Hindawi
Mobile Information Systems
Volume 2021, Article ID 6784026, 10 pages
https://doi.org/10.1155/2021/6784026

Research Article

Individual Driver Crash Risk Classification Based on IoV Data and Offline Consumer Behavior Data

Xuemei Zhao,[1,2] Ting Lu ,[1] and Yonghui Dai[3]

[1]*School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China*
[2]*Shandong Management University, Jinan, Shandong 250357, China*
[3]*School of Management, Shanghai University of International Business and Economics, Shanghai 201620, China*

Correspondence should be addressed to Ting Lu; luting@189.cn

Received 19 April 2021; Accepted 1 June 2021; Published 12 June 2021

Academic Editor: omar cheikhrouhou; cheikhrouhou@gmail.com

With the development of big data technologies, usage-based insurance (UBI) has received considerable attention from insurance companies. UBI products focus on identifying the relationship between the individual driver's risk and online channel behavior variables from Internet of Vehicles (IoV) data. Although omnichannel information integration has promoted the development of many industries, it has not been used to improve the accuracy of driver risk classification models in insurance industries. This paper investigates the role of combining different channel variables in improving the classification of driver's risk. Specifically, several models, including logistic regression and three different data mining techniques (neural networks, random forests, and support vector machines), augmented with driving behavior data based on the IoV and offline consumer behavior data collected from 4S (Sale, Spare part, Service, Survey) dealers, are applied to the classification model of risk. The empirical results show that the inclusion of online and offline channel data improves the different risk assessments; results also demonstrate the importance of offline consumer behavior variables in different models. These insights have important implications for insurance companies on UBI pricing strategy and cost management.

1. Introduction

Individual driving risk is characterized by substantial variation every year [1]. According to the World Health Organization, approximately 1.35 million people die in traffic accidents every year, which means nearly 3,700 people die in traffic accidents every day [2]. At the same time, because insurance is an important risk transfer tool, traffic accidents damage the benefits of insurance companies [3]. Therefore, predicting crashes and identifying the factors related to individual driving risk classification will have great value for insurance companies.

Due to limitations of data collection, early studies about crash risk classification focused on demographic variables, such as driver age and gender, and vehicle characteristics, such as vehicle age and color [4]. Recently, sensor technology has been widely adopted in the automobile industry [5]. With the development of the Internet of Things (IoT), a massive flow of online channel data reflecting driving behavior has been generated, and this provides new opportunities to classify crashes. The new opportunities offer strong business decision-making support for insurance companies, especially in relation to usage-based insurance (or user behavior insurance) (UBI) [6, 7]. Nowadays, the price of UBI products is determined by individuals' driving behavior collected from in-vehicle data records (IVDRs), which differs from the original UBI that used only vehicle usage information [8, 9].

On the other hand, with the development of new technologies such as 5G and mobile devices, omnichannel retailing has accelerated the combination of online and offline channel information. Therefore, many new business models have been developed, such as online and offline self-order, and buy online and pick up in store (BOPS). Nowadays, UBI mainly considers online rather than offline channel information in constructing various models. Can

the insurance industry benefit from using omnichannel information? Offline consumer behavior includes several psychological and personality-related features, which may have relationships with crash risk. For instance, if a driver chooses expensive auto parts, this may be an indication that they cherish their car, which could mean that the possibility of them having a car accident is relatively low. However, driving behavior variables collected from online channels do not include such information. Therefore, this paper explores whether offline consumer behavior data can improve the accuracy of crash risk classification. First, we combine prior studies and expert domain knowledge to extract more variables from the IoT data, which can reflect driving behavior. Subsequently, we obtain offline data from car dealers to construct offline consumer behavior variables. Detailed models for classifying different crash risks are run using these online and offline channel variables. Second, in line with the existing literature, we categorize the variables as basic or new, and we verify the power of the new variables.

To sum up, this study establishes a new framework that combines IoT data and offline consumer behavior data to classify different levels of crash risk. Furthermore, for the two new category variables, we find that adding offline consumer behavior variables to the basic model can significantly improve the accuracy of the crash risk classification, though the power of vehicles' turning variables is relatively weak. To our knowledge, no research to date has been conducted on offline consumer behavior in relation to the classification of driver crash risk. Importantly, then, our framework can enrich the business practice of insurance companies.

The rest of this paper is organized as follows. Section 2 presents an overview of crash risk, UBI pricing, and omnichannel information integration. Section 3 describes the data sources, including vehicle usage information, and online and offline channel variables. Our methodology, based on a combination of different data and models, is explained in Section 4. Section 5 presents the empirical results and discussion. In Section 6, the conclusions are set out.

2. Related Work

2.1. Crash Risk. Traffic accidents are one of the world's major problems, and they affect social and economic development. In order to reduce crash risk, scholars have paid considerable attention to the factors that cause vehicle collisions. Research suggests that three factors have a major impact on traffic accidents: driver behavior, environment factors, and vehicle factors [10]. Of these, driver behavior is reported as the most important factor in traffic accidents [11–13], and Singh [14] showed that driving errors are the major cause of 74% of traffic accidents. Therefore, to reduce traffic accidents, it is important to understand the effects of driver behavior on crash risk.

