Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Predicting and Analysing Road Accident Severity with Machine learning Models and Resampling

Xin Wang

School of Computing, Australian National University, Canberra, Australia

*Corresponding author: Xin.Wang1@anu.edu.au

Abstract. Road accident threaten people's life safety as well as wealth seriously, therefore predicting and analysing car accidents have great significance. This research first compared the performance of four machine learning models in analysing the severity of traffic accidents, including Random Forest, Naïve Bayes, Logistic Regression and Multi-Layer Perceptron. Naïve Bayes performs badly with a low accuracy while other three models have equal level of performance. Multi-Layer Perceptron performs better in minority classes than Random Forest and Logistic Regression. Among all features, Random Forest focus on geographical and time features, Logistic Regression and Multi-Layer Perceptron focus on driving features, including lighting and road surface. Then resampling method is applied to imbalanced data. After trained in resampled data, Random Forest and Logistic Regression perform better in minority classes with higher precision and recall. In summary, this research compares the performance of machine learning models in road accident severity predicting and analysing, and addresses the challenge of imbalanced data with resampling method.

Keywords: Road accident; machine learning; resampling.

1. Introduction

Traffic accidents pose a severe threat to our society. According to the World Health Organization, there are approximately 1.19 million deaths each year because of traffic accidents, costing approximately 3% of most countries' GDP, which not only shows a negative impact on people's safety, but also lead to a substantial economic burden [1]. In Australia alone, there were 1194 road crash deaths in 2022, 5.8% more compare to the previous year, indicating an increasing trend [2]. Furthermore, autonomous driving technologies introduce various ethical problems and potentially unpredictable influences on traffic accidents [3]. Therefore, gaining deeper knowledge about traffic accidents and predicting as well as analysing them have a great significance.

Given the great significance of this problem, the prediction and analysis of road accident is an important topic. If the risk of traffic accidents can be predicted and what caused accidents can be find out, improvements can be made to avoid them. With their remarkable predictive performance, machine learning models offer a viable approach for this type of task. A variety of machine learning models has been applied in previous researches. For example, Chen TY.et al. used K-Means and Random Forest to predict the risk of lane-changing behaviors [4]. Chen C .et al have predicted collision risk using a Bayesian Neural Network [5]. Shangguan QQ .et al predicted the risk of driving using a multi-layer perceptron with a changing time window length [6]. All of these researches indicate that machine learning algorithms are suitable for this specific topic.

However, severe accidents only represent a small fraction among the total number of all accidents, the more severe a kind of accident is, the fewer its amount would be. According to the data collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police, fatal accidents are only 0.09% of all accidents [7]. This kind of unbalanced data would result in models with bias. Those models would ignore minor classes, but the accuracy is still high. To improve this, resampling methods can be applied. Resampling methods would increase the minority data or decrease the majority data to generate a balanced dataset [8].

This paper aims to predict the severity of traffic accidents, analyse the contributing factors behind those accidents, and try to solve the imbalanced data problem. Four models, including Random Forest, Naïve Bayes, Logistic Regression and Multi-Layer Perceptron were applied to achieve that goal, and some comparisons and discussions have been made. After that, some resampling methods are used to discuss whether resampling would improve the performance of those models.

2. Organization of the Text

2.1. Data description

The data is collected and published by Transport of NSW from New South Wales Government [9]. It contains 95874 accidents happened in New South Wales of Australia from 2018 to 2022. This data set contain following features: (1) Accident-related features: severity level, number of injured people, time and other features. (2) Weather features: weather, lighting condition, road surface and more. (3) Location features: Street type, latitude, longitude, town and other location information.

The target variable is 'Degree of crash - detailed', denotes the severity of accident, its number of unique values can be viewed in Fig. 1. The severity is highly-imbalanced, where fatal accidents only count for 1% while non-casualty accidents count for 32%.

