4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15041

Cape Town road traffic accident analysis: Utilising supervised learning techniques and discussing their effectiveness

Christo du Toit, Sulaiman Salau, Sebnem Er
Statistical Sciences Department, University of Cape Town, South Africa.

Abstract

Road traffic accidents (RTA) are a major cause of death and injury around the world. The use of Supervised Learning (SL) methods to understand the frequency and injury-severity of RTAs are of utmost importance in designing appropriate interventions. Data on RTAs that occurred in the city of Cape Town during 2015-2017 are used for this study. The data contain the injury-severity (no injury, slight, serious and fatal injury) of the RTAs as well as several accident-related variables. Additional locational and situational variables were added to the dataset. Four training datasets were analysed: the original imbalanced data, data with the minority class over-sampled, data with the majority class under-sampled and data with synthetically created observations. The performance of different SL methods were compared using accuracy, recall, precision and F1 score evaluation metrics and based on the average recall the ANN was selected as the best performing model on the validation data.

Keywords: *Road traffic accidents; supervised learning methods; imbalanced data..*

1. Introduction

A road traffic accident (RTA) can be defined as a rare, random, multi-factor event always preceded by a situation in which one or more road users fail to cope with the road environment (Rospa, 2002). In 2018, there were 12,921 fatalities recorded in South Africa as a result of RTAs. In addition to the social cost, RTAs also have significant economic costs for South Africa. In order to effectively reduce the number and injury-severity of RTAs in South Africa, a better understanding of the relationship between RTA injury-severity and accident-related factors is needed. The use of supervised learning (SL) methods can be very useful for informing future road safety campaigns and potentially reducing the frequency and injury-severity of RTAs.

Several statistical models such as logistic regression, classification and regression trees (CART), random forests (RF) and artificial neural networks (ANNs) have been effectively employed in previous research in different countries including Saudi Arabia, the United States, Italy and Canada, (Al-Ghamdi, 2002; Kong & Yang, 2010; Chang & Wang, 2006; Montella, *et al.*, 2012; Akın & Akbaç, 2010; Chong, *et al.*, 2005; Olutayo & Eludire, 2014) to predict injury-severity of RTAs. These SL methods were shown to regularly outperform logistic regression methods.

There are very few studies conducted on RTAs in South Africa predicting RTA injury-severity. South African literature mostly consists of identifying significant contributors to RTAs. The literature suggests that the quality of South African RTA data is generally poor due to issues such as underreporting, duplication as well as missing values in the data. There are limited studies modeling RTA injury-severity using SL methods, with the focus area mainly in the province of Gauteng (Govender, *et al.*, 2020; Mokoatle, *et al.*, 2019; Twala, 2013; Saar-Tsechansk & Provost, 2007; Moyana & Chibira, 2016). There is a definite need for more research focusing on South African RTA injury-severity prediction especially in the Western Cape province, which is one of the few provinces in South Africa with comprehensive and easily accessible RTA data. This research aims to contribute to the literature by including variables generated from external resources (ie. weather-related variables and geolocation related variables) in addition to the variables obtained from the provincial database. Additionally, it is aimed to highlight the best SL modeling and data sampling approaches for addressing the issue of class imbalance in RTA data.

2. Data and Methods

2.1. Data

The dataset used for this study contains records of more than 82,000 RTAs that occurred during the 2015-2017 period in Cape Town. The data were sourced from the City of Cape

Town, one of the major cities in South Africa and located in the Western Cape Province. The city has a well developed and managed road network and provides researchers with access to its comprehensive RTA database. The dataset contains several variables related to the accident such as street name, crash date, weekday, time of day, alleged cause, crash type, vehicle type, number of vehicles, number of passengers, number of pedestrians involved in the accident as well as the worst injury-severity sustained during the accident. Additionally, the data was enriched with weather-related variables such as temperature, precipitation, wind speed, visibility and cloud cover. Several other variables such as those relating to whether an accident occurred on a public holiday, on a weekend, the season the accident occurred, the number of vehicles involved, whether the accident occurred during peak traffic times as well as whether an accident occurred at an intersection or non-intersection were also added. The location of an accident (amongst other variables) was geocoded in order to obtain geographical coordinates for each accident. After inspecting that valid coordinates were returned for each accident's street address, the longitude and latitude coordinates were added as variables to the dataset.