One of the most common methods of collecting driving behavior data to study crashes is to use driver self-reports in the form of questionnaires. De Winter and Dodou [15] found that driving errors and violations correlated negatively with age and that males reported more violations than females. Rowe et al. [13], using a 12-item version of the driver behavior questionnaire, found that different factors were linked to crashes. However, many studies have indicated that this method is not always reliable [16, 17], and a second approach is to investigate crash data. Lombardi et al. [18] analyzed 120,809 fatal accidents in the US during the period 2011–2014 and found that most drivers in fatal intersection crashes were either teenage drivers or older drivers. Li et al. [19] found that different crash types were affected by different factors. For instance, the drivers in crashes that took place on rainy days were in many cases young drivers or old drivers. Likewise, Li et al. [20] studied the relationship between land-use patterns and vehicle crashes. However, this method does not provide a full understanding of driving behavior [21]. A third method is to use driving simulators to collect driving behavior data. Choudhary and Velaga [22] analyzed distraction effects when drivers of different age groups used a mobile phone while driving in a simulator, finding that the probability of a crash increased threefold or fourfold when using a phone. Again using driving simulator experiments, Pawar et al. [23] found that the braking behavior of drivers under time pressure was affected by many factors, including approach speed and driving history. Zhao et al. [24] found that swerving without braking was not effective in avoiding collisions in critical crashes. In addition, driving simulator data have been used to research the risk assessment of autonomous vehicles [25]. However, driver behavior in the experimental environment differs from naturalistic driving [26, 27].

The remaining method is to use naturalistic driving data to explore driving risk. One of the earliest studies used 100 instrumented camera cars to collect driving data in the US [28]. Important findings were established, namely, that inattention was the main factor in about 93% of traffic accidents and that sleepiness was associated with 12% of all crashes [28]. Using the driving data of vehicles equipped with cameras, a novel approach based on interrelationships of different maneuver states was proposed to identify different levels of driving risk [29].

The Internet of Vehicles (IoV) is a special domain in IoT that is now becoming a popular platform with various kinds of vehicle information. With the development of the IoV, many researchers have used global positioning system (GPS) data as a proxy for driving behavior data to study crash risk in terms of location data [30]. For example, Toledo et al. [31] used speeding variables collected from GPS devices as an indicator of crash risk. Ellison et al. [30] used GPS to collect acceleration and location data to analyze drivers' safety. Onboard diagnostic (OBD) loggers have also been used to collect driving behavior information such as speed, braking, and acceleration, and this information tends to be more reliable than GPS data [32, 33]. Cao et al. [34] used driving behavior data (including deceleration, braking, location, and acceleration) collected from OBD loggers, combined with cluster techniques, to forecast crash events.

We found that many studies have begun to explore crash risk based on IoV data, with the most common variables

being based on speed and acceleration. However, few studies have considered vehicles' turning variables, which are also correlated with crash risk [24]. Therefore, one of this study's aims is to explore crash risk using more IoT information.

2.2. UBI Pricing. Insurance companies want to differentiate their product prices by identifying a driver's risk [16]. Therefore, predicting the driver's risk is one of the insurance companies' core tasks [35]. Traditional car insurance pricing includes vehicle information and driver characteristics, such as car color and driver gender, but driving behavior is the most important factor in driver's risk. Accordingly, UBI pricing research on risk classification has recently attracted much attention.

Ma et al. [36] incorporated real-time GPS vehicle trajectories and accident data into their generalized linear model to explore the probability of auto accidents. Their GPS data included unique contextual-based risk measurements, which differ from traditional UBI factors. Paefgen et al. [37, 38] subdivided different kinds of vehicle mileage data based on GPS from 1,600 cars, including driving speed and time of day, to explore novel UBI premium models and to predict accidents. In their pioneering work on the application of machine learning to UBI pricing, they concluded that logistic regression (LR) models are more suitable than neural network (NN) models for that purpose [37]. However, in the classification of accidents, NN models showed the best performance. Baecke and Bocca [39] used telematics data (total distance, total trip time, location distance, daytime distance, and crash information) and rush hour trip times to assess driving risk for insurance. However, the drivers in their data set were all under the age of 30. Huang and Meng [40] collected 30 driving behavior variables derived from telematics data, combining different models to classify risk and predict claim frequency. They verified the potential of driving behavior variables and machine learning in the UBI context. By examining driving data from a naturalistic driving experiment, new unsupervised algorithms were applied to different driving patterns, which could help companies to develop reasonable strategies for UBI [41].

Although previous studies have described the kinds of telematics data available for UBI, further combinations with other types of variables remain to be investigated. Offline consumer behavior is important to companies, as this can reflect, to a certain degree, consumer psychological states or personality [42]. For example, consumers who buy high-quality auto parts are likely to take their cars seriously, which may reduce the probability of risk happening. In addition, some information that is omitted from online channel information, such as telematics data, likely exists in offline consumer behavior. Guo and Fang [43] used logistic regression to analyze the relationship between driver personality and driving accidents and verified that data on characteristics were closely connected to the prediction of vehicle collisions. Similarly, our study explores the power of offline consumer behavior in UBI pricing and is, to our knowledge, the first to do so.

2.3. Omnichannel Information Integration. New technologies, such as 5G, smart mobile devices, and websites, are widely used in various business scenarios, which promotes the development of omnichannel information integration. Such integration can provide customers with an enhanced shopping experience [44, 45]. Recent studies on omnichannel information integration have mainly focused on the retail and restaurant industries. In relation to the retail, Gallino and Moreno [46] found that when an online store offered inventory information about brick-and-mortar (B&M) stores and allowed customers to BOPS, sales at B&M stores increased, while sales via the online store decreased. Gao and Su [47] derived similar results using an analytical model. In addition, they found that a decentralized retail system might be more effective for increasing revenue from BOPS. Li et al. [48] found that offline channels offered advantages when online review information was integrated into those offline channels. In the restaurant industry, Gao and Su [49] found that when online and offline order information was integrated, restaurants could reduce customers' wait time and increase demand.

However, less attention has been paid to proposing new business models that entail integrating different channel information in the insurance industry. UBI is a very large market, but the lack of effective analysis tools has resulted in many lost opportunities for insurance companies. Therefore, in this study we aim to develop new business models by integrating various channel information, such as IoV data and offline consumer behavior data, to improve the UBI industry.

3. Data Description

The data set used in this paper came from a Chinese automobile company. To protect customer privacy and commercial secrets, the data were desensitized. All vehicles used in this study were equipped with OBD loggers, and the area of activity of the vehicles was Shanghai, one of the four biggest cities in China.