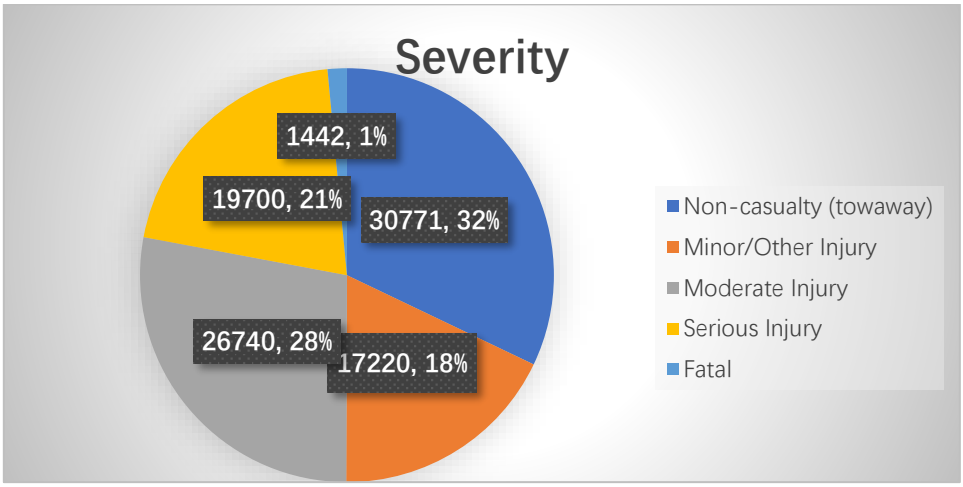

Fig. 1 The severity distribution of data

2.2. Data preprocessing

Most features in the dataset cannot be directly used for model training. For instance, Logistic Regression and Multi-Layer Perceptron requires numerical input. Furthermore, some features contain information about the severity and injury condition of the accident, which cannot be known before accident occurs. These features should be deleted to prevent data leakage. To address such issues, data-preprocessing is conducted by data transformation and deletion, which is explained detailed below:

(1) Feature deletion: Features meeting the following criteria are removed: Firstly, irrelated features such as Crash ID are deleted. Secondly, features containing too many missing values that cannot be imputed are deleted. For example, features like the number of route or other transport unit type contain too many missing values, therefore they are deleted. Lastly, features highly related to target value and unknown before the accident are deleted. For instance, the number of people that is seriously injured, or the kind of collusion, are deleted from dataset.

(2) Feature encoding: categorized features are transformed to numerical values using encoding methods. Features with large number of unique values, such as town, are transformed using frequency encoding. Features containing a small number of missing values and lacking sequential order are transformed using one-hot encoding, which generates the same number of features according to the

number of unique values in the original feature. Features contain sequential order, such as surface condition, are transformed using ordinal encoding.

(3) Normalization: all features are normalized to [0, 1] scale. This aim of this process is to stabilize model and accelerate the convergence of models, and make sure feature importance is comparable [10].

2.3. Models and implementation

Four models have been implemented: Random Forest, Naïve Bayesian, Logistic Regression and Multi-Layer perceptron. Each model has its own advantages and characteristics.

(1) Random Forest is proposed by Leo Breiman in 2001 [11]. It's a combination of decision trees, where trees are different due to random selection of features, which makes the model robust to noise while keep a satisfying performance level. Its ensemble learning usually performs well when dealing with imbalanced data.

(2) Naïve Bayes is a classification algorithm based on probability and bayes theory. It assumes features are independent, which is not true in most cases but the model performs well [12]. The model can output predictions by direct calculation without a learning process, thus the model runs faster than other models.

(3) Logistic Regression is proposed by David Cox in 1958, which is a widely-applied machine learning model nowadays [13]. It assumes the relationship between input and output is linear. It's a simple and understandable model with outstanding performance and good interpretability.

(4) Multi-Layer Perceptron designed by Frank Rosenblat is the foundation of future Neural Networks [14]. It has the ability of learning useful representations and hidden patterns from data. Multi-Layer Perceptron is a popular model with wide application.

