Hindawi
Journal of Advanced Transportation
Volume 2023, Article ID 2780961, 16 pages
https://doi.org/10.1155/2023/2780961

Research Article

Exploring the Behavior-Driven Crash Risk Prediction Model: The Role of Onboard Navigation Data in Road Safety

Xiao-chi Ma[1,2,3,4] **Jian Lu**[1,2,3] **and Yiik Diew Wong**[4]

[1]*Jiangsu Key Laboratory of Urban ITS, Southeast University, Nanjing, China*
[2]*Jiangsu Province Collaborative Innovation Centre of Modern Urban Traffic Technologies, Southeast University, Nanjing, China*
[3]*School of Transportation, Southeast University, Nanjing, China*
[4]*School of Civil and Environmental Engineering, Nanyang Technological University, Singapore*

Correspondence should be addressed to Jian Lu; lujian_1972@seu.edu.cn

Received 20 September 2023; Revised 1 November 2023; Accepted 14 December 2023; Published 26 December 2023

Academic Editor: Jaehyun (Jason) So

Driving behavior has frequently been overlooked in previous road traffic crash research. Hereby, abnormal (extreme) driving behavior data transmitted by the onboard navigation systems were collected for vehicles involved in traffic crashes, including sharp-lane-change, sharp-acceleration, and sudden-braking behaviors. Using these data in conjunction with expressway crash records, multiple classification learners were trained to establish a behavior-driven risk prediction model. To further investigate the influence of driving behavior on crash risk, partial dependence plots (PDPs) were applied. Regression analyses indicate that models have a stronger effect when derivative features such as frequency of specific deviant behavior, speed, and acceleration in the behavior process are included. The behavioral RUSBoost model surpasses other models, achieving an AUC prediction metric of 0.782 and outperforming traditional traffic-flow-driven machine learning models. PDP analysis demonstrates that the sudden-braking behavior is the leading contributory factor of expressway crashes, particularly when the acceleration exceeds 0.5 G. This study confirms the potential of predicting crash risks through augmenting behavior data from navigation software; the findings lay a foundation for countermeasures.

1. Introduction

Traffic is a complex system that consists of the driver, vehicle, road, and environment, in which the abnormality of one or more components will heighten crash risk. Researchers have studied extensively the relationship between vehicles, roads, and environment with crash risk and built real-time risk prediction models. However, there are few studies that evaluate the impact of driver behavior on crash risk. From the general rational point of view, there should exist some close connection between driver behavior and crashes. Most existing research on driving behavior has been based on perception surveys [1] or vehicle trajectories [2] captured by video cameras. On the one hand, the response rate of questionnaire surveys tends to be low, yet the research result is greatly affected by the quality of respondents' answers. On the other hand, the extraction process of vehicle trajectory data is cumbersome, and often times only a short distance of trajectory is filmed. In earlier studies, given the difficulty of obtaining real-time driving behavior data, behavior was often treated as being heterogeneous in the analysis process [3] and dealt with via a random effect model to clarify the mechanism of the crash [4]. Distraction [5], fatigue [6], and drunk driving [7] were the familiar heterogeneities considered. Later, by using eye trackers or simulators, different driver attributes were also fully considered [3]. An ideal method to obtain driving behavior is to utilize onboard sensors and GPS to transmit real-time vehicle position and operating condition information, including longitude and latitude coordinates, vehicle speed, acceleration, and steering angle, and use such data to assess driving behavior and crash risk. An easily accessible source for this type of data can be streamed from online navigation software operating in conjunction with high-precision maps.

Hereby, in order to fill the gap of the role of driving behavior in crash risk theory and hence allow timely implementation of safety remediation measures, this study dealt with extreme driving behavior (termed herewith as abnormal driving behavior) obtained for the China G25 expressway. The data were provided by AutoNavi over a span of 7 days and 120 km and the crashes that occurred during this same period. This facilitated the development of a crash risk prediction model based on real-time driving behavior. AutoNavi first screened out abnormal driving behavior on the expressway, including sharp-left-lane-change, sharp-right-lane-change, sharp-acceleration, and sudden-braking behaviors. For each crash, the features of risky driving behavior occurring before the crash were attached, and a case-control method was applied to generate datasets with different proportions of positive and negative cases. The crash risk prediction models are established via various machine learning methods, and the role of driving behavior is explored via logit regression and partial dependency plots.

A key contribution of this study is to verify the feasibility of predicting crash risk using abnormal driving behavior collected by the onboard navigation system. Regression using feature variables generated from original data is presented, and these generated variables are found to be interpretable in improving the performance of the prediction model. Another contribution is that through the behavior-driven risk prediction model, the relationship between driving behaviors and crash risk is studied in depth, which can provide drivers with risk avoidance advice.

The structure of the paper is organized as follows. This section is the introduction, followed by Section 2 which covers the literature review. In Section 3, processing of the dataset and the methodology are described. Section 4 gives the regression model results and an in-depth analysis of the variables. Limitations of the study and future work are presented in Section 5, along with a conclusion summary of the research study.

2. Literature Review

Many scholars have explored the contributory factors of traffic crashes from various prospects. Corresponding risk prediction models are also established for real-time risk prediction [8]. Macroscopic traffic-flow information such as traffic volume, speed, and density are collected to assess their impact on crash risk [9, 10]. Microscopic traffic data, such as headway [11] collected via onboard cameras and radar sensors, are used to reconstruct the process of crash formation. Later, scholars introduced more factors about road alignment and the environment, including horizontal and vertical curves [12], pavement surface [13], illumination conditions [14], and the weather [15].

When it comes to the effects of driving behavior on crash risk, many data collection methods have been proposed. The most economical way to collect driving behavior data is to conduct surveys by using questionnaires which have found significant driver-related crash factors on drinking [16], fatigue, and distracted driving [17]. However, the response rate of questionnaire surveys tends to be low and it is not possible to do real-time tracking. Another widely used method to analyze driving behavior is by examining the trajectory of vehicles. Researchers have used onboard cameras or drones to record vehicle trajectories and explore crash risk factors combined with microscopic traffic-flow indicators [18]. Meanwhile, more publicly available datasets of naturalistic driving studies are used to assess driving risk [19]. Furthermore, researchers found that combining driving behavior and traffic-flow parameters to predict risk can greatly improve prediction performance. Ma et al. [20] used the frequency of risky driving behaviors and traffic-flow data to establish a short-term crash risk prediction model. However, given the difficulty and cost of data acquisition, large-scale and long-term collection of driving behavior and traffic-flow parameters is still difficult to apply in practice. The information silo effect of roadside and in-vehicle devices also makes the collaborative evaluation of multisource data challenging.

Nowadays, modern onboard navigation software can effectively capture vehicle movement information [21]. In China, Amap and Baidu Map together hold more than 50% of the market share. In the U.S., Google Map has a wide user base. There are also a number of vehicle manufacturers that have partnered with navigation companies, e.g., Tesla®, Cadillac®, and BYD®, where the navigation system provides drivers with lane-level positioning, as well as a variety of speed and traffic limit information for the vehicle. Herein, the use of data from navigation software that has high market penetration to assess safety risks on the roads is very promising and commercially viable. There is a diversity of vehicle brands on the roads but utilizing in-vehicle high-performance sensors installed on many brands of vehicle to collect data will inevitably magnify the cost. Hereby, collecting data by using the navigation software will be the most economical and also the most suitable way for relevant companies to develop this function.

For the research to explore the influencing factors of crash risk, machine learning methods can easily improve the accuracy of the prediction model, but the black-box model will complicate the interpretation [22]. Therefore, researchers often use agent models (e.g., logistics regression) [23] and visualization methods (e.g., SHAP and PDP) [24] to interpret machine learning models.

To sum up, as illustrated by Figure 1, driving behavior plays an important role along with other factors in the formation of a crash, while noting that current behavior data collection methods require substantive costs and are also constrained by the information silo effect. In this research, abnormal driving behavior data provided by AutoNavi® are utilized to generate datasets based on crash data via the case-control method. Multiple machine learning methods are used to establish the behavior-driven risk prediction model and explore the impact of risky driving behavior on traffic crashes, which will help in the development of appropriate risk prediction functions to be applied on a large scale.