We collected collision samples from January 1, 2019, to March 31, 2020, and then collected the corresponding vehicle data from the time of the collision to a prior year. To assess driver's risk precisely, we categorized collisions as severe or general, unlike previous studies, which made no such distinction [40]. A severe collision is defined as one in which airbags were deployed during the crash, which can be detected using IoV information. A general collision is defined as a slight collision during which airbags were not deployed, such as a minor vehicle scratch. We confirmed our categorization with the corresponding maintenance information, including the use of collision-related accessories and repair charges lower than RMB 10,000. According to the positive/negative sample ratio of one to five, we sampled collision-free vehicles at random between January 1, 2019, and March 31, 2020, and we took one year forward as the data sampling time. We failed to get IoV data of some samples and deleted that samples. Table 1 shows the final numbers of different collision vehicles. Our sample is larger than those of many previous similar studies, which enhances the reliability of our findings.

Table 1: Collisions by type.

Incident type	Number
Collision-free	22542
Severe collision	811
General collision	3369

Table 2: Vehicle information of different samples.

Vehicle characteristic	Mean		
	Collision-free	Severe collision	General collision
Price (CNY)	233158.318	246136.650	228163.719
Age (years)	2.062	2.453	2.106

The independent variables are multisource data. Basic vehicle information is widely used in crash risk studies [40]. The information we used, taken from sales records, includes car price and age. Table 2 presents the vehicle information and descriptive statistics.

The driving behavior variables were collected from GPS data and OBD-based data. GPS usually collects trajectory data, including time, longitude, latitude, speed, and direction, once every 15 seconds. The OBD information used in this study concerns acceleration and deceleration. In order to reduce storage pressure, the OBD adopts a trigger collection rule; that is, it starts to collect data second-by-second when the acceleration reaches a threshold, and it stops collecting when the acceleration drops below the threshold. From these online real-time data, we can obtain information about driving behavior that allows us to assess certain driving risks.

These online channel data include three types: speed, acceleration, and vehicle's turning. First, speed can reflect driving habits and road conditions. Both high and low vehicle speeds may cause collision accidents. Therefore, speed is widely used in crash risk research. In addition to the averages and quantiles used in prior studies, our speed data include speed segment proportions. Second, acceleration variables mainly reflect the behavior of stepping on the accelerator pedal or the brake pedal. Individuals that have good driving habits typically maintain a safe distance from the vehicle in front and always pay attention to the surrounding environment, which can effectively reduce their crash risk. Most studies have used absolute values instead of combining acceleration with mileage and severity. However, acceleration data in combination with mileage and severity data can better reflect driving behavior, and different acceleration behaviors cause different types of collisions. Therefore, in this paper we include more acceleration variables. Finally, many accidents are related to vehicles' turns. Compared with driving on a straight road, drivers need to consider pedestrians and turning lanes when turning. For example, high-speed turns may make the vehicle body unstable and cause an accident. Prior studies have found that swerving relates to driving behavior, but few studies have characterized this variable in detail [24, 40]. We therefore include the turning-related features of different types of turns per 100 kilometers, percentage of different turns, and speeds of different turns.

Table 3 presents an overview of the driving behavior variables. A relationship between driving behavior and different types of collisions is evident; for instance, the average speed of vehicles in collisions is lower than that of vehicles in collision-free incidents. However, the proportion of vehicles in severe collisions with speed values greater than 90 is higher than for the other two types of collision. In terms of acceleration behavior, the frequency of severe acceleration/deceleration per 100 km is the lowest for vehicles involved in collision-free incidents. The proportion of right turns in severe collisions is relatively low, and the proportion of turning around is relatively high.

For the offline channel information, this study uses car maintenance behavior in B&M stores. These consumer behavior data can reflect driver personality and behavior, which may play a role in predicting driver risk. Good vehicle maintenance behavior may keep vehicles in good condition and avoid traffic accidents caused by vehicle problems. Similarly, high-quality maintenance accessories may reflect the owner's concern about their vehicle, which can decrease the probability of them being involved in car collisions. Table 4 shows all the consumer behavior variables and their descriptive statistics. Those in the collision-free sample preferred to go to 4S (Sale, Spare part, Service, Survey) dealer shops for maintenance, while those in severe collisions used other types of shops. The average maintenance interval in miles was largest among vehicles that were involved in severe collisions versus the other types of collisions, although the interval in days in that group was shortest. Therefore, models for classifying driver's risk may achieve better performance when supplemented with consumer behavior data.

4. Methodology

4.1. Data Binning. Figure 1 presents the structure of our research. Data binning is a common preprocessing method that can reduce the risk of overfitting. Data binning methods fall into two categories: supervised and unsupervised. Supervised binning considers the value of the dependent variable when binning and can achieve the minimum entropy after binning. This method, combined with the dependent variable when binning, improves prediction accuracy, whereas unsupervised binning does not offer the same advantage. Therefore, we choose supervised binning, and we employ the classic decision trees binning method, which sorts the variables into ascending order and then calculates the average between two adjacent variables. After that, it chooses the largest Gini value from all the averages as the threshold for dividing the variables, and it iterates until the termination criterion is reached. Unlike previous studies that used continuous or discrete variables, we use both types of data with different models, which helps us to understand the relationship between different behaviors and predictions of driver's risk [50, 51].

We choose 70% of the data set at random as the training set, and the remaining 30% constitute the test set. We use the training set to make sure that the nodes of different variables are divided uniformly into six bins. The test set is divided into bins according to the nodes.

Table 3: Driving behavior variables and descriptive statistics of different samples.

