Journal of Electronic Research and Application, 2024, Volume 8, Issue 1
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Analysis of Traffic Accidents Based on the Integration Model

Yanshun Ma*, Yi Shi, Yihang Song, Chenxiao Wu, Yuanzhi Liu

The School of Civil Engineering and Transportation, Northeast Forestry University, Harbin 150040, Heilongjiang Province, China

Corresponding author: Yanshun Ma, dljtxyyhb@163.com

Abstract: To enhance the safety of road traffic operations, this paper proposed a model based on stacking integrated learning utilizing American road traffic accident statistics. Initially, the process involved data cleaning, transformation, and normalization. Subsequently, various classification models were constructed, including logistic regression, k-nearest neighbors, gradient boosting, decision trees, AdaBoost, and extra trees models. Evaluation metrics such as accuracy, precision, recall, F1 score, and Hamming loss were employed. Upon analysis, the passive-aggressive classifier model exhibited superior comprehensive indices compared to other models. Based on the model's output results, an in-depth examination of the factors influencing traffic accidents was conducted. Additionally, measures and suggestions aimed at reducing the incidence of severe traffic accidents were presented. These findings served as a valuable reference for mitigating the occurrence of traffic accidents.

Keywords: Stacking integrated learning; Data analysis; Traffic safety

Online publication: January 18, 2024

1. Introduction

The elements of road traffic, including people, vehicles, and infrastructure, are experiencing rapid growth, demanding higher standards for road traffic safety. According to the 2022 Statistical Communiqué on National Economic and Social Development by the National Bureau of Statistics, China reported approximately 14,527 deaths due to road traffic accidents in 2022. These accidents have inflicted significant losses to both lives and property, posing serious obstacles to the rapid development of China's transportation industry. Consequently, identifying crucial influencing factors of traffic accidents and predicting their occurrences are of paramount importance and warrant thorough exploration.

In recent years, scholars globally have employed various methods to study influencing factors and predict traffic accidents, yielding notable results. For instance, scholars proposed a logistic regression analysis model, elucidating a significant variable affecting collision severity and analyzing the primary factor contributing to the severity of traffic accidents in Riyadh. Wang *et al.* comprehensively analyzed the influencing factors of urban road traffic accidents from 2014 to 2019, considering human, vehicle,

road, and environmental aspects, and predicted the severity of road traffic accidents using the XGBoost algorithm [1]. Cao *et al.* focused on 6,891 urban road traffic accidents in Heilongjiang Province from 2017 to 2019, introducing the impact of traffic surrounding roads. They employed the Bayesian network to accurately predict severity and influencing factors [2]. Meng *et al.* utilized a logistic regression analysis model to screen factors with significant influence on the consequences of highway traffic accidents, proposing comprehensive measures to mitigate their severity [3]. Chen *et al.* separately applied the C5.0 decision tree algorithm, multivariate logistics regression, and multilayer perceptron (MLP) neural network for prediction. They identified the main influencing factors of traffic accident patterns, providing a basis for decision-making in traffic management departments [4].

2. Data preprocessing

2.1. Selection of characteristic variables

This paper, drawing insights from United States road traffic accident statistics, identifies key variables pertaining to traffic accidents across multiple perspectives. Through a literature research approach, these variables are categorized into three groups: the impact of accidents, the influence of natural factors, and the effects of nearby location and road conditions. Faced with an extensive dataset of quantitative feature variables during the data analysis phase, it becomes imperative to segment this data into non-uniform intervals. Recognizing the limitations of traditional division methods, the study leverages the Pandas and NumPy open-source libraries in Python. The 'pandas.cut( )' method is employed to globally partition the data range, and the 'iloc( )' method visualizes these intervals, generating a set of interval values that constitute the characteristic variable set for the influencing factors of road accidents.

2.2. Data cleansing

To enhance the applicability of the data in model establishment and subsequently improve prediction accuracy, this paper rigorously validates and excludes missing and aberrant data. Additionally, some data undergoes derivative processing. Generally, characteristic variable data can be classified into qualitative and quantitative types. Based on the nature of the data, feature variables related to road accidents are categorized into categorical and numerical variables. To facilitate representation and model application, categorical variables are quantified, and **Table 1** categorizes all feature variables.

