Learning Model

We want to detect the average insurance cost for different types of driving behaviors. Thus in this paper, the interest variable (Y) denotes the actual insurance costs of the driver per mile, and it is equal to the insurance payouts divided by the mileage in the records. A list of explanatory variables which indicates the driver behavior such as the driver's demographic features, vehicle features, driving behavior, the geographical features and the driver's personality are extracted from the dataset. Table 1 summarized the most influential driving variables in prior studies and some novel driving behavior features collected by our in-car sensors.

Data Category	Variables	Description
Driver information	Age	The driver's age (years)
	Gender	1, Male; 2, Female
	Salary	The salary level of driver
	Profession	The main work of driver
	Education	The education level of driver
	Time span with driving license	The time period that owning the valid driving license

	Endorsement	Traffic violation record
	Family status	No. of driver's families
	Physical status	The degree of health and weight
Vehicle status	Time length of vehicle use	Using year of the vehicle
	Purchase price	Initial purchase price of vehicle
	Intended use	Private or commercial use
Driving behavior	Mileage	Mileage per trip
	Time	Driving time per trip (when/length)
	Ave speed	Average speed per trip
	Over speed	Times of exceed the speed limit
	Acceleration	Times of acceleration
	Maximum deceleration	The maximum deceleration during the emergency braking process
	Sharp turn	Times of sharp turn
Geographical	Road type	1, Structure road; 2, Normal road; 3, Hybrid road; 4, Rural road
	Near crash location	1, Intersection; 2, Non-intersection
	Crash Object Type	1, Vehicle; 2, Single-track vehicle (motorcycle and bicycle); 3, Pedestrian
	Potential crash type	1, Rear end; 2, Conflict during intersection; 3, Jump out; 4, Opposite driving conflict; 5, Cut-in conflict
	Triggering factors	0, Non-host vehicle factors; 1, Traffic light; 2, Lane reduction; 3, Lane change; 4, Collision avoidance
Driver personality	Degree of risk preference	[1，Strongly low; 2, low; 3，Neutral; 4, High; 5, Strongly high] level of …
	Degree of aggressive	
	Degree of accident prediction ability	
	Degree of assistive technology using ability	
	Degree of carefulness	
	Degree of safety awareness	
	Degree of driving enjoyment	

Table 1 Influential Variables for UBI

As shown in Table 1, there are two main types of vehicle sensor data: numeric data and character data. In this study the numeric data includes age, time span with driving license, time length of vehicle use, purchase price, mileage, time, average speed, over

speed, times of acceleration, maximum deceleration, times of sharp turn, degree of risk preference, degree of aggressive, degree of accident prediction ability, degree of assistive technology using ability, degree of carefulness, degree of safety awareness, degree of driving enjoyment. The character data includes the features of gender, salary, profession, education, endorsement, family status, physical status, intended use, road type, near crash location, crash object type, potential crash type, triggering factors, and so on. For the character data, we need to convert them into virtual numeric variables. For example, in the gender data, male can be represented by 1 and female can be expressed by 0. To detect the effect of each feature, it is necessary to normalize the different features to the same interval. Here we adopt the well-known min–max normalization approach, and the processed values range from 0 to 1.

To learn the liner weight of these explanatory variables, a typical Support Vector Machine (SVM) based classification method is used in this paper[2]. The SVM-Rank approach is a common supervised method that is frequently employed to learn liner weights from the given training set (Joachims 2006). As traditional SVM based methods, it could train a hyper-plane to separate the different results by labels in the classification process. The difference is that, during resolving the optimization problem, the learner algorithm will sort out a list of linear ranking functions from the set of ranking functions, which could guarantee the maximization of the interest variable (Joachims 2002).

After the feature extraction and normalization process, 31 lists of numerical values can be generated according to the five main streams of driving behavior related measures in our study. Here, we use the sign r_1, r_2, …, r_{31} to denote the score of the normalized value lists. So the input data for the weight learning of driving risk is Input = $<r_1, r_2, ..., r_{31}, label>$, label is the quotient of the insurance payouts and the mileage in the records, which indicates the average insurance cost of a driver for one mile. The supervised method could make iterations with the training data, and obtain the final convergent liner parameter of the 31 lists, as W = $< w_1, w_2, ..., w_{31}>$.

