Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms

Meghna Chakraborty[1], **Timothy Gates, Jhelum Chakravorty**

Department of Civil and Environmental Engineering, Michigan State University, 428 South Shaw Lane, East Lansing, MI 48824, USA, Email: chakra43@msu.edu
Department of Civil and Environmental Engineering, Michigan State University, 428 South Shaw Lane, East Lansing, MI 48824, USA, Email: gatestim@egr.msu.edu
Hitachi Energy Research, 800 Boulevard Hymus, Saint Laurent, QC H4S0B5, jhelum.chakravorty@hitachienergy.com

Extended ABSTRACT

Introduction

In traffic safety research, the development of reliable methodologies to predict and classify crash injury severity based on various explanatory variables has been crucial. Previous research focusing on traffic safety analyses has widely utilized classical statistical techniques [1–4]. While these parametric methods have undoubtedly provided insights, the fundamental characteristics of crash data often result in methodological limitations that are not fully accounted for. The parametric modeling techniques used for crash severity analysis are subject to rather strict assumptions about the distribution of data, and usually, a linear functional form between the dependent and explanatory variables. However, these assumptions may not always hold true. When basic assumptions of traditional statistical models were violated, erroneous estimations and incorrect inferences could be produced [5, 6]. To overcome the limitations associated with classical statistical models and the proficiency in capturing the nonlinear relationship between input and output data, more recently, researchers have proposed non-parametric methods and artificial intelligence algorithms for crash analysis. Furthermore, as an emerging analytics technique, deep learning is increasingly being introduced into safety research [7]. Also, the increasing availability of large-scale data from various sources, including connected and autonomous vehicles, and naturalistic driving studies, calls for a thorough understanding of the causal relationships between safety and various associated factors with the help of sophisticated methodological approaches.

In recent times, though there have been extensive research using big datasets, causal analysis for identifying causal structure and influential variables has received limited attention. Although there are different measures of causality, Granger causality is one of the most widely used methods, especially for static data, and it has been popular for identifying influential factors and prediction purposes in econometric studies and transportation research [8].

To this end, this study presents a methodological framework to model the severity of motor vehicle crashes on interstates. The background of this study is premised in Texas as it has historically been among the top states in terms of statewide fatalities in the U.S. The analysis involves causal inference, using Granger causality tests and injury severity classification using different machine learning and deep learning approaches including decision trees (DT), random forest (RF), extreme gradient boosting (XGBoost), and deep neural network (DNN). While Granger Causality helped identify the important factors affecting crash severity, the learning-based models predicted the severity classes with varying performance. The output of the proposed crash severity classification approach includes three classes: fatal and severe injury (KA) crashes, non-severe and possible injury (BC) crashes, and property damage only (PDO) crashes.

Causal Analysis

Granger causality test is a statistical hypothesis test which determines whether one time series is helpful in predicting another time series. In particular, a variable *X* Granger causes a variable *Y* if the prediction of *Y*, based on its own past and past of *X* is different than the prediction of *Y* based on its own past alone.

The notion of Granger causality assumes that cause happens before the effect and the cause has unique information about the future of the effect. With this, Granger causality can be formally defined as,

Definition: The hypothesis for Granger causality of *X* on *Y* is

$$P[Y(t+1) \in \Omega | I(t)] \neq P[Y(t+1) \in \Omega | I_{\hat{x}}(t)]$$

where $P[Y(t+1) \in \Omega | I(t)]$ is the probability of *Y(t+1)* belonging to the set Ω when the entire information till time *t* is considered (I(t)) and $P[Y(t+1) \in \Omega | I_{\hat{x}}(t)]$ is the probability of *Y(t+1)* belonging to the set Ω when *X* is

[1] * Corresponding author. Tel.: +1-480-634-3713;
E-mail address: chakra43@msu.edu

removed from the information set (denoted by $I_{\bar{x}}(t)$). When above hypothesis is satisfied, it is said *X* Granger causes *Y*.

Though the above definition is for a bivariate time-series, the naïve way to perform Granger causality test for causal discovery among variables of a multivariate time-series data is to consider two variables at a time, the relevant idea is the concept of conditional Granger causality.

