IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 31, 2020, accepted August 17, 2020, date of publication August 20, 2020, date of current version September 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3018183

Comparing Factors Affecting Injury Severity of Passenger Car and Truck Drivers

BEI ZHOU[1,2], XIQING WANG[3], SHENGRUI ZHANG[1,2], ZONGZHI LI[4], SHAOFENG SUN[1], KUN SHU[1], AND QING SUN[1]

[1]College of Transportation Engineering, Chang'an University, Xi'an 710064, China
[2]Key Laboratory of Transport Industry of Management, Control and Cycle Repair Technology for Traffic Network Facilities in Ecological Security Barrier Area, Chang'an University, Xi'an 710064, China
[3]School of Highway, Chang'an University, Xi'an 710064, China
[4]Department of Civil, Architectural, and Environmental Engineering, Illinois Institute of Technology, Chicago, IL 60616, USA

Corresponding author: Bei Zhou (bzhou3@chd.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 71871029, in part by the Fundamental Research Funds for the Central Universities, CHD, under Grant 300102210301 and Grant 300102219306, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2020JM-222, and in part by the 111 Project of Sustainable Transportation for Urban Agglomeration in Western China under Grant B20035.

ABSTRACT This study aims to explore factors affecting passenger car and truck driver injury severity in passenger car-truck crashes. Police-reported crash data from 2007 to 2017 in Canada are collected. Two-vehicle crashes involving one truck and one passenger car are extracted for modeling. Different injury severities are not equally represented. To address the data imbalance issue, this study applies four different data imbalance treatment approaches, including over-sampling, under-sampling, a hybrid method, and a cost-sensitive learning method. To test the performances of different classifiers, five classification models are used, including multinomial logistic regression, Naive Bayes, Classification and Regression Tree, support vector machine, and eXtreme Gradient Boosting (XGBoost). In both the passenger car driver and truck driver injury severity analysis, XGBoost combined with cost-sensitive learning generates the best results in terms of G-mean, area under the curve, and overall accuracy. Additionally, the Shapley Additive Explanations (SHAP) approach is adopted to interpret the result of the best-performing model. Most of the explanatory variables have similar effects on passenger car and truck driver fatality risks. Nevertheless, six variables exhibit opposite effects, including the age of the passenger car driver, crash hour, the passenger car age, road surface condition, weather condition and the truck age. Results of this study could provide some valuable insights for improving truck traffic safety. For instance, properly installing traffic control devices could be an effective way to reduce fatality risks in passenger car-truck crashes. Besides, passenger car drivers should be extremely cautious when driving between midnight to 6 am on truck corridors.

INDEX TERMS Driver injury severity, data imbalance, interpretable machine learning, truck crashes.

I. INTRODUCTION

As the dominant mode of freight transportation in North American, trucking plays an important role in commodity flow and economic vitality. According to the 2017 Commodity Flow Survey conducted by the U.S. Department of Transportation, trucks move 73.0% of all goods by value, 71.5% by weight, and 41.6% by ton-miles [1]. Unfortunately, the considerable volume of truck traffic has also brought some regrettable safety issues. Compared to other types of vehicles, trucks have some unique characteristics, such as heavier gross weight, larger vehicle size, and larger blind spot area, which might increase the risk of severe crashes. According to the National Highway Traffic Safety Administration (NHTSA), there were 4,761 people killed in crashes involving large trucks in 2017, a 12% increase from 2008 [2]. It should be noted that 72% of these fatalities were occupants of other vehicles. Additionally, the involvement rate of large trucks in injury crashes was 31 per 100 million large-truck miles traveled in 2015, a 48% increase from 2008.

Compared to crashes involving other types of vehicles, truck crashes usually result in more severe economic losses and crash severity. As such, a significant amount of research has been conducted to explore factors affecting injury severity

The associate editor coordinating the review of this manuscript and approving it for publication was Rashid Mehmood.

in truck-involved crashes. However, a review of the literature indicates that comparatively few studies have compared factors affecting injury severity of passenger car drivers and truck drivers in truck-involved crashes. Since the crash outcomes of passenger car drivers and truck drivers are significantly different, lacking such information could affect the effectiveness of safety improvement countermeasures.