Variable type	Predictor variable	Mean		
		Collision-free	Severe collision	General collision
Speed specific	Average recorded speed (km/h)	26.869	25.544	25.323
	1/4 quantile of recorded speed (km/h)	0.329	0.316	0.293
	Median quantile of recorded speed (km/h)	17.541	16.991	16.294
	3/4 quantile of recorded speed (km/h)	46.433	44.159	43.792
	Maximum recorded speed (km/h)	131.850	133.457	130.390
	Standard deviation of recorded speed (km/h)	39.554	37.497	37.424
	Proportion of idling speed	0.324	0.342	0.329
	Proportion of recorded speed between 0 and 15 km/h	0.157	0.145	0.161
	Proportion of recorded speed between 15 and 30 km/h	0.146	0.140	0.149
	Proportion of recorded speed between 30 and 60 km/h	0.218	0.211	0.213
	Proportion of recorded speed between 60 and 90 km/h	0.106	0.109	0.106
	Proportion of recorded speed between 90 and 120 km/h	0.047	0.050	0.040
	Proportion of recorded speed above 120 km/h	0.002	0.003	0.002
Acceleration specific	Frequency of acceleration/deceleration per 100 km	233.894	240.643	277.220
	Frequency of mild acceleration/deceleration per 100 km	148.843	149.734	174.904
	Frequency of moderate acceleration/deceleration per 100 km	71.046	74.322	85.141
	Frequency of severe acceleration/deceleration per 100 km	14.005	16.587	17.175
	Proportion of mild acceleration/deceleration	0.651	0.616	0.646
	Proportion of moderate acceleration/deceleration	0.294	0.306	0.297
	Proportion of severe acceleration/deceleration	0.055	0.079	0.057
Turning specific	Proportion of right turns	0.719	0.688	0.713
	Proportion of left turns	0.212	0.202	0.214
	Proportion of turning around	0.069	0.111	0.073
	Average speed of right turns	17.709	17.821	17.638
	Standard deviation of right-turn speed	13.210	13.457	13.276
	Average speed of left turns	17.915	17.940	17.901
	Standard deviation of left-turn speed	13.165	13.415	13.226
	Average speed of turning around	14.094	12.559	14.088
	Standard deviation of turning around speed	10.882	10.724	10.892
	Frequency of left turns per 100 km	12.637	11.446	12.874
	Frequency of right turns per 100 km	42.833	39.412	42.793
	Frequency of turning around per 100 km	4.347	4.575	4.342

Table 4: Consumer behavior variables and descriptive statistics of different samples.

Consumer behavior	Mean		
	Collision-free	Severe collision	General collision
Proportion of maintenance carried out in 4S dealer shops	0.780	0.462	0.624
Average price of maintenance carried out in 4S dealer shops	587.723	416.948	836.425
Average maintenance interval (days)	196.386	168.579	205.520
Average maintenance interval (miles)	6216.902	7503.283	7373.752
Average number of products used in one maintenance episode	1.321	1.166	1.884

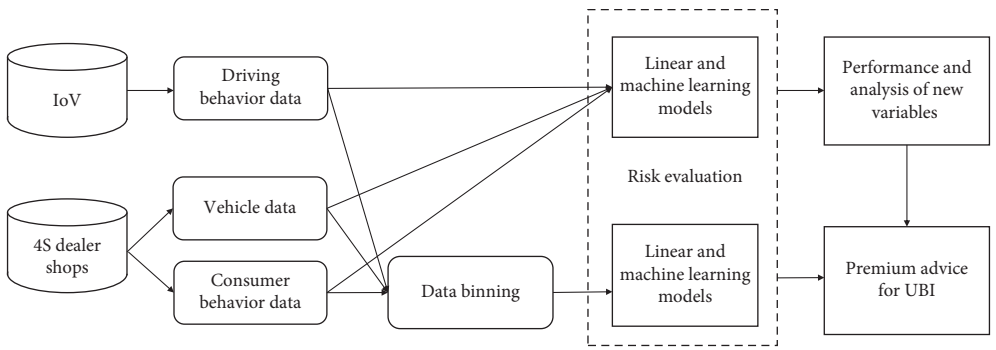

Figure 1: Individual driver crash risk classification process.

4.2. Models of Driver's Risk. In UBI research, LR models are widely used to estimate car crashes because of their high stability [37, 39, 40]. To classify different risks precisely, we first distinguish collision-free and collision incidents and then severe and general collisions. LR is a well-known data mining method for classification problems [52], and its stability is an important requirement of insurance companies. With this method, the prediction model of driver's risk is as follows:

$$\frac{\Pr\left(y_i = 1|X_i\right) = \exp\left(\beta x_i^T\right)}{\left(1 + \exp\left(\beta x_i^T\right)\right)}, \quad (1)$$

where X_i represents a vector containing values of selected independent variables for customer i, and β is a vector of corresponding coefficients.

Although the linear model performs well in some situations, companies sometimes prefer machine learning models because of their accurate performance in classification. In this study, we choose NN, RF, and support vector machine (SVM) models to construct nonlinear models. These models have been used widely in previous studies of driver's risk [53, 54]. First, for the NN model, we choose a feedforward NN consisting of an input layer, a hidden layer, and an output layer. The neurons between different layers are fully connected. The number of neurons in the input layer equals the number of dependent variables, and the number of neurons in the output layer equals 1. The number of neurons in the hidden layer is set to 10.

RF consists of an ensemble of decision trees and is a powerful machine learning method for prediction and classification [55]. RF selects feature variables at random and samples a large number of trees by bootstrapping, which can compensate for the limitations of single decision trees. As the computational source of RF is relatively low, it can process big data modeling at high speeds. In this study, we employ a CART-based RF model and put different variables into the RF model to classify different driver's risks.

SVM, a supervised learning technique, derives from the generalized linear classifier. With the development of nonlinear methods, Boser et al. proposed nonlinear SVM using the kernel method [56], which has performed well in many research fields, such as life insurance and banking insurance [57, 58]. Accordingly, we use SVM to classify driver's risk and to choose the polynomial kernel functions.

