Investigating Injury Severity of Occupants at Highway-Rail Grade Crossings: An Application of Machine Learning Techniques and A Partial Least Square Ordinal Logit Model

Anae Sobhani[1], Hamed Kianmehr[2], Samira Soleimani [3], Ahmad Sobhani[4]

1 INTRODUCTION AND STUDY CONTEXT

Highway-railroad grade crossings (HRGCs) are critical locations where the railway and the roadway intersect. According to Yan et al., vehicle-train crash collisions are the most dangerous traffic accidents at highway-rail grade crossings since the average weight ratio of a train to a motor vehicle is about 4000 to 1 (Yan, Han, Richards, & Millegan, 2010) . Collisions between trains and vehicles usually cause severe injuries to vehicle passengers. There were 35,751 reported collisions between vehicles and trains in the United States from 2001 to 2015 (Federal & Railway Administration, 2017). According to statistical data obtained from the Federal Railway Administration, those collisions include 2849 fatal 9003 injury collisions (Federal & Railway Administration, 2017).

Although grade separation claimed to be the safest solution to reduce the accident rate, it is considered as a very expensive and not always possible solution. Instead, active and passive grade crossing control devices have been widely used as countermeasures to prevent accidents or at least reduce their negative consequences (i.e. severe injuries). Active control devices detect approaching trains and react by initiating sequences of flashing lights, bells, and closing gates. While, passive control devices only provide signs regarding grade crossing locations, and speed around grade crossings. It is interesting to note that based on studies on pre-crash behavior of vehicle drivers and severity of injury, over 80% of drivers do not fully understand the meaning of signs placed at railroad crossings (Liu et al., 2016; Rifaat, Tay, Perez, & Barros, 2009). Furthermore, disobeying the rules by vehicle drivers is another major cause of accident at grade crossings. Recent studies showed that around 10-20% of drivers choose to “beat the train” even when traffic lights are flashing and the train is within sight see ((Rifaat et al., 2009)). Hence, traffic controls affect driver behavior at the railroad crossings and surprisingly not always in an expedited positive way. In short, pre-crash driver behaviors, defined as motorist actions prior to the event of a crash, were found to be a good indicator of the effectiveness of crossing control devices and can result in higher or lower injury severity. There are limited studies conducted on investigate factors influencing crash severity at highway-railway grade crossings while considering drivers’ pre-crash behavior toward crossing control devices.

Machine learning is the practice of bringing quantitative datasets, analyzing and visualizing them in ways to bear on decision making and predicting futures by finding patterns from existing data. Machine Learning techniques (ML) employ different algorithms to extract knowledge/information from large datasets. Among MLs, Decision Tree (DT), Support Vector Machine (SVM), and Random Forest (RF) are one of the most popular and strong non-parametric methods that predict future responses within a black box framework. As an advantage of ML analytics, the introduced methods can handle complex data collected from different resources such as videos, pictures, surveys, text in efficient ways. The application of ML in different transportation domains has become popular recently (e.g. (Wong, Farooq, & Bilodeau, 2017)). However, the application of ML in crash severity analysis is still limited mostly to predict injury severity patterns.

[1] Assistant Professor, Department of Geography and Planning, Utrecht University, The Netherlands.
[2] Ph.D. Candidate, Department of Systems Science and Industrial Engineering, Binghamton State University, USA.
[3] Ph.D. Student, Department of Geography and Anthropology, Louisiana State University, USA.
[4] Assistant Professor, Department of Decision and Information Sciences, Oakland University, USA.

Hence, it is essential to use ML methods as data-oriented techniques in identifying and ranking factors reducing the injury severity of vehicle-train crashes.

This paper aims to study and compare the prediction accuracy and variables' detection of ML techniques and a regression logistic model. For the case study, this paper analyzes the effects of different factors (especially crossing control devices and pre-crash behavior of drivers) on drivers' injury severity at highway-railway grade crossings. The Partial Least Square ordinal Logit model (PLSL) is adopted in this paper; while its model estimation and prediction results are compared with DT, SVM, and RF techniques. To the best of our knowledge, this combination of analyzing techniques has not been used for assessing injury severity of crashes at highway-railroad crossings. Our study employed the crash data occurred in Texas State from 1990 to 2018. In addition to the crash data attributes, drivers' demographics, seasonal data, the attributes of highway-rail grade crossing (crossing control devices, land use, population density, etc.) (total of 45 attributes) were generated and added to the crash dataset.