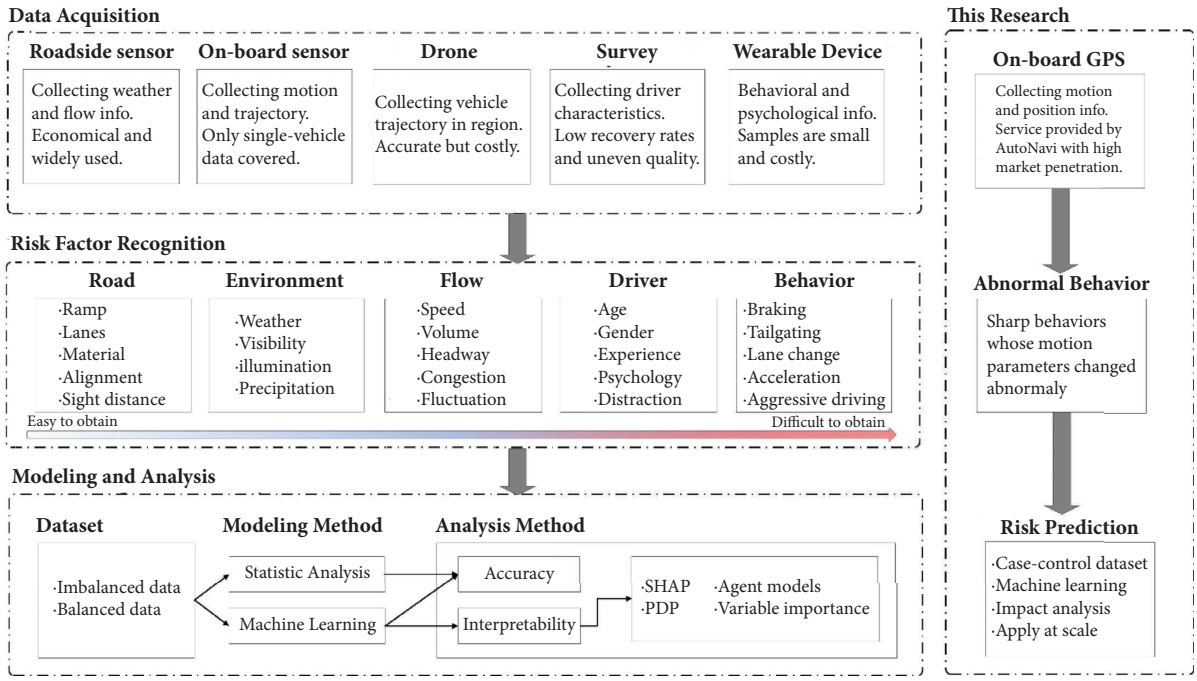

Figure 1: Framework of crash risk factor acquisition and prediction models.

3. Data Preparation and Methodology

3.1. Data Preparation. The data on incidences of abnormal driving behavior on a 120-kilometer section of the G25 expressway in China from September 26 to October 3 in 2020 were provided by AutoNavi Software Co. Ltd. The study area is the section from Liyang to Changxing of the G25 expressway, which has 3 lanes in both directions, with an average annual daily traffic volume of 56,772 vehicles. The alignment is gentle, and the fluctuation is small. There are no large- and medium-sized cities or large entrances along the road, and the traffic volume of each section along the road does not change significantly. Crash data were provided by the authority responsible for managing the G25 expressway. Statistics showed that on the above-mentioned dates at the expressway sections, over 140,000 incidents of abnormal driving behavior data and 284 crashes were collected accordingly. Original data collected by AutoNavi were sampled via onboard GPS once per second, including acceleration, speed, course angle, latitude, and longitude changes during vehicle movement. AutoNavi classifies the vehicle motion state with these parameters at the 90th percentile as an "abnormal" behavior and defines four incidents based on the specific weighting parameter. Incidents of abnormal driving behaviors were divided into four categories, namely, sharp-left-lane-change, sharp-right-lane-change, sharp-acceleration, and sudden-braking. The specific technical details including the threshold and weights are not allowed to be disclosed due to confidential agreement. However, the threshold is not the focus of this research. This study is aimed at proposing a comprehensive methodology to study the feasibility of establishing a risk prediction model based on extreme driving behavior. Kinematic parameters of driving behaviors provided in the processed dataset are sufficient for the research. The features of the 4 abnormal behavior data are listed in Table 1.

Crash data and behavior data were connected with coordinates and time as key values. Previous studies have shown that behaviors around the crash site will show some high-risk characteristics before the crash occurs, Ma et al. [25] pointed that there is a generalized linear relationship between the length of the road section and the number of crashes, and with a certain length range, the heterogeneity caused by road alignment and section location can be eliminated to a certain extent. Hereby, the mean value 120 km/284 = 422 m is calculated as the theoretical length. While the actual crash coordinate is not so precise, considering a redundancy here, a spatial range of 500 m is chosen as the practical target section. That is, for each crash coordinate, abnormal driving behaviors within 250 meters before and after the crash position were extracted via GIS software, as shown in Figure 2. The red cross dot represents a crash, and the solid pink point is driving behavior.

Compared to normal traffic operation status, a crash is actually a rare event. The case-control method used for constructing datasets for regression can help analyze risk factors that deviate from normal operating status. The dependent variable of crash data (positive case) is regarded as 1, and the noncrash data (negative case) with dependent variable 0 can be obtained by the following two rules (also shown in Figure 2):

Table 1: Elements of each category of abnormal driving behavior.

Name	Variable description					Unit
Category	Types of abnormal behavior (4 types)					
max_a	Maximum acceleration during the abnormal behavior					g
max_v	Maximum velocity during the abnormal behavior					m/s
Time	Date and time of the abnormal behavior. Accuracy to the second					s
Longitude	Longitude of the abnormal behavior. Accuracy to the centimeter					cm
Latitude	Latitude of the abnormal behavior. Accuracy to the centimeter					cm
Behavior	Feature	Min	Max	Mean	Std	No. of cases
Sharp-left-lane-change behavior	max_a	0.171	0.935	0.330	0.166	996
	max_v	0.000	37.392	21.705	8.034	996
Sharp-right-lane-change behavior	max_a	0.170	0.949	0.382	0.202	861
	max_v	0.000	38.742	22.004	8.296	861
Sharp-acceleration behavior	max_a	0.107	0.992	0.252	0.083	55317
	max_v	0.000	49.323	20.937	9.904	55317
Sudden-braking behavior	max_a	0.198	1.370	0.325	0.104	60056
	max_v	0.000	55.633	21.552	8.253	60056

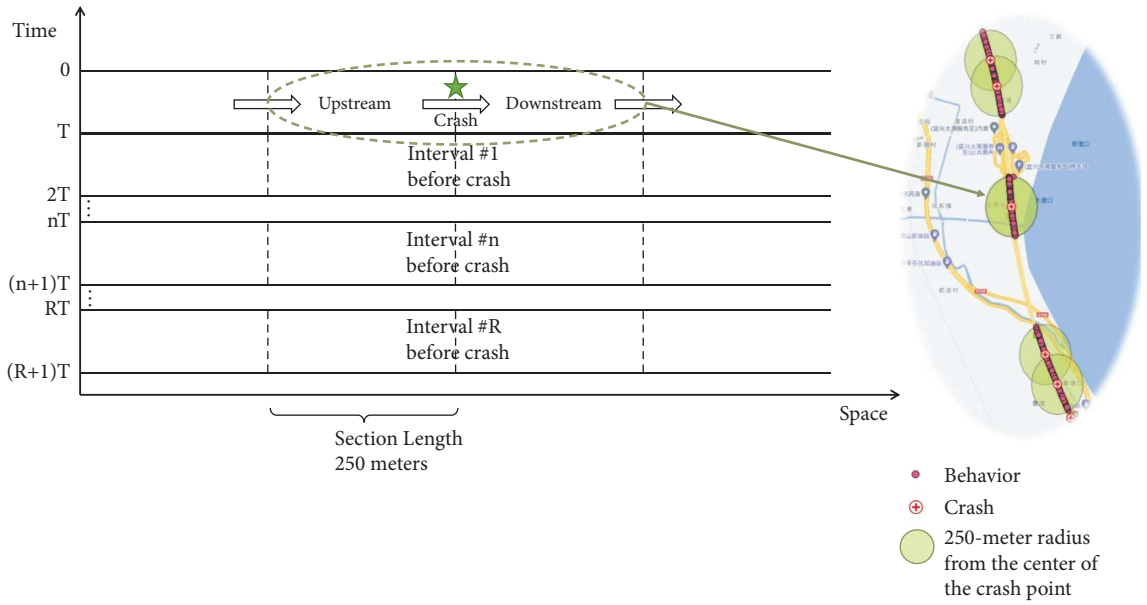

Figure 2: Schematic diagram of driving feature aggregation and sampling process.

