Analysis of Factors Causing Urban Road Traffic Accidents Based on XGboost Algorithm

Jiaqi Yang

Civil Aviation University of China, Air Traffic Management College, Tianjing, China

1320137085@qq.com

Abstract. To further strengthen road traffic safety management and improve the accuracy of early warning systems for road traffic safety, a model for factors causing urban road traffic accidents based on the XGBoost algorithm is proposed. Firstly, SMOTE is used to process the unbalanced data, including supplementing the missing values, deleting the duplicate values, and visualizing the data. Then the prediction model is built by the XGBoost algorithm. By comparing and analyzing the results with the LR model, linear SVM model, DT model and Lightgbm model, the average accuracy rate of the XGBoost model reaches 0.979. Based on the XGBoost algorithm, the analysis of the factors causing urban road traffic accidents has better prediction performance, which can provide a reliable reference for preventing traffic accidents.

Keywords: Traffic Safety; XGBoost; Urban Roads; SMOTE; Machine Learning.

1. Introduction

With the continuous development of urbanization in China, the number of vehicles in China is increasing and the urban road traffic accident has always been a major challenge faced by urban road traffic managers and traffic participants. According to data from the National Bureau of Statistics in 2019 [1], there were more than 240,000 traffic accidents. From 2001 to 2006, more than 20,000 traffic accidents occurred on urban roads in China every year on average. In 2006, the death rate of traffic accidents in China reached 2.5%, while that of developed countries such as Europe and the United States in the same year was below 0.5%. Hence, it is of great significance to conduct research on urban road traffic accidents in China.

In recent years, scholars have analyzed urban road traffic accident datasets based on data statistics or machine learning algorithms to discuss factors that cause traffic accidents. Olutayo [2] used decision trees and neural network algorithms to predict the historical traffic accident set, which concluded that the main influencing factors of traffic accidents were vehicle tire blowout, vehicle out of control and speeding. Zheng et al. [3] converted the single feature relationship of traffic accident data into gray-level images including combined relationships, and proposed a traffic accident prediction model based on convolutional neural networks. Zhang et al. [4] established a traffic accident model based on long-term and short-term memory networks, predicted traffic safety level indicators, and captured the temporal dependence in the data through the LSTM model. Yan et al. [5] optimized the prediction method of the LSTM network model. Guo et al. [6] trained and applied the model empirically based on the traffic accident prediction of the ConvLSTM network model.

However, the current analysis of traffic accident factors is relatively simple. From a certain aspect, the consideration of influencing factors in many studies is not comprehensive enough, and the accuracy of model prediction needs improvement. In addition, the input and prediction in the modeling process need to be further selected. On this basis, this paper considers the influencing factors of traffic accidents from four aspects, including vehicles, drivers, roads and environments. The random forest model is used to select important influencing factors, based on which the XGBoost algorithm is to predict the incidence of traffic accidents.

2. Description of Road Traffic Data

2.1 Data Sources

This paper selects the urban road traffic accident data in the UK in the past decade for research. Based on the principle of complete and accurate data records, Python is used to detect and delete missing data in the dataset, which finally gets 22896 traffic accident data. The research is conducted based on traffic accident data including four influencing factors: vehicle, driver, road and environment.

2.2 Classification of Factors Causing Urban Traffic Accidents

Different countries have various classification standards for road traffic accidents. In China, it is divided into four types: minor accidents with 1-2 minor injuries, general accidents with 1-2 serious injuries, major accidents with 1-2 deaths, and serious accidents with more than 3 deaths. According to the analysis of traffic accident datasets in this paper, there are many minor accidents and general accidents, while only a few records of major accidents and serious accidents exist. Thus, minor accidents and general accidents are merged into general accidents, while major accidents and serious accidents are merged into serious accidents.

3. Data Processing

3.1 Preprocessing

Dummifying variables: Dummy variables are called indicator variables, usually with a value of 0 or 1 to reflect the different attributes of a certain variable. For an independent variable with n classification attributes, it is necessary to select a classification as a reference, so n-1 dummy variables are generated. In this paper, dummifying variable is used to transform values not related to each other in a sense into dummy variables for quantification. For example, when Legacy-collision-severity is converted into a dummy variable, its three cases are Legacy-collision-severity-1, Legacy-collision-severity-2, and Legacy-collision-severity-3. In their corresponding data, 0 means that such a situation does not occur, and 1 means the opposite.