4.3. Model Construction and Evaluation. We divide the data set at random, taking 70% as the training set and 30% as the test set. The training set is used to determine the optimal models. For LR models, it is necessary to ascertain the coefficients. In the process of constructing a machine learning model, some hyperparameters need to be tuned, such as the learning rate of the RF models. Therefore, we use the grid research method combined with fivefold cross-validation in the training set to choose these parameters and the selection rule to ensure the optimal value of AUC. For the grid search, we use the Gridsearch function in the Python platform.

We use the test set to evaluate the different models and the performance of certain variables. We choose AUC, the area under the receiver operating characteristic (ROC) curve, as the performance criterion. The ROC curve has as its vertical axis the true positive rates (such as the ratio of collision samples correctly classified) and as its horizontal axis the false positive rates (such as the ratio of collision-free samples erroneously classified as collision samples). If a classification model performs ideally, the curve is close to the upper left. This criterion is insensitive to the class imbalance, which is more appropriate for our study than other metrics such as accuracy or F-measures [59].

5. Empirical Results and Discussion

5.1. Results of Online and Offline Channel Information-Based Models. In this section, we construct a model using all the dependent variables. The results when using nonbinning and binning data are presented in Tables 5 and 6, respectively. For models using nonbinning data, all the results were better than random classification (AUC = 0.50), which proves that the three types of collisions can be identified to a high level of performance based on these data. Generally, the models performed better in identifying severe collision and general collision samples than collision-free and collision samples. Finally, compared with the NN models, the LR models have better performance and stability. The performance of the RF and SVM models is not robust, but they have the optimal performance in predicting collision-free versus collision and severe collision versus general collision incidents, respectively.

In Table 6, we see that the performance of all the models using binning data is better than that of models using nonbinning data. The highest increased ratio is 32.74%. In all situations, the performance of the LR models is better than that of the other models. These results show that data binning can improve the performance of the models, and binning data are therefore adopted in the experiments that follow.

5.2. Identification and Analysis of High-Power Factors in New Variables. The results in Section 5.1 establish the models using various information to improve the accuracy of classifying different collisions. In previous studies, the impacts of speed, acceleration, driving time, and mileage on driver's risk were investigated, but no research to date has explored the effects of turning and offline consumer behavior variables. Therefore, we use vehicle information, all speed variables, and all acceleration variables to construct the basic model. We add turning variables and offline consumer behavior variables to the basic model and identify high-power variables. The experimental results are shown in Tables 7 and 8.

For the classification of collision-free and collision incidents, the turning variables improve the performance of the SVM model only. However, offline consumer behavior variables greatly improve the results of each model. Specifically, the increase in the AUC for each model exceeds

Table 5: Comparison of risk classification models (nonbinning data).

	LR	NN	RF	SVM	Average
$AUC_{0\text{-}1\text{-}2}$	0.748	0.741	0.770	0.701	0.740
$AUC_{1\text{-}2}$	0.783	0.779	0.627	0.780	0.743

Note 1: $AUC_{0\text{-}1\text{-}2}$ and $AUC_{1\text{-}2}$ represent the AUCs for collision-free and collision, and severe collision and general collision incidents, respectively.

Table 6: Comparison of risk classification models (binning data).

	LR_{box}	NN_{box}	RF_{box}	SVM_{box}	Average
$AUC_{0\text{-}1\text{-}2}$	0.832	0.798	0.796	0.798	0.806
IR (%)	11.25	7.72	3.39	13.82	8.93
$AUC_{1\text{-}2}$	0.850	0.823	0.833	0.829	0.834
IR (%)	8.45	5.71	32.74	6.27	12.28

Note 1: LR_{box}, NN_{box}, RF_{box}, and SVM_{box} represent LR, NN, RF, and SVM models using binning data, respectively. Note 2: $AUC_{0\text{-}1\text{-}2}$ and $AUC_{1\text{-}2}$ represent the AUCs for collision-free and collision, and severe collision and general collision incidents, respectively. Note 3: IR represents the increase in the AUC compared with the corresponding nonbinning data model.

Table 7: Comparison of collision-free and collision risk classification models (binning data).

Dependent variable	LR_{box}		NN_{box}		RF_{box}		SVM_{box}	
	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)
Basic features	0.747	—	0.752	—	0.722	—	0.690	—
Basic features + turning specific	0.745	−0.21	0.712	−5.31	0.696	−3.63	0.714	3.43
Basic features + consumer behavior	0.832	11.37	0.833	10.83	0.814	12.75	0.792	14.76

Note 1: LR_{box}, NN_{box}, RF_{box}, and SVM_{box} represent LR, NN, RF, and SVM models using binning data, respectively. Note 2: IR represents the increase in the AUC compared with the basic feature-based model.

Table 8: Comparison of severe collision and general collision risk classification models (binning data).

Dependent variable	LR_{box}		NN_{box}		RF_{box}		SVM_{box}	
	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)
Basic features	0.800	—	0.729	—	0.785	—	0.778	—
Basic features + turning specific	0.779	−2.60	0.725	−0.58	0.761	−3.07	0.779	0.12
Basic features + consumer behavior	0.861	7.62	0.808	10.80	0.830	5.67	0.833	7.04

Note 1: LR_{box}, NN_{box}, RF_{box}, and SVM_{box} represent LR, NN, RF, and SVM models using binning data, respectively. Note 2: IR represents the increase in the AUC compared with the basic feature-based model.

10%, with an average of 12.42%. Table 8 presents the classification results for severe collisions and general collisions. As before, the turning variables do not contribute to the models, but the offline consumer behavior variables are conducive to their improvement. Although performance is not improved as much as for the classification of collision-free and collisions, the increase in the AUC is still 7.78% on average.