Besides, although a wide variety of modeling approaches have been adopted to study injury severity of truck-involved crashes, relatively little attention has been paid to the data imbalance issue. A dataset is considered as imbalanced when one class has a much greater number of instances than the other classes [3]. In a typical traffic crash dataset, the number of fatal crashes (minority instances) is considerably outnumbered by non-fatal crashes (majority instances), leading to a data imbalance problem. Without proper treatments, data imbalance could severely undermine the performance of classification models. This is mainly due to standard classifiers (such as logistic regression, decision tree and Naive Bayes) are designed for balanced training data. When the data imbalance is present, these classifiers often provide suboptimal results by classifying majority instances more accurately while misclassifying the minority instances [4]. Another reason is the learning process of standard classifiers is guided by achieving the highest overall accuracy, inducing a bias towards the majority instances [5]. The value of a crash classification model depends largely on its accuracy in predicting more severe crashes, which happen to be minority instances in a crash dataset [6]. As such, to make crash classification models more informative, the data imbalance problem needs to be properly handled.

Another issue worth studying is the model results interpretability. In recent years, various machine learning techniques have been used to study traffic injury severities, such as classification and regression tree [7], support vector machine [8], and gradient boosting model [9]. Compared to traditional safety models, these more sophisticated data-driven models have been shown to predict crash severity with relatively high accuracy [10]. Nevertheless, these models are often considered as ''black-box'' methods as lacking the inference ability. To unravel how specific variable influences model prediction results, the current study adopts a recently proposed approach called SHAP (Shapley Additive Explanations), which is a unified approach for interpreting the output of any machine learning model [11]. Based on coalitional game theory, SHAP is able to explain a prediction by computing the contribution of each variable to the prediction.

The rest of this article is organized as follows: Section 2 reviews relevant literature in related domain; Section 3 introduces the dataset used in this study, and the descriptive statistics of variables are provided; Section 4 describes the proposed methodology in detail; Section 5 discusses the model results; Section 6 concludes the current study and Section 7 points out the study limitations.

II. LITERATURE REVIEW

Since the available traffic crash datasets typically report injury levels as discrete variables, many previous studies on injury severities of truck-involved crashes have adopted discrete outcome regression models. In an early study, Khattak and Targa applied the ordered probit model to examine factors affecting truck-involved crash severities in work zones [12]. Based on a unique dataset collected from North Carolina, the effects of various variables were tested. The results suggested that following variables significantly affected the severities of multivehicle crashes involving trucks: the roadway configuration, posted speed limits, adjacent to the work zone, and whether a bypass was required on the opposite side. To study the impact of vehicle, driver, occupant and environmental attributes on injury levels of crashes involving heavy-duty trucks, Lemp *et al.* established an ordered probit model based on datasets consolidated from various data sources [13]. Results suggested that increasing the number of trailers could increase the likelihood of more severe crashes. Chu used a binary logit model to study factors contributing to severities of crashes involving gravel trucks [14]. This study found that lacking driver awareness, geometric improvement of roadways, and the desire to make more runs in a day significantly increased the likelihood of severe injury crashes. Choi *et al.* applied a binary logistic regression model to identify factors affecting truck-involved crash severities under normal and adverse weather conditions [15]. Based on the model results, speed-related variables were identified as the most important factors affecting crash severities. To explore factors affecting the frequency and severity of large-truck involved crashes, Dong *et al.* [16] proposed multinomial logit and negative binomial models. This study concluded that truck percentage, annual average daily traffic, weather condition and driver condition significantly affected both the severity and frequency of crashes involving large trucks. Using 2009-2013 crash data in Ohio, Uddin and Huynh [17] developed six mixed logit models considering three lighting conditions and two area types to investigate factors affecting injury severities of truck-involved crashes. Results revealed the impacts of variables on injury severity were quite different in different models, highlighting the necessity of investigating crashes based on different lighting conditions and area types. Newnam *et al.* [18] studied a unique safety issue: whether older truck drivers give rise to an increased safety risk. Chi-square statistics were used to explore differences in injury levels in middle-aged and older driver groups. Based on the results of this study, compared with middle-aged drivers, older drivers presented some safer driving behaviors. Moomen *et al.* [19] utilized a logistic regression model to analyze factors affecting truck crashes on Wyoming downgrades. Several countermeasures were identified to prevent such crashes. Osman *et al.* [20] analyzed injury severity of large truck crashes in work zones by using a generalized ordered response logit model. It was concluded that following factors had higher elasticity: lower AADT, higher speed