Filling missing values: In the dataset, variables without missing values are complete variables, while variables with missing values are incomplete variables. There are two methods to deal with missing values: deleting and filling. This paper adopts deleting. Typically, the missing value is replaced with a null value, and then the row with the null value is deleted.

Data imbalance processing: Aiming at the data label imbalance, oversampling and undersampling techniques are proposed to generate a balanced dataset to train the prediction model. The synthetic minority oversampling technique (SMOTE) is superior to random oversampling and oversampling based on previous studies. The key of SMOTE is to increase the number of minority class samples in the dataset by synthesizing new minority class samples, so as to balance the dataset. First, the distance between each adjacent sample in the dataset is calculated. Then for each minority sample, one of the adjacent samples is randomly selected, and a new sample is generated by using the formula. Finally, the generated samples are added to the original dataset to increase the number of samples and balance the dataset.

3.2 Fundamental Analysis of Data

This paper uses data imbalance processing, duplicate value deletion, dummifying variables and missing value filling to discretize the variables. Taking Pedestrian-crossing-human-control as an example, dummifying data are divided into five rows, each row corresponding to the situation it represents, with its discrete results shown in Table 1. None-within-50-metres has the most frequency, which indicates that it is most likely to cause traffic accidents in this case and can be analyzed as a vital factor in the following analysis.

Table 1 Description of Variables

Variable	Type	Assignment	Remarks (Information Represented by Figures)
number_of_vehicles	Integer	(Unit in Vehicles) 1: 24.11%, 2: 70.27%, others: 5.62%	None
number_of_casualties	Integer	(Unit in Pieces) 1: 98.38%, 2: 1.61%, 3: 0.01%	None
speed_limit	Float	Minimum 20, maximum 70	None
junction_detail	Integer	All high precision	None
junction_control	Float	1.0: 311, 2.0: 5519, 3.0: 355, 4.0: 21402, 9.0: 1037	None
legacy_collusion_ severity_1(2)(3)	Integer	1: 249, 2: 5878, 3: 6077	1: fatal, 2: serial, 3: slight
pedestrian_crossing_ human_control_ -1(0)(1)(2)(9)	Integer	1: 154, 2: 28470	-1: Data missing or out of range 0: None within 50 metres 1: Control by crossing patrol 2: Control by other authorized person 9: unknown
pedestrian_crossing_ physical_facilities_ -1(0)(1)(4)(5)(7)(8)(9)	Integer	1: 82, 2: 28542	-1: Data missing or out of range 0: No physical crossing facilities within 50 metres 1: Zebra 4: Pelican puffing toucan or similar non-junction pedestrian light crossing 5: Pedestrian phase at traffic signal junction 7: Footbridge or subway 8: Central refuge 9: unknown
light_conditions_ -1(1)(4)(5)(6)(7)	Integer	0: 7410, 1: 21214	-1: Data missing or out of range 1: Daylight 4: Darkness-light lit 5: Darkness-light unit 6: Darkness-no lighting 7: Darkness-lighting unknown
weather_conditions _ -1(1)(2)(3)(4)(5)(9)	Integer	0: 28482, 1: 142	-1: Data missing or out of range 1: Fine no high winds 2: Raining no high winds 3: Snowing no high winds 4: Fine + high winds 5: Raining + high winds 9: unknown
road_surface_conditions_-1(1)(2)(3)(4)(5)(9)	Integer	0: 28482, 1: 142	-1: Data missing or out of range 1: Dry 2: Wet or Damp 3: Snow 4: Frost or ice 5: Flood over 3cm deep 9: unknown

4. Establishment of Prediction Model

4.1 XGBoost Model Theory

Extreme gradient boosting (XGBoost) is developed on the traditional GBDT model. In the XGBoost algorithm, not only CART decision tree can be used, but also a linear foundation model can be supported. In the design of the loss function, a regular term is added to the algorithm to prevent the model from overfitting and control the complexity of the model. In addition, to reduce the calculation of the model and prevent overfitting, XGBoost uses a random forest to sample the fields.