Table 1. Summary table of characteristic variable division

Type	Characteristic
Categorical variable	The severity of the accident, weather, day or night, near a speed bump or hump, near the intersection, near the junction, near the location with a no-exit sign, near the railway, near the roundabout, near the station (bus, train, etc.), near stop signs, near a traffic calming device, near traffic signs, whether the position near turning tips, nearby facilities
Numerical variable	Temperature, wind chill, humidity, air pressure, visibility, wind speed, and precipitation

2.3. Data forwarding and normalization

Quantitative data necessitates further processing to eliminate unit influence on model establishment. This paper forwards the quantitative data to standardize dimensions, providing greater convenience in model construction and minimizing the risk of errors. Building on this, this study normalizes the dimensional data, restricting it within the 0–1 range. This ensures more convenient data processing, accelerates program execution, and

enhances the overall efficiency of the entire process.

3. Model building

3.1. Basic principle of a single model

3.1.1. Logistic regression

The logistic regression model serves as a probability model, typically utilizing the likelihood of a variable causing a state as an independent variable that influences factors related to that state. In the context of binary classification logistic regression analysis, where the dependent variable is a binary classification variable (1 indicating occurrence and 0 indicating non-occurrence), if there are N factors associated with it, the probability of the state occurring is calculated as follows:

$$P(Y=1|X)=\frac{1}{1+exp\left[-\left(B_0+\sum_{i=1}^{n}B_ix_i\right)\right]} \quad (1)$$

3.1.2. K-nearest neighbors classifier

The k-nearest neighbors (KNN) classifier relies on data-driven principles. Through statistical screening of the database, it identifies a number of near neighbors with similar features to the current data, conducts statistical analysis on these neighbors, and integrates the results with weights to achieve predictive functions. The KNN method typically begins by structuring the status vector and determining the selection indicator of the nearest neighbor. A distance threshold is set, and Euclidean distance is employed to match the status vector of the current status with the data in the database, thereby obtaining the predictive result of the model. As a classification algorithm, the KNN algorithm boasts advantages such as a clear structure and easy establishment.

3.1.3. Gradient boosting classifier

The gradient boosting classifier is well-suited for regression problems encountered in practical applications. The process of gradient-boosted decision trees (GBDT) is similar to the boosting tree algorithm. However, the distinction lies in the fitting of the loss function. The boosting tree uses the square loss for fitting, while GBDT fits the approximation of the loss in each round by utilizing the negative gradient of the loss function. Subsequently, it fits the regression tree. If the dataset has p data points and q features, and the training dataset is represented as $T=\{(x_1, y_1), (x_2, y_2), \ldots (x_p, y_p)\}$, its loss function can be expressed as $L(y, f(x))$. Initialization involves calculating the negative gradient of the loss function (r_{pi}):

$$\begin{cases} c_{pj}=\sum_{x_i\in R_{pj}} L\left(y_i,\ f_{p-1}(x_i)+c\right) \\ f_p(x)=f_{p-1}(x)+\sum_{j=1}^{J} c_{pj}I\left(x\in R_{pj}\right) \end{cases} \quad (2)$$

Obtaining the regression tree:

$$\widehat{f_{(x)}}=f_P(x)=\sum_{i=1}^{P}\sum_{j=1}^{J} c_{pj}I\left(x\in R_{pj}\right) \quad (3)$$

The gradient boosting tree utilizes the negative gradient of the loss function to fit the approximation of the current round loss, enhancing the model's accuracy and making it more precise than the boosting tree.

3.1.4. Decision tree classifier

The decision tree classifier utilizes extensive datasets to train sample data and articulates classification rules by constructing a tree structure. The decision tree classification model selects the partition attribute by introducing the Gini(D).

$$Gini(D) = 1 - \sum_{k=1}^{K} \left(\frac{|C_k|}{D} \right)^2 \qquad (4)$$

Here, Gini(D) represents the impurity of the data classified according to a specific feature attribute. The smaller the Gini(D), the higher the purity of the data set, indicating less impurity.