limits, and daytime. Useche *et al.* [21] examined the effect of various factors associated with serious injuries and fatalities among Spanish professional drivers. Results of the study indicated that the type of road and crash, light and vehicle conditions, along with individual driver's characteristics are significant factors for predicting serious injuries and fatalities of professional drivers. Unlike previous literature focusing on truck crashes in developed countries, Wang and Prato [22] analyzed injury severities of truck crashes on mountainous expressways in China. A total of 2,695 truck crashes occurring on four mountain expressways were analyzed with a partial proportional odds model. This study focused on the geometric characteristics of expressways and proposed several road design suggestions to alleviate truck crash severity. Rahimi *et al.* [23] studied the injury severity of single-vehicle truck crashes in Iran. A random thresholds random parameter hierarchical ordered probit model was used to consider the heterogeneity across crashes. Several safety countermeasures were also proposed. A recent study conducted by Behnood and Mannering [24] studied the temporal instability of factors affecting injury severities of truck-involved crashes. Based on the results of random parameters logit models, this study found that the effects of factors influencing injury severities in truck crashes were unstable from year to year and across daily time periods. Behnood and AI-Bdairi [25] analyzed the weekly instability of factors affecting injury severities in large truck crashes. It was revealed that model estimation results were not transferable across weekends and weekdays. Haq *et al.* [26] investigated occupant injury severity of truck-related crashes based on vehicle types. It was found that sperate models should be used for each occupant of each vehicle type. Besides, the actions of drivers had more significant impacts on crash severity. Although most previous studies have applied regression models to study injury severity of truck crashes, the application of non-parametric machine learning techniques has also attracted some attention. For instance, Chang and Chien [7] developed a classification and regression tree (CART) model to uncover the relationship between truck crash severities and various driver, roadway, environment and crash characteristics. The results revealed that the following variables were the key determinants of truck crashes severities: seatbelt use, crash type, vehicle type, driver action, crash location and number of vehicles involved in the crash. In another study, a more advanced gradient boosting model was developed to analyze commercial truck crash severities [9]. The model revealed that 22 variables significantly contributed to injury severities and 11 of them could explain more than 80% of the model forecasting.

Regarding the data imbalance embedded in traffic crash datasets, several studies have proposed corresponding treatments. To analyze factors affecting crash severity in Jordan, Mujalli *et al.* [27] used three different resampling techniques to address the data imbalance issue. It was found that using the balanced data set to train the classifier could improve the classification accuracy of killed and severe injuries crashes. In another study, Goh *et al.* [28] applied logistic regression and six popular machine learning algorithms to uncover the relationship between different cognitive factors and unsafe behaviors. Since the unsafe behaviors are highly imbalanced, this study used an over-sampling technique to rebalance the training data. It was concluded that the decision tree algorithm achieved the best classification performance when training on the rebalanced dataset. Jeong *et al.* [6] proposed a hybrid approach for imbalanced traffic crash data analysis. They used two resampling techniques and five classification algorithms to classify injury severities in motor vehicle crashes. It was revealed that the best classification performance was achieved when Bootstrap aggregation was used with the decision tree, with over-sampling technique to treat data imbalance.

The current study proposes a threefold contribution to existing literature. Firstly, by comparing factors affecting injury severity of passenger car drivers and truck drivers in truck-involved crashes, this article provides some valuable insights for stakeholders to alleviate crash severity. Secondly, by implementing several algorithms to deal with imbalanced crash datasets, the current study significantly improves the classification accuracy of more severe crashes. Thirdly, to the best of our knowledge, this is the first study to apply SHAP to improve the interpretability of traffic crash classification models.

III. DATA PREPARATION

The data used in the current study is extracted from Canada National Collison Database (NCDB). NCDB contains all police-reported vehicle crashes on public roads in Canada since 1999 [29]. For the modeling purpose, this study collects the 2007-2017 two-vehicle crashes involving one truck and one passenger car. According to NCDB, a truck is defined as a heavy vehicle with GVWR (Gross Vehicle Weight Rating) of more than 4,536 kg. From 2007 to 2017, there are 28,605 two-vehicle crashes involving one truck and one passenger car. Among these crashes, 1,274 passenger car drivers suffer from fatal injuries, accounting for 4.45% of total crashes. Contrastingly, 27 truck drivers are killed in these crashes, accounting for 0.09% of total crashes. Additionally, 80.46% of passenger car drivers are injured and only 15.36% of truck drivers suffer from injuries in these crashes

In the context of the current study, 16 variables are selected for the modeling purpose, including crash characteristics (e.g., crash month and day), infrastructure characteristics (e.g., roadway alignment and surface condition), vehicle characteristics (e.g., age of the passenger car and truck), and driver characteristics (e.g., driver's gender and age). Please refer to Table 1 for the detailed variable description and distribution.

IV. METHODOLOGY

A. TREATING DATA IMBALANCE

As shown in Table 1, the injury levels of passenger car drivers and truck drivers are both highly imbalanced. Without proper

TABLE 1. Variable summary.