Suppose *X, Y* and *Z* are three jointly distributed multivariate stochastic processes and consider the regression models

$$X_t = \alpha_t + \left(X_{t-1}^{(p)} \oplus Z_{t-1}^{(r)}\right) \cdot A + \epsilon_t \quad (1)$$

$$X_t = \alpha_t^{'} + \left(X_{t-1}^{(p)} \oplus Y_{t-1}^{(q)} \oplus Z_{t-1}^{(r)}\right) \cdot A^{'} + \epsilon_t^{'} \quad (2)$$

where *A* and *A'* are the regression coefficients, α and α' are constants and ϵ_t and $\epsilon_t^{'}$ are the residuals. The predictee variable *X* is first regressed on previous *p* lags of itself and *r* lags of the conditional variable *Z* and second on previous *p* lags of itself, *q* lags of *Y* and *r* lags of *Z*. With this, the Granger causality of *Y* on *X*, given *Z*, is

$$G_{Y \to X|Z} = \ln \frac{var(\epsilon_t)}{var(\epsilon_t')}$$

where $var(\cdot)$ denotes the variance, and $G_{Y \to X|Z}$ is a measure of the extent to which the inclusion of *Y* in the model (2) reduces the prediction error of (1).

Machine Learning Algorithms for Classification

In this study, a multi-class classification problem has been addressed using four different methods, namely, Decision Trees, Random Forests, XGBoost Classifier and finally Deep Neural Nets.

Decision Trees: Decision tree is a supervised learning algorithm which has a tree structure where at each node a yes-no kind of question is answered. Hence, unlike probabilistic classification algorithms, decision tree is a decision or rule-based algorithm. For decision tree classification, one uses the dataset features to create yes-no questions and this process continuously splits the dataset until all the data-points belonging to each class are isolated. The decision tree algorithm tries to completely divide the dataset such that to each leaf node the algorithm assigns the most common class among all the data points in that node.

Random Forest: Random Forest is an ensemble classification algorithm which builds a collection of decision trees and is usually trained using a bagging algorithm. Multiple decision trees are trained separately, and the output of the random forest is obtained as an average of the outputs of individual decision trees.

XGBoost Classifier: XGBoost, an ensemble algorithm, stands for extreme boosting in the sense that it *boosts* the performance of a regular gradient boosting algorithm. XGBoost creates a sequence of models that sequentially corrects the model obtained from the previous step. The main advantage XGBoost classifier over other ensemble classifiers is the fact that it is comparatively faster than other methods. Furthermore, the core algorithm can be parallelized and hence it can take advantage of multi-core processors. It can also be parallelized over GPUs and network of computers, thus making it feasible for really large datasets.

Deep Neural Network Classifier: Deep Neural Network is an artificial neural network with multiple *hidden* layers between the inputs and outputs. One of the main advantages of a Neural Net is the fact that it can approximate a nonlinear function to arbitrary accuracy, and this can be achieved by properly choosing the activation functions. Although the choice of activation function depends on the task, Rectified Linear Unit, Hyperbolic Tangent and Sigmoid are among the mostly used activation functions.

Analysis and Results

Data

The data on traffic crashes were obtained from the crash database maintained by Texas Department of Transportation (TxDOT), known as Crash Records Information System (CRIS) [9]. The traffic volumes in the crash database were from 2018. Therefore, the growth factors were applied from the Traffic Count Database System (TCDS) maintained by TxDOT to obtain traffic volumes of other years [10]. For this research, crash data from all interstates across the state were collected for a period of six years, between 2014 and 2019. For the purpose of this study, crashes occurring only on interstates in mostly urban areas (population > 50,000) and a few suburban areas on the fringes of major cities with population between 25,000 and 49,000 were evaluated. This resulted in a total of 156,166 crashes occurring across 59 counties statewide during the analysis period. The summary statistics of the data prior to addressing the data imbalance issue shows the traffic volumes vastly vary across the freeways. The crash locations also include work zones, hence, for approximately 2% of crashes, the speed limit was 45 mph, the minimum for this factor. Consistent with urban nature of roadways, over a quarter of observations have HOV lanes, and in most cases, more than 4 lanes, considering both traffic directions.