(1) *Step 1*. The extraction interval T (e.g., 5 minutes) of additional features is determined, and then the crash data will include abnormal driving behaviors within the time window ranging from 0 to T before the crash.

(2) *Step 2*. The ratio R of crash and noncrash data (e.g., 3) is determined, and then the location of the crash in step 1 is kept unchanged. The abnormal driving behaviors are taken within multiple time windows ranging from $n \times T$ to $(n+1) \times T$ before the crash as the nonaccident data, where n is traversed from 1 to R.

Thus, multiple case-control datasets with different extraction intervals and different data ratios are obtained. In the past research on crash prediction driven by traffic-flow data, multiple sets of time intervals were used to collect data on traffic-flow characteristics and the corresponding prediction models were established [26]. Guo et al. [27] collected the frequency of abnormal driving behavior in a 1-hour interval for modeling, which achieved good performance. According to the literature, in this paper, five intervals of 5, 15, 30, 45, and 60 minutes were used to construct datasets and to select the optimal solution according to their performance. Four ratios, 1 : 1, 1 : 3, 1 : 10, and 1 : 20, were selected to investigate the influence of the ratio of positive and negative cases on the regression results [28]. Due to the loss of behaviors near some crash points, the ratios in the actual constructed datasets would not be accurate, but basically they met the proportion requirements. Thus, 20 datasets were obtained, and the actual number of positive and negative cases in each dataset is shown in Table 2.

Table 2: Real ratio of positive to negative cases in each dataset.

Time interval (min)	Theoretical ratio			
	1 : 1	1 : 3	1 : 10	1 : 20
5	151 : 134	151 : 433	151 : 1420	151 : 2873
15	213 : 232	213 : 655	213 : 2001	213 : 4085
30	265 : 249	265 : 736	265 : 2365	265 : 4867
45	276 : 275	276 : 775	276 : 2534	276 : 5272
60	284 : 282	284 : 815	284 : 2661	284 : 5480

For each crash or noncrash data, three derivative variables are calculated: the number of certain behaviors occurring, the average value of the maximum speed of certain behaviors, and the average value of the maximum acceleration of certain behaviors, which are calculated by using the following equation:

$$\overline{a} = \frac{\sum \max\_a_{c,i}}{N}, \quad \overline{v} = \frac{\sum \max\_v_{c,i}}{N}, \tag{1}$$

where max_a and max_v are the features illustrated in Table 1, subtitle i is the behavior data id, and N is the total number of certain behavior category c.

The naming principles and meanings of the variables in the dataset are given in Table 3. In summary, each piece of data includes a dichotomous dependent variable and 24 behavioral characteristics as independent variables.

3.2. Methodology

3.2.1. Logistic Regression (LR). Currently, there are many methods to assess the contributory factors of a crash [29]. Logistic regression is a mature statistical model, which is a simple but effective binary classifier, belonging to a kind of generalized linear model. We consider a pair of data points set $\{(x, y)\}$, where x is the N-dimensional feature variable and y is the binary dependent variable. The mathematical form of LR is as follows:

$$r = P(y = 1), \quad \text{Logit}\, r = \ln \frac{r}{1 - r} = \alpha + \beta x, \tag{2}$$

where r is the possibility index, which in this study denotes the risk of a crash, α is the constant term, x is the independent variable, and β is the estimated coefficient.

The analysis of variance (ANOVA) test is applied to judge the significance level of variables in the model. During the variance test, all highly significant variables with a p value below 0.05 should be retained as far as possible.

3.2.2. Machine Learning Methods. Multiple machine learning methods are used to obtain a more accurate classification model, including artificial neural network, Naive Bayesian model, and RUSBoost. Feature selection is performed before machine learning regression on the dataset. LR itself is an effective feature extraction method, and hence the high-significance variables (p value <0.05) retained in LR are used as input variables.

The artificial neural network (ANN) is a machine learning algorithm based on the perceptron model. ANN contains an input layer, an output layer, and several hidden layers, and the connections between neurons in all layers have learnable weight parameters. Through model learning, the connection weights between neurons can be continuously adjusted, so that the model output is gradually close to the real situation.

Naive Bayes is a well-established classifier whose core idea is to compute the probability of each categorical value under given conditions and use the class with the highest probability as the output. The parameter estimation of Naive Bayes uses maximum likelihood estimation methods and shows robust performance in noisy datasets.

RUSBoost (random undersampling boost) is a combination of undersampling and AdaBoost algorithm aimed to solve unbalanced data. The weak learner is trained by constructing a balanced dataset through random sampling, and then the integration algorithm is used to obtain a classifier with higher accuracy. Although the SMOTE algorithm also solves the imbalance problem [30], it enlarges the dataset and amplifies the training time, and studies have shown no significant difference in performance between oversampling and random sampling [31].

Each dataset in Table 2 is divided into a training set and a validation set to conduct 5-fold cross-validation in the regression process. Then, a test set with 15% samples of the original dataset will be applied to verify the model performance.

3.2.3. Performance Evaluation Metrics. The datasets obtained in Section 3.1 are a typical unbalanced dataset of classification problems, so comprehensive indicators should be selected to evaluate the performance of the model. The confusion matrix of the predicted results is shown in Table 4. Then, the true positive rate (TPR) and false positive rate (FPR) can be calculated by using the following equations. Multiple (FPR and TPR) coordinates can be obtained by adjusting the classification threshold, which is connected successively to draw the receiver operating characteristic curve (ROC) and calculate the area under the curve (AUC). TheAUC is a comprehensive evaluation index, which is greater than 0.5 and less than 1. The closer it is to 1, the better the classifier effect is.

Table 3: Independent variable description of each data at the given time interval.

Variable format and subscript meaning
Capital letter: the category of abnormal behavior
L: sharp-left-lane-change behavior
R: sharp-right-lane-change behavior
A: sharp-acceleration behavior
B: sudden-braking behavior
Letter after the 1st underscore: the derivative feature
n: number of abnormal behaviors
a: the average max_accelerate during all abnormal behaviors
v: the average max_velocity during all abnormal behaviors
Letter after the 2nd underscore: the position relative to the accident site
up: within 250 m upstream
dn: within 250 m downstream
e.g., B_v_dn means "average max_v of all sudden-braking within 250 m downstream of the crash coordinate"

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

$$\mathrm{FPR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{4}$$

3.2.4. Model Interpretation Method. Partial dependence plot (PDP) is a visualization method widely used to explain the joint effect of two independent variables on the dependent variable in a model. For the variable $\mathbf{x}_E$ to be explained, the following equation is generally used to calculate their partial dependence effects:

$$y^* = \frac{1}{N}\sum_{S=1}^{N} f\left(\mathbf{x}_E, \mathbf{x}_S\right), \tag{5}$$

where N is the number of samples in the dataset and $\mathbf{x}_S$ is the variable in the sample data other than $\mathbf{x}_E$.

4. Results and Discussion

4.1. Model Accuracy Performance. LR and machine learning were performed on the case-control datasets obtained in Section 3, respectively, and ANOVA was conducted to ensure that the significance (p value) of variables that remained in the model was all below 0.05. All the models have been verified by cross-validation, and their AUCs in diverse datasets are shown in Figure 3.

From Figure 3, on the one hand, under the condition of the same sampling time interval, we see that if the unbalanced proportion of the dataset is larger, then the regression effect of the models gradually increases, except for ANN. Too little data can lead to very poor model performance, as can be seen with data ratios of 1 and 3 for each model and a time interval of 5 minutes. On the other hand, under the condition of the same sampling ratio, we see that the effect of the model engendered apparent fluctuation, except for Naive Bayesian, which indicates that blindly expanding the data scale and time interval is not conducive to improving the effect of the regression model. Note that the AUC reaches the best performance of 0.782 in the RUSBoost model, with the ratio of 1 : 20 and the time interval of 15 minutes; therefore, this model is selected finally and interpreted.

Table 4: Confusion matrix of the prediction model.