Variable	Description	No. of Crashes	Distribution
Crash severity			
P_ISEV_p	Injury severity of passenger car drivers		
	No Injury = 1	4315	15.08%
	Injury = 2	23016	80.46%
	Fatality = 3	1274	4.45%
P_ISEV_t	Injury severity of truck drivers		
	No Injury = 1	24185	84.55%
	Injury = 2	4393	15.36%
	Fatality = 3	27	0.09%
Crash-related			
C_MNTH	Crash month		
	Spring: March to May = 1	5857	20.48%
	Summer: June to August = 2	7020	24.54%
	Autumn: September to November = 3	7634	26.69%
	Winter: December to February = 4	8094	28.30%
C_WDAY	Crash day		
	Weekday = 1	24918	87.11%
	Weekend = 2	3687	12.89%
C_HOUR	Crash hour		
	0:00~6:59 = 1	2680	9.37%
	7:00~9:59 = 2	5437	19.01%
	10:00~15:59 = 3	13028	45.54%
	16:00~19:59 = 4	5525	19.31%
	19:00~23:59 = 5	1935	6.76%
C_CONF	Crash configuration		
	Same Direction of Travel = 1	15619	54.60%
	Different Direction of Travel = 2	12986	45.40%
Infrastructure-related			
C_RCFG	Roadway configuration		
	Non-intersection = 1	14614	51.09%
	Intersection = 2	13991	48.91%
C_RSUR	Road surface		
	Dry, normal = 1	20688	72.32%
	Wet = 2	4361	15.25%
	Snow = 3	1530	5.35%
	Slush, wet snow = 4	439	1.53%
	Ice = 5	1587	5.55%
C_RALN	Road alignment		
	Straight and level = 1	22965	80.28%
	Straight with gradient = 2	2584	9.03%
	Curved and level = 3	1682	5.88%
	Curved with gradient = 4	836	2.92%
	Top of hill or gradient = 5	302	1.06%
	Bottom of hill or gradient = 6	236	0.83%
C_TRAF	Traffic control		
	Without traffic control devices = 1	16576	57.95%
	With traffic control devices = 2	12029	42.05%
Environment-related			
C_WTHR	Weather condition		
	Clear and sunny = 1	21616	75.57%
	Cloudy = 2	2137	7.47%
	Raining = 3	2137	7.47%
	Snow = 4	2715	9.49%
Vehicle-related			
V_YEAR_p	Passenger car age		
	1~5 = 1	7326	25.61%
	6~10 = 2	9756	34.11%
	>=10 = 3	11523	40.28%
V_YEAR_t	Truck age		
	1~5 = 1	9672	33.81%
	6~10 = 2	6518	22.79%
	>=10 = 3	12415	43.40%
Passenger car driver-related			
P_SEX_p	Passenger car driver's gender		
	Male = 1	15861	55.45%
	Female = 2	12744	44.55%
P_AGE_p	Passenger car driver's age		
	<=20 = 1	2636	9.22%
	21~30 = 2	6327	22.12%
	31~40 = 3	5344	18.68%
	41~50 = 4	5233	18.29%
	51~60 = 5	4313	15.08%
	>=61 = 6	4752	16.61%
P_SAFE_p	Passenger car safety device usage		
	Not wearing safety belt = 1	725	2.53%
	Wearing safety belt = 2	27880	97.47%
Truck driver-related			
P_SEX_t	Truck driver's gender		
	Male = 1	27467	96.02%
	Female = 2	1138	3.98%
P_AGE_t	Truck driver's age		
	<=20 = 1	589	2.06%
	21~30 = 2	4714	16.48%
	31~40 = 3	6572	22.98%
	41~50 = 4	7572	26.47%
	51~60 = 5	6540	22.86%
	>=61 = 6	2618	9.15%
P_SAFE_t	Truck safety device usage		
	Not wearing safety belt = 1	477	1.67%
	Wearing safety belt = 2	28128	98.33%

treatment, this could severely compromise the performance of the classifier. In the past few years, hundreds of algorithms have been proposed to address the data imbalance issue. Basically, these techniques can be divided into two groups: resampling and cost-sensitive learning [3]. To compare the performance of different data imbalance treatment approaches, the current study adopts three resampling algorithms (over-sampling, under-sampling, and a hybrid method combining under-sampling and over-sampling) and one cost-sensitive learning method.