Tables 7 and 8 show that offline consumer behavior variables can greatly improve the performance of all models. Therefore, to investigate the power of different offline consumer behavior variables, we construct different offline consumer behavior-based models, and the different driver's risk classification results are shown in Tables 9 and 10. For the classification of collision-free and collision samples (Table 9), all the variables improve the classification results. The average increase in the AUC ratio for the average maintenance interval in days is 10.62%, but the power of the average number of products used for one maintenance episode is not robust, especially in the NN and RF models. The other three variables also contribute to the classification models. The performance of consumer behavior in detecting severe collisions and general collisions is similar to the results obtained above (see Table 10). The average maintenance interval in days is still the most effective factor among the offline consumer behavior variables, and the power of the average maintenance interval in miles takes second place to that of the average maintenance interval in days.

These results indicate the effectiveness of our framework for identifying different drivers' risks. It is helpful to preprocess these variables using the data binning technique. We find that the turning behavior variables are not very important in the risk classification models; however, offline consumer behavior is proven to be effective in classifying driver's risk. Neither of these factors has been explored in previous studies, and therefore our findings are useful for both insurance practice and academic research.

Table 9: Comparison of different consumer behavior-based models for collision-free and collision risk classification (binning data).

Dependent variable	LR_{box}		NN_{box}		RF_{box}		SVM_{box}	
	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)
Basic features	0.747	—	0.752	—	0.722	—	0.690	—
Basic features + proportion of maintenance carried out in 4S dealer shops	0.771	3.23	0.768	2.14	0.750	3.92	0.729	5.56
Basic features + average price of maintenance carried out in 4S dealer shops	0.751	0.55	0.750	−0.25	0.735	1.77	0.718	4.01
Basic features + average maintenance interval (days)	0.829	10.92	0.811	7.87	0.808	11.92	0.772	11.75
Basic features + average maintenance interval (miles)	0.765	2.40	0.755	0.41	0.742	2.83	0.714	3.43
Basic features + average number of products used for one maintenance episode	0.750	0.41	0.737	−1.96	0.718	−0.51	0.707	2.33

Note 1: LR_{box}, NN_{box}, RF_{box}, and SVM_{box} represent LR, NN, RF, and SVM models using binning data. Note 2: IR represents the increase in the AUC compared with the basic feature-based model.

Table 10: Comparison of consumer behavior-based models for severe collision and general collision risk classification (binning data).

Dependent variable	LR_{box}		NN_{box}		RF_{box}		SVM_{box}	
	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)	AUC	IR (%)
Basic features	0.800	—	0.729	—	0.785	—	0.778	—
Basic features + proportion of maintenance carried out in 4S dealer shops	0.801	0.07	0.739	1.40	0.779	-0.75	0.783	0.64
Basic features + average price of maintenance carried out in 4S dealer shops	0.804	0.44	0.733	0.52	0.781	-0.61	0.785	0.89
Basic features + average maintenance interval (days)	0.855	6.91	0.820	12.47	0.824	4.95	0.829	6.57
Basic features + average maintenance interval (miles)	0.824	3.00	0.764	4.84	0.819	4.32	0.796	2.35
Basic features + average number of products used in one maintenance episode	0.801	0.07	0.731	0.25	0.789	0.42	0.780	0.24

Note 1: LR_{box}, NN_{box}, RF_{box}, and SVM_{box} represent LR, NN, RF, and SVM models using binning data, respectively. Note 2: IR represents the increase in the AUC compared with the basic feature-based model.

5.3. Practical Implications, Limitations, and Future Work. The results found in this paper suggest that the following UBI pricing and cost-control strategies could be employed by insurance companies. First, insurance companies can use the classification models for different types of collisions to fine-tune the pricing of auto insurance and reduce their costs. By using a more accurate classification model for general and severe collisions, insurance companies can charge higher premiums to customers that have a high probability of severe collisions, thereby reducing their losses. For customers with a high probability of general collisions, insurance companies can increase premiums slightly, thus not only reducing their loss in relation to general collisions but also attracting valuable customers from insurance companies that do not distinguish between severe and general collisions. Second, insurance companies can use offline consumer behavior variables to develop positive cost-control strategies. Consumer behaviors, as very important vehicle collision classification variables, can distinguish between collision-free, general collision, and severe collision incidents. This brings a fresh approach to auto insurance cost-control strategies, enabling insurance companies to cooperate with car maintenance shops to give consumers, for instance, time-limited car maintenance vouchers to cultivate their vehicle maintenance behavior, thereby reducing the probability of collisions and controlling compensation costs.

Although this paper provides some valuable UBI business strategies, it also has some limitations that future work can address. First, this paper has studied vehicles in one city, but regional differences may have an impact on traffic crashes. Future research could seek to verify our conclusions using a nationwide sample. Second, this paper mainly considered variables that reflect drivers' behavior. In the future, studies could investigate the impact of additional external variables, such as weather, on the classification of different driving risks.

6. Conclusions

In recent years, with the rise of UBI, insurance companies have paid greater attention to factors that can improve the classification of driver's risk, while drivers and the insurance market have called for more effective UBI products. Therefore, this paper proposes new models for detecting different drivers' risks by combining IoT data and offline consumer behavior data. Our exploration of the added power of different variables provides surprising advice for the improvement of UBI products. Our main conclusions are as follows. First, the proposed classification models in this paper can verify different levels of vehicle collision (collision-free, severe collision, and general collision) to a satisfactory level of performance. Second, offline consumer behavior variables are strongly capable of improving the performance of classification models. Of these, the average maintenance interval in days is the most effective for improving the accuracy of classification results. Vehicles that are maintained frequently are less likely to be involved in a collision incident.

Data Availability

The data used to support the findings of this study have not been made available because the authors have signed a confidentiality agreement with the company that provided the data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by the National Social Science Fund (20BJY180) and the Graduate Innovation Fund of Shanghai University of Finance and Economics (CXJJ-2017-418).