Here the SVM-Rank method serves as the liner weight learning approach. In the experiment, a fraction of data could be selected from the dataset as the training set and to be used for the weight learning of the explanatory variables.

3.2 The Insurance Pricing Model

Based on the obtained prediction model of average vehicle insurance cost per mile, we can estimate the unit cost of each driver based on his/her driving historical behavior data. It is obvious that the mileage is another key issue for vehicle insurance premium and is positively correlated with the insurance cost during the insured period. Thus in our final pricing model, the mileage during the insured period is multiplied by the estimated average insurance cost of the driver per mile to obtain the final price. The product is used to reflect the estimation of the insurance cost per year, which is shown in the following equation.

$$Insurance = C_0 + (w_1 * r_1 + w_2 * r_2 + ... + w_{31} * r_{31}) * M \quad (1)$$

[2]The SVM-Rank approach and its instance can be found at http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html.

where, C_0 denotes the basic cost of a vehicle per year. M represents the mileage of the vehicle. We can leverage this simple function to capture the driver's behavior and make personalized pricing strategy for different types of drivers. In the future we will carry out experiments on the real data and test the effectiveness and reliability of our pricing model.

4. Conclusions and Future Work

Thanks to the technological advances of the in-car sensor, the drivers (insurance applicant) can share more personal information with the insurance company, which would inevitably affect the future development of insurance premium model. In this paper, we propose a novel pricing model for vehicle insurance that employs data mining approaches to analysis and utilize the driver's behavior data. Dozens of driving behavior features are extracted from the massive amounts of sensor data. In particular, driver's demographic features, vehicle features, driving behavior, the geographical features and the driver's personality are obtained. A supervised learning model is established to obtain the weighting degree of these features. Our proposed pricing model leverages the driver's estimated average vehicle insurance cost per mile and the driver's mileage to make prediction. The proposed behavior-based pricing model can extend our understanding of traditional UBI models, and provide pricing references for insurers in practice. Further more, our approach can be extended to predict potential driving risks for car driver, which is beneficial to regulate driver's behavior and reduce their driving risk. In the future, several rounds of experiments will be conducted to demonstrate the effectiveness of our proposed approach. The proposed approach will be compared with several state of the art approaches such as logistic regression, neural network, generalized linear models and decision tree classifiers models in previous auto insurance premium models (David 2015; Hamdar et al. 2008; Paefgen et al. 2013).

References

Aarts, L., and Van Schagen, I. 2006. "Driving speed and the risk of road crashes: A review," *Accident Analysis & Prevention* (38:2), pp 215-224.

Azzopardi, M., and Cortis, D. 2013. "Implementing Automotive Telematics for Insurance Covers of Fleets," *Journal of technology management & innovation* (8:4), pp 59-67.

Bomberg, M., Baker, R. T., and Goodin, G. D. 2009. "Mileage - Based User Feesker,

R. T., and Goodin, G. D. 2009. "Mileagee Telematics for Insurance Cover

Boucher, J.-P., Pérez-Marín, A. M., and Santolino, M. 2013. "Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident," *Anales del Instituto de Actuarios Españoles, 3ª Época* (19), pp 135-154.

Chipman, M. L., MacGregor, C. G., Smiley, A. M., and Lee-Gosselin, M. 1993. "The role of exposure in comparisons of crash risk among different drivers and driving environments∗," Accident Analysis & Prevention (25:2), pp 207-211.

Clint Boulton. 2013, “Auto Insurers Bank on Big Data to Drive New Business,” *The Wall Street Journal*, Feb 20, 5:03 pm ET

David, M. 2015. "Auto Insurance Premium Calculation Using Generalized Linear Models," *Procedia Economics and Finance* (20), pp 147-156.

Desyllas, P., and Sako, M. 2013. "Profiting from business model innovation: Evidence from Pay-As-You-Drive auto insurance," *Research Policy* (42:1), pp 101-116.

