Research and Analysis of Traffic Accident Severity Prediction Based on Data Augmentation and Feature Interpretation

Yuhan Chen

International Institute of Excellence for Engineering, East China University of Science and Technology, Shanghai, China
22013718@mail.ecust.edu.cn

Abstract. Predicting the severity of traffic accidents is crucial to traffic safety management and accident prevention. However, in the actual data, the number of minor accidents far exceeds that of serious accidents, resulting in deviations in most types of predictions. In order to alleviate the category imbalance in traffic accident data, this paper proposes an analytical framework that combines SMOTE data enhancement and SHAP feature interpretation. Based on a traffic accident data set, this paper compares the performance differences between the four models of logical regression, support vector machine, random forest and XGBoost before and after data enhancement. The results show that SMOTE has effectively improved the model's ability to identify a few classes, and the F1 value of SVM has been increased from 0.31 to 0.39. SHAP analysis shows that after sampling, the importance of characteristics such as "number of casualties", "lighting conditions" and "road conditions" has increased significantly, indicating that the model pays more attention to key risk factors after data balance. This study validates the effectiveness of combining data augmentation and feature interpretation, providing a reference for traffic accident risk assessment.

Keywords: Traffic accident severity prediction, Data augmentation, SMOTE, Interpretability, SHAP.

1 Introduction

Traffic accident prediction, a core component of intelligent traffic safety systems, aims to estimate the likelihood and severity of future accidents. This process plays a vital role in evaluating accident trends and supporting timely preventive decision-making under current road conditions [1].

The data has a problem of extreme imbalance in class distribution, which causes the model to be biased in predicting the majority class during training. For ensemble learning methods (such as random forest, XGBoost, etc.), the adoption of many DTs results in the user losing access to the internal decision logic, which may be the primary

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_74

disadvantage, because the user should at least be aware of the variables that influence forecasts [2].

In recent years, more and more studies have focused on using machine learning methods to predict the severity of traffic accidents, and many of them have paid attention to the impact of imbalanced data class distribution on model performance. Fiorentini and Losa used the Random Undersampled Majority Class (RUMC) method to balance the minority class and improve the minority class prediction ability [3]. Li et al. used Wasserstein Generative Adversarial Network with Gradient Penalty (WGAN-GP) to generate minority class samples and combined it with the SHAP method to improve the interpretability of the model [4]. Aziz et al. proposed a prediction framework of Dynamic Ensemble Selection (DES-MI) and Ensemble Imbalanced Learning (EIL) for multi-class imbalance problems, which significantly improved the classification performance of multi-class severity [5]. At the same time, Aboulola combined transfer learning models with SHAP technology to improve prediction accuracy and interpretability from the perspective of deep learning [6]. These studies show that in the field of traffic accident severity prediction, balancing sample distribution and enhancing model interpretability are the current research focus. In addition, some studies have improved prediction performance through model optimization. For example, Yan and Shen used Bayesian optimization to adjust the parameters of random forest (BO-RF), which improved prediction accuracy and enhanced model interpretability [7].

This paper proposes a comprehensive research framework combining SMOTE data augmentation and SHAP feature interpretation, aiming to improve the model's predictive performance on minority classes and reveal the model's decision-making mechanism. Subsequent chapters will describe the dataset features, preprocessing methods, and model design, present experimental results and feature importance analysis, summarize research conclusions, and propose future research directions.

2 Method

2.1 Dataset description

The traffic accident dataset used in the study comes from the Kaggle RTA Dataset, which contains 32 feature variables and 12,316 accident instances. The data covers multiple dimensions of factors, including weather, lighting conditions, road type, driver and vehicle information. The target variable is the severity of the accident (0: minor, 1: moderate, 2: severe).

2.2 Data preprocessing

Missing values were filled using the mode, and weakly correlated or highly missing features such as time and gender were removed. Label encoding was used to transform the accident severity features into numerical features, and the SMOTE method was applied to create new minority class data samples by analyzing the distribution of nearby samples and generating synthetic instances positioned within the feature space between them, thereby improving the classifier's generalizability by increasing the

number of minority class samples in an unbalanced dataset [8]. Finally, Min-Max normalization was performed before training each model.

