applied sciences

MDPI

Article

Analyzing the Risk Factors of Traffic Accident Severity Using a Combination of Random Forest and Association Rules

Jianyu Wang [1], Shuo Ma [1], Pengpeng Jiao [1,*], Lanxin Ji [1], Xu Sun [1] and Huapu Lu [2]

[1] Beijing Laboratory of General Aviation Technology, Beijing University of Civil Engineering and Architecture, Beijing 100044, China; wangjianyu@bucea.edu.cn (J.W.); mashuo@stu.bucea.edu.cn (S.M.); jilanxin@stu.bucea.edu.cn (L.J.); sunxu@bucea.edu.cn (X.S.)

[2] Institute of Transportation Engineering and Geomatics, Tsinghua University, Beijing 100084, China; luhp@tsinghua.edu.cn

* Correspondence: jiaopengpeng@bucea.edu.cn; Tel.: +86-186-1192-0632

Abstract: This study explores risk factors influencing the at-fault party in traffic accidents and analyzes their impact on traffic accident severity. Based on the traffic accident data of Shenyang City, Liaoning Province, China, from 2018 to 2020, 19 attribute variables including road attributes, time attributes, environmental attributes, and characteristics of the at-fault parties with either full responsibility, primary responsibility, or equal responsibility of the traffic accidents were extracted and analyzed in conjunction with the built environment attributes, such as road network density and POI (points of interest) density at the sites of traffic accidents. Using the RF-SHAP method to determine the relative importance of risk factors influencing the severity of traffic accidents with either motor vehicles or vulnerable groups at-fault, the top ten risk factors influencing the severity of traffic accidents with vulnerable road users as the at-fault parties are: functional zone, density of shopping POI, density of services POI, cause of accident, travel mode, collision type, season, road type, age of driver, and physical isolation. Travel mode, season, and road speed limit are more important risk factors for traffic accidents, with motor vehicle drivers as the at-fault parties. The density of service POI and cause of the accident are less critical for traffic accidents with motor vehicle drivers than traffic accidents with vulnerable road users who are at-fault. Subsequently, the Apriori algorithm based on association rules is used to analyze the important causal factors of traffic accidents, so as to explore the influence mechanism of multiple causal factors and their implied strong association rules. Our results show that most combined factors are associated with the matched Service and Shopping POI features. This study provides valuable information on the perceived risk of fatal accidents and highlights the built environment's significant influence on fatal traffic accidents. Management strategies targeting the most typical combinations of accident risk factors are proposed for preventing fatalities and injuries in serious traffic accidents.

Keywords: crash severity; random forest; shapley additive explanation; Apriori

Citation: Wang, J.; Ma, S.; Jiao, P.; Ji, L.; Sun, X.; Lu, H. Analyzing the Risk Factors of Traffic Accident Severity Using a Combination of Random Forest and Association Rules. *Appl. Sci.* **2023**, *13*, 8559. https://doi.org/10.3390/app13148559

Academic Editor: Luís Picado Santos

Received: 29 May 2023
Revised: 12 July 2023
Accepted: 20 July 2023
Published: 24 July 2023

1. Introduction

Road traffic safety have been an ongoing topic of global debate for many years. The number of deaths caused by motor vehicle collisions are increasing, with the World Health Organization reported in 2020 that around 1.35 million people die of road traffic accidents in the world each year [1]. It is important to understand factors influencing the severity of injuries sustained by participants in road traffic crashes. In previous studies, regression models have been widely used in traffic accident research. Mohammad Abrari Vajari et al. [2] analyzed a dataset of 7714 motorcycle crashes which occurred at intersections in Victoria, Australia, between 2006 to 2018, and used the multinomial logit models to study the accident severity. Jiao et al. [3] used latent clustering analysis and stochastic parametric logit models to investigate factors influencing the crash severity of traffic crashes involving

elderly pedestrians in North Carolina from 2007 to 2019. For exploring the influencing factors affecting the severity of road traffic accidents, we have mainly referred to the selection of variables in many relevant papers, with the academic literature emphasizing five key dimensions, namely the at-fault party characteristics, road characteristics, crash characteristics, time of day, and environmental conditions [4]. Among these, the at-fault party characteristics include the driver's age [5], gender [6], driving experience [7], and mode of transport [8]. Road characteristics include physical separation of the road [9], and road type [10]. Crash characteristics include cause of accident [11], cross-sectional location [12], and whether the crash occurred at an intersection [13]. Temporal and environmental conditions include seasonality [14], weather [15], climatic factors [16], and built environment related variables.

With the rapid development of machine learning, more than 20 machine learning-related algorithms have been adopted to explore factors influencing traffic accident severity, such as artificial neural networks, support vector machines, decision trees, Bayesian networks, and K-Nearest Neighbors. Compared with traditional regression models, machine learning can better handle different types of data, and are capable of analyzing the complex nonlinear relationships among variables [17]. In a study by Shakil Ahmed et al. [18], six different machine learning algorithms, namely Random Forest (RF), Decision Jungle (DJ), Adaptive Boost (AdaBoost), Extreme Gradient Boost (XGBoost), Light Gradient Boost Machine (LGBM), and Classification Boost (CatBoost), were used to analyze road accidents in New Zealand from 2016 to 2020, and they showed RF has the best prediction results. Fatma Outay et al. [19] built a dataset of traffic accidents using motorcycle track data from GPS and road accident feature data collected through questionnaires and police inquiries, and ranked and predicted feature importance using a random forest algorithm. Masello et al. [20] investigated the effect of the driving environment on attempted accidents, speeding and distraction events based on RF and XGBoost combined with SHAP, aiming to explore the factors affecting the frequency of road accidents and driving risks. The most important contextual factors in predicting these risky events are identified and ranked through Shapley Additive Explanations. Wen et al. [21] used the LGBM-SHAP methodology was used to explore the impact of the importance of accident risk factors on different types of accidents based on accident data in Texarkana. The SHAP methodology quantified the results and yielded more evident results. Wang et al. [22] used XGboost and SHAP's LCA-based clustering results to identify the main influencing factors in the potential categories and analyzed the interactions between the factors. The SHAP visualization of their results shows the top 15 accident duration influencing factors regarding importance ranking. All these studies have shown that SHAP is a very effective interpretation method for machine learning and is widely used in various industries.