A common issue with RTA datasets is that the classification categories are imbalanced. The target variable consists of four injury classes, namely: "fatal" (0.27%), "serious" (2.54%), "slight" (10.33%) and "no injury" (86.86%) and is severely imbalanced in Cape Town for the period of 2015-2017 (N= 82,363). Imbalanced data can negatively affect the performance of certain classification methods, especially with regards to predicting the minority class (Weiss & Provost, 2001). This is an issue since the minority class is often the class researchers are most interested in predicting correctly.

Three common data sampling approaches used by researchers to address imbalanced data are utilised in this study, namely (i) undersampling of the majority class, (ii) oversampling of the minority class and (iii) Synthetic Minority Oversampling Technique (SMOTE). SMOTE is a popular over-sampling method developed by Chawla, *et al.* (2002), that creates artificial data examples of the minority class in order to improve the imbalanced distribution of the target variable. While random over-sampling methods simply duplicate existing minority class examples, SMOTE creates artificial minority examples by extrapolating between existing minority examples by finding the k-nearest neighbours of the minority class for each minority example and then generating artificial examples in the feature space of the nearest neighbours. The artificial examples cause the classifier to create larger and less specific decision boundaries resulting in decision region for the minority class to become more general (Chawla, *et al.*, 2002).

The original imbalanced data as well as the data sets generated under different sampling schemes are analysed using multinomial logistic regression, CT, RF, Gradient Boosted Machine (GBM) and ANN methods to predict the target variable, the worst injury-severity resulting from a RTA. The next section briefly discusses the methods, therefore the authors recommend reading the resources such as Hastie, *et al.*, (2009) and Gareth, *et al.*, (2013) for further details of the various methods.

2.2. Methods

The multinomial logistic regression (MLR) model calculates the probability of an RTA belonging to each injury-severity category relative to a reference category, "no injury" in this case (Yasmin & Eluru, 2013). Classification trees (CT) are a non-parametric classification method that do not require any pre-defined underlying relationship between the predictor variables and the target variables to be specified. CTs use a tree-like structure in order to model the relationship between the predictor variables and the target variable. While a CT might be easily interpretable, it comes at the expense of predictive accuracy as well as high sampling variability (Chang & Wang, 2006). Random forests (RFs) can be used to reduce the variance of a SL method such as CTs (Hastie, *et al.*, 2009). This method is built on the idea that averaging a set of predictions reduces the variance. RF is an ensemble learning method, meaning that many CTs are combined/ensembled into one, better model. An RF model is built by growing a multiple number of trees, $B > 0$, on bootstrapped samples (random sub-samples of data with replacement). Gradient boosting machines (GBM), similar to RF, is an ensemble learning method. Unlike RF, which grows trees independently, GBM grows trees sequentially. This means each tree can learn from the errors made by the previous trees. Artificial neural networks (ANN) are another SL method that can be used for both regression and classification problems. ANNs are especially useful when one does not need interpretable results and when there are non-linear relationships present in the data (Hastie, *et al.*, 2009).

2.3. Evaluation Metrics

To identify the "best" performing model with regards to predicting RTA injury-severity, evaluation metrics are needed to compare the performance of the different models (multinomial logistic regression, CT, RF, GBM and ANN). While accuracy might be the most simple metric to understand and calculate, it can be misleading especially when dealing with imbalanced data. A model could classify all observations as the majority class and still achieve a relatively high accuracy despite it failing to identify any observations belonging to the minority class. For this reason, the validation data performance of the models are evaluated using the average (i) accuracy, (ii) recall, (iii) precision and (iv) F1 score of each model (Gareth, *et al.*, 2013, p. 149, Grandini, *et al.*, 2020).

3. Results

The analyses were conducted using R (R Team, 2013) and all plots were created using the "ggplot2" package (Wickham, et al., 2016) and models were built with the "caret" (Kuhn,

2008) package. Hyperparameter tuning for RF, GBM and ANN models was performed under University of Cape Town’s ICTS High Performance Computing cluster: hpc.uct.ac.za. Different hyperparameters were tested with cross validation. For RFs, the number of predictors at each split considered were (2, 3 and 7), and the number of trees were (500 and 1000); in the case of the GBM models, the number of trees considered were (50, 100, 150, 500), the different shrinkage parameter or learning rates were (0.005, 0.01, and 0.05), and the number of splits in each tree were (2, 10, and 20); for NNs, three hidden layers with varying number of neurons (layer1: 30, 40, 55, layer2: 0, 15, 25, 35, layer3: 0, 5, 15, 25) were tested with different weight decay (0.0001, 0.001, 0.1) settings.