The prediction model of traffic accident severity based on the XGBoost algorithm has the following prediction results:

$$y_j^{(m)} = \sum_m^M f_m\,(x_j) \tag{2}$$

$f_{(m)}$ means the m-th decision tree. x_j means the eigenvector of the j-th sample. $f_{(m)(x_j)}$ means the prediction score of the m-th decision tree on the j-th sample, that is, the leaf weight. $y_j^{(m)}$ means the sum of the leaf weights of m decision trees, that is, the predicted results of XGBoost.

The loss function of the XGBoost algorithm consists of the actual loss value and the regular term. The specific expression of the loss function is as follows:

$$obj = \sum_{j=1}^N l(y_j, \hat{y}_j) + \sum_{m=1}^M \Omega(f_m) \tag{3}$$

j is the sample index. N is the total number of samples. y_j represents true values. $\hat{y}_j$ represents predicted values. $\sum_{j=1}^N l(y_j, \hat{y}_j)$ represents the loss value used to measure the difference between y_j and $\hat{y}_j$; $\sum_{m=1}^M \Omega(f_m)$ represents the regular term, that is, the sum of all decision tree complexity to reduce overfitting.

$\hat{y}_j^{(k)} = \Sigma_m^k f_m(x_j) = \hat{y}_j^{(k-1)} + f_k(x_j)$, thus the loss function is:

$$\widehat{Obj}(k) = \sum_{j=1}^N l(y_j^k, \hat{y}_j^{(k-1)} + f_k(x_j)) + \sum_{m=1}^{k-1} \Omega(f_m + \Omega(f_k)) \tag{4}$$

The second-order Taylor is used to obtain Obj(k):

$$(Obj)\hat{}(k) = \sum_{j=1}^N [l(y_j^k, \hat{y}_j^{(k-1)}) + f_k(x_j)p_j + \tfrac{1}{2}(f_k(x_j))^2 q_j] + \sum_{m=1}^{k-1} \Omega(f_m) + \Omega(f_k) \tag{5}$$

At the same time, there is definition:

$$p_j = \partial_{\hat{y}_j^{(k-1)}} l(y_j^k, \hat{y}_j^{(k-1)}) \tag{6}$$

$$q_j = \partial^2_{\hat{y}_j^{(k-1)}} l(y_j^k, \hat{y}_j^{(k-1)}) \tag{7}$$

p_j and q_j refer to the first-order and second-order derivatives of the loss value l$(y_j^k, \hat{y}_j^{(k-1)}$ to $\hat{y}_j^{(k-1)}$ respectively.

After removing the constant in Equation (5), the loss function is:

$$\widehat{Obj}(k) = \sum_{j=1}^N [f_k(x_j)p_j + \tfrac{1}{2}(f_k(x_j))^2 q_j] + \Omega(f_k) \tag{8}$$

Meanwhile, the regular term can be expressed as:

$$\Omega(f_k) = \gamma K + \tfrac{1}{2}\lambda ||\omega||^2 = \gamma K + \tfrac{1}{2}\lambda \sum_{h=1}^K \omega_h^2 \tag{9}$$

λ and γ represent the hyperparameters of the model. K represents the number of leaf nodes of the tree f; γK represents the structure of the control number. $\frac{1}{2}\lambda||\omega||^2$ represents the regular term used to control the complexity of the

model. ω_h represents the sample weight of the current corresponding leaf node.

Then the final objective function is:

$$\text{Obj(k)} = -\tfrac{1}{2}\sum_{h=1}^K \frac{P_h^2}{Q_h + \lambda} + \lambda K \tag{10}$$

At the same time, there is definition:

$$p_h = \sum_{j=l_k} p_j \quad (11)$$

$$Q_h = \sum_{j=l_k} q_j \quad (12)$$

Equation (10) is a new objective function for model optimization, relying on p_h and Q_h. Because p_h and Q_h is determined by the loss function and the predicted result $\hat{y}_j^{(k-1)}$ of the tree under the structure, and K is determined by thetree structure, minimization of the objective function Obj to solve the optimal

tree structure.