References

[1] P. Ulleberg, "Personality subtypes of young drivers. Relationship to risk-taking preferences, accident involvement, and response to a traffic safety campaign," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 4, no. 4, pp. 279–297, 2001.

[2] WHO, *Global Status Report on Road Safety 2018*, WHO, Geneva, Switzerland, 2018.

[3] A. Cohen and P. Siegelman, "Testing for adverse selection in insurance markets," *Journal of Risk and Insurance*, vol. 77, no. 1, pp. 39–84, 2010.

[4] M.-J. Segovia-Vargas, M.-d.-M. Camacho-Miñano, D. Pascual-Ezama, and D. Pascual-Ezama, "Risk factors selection in automobile insurance policies: a way to improve the bottom line of insurance companies," *Review of Business Management*, pp. 1228–1245, 2015.

[5] L. D. Xu, W. He, and S. Li, "Internet of Things in industries: a Survey," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2233–2243, 2014.

[6] G. Jayavardhana, B. Rajkumar, M. Slaven, and P. Marimuthu, "Internet of Things (IoT): a vision, architectural elements, and future directions," *Future Generation Computer Systems*, vol. 29, pp. 1645–1660, 2013.

[7] D. Panos and S. Mari, "Profiting from business model innovation: evidence from pay-as-you-drive auto insurance," *Research Policy*, vol. 42, no. 1, p. 101, 2013.

[8] S. Husnjak, D. Peraković, I. Forenbacher, and M. Mumdziev, "Telematics system in usage based motor insurance," *Procedia Engineering*, vol. 100, pp. 816–825, 2015.

[9] G. Vaia, C. Erran, W. DeLone, H. trautsch, and F. menichetti, "Vehicle telematics at an Italian insurer: new auto insurance products and a new industry ecosystem," *MIS Quarterly Executive*, vol. 11, no. 3, p. 113, 2012.

[10] J. R. Treat, N. S. Tumbas, and S. T. McDonald, "Tri-level study of the causes of traffic accidents," *Vision Research*, vol. 42, no. 21, pp. 2419–2430, 1979.

[11] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, *The Impact of Driver Inattention on Near Crash/Crash Risk: An Analysis Using the 100-Car Naturalistic Driving Study Data*, National Highway Traffic Safety Administration, Washington, DC, USA, 2006.

[12] E. Petridou and M. Moustaki, "Human factors in the causation of road traffic crashes," *European Journal of Epidemiology*, vol. 16, no. 9, pp. 819–826, 2000.

[13] R. Rowe, G. D. Roman, F. P. McKenna, E. Barker, and D. Poulter, "Measuring errors and violations on the road: a bifactor modeling approach to the driver behavior questionnaire," *Accident Analysis & Prevention*, vol. 74, pp. 118–125, 2015.

[14] S. Singh, *Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey*, National Highway Traffic Safety Administration, Washingston, DC, USA, 2015.

[15] J. C. F. De Winter and D. Dodou, "The driver behaviour questionnaire as a predictor of accidents: a meta-analysis," *Journal of Safety Research*, vol. 41, no. 6, pp. 463–470, 2010.

[16] A. E. af Wåhlberg, P. Barraclough, and J. Freeman, "The driver behaviour questionnaire as accident predictor; A methodological Re-Meta-Analysis," *Journal of Safety Research*, vol. 55, pp. 185–212, 2015.

[17] R. A. Blanchard, A. M. Myers, and M. M. Porter, "Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 523–529, 2010.

[18] D. A. Lombardi, W. J. Horrey, and T. K. Courtney, "Age-related differences in fatal intersection crashes in the United States," *Accident Analysis & Prevention*, vol. 99, pp. 20–29, 2017.

[19] G. Li, W. Lai, and X. Qu, "Association between crash attributes and drivers' crash involvement: a study based on police-reported crash data," *International Journal of Environmental Research and Public Health*, vol. 17, no. 23, pp. 1–18, 2020.

[20] G. Li, Y. Liao, Q. Guo, C. Shen, and W. Lai, "Traffic crash characteristics in shenzhen, China from 2014 to 2016," *International Journal of Environmental Research and Public Health*, vol. 18, no. 3, pp. 1–24, 2021.

[21] I. van Schagen and F. Sagberg, "The potential benefits of naturalistic driving for road safety research: theoretical and empirical considerations and challenges for the future," *Procedia - Social and Behavioral Sciences*, vol. 48, pp. 692–701, 2012.

[22] P. Choudhary and N. R. Velaga, "Mobile phone use during driving: effects on speed and effectiveness of driver compensatory behaviour," *Accident Analysis & Prevention*, vol. 106, pp. 370–378, 2017.

[23] N. M. Pawar, R. Kaur Khanuja, P. Choudhary, and R. Nagendra, "Velaga. "Modelling braking behaviour and accident probability of drivers under increasing time pressure conditions," *Accident Analysis and Prevention*, vol. 136, 2020.

[24] Y. Zhao, T. Miyahara, K. Mizuno, D. Ito, and Y. Han, "Analysis of car driver responses to avoid car-to-cyclist perpendicular collisions based on drive recorder data and driving simulator experiments," *Accident; Analysis and Prevention*, vol. 150, p. 105862, 2021.

[25] G. Li, Y. Yang, T. Zhang et al., "Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios," *Transportation Research Part C: Emerging Technologies*, vol. 122, 2021.

[26] J. Santos, N. Merat, S. Mouta, K. Brookhuis, and D. De Waard, "The interaction between driving and in-vehicle information systems: comparison of results from laboratory, simulator and real-world studies," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 2, pp. 135–146, 2005.

[27] I. Zöller, B. Abendroth, and R. Bruder, "Driver behaviour validity in driving simulators - analysis of the moment of initiation of braking at urban intersections," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 61, pp. 120–130, 2019.