		Predictive value (y^*)	
		Positive (1)	Negative (0)
Real value (y)	Positive (1)	TP	FN
	Negative (0)	FP	TN

If only the frequency of occurrence of abnormal driving behavior is used, the prediction results are shown in Figure 4. Models using only behavioral frequency variables performed worse than models with speed and acceleration variables attached under all datasets. The two additional variables do not directly characterize the traffic-flow parameters, but still reflect the traffic operation state to some extent, which is an enhancement to the behavior-driven crash prediction model.

Compared with other data sources, the abnormal driving behavior data collected by the navigation system also have advantages in accuracy and usability. As shown in Table 5, the performance of risk prediction models using other data sources is enumerated. The performance of the abnormal behavior data-driven RUSBoost model represented by the AUC value is better than that of the machine learning models using traditional traffic-flow data [20, 32]. In previous studies, it has been demonstrated that the simultaneous use of driving behavior and other traffic information can substantially improve the prediction results [33, 34]. Exactly, the more detailed the data, the better the performance of the model, but the cost and acquisition difficulty also increase. The abnormal driving behavior data used in this study are easy to obtain by navigation enterprises, and the performance of the obtained model is greatly improved compared with the model driven by traditional traffic-flow data, which achieves a balance between accuracy and the difficulty of data acquisition. In summary, it is technically and economically feasible for navigation companies to use abnormal driving behavior data to predict crash risk.

4.2. Discussion of Factor Impact. LR is a statistical analysis method with good interpretability. According to the definition of equation (2), the parameter estimation β is the contribution of the behavior feature to the crash risk, and its exponent value demonstrates the fact that when the variable is increased by 1, the probability of an accident becomes e^{β} times greater than what it would otherwise be. The regression results for each LR experiment are shown in Tables 6–9. Here, the focus is more on common findings across multiple sets of experiments. The first three variables with the largest absolute values of the estimation in each LR result of the 20 experiments were retained, and their frequency of occurrence was counted. The frequency histogram is shown in Figure 5(a). Clearly, the occurrence frequency of

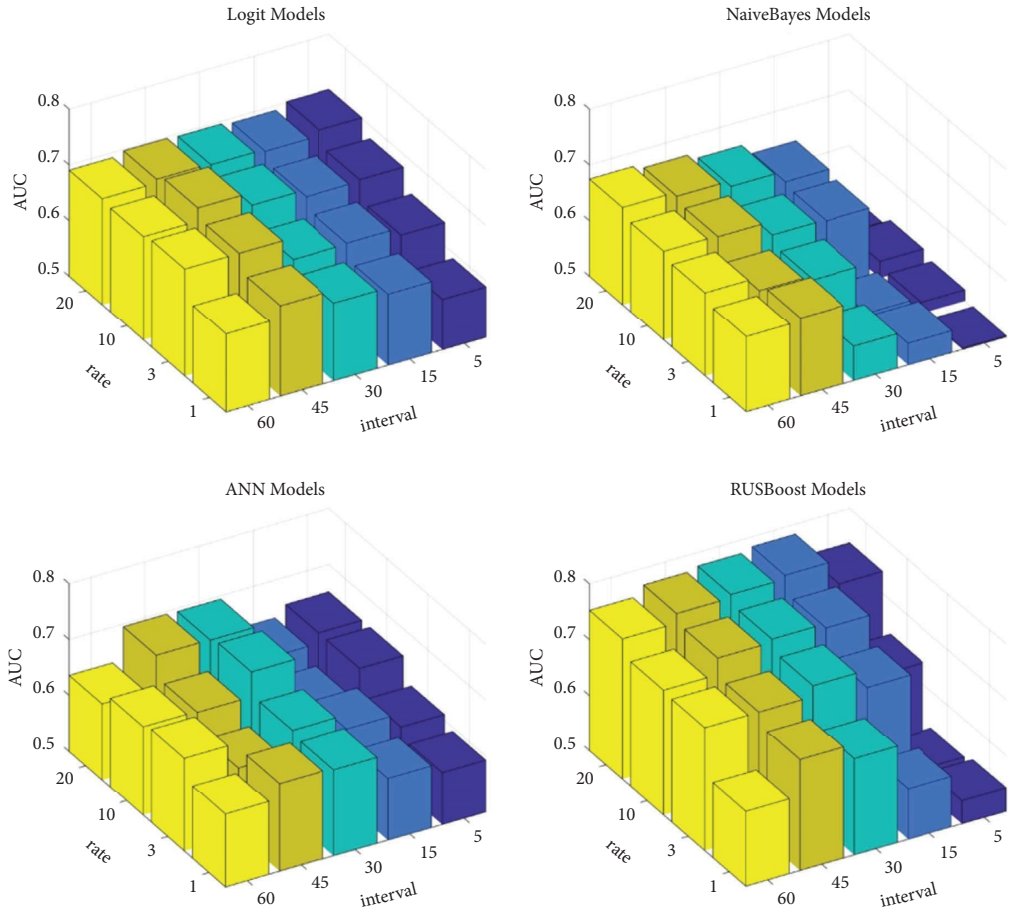

Figure 3: Models' AUC performance of different datasets (with speed and acceleration features).

acceleration features of sharp-acceleration and sudden-braking behaviors occupies the forefront in an overwhelming number, while that of the features of sharp-lane-change behavior are very few, in fact only once. Consequently, subsequent PDP analysis is also dominated by the interaction of the characteristics of these two behaviors.

Figure 5(b) describes in more detail the distribution characteristics of the estimation values of these variables. The vertical axis is the name of the variables and is also sorted in the same frequency order as Figure 5(a), and the horizontal axis is the estimation values. Each dot in Figure 5(b) records one occurrence of the corresponding variable and the exponent value of its coefficient estimation. The blue line perpendicular to the X-axis demarcates the boundaries of influence, with dots to the left of the line representing a negative impact on risk and to the right representing a positive effect. It can be clearly seen that the variables that have a strong impact on crash risk whose estimation is much greater than 0 are the four acceleration variables of sharp-acceleration and sudden-braking behaviors (A_a_up, A_a_dn, B_a_up, and B_a_dn). Even if other features appear in the model, they have little influence on crash risk. Hence, it can be considered that the acceleration features play a dominant role in the formation of expressway crashes, which reflect the intensity of driving behavior. In addition, four speed-related variables (B_v_up, L_v_dn, B_v_dn, and A_v_up) showed negative correlations. The number of sharp-acceleration behaviors was negatively correlated, and the number of sudden-braking behaviors was positively correlated (for more detailed numerical results, one can refer to the contents in Tables 6–9, which will not be repeated here).

In order to deeply analyze the risk factors in driving behavior and derive quantitative risk aversion opinions, PDP analysis was conducted on acceleration and times and speed and times in the two behaviors, respectively. The results are shown in Figures 6 to 8. Not all variables for sharp-acceleration behavior and sudden-braking behavior were retained in the LR regressions, resulting in a total of three PDP results.

Figure 6 illustrates the interaction between acceleration and the number of sharp-acceleration behavior on risk. It can be seen that although the estimation of acceleration in the LR results is large, the magnitude of the change in risk is very small and not outstanding, at only 0.02. We consider the interaction of speed and times which is shown in Figure 7. In general, the speed and times of sharp-acceleration behavior are both negatively correlated with risk. When the speed exceeds 12 m/s (43.2 km/h), the risk depends mainly on the speed of the sharp-acceleration, while the number has little effect.