Most scholars have found that traffic accident severity is influenced by a multitude of factors, hence multi-factor interactions should be explored. Jiang et al. [23] proposed a framework based on association rule mining to identify critical factors influencing motorcycle accident severity, parameter optimization of parameter thresholds in association rule mining is proposed to identify individual key factors from two-item rules and boosting factors from multiple-item rules. Samerei et al. [24] conducted a study on bus accidents which occurred in Victoria, Australia between 2006 and 2019, and discovered effective chains of factors influencing fatalities in bus accidents using association rules. Kong et al. [25] applied an association rule mining method on a natural driving dataset to study the correlations between accidents and road features. The item set frequencies after the Apriori algorithm showed that the accidents were strongly associated with roadways without access control, driving during non-peak hours, roadways without a shoulder or a median, roadways with the minor arterial functional class, and roadways with a speed limit between 30 and 60 mph. Xu et al. [26] applied association rule mining techniques on traffic accidents with high number of casualties that occurred in China between 2009 and 2013, and the results

showed that traffic accidents with high casualties are affected by the interactions of different complex factors.

Based on the relevant research by previous scholars, it is clear that the Random Forest algorithm is superior in exploring the influence of a single factor and the association rule algorithm shows advantages in exploring the multi-factor interaction effects. Therefore, this study adopts an integrated approach of RF and Apriori algorithm to explore the mechanisms of single factors and multi-factor interactions. The remainder of this paper is organized as follows: Section 2 provides an overview of the processing of the data and methodology, Section 3 shows and discusses the results, and Section 4 summarizes the implications of the results for decisions on enhancing road traffic safety.

2. Materials and Methods

2.1. Data

This paper is based on data on traffic accidents in Shenyang City, Liaoning Province, China, 2018–2020, extracted from the at-fault parties (full responsibility, main responsibility, and equal responsibility). The Shenyang City Traffic Police provided the data, and certain sensitive data with irrelevant features or privacy implications to this study were removed in advance, such as property damage and driver identification numbers. A total of 2022 accident data were processed after data cleaning. According to the classification of the severity of road traffic accidents by China's traffic management authorities, traffic accidents are classified into three categories: accidents causing only property damages, accidents causing injuries, and accidents causing fatalities. Many scholars classify road traffic accidents into two categories: fatal and non-fatal. Since the focus of this study is to reduce accident fatalities, the scholars' definition of accident severity is used. The dependent variable in this study is accident severity, coded as fatal accidents and non-fatal accidents, where non-fatal accidents are coded as 0 and fatal accidents are coded as 1, accounting for 60.4% and 39.6% of the traffic accidents, respectively.

The independent variables include five aspects: Crash Attributes at Fault, Infrastructure Attributes, Road Attributes, Environmental and Time Attributes. The study in this paper considers the characteristics of the land surrounding the road as a potential influence on the severity of accidents as well, and therefore in the context of environmental factors, in addition to the general environmental features provided by the traffic police, such as weather and temperature, this paper uses ArcGIS 10.6 software to establish a buffer zone with a radius of 500 m by combining the latitude and longitude of the accident site and the Point of Interest in Shenyang. It calculates the road network density, commercial and residential density, scientific and cultural density, restaurant and shopping density and living service density in the buffer zone of the accident site as the built environment features. Each data sample contains 24 features, as shown in Table 1.

A total of 401 traffic accidents has vulnerable road users as the at-fault party, including pedestrians, bicycle drivers, electric bicycle drivers, and motorcycle riders. This constitutes 33.9% of all fatal accidents. In this study, we model and analyze traffic accidents with vulnerable road user and motor vehicle drivers as the at-fault parties separately.

Table 1. Variable summary.

Variable Type	Variable	Description	Description	Proportion (%)
Dependent Variables	Crash Injury levels	The severity of a crash based on the most severe injury to any person involved in the crash.	0 = Non-fatal 1 = Fatal	60.4% 39.6%
Crash Attributes at Fault	Gender of driver	The sex of person involved in a crash.	0 = Male 1 = Female	89.3% 10.7%
	Age of driver	The age of driver involved in a crash. If it not available, the approximate age.	0 $\leq$ 25 years 1 = 26–45 years 2 = 46–60 years 3 > 60 years	8.8% 56.7% 26.7% 7.9%
	Driving experience	The number of years a driver has been licensed to drive	0 = 0–6 years 1 = 7–16 years 2 > 16 years	45.4% 36.8% 17.8%
	Liability	The liability of driver in the accident is determined	0 = Full Liability 1 = Primary Liability 2 =Equal Liability	47.7% 29.2% 23.1%
	Travel mode	The type of vehicle by the driver	0 = Pedestrian 1 = Non-Motorized Vehicle 2 = Motorcycle 3 = Motorcar 4 = Buggy 5 = Large Passenger Truck. 6 = Other	1.0% 12.0% 6.8% 55.0% 7.4% 16.5% 1.2%
Infrastructure Attributes	Cause of Accident	The cause of the accident (these data are generally determined by the police at the time of the accident determination)	0 = Improper operation of the driver 1 = Overspeed or overloading 2 = Drunk or fatigued driving 3 = Failure to give way as required 4 = Hit-and-run 5 = Failure to follow signal instructions 6 = Other violations	11.0% 5.8% 9.3% 9.4% 1.4% 4.0% 59.1%
	Collision Type	The types of participants in accident.	0 = Single vehicle accident 1 = Person-vehicle accident 2 = Vehicle-vehicle accident	6.8% 25.1% 68.2%
	Position	The location of the road cross-section of the accident.	0 = Non-motor vehicle lane 1 = Motor vehicle lane 2 = Mixed lane of motor vehicles and non-motor vehicles 3 = Other	5.6% 72.0% 13.9% 8.5%
	Crossing or not	Whether the accident occurred in intersections.	0 = No 1 = Yes	65.4% 34.6%
	Functional Zone	Indicates if the crash occurred within a municipality (Urban) or in a Rural location.	0 = Urban District 1 = Suburban District 2 = Rural District	57.1% 25.8% 17.1%
Road Attributes	Road Type	Route class of the On Road	0 = Other 1 = Trunk Road 2 = Secondary and Tertiary Roads 3 = Primary Roads and Highways. 4 = Urban Expressways	5.0% 67.7% 13.8% 8.0% 5.5%
	Speed Limit	Authorized speed limit for the vehicle at the time of the crash. (km/h)	0 $\leq$ 20 km/h 1 = 20–40 km/h 2 = 40–60 km/h 3 = 60–80 km/h 4 $\geq$ 80 km/h	37.9% 41.8% 10.3% 8.4% 1.6%
	Physical Isolation	The type of physical isolation facilities set up at the point of accident.	0 = No Isolation 1 = Isolation Only Between Motor and Non-motor Vehicle 2 = Only Central Isolation 3 = Full Isolation	68.9% 2.3% 22.7% 6.1%