Models are implemented with python. Random Forest, Naïve Bayes and Logistic Regression were implemented using scikit-learn and Multi-Layer Perceptron is implemented by pytorch [15,16]. Models are tuned to make sure models reach convergence, and the training time are similar to ensure the comparison process is valid and fair. Naïve Bayes does not need training and runs much faster than other models. The detailed training time is shows in Table 1.

Table 1. Model training time (seconds)

Random Forest	Naïve Bayes	Logistic Regression	Multi-Layer perceptron
25.00	1.17	28.45	24.24

3. Results and Analysis

3.1. Results without resampling

Table 2. Result of models. (0 - Non-casualty, 1 - Minor/Other Injury, 2 - Moderate Injury, 3 - Serious Injury, 4 - Fatal)

	Random Forest	Naïve Bayes	Logistic Regression	MLP
Accuracy	0.4612	0.2020	0.4607	0.4578
Precision 0	0.5497	0.6234	0.5369	0.5520
Recall 0	0.7080	0.3625	0.7377	0.6948
Precision 1	0.4091	0.3228	0.4291	0.4001
Recall 1	0.1314	0.3822	0.0643	0.1713

Precision 2	0.3978	0.3226	0.3909	0.3939
Recall 2	0.4614	0.0037	0.5309	0.4468
Precision 3	0.4001	0.1587	0.4211	0.3936
Recall 3	0.4068	0.0026	0.3234	0.3943
Precision 4	0.0000	0.0247	0.0000	0.5000
Recall 4	0.0000	0.9688	0.0000	0.0035

The performance of our models on all classes are shown in Table 2. Given the numerical results from Table 2, some comparison and discussion can be made:

(1) The overall accuracy of Random Forest, Logistic Regression and MLP are in the same level, approximately 0.457 to 0.461. Naïve Bayesian has the lowest accuracy 0.20, it has a high recall of 0.97 and low precision 0.02 for class 4, which is the class with lowest frequency. This indicates it has mis-classified plenty of instances as fatal accident.

(2) For the minority class, Random Forest and Logistic Regression gives 0 precision and recall, which means they did not classify any instances as fatal accident. Multi-Layer Perceptron performs much better with a 0.5 precision, but the recall of 0.0035 means it still misses most instances of fatal accidents.

(3) Random Forest, Logistic Regression and MLP performs at the same level for majority classes, with precision from 0.4 to 0.5 and recall from 0.3 to 0.7. Naïve Bayes performs badly for majority classes with a lower precision and recall.

To further analyse the relationship between models, we can use Person Correlation Coefficient for the predictions of models on test dataset. Through this, we can infer whether models give similar predictions or they focus on different classes. The coefficient values are shown in Table 3. Overall, models have a coefficient from 0.7 to 0.8, which indicates high positive relations. Among all models, Random Forest, Logistic Regression and MLP are highly related to each other with a coefficient around 0.8, while Naïve Bayes are slightly unrelated with a coefficient of 0.70 to 0.76.

Table 3. Person Correlation Coefficient

	Random Forest	Naïve Bayes	Logistic Regression	MLP
Random Forest	1	0.7297	0.8197	0.8000
Naïve Bayes	——	1	0.7611	0.7055
Logistic Regression	——	——	1	0.8282
Multi-Layer Perceptron	——	——	——	1

3.2. Model interpretability

Model interpretability is how human understand the deep mechanism of models, and one approach analysing feature importance, so human can understand with feature contributes the most to prediction results. Among four models above, feature importance can be extracted from Random Forest naturally due to its rule-based learning approach, by calculating the Gini Impurity. The feature importance of Logistic Regression can be inferred by the absolute value of coefficient, because the features has been normalized before.