1) OVER-SAMPLING

Over-sampling aims to eliminate the adverse impact of skewed class distribution by creating synthetic minority instances. A popular over-sampling technique called SMOTE (Synthetic Minority Over-sampling Technique) is used in the current study. Originally proposed by Chawla *et al.* [30], SMOTE is widely used in previous imbalanced learning studies [3], [27], [28], [31]. SMOTE aims to create a more balanced dataset by randomly generating artificial minority samples along the line segments joining each minority sample with its k nearest neighbors (in our case, $k = 5$). Depending

on the amount of over-sampling instances required, neighbors from the k nearest neighbors are randomly chosen and one synthetic sample is created in each direction. This is done as follows. Firstly, SMOTE measures the difference between the feature vector (an n-dimensional vector representing the sample) under consideration and its nearest neighbor. Secondly, this measured difference is multiplied by a random number between 0 and 1, which is then added to the feature vector. This forces the selection of a random point and creates an artificial instance along the line segment joining two feature vectors.

2) UNDER-SAMPLING

Unlike over-sampling, under-sampling tries to create better-defined class clusters by removing samples according to a specific selection criterion. The current study applies the Edited Nearest Neighbor (ENN) method to perform under-sampling [32]. For each sample in the dataset, its three nearest neighbors are located. If this sample pertains to the minority class, and at least two of its three nearest neighbors belong to the majority class, then this sample is eliminated. Likewise, if this sample belongs to the majority class, and at least two of its three nearest neighbors pertain to the minority class, then this sample is also deleted. Through this in-depth data cleaning, the ENN method could generate a more balanced class distribution.

3) THE HYBRID METHOD

The hybrid method combines the over- and under-sampling method. Although the SMOTE method could generate synthetic samples by interpolating new points between existing feature vectors, it can also bring on other problems. As shown in Figure 1(b), the interpolation of minority samples could generate artificial samples too deeply in the majority class cluster. To this end, the classification algorithm might be overfitted and less informative. The hybrid method tries to solve this problem by applying the SMOTE and ENN methods in sequence. The SMOTE method is firstly applied to generate artificial minority instances, resulting in a more balanced dataset. Then, the ENN method is called for the data cleaning purpose. This would create better-defined class clusters.

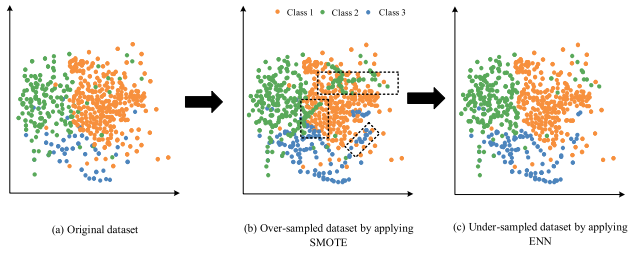

FIGURE 1. Illustration of the hybrid method for treating data imbalance.

In the current study, the SMOTE method, ENN method and hybrid method are all coded in Python based on the imbalanced-learn library [33].

4) COST-SENSITIVE LEARNING METHOD

In addition to the aforementioned three resampling methods, this study also tests a cost-sensitive learning method. Cost-sensitive learning assumes a higher cost for misclassifying minority instances with respect to majority instances. To this end, a weight is calculated for each sample in the dataset according to the class frequency of this sample. For a sample S_i belonging to class i, the weight is calculated as:

$$Weight_{Si} = \frac{the\ total\ number\ of\ samples}{number\ of\ classes \times number\ of\ class\ i\ samples} \tag{1}$$

Obviously, minority instances have higher weights relative to majority instances. And this would force the classifier to put more emphasis on correctly classifying minority instances. Unlike data resampling methods which are incorporated at the data level, the cost-sensitive learning method is incorporated at the algorithmic level by modifying the loss function. Compared to data resampling methods, the cost-sensitive learning method is more computationally efficient, which makes it more suitable for large-size datasets. The current paper adopts the scikit-learn Python library to computer each sample's weight [34].

B. CLASSIFICATION MODELS

To compare factors contributing to crash severity of passenger car and truck drivers, the crash outcomes of both drivers are modelled separately. To examine the performances of different classifiers on predicting crash severity, this study uses five classification models, including multinomial logistic regression, Naive Bayes, Classification and Regression Tree (CART), support vector machine, and eXtreme Gradient Boosting (XGBoost). This section briefly elaborates on each model, as well as the classification performance evaluation metrics.

1) MULTINOMIAL LOGISTIC REGRESSION

In the current study, the crash severity is divided into three categories: no injury, injury, and fatality. As a traditional unordered discrete outcome model, the multinomial logistic regression (MNL) model is suitable for exploring the potential relationship between contributing factors and three or more injury outcomes. Besides, the MNL model does not impost sometimes unrealistic restrictions on parameters, such as normality or homoscedasticity, which makes it a popular choice in crash severity analysis. The MNL works by selecting one injury outcome as the base category and the other injury outcomes are estimated relative to this base category. A standard MNL model is expressed as [35]:

$$P_n(i) = \frac{EXP(\beta_i X_{in})}{\sum \forall_I EXP(\beta_I X_{In})} \tag{2}$$

where β_i is the estimated coefficients for the injury outcome i, and X_{in} stands for independent variables which impact the injury outcome i sustained by crash n. I represents a set of possible injury outcomes.