Table 1. *Cont.*

Variable Type	Variable	Description	Description	Proportion (%)
Environmental and Time Attributes	Weekday or not	Whether the accident occurred on a weekday.	0 = No 1 = Yes	39.4% 60.6%
	Rush Hour or not	Whether the accident occurred during rush hour. (Peak hours are set from 7:00–9:00; 17:00–19:00)	0 = No 1 = Yes (7:00–9:00; 17:00–19:00)	65.3% 34.7%
	Night Time or not	Whether the accident occurred at night.	0 = No 1 = Yes	80.1% 19.9%
	Extreme Temperatures	Whether the temperature is higher than 30 °C or lower than 0 °C on the day of the accident	0 = No (0–30 °C) 1 = Yes (<0 °C or >30 °C)	78.2% 24.4%
	Season	The season in which the accident occurred. (Due to the special geographical location of Shenyang, spring and autumn are shorter, while winter is longer)	0 = Spring (4–5) 1 = Summer (6–8) 2 = Autumn (9–10) 3 = Winter (1–3; 11–12)	26.5% 29.5% 8.9% 35.1%
	Weather	The general atmospheric conditions that existed at the time of a crash.	0 = Sunny 1 = Cloudy 2 = Rain 3 = Fog 4 = Snow	90.8% 4.4% 4.1% 0.1% 0.6%
	Network Density	The density of road network in the buffer zone (km/km^2)	0 $\leq$ 10 km/km^2 1 = 10–20 km/km^2 2 > 20 km/km^2	75.0% 23.9% 1.1%
	Shopping-POI	The density of restaurant and shopping centers in the buffer zone (pcs/km^2)	0 $\leq$ 50 pcs/km^2 1 = 50–500 pcs/km^2 2 > 500 pcs/km^2	43.0% 41.9% 15.1%
	Education-POI	The density of scientific, educational and cultural facilities in the buffer zone (pcs/km^2)	0 $\leq$ 50 pcs/km^2 1 = 50–500 pcs/km^2 2 > 500 pcs/km^2	80.2% 19.8% 0.0%
	Commercial-POI	The density of commercial and residential facilities in the buffer zone (pcs/km^2)	0 $\leq$ 50 pcs/km^2 1 = 50–500 pcs/km^2 2 > 500 pcs/km^2	97.2% 2.8% 0.0%
	Service-POI	The density of living service in the buffer zone (pcs/km^2)	0 $\leq$ 50 pcs/km^2 1 = 50–500 pcs/km^2 2 > 500 pcs/km^2	57.5% 42.4% 0.1%

2.2. Random Forest

The Random Forest method is a non-parametric supervised learning method that belongs to the Bagging type of ensemble learning. By combining multiple weak classifiers, the final result is averaged by voting or taking the mean, which makes the result of the overall model have high accuracy and generalization performance [27]. Generally, there is no overfitting on noisy data. It can also obtain better results, has better classification accuracy, and does not produce overfitting problems. So, the random forest is chosen as the classification model, often applied in traffic accident injury studies to rank the importance of accident severity risk factors.

When using a dataset to train a model, the database is first divided into a training dataset (70%) and a test dataset (30%) using a stratified sampling approach, which reduces errors due to sample distribution and ensures that the imbalance rate is consistent between the training and test datasets. Random Forest generates a new training set (N_t) by repeatedly randomly selecting T samples from the original set N by a self-service resampling technique with put-back. Subsequently, T new training sets and corresponding decision trees are used to form the random forest model. The random forest involves many parameters, the most important of which are as follows:

(1) n_estimators: This is the number of decision trees in the random forest, i.e., the number of base evaluators. Theoretically, more trees create better model, but it is also more likely for models with more trees to encounter problems such as model overfitting and long model computation time. Hence, a reasonable number of decision trees will often achieve good results;

(2) max_depth: this parameter indicates the maximum tree depth. When the decision tree splits and reaches the depth set by this parameter, it will stop splitting, i.e., branches exceeding max_depth will be cut off;
(3) max_features: this parameter reflects the number of features to be randomly selected by each base evaluator in the random forest when generating the tree, and its default value is the squared-off integer of the total number of features in the dataset;
(4) min_samples_leaf: in the decision tree splitting process, if the number of samples in a child node generated by a node after splitting is less than this value, the node will not be split;
(5) min_samples_split: if the number of samples in a node is less than this value, the node is a leaf node and will not be split.

In the actual modeling process, the parameters do not need significant adjustments to achieve good classification prediction results. In this study, two main parameters affecting the tuning performance of random forest are selected, including the total number of decision trees (n_estimators) and the maximum tree depth (max_depth).

The steps of the algorithm are as follows:

Step 1: For a given sample set N consisting of X_1, X_2, ..., X_k, construct a set of random vectors N_1, N_2, ..., N_T through T random repeatable samples.
Step 2: Construct a decision tree based on each random vector N_t.
Step 3: Repeat steps 1 and 2 to obtain T decision trees.
Step 4: Use the obtained T decision trees to vote on the input variables X_k.
Step 5: Calculate all the votes and find out the value with the highest number of votes among all the predictions as the classification label of the input variable X_k. When generating each decision tree, calculate the out-of-bag error rate, denoted as E_{OOB1}, and at the same time, after adding random noise for feature X_k, calculate the value again, denoted as E_{OOB2}, then the importance of feature X_k is:

$$I_{X_k} = \frac{1}{\mathrm{T}}\sum (E_{OOB2} - E_{OOB1})_t \quad (1)$$

Repeating Equation (1), then the importance of all features can be calculated and ranked.

2.3. SHAP (Shapley Additive Explanations)

The problem of interpretability still needs to be solved for ensemble learning methods. The game theory-SHAP (SHapley Additive exPlanations) method solves this problem well. For integrated tree models, the model output is a probability value when doing a classification task. Thus, SHAP attributes the output value to the shapely value of each feature to measure the effect of the feature on the final output value.

SHAP value can be used as a unified approach to interpret the output of any machine learning model [28]. Traditional feature importance can find the most influential features among hundreds of features, but it is impossible to know how each feature influences the dependent variable. The most significant advantage of SHAP value is that it can reflect the influence of the features in each sample and indicate the direction of the influence (positive/negative). Therefore, to improve the interpretability of the machine learning model, this study analyzes the risk factors of accident severity using the RF-SHAP method.

In SHAP, the importance of each feature is assigned according to its respective marginal contribution. A positive SHAP value of a feature influences the final prediction result positively; a negative SHAP value of a feature influences the final prediction result negatively. The larger the SHAP value of a feature, the greater its influence.

2.4. Association Rules

Currently, there are two relatively developed association rule mining algorithms: Apriori algorithm and FP-Growth algorithm. This study uses the Apriori algorithm to mine association rules for rear-end accident data.