3.1.5. AdaBoost classifier

The adaptive boosting (AdaBoost) classifier is an ensemble model that harnesses the idea of the boosting framework. At a deeper level, AdaBoost is an additive model, a method of combining base learners in an additive form. The AdaBoost classification model employs a forward stagewise algorithm and an index loss function algorithm to address the optimization problem associated with the method's addition. The forward stagewise learning algorithm involves continuous iteration through repeated cycles. At each step, it gradually approaches the optimization target formula by considering only one base function, $h_t(x)$, and its coefficient α_t. The index loss function, $E(h(x), y, i) = e^{-yih(xi)}$, contributes to the AdaBoost classification model formula: $H_t(x) = H_{t-1}(x)+\alpha_t h_t(x)$, where h_t represents the t-th base learning device, and α_t is the weight coefficient for each base learning device.

3.1.6. Extra trees classifier

The extra trees (short for extremely randomized trees) classifier is an ensemble learning algorithm that produces classification results by aggregating multiple related outcomes in random forests. Each decision tree is constructed using the original training sample, corresponding to a random sample with each sample having 'k' features. In building each decision tree, the algorithm selects the best features from these characteristic samples and splits the data based on certain indicators. This results in generating random and unrelated decision trees.

3.2. Basic principles of the integration model

The stacking algorithm adopts a layered model integration framework, similar to the bagging algorithm, aiming to enhance accuracy by reducing variance. However, the stacking algorithm achieves improved data fitting by utilizing multiple different models to mitigate variance. Unlike bagging, which trains different data with the same model, stacking employs diverse models to train the same data, resulting in higher accuracy.

The stacking algorithm process begins with the input dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$. This dataset is divided into training and test sets. On the first floor, multiple base learning devices are created using the aforementioned training set. Simultaneously, k-fold cross-validation is performed on the training set, and predictions are made. The features corresponding to each base learning device are collected to form a secondary training set. The test set is then input into the same learning devices to generate the secondary test set. The feature training set is added to the meta-learning device along with the sub-test set. Classification results are predicted, yielding the final prediction model.

The stacking algorithm enhances accuracy by leveraging multiple models for repeated training, resulting in more precise predictions compared to the bagging algorithm. Using the stacking integrated learning method, the seven single models mentioned earlier serve as sub-models, while three linear models function as component models. Three distinct integrated models are derived, and based on the output results, the optimal model is

determined.

4. Model evaluation

4.1. Importance assessment of single model influencing factors

To evaluate the severity recognition rate of traffic accidents using a single model, this article selects various models, including logical regression, KNN classifier, gradient boosting classifier, decision tree classifier, AdaBoost classifier, and extra trees classifier. Five indicators – accuracy, precision, recall, F1-score, and Hamming loss – are employed to determine the optimal solutions among these single models. Higher accuracy rates, precision rates, recall rates, F1-score values, and lower Hamming loss values indicate a better recognition rate for traffic accident severity. To facilitate comparison, line diagrams are drawn for the first four indicators. Each folding line represents a different single model. The horizontal coordinates correspond to the five indicators, while the vertical axis remains within the 0–1 dimensional, as illustrated in **Figure 1**.

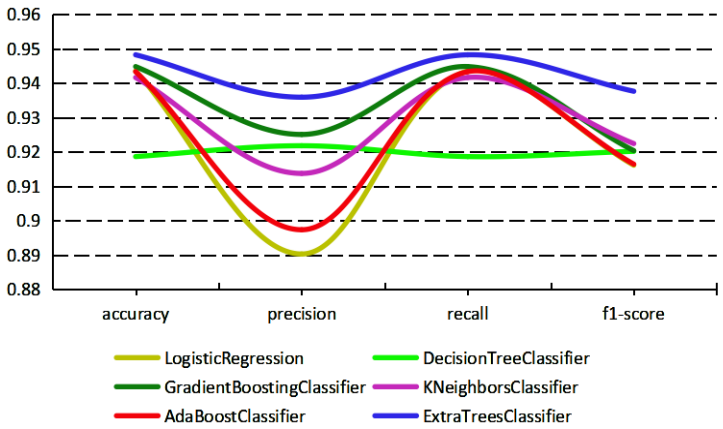

Figure 1. Single model evaluation index line chart

Subsequently, a separate curve chart is drawn for Hamming loss, where the abscissa represents different single models, and the ordinate remains within the 0–1 dimension, as depicted in **Figure 2**.