Based on the results of the MNL model, the impact of each variable on the injury outcome can be easily interpreted by the estimated coefficient or the odds ratio (exponent of the coefficient). Nevertheless, the MNL model does require careful consideration of the correlation between each crash contributing factor and the crash outcome, as well as the possible multicollinearity among contributing factors.

2) NAIVE BAYES

A Naive Bayes (NB) classifier is a popular supervised learning algorithm based on the Bayes theorem. It is called ''Naive'' because a NB classifier has a strong assumption of conditional independence between each pair of explanatory variables, given the class variable value. In other words, a NB classifier assumes that each explanatory variable contributes independently and equally to the class variable. Based on the Bayes theorem, the probability of class variable $Y = y$ given that the explanatory variable $X = (x_1, x_2, \ldots, x_n)$ can be describes as:

$$P(Y = y|X = (x_1, \ldots x_n)) = \frac{P(Y = y)\prod_{i=1}^{n} P(x_i|Y = y)}{P(X = (x_1, \ldots x_n))} \quad (3)$$

Then, the Maximum A Posteriori (MAP) probability could be used to estimate $P(Y = y)$ and $P(x_i|Y = y)$. It should be noted that NB classifiers are a set of classification algorithms but not a single classifier. Different NB classifiers are mainly different due to the assumptions regarding the distribution of $P(x_i|Y = y)$. The current paper adopts a Gaussian NB algorithm which assumes that $P(x_i|Y = y)$ follows the Gaussian distribution.

3) CART MODEL

The Classification and Regression Tree (CART) model is one of the most popular machine learning models, which has been widely used in traffic safety analysis [7], [31], [36], [37]. Compared with most regression models, the CART model does not impose any predefined relationship between explanatory variables and the class variable. As indicated by the model name, the CART model could handle both classification and regression tasks depending on the nature of the target variable. In the current study, the target variable is the injury severity of drivers, which is a discrete variable. Hence, a classification tree is developed. The CART modeling procedure includes two major steps: tree growing and tree pruning. Starting at the root node, tree growing aims to recursively partition the class variable to minimize the impurity of two child nodes. To this end, during each step, the CART model needs to select an explanatory variable as the splitter which can improve the purity of two child nodes most significantly. There are several indicators to measure the purity improvement, of which the Gini index is most commonly used. And this study selects the Gini index to measure the impurity of any child node. The tree keeps growing by recursively partitioning the class variable based on the Gini index. At some point, all samples within each child node belong to the same class and a saturated tree is generated. This saturated tree is most probably overfitting and could lead to high misclassification rate when classifying a new dataset. As such, this saturated tree should be pruned by adjusting parameters which control the tree growing, such as the maximum depth of the tree, the maximum number of leaf nodes in a tree, and the minimum number of samples required to be at a leaf node.

4) SUPPORT VECTOR MACHINE

The support vector machine (SVM) model is a widely used non-parametric machine learning model of the recent years, mostly because of its sound theoretical foundation and superior predictive performance. Originally proposed by Cortes and Vapnik [38], the SVM model is based on the structural risk minimization principle and the statistical learning theory. Similar to the CART model, the SVM model can also handle both classification and regression problems. For classification problems, the SVM model can map the input vector into a high dimensional feature space. Generally speaking, many hyperplanes can separate the data into different groups in the feature space. The purpose of the SVM model is to construct an optimal hyperplane which can maximize the margin between these groups. The optimal hyperplane is known as the maximum-margin hyperplane, and it can be represented by quadratic optimization modeling.

Although the SVM model was originally designed for two-category classification problems, it can be extended for dealing with multi-category classification problems after some modifications. In the current paper, the prevailing one-versus-one approach is used. For a classification problem with N classes, the one-versus-one approach trains $N(N-1)/2$ binary SVM models for all possible pairs of classes. Each binary model may predict one class label and the label with the most predictions or votes is determined as the severity level of the crash.