The association rule algorithm can mine the intrinsic connection between elements, represented by $(X \cap Y)$, where X is the event that occurs in the preceding item and Y is the event that occurs in the following item. Generally, the association rule algorithm has three metrics: support, confidence, and lift. Support is the probability of event X and event Y occurring simultaneously, denoted Support$(X \cap Y)$:

$$\text{Support}(X \cap Y) = \frac{\text{Freq}(X \cap Y)}{N} \quad (2)$$

where N is the total number of events and Freq$(X \cap Y)$ is used to denote the number of events occurring at the same time as X and Y. Confidence is the probability of Y occurring after the occurrence of event X, denoted Confidence$(X \cap Y)$:

$$\text{Confidence}(X \cap Y) = \frac{\text{Freq}(X \cap Y)}{\text{Freq}(X)} \quad (3)$$

The degree of elevation indicates the elevating effect that event X has on the probability of the occurrence of event Y. It is used to determine whether the rule has practical value and is denoted Lift$(X \cap Y)$:

$$\text{Lift}(X \cap Y) = \frac{\text{Support}(X \cap Y)}{\text{Support}(X) \times \text{Support}(Y)} \quad (4)$$

A higher support indicates a higher probability of occurrence of the antecedent. A higher confidence indicates a higher probability of occurrence of the antecedent and a higher probability of occurrence of the consequent. An elevation greater than one indicates a more positive correlation between the antecedent and the consequent. However, a lift > 2 is a recognizable correlation rule in practice. If lift = 1, it means that there is no significance, A and B are independent of each other and do not affect each other; if lift < 1, it means that there is an opposite effect. If A event occurs, B event is less likely to coincide.

3. Results and Discussion

3.1. Feature Importance Ranking

In this study, we use Python 3.9 software to build and interpret the model through the 'Random Forest' package and 'SHAP' package. The finding ranges of n_estimators and max_depth are determined by grid search and cross-validation, the total number of decision trees in the model and the two parameters of max_tree_depth are optimally tuned, and the rest of the parameters are set as follows: max_features = 'auto'; min_ samples_leaf = 1; min_samples_split = 2.

The most accurate training set of 0.813 is obtained when max_depth = 6, n_estimators = 25 for the RF model of vulnerable road user. For the RF model on traffic accidents with motor vehicle drivers at-fault, when max_depth = 6 and n_estimators = 50, the most accurate rate of 0.742 is obtained for the training set. The risk factor importance ranking results of accident severity obtained by RF-SHAP method that are ranked according to their average absolute SHAP values.

Figure 1a,b shows the average absolute SHAP values of each characteristic in traffic accidents with either vulnerable road user or motor vehicle driver at-fault. Figure 1a shows that the top ten characteristics with the highest to lowest degree of influence on vulnerable road user are area division, restaurant and shopping POI density, life service POI density, cause of accident, traffic mode, collision type, season, road type, age, and physical separation. Figure 1b shows that the top ten characteristics with the highest to lowest degree of influence on motor vehicle driver are traffic mode, restaurant and shopping POI density, area division, season, road speed limit, crash type, life service POI density, cause of accident, road network density, and driving experience.

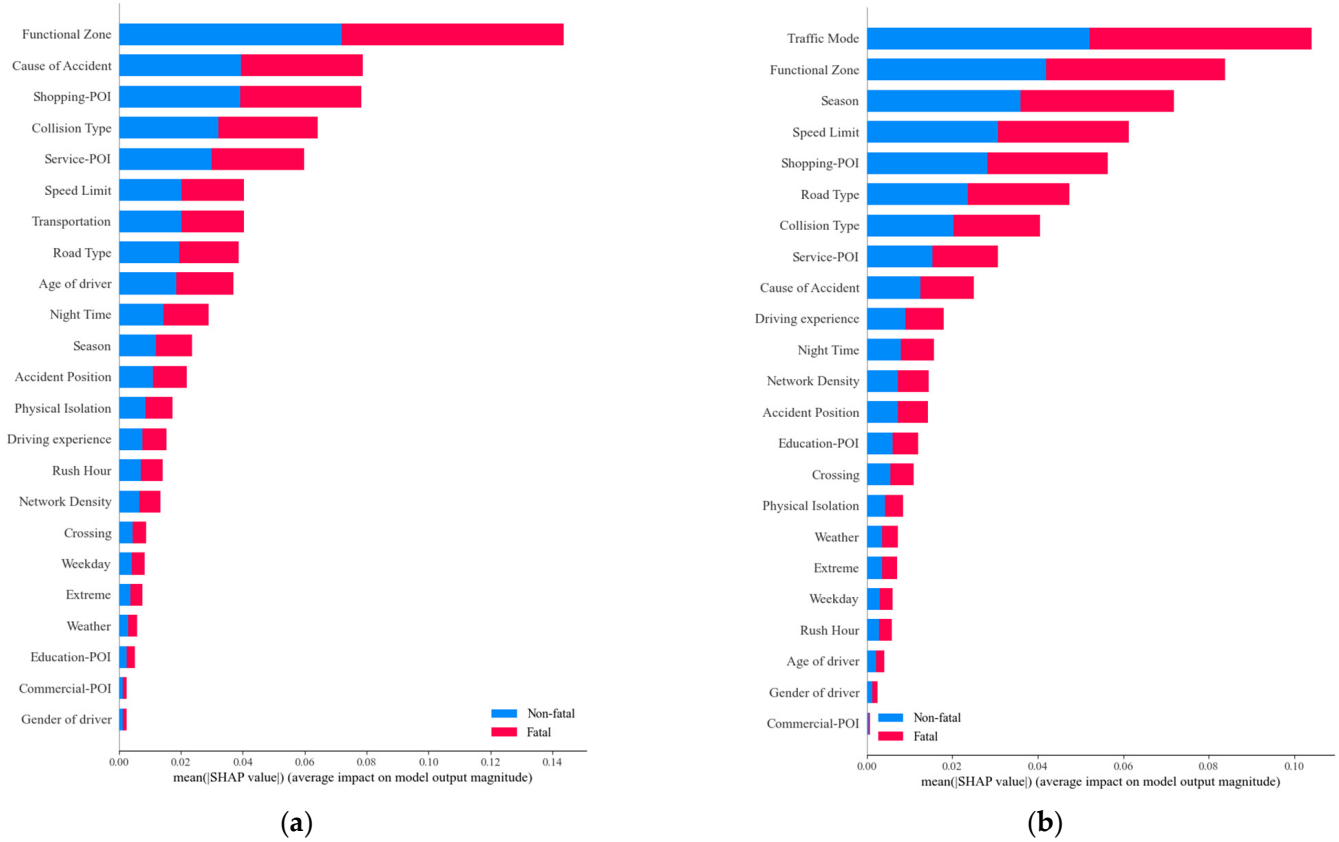


Figure 1. SHAP bar plots. (**a**) vulnerable road user; (**b**) motor vehicle driver.

The comparison found that for traffic accidents with motor vehicle drivers at-fault, traffic mode, season and road speed limit are more critical, whereas POI density of commercial and residential establishments have minor to negligible importance. The POI densities of living services and cause of accident have lower influence on traffic accidents with motor vehicle drivers at fault than traffic accidents with vulnerable road users at fault. The POI densities of science, education and culture and commercial housing were lower than those of restaurants, shopping and living services, and they also have lesser importance over accident severity.