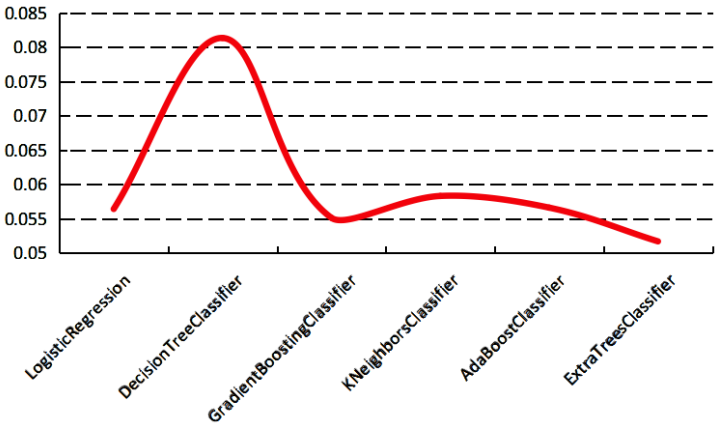

Figure 2. Single model Hamming loss line diagram

Therefore, based on the figures presented, it was concluded that among the single models, the extra trees classifier exhibits the highest recognition rate and the most effective evaluation.

4.2. Assessment of the importance of influencing factors in ensemble models

For evaluating the recognition rate of traffic accident severity using integrated models, this article selects three relatively simple and computationally convenient integrated linear models: passive-aggressive classifier, logistic regression, and ridge classifier. These models serve as the meta-model algorithms in the stacking algorithm, and their process structure is depicted in **Figure 3**.

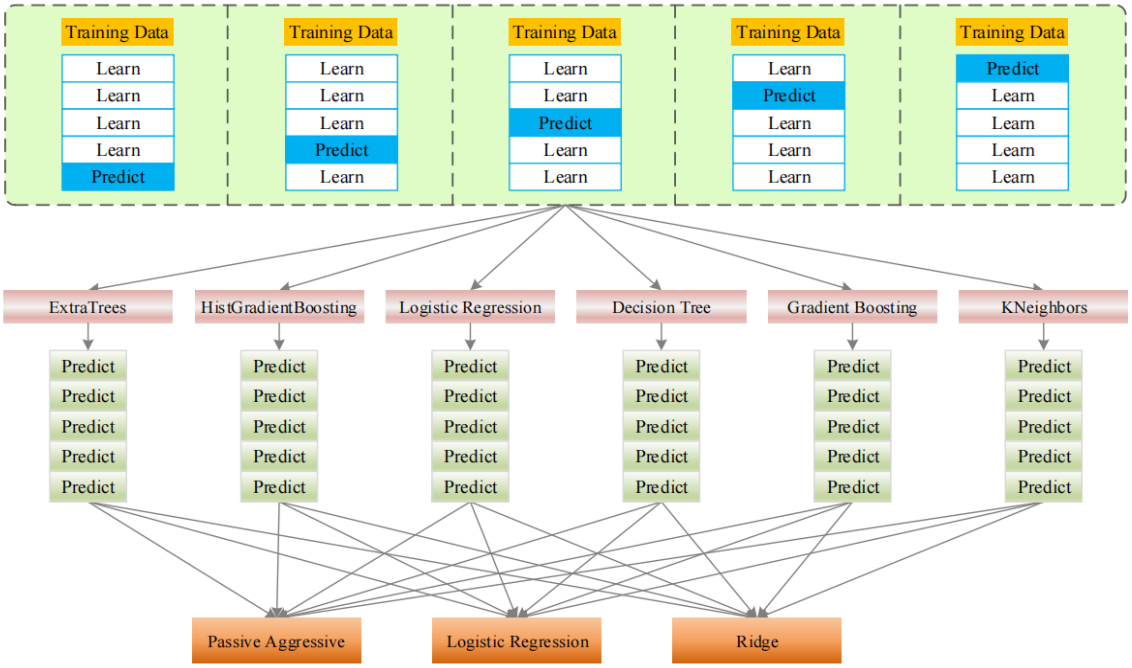

Figure 3. A schematic diagram of the model process structure

Similar to the indicators selected for the impact assessment of a single model, the evaluation of the best scheme among these models is conducted using accuracy, precision, recall, F1 score, and Hamming loss. The Hamming loss is treated separately from other indicators. Using three different lines to represent different integration models, a line chart is drawn for the indicators with values proportional to the recognition rate. The line chart is presented in **Figure 4**.