The average recall of the different SL methods are compared in order to identify the best model based on its performance on the validation data. The comparisons of the average recall of all SL methods are shown in Figure 1 (a).

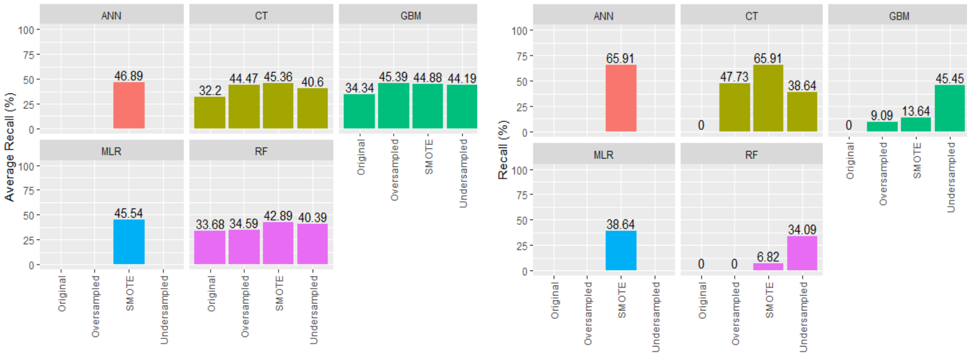

Figure 1. (a) Comparison of average recall *(b) Comparison of recall for “fatal” RTAs*

There is a noticeable difference in average recall between the different SL methods. As shown in Figure 1 (a), the model with the highest average recall is the ANN trained on the SMOTE training data (46.89%). The multinomial logistic regression model trained on the SMOTE data achieved the second highest average recall. The GBM model trained on the oversampled data and the CT trained on the SMOTE data have the next highest average recall respectively with RF model trained achieving the lowest value. It is important to note that some models, ie. ANN in the original dataset, failed to predict the true positive cases (fatal class), hence resulting in evaluation metrics of 0%.

Since RTAs that result in “fatal” or “serious” injuries carry the highest social and economic impact, the misclassification of these classes is undoubtedly the most important issue in predictions. Therefore, a comparison of the recall for the “fatal” RTAs of the different SL methods is shown in Figure 1 (b). The results show that ANN and CT trained on the SMOTE data achieved the highest recall for “fatal” RTAs compared to the other models. The results show that the ANN and CT models achieved the highest recall for “fatal” RTAs overall, followed by the GBM, multinomial logistic regression and finally the RF models respectively. The CT, RF and GBM models trained on the original data could not identify any “fatal” RTAs. The RF model trained on the oversampled data also failed to identify any “fatal” RTAs.

The ANN model trained on the SMOTE data was selected as the “best” performing model and its performance on the test data, also known as “unseen” data, is assessed. The confusion matrix of the ANN’s performance on the test data is shown in Table 1, while the evaluation metrics are shown in Table 2.

Table 1. Confusion matrix of ANN on test data

		Actual Category			
		Fatal	No Injury	Serious	Slight
Predicted Category	**Fatal**	38	1011	223	574
	No Injury	2	12257	64	698
	Serious	1	132	80	157
	Slight	3	908	52	272
				Overall Accuracy: 76.78%	

Table 2. Evaluation metrics of ANN on test data by class

Class	Recall (%)	Precision (%)	F1 (%)
Fatal	86.36	2.06	4.02
No Injury	85.67	94.13	89.70
Serious	19.09	21.62	20.28
Slight	15.99	22.02	18.53
Average	**51.78**	**34.96**	**33.13**

The model has an overall accuracy of 76.78% while being able to correctly identify some RTAs belonging to each of the four different injury-severity categories. The model has a higher average recall (51.78%) than any of the SL methods manage to achieve on the validation data. As shown in Table 1, the model also managed to correctly identify a large number of “fatal” and “no injury” RTAs, in contrast to fewer correctly identified “slight” and “serious” RTAs. Table 2 also shows that the ANN model has a very high recall for both “fatal” (86.36%) and “no injury” (85.67%) RTAs and a comparatively low recall for “slight” (15.99%) and “serious” (19.09%) RTAs. This is consistent with the findings of Chong, *et al.* (2005), who found that several SL methods applied in their study predicted “no injury” and “fatal” RTAs most accurately out of all the injury-severity categories. The model also has a high precision for “no injury” RTAs, indicating that it is very precise at correctly identifying “no injury” RTAs. This is in contrast with “fatal” RTAs, for which the model has a low precision score. This suggests that although the model correctly identifies the vast majority of “fatal” RTAs, it also results in a large number of false positives for “fatal” RTAs.