Table 4. Top 10 important features of three models

	Random Forest	Logistic Regression	Multi-Layer Perceptron
1	Longitude	Pedal cycle	Street lighting unknown
2	Latitude	Motorcycle	Motorcycle
3	Town frequency	Speed	Street lighting unknown
4	Feature frequency	Multiple intersection	Speed
5	Street frequency	Motor vehicle unknown	Street lighting nil
6	Month	Street lighting unknown	Longitude
7	LGA frequency	School zone unknown	Street lighting off
8	Time	Urbanization unknown	4 wheel drive
9	Day of week	Single limited access	Latitude
10	Year of Crash	Ped crossing	Pedal cycle

For Multi-Layer Perceptron, feature importance can be inferred by Permutation technique. Permutation is a method that randomly shuffle a feature each time to see how the result is affected. If the output is significantly affected, then the feature has a greater significance. In summary, the feature importance of above three models can be extracted, which is shown in Table 4. From above table, we can analyse models by feature importance:

(1) Random Forest focus on geographical and time features: longitude, latitude, month, day and time. Features that encoded using frequency encoding are also important, indicating that Random Forest believes that places with large number of accidents before are more likely to have new accidents, while Logistic Regression and MLP do not focus on frequency encoded features.

(2) Logistic Regression focus on more specific features instead of geographical and time features: the vehicle type, speed and road lighting are important. Multi-Layer Perceptron focus on similar features, while geographical features like longitude and latitude are also important.

In summary, Logistic Regression and MLP focus on specific details while Random Forest focus on time and place.

3.3. Resampling

The resampling method applied here is SMOTE, implemented using imbalanced-learn, a python library offering methods for dealing with imbalanced data [17]. SMOTE method generates new instances of minority classes by interpolating new instances between existing data samples. After resampling, all classes contain 24660 instances. The comparison can be viewed in Fig. 2., the data are equally distributed after resampling. Resampling is only applied to the training data but not the test data, as the models should be tested in real-life data, which is data without resampling in this case.

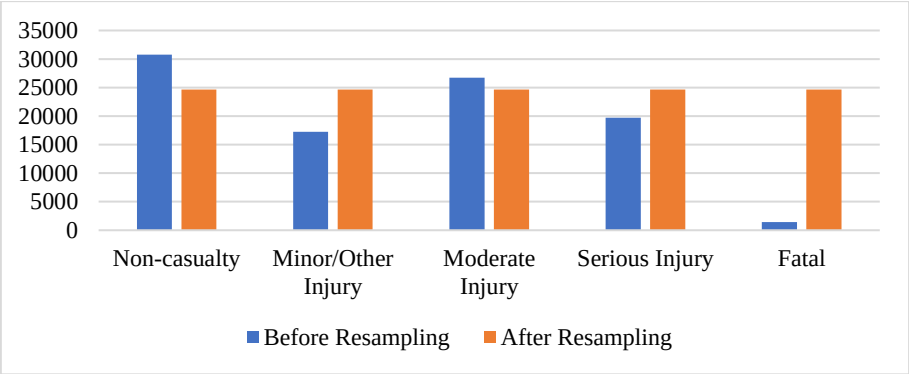

Fig. 2. Data distribution before and after resampling

Table 5. Results after resampling

	Random Forest	Naïve Bayesian	Logistic Regression	MLP
Accuracy	0.4519	0.2356	0.4561	0.4569
Precision 0	0.5593	0.4868	0.5482	0.5487
Recall 0	0.6655	0.3021	0.6873	0.6996
Precision 1	0.3763	0.2596	0.4037	0.3940
Recall 1	0.1620	0.4963	0.1425	0.1147
Precision 2	0.3968	0.4127	0.3967	0.3912
Recall 2	0.4171	0.0571	0.4751	0.5330
Precision 3	0.3788	0.3612	0.3991	0.4158
Recall 3	0.4598	0.1095	0.3820	0.3140
Precision 4	0.1154	0.0279	0.1045	0.0752
Recall 4	0.0104	0.6493	0.0486	0.0347

The performances of models after resampling are shows in Table 5. By comparing Table 2 and Table 5, we can discuss the influence of resampling on models:

(1) The overall accuracy of Random Forest, Logistic Regression and MLP is slightly decreased, where Random Forest decreased by 0.01, Logistic Regression decreased 0.004 and MLP decreased 0.0009. For Naïve Bayes, the performance is increased by 0.03.