The SHAP value plot illustrates the positive and negative relationship between the explanatory variable and target variables. Each point represents a sample, and the colors in the plot from blue to red denote low to high SHAP values of the different factors, respectively. The clustering of points with the same color indicates that more data correspond to that characteristic value. The greater the SHAP value corresponding to the eigenvalue, the more serious the traffic accident. Figure 2a,b shows the top ten important risk factors for traffic accidents with either vulnerable road user or motor vehicle driver at fault, and the positive or negative impact of each feature on the severity of traffic accident. From both the perspectives of vulnerable road users and motor vehicle driver, factors such as traffic mode, functional zone, road type, season, cause of accident, collision type, density of shopping POI and density of service POI have significant impact on accident severity. Therefore, the analysis of the impact of each characteristic for the different perspectives shows that:

(1) For the mode of transport, vulnerable road users are transported on foot, by bicycle (including tricycles and e-bikes), and by motorbike. The aggregation in the image shows that motorbike driving has a higher number of accidents and a positive SHAP value, which corresponds to a higher probability of fatal accidents. At the same time, motor vehicle drivers drive motorcars, minivans, large passenger trucks, or other models. Although more drivers were on the road in minibuses, vehicle type

significantly affected fatal accidents when the at-fault driver's mode of transport was a minivan or large passenger truck than when the at-fault driver was driving a motor vehicle.

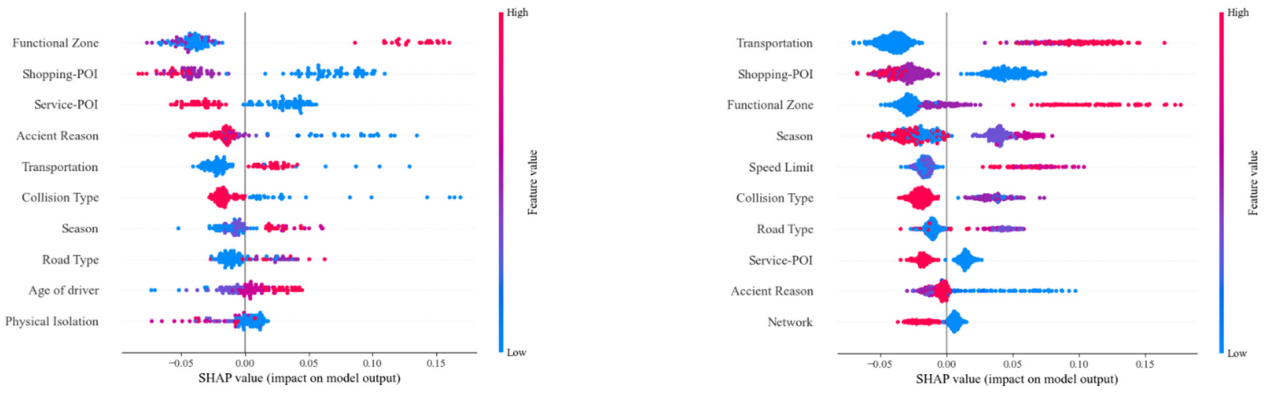
(2) The regional division of the location of the accident point (0 for urban, 1 for peri-urban, and 2 for far-urban), for both vulnerable road user or motor vehicle driver primary responsibility accidents showed that the urban area hurt the severity of the accident, and the far-urban area had a positive effect on it. That may be related to differences in design standards between urban and rural roads. Through the understanding of Shenyang's urban development, the main urban and peri-urban areas have a higher level of economic development than the far suburban areas, with a relatively high level of infrastructure protection, and thus are more inclined to have minor accidents. However, more non-serious crashes occur on urban roads. The risk of death is higher on rural roads, the same conclusion also obtained by Cabrera-Arnau et al. [29].

(3) For road types, it can be seen that lower eigenvalues have more significant accident aggregation and hurt accident severity. As its eigenvalue increases, it positively affects accident severity. In other words, more accidents occur on trunk roads or low-grade roads, but their severity is usually lower. In contrast, urban motor and expressway accidents are usually more severe, the same as the results obtained in previous studies from Goswamy et al. [30].

(4) For seasons, the severity of traffic accidents with vulnerable road user at-fault corresponds to several eigenvalues. Samples with positive SHAP values are mainly composed of red and pink dots, indicating accidents in autumn and winter are usually more severe. Shenyang has a long winter with low temperatures and heavy snowfall lasting from November to March, resulting in icy roads and reduced visibility that leads to more severe traffic accidents.

(5) As the second most important feature, the higher the value of restaurant and shopping POI density (exceeding 50 pcs/km^2), the lower the probability of a fatal accident. The same pattern is observed for life service POI density. This indicates that drivers are more likely to concentrate and drive their vehicles cautiously in densely populated areas.

(6) For the vulnerable road user, the age of the driver and the setting of physical separation of the road are equally important, whereas road speed limits and the density of the road network have essential influences on traffic accidents with motor vehicle drivers at-fault.



(**a**) (**b**)

Figure 2. SHAP summary plots. (**a**) vulnerable road user; (**b**) motor vehicle driver.

3.2. Association Rules

Association rules are used to identify a set of risk factors that are often present concurrently in traffic accidents and examine how various accident characteristics are related [31].

This study investigated the effects of interactions between different influencing factors on specific outcomes. The top ten characteristic variables from RF-SHAP's importance ranking results were selected for multifactor interaction analysis. Significant features between different factors were found using association rule analysis.

Our study focuses on reducing fatal accidents; this section selects 137 and 665 fatal accidents with vulnerable road users and motor vehicle drivers as parties of primary responsibility, respectively. The classical Apriori algorithm in association rules is used to perform the analysis and the support, confidence and boosting thresholds were set. According to past research which applied the association rules to analyze traffic safety and in order to improve the strength and accuracy of association rules, the thresholds of the three indicators in this study were set as support $\geq$ 10%, confidence $\geq$ 80%, and boost $\geq$ 1.5, respectively. Two, three and four association rule results of traffic accidents and the corresponding support, confidence, and boost values are calculated. The association rules were ranked according to the elevation values in each table from highest to lowest. It should be emphasized that some association rules may have low support in the bumper-to-bumper database, but they may be necessary because the features contained in the rules are rarely present. This means the elevation value is more important than the other two criteria for determining the strength of association rules.