Dickerson, A., Peirson, J., and Vickerman, R. 2000. "Road accidents and traffic flows: an econometric investigation," *Economica* (67:265), pp 101-121.

Dijksterhuis, C., Lewis-Evans, B., Jelijs, B., de Waard, D., Brookhuis, K., and Tucha, O. 2015. "The impact of immediate or delayed feedback on driving behaviour in a simulated Pay-As-You-Drive system," *Accident Analysis & Prevention* (75), pp 93-104.

Ferreira Jr, J., and Minike, E. 2010. "Pay-As-You-Drive Auto Insurance In Massachusetts: A Risk Assessment And Report On Consumer," *Industry And Environmental Benefits, by the Department of Urban*).

Guelman, L. 2012. "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Systems with Applications* (39:3), pp 3659-3667.

Guo, F., and Fang, Y. 2013. "Individual driver risk assessment using naturalistic driving data," *Accident Analysis & Prevention* (61), pp 3-9.

Hamdar, S., Treiber, M., Mahmassani, H., and Kesting, A. 2008. "Modeling driver behavior as sequential risk-taking task," *Transportation Research Record: Journal of the Transportation Research Board*:2088), pp 208-217.

Hunstad, L., Bernstein, R., and Turem, J. S. 1994. *Impact analysis of weighting auto rating factors to comply with Proposition 103*, (Office of Policy Research, California Department of Insurance.

Husnjak, S., Peraković, D., Forenbacher, I., and Mumdziev, M. 2015. "Telematics System in Usage Based Motor Insurance," *Procedia Engineering* (100), pp 816-825.

Insurance, P. 2005. "Texas Mileage Study: Relationship Between Annual Mileage and Insurance Losses," Report.

Joachims, T. Year. "Optimizing search engines using clickthrough data," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM2002, pp. 133-142.

Joachims, T. Year. "Training linear SVMs in linear time," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM2006, pp. 217-226.

Joseph B. White. 2012. “Auto Insurance Enters the 'Pay-per-View' Era,” *The Wall Street Journal*, June 12, 7:01 p.m. ET

Jun, J., Guensler, R., and Ogle, J. 2011. "Differences in observed speed patterns between crash-involved and crash-not-involved drivers: Application of in-vehicle monitoring technology," *Transportation research part C: emerging technologies* (19:4), pp 569-578.

Litman, T. 2001. "Distance-Based Vehicle Insurance Feasibility," *Benefits and Costs:*

Comprehensive Technical Report, Victoria Transport Policy Institute (www. vtpi. org)).

Litman, T. 2005. "Pay-as-you-drive pricing and insurance regulatory objectives," *Journal of Insurance Regulation* (23:3), p 35.

Litman, T. 2008. "Distance-based vehicle insurance: feasibility, costs and benefits," *Victoria Transport Policy Institute, British Columbia, Canada. www. vtpi. org/dbvi_com. pdf. Accessed Dec* (22).

Lonczak, H. S., Neighbors, C., and Donovan, D. M. 2007. "Predicting risky and angry driving as a function of gender," *Accident Analysis & Prevention* (39:3), pp 536-545.

Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., and Itakura, F. 2007. "Driver modeling based on driving behavior and its evaluation in driver identification," *Proceedings of the IEEE* (95:2), pp 427-437.

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), pp 27-40.

Paefgen, J., Staake, T., and Thiesse, F. 2013. "Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach," *Decision Support Systems* (56), pp 192-201.

SCHOON, C., and SCHREUDERS, M. 2006. "Road safety in the Netherlands up to 2003: analysis of size, features, and development,").

Ubbels, B., and Knockaert, J. 2006. "Pay as you drive insurance: Issues affecting charge design," *Vrije Universiteit, Amsterdam*).

Vlakveld, W. P. 2005. "Jonge beginnende automobilisten, hun ongevalsrisico en maatregelen om dit terug te dringen: een literatuurstudie,").

Wenzel, T. 1995. "Analysis of national pay-as-you-drive insurance systems and other variable driving charges," Energy Analysis Program, Energy and Environment Division, Lawrence Berkeley National Laboratory, University of California.