(2) For minority classes, the performance of Random Forest and Logistic Regression is improved. The precision increased from 0 to about 0.1 and recall increased from 0 to 0.01 or 0.04. However, the performance in those classes is still worse than majority classes.

(3) MLP has a lower precision in class 4 with a higher recall, indicating it classified more instances as fatal accidents.

Overall, resampling has improved the performance of models in minor classes.

4. Conclusion

In this research, after data preprocessing, four models are applied in a car accident database to predict the severity level of accidents, including Random Forest, Naïve Bayes, Logistic Regression and Multi-Layer Perceptron. Among all models, Random Forest, Logistic Regression and Multi-Layer Perceptron have same level of performance while Naïve Bayes performs badly. Multi-Layer

Perceptron shows a better performance in majority classes than Random Forest and Logistic Regression. Model predictions have a high positive relation. Random Forest focus on time and geographical After applying resampling, Naïve Bayes has a better overall performance. Random Forest and Logistic Regression perform better in minor classes, while MLP is not significantly influenced.

Although four models have been attempted, other various models might have the potential of achieving a better performance, especially models with more complicated structures. Other resampling methods may also be attempter in the future. The dataset contains around 50 features, containing time and geographical data, road data, location data and other types of features. However, other features can be introduced, to gain a better performance. For example, vehicle engine status or mileage might affect car condition directly, but collecting those data requires more sensors, and there are ethical issues like privacy to be consider.

References

[1] World Health Organization, Road traffic injuries, 2023.12.13, accessed 2024.4.15, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2] Bureau of Infrastructure and Transport Research Economics, Road Trauma Australia—Annual Summaries, 2023.5.10, accessed 2024.3.29, https://www.bitre.gov.au/publications/ongoing/road_deaths_australia_annual_summaries

[3] L. Chen et al., "Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys," in IEEE Transactions on Intelligent Vehicles, vol. 8, no. 2, pp. 1046-1056, Feb. 2023

[4] Chen, Tianyi, et al. "Key Feature Selection and Risk Prediction for Lane-Changing Behaviors Based on Vehicles' Trajectory Data." Accident Analysis and Prevention, vol. 129, pp. 156–69, 2019

[5] Chen, Chen, et al. "A Rear-End Collision Risk Evaluation and Control Scheme Using a Bayesian Network Model." IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 264–84, 2019

[6] Shangguan, Qiangqiang, et al. "An Integrated Methodology for Real-Time Driving Risk Status Prediction Using Naturalistic Driving Data." Accident Analysis and Prevention, vol. 156, pp. 106122–106122, 2021

[7] United States Government's open data site, Crash Reporting - Drivers Data, 2024.2.9, accessed 2024.2.18, https://catalog.data.gov/dataset/crash-reporting-drivers-data

[8] Chen, Tianyi, et al. “Key Feature Selection and Risk Prediction for Lane-Changing Behaviors Based on Vehicles’ Trajectory Data.” Accident Analysis and Prevention., vol. 129, pp. 156–69, 2019

[9] Transport for NSW, NSW Road Crash Data, 2024.1.11, accessed 2024.2.18. https://data.nsw.gov.au/data/dataset/2-nsw-crash-data

[10] LeCun, Yann, Leon Bottou, Genevieve B. Orr, and Klaus -Robert Müller. “Efficient BackProp.” In Lecture Notes in Computer Science, 9–50. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[11] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001)

[12] Rish, Irina, An Empirical Study of the Naïve Bayes Classifier, IJCAI 2001 Work Empir Methods Artif Intell. 3, 2001

[13] Cox, D. R. “The Regression Analysis of Binary Sequences.” Journal of the Royal Statistical Society. Series B (Methodological), vol. 20, no. 2, pp. 215–42. JSTOR, 1958

[14] Rosenblatt, F, The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408, 1958

[15] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

[16] Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” arXiv.org (2019).

[17] Lemaitre, Guillaume, Fernando Nogueira, and Christos K Aridas. “Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning.” Journal of machine learning research 18 (2017): 1–5.