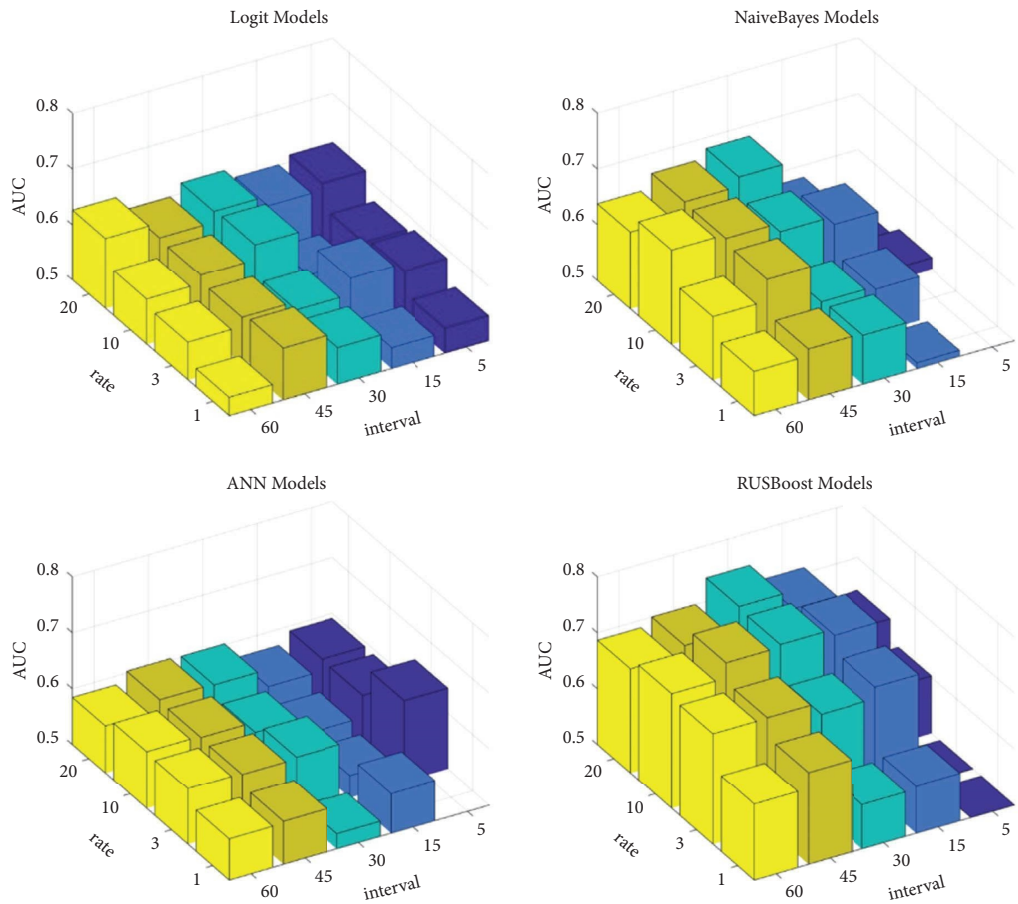

Figure 4: Models' AUC performance of different datasets (frequency feature only).

When it comes to the interaction of acceleration and the number of sudden-braking behavior, the PDP heatmap in Figure 8 shows a typical convex curve when the acceleration reaches 0.5 g and the number is from 9 to 21. Meanwhile, the risk index ranging from 0.24 to 0.36 is greater than that of sharp-acceleration behavior. Therefore, compared with the sharp-acceleration behavior, the sudden-braking behavior is a more critical factor leading to the surge of crash risk.

By comparing the characteristics of the interaction heatmap of sharp-acceleration and sudden-braking behaviors, it is intuitively reflected that the sudden-braking behavior is a typical risk factor, wherein the upper limit of risk in Figure 8 is much higher than that in Figures 6 and 7. The phenomenon that the risk effect of sudden-braking behavior is greater than that of sharp-acceleration behavior can be explained by recreating the scenario in which such risky driving behavior occurred. There are usually three kinds of sharp-acceleration behavior on the expressway. One is that the driver decides to increase its cruising speed to a higher value, the second is to change lanes and overtake, and the third is the vehicle starting off on a congested road. The first two behaviors are allowed to occur on the premise that the current road traffic flow is stable and the driving conditions are good. Generally, the driving speed in this scenario is high, reaching or approaching free flow. In contrast, the third situation occurs on congested road sections whereby vehicles will frequently start and accelerate in a rapid manner, with low intensity but high frequency of occurrence, and crashes are more likely to occur. This finding is consistent with the pattern in Figure 7 and the findings in some previous research [35]. In other words, the features at the time when sharp-acceleration behavior occurs also reflect the current road level of service. When sharp-acceleration occurs, most of the surrounding vehicles are in the similar traffic state and generally follow the driving rules, releasing signals in advance (such as turning on the turn light) to inform the surrounding vehicles of the upcoming action to avoid risk. The occurrence of the sudden-braking behavior is different from the sharp-acceleration behavior. It may occur when an obstacle is met suddenly in front of the car, or drivers find their routes are wrong while on the ramp, being full of randomness, suddenness, and uncertainty, which is difficult to be predicted by other drivers, eventually leading to an accident. Although both behaviors are the main causes of crash risk, they also show great differences in formation mechanisms. Therefore, when using behavior-driven risk prediction models, navigation companies should focus on congested environments and driving conditions with frequent hard braking, specifically represented in this dataset by speeds less than 12 m/s during

Table 5: Comparison of prediction performance using different data sources.

Features in dataset	Model	Road type	AUC	Reference
Traffic flow	Random forest	Highway	0.684	Zhang et al. [32]
Traffic flow and congestion index	Genetic programming	Expressway	0.702	Ma et al. [25]
Abnormal behavior (frequency only)	RUSBoost	Highway	0.741	This study
Abnormal behavior (with movement info)	RUSBoost	Highway	0.782	This study
Traffic flow, aerial photograph, and vehicle trajectory	XGBoost	Highway	0.871	Yuan et al. [33]
Vehicle kinematics, driver input, weather, and physiological signal	Fusion method	Simulation	0.930	Elamrani Abou Elassad et al. [34]

Table 6: Logit regression and ANOVA results in datasets with rates of 1 : 1.

Interval (min)	Variable	Estimation	Std	P value	Exp (estimation)
5	A_a_up	2.399	0.85	0.005	11.007
	Constant	−0.203	0.165	0.218	0.816
15	B_a_up	2.48	0.9	0.006	11.943
	B_v_up	−0.055	0.016	0	0.947
	B_n_dn	0.114	0.04	0.005	1.121
	Constant	−0.21	0.14	0.133	0.81
30	A_a_up	2.064	0.88	0.019	7.878
	A_v_up	−0.031	0.012	0.011	0.969
	A_v_dn	−0.029	0.009	0.001	0.972
	B_a_dn	1.549	0.527	0.003	4.708
	Constant	−0.016	0.2	0.935	0.984
45	A_n_dn	−0.076	0.037	0.04	0.927
	A_a_dn	3.136	1.041	0.003	23.016
	A_v_dn	−0.035	0.01	0.001	0.965
	B_a_dn	2.794	0.592	0	16.344
	Constant	−0.633	0.199	0.001	0.531
60	A_a_up	2.757	0.918	0.003	15.753
	A_v_up	−0.045	0.01	0	0.956
	B_a_dn	1.337	0.588	0.023	3.808
	Constant	−0.275	0.203	0.176	0.759

Table 7: Logit regression and ANOVA results in datasets with rates of 1 : 3.

Interval (min)	Variable	Estimation	Std	P value	Exp (estimation)
5	A_a_up	1.859	0.667	0.005	6.414
	B_v_up	−0.023	0.011	0.035	0.978
	B_n_dn	0.177	0.062	0.004	1.193
	B_v_dn	−0.035	0.013	0.008	0.966
	Constant	−1.049	0.175	0	0.35
15	A_a_up	2.183	0.767	0.004	8.872
	A_v_up	−0.035	0.012	0.004	0.966
	B_v_up	−0.028	0.008	0.001	0.973
	A_v_dn	−0.018	0.008	0.031	0.982
	B_n_dn	0.137	0.031	0	1.147
	Constant	−0.958	0.154	0	0.384
30	A_a_up	1.932	0.733	0.008	6.904
	A_v_up	−0.038	0.01	0	0.963
	A_n_dn	−0.077	0.044	0.078	0.926
	A_a_dn	1.665	0.851	0.05	5.285
	A_v_dn	−0.033	0.01	0.001	0.968
	B_n_dn	0.046	0.024	0.06	1.047
	B_a_dn	0.96	0.498	0.054	2.612
	Constant	−1.085	0.165	0	0.338
45	L_v_up	0.043	0.024	0.076	1.044
	A_a_up	2.338	0.769	0.002	10.359
	A_v_up	−0.032	0.01	0.001	0.968
	B_n_up	0.048	0.019	0.01	1.049
	R_v_dn	0.102	0.037	0.006	1.108
	A_n_dn	−0.118	0.033	0	0.888
	A_a_dn	1.645	0.827	0.047	5.181
	A_v_dn	−0.02	0.009	0.031	0.98
	B_a_dn	2.501	0.528	0	12.2
	Constant	−1.645	0.191	0	0.193

Table 7: Continued.