4. Recommendations and Future Work

Imbalanced data is a common issue found with RTA data. The comparison of the CT, RF, GBM and ANN models trained on the four different training datasets indicate that the best data sampling technique to address class imbalance in RTA datasets is the SMOTE technique with regards to maximising average recall. It is therefore recommended that future researchers use the SMOTE technique to address imbalanced RTA datasets when predicting RTA injury-severity.

It is recommended that the City of Cape Town expand and improve the quality of their RTA data. This study added several new predictor variables to the dataset obtained from the City of Cape Town, several of which were found to be significantly associated with RTA injury-severity. This will allow future researchers to analyse and model RTA injury-severity more comprehensively and identify the most comprehensive set of predictors that will help reduce the frequency and injury-severity of RTAs.

The data used for this study was sourced from the City of Cape Town, who collected and processed the data from the SAPS. The data contained several accident-related variables along with the RTA injury-severity. However, compared to RTA data used in similar international studies, the data used for this study contains relatively few accident-related variables. This can negatively affect the performance of the SL methods as the models are trained on data that is potentially missing some important accident-related variables.

The geographical coordinates of an RTA were added as predictor variables to the data. The use of models that explicitly take the spatio-temporal nature of RTA data into account could be beneficial since this study determined that the geographical location of an accident is significantly associated with RTA injury-severity. Reducing the cardinality of predictor variables in RTA data may also result in more interpretable models.

References

Akın, D. & Akbaç, B. (2010). A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics. *Scientific Research and Essays*, 5(19), 2837-2847.

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, 34(6), 729-741.

Chang, L.Y. & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis & Prevention*, 38(5), 1019-1027.

Chawla, N.V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chong, M., Abraham, A. & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica*, 29, 89-98.

Gareth, J., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*, Springer.

Govender, R., Sukhai, A. & van Niekerk, A. (2020). Driver intoxication and fatal crashes. Road Traffic Management Corporation Research and Development.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*, Springer.

Kong, C. & Yang, J. (2010). Logistic regression analysis of pedestrian casualty risk in passenger vehicle collisions in China. *Accident Analysis & Prevention*, 42(4), 987-993.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5), 1-26. doi:http://dx.doi.org/10.18637/jss.v028.i05

Mokoatle, M., Marivate, V. & Bukohwo, M. E. (2019). Predicting road traffic accident severity using accident report data in South Africa. Proceedings of the 20th Annual International Conference on Digital Government Research, 11-17.

Montella, A., Aria, M., D'Ambrosio, A. & Mauriello, F. (2012). Analysis of powered two-wheeler crashes in Italy by classification trees and rules discovery. *Accident Analysis & Prevention*, 49, 58-72.

Moyana, H. & Chibira, E. (2016). Improving safety in the road transport sector through road user behaviour changing interventions: a look at challenges and prospects. *Proceedings of the 35th Southern African Transport Conference (SATC 2016)*, 516-528.

Olutayo, V. A. & Eludire, A. A. (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. International Journal of Information Technology and Computer Science, 6(2), 22-28.

Rospa. (2002). The Royal Society for Prevention of Accidents (ROSPA) Road Safety Engineering Manual. Retrieved from hhttps://trid.trb.org/view/730321 (2021/07/19)

Saar-Tsechansky, M. & Provost, F. (2007). Handling missing values when applying classification models. Journal of Machine Learning Research, 8, 1625-1657.

Team, R. C. (2013). R: A language and environment for statistical computing.

Twala, B. (2013). Extracting grey relational systems from incomplete road traffic accidents data: the case of Gauteng Province in South Africa. *Expert Systems*, 31(3), 220-231.

Weiss, G. M. & Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study. DOI: https://doi.org/10.7282/t3-vpfw-sf95

Wickham, H., Chang, W. & Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version, 2(1), 1-189.

Yasmin, S. & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention, 59, 506-521.