It can be seen from Table 2a,b that among the dichotomous association rules, significant association rules in traffic accidents with vulnerable road users at-fault are related to road type, road physical isolation, collision type, and functional zone. Rules #1 and #3 have the highest elevation values in Table 2a, which indicate that if a fatal accident occurred on a median-separated road, it is likely to have happened in the urban area and on a trunk road. Rules #2 and #5 indicate that fatal accidents are more likely to occur on a trunk road when the density of life service POI is in the range of 50 to 500 pcs/km^2 or in urban areas. Compared with high-grade roads, trunk roads in urban areas are generally connected to residential and commercial areas, with more blind spots for drivers and more chaotic road appurtenances. The higher density of life service POI also equates to more vulnerable road users and more random travel paths, which are likely to result in fatal accidents on roads with only middle separation and no guarantees of motor vehicle right-of-way. Therefore, the road environment should be analyzed, with appropriate additional machine-non-motorized separation zones and clear signage. Traffic control should be strengthened in accident-prone areas, and the safety awareness of non-motorized plans and pedestrians should be enhanced to avoid illegal crossings to prevent fatal accidents.

Table 2. Binomial association rules.

No.	head_set	tail_set	Support	Confidence	Lift
(**a**) vulnerable road user					
1	['Physical Isolation is Only Central Isolation']	['Functional Zone is Urban District']	0.132	0.857	2.199
2	['Service-POI is 50–500 pcs/km^2']	['Road Type is Trunk Road']	0.199	0.9	1.827
3	['Physical Isolation is Only Central Isolation']	['Road Type is Trunk Road']	0.132	0.857	1.74
4	['Collision Type is Single vehicle accident']	['Road Type is Trunk Road']	0.125	0.85	1.725
5	['Functional Zone is Urban District']	['Road Type is Trunk Road']	0.316	0.811	1.647
(**b**) motor vehicle driver					
1	['Road Type is Secondary and Tertiary Roads']	['Shopping-POI is $\leq$50 pcs/km^2']	0.145	0.892	2.022
2	['Functional Zone is Rural District']	['Shopping-POI is $\leq$50 pcs/km^2']	0.185	0.851	1.927
3	['Road Type is Secondary and Tertiary Roads']	['Service-POI is $\leq$50 pcs/km^2']	0.157	0.969	1.633
4	['Functional Zone is Rural District']	['Service-POI is $\leq$50 pcs/km^2']	0.195	0.897	1.511

The essential association rules in Table 2b from the viewpoint of motor vehicle drivers are related to the POI density of shopping, POI density of service, road type, and functional zone. Four association rules are obtained by excluding rules which the preceding and following items are POI density. The rules show that when a fatal accident take place in urban areas or on secondary and tertiary roads, these are also more likely to be areas with

POI densities less than 50 pcs/km^2 for services and shopping. Motor vehicle drivers are more likely to drive at faster speeds and lower vigilance on high-grade roads due to more space, flatter road surfaces and higher speed limits, which constitute critical reasons for severe and even fatal accidents.

If Lift = 2, indicating that the antecedent has already occurred, the probability of the posterior occurring is twice the probability of the posterior occurring in the database. The higher the Lift value, the more likely it is that the simultaneous occurrence of antecedent and consequent terms in an event is not coincidental. Therefore, to demonstrate that the antecedent and posterior terms of the rules are interdependent, we extracted three association rules with lift values exceeding 2, containing eight and six rules, respectively, in Table 3a,b. All posterior terms are associated with density of Service-POI and Shopping-POI. As shown in Table 3a, for traffic accidents with vulnerable road users at-fault, the posterior terms of rules #1–#4 are the density Service POI area of 50–500 pcs/km^2, and the antecedent terms contain the high-density Shopping-POI densities. All four rules have the same confidence and lift values of 1 and 2.46, indicating strong correlation. Thus, several types of traffic accidents are more likely to be fatal, including crashes which occurred in urban areas with high restaurant shopping POI densities, crashes which occurred on trunk roads, crashes where non-motorized vehicles are primarily responsible for vehicle–vehicle collisions, and crashes which occurred in areas with medium to high-density Service POI. It is similar to the conclusion of the two association rules above. Therefore, for traffic accidents where the vulnerable road users are primarily responsible, separating the non-motorized lane from the pavement area in urban areas with high restaurant and shopping POI densities is necessary for minimizing the likelihood of vehicle–pedestrian collisions. Due to the instability of two-wheelers, increased efforts should be made to remind motorbike and electric bike riders to wear helmets on the road. Rules #5 and #6 point out that when the traffic accidents occur in low-density Service POI areas and Rural District or on Secondary and Tertiary Roads, the density of Shopping POI around the site of the fatal accident is low. Therefore, safety signage and surveillance should be enhanced in remote suburban areas and on low-grade roads in areas with low Service and Shopping POI densities. Consideration should be given to reduce speed limit on roads with high accident rates to prevent fatal accidents.

In addition, we show that the number of three-association rules with a lift value above 2.0 is higher than the number of two-association rules with high lift values, which indicates that traffic accidents are the products of multiple influencing factors [32]. Moreover, traffic accidents with vulnerable road users at-fault have more association rules with lift values above 2.0, as vulnerable road users are associated with higher crash rates due to their travel patterns that increase the likelihood of exposure [33]. Therefore, to investigate factors influencing fatal crashes for which the vulnerable road users are primarily responsible, it is more important to emphasize analysis on the mechanisms of multi-factor interactions.

Table 3. Three association rules.

No.	head_set	tail_set	Support	Confidence	Lift
(**a**) vulnerable road user					
1	['Collision Type is Vehicle-vehicle accident' and 'Shopping-POI is >500 pcs/km^2']	['Service-POI is 50–500 pcs/km^2']	0.122	1	2.46
2	['Functional Zone is Urban District' and 'Shopping-POI is >500 pcs/km^2']	['Service-POI is 50–500 pcs/km^2']	0.115	1	2.46
3	['Shopping-POI is >500 pcs/km^2'and 'Road Type is Trunk Road']	['Service-POI is 50–500 pcs/km^2']	0.15	1	2.46
4	['Traffic Mode is Non-Motorized Vehicle' and 'Shopping-POI is > 500 pcs/km^2']	['Service-POI is 50–500 pcs/km^2']	0.11	1	2.46
5	['Functional Zone is Rural District' and 'Service-POI is $\leq$50 pcs/km^2']	['Shopping-POI is $\leq$50 pcs/km^2']	0.185	0.949	2.149
6	['Road Type is Secondary and Tertiary Roads' and 'Service-POI is $\leq$50 pcs/km^2']	['Shopping-POI is $\leq$50 pcs/km^2']	0.145	0.921	2.086
7	['Traffic Mode is Motorcycle' and 'Functional Zone is Rural District']	['Shopping-POI is $\leq$50 pcs/km^2']	0.12	0.906	2.052
8	['Physical Isolation is No Isolation' and 'Road Type is Secondary and Tertiary Roads']	['Shopping-POI is $\leq$50 pcs/km^2']	0.145	0.892	2.022
(**b**) motor vehicle driver					
1	['Traffic Mode is Motorcar' and 'Shopping-POI is >500 pcs/km^2']	['Service-POI is 50–500 pcs/km^2']	0.109	0.994	2.323
2	['Road Type is Trunk Road' and 'Shopping-POI is >500 pcs/km^2']	['Service-POI is 50–500 pcs/km^2']	0.141	0.991	2.315
3	['Functional Zone is Urban District' and 'Shopping-POI is >500 pcs/km^2']	['Service-POI is 50–500 pcs/km^2']	0.126	0.99	2.313
4	['Functional Zone is Rural District' and 'Service-POI is $\leq$50 pcs/km^2']	['Shopping-POI is $\leq$50 pcs/km^2']	0.13	0.934	2.187
5	['Service-POI is $\leq$50 pcs/km^2' and 'Road Type is Secondary and Tertiary Roads']	['Shopping-POI is $\leq$50 pcs/km^2']	0.112	0.879	2.06
6	['Road Type is Secondary and Tertiary Roads' and 'Network is $\leq$10 km/km^2']	['Shopping-POI is $\leq$50 pcs/km^2']	0.11	0.861	2.016