2 METHODOLOGY

Employing the three Machine Learning techniques in our study follows three main steps: First, the dataset was preprocessed the data extensively to remove redundant attributes, retrieve missing values by a sequential series of univariate imputation models, and improve the unbalanced characteristics of the data by applying Synthetic Minority Over-sampling Technique which enabled us to construct ML classifiers from imbalanced dat. Second, the recursive Decision Tree Regression model was used to select attributes (e.g. crossing controls, drive pre-crash behavior indicators, type of vehicles, driver demographics etc.)) with a 95% level of confidence that have significant contributions on predicting injury severity levels (i.e. non-significant injury (level 1), major injury (level 2), and fatality (level 3)). Third, The Decision Tree Classification, Random Forest, and Support Vector Machine techniques were applied separately to mind the pattern of injury severities. These ML techniques are applied to random training datasets to determine the decision-making rules in predicting injury severity levels with respect to selected features/attributes from the step two. Then, DT, RF, and kernel SVM rules were evaluated by adopting random test sets.

Decision Tree is a supervised Machine Learning technique generates decision rules for selecting/ predicting a choice (called class in ML domain) according to input/independent attributes. DT builds regression as well as classification models in the form of a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. According to Classification And Regression Tree (CART), and CHi-squared Automatic Interaction Detector (CHAID) algorithms, the essence of the tree is that the features (attributes) are partitioned, starting with the first split. Splitting continue until the termination of the tree. Each subsequent split/partition is not done on the entire dataset but only on the portion of the prior split that it falls under. This top-down process is referred to as recursive partitioning. When class levels are discrete, DT uses Gini index, Equation (1), to select an attribute for partitioning the dataset at each node. Gini index measures the impurity of data partition. That is, a feature (attribute) is selected for partitioning at each node of the DT, if a dataset (D) after partitioning having the most number of same class levels. In other words, DT use Gini index to select an attribute that minimizes the information needed to classify observed cyclists, according to their class levels (alternatives) in the resulting partitions and reflects the least randomness or "impurity" in these partitions.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad (1)$$

p_i^2 is the probability that an observed cyclist in dataset D belongs to class level C_i. m is the total number of class levels (alternatives). For regression decision tree, use the residual sum of squares (RSS) at each node to select the appropriate attribute for partitioning. In other words, the selected attributes reduce the RSS after partitioning. A small RSS indicates a tight fit of the model to the data.

Decision Tree algorithms follow a kind of greedy approach in selecting and splitting attributes to develop the tree structure introduced above. Greedy means that, during each split in the process, the algorithm looks for the greatest reduction in the RSS or Gini index without any regard to how well it will perform on the latter partitions. The result is that one may end up with a full tree of unnecessary branches leading to a low bias, but a high variance that reduces predictability power. Random Forest technique is dealing with this issue by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. According to the Random Forest algorithm, an individual tree is built on a random sample of dataset, selected each time. This tree construction is repeated dozens or hundreds of times and the results are averaged. Each of generated trees is fully grown. By averaging the results, one can reduce the variance of prediction without increasing the bias. Random Forest also takes a random sample of the input features (attributes) at each split for each tree to mitigate the effect of a highly correlated predictor on developed decision rules and variance of prediction. Therefore, the subsequent averaging of the trees that are less correlated to each other and is more generalizable.

Support vector machines are a set of related supervised learning methods use ML theory to maximize predictive accuracy while automatically avoiding over-fit to the data. SVMs can be defined as systems which use hypothesis space of a linear or non-linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory.

After completing the ML analysis, we applied the PLSL model in to our data. To be sure, since in the focus of the case study is especially on studying the effect of crossing control devices and drivers' pre-crash behaviors jointly which raised the multicollinearity problem, adaptation of the Partial Least Square ordinal Logit model (PLSL) that deals with classification problems (i.e. multicollinearity issues) was appropriate for our analyzing purposes. This modeling approach enables us to bypass multicollinearity problem while specifying the contributions of predictor variables (crossing control devices, pre-crash behavior of drivers,

3 PRELIMINARY RESULTS

The Partial Least Square ordinal Logit prediction results based on the estimated parameters, resulted in the prediction accuracy of 63% of the observed injury severity levels of crashes at highway-railroad crossings. Table 1 presents the estimated variables' ranking which have statistically significant contribution to the severity of injuries (95% level of confidence) at highway-rail grade crossings. This ranking was generated according to the predictors' (absolute) highest marginal effects for the given injury severity levels level 1, 2, 3). Table 2 provides the predictors' ranking based on the employed ML methods along with each method prediction accuracy at the end of the table. Comparing the results confirmed that the Random Forest improves the prediction by 15%. Moreover, comparing the PLSL estimated parameters with RL predictors' rankings, it is observed that the RF and PLSL overlap by 73% in detecting important/significant variables affecting crash injury severities at highway-railroad crossings.