2.3 Model structure

Four classification models were selected: Logistic Regression, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Random Forest. Logistic Regression is a probabilistic model based on linear discrimination, which achieves binary classification by weighted summation of input features and mapping using the sigmoid function. Support Vector Machine constructs a maximum margin hyperplane to distinguish different classes, exhibiting good generalization performance, especially suitable for high-dimensional feature spaces. Random forests use the integrated results of multiple decision trees for prediction, which effectively reduces the risk of overfitting and improves the stability of the model. Based on the gradient enhancement framework, XGBoost improves the overall performance by iteratively optimizing the residuals of weak classifiers, and excels in handling complex nonlinear relationships. These four models have their own advantages in interpretability and predictability, which provides a multi-faceted basis for the comparison of subsequent experiments.

Data were split into training and test sets in a 7:3 ratio with a fixed random seed to ensure experiment consistency. Performance was evaluated using metrics such as F1-score, recall, and precision.

Precision reflects how many of the samples labeled positive by the model are truly positive, measuring the reliability of the prediction. Recall represents how well the model detects real positive samples, reflecting its effectiveness in identifying minority classes. The F1-score, calculated as the harmonic mean of precision and recall, comprehensively reflects the model's balance between accuracy and coverage. Since this paper focuses on class imbalance, the F1-score provides a more complete view of performance variation before and after data augmentation, and therefore serves as the primary reference metric.

3 Analysis of experimental results

3.1 Model training and evaluation

Table 1. Performance evaluation of each model before and after SMOTE.

Model	Dataset	Accuracy	F1-macro	F1-weighted
LogisticRegression	Before SMOTE	0.8365	0.3037	0.7621
RandomForest	Before SMOTE	0.8395	0.3260	0.7730
XGBoost	Before SMOTE	0.8382	0.3500	0.7827
SVM	Before SMOTE	0.8376	0.3086	0.7647
LogisticRegression	After SMOTE	0.4146	0.2891	0.5086
RandomForest	After SMOTE	0.8357	0.3358	0.7758
XGBoost	After SMOTE	0.8338	0.3436	0.7782

SVM	After SMOTE	0.6539	0.3929	0.6915

Because the amount of data is small, the experiment was conducted in a CPU environment.

It can be seen that SMOTE has different effects on different models. As shown in Table 1, SVM showed the most significant improvement in F1 score, increasing from 0.31 to 0.39. Random Forest remained relatively stable with a slight increase, as it is less affected by outliers and its predictive ability outperforms other algorithms even in cases of class imbalance and very small sample sizes. XGBoost maintained high stability, with a slight decrease in F1 score. Logistic Regression showed a significant decrease in performance, indicating that SMOTE performs better for nonlinear models.

3.2 Feature contribution analysis

This paper selects two models that improved performance after SMOTE—random forest and SVM—to analyze SHAP values, and uses the Spearman correlation coefficient (ρ) to measure the consistency of feature ranking before and after data augmentation.

SHAP calculates feature importance using ideas from game theory [9]. All of the data from the first interaction are used to train a classification model, which then calculates SHAP values for each feature and ranks them to determine which features are most important for modeling the target problem [9].

It is commonly stated that Spearman's correlation coefficient, which is rank-based, quantifies the degree and direction of a monotonic relationship between X and Y [10]. It is frequently used as a measure of association between two metrics because of these benefits [10].

Figure 1 is a scatter plot of the changes in feature importance before and after data augmentation in random forest. The closer the point is to the diagonal, the smaller the change. The corresponding Spearman correlation coefficient $\rho=0.98$ indicates that the overall feature ranking is relatively stable. As shown in Figure 2, the importance of features such as "number of casualties", "lighting conditions" and "road surface conditions" increased significantly after SMOTE processing, indicating that the model pays more attention to the core factors that have a greater impact on the risk of traffic accidents after data balancing.