Interval (min)	Variable	Estimation	Std	P value	Exp (estimation)
60	L_v_up	0.059	0.022	0.008	1.061
	R_v_up	−0.09	0.054	0.098	0.914
	A_a_up	2.74	0.766	0	15.486
	A_v_up	−0.046	0.009	0	0.955
	B_n_up	0.072	0.018	0	1.075
	L_v_dn	−0.092	0.045	0.04	0.912
	R_v_dn	0.062	0.029	0.034	1.064
	A_n_dn	−0.081	0.026	0.002	0.922
	B_a_dn	1.614	0.524	0.002	5.023
	Constant	−1.407	0.173	0	0.245

Table 8: Logit regression and ANOVA results in datasets with rates of 1 : 10.

Interval (min)	Variable	Estimation	Std	P value	Exp (estimation)
5	A_a_up	1.376	0.598	0.021	3.958
	B_a_up	1.447	0.783	0.065	4.249
	B_v_up	−0.044	0.016	0.008	0.957
	A_a_dn	1.794	0.787	0.023	6.014
	A_v_dn	−0.038	0.014	0.009	0.963
	B_n_dn	0.229	0.059	0	1.257
	B_v_dn	−0.027	0.012	0.028	0.973
	Constant	−2.334	0.185	0	0.097
15	A_a_up	2.336	0.637	0	10.342
	A_v_up	−0.042	0.011	0	0.959
	B_n_up	0.064	0.036	0.077	1.066
	B_v_up	−0.028	0.009	0.001	0.973
	A_v_dn	−0.022	0.008	0.004	0.978
	B_n_dn	0.123	0.03	0	1.131
	Constant	−2.058	0.141	0	0.128
30	A_a_up	1.737	0.677	0.01	5.678
	A_v_up	−0.04	0.009	0	0.96
	B_n_up	0.078	0.026	0.002	1.081
	B_v_up	−0.014	0.007	0.044	0.986
	A_n_dn	−0.112	0.04	0.005	0.894
	A_a_dn	1.477	0.756	0.051	4.382
	A_v_dn	−0.028	0.009	0.002	0.973
	B_n_dn	0.051	0.027	0.055	1.053
	B_a_dn	1.197	0.465	0.01	3.31
	Constant	−2.187	0.172	0	0.112
45	L_n_up	3.483	1.344	0.01	32.55
	L_a_up	−30.421	11.702	0.009	0
	L_v_up	0.17	0.057	0.003	1.185
	A_a_up	2.522	0.678	0	12.455
	A_v_up	−0.036	0.009	0	0.964
	B_n_up	0.06	0.016	0	1.062
	R_v_dn	0.041	0.021	0.049	1.042
	A_n_dn	−0.091	0.03	0.002	0.913
	A_a_dn	1.626	0.725	0.025	5.086
	A_v_dn	−0.02	0.008	0.017	0.98
	B_a_dn	2.473	0.488	0	11.861
	Constant	−2.863	0.176	0	0.057
60	A_a_up	2.576	0.661	0	13.149
	A_v_up	−0.044	0.008	0	0.957
	B_n_up	0.054	0.013	0	1.055
	A_n_dn	−0.041	0.021	0.05	0.96
	B_a_dn	2.22	0.486	0	9.207
	Constant	−2.765	0.158	0	0.063

Table 9: Logit regression and ANOVA results in datasets with rates of 1 : 20.

Interval (min)	Variable	Estimation	Std	P value	Exp (estimation)
5	A_a_up	2.549	0.725	0	12.79
	A_v_up	−0.027	0.013	0.046	0.974
	B_v_up	−0.022	0.01	0.022	0.978
	A_a_dn	2.127	0.767	0.006	8.389
	A_v_dn	−0.045	0.014	0.002	0.956
	B_n_dn	0.267	0.056	0	1.306
	B_v_dn	−0.033	0.012	0.008	0.968
	Constant	−2.858	0.185	0	0.057
15	A_a_up	2.245	0.612	0	9.441
	A_v_up	−0.043	0.01	0	0.957
	B_n_up	0.097	0.036	0.007	1.101
	B_v_up	−0.027	0.008	0.001	0.973
	A_v_dn	−0.024	0.008	0.002	0.976
	B_n_dn	0.142	0.03	0	1.152
	Constant	−2.739	0.138	0	0.065
30	A_v_dn	−0.021	0.007	0.001	0.979
	B_a_dn	1.681	0.43	0	5.371
	A_a_up	2.468	0.604	0	11.795
	A_v_up	−0.049	0.009	0	0.952
	B_n_up	0.08	0.019	0	1.083
	B_v_up	−0.013	0.007	0.05	0.987
	Constant	−2.899	0.166	0	0.055
45	A_a_up	3.114	0.583	0	22.508
	A_v_up	−0.044	0.008	0	0.957
	B_n_up	0.053	0.015	0	1.054
	A_n_dn	−0.067	0.026	0.008	0.935
	B_a_dn	2.65	0.461	0	14.158
	Constant	−3.562	0.152	0	0.028
60	A_a_up	3.043	0.63	0	20.973
	A_v_up	−0.044	0.008	0	0.957
	B_n_up	0.04	0.012	0.001	1.041
	A_n_dn	−0.041	0.021	0.045	0.96
	B_a_dn	2.437	0.474	0	11.438
	Constant	−3.528	0.154	0	0.029

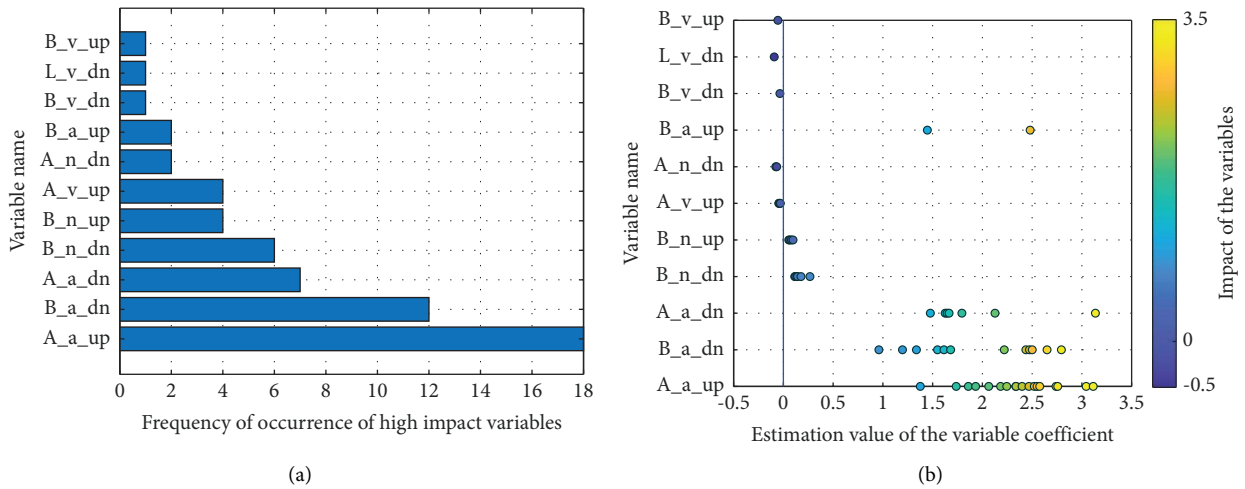

Figure 5: Distribution characteristics and the coefficient estimation of high-impact variables. (a) Frequency of occurrence of high-impact variables in 20 LR results. (b) Coefficient estimation of high-impact variables.

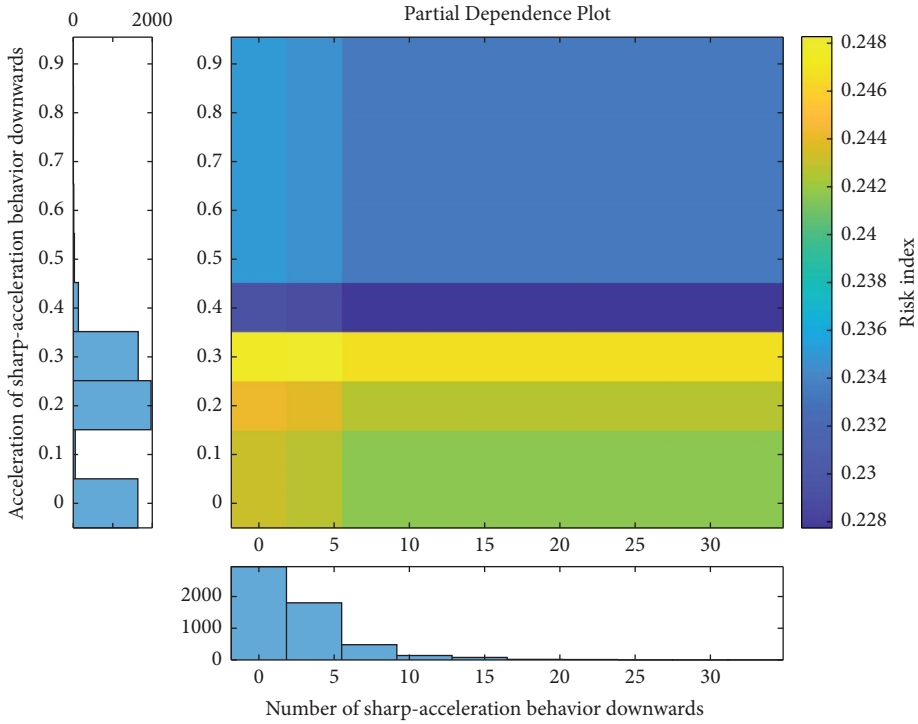

Figure 6: Interaction for acceleration and number of sharp-acceleration behavior.