4. Conclusions

This study uses police-reported traffic accident data from Shenyang, Liaoning Province, China. The accidents were separated into two categories based on the at-fault parties: vulnerable road users or motor vehicle drivers. A random forest model was developed to analyze the importance of factors influencing traffic accident severity. Then the association rule analysis is used to study the combination mechanism of traffic accident influencing factors. This paper introduces Random Forest and SHAP to determine the critical risk factors. RF algorithm improves model accuracy using less training time when dealing with categorical variables and about a thousand datasets. Introducing a priori algorithms helps us find the association between different factors more clearly. The main findings are summarized as follows:

(1) Descriptive statistical analysis and classification of variables were performed on the dataset, and the importance of 24 characteristic factors was assessed using RF-SHAP. The results show that for accidents in the vulnerable road user category, factors such as area division, restaurant and shopping POI density, life service POI density, cause of accident and traffic mode exert a key influence on accident severity. For accidents in the motor vehicle driver group, factors such as mode of transport, the density of restaurant and shopping POIs, zoning, season, road speed limit, and type of collision significantly influence the fatality of traffic accidents.

(2) This paper focuses on the first ten characteristic variables for the critical influencing factors under the dual perspective of accidents. The Apriori algorithm was used to delve into the mechanism of multi-factor interactions in fatal accidents. Our results

show that most combinations of the factors that occur contain Service and Shopping POI density features. Therefore, it is essential to pay more attention to the vital influence of the built environment around the accident site on fatal accidents and to increase the planning and management of land use to propose more detailed measures.

(3) Areas with High POI density are more common in urban regions, with more non-motorized vehicles and pedestrians. Isolation between motor and non-motor vehicles on high-grade road sections, enhanced management of the road speed limits, and clarifications on the right of way can reduce the likelihood of fatal accidents. For sub-urban roads, fewer pedestrians and non-motorized vehicles make it easier for drivers to increase their speed and relax their vigilance whilst driving. Therefore, reducing speed limits on roads with high accident rates, and increase efforts on reminding motorbike and electric bike riders to wear helmets can prevent fatal accidents.

Worth noting that article collects data from a single city in Shenyang for three years to analyze the data, and our research framework and methodology are transferable to other time zones. However, the results and conclusions obtained may only apply equally to cities of similar size, population size, and geographic location, which is inconvenient for road managers and policy makers.

For future studies, we can perform multi-year, multi-city comparisons of accident characteristics, consider the heterogeneity of the unobserved factors, and perform cluster segmentation followed by a discussion of the importance ranking of the factors and a combination analysis. In summary, this study uses machine learning techniques to explore the interaction mechanisms of single risk factors on the severity of traffic accidents with different at-fault parties. The poor interpretability of machine learning models was compensated using SHAP, and Apriori algorithm was utilized to explore the combination effects of multiple factors. The integrated use of tree models and association rules should not be seen as a substitute for other modelling techniques, but rather as a complementary descriptive method for conducting road safety research.

