Application of Machine Learning to Severe Injury Prediction in Rural Run-off-road Crashes

Chris Jurewicz[a] and Dr Farhana Ahmed[a]

[a]Australian Road Research Board, Melbourne, Australia

Abstract

This paper describes how Machine Learning (ML) techniques were used gain a deeper insight into the factors leading to rural run-off-road casualty crashes being severe. The ML findings were compared with a conventional binary logistic modelling approach.

The findings showed that roadside objects hit, road curvature, vehicle type and age, and the number of persons in vehicle were strong predictors of run-off-road crash severity. More importantly, ML highlighted specific combinations of risk factors which were linked to high risk of severe injury in a run-off-road casualty crash. ML may enable a more synergistic approach to risk and Safe System assessment.

Background

Run-off-road crashes are the cause of nearly half of all road-related severe injuries on rural roads in Victoria. Significant funds are being spent on Safe System road environment treatments to reduce occurrence and severity of run-off-road crashes. Yet, there has been limited published research linking road, vehicle and road user contributions to the problem.

Machine learning is an analytical branch of computer science which automates pattern-recognition to generate data models. Models (i.e. algorithms) learn from large volume of training data by iteratively recognising connections between variables and defining strength of these connections (Mannila & Heikki 1996, Domingos 2015). Thus, generalizing from training data, machine learning allows to find hidden data insights and to make predictions without being explicitly programmed.

This approach is often feasible where descriptive or inferential statistics would be of limited value, or time consuming due to scale and complexity of data (SAS 2016). As more data becomes available, more ambitious problems can be tackled with machine learning. This includes predicting rare events such as equipment failure, or deeper understanding of human behaviour e.g. web searches or new pricing models. As a result, this branch of computing science is becoming widely used in scientific and commercial fields.

There are multiple machine learning challenges which require a model to predict an output variable given a number of input variables. These problems can be divided into classification and regression problems (BigML Team 2018). The first is used for categorical output variables. The latter for continuous output variables. This paper focussed on a classification prediction whether a casualty crash on a rural state road in Victoria would be severe casualty crash (fatal or serious), or other injury (minor).

Method

Three machine learning techniques: Association Rules, Random Decision Forests and Deep Neural Networks, were used to identify a range of interlinked road, vehicle and road user risk factors contributing to severe injury outcomes in run-off-road casualty crashes on Victorian rural undivided roads. The machine learning findings were compared with those derived using a conventional binary logistic modelling approach.

Machine learning in non-parametric and makes no assumptions about correlation of input variables. Machine learning captures interactions between input variables, while standard binary logistic assumes variable independence. This variable interaction capability is important when seeking to recognise positive or negative synergies.

As in binary logistic modelling, validation and model performance metrics are used to manually optimise machine learning model designs. In both approaches, the process started with 26 input (independent) variables, with the output (dependent) variable being a probability of binary severe or non-severe crash. Using backward elimination, variables with little contribution/lack of significance were iteratively removed to optimise the models.

Victorian crash data obtained from VicRoads (2012-16) and ANRAM rural undivided road inventory data were used to develop training (80%) and validation data (20%) for the models.

Results

The findings of machine learning models showed that: roadside objects hit, road curvature, vehicle type and age, age of the most severely injured person, and the number of persons in vehicle were strong predictors of run-off-road crash severity. More importantly, machine learning techniques highlighted the specific combinations of factors linked to high risk of severe injury in a run-off-road casualty crash.

Association Rules, a data mining technique, was able to show groupings of variables which were more strongly associated with severe injury outcome when present together. Figure 1 shows one of such associations showing links between severe run-off-road crash outcome (FSI=1) and hitting a tree on a straight section (CurveBinary=0). Severe outcome is also associated with vehicle category 'car' made before 2012. Female gender of the most severely injured occupant was indirectly associated with severe outcome.