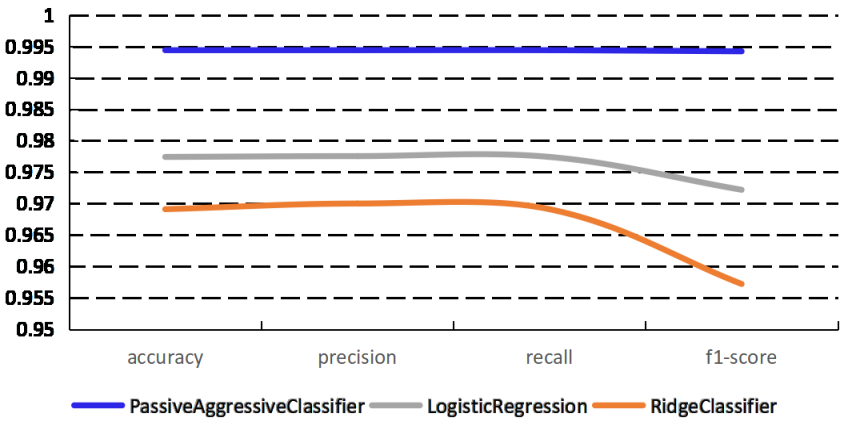

Figure 4. Line diagram of the integrated model evaluation index

Based on **Figure 4**, it is concluded that in the integrated model, the passive-aggressive classifier exhibits the highest recognition rate and the most effective evaluation. Subsequently, the model incorporates the five indicators of weather, time, visibility, nearby convenience facilities, and nearby intersections using permutation feature importance to calculate the importance of the arrangement features that affect traffic accidents for the optimal model.

Table 2. Characteristics and corresponding importance weights

Weight	Feature	Weight	Feature
0.0473 ± 0.0004	Humidity (%)	0.0047 ± 0.0002	Station
0.0453 ± 0.0002	Pressure (in)	0.0045 ± 0.0001	Precipitation (in)
0.0441 ± 0.0002	Temperature (F)	0.0034 ± 0.0001	Stop
0.0429 ± 0.0001	Wind_Chill (F)	0.0012 ± 0.0001	Railway
0.0397 ± 0.0003	Wind_Speed (mph)	0.0008 ± 0.0000	Junction
0.0336 ± 0.0002	Weather_Condition	0.0003 ± 0.0000	No_Exit
0.0225 ± 0.0004	Sunrise_Sunset	0.0003 ± 0.0000	Give_Way
0.0203 ± 0.0003	Traffic_Signal	0.0001 ± 0.0000	Traffic_Calming
0.0144 ± 0.0001	Crossing	0.0000 ± 0.0000	Bump
0.0128 ± 0.0002	Visibility (mi)	0.0000 ± 0.0000	Roundabout

5. Countermeasures analysis

5.1. Existing problems

The road traffic system, comprising people, vehicles, roads, and the environment, forms a complex dynamic system. The causes of traffic accidents primarily encompass subjective factors, such as psychological or physiological aspects of the involved parties, and objective factors, including inherent attributes of the vehicle like weight and configuration, road geometric composition (e.g., number of lanes, driving sight distance), changing areas of road environmental characteristics, and management factors.

Based on the analysis of the significance of arrangement characteristics affected by the above outputs, environmental factors emerge as having a substantial impact on the severity of traffic accidents. These factors include temperature, humidity, weather conditions, pressure, light, and visibility. In recent years, the frequency of traffic accidents caused by meteorological disasters has risen, with adverse weather conditions being a primary factor affecting the safe operation of the highway network. Results indicate that the structural composition, road facilities, and the road's environment significantly influence road traffic safety. Factors such as the presence of intersections, signal lights, lighting settings, and traffic signs near the road all contribute to the occurrence of traffic accidents.