4.2 Model Evaluation Indicators

To evaluate the model's predictive performance, the model evaluation indicators include accuracy, precision, recall, etc., with their specific meanings and evaluation criteria shown in Table 2. The accuracy of the XGBoost model is compared with the other 4 popular machine learning models. To ensure fairness, the training and testing of 5 machine learning models are based on the same dataset, with all data selected for model training. Then, the data are used to test the model accuracy. For the training data, SMOTE data balance technology is used to generate positive and negative samples with balanced proportions. The balanced distribution of samples can avoid uneven distribution of samples for the model training, with default parameters used in the data balancing process. The distribution characteristics of the samples before and after the data balancing process are shown in Figure 2. In this study, it is hoped that the model has a higher accuracy rate to analyze all the causes of traffic accidents as much as possible to reduce losses, so the accuracy is used as an evaluation index to compare the accuracy of five machine learning models as seen in Figure 1.

Table 2 Evaluation Index and Their Meanings

Evaluation Index	Index Meaning	Evaluation Basis	Evaluation Criteria
Accuracy	Ratio of samples with correct predictions to total samples	TP+FN/TP+TN+FP +FN	The higher, the better
Precision	Ratio of true positive samples predicted to be positive	TP/TP+FP	The higher, the better
Recall	Ratio of true positive cases predicted to be positive	TP/TP+FN	The higher, the better
False Alarm Rate	Ratio of true countersamples predicted to be positive	FP/TN+FP	The lower, the better
F1-Socre	Harmonic average of precision and recall	2*Precision*Recall / (Precision+Recall)	The closer to 1, the better

Note: TP means that the prediction to be positive is correct. FP means that the prediction to be positive is wrong. TN indicates that prediction to be negative is correct. FN indicates that the prediction to be negative is wrong.

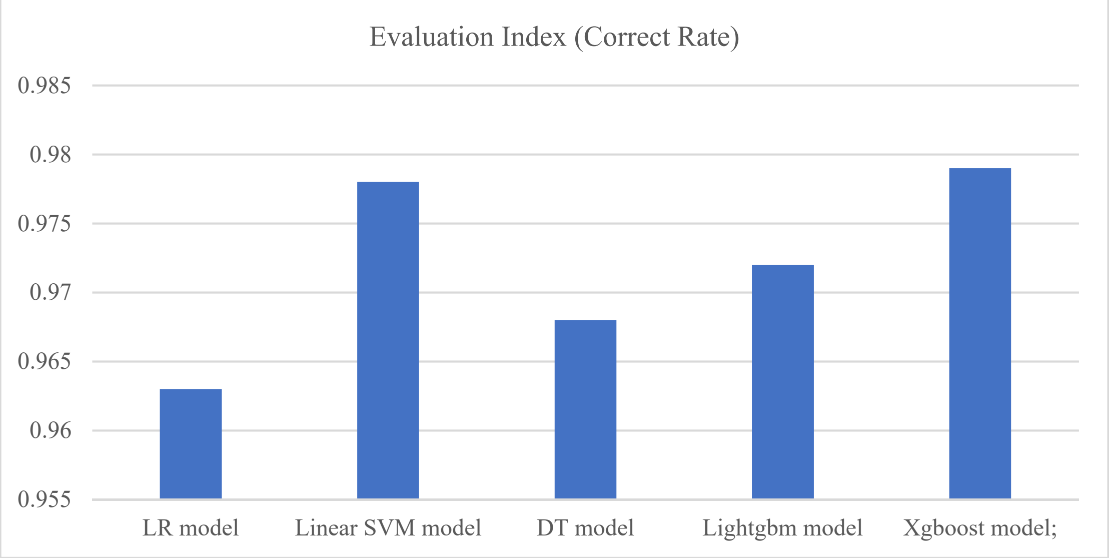

Figure 1 Model Evaluation Index.

Among the five machine learning models, the XGBoost model has the highest correct rate, so its comprehensive analysis performance is the best.

4.3 XGboost Model Building

In the XGBoost model, several parameters are selected to maximize the model's

analytical performance. Parameter optimization can prevent model overfitting and overcomplexity. In this paper, the grid search is used to optimize the XGBoost model parameters with the results shown in Table 3. Default values are used for other parameters not mentioned.