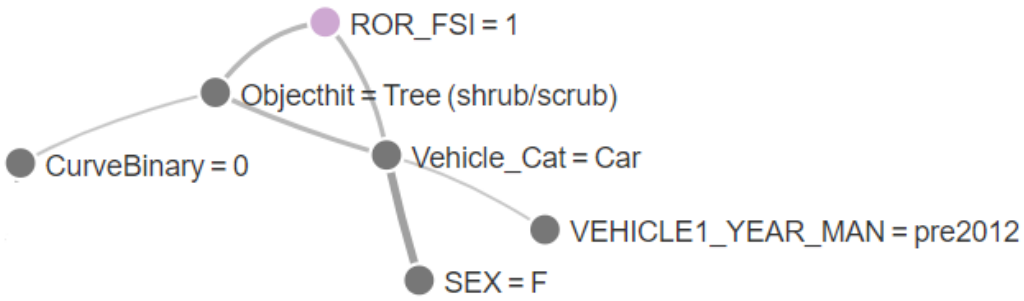

Figure 1. Example of an association rule between severe outcome and other variables

Figure 2 is an output of a Random Decision Forest ensemble and shows that children and seniors were at most risk of severe injury in run-off-road casualty crashes (lighter shade). It also shows that crashing in pre-2012 vehicle increased these chances considerably.

Overall, Deep Neural Network model validation against a random 20% sample not used in training provided slightly better Accuracy, Precision and Spearman's Rho (0.23), than Random Decision Forests model ensemble and the binary logistic model.

Binary logistic regression provided a statistically significant model with the best AIC chosen to promote parsimony, although with fewer variables. Object hit=tree/pole, curvature=present, occupant=senior, number of persons, vehicle category=motorcycle, and AADT, were retained, still providing a useful understanding of severe outcome risk.

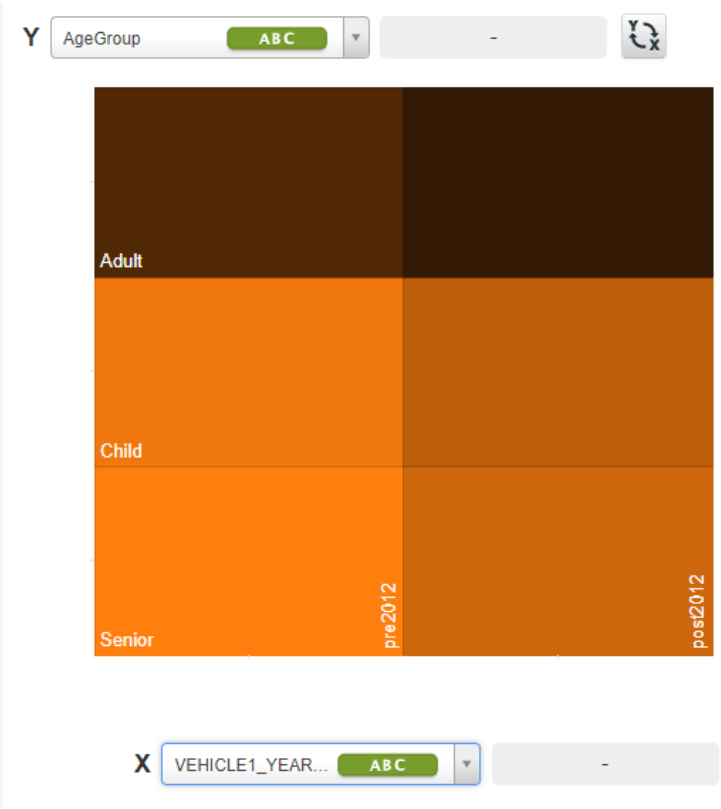

Figure 2. Example of machine learning insights: effect of vehicle age and most severely injured occupant age on probability of sever injury (lighter colour = higher)

Discussion and Conclusions

The paper discussed how selected machine learning techniques might contribute to a more systemic understanding of road safety by recognising and demonstrating interactions between risk factors, while providing comparable predictive strength to conventional methods. Also, machine learning methods are more automated and require less assumptions being made by the modeller (i.e. less source of bias, error).

This new knowledge may offer an opportunity for a more synergistic approach to road safety and its strategic improvement.

References

Domingos, P. 2015. A Few Useful Things to Know about Machine Learning, viewed 17 February 2018, <https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>

SAS 2016. Machine Learning: What it is and why it matters? Viewed 17 February 2018 <www.sas.com>

Mannila, Heikki 1996. Data mining: machine learning, statistics, and databases. Int'l Conf. Scientific and Statistical Database Management. IEEE Computer Society.

BigML Team (2018) Classification and Regression with the BigML Dashboard The BigML Team, Version 2.1. viewed January 2018 <www.bigml>