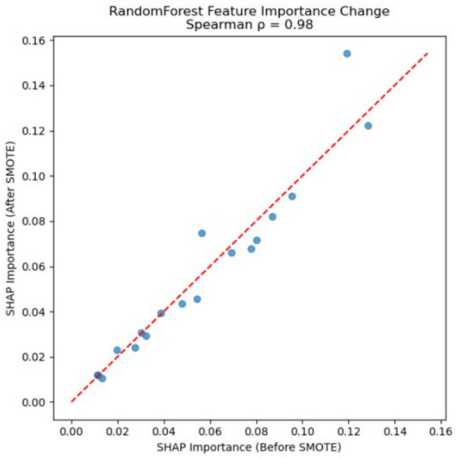

Fig. 1. Scatter plot of feature importance changes before and after SMOTE in random forest.

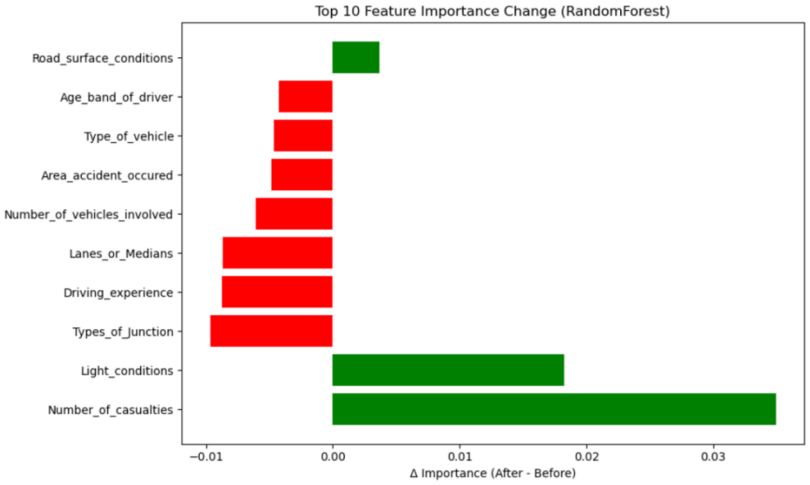

Fig. 2. Changes in feature importance before and after SMOTE in random forest.

Figure 3 shows a scatter plot of feature importance before and after data augmentation in SVM. The Spearman correlation coefficient $\rho = 0.78$ indicates that the sequence of features has changed significantly. Figure 4 further shows the changes in various characteristics, among which the importance of "number of casualties", "type of vehicle", "pedestrian movement" and "lighting conditions" has been significantly enhanced. This shows that SVM is more sensitive to data enhancement, and the weight of different characteristics has also been significantly adjusted.

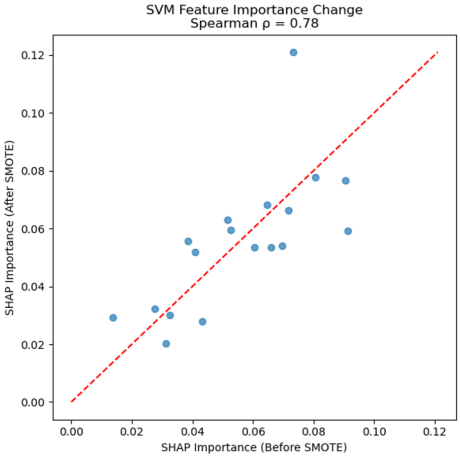

Fig. 3. Scatter plot of feature importance changes before and after SMOTE in SVM.