Table 3 Parameter Optimization Results of XGBoost Model

Parameter	Explanation	Optimization Result
n_estimators	Number of weak learners	2500
max_depth	Maximum depth of number	10
eta/leanring_rate	Shrink step used during update	0.05
min_child_weight	Minimum leaf node sample weight	1
gamma/min_split_loss	Node splitting threshold	0
subsample	Ratio of randomly sampled per tree	1
colsample_bytree	Ratio of columns randomly sampled per tree	1
reg_lambda/lambda	Weight coefficient of L_2 regularization penalty term	1
reg_alpha/alpha	Weight coefficient of L_1 regularization penalty term	0

In addition, the stable low false alarm rate of the XGBoost model has been tested and verified in other studies [7].

5. Analysis of Experimental Results

5.1 Accuracy Analysis

In Figure 2, the XGBoost model has the best classification performance among the five machine learning models. After optimizing model parameters, the accuracy is 98%, the precision is 97%, the recall is 98%, and the F1-Score is 97%. The model has good predictive performance with the accuracy of XGBoost reaching the highest 0.979, so it is the most vital core in the factor analysis of urban road traffic accidents.

5.2 Assessment of Influencing Factors

The F-score model interpreter is adapted with the XGBoost classification model, which is then sorted and visualized according to the F-score average absolute value of the accident characteristic variables in the dataset. Thus, the F-score average influence ranking chart that reflects the contribution of each accident characteristic variable to the analysis of the accident cause can be obtained as shown in Figure 2.

According to Figure 2, characteristic variables such as junction detail, speed limit, and number of vehicles are vital factors affecting urban road traffic accidents. The chart of the F-score effect in Figure 5 qualitatively describes the overall relationship between the characteristic variables of urban road traffic accident impact. First of all, the F-score value of junction detail is the highest. Hence, when a traffic accident occurs on an urban road, the safety function of the vehicle is the most important, followed by the speed limit and the number of cars on the road.

Based on the F-score values of the lighting state in Figure 2 and their comparison, traffic accidents are more likely to occur at night than during the day, and traffic accidents are more likely to occur without lighting at night than with lighting at night. When exploring the causes of urban road traffic accidents, the lighting state is one of the crucial factors. When improving the state of road lighting, for example, the lighting facilities should be overhauled on road sections prone to traffic accidents, so as to ensure that the lighting is in good condition at night.

Based on the F-score values of the weather state in Figure 2 and their comparisons, snow is more likely to cause traffic accidents than rain and clear sky. In snowy weather, the road surface is prone to snow accumulation, causing the road to be slippery, and vehicles cannot maintain smooth driving, which leads to traffic accidents. In rainy weather, the driver should turn on the wiper, always keep a good line of sight, and reduce the driving speed to undermine traffic risk. When exploring the causes of urban road traffic accidents, weather conditions are also one of the crucial factors. In case of snowy weather, timely notification should be made through the Internet and radio, and the snow on the road surface should be cleared in time.