[28] T. a. Dingus, S. G. Klauer, V. L. Neale et al., *The 100-Car Naturalistic Driving Study Phase II – Results of the 100-Car Field Experiment*, National Highway Traffic Safety Administration, Washingston, DC, USA, 2006.

[29] G. Li, S. E. Li, B. Cheng, and P. Green, "Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities," *Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 113–125, 2017.

[30] A. B. Ellison, S. P. Greaves, and M. C. J. Bliemer, "Driver behaviour profiles for road safety analysis," *Accident Analysis & Prevention*, vol. 76, pp. 118–132, 2015.

[31] T. Toledo, O. Musicant, and T. Lotan, "In-vehicle data recorders for monitoring and feedback on drivers' behavior," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 3, pp. 320–331, 2008.

[32] H. Singh and A. Kathuria, "Analyzing driver behavior under naturalistic driving conditions: a review," *Accident; Analysis and Prevention*, vol. 150, p. 105908, 2021.

[33] J. Sangster, H. Rakha, and J. Du, "Application of naturalistic driving data to modeling of driver car-following behavior," *Transportation Research Record*, vol. 2390, pp. 20–33, 2013.

[34] G. Cao, J. Michelini, K. Grigoriadis, B. Ebrahimi, and M. A. Franchek, "Cluster-based correlation of severe driving events with time and location," *Journal of Intelligent Transportation Systems*, vol. 20, no. 6, pp. 516–531, 2016.

[35] J. Beirlant, V. Derveaux, A. M. De Meyer, M. J. Goovaerts, E. Labie, and B. Maenhoudt, "Statistical risk evaluation applied to (Belgian) car insurance," *Insurance: Mathematics and Economics*, vol. 10, no. 4, pp. 289–302, 1992.

[36] Y.-L. Ma, X. Zhu, X. Hu, and Y.-C. Chiu, "The use of context-sensitive insurance telematics data in auto insurance rate making," *Transportation Research Part A: Policy and Practice*, vol. 113, pp. 243–258, 2018.

[37] J. Paefgen, T. Staake, and F. Thiesse, "Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach," *Decision Support Systems*, vol. 56, no. 1, pp. 192–201, 2013.

[38] J. Paefgen, T. Staake, and E. Fleisch, "Multivariate exposure modeling of accident risk: insights from pay-as-you-drive insurance data," *Transportation Research Part A: Policy and Practice*, vol. 61, pp. 27–40, 2014.

[39] P. Baecke and L. Bocca, "The value of vehicle telematics data in insurance risk selection processes," *Decision Support Systems*, vol. 98, pp. 69–79, 2017.

[40] Y. Huang and S. Meng, "Automobile insurance classification ratemaking based on telematics driving data," *Decision Support Systems*, vol. 127, 2019.

[41] G. Li, Y. Chen, D. Cao, X. Qu, Bo Cheng, and K. Li, "Extraction of descriptive driving patterns from driving data using unsupervised algorithms," *Mechanical Systems and Signal Processing*, vol. 156, 2021.

[42] P. J. Peter, J. C. Olson, and J. A. Stuart, "Consumer behaviour and marketing strategy," *Journal of Marketing Research*, pp. 18–21, 2005.

[43] F. Guo and Y. Fang, "Individual driver risk assessment using naturalistic driving data," *Accident Analysis & Prevention*, vol. 61, pp. 3–9, 2013.

[44] D. K. Rigby, "The future of shopping," *Harvard Business Review*, vol. 89, no. 12, pp. 65–76, 2011.

[45] G. Li, T. Zhang, and G. K. Tayi, "Inroad into omni-channel retailing: physical showroom deployment of an online retailer," *European Journal of Operational Research*, vol. 283, no. 2, pp. 676–691, 2020.

[46] S. Gallino and A. Moreno, "Integration of online and offline channels in retail: the impact of sharing reliable inventory availability information," *Management Science*, vol. 60, no. 6, pp. 1434–1451, 2014.

[47] F. Gao and X. Su, "Omnichannel retail operations with buy-online-and-pick-up-in-store," *Management Science*, vol. 63, no. 8, pp. 2478–2492, 2017.

[48] Y. Li, G. Li, G. K. Tayi, and T. C. E. Cheng, "omni-channel retailing: do offline retailers benefit from online reviews?" *International Journal of Production Economics*, vol. 218, pp. 43–61, 2019.

[49] F. Gao and X. Su, "Omnichannel Service operations with online and offline self-order technologies," *Management Science*, vol. 64, no. 8, pp. 3595–3608, 2018.

[50] R. Henckaerts, K. Antonio, M. Clijsters, and R. Verbelen, "A data driven binning strategy for the construction of insurance tariff classes," *Scandinavian Actuarial Journal*, vol. 2018, no. 8, pp. 681–705, 2018.

[51] G. Gao, M. V. Wüthrich, and H. Yang, "Evaluation of driving risk at different speeds," *Insurance: Mathematics and Economics*, vol. 88, pp. 108–119, 2019.

[52] M. Ayuso, M. Guillen, and J. P. Nielsen, "Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data," *Transportation*, vol. 46, no. 3, pp. 735–752, 2019.

[53] X. Xiong, L. Chen, and J. Liang, "A new framework of vehicle collision prediction by combining SVM and HMM," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 699–710, 2018.

[54] Y. Lin and R. Li, "Real-time traffic accidents post-impact prediction: based on crowdsourcing data," *Accident; Analysis and Prevention*, vol. 145, p. 105696, 2020.

[55] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[56] B. E. Boser, M. G. Isabelle, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, New York, NY, USA, July 1992.

[57] G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Engineering Applications of Artificial Intelligence*, vol. 37, pp. 368–377, 2015.

[58] Md R. Islam, S. Liu, R. Biddle et al., "Discovering dynamic adverse behavior of policyholders in the life insurance industry," *Technological Forecasting and Social Change*, vol. 163, 2021.

[59] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.