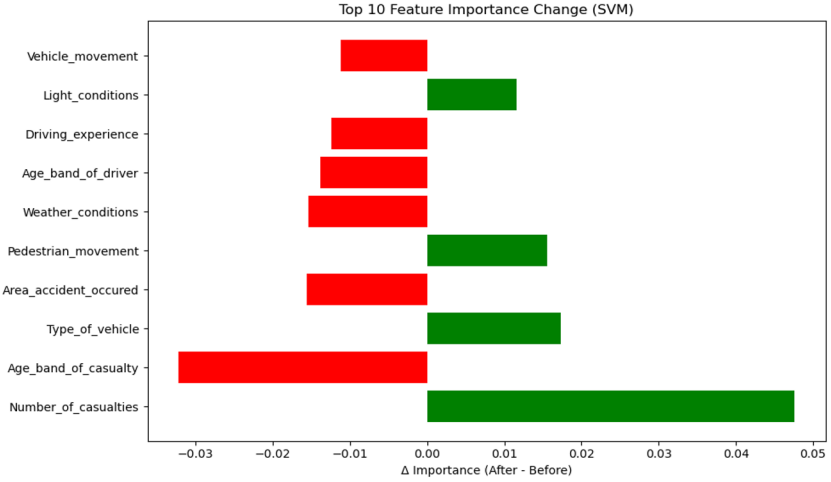

Fig. 4. Changes in feature importance before and after SMOTE in SVM.

Comparing the two models, it can be seen that the response of random forests to data enhancement is more stable, while the changes in support vector machines (SVM) are more significant, indicating that different models have different degrees of dependence on features in the face of category imbalance. In general, after SMOTE processing, the importance of core influencing factors has been enhanced, which verifies the positive role of data enhancement on model feature learning and provides a more reliable interpretation basis for traffic accident risk analysis.

4 Conclusion

The proposed "data enhancement-feature interpretation" framework performs well in predicting the severity of traffic accidents. Experimental results show that SMOTE effectively alleviates the problem of category imbalance and significantly improves the ability of some models to identify minority classes, among which the F1 score of SVM increased from 0.31 to 0.39. Combined with SHAP analysis, it is found that after data enhancement, the model's attention to key features such as "number of casualties", "lighting conditions" and "road conditions" has been significantly increased, which verifies the positive role of data balance in model feature learning. In a word, the combination of SMOTE and SHAP not only improves the prediction performance, but also enhances the interpretability of the model.

Future research can further explore the comparative effects of various oversampling and undersampling algorithms, and combine more advanced interpretable methods to improve the transparency of the model. At the same time, we can consider introducing space-time characteristics or external environmental variables to build a more universal prediction model and provide stronger technical support for traffic safety management and risk early warning.

References

1. Gan, J., Li, L., Zhang, D., Yi, Z., Xiang, Q.: An alternative method for traffic accident severity prediction: using deep forests algorithm. Journal of Advanced Transportation **2020**(1), 1257627 (2020)
2. Madushani, J.P.S.S., Sandamal, R.M.K., Meddage, D.P.P., Pasindu, H.R., Gomes, P.I.A.: Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers. Transportation Engineering **13**, 100190 (2023)
3. Fiorentini, N., Losa, M.: Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures **5**(7), 61 (2020)
4. Li, Y., Yang, Z., Xing, L., Yuan, C., Liu, F., Wu, D., Yang, H.: Crash injury severity prediction considering data imbalance: a Wasserstein generative adversarial network with gradient penalty approach. Accident Analysis and Prevention **192**, 107271 (2023)
5. Aziz, K., Chen, F., Ahmad, M., Khan, M.S., Sabri Sabri, M.M., Almujibah, H.: An interpretable dynamic ensemble selection multiclass imbalance approach with ensemble imbalance learning for predicting road crash injury severity. Scientific Reports **15**(1), 24666 (2025)
6. Aboulola, O.I.: Improving traffic accident severity prediction using MobileNet transfer learning model and SHAP XAI technique. PLOS ONE **19**(4), e0300640 (2024)
7. Yan, M., Shen, Y.: Traffic accident severity prediction based on random forest. Sustainability **14**(3), 1729 (2022)
8. Joloudari, J.H., Marefat, A., Nematollahi, M.A., Oyelere, S.S., Hussain, S.: Effective class-imbalance learning based on SMOTE and convolutional neural networks. Applied Sciences **13**(6), 4006 (2023)
9. Wang, H., Liang, Q., Hancock, J.T., Khoshgoftaar, T.M.: Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. Journal of Big Data **11**, 44 (2024)

10. Yu, H., Hutson, A.D.: A robust Spearman correlation coefficient permutation test. Communications in Statistics – Theory and Methods **53**(6), 2141–2153 (2022)