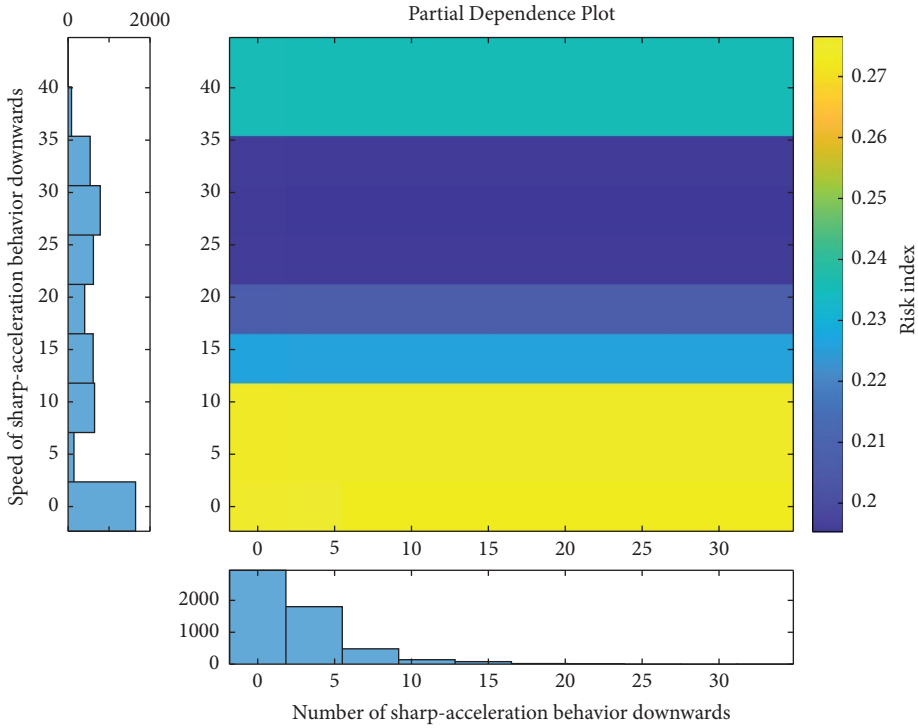

Figure 7: Interaction for speed and number of sharp-acceleration behavior.

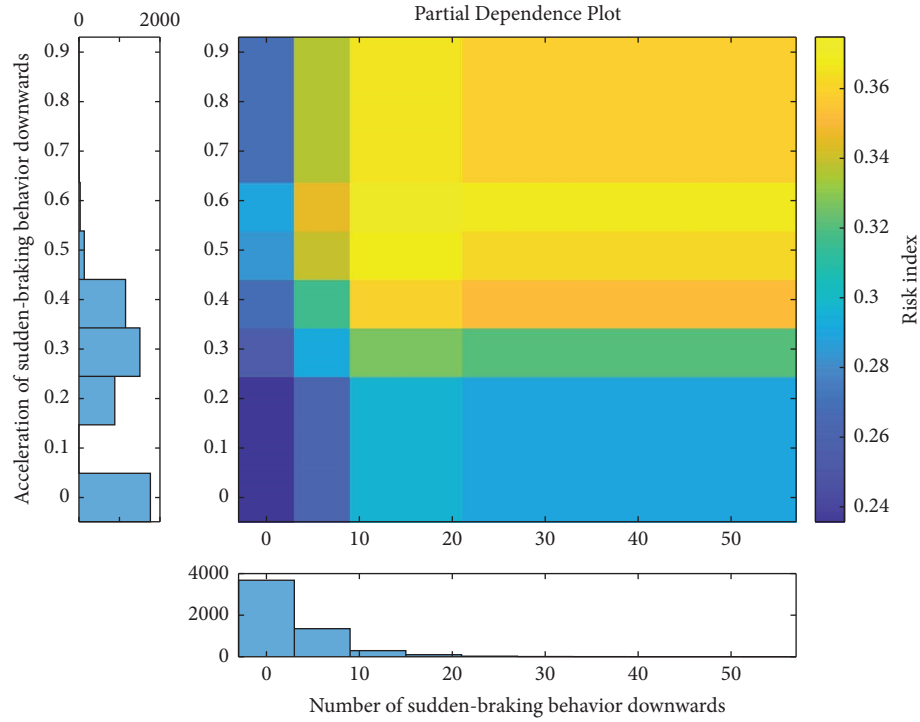

Figure 8: Interaction for acceleration and number of sudden-braking behavior.

acceleration, and acceleration near 0.5 gravitational acceleration during braking. Drivers are promptly reminded to drive carefully in such situations.

5. Conclusion

In order to predict crash risk on a large scale and at a lower cost, real-time driving behavior data provided by AutoNavi onboard GPS were utilized to establish a behavior-driven risk prediction model. The generated datasets contained sharp-left-lane-change, sharp-right-lane-change, sharp-acceleration, and sudden-braking behaviors within 250 meters upstream and downstream of the crash site within a certain time interval. The frequency, speed, and acceleration in the process of these behaviors were calculated as supplementary features. Multiple classification learners were applied to regress the dataset, and PDP was applied to determine the main contributory factors of expressway crash risk.

Primarily, the behavior-driven risk prediction model is established through RUSBoost, with the AUC index reaching 0.782, which overperforms various machine learning models which use traditional traffic-flow data. It is demonstrated that the behavior-driven model has more advantages than the traffic-flow-driven model in risk prediction. Herein, navigation systems can provide corresponding safety monitoring services by using the real-time behavior data.

Furthermore, the results of LR and PDP show that sharp-acceleration and sudden-braking behaviors are the main factors of expressway crash risk. Further study of the interaction of these two behaviors' features demonstrates that sudden-braking is the most critical source of risk. The risk of sharp-acceleration behavior is found on the congested roads with high frequency. When sudden-braking behaviors occur in excess of 0.5 g acceleration, it is imperative to inform drivers of prospective upcoming risks.

In addition, when constructing a dataset, the interval before a crash in collecting real-time driving behavior has a great influence on the regression effect and so does the ratio of the crash and noncrash data to be chosen. Blindly expanding the sampling interval and data ratio would not achieve the best regression effect. After comparing multiple groups of experiments, in this case, real-time driving behavior features within 15 minutes before the crashes are collected and the unbalanced ratio of the dataset is about 1 : 20. The speed and acceleration characteristics in the driving behavior can also effectively influence the prediction accuracy, while data of vehicle motion conditions and traffic operation patterns should be retained as much as possible.

To sum up, this research verifies the feasibility of the behavior-driven risk prediction model by using onboard navigation data and provides a rational basis to apply active countermeasures. Future research will focus on supplementing other traffic-flow data to obtain a highly accurate and interpretable model and developing risk avoidance measures. Last but not least, the legal risk on personal privacy brought about by the use of navigation software to capture driving behavior should also be taken seriously, so as to safeguard traffic safety while maintaining the privacy of users.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

J.L., Y.D.W., and X.C.M. conceptualized the study. X.C.M proposed the methodology, provided the software, and prepared the original draft. J.L and X.C.M. performed data curation and funding acquisition. Y.D.W. reviewed and edited the manuscript. Y.D.W. and J.L. supervised the study. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This study was sponsored by the National Natural Science Foundation of China (52072071), the Jiangsu Province Transportation Science and Technology Project (2022G02), and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (SJCX230081).