5) XGBoost MODEL

XGBoost stands for eXtreme Gradient Boosting. Originally proposed by Chen and Guestrin [39], XGBoost has been widely used in various machine learning competitions to achieve state-of-the-art results. XGBoost is a scalable tree-boosting system with the purpose of achieving extreme execution speed and model performance. As an advanced implementation of gradient boosting machines, XGBoost is also an ensemble tree method that aims to create a strong classifier based on a series of weak learners. The most commonly used weak learners are CARTs. A single CART might fail to incorporate predictive power from multiple feature space regions, which is why it is called a weak learner. In contrast, by iteratively training a set of weak classifiers, ensemble methods have been proven to be much more accurate than a single classifier [40]. The objective function of XGBoost consists of training loss and regularization term, which can

be written as:

$$obj(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega(f_k) + \text{constant} \quad (4)$$

where θ stands for model parameters which need to be learned from the training data; $\hat{y}_i$ is model prediction for the i_{th} data sample; y_i is the actual label of the i_{th} data sample; l is the loss function of the i_{th} data sample, measuring how well the model can fit the training data; $\Omega(f_k)$ stands for the regularization term, which is used to control the complexity of the model and avoid overfitting; f_k is a scoring function to estimate the output in the k_{th} tree, and t is the total number of trees.

Training all CARTs at once is very difficult. Instead, XGBoost adopts an additive training strategy. At training step t, the model prediction $\hat{y}_i^t$ is the summation of the prediction at step $t-1$ and the score of a new tree, which can be written as:

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \quad (5)$$

To this end, the objective function at step t is:

$$\begin{aligned} obj^t(\theta) &= \sum_{i}^{n} l(y_i, \hat{y}_i^t) + \sum_{k=1}^{t} \Omega(f_k) + \text{constant} \\ &= \sum_{i}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + \text{constant} \quad (6) \end{aligned}$$

The model parameter θ is updated at each step t according to the new objective function. The loss function in XGBoost can take various forms, such as mean squared error or logistic loss. Besides, XGBoost supports custom loss functions. The regularization term is a major contribution of XGBoost, which is given as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (7)$$

where γ stands for the complexity parameter of each leaf; T is the total number of leaves; λ is used to scale the penalty; w is the vector of scores on leaves.

6) CLASSIFICATION PERFORMANCE EVALUATION

Probably the most intuitionistic metric in evaluating classification model performance is the overall accuracy, which can be derived from the confusion matrix in Table 2.

TABLE 2. Confusion matrix for 2-class classification.

	Predicted positive	Predicted negative
Actual positive	True positive (TP)	False negative (FN)
Actual negative	False positive (FP)	True negative (TN)

The overall accuracy is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (8)$$

In general, the overall accuracy can be used to evaluate how accurately the classification model can predict the testing data. However, when the dataset is imbalanced, relying merely on the overall accuracy might produce biased evaluation. For instance, when a model tries to classify a dataset with 95 negative instances and 5 positive instances, it can easily achieve a 95% accuracy by classifying all instances as negative. Apparently, this high accuracy is doubtable, and the corresponding classification model might fail to be informative. In this article, geometric mean (G-mean) is selected to evaluate the classification performance of the proposed models together with the overall accuracy. As a widely used metric in imbalanced learning field [3], [6], [41], G-mean aims to maximize the accuracy of each class while keeping these accuracies balanced. For a n-class classification problem, G-mean is calculated as:

$$G - mean = \sqrt[n]{class\,1\,accuracy \times class\,2\,accuracy \times \cdots \times class\ n\ accuracy} \quad (9)$$

As shown in Equation (9), G-mean is not affected by the number of instances within each class. In addition, the area under the curve (AUC) is also calculated to further compare the classification performance of each modeling scenario.

C. MODEL RESULTS INTERPRETATION

The purpose of developing a crash severity analysis model is to uncover the relationship between various features and crash outcomes. Subsequently, corresponding administrative and engineering countermeasures could be implemented to alleviate the crash severity. As such, the interpretability of the model output is as important as its accuracy. The output of a linear model, such as MNL, is straightforward and easy to understand: the parameter value of each feature could be used to measure the impact of this feature on the model outcome. However, such models are only able to uncover linear relationships. On the other hand, more sophisticated machine learning models, such as XGBoost or random forest, are able to uncover more complicated relationships and predict crash severities with relatively high accuracy [10]. Nevertheless, these models could be difficult to interpret. A common approach to explain these models' results is to calculate the importance of features based on gain or split counts. But this approach could suffer from inconsistency, i.e., the order in which a feature is added to the model could significantly affect the importance of this feature [11], [42].