Granger Causality

A total of 24 variables were evaluated using Granger causality and a set of most influencing predictors were identified and ranked, based on the causality scores. The optimal lag for the VAR model for Granger causality

analysis was evaluated using AIC criterion and was determined to be 4. Ultimately, a total of 17 predictors with the highest scores were selected for the classification analysis and compared that with the full set of variables. For the most variables, the rank of importance makes sense and compare favorably with the earlier research [11, 12], indicating speed limits, traffic volumes, percent of heavy vehicles, or presence of workzones to be some of the most important factors in classifying crash injury severity. The Granger causality analysis was performed on R statistical software version 4.0.2.

Data Balancing and Crash Severity Classification

The classification of crash injury severity utilizing different machine learning algorithms is carried out with the 17 most important predictors as selected based on the causality score and its prediction performance is compared with that using all variables. The classification analysis was carried out in Python 3.8 and Tensorflow 2.7.

Data Balancing and Pre-processing: The classification of crash injury severity in this study considered three different injury classes including fatal and severe injury (KA), non-severe and possible injury (BC), and no injury or property damage only (O or PDO) from the traditional KABCO scale. For training and testing purposes, the entire dataset was split into two parts, where 80% of the data was used for training the models, while the remaining 20% data was used for testing. As is commonly the case with most crash data, the data analyzed in this study was highly skewed, in which the number of observations for no injury class is more than 30 times the number of observations in fatal and severe injury class. To address the issue with biased predictions due to imbalanced data, the training data was balanced using random under-sampling of no injury class observations and over-sampling of the data of the remaining classes. This was done by using the Synthetic Minority Oversampling Technique (SMOTE) technique in python.

Classification using Machine Learning Algorithms: This study utilized four different machine learning algorithms, namely, Decision Trees, Random Forest, Extreme Gradient Boosting, and Deep Neural Network for classifying the crash injury severity. Since the data analyzed here has a high degree of imbalance, the confusion matrix for each classifier is considered as a measure of its efficiency. In all cases, the normalized confusion matrices for the models using the 17 most influential variables selected through Granger causality (reduced classifier) are compared with those using all the independent variables (full classifier).

The normalized confusion matrices of the Decision Tree (DT) classifier show that the true positives for both KA (Label 2) and BC (Label 1) classes decrease with reduced classifier, while that for PDO (Label 0) class increases. Additionally, the computational complexity and training time are lower when fewer number of features are considered compared to all predictors. These results also confirm that causal analysis of a dataset not only identifies the most influential variables appropriately, thus providing a deeper insight into the data and the problem at hand, it can also be used to select a subset of independent variables while developing a model which reduces the computational complexity, without compromising much on the predictive performance.

Similar observations can be made from the normalized confusion matrices of the Random Forest (RF) classifier where we trained the data with the number of estimators set to 1,000. As with the DT classifier, for the RF classifier the classification performance for the PDO crashes improves for the reduced classifier, whereas for the BC class, it remains the same and for KA class, the performance of the reduced classifier degrades.

Next, the normalized confusion matrices for the Extreme Gradient Boosting (XGBoost) classifier shows a considerable improvement in prediction performance for KA class compared to DT and RF classifiers. Moreover, for the XGBoost classifier, the predictive performance of the reduced classifier improves substantially for PDO class, while for BC class, it remains almost the same for both XGBoost models.

The final classifier considered in this study was a Deep Neural Network (DNN) with all-to-all connections between the consecutive layers. Four hidden layers were considered with 128 nodes for the first three hidden layers and 64 nodes for the layer. The activation functions used were Rectified Linear Units (ReLU) for the hidden layers and softmax for the output layer. With this, the DNN was trained for 150 epochs with a batch size of 2,048.

Note that the performance of different classifiers varied across the different injury severity classes, a finding consistent with previous studies [13]. When considering DT, RF, and XGBoost, for PDO class, DT performs the best on the test set, followed by similar performances of RF and XGBoost classifiers. For BC class, RF performs the best, followed by XGBoost and then DT classifiers. Among DT, RF, and XGBoost classifiers, the highest true positives for KA class are provided by the XGBoost classifier outperforming both DT and RF classifiers. However, the prediction performance for KA class degrades considerably for the reduced XGBoost classifier, most likely due to the high imbalance in the data. In all the classifier models, the predictive performance for PDO class increases when only the most influential variables are considered. This strongly affirms that the factors which *causes* more severe crashes are different that those causing lower or no injury crashes. Finally, based on the number of true positives, although the DNN classifier compares favorably with DT, RF, and XGBoost classifiers for BC and PDO classes, it outperforms all other classifiers as far as correct prediction for KA class is concerned.