References

[1] W. Han and J. Zhao, "Driver behaviour and traffic accident involvement among professional urban bus drivers in China," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 74, pp. 184–197, 2020.

[2] Y. Ma, H. Meng, S. Chen, J. Zhao, S. Li, and Q. Xiang, "Predicting traffic conflicts for expressway diverging areas using vehicle trajectory data," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 3, 2020.

[3] G. Fountas, S. S. Pantangi, K. F. Hulme, and P. C. Anastasopoulos, "The effects of driver fatigue, gender, and distracted driving on perceived and observed aggressive driving behavior: a correlated grouped random parameters bivariate probit approach," *Analytic Methods in Accident Research*, vol. 22, Article ID 100091, 2019.

[4] F. Mannering, C. R. Bhat, V. Shankar, and M. Abdel-Aty, "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis," *Analytic Methods in Accident Research*, vol. 25, Article ID 100113, 2020.

[5] A. Hossain, X. D. Sun, S. Islam, S. Alam, and M. Mahmud Hossain, "Identifying roadway departure crash patterns on rural two-lane highways under different lighting conditions: association knowledge using data mining approach," *Journal of Safety Research*, vol. 85, pp. 52–65, 2023.

[6] E. K. Adanu, A. Lidbe, J. Liu, and S. Jones, "A comparative study of factors associated with motorcycle crash severities under different causal scenarios," *Journal of Transportation Safety and Security*, vol. 15, no. 4, pp. 376–396, 2023.

[7] S. H. Wang, J. Z. Liu, N. Chen, J. J. Xiao, and P. Y. Wei, "How does the built environment affect drunk-driving crashes? A spatial heterogeneity analysis," *Applied Sciences*, vol. 13, no. 21, p. 11813, 2023.

[8] M. Abdel-Aty, O. Zheng, Y. Wu, A. Abdelraouf, H. Rim, and P. Li, "Real-time big data analytics and proactive traffic safety management visualization system," *Journal of Transportation Engineering, Part A: Systems*, vol. 149, no. 8, 2023.

[9] Q. Shi and M. Abdel-Aty, "Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 380–394, 2015.

[10] K. Yang, X. Wang, and R. Yu, "A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation," *Transportation Research Part C: Emerging Technologies*, vol. 96, pp. 192–207, 2018.

[11] X. Shi, Y. D. Wong, M. Z. F. Li, and C. Chai, "Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory," *Accident Analysis and Prevention*, vol. 117, pp. 346–356, 2018.

[12] L. Eboli, G. Mazzulla, and G. Pungillo, "How to define the accident risk level of car drivers by combining objective and subjective measures of driving style," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 49, pp. 29–38, 2017.

[13] A. Kassu and M. Anderson, "Analysis of nonsevere crashes on two- and four-lane urban and rural highways: effects of wet pavement surface condition," *Journal of Advanced Transportation*, vol. 2018, Article ID 2871451, 10 pages, 2018.

[14] M. M. Hossain, H. G. Zhou, S. Das, X. D. Sun, and A. Hossain, "Young drivers and cellphone distraction: pattern recognition from fatal crashes," *Journal of Transportation Safety and Security*, vol. 15, no. 3, pp. 239–264, 2023.

[15] F. L. Wei, Z. G. Cai, P. Liu, Y. Q. Guo, X. Li, and Q. Y. Li, "Exploring driver injury severity in single-vehicle crashes under foggy weather and clear weather," *Journal of Advanced Transportation*, vol. 2021, Article ID 9939800, 12 pages, 2021.

[16] M. B. Johnson, "The prevalence of alcohol-involved crashes across high and low complexity road environments: does knowing where drinking drivers crash help explain why they crash?" *PLoS One*, vol. 17, no. 4, Article ID e0266459, 2022.

[17] T. A. Dingus, F. Guo, S. Lee et al., "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proceedings of the National Academy of Sciences*, vol. 113, no. 10, pp. 2636–2641, 2016.

[18] Y. Lu, K. Cheng, Y. Zhang, X. Chen, and Y. Zou, "Analysis of lane-changing conflict between cars and trucks at freeway merging sections using UAV video data," *Journal of Transportation Safety and Security*, vol. 15, no. 9, pp. 943–961, 2022.

[19] Q. Q. Shangguan, T. Fu, J. H. Wang, R. Jiang, and S. E. Fang, "Quantification of rear-end crash risk and analysis of its influencing factors based on a new surrogate safety measure," *Journal of Advanced Transportation*, vol. 2021, Article ID 5551273, 15 pages, 2021.

[20] X. C. Ma, J. Lu, X. Liu, and W. B. Qu, "A genetic programming approach for real-time crash prediction to solve trade-off between interpretability and accuracy," *Journal of Transportation Safety and Security*, vol. 15, no. 4, pp. 421–443, 2023.

[21] F. Rahman, X. Zhang, and M. Chen, "Evaluating effect of operating speed on crashes of rural two-lane highways," *Journal of Advanced Transportation*, vol. 2023, Article ID 2882951, 13 pages, 2023.

[22] A. Chand, S. Jayesh, and A. B. Bhasi, "Road traffic accidents: an overview of data sources, analysis techniques and contributing factors," *Materials Today: Proceedings*, vol. 47, pp. 5135–5141, 2021.

[23] J. Lee, X. Li, S. Y. Mao, and W. Fu, "Investigation of contributing factors to traffic crashes and violations: a random parameter multinomial logit approach," *Journal of Advanced Transportation*, vol. 2021, Article ID 2836657, 11 pages, 2021.

[24] J. Li, Y. Yang, Y. R. Hu, X. Y. Zhu, N. X. Ma, and X. J. Yuan, "Using multidimensional data to analyze freeway real-time traffic crash precursors based on XGBoost-SHAP algorithm," *Journal of Advanced Transportation*, vol. 2023, Article ID 5789573, 18 pages, 2023.

[25] Y. Ma, J. Zhang, J. Lu, S. Chen, G. Xing, and R. Feng, "Prediction and analysis of likelihood of freeway crash occurrence considering risky driving behavior," *Accident Analysis and Prevention*, vol. 192, 2023.

[26] M. Hossain, M. Abdel-Aty, M. A. Quddus, Y. Muromachi, and S. N. Sadeek, "Real-time crash prediction models: state-of-the-art, design pathways and ubiquitous requirements," *Accident Analysis and Prevention*, vol. 124, pp. 66–84, 2019.

[27] M. Guo, X. Zhao, Y. Yao et al., "A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data," *Accident Analysis and Prevention*, vol. 160, Article ID 106328, 2021.

[28] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Analysis and Prevention*, vol. 136, Article ID 105405, 2020.

[29] C. Gutierrez-Osorio and C. Pedraza, "Modern data sources and techniques for analysis and forecast of road accidents: a review," *Journal of Traffic and Transportation Engineering*, vol. 7, no. 4, pp. 432–446, 2020.

[30] Q. Cai, M. Abdel-Aty, J. Yuan, J. Lee, and Y. Wu, "Real-time crash prediction on expressways using deep generative models," *Transportation Research Part C: Emerging Technologies*, vol. 117, Article ID 102697, 2020.

[31] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "RUSBoost: improving classification performance when training data is skewed," in *Proceedings of the 2008 19th International Conference on Pattern Recognition*, pp. 1–4, Tampa, FL, USA, December, 2008.

[32] Z. Zhang, Q. Nie, J. Liu, A. Hainen, N. Islam, and C. Yang, "Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data," *Journal of Intelligent Transportation Systems*, pp. 1–19, 2022.

[33] C. Yuan, Y. Li, H. L. Huang, S. Q. Wang, Z. H. Sun, and Y. Li, "Using traffic flow characteristics to predict real-time conflict risk: a novel method for trajectory data analysis," *Analytic Methods in Accident Research*, vol. 35, Article ID 100217, 2022.

[34] Z. Elamrani Abou Elassad, H. Mousannif, and H. Al Moatassime, "A real-time crash prediction fusion framework: an imbalance-aware strategy for collision avoidance systems," *Transportation Research Part C: Emerging Technologies*, vol. 118, Article ID 102708, 2020.

[35] L. P. Zhao, F. Li, D. Y. Sun, and F. Dai, "Highway traffic crash risk prediction method considering temporal correlation characteristics," *Journal of Advanced Transportation*, vol. 2023, Article ID 9695433, 13 pages, 2023.