4 CONCLUSION

This study is a preliminary phase in comparing ML methods versus the Partial Least Square ordinal Logit as one of the well-known discrete choice modeling approaches that addresses attributes multicollinearity problem in the model. The results indicated that the Random Forest method outperformed the Partial Least Square ordinal Logit model in predicting crash injury severities at highway-railroad crossings. The results conclude that Random Forest has 73% overlap in detecting important variables affecting injury severities in compared with the Partial Least Square ordinal Logit (comparing Table 1 and 2). However, it cannot rank important predictors as precisely as the Partial Least Square ordinal Logit known as a parametric statistical model. Hence, preparing a more advanced systematic approach to set up ML methods is essential.

Table 1 Partial Least Square Ordinal Logit Model Estimation Results

Rank	**Variables/ Predictors**	**Highest Marginal Effects** *(crash severity level)*
1	Crossing (e.g. gates, flash lights, traffic Signal, Audible)	-0.50 *(level 1)*
2	Pre-crash behavior (e.g. went around/Thru due to temporarily or permanent barriers)	+ 0.39 *(level 1)*
3	Type of user involved	-0.47 *(level 1)*
4	Position of user involved (e.g. moving over crossing)	-0.25 *(level 2)*
5	Type of highway/road way at crossings (e.g. interstate)	-0.15 *(level 2)*
6	Train speed at the time of accident	0.13 *(level 3)*
7	Total number of vehicle occupants	0.08 *(level 1)*
8	Vehicle speed	-0.06 *(level 2)*
9	Type of land use (e.g. industrial – commercial)	0.01 – 0.03 *(level 2)*
10	Driver age	-0.03 *(level 2)*
11	Visibility (e.g. dark)	0.01 *(level 2)*
12	Number of traffic lanes crossing railroad	0.01 *(level 1)*

Table 2 Machine Learning Techniques Predictors' Ranking and Prediction Results

Random Forest		**Support Vector Machine**		**Decision Tree**	
Rank	**Variables/ Predictors**	**Rank**	**Variables/ Predictors**	**Rank**	**Variables/ Predictors**
1	Train speed	**1**	Train speed	**1**	Train Speed
2	Average daily traffic	**2**	Type of Highway/Roadway	**2**	Number of vehicle occupants
3	County population	**3**	Type of land Use	**3**	Type of highway/road way at crossings
4	Driver age	**4**	Functional classification of roads	**4**	Functional classification of roads
5	Crossing controls	**5**	Type of user involved	**5**	Crossing controls
6	Vehicle speed	**6**	Road way paved or not	**6**	Average daily traffic
7	Type of user involved	**7**	Number of vehicle occupants	**7**	Pre-crash behavior
8	Number of vehicle occupants	**8**	Crossing controls	**8**	Position of user involved
9	Type of land use	**9**	Number of main tracks	**9**	Vehicle speed
10	Season	**10**	Average daily traffic	**10**	Type of Land Use
11	Drive pre-crash behavior	**11**	County population	**11**	Type of user involved
12	Type of Land Use	**12**	Number of traffic lanes	**12**	Number of main tracks
Prediction Accuracy	**72%**		**58%**		**54%**

Note: The predictor rankings were generated based on series of systematic sensitive analysis.

This can be done by combining the Partial Least Square ordinal Logit and ML modeling approach (PLSL-ML) which might improve not only the prediction results, but also the important/significant variable detection. To be sure, the PLSL-ML modeling approach, as an extension of this research project, is in progress by the authors.

Keywords: Injury Severity, Highway-railroad grade crossings, Machine Learning techniques, Decision Tree, Support Vector Machine, Random Forest, Partial Least Square ordinal Logit model

REFERENCE

Federal, & Railway Administration, O. of S. A. (2017). Highway-Rail Accidents Statistical Data.

Liu, J., Wang, X., Khattak, A. J., Hu, J., Cui, J. X., & Ma, J. (2016). How big data serves for freight safety management at highway-rail grade crossings? A spatial approach fused with path analysis. *Neurocomputing*, *181*, 38–52. https://doi.org/10.1016/j.neucom.2015.08.098

Rifaat, S., Tay, R., Perez, A., & Barros, A. De. (2009). Effects of neighborhood street patterns on traffic collision frequency. *Journal of Transportation Safety and Security*, *1*(4), 241–253. https://doi.org/10.1080/19439960903328595

Wong, M., Farooq, B., & Bilodeau, G.-A. (2017). Latent Class model using discriminative restricted Boltzmann Machine. In *International Choice Modelling conference*. Retrieved from http://www.icmconference.org.uk/index.php/icmc/ICMC2017/paper/view/1134

Yan, X., Han, L. D., Richards, S., & Millegan, H. (2010). Train-vehicle crash risk comparison between before and after stop signs installed at highway-rail grade crossings. *Traffic Injury Prevention*, *11*(5), 535–542. https://doi.org/10.1080/15389588.2010.494314