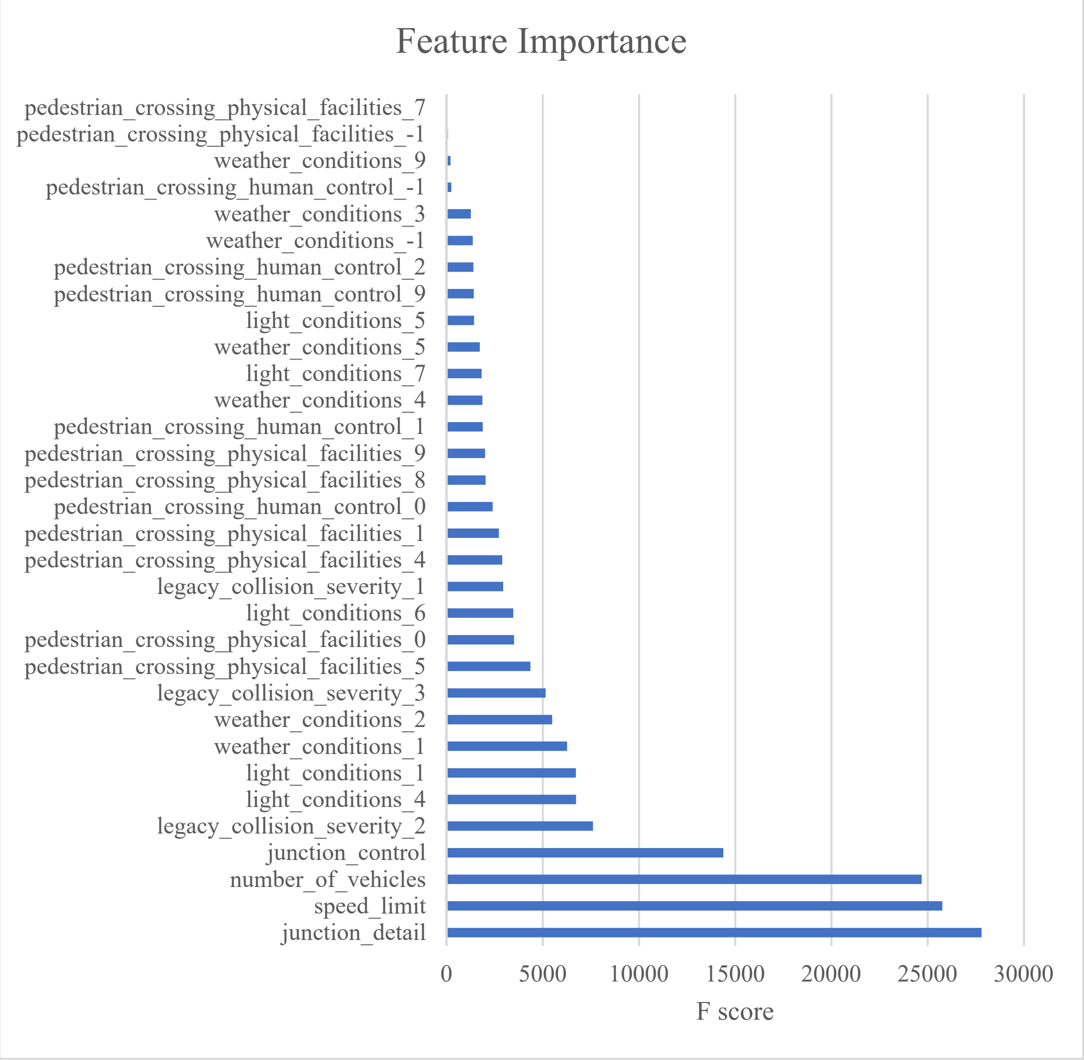

Figure 2 Feature Importance

6. Conclusion

Based on the dataset of a foreign city road traffic accidents, this paper extracts 41 related influencing factors of traffic accidents, establishes the influencing factor model of urban road traffic accident with the XGBoost method, and analyzes the nonlinear relationship between urban road traffic accidents and each influencing factor with SHAP interpretability method. Through the analysis, it is concluded that driving speed, the number of vehicles, and the safety factor of vehicles are crucial factors affecting traffic accidents. Thus, in future road traffic safety work, we should focus on the above factors to prevent traffic safety accidents and further reduce the probability of traffic accidents.

In this paper, the XGBoost single algorithm is used for predictive analysis, which can be combined with other single models to improve the prediction accuracy.

References

[1] Statistical Communiqué of the People's Republic of China on the 2019 National Economic and Social Development.

[2] Olutayo, V. A. & Eludire, A. A. (2014). Traffic accident analysis using decision trees and neural networks. International Journal of Information Technology & Computer Science, 6(2): 207-237.

[3] Zheng, M., Li, T., Zhu, R. et al. (2019). Traffic accident’s severity prediction: A deep-learning approach-based CNN network. IEEE Access, 7: 39897-39910.

[4] Zhang, Z. H., Yang, W. Z., Zhong, T. T. et al. (2019). Traffic accident prediction based on LSTM neural network model. Computer Engineering and Application, 55(14): 249-253.

[5] Yan, Z., Yu, C. C., Han, L. et al. (2019). Short-term traffic flow forecasting methods based on CNN+LSTM. Computer Engineering and Design, 40(9): 2620-2624.

[6] Guo, X. & Hu, Z. H. (2023). Traffic accident prediction method based on ConLSTM. Journal of Ningbo University of Technology, 35(01): 9-15.

[7] Zhao, W. & Ma, K. (2013). Fire loss prediction based on gray prediction GM (1,1) model. Fire Science and Technology, 32(3): 324-327. DOI:10.3969/j.issn.1009-002 9.2013.03.028.