5.2. Countermeasures to promote traffic safety

5.2.1. Environmental aspects

Survey data and statistics reveal that 25% of highway accidents in China are attributed to foggy weather, with nearly one-third of major traffic accidents linked to such conditions. In response, some countries may close highways during heavy fog, and drivers on other highways should promptly turn on their fog lights, maintain a safe following distance, and reduce driving speed. In addition to foggy weather, ice and snow conditions can significantly impact road operations. Drivers should stay informed about road conditions during snowy weather,

avoid forcing passage in heavy snow, and attempt detours if necessary. Transportation departments should swiftly organize resources to clear snow and implement vehicle deceleration facilities to ensure road traffic safety in extreme weather. Precipitation also affects the severity of traffic accidents, so road structures should accommodate maximum drainage capacity to ensure driving safety during rainy and snowy weather.

5.2.2. Road aspects

The characteristics of road sections also influence the severity of traffic accidents. Environmental factors near road sections, such as intersections, stations, and railways, have a notable impact on accident severity. Intersection sections show a higher frequency of serious accidents compared to general sections, and sections near stations and railways also exhibit a more prominent frequency of serious accidents. Installing traffic management facilities and equipment in high-risk road sections is significant for reducing the incidence of serious traffic accidents.

5.2.3. Management aspects

To ensure the safe operation of road traffic, traffic management departments, public security organizations, highway bureaus, and other relevant units employ education, technology, and other means to restrict traffic reasonably, organize and command traffic scientifically, and manage the relationship between people, vehicles, and roads correctly. Improving road traffic management facilities, such as installing traffic signs and lights, aids road users in safely navigating through road sections. With the advancement of science and technology, the utilization of information-based intelligent transportation meets the evolving needs for safe, orderly, smooth, and sustainable transportation.

6. Conclusion

With the acceleration of urbanization, the complexity of traffic composition is on the rise, leading to a significant increase in the probability of traffic accidents compared to previous times. This article employs ensemble learning methods in machine learning to stack multiple models that are relatively effective in predicting traffic accident risk and applies a simple linear method. The result is a fusion model with high-performance prediction for traffic accident risk. It achieves an accuracy of 0.994 for public data, a recall of 0.994, an F1 score of 0.99423, and a Hamming loss of 0.0056. Conducting a permutation feature importance analysis on the model identifies the characteristics that have a significant impact on traffic accident risk. Based on these important features, corresponding suggestions and improvement measures are proposed. The research conclusions can serve as a reference for further studies.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Wang Y, Liu Q, Cheng W, 2022, Prediction of Road Traffic Accident Severity Based on the XGBoost Algorithm. Software Guide, 21(5): 84–88.

[2] Cao Y, Zhang B, Li S, 2022, Analysis of the Factors Influencing the Severity of Urban Road Traffic Accidents in the Ice and Snow Season. Journal of Dalian Jiao Tong University, 43(4): 8–13.

[3] Meng Y, Zhang X, Qing G, et al., 2022, Analysis of Influencing Factors on the Consequences of Highway Traffic Accidents Based on Logistic Regression. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 46(1): 12–16.

[4] Chen L, Li C, Zhan L, et al., 2022, Analysis and Prediction of Urban Road Traffic Accidents. Journal of Chang'an University (Natural Science Edition), 42(4): 98–107.

[5] Zhu H, Li H, Chi Y, et al., 2022, A Method for White Layer Prediction of Hard Car Surface Based on Gradient Lifting Decision Tree. Shanghai Municipality, CN115099266A.

[6] Liu G, Gao Y, 2021, Review of Intrusion Detection Methods Based on Machine Learning. Information and Computer (Theoretical Edition), 33(10): 34–37.

[7] Zhang P, Xu S, 2008, Exploring the Method of Road Traffic Safety Management and Management Planning in Heilongjiang Reclamation Area. Heilongjiang Science and Technology Information, 2008(19): 82.