Author Contributions: Conceptualization, J.W. and X.S.; methodology, J.W.; software, S.M. and L.J.; validation, X.S. and J.W.; formal analysis, S.M. and J.W.; investigation, S.M.; resources, J.W.; data curation, L.J.; writing—original draft preparation, S.M. and L.J.; writing—review and editing, J.W. and S.M.; visualization, L.J.; supervision, P.J. and X.S.; project administration, H.L., J.W., X.S. and L.J.; funding acquisition, P.J., J.W., X.S. and L.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Beijing Natural Science Foundation (Grant No. 9234025), the Research Capacity Enhancement Program for Young Teachers of Beijing University of Civil Engineering and Architecture (No. X22006), the R&D Program of Beijing Municipal Education Commission (Grant No. KM202110016013), the Humanity and Social Science Youth Foundation of the Ministry of Education of China (Grant No. 19YJC630148), the National Natural Science Foundation of China (Grant No. 52172301), the Key Program of Beijing Social Science Foundation (Grant No. 21GLA010) and the Major Program of the National Social Science Foundation of China (Grant No. 21ZDA029) and the BUCEA Post Graduate Innovation Project (Grant No. PG2023050).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data in this paper are not publicly available for the time due to the relevant policy regulations in China. If you would like to access the data source, please contact the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. *Global Status Report on Road Safety 2020*; World Health Organization: Geneva, Switzerland, 2020.
2. Abrari Vajari, M.; Aghabayk, K.; Sadeghian, M.; Shiwakoti, N. A multinomial logit model of motorcycle crash severity at Australian intersections. *J. Saf. Res.* **2020**, *73*, 17–24. [CrossRef] [PubMed]
3. Jiao, P.; Li, R.; Wang, J.; Ge, H.; Chen, Y. Causes Analysis on Severity of Elderly Pedestrian Crashes Considering Latent Classes. *J. Transp. Syst. Eng. Inf. Technol.* **2022**, *5*, 328–336.
4. Yang, Y.; Yuan, Z.; Meng, R. Exploring Traffic Crash Occurrence Mechanism toward Cross-Area Freeways via an Improved Data Mining Approach. *J. Transp. Eng. Part A Syst.* **2022**, *148*, 04022052. [CrossRef]
5. Ahmad, N.; Ahmad, A.; Wali, B.; Saeed, T.U. Exploring factors associated with crash severity on motorways in Pakistan. *Proc. Inst. Civ. Eng. Transp.* **2022**, *175*, 189–198. [CrossRef]
6. Se, C.; Champahom, T.; Jomnonkwao, S.; Karoonsoontawong, A.; Ratanavaraha, V. Temporal stability of factors influencing driver-injury severities in single-vehicle crashes: A correlated random parameters with heterogeneity in means and variances approach. *Anal. Methods Accid. Res.* **2021**, *32*, 100179. [CrossRef]
7. Li, Y.; Zhang, X.; Wang, W.; Ju, X. Factors Affecting Electric Bicycle Rider Injury in Accident Based on Random Forest Model. *J. Transp. Syst. Eng. Inf. Technol.* **2021**, *1*, 196–200.
8. Zhu, X.; Srinivasan, S. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accid. Anal. Prev.* **2011**, *43*, 49–57. [CrossRef]
9. Yang, Y.; Wang, K.; Yuan, Z.; Liu, D. Predicting Freeway Traffic Crash Severity Using XGBoost-Bayesian Network Model with Consideration of Features Interaction. *J. Adv. Transp.* **2022**, *2022*, 4257865. [CrossRef]
10. Adanu, E.K.; Agyemang, W.; Islam, R.; Jones, S. A comprehensive analysis of factors that influence interstate highway crash severity in Alabama. *J. Transp. Saf. Secur.* **2022**, *14*, 1552–1576. [CrossRef]
11. Sun, Z.; Wang, D.; Gu, X.; Xing, Y.; Wang, J.; Lu, H.; Chen, Y. A hybrid clustering and random forest model to analyse vulnerable road user to motor vehicle (VRU-MV) crashes. *Int. J. Inj. Control Saf. Promot.* **2023**, *ahead-of-print*. 1–14. [CrossRef]
12. Kullgren, A.; Stigson, H.; Ydenius, A.; Axelsson, A.; Engström, E.; Rizzi, M. The potential of vehicle and road infrastructure interventions in fatal bicyclist accidents on Swedish roads—What can in-depth studies tell us? *Traffic Inj. Prev.* **2019**, *20*, S7–S12. [CrossRef] [PubMed]
13. Tay, R.; Rifaat, S.M. Factors contributing to the severity of intersection crashes. *J. Adv. Transp.* **2007**, *41*, 245–265. [CrossRef]
14. Jiang, C.; He, J.; Zhu, S.; Zhang, W.; Li, G.; Xu, W. Injury-Based Surrogate Resilience Measure: Assessing the Post-Crash Traffic Resilience of the Urban Roadway Tunnels. *Sustainability* **2023**, *15*, 6615. [CrossRef]
15. Yang, Y.; Tian, N.; Wang, Y.; Yuan, Z. A Parallel FP-Growth Mining Algorithm with Load Balancing Constraints for Traffic Crash Data. *Int. J. Comput. Commun. Control.* **2022**, *17*, 4806. [CrossRef]
16. Zeng, Q.; Wang, X.; Zhang, X.; Wen, H. Seasonal Analysis of Contributing Factors to Freeway Crash Frequency Using a Spatio-temporal lnteraction Model. *China J. Highw. Transp.* **2020**, *33*, 255–263.
17. Wang, Z.; Jiao, P.; Wang, J.; Huang, Q.; Li, R.; Lu, H. The level of delay caused by crashes (LDC) in metropolitan and non-metropolitan areas: A comparative analysis of improved Random Forests and LightGBM. *Int. J. Crashworthiness* **2022**, 1–15. [CrossRef]
18. Ahmed, S.; Hossain, M.A.; Ray, S.K.; Bhuiyan, M.M.I.; Sabuj, S.R. A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and performance. *Transp. Res. Interdiscip. Perspect.* **2023**, *19*, 100814. [CrossRef]
19. Outay, F.; Adnan, M.; Gazder, U.; Baqueri, S.F.A.; Awan, H.H. Random forest models for motorcycle accident prediction using naturalistic driving based big data. *Int. J. Inj. Control Saf. Promot.* **2023**, *30*, 282–293. [CrossRef]
20. Masello, L.; Castignani, G.; Sheehan, B.; Guillen, M.; Murphy, F. Using Contextual Data to Predict Risky Driving Events: A Novel Methodology from Explainable Artificial Intelligence. *Accid. Anal. Prev.* **2023**, *184*, 106997. [CrossRef]
21. Wen, X.; Xie, Y.; Wu, L.; Jiang, L. Quantifying and Comparing the Effects of Key Risk Factors on Various Types of Roadway Segment Crashes with Lightgbm and Shap. *Accid. Anal. Prev.* **2021**, *159*, 106261. [CrossRef]
22. Wang, Z.; Jiao, P.; Wang, J.; Luo, W.; Lu, H. Contributing factors on the level of delay caused by crashes: A hybrid method of latent class analysis and XGBoost based SHAP algorithm. *J. Transp. Saf. Secur.* **2023**, 1–33. [CrossRef]
23. Jiang, L.; Xie, Y.; Wen, X.; Ren, T. Modeling highly imbalanced crash severity data by ensemble methods and global sensitivity analysis. *J. Transp. Saf. Secur.* **2022**, *14*, 562–584. [CrossRef]
24. Samerei, S.A.; Aghabayk, K.; Mohammadi, A.; Shiwakoti, N. Data mining approach to model bus crash severity in Australia. *J. Saf. Res.* **2021**, *76*, 73–82. [CrossRef]
25. Kong, X.; Das, S.; Tracy Zhou, H.; Zhang, Y. Patterns of near-crash events in a naturalistic driving dataset: Applying rules mining. *Accid. Anal. Prev.* **2021**, *161*, 106346. [CrossRef]
26. Xu, C.; Bao, J.; Wang, C.; Liu, P. Association rule analysis of factors contributing to extraordinarily severe traffic crashes in China. *J. Saf. Res.* **2018**, *67*, 65–75. [CrossRef] [PubMed]
27. Cao, Z.F. *Research on Random Forest Algorithm Optimization*; Capital University of Economics and Business: Beijing, China, 2014.
28. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
29. Cabrera-Arnau, C.; Prieto Curiel, R.; Bishop, S.R. Uncovering the behaviour of road accidents in urban areas. *R. Soc. Open Sci.* **2020**, *7*, 191739. [CrossRef] [PubMed]

30. Goswamy, A.; Abdel-Aty, M.; Islam, Z. Factors affecting injury severity at pedestrian crossing locations with Rectangular RAPID Flashing Beacons (RRFB) using XGBoost and random parameters discrete outcome models. *Accid. Anal. Prev.* **2023**, *181*, 106937. [CrossRef]
31. Yu, S.; Jia, Y.; Sun, D. Identifying Factors that Influence the Patterns of Road Crashes Using Association Rules: A case Study from Wisconsin, United States. *Sustainability* **2019**, *11*, 1925. [CrossRef]
32. Jin, X. *Influencing Factors Modeling and Analysis of Extraordinarily Severe Traffic Crashes*; Beijing Jiaotong University: Beijing, China, 2021.
33. Louis, A.M.; Erick, G.; Eric, D. Crash risk, crash exposure, and the built environment: A conceptual review. *Accid. Anal. Prev.* **2019**, *134*, 105244.