Conclusions

This study presents a methodological framework involving the development of causal inference and injury severity classification for freeway traffic crashes. Granger causality test was used to identify and rank-order the influential features causing the varying injury severity that were further utilized to build four different classifiers, including Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Deep Neural Network (DNN) classifiers to classify the crashes according to their severity. The output of the proposed classification approach includes three severity classes for fatal and severe injury (KA) crashes, non-severe and possible injury (BC) crashes, and property damage only (PDO) crashes.

The most influencing factors identified by Granger causality include speed limit, surface and weather conditions, traffic volume, presence of workzones, workers in workzones, and HOV lanes, among others. Granger causality test provides a systematic procedure for selecting the influential variables and as shown in this study, the efficacy of the Granger causality was demonstrated by achieving comparable results between reduced order and full order models. In terms of the prediction performance of the classifiers, decision tree and random forest classifiers provided the greatest performance for PDO and BC crash severities, respectively. For the KA class, the rarest class in the data, deep neural net classifier performed superior to all other algorithms, most likely due to its capability in approximating nonlinear models.

The study identifies some limitations that can be addressed in future research. Firstly, the factors affecting injury severity often differ based on severity level and this was not considered in this study wherein the influential factors were identified using all severity classes together. In future, identification of important factors separately for different injury levels would be more insightful. Additionally, to improve the prediction performance further, a subsequent analysis can be carried out by using binary classifiers for fatal and injury, and PDO crashes. Nevertheless, overall, this study contributes to the limited body of knowledge pertaining to causal analysis and classification prediction of traffic crash injury severity using non-parametric approaches.

References

1 Chakraborty, M., Gates, T.J. Association between Driveway Land Use and Safety Performance on Rural Highways. Transportation Research Record: Journal of the Transportation Research Board, 2020, 2675, (1), pp. 114–124

2 Chakraborty, M., Stapleton, S.Y., Ghamami, M., Gates, T.J. Safety Effectiveness of All-Electronic Toll Collection systems. Advances in Transportation Studies, 2020, 2, (Special Issue), pp. 127–142

3 Chakraborty, M., Singh, H., Savolainen, P.T., Gates, T.J. Examining Correlation and Trends in Seatbelt Use among Occupants of the Same Vehicle using a Bivariate Probit Model. Transportation Research Record, 2021

4 Chakraborty, M., Mahmud, S., Gates, T. Analysis of Trends and Correlation in Child Restraint Use and Seating Position of Child Passengers in Motor Vehicles: Application of a Bivariate Probit Model. Transportation Research Record: Journal of the Transportation Research Board, 2022

5 Mussone, L., Ferrari, A., Oneta, M. An analysis of urban collisions using an artificial intelligence model. Accident Analysis & Prevention, 1999, 31, (6), pp. 705–718

6 Chakraborty, M., Mahmud, S., Gates, T., Sinha, S. Linear Regularization-based Analysis and Prediction of Human Mobility in the U.S. during the COVID-19 Pandemic. engrXiv, 2020

7 Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z., Wang, Z. Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network. IEEE Access, 2019, 7, pp. 39897–39910

8 Sinha, S., Chakraborty, M. Causal Analysis and Prediction of Human Mobility in the U.S. during the COVID-19 Pandemic. arXiv:2111.12272 [cs, stat], 2021

9 CRIS Query, https://cris.dot.state.tx.us/public/Query/app/public/welcome, accessed April 2019

10 Traffic Count Database System (TCDS), https://txdot.ms2soft.com/tcds/tsearch.asp?loc=Txdot&mod=TCDS, accessed April 2019

11 Elvik, R., Vaa, T., Hoye, A., Sorensen, M. The Handbook of Road Safety Measures. Emerald Group Publishing, 2009

12 Osman, M., Paleti, R., Mishra, S. Analysis of passenger-car crash injury severity in different work zone configurations. Accident Analysis & Prevention, 2018, 111, pp. 161–172

13 Ahmadi, A., Jahangiri, A., Berardi, V., Machiani, S.G. Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. Journal of Transportation Safety & Security, 2020, 12, (4), pp. 522–546