This study adopts a novel approach called SHAP (Shapley Additive Explanations) to explain the output of machine learning models. Originally proposed by Lundberg and Lee [11], SHAP is designed to explain the output of any machine learning model in a consistent and accurate way based on game theory and local explanation. Generally speaking, SHAP measures the importance of a feature by comparing model predictions with and without this particular

feature. Unlike other feature attribution approaches, SHAP is able to compute the exact SHAP value of each feature for each individual instance. As an additive feature attribution method, SHAP develops a linear explanation model g for each instance within the dataset:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \tag{10}$$

where g is the explanation model used to explain the model prediction on an instance; M is the number of features in the model; ϕ_i is the SHAP value for a feature i; $z_i' = 1$ if a feature i is present and $z_i' = 0$ otherwise. The SHAP value for a feature i is calculated by comparing the model predictions with and without this feature. Since the order in which features are added to the model could affect the model prediction, all possible orders are permuted, and the SHAP value is calculated as a weighted summation. This can be described in the following equation:

$$\phi_i = \sum_{S \subseteq M/i} \frac{|S|!(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{11}$$

where S is the subset of features used in the model; M is the number of features; $f_x(S \cup \{i\})$ and $f_x(S)$ are the model predictions with and without feature i. In this way, the individual prediction in the model could be accurately explained. The second term in Equation (11) indicates that the ϕ_i could be negative, meaning feature i could have a negative impact on the model output. For a classification problem, a SHAP value matrix with the same size of the input data could be obtained for each possible model output. The average of the absolute SHAP value of feature i is used to measure the impact of this feature on the model output. The current study uses Python library SHAP to calculate SHAP values [11].

V. RESULTS AND DISCUSSION

The dataset is randomly divided into training and testing subsets according to a 7:3 ratio. The passenger car driver models and truck driver models are separated developed. The classification performances are reported in Table 3 and Table 4, respectively. These tables include overall accuracy, per-class accuracy, G-mean and AUC for each classification model and each data imbalance treatment approach. As shown in Table 3, for the passenger car driver crash severity analysis, the highest G-mean is achieved when XGBoost is used with the cost-sensitive learning approach (G-mean = 58.23%). And the associated overall accuracy is 60.37%. The second-best result of G-mean is achieved when XGBoost is combined with the hybrid data preprocessing approach (G-mean = 56.82%). Nevertheless, the overall accuracy in this scenario is significantly lower (overall accuracy = 46.94%). Besides, XGBoost combined with cost-sensitive learning achieves the highest AUC (0.72) as well. As for the truck driver crash severity analysis, results in Table 4 reveal that the highest G-mean is also achieved when XGBoost is combined with the cost-sensitive learning approach (G-mean = 55.55%). And the associated overall accuracy is 63.57%. Although the AUC of this modeling scenario is lower than that of XGBoost model trained on the imbalanced dataset, the per-class accuracy and G-mean are greatly improved. The second-best G-mean is achieved when the decision tree model is combined with the hybrid data preprocessing approach (G-mean = 51.84%).

Then, SHAP is adopted to explain the results of the best-performing modeling scenarios. Figure 2 and Figure 3 present the impact of factors on passenger car driver and truck driver crash severity, respectively. The factors are sorted in descending order based on the average of the absolute SHAP values. It's worth mentioning that units of SHAP values depend on the selected classification model. For XGBoost, SHAP values have log odds units.

As shown in Figure 2, the crash configuration is the strongest predictor for passenger car driver injury severity. Besides, the gender of the passenger car driver, the traffic control device, roadway configuration and the age of the passenger car driver also have significant impacts on crash outcomes. On the other hand, weather condition, road surface condition and the truck age have the least impact on passenger car driver crash outcomes.

Turning to factors affecting crash severity of truck drivers (Figure 3), crash configuration is also the strongest predictor, followed by the gender of the passenger car driver, roadway configuration, the age of the truck driver, and the traffic control device. Meanwhile, the passenger car age, the gender of the truck driver and road alignment have the least impact on truck driver crash outcomes.

Although feature importance figures are useful, they contain no information beyond average impacts of features on model output magnitude. For more informative explanations, the contribution of each feature on a specific crash outcome should be illustrated. Figure 4 and Figure 5 presents the impact of features on passenger car driver and truck driver fatal crashes, respectively. In SHAP, these figures are called summary plots, which combine feature importance with feature effects. The features (y-axis) are sorted in descending order according to their global impact on the model output (in this case, fatal crashes). Each point on the summary plot represents a SHAP value (x-axis) for a feature in a crash. The color represents the value of a feature. Overlapping points are plied up to show density. Again, for XGBoost, SHAP values on the x-axis have units of log odds.

As shown in Figure 4, the crash configuration is the single most important predictor for passenger car driver fatalities. In crashes involving vehicles travelling in the same direction, passenger car drivers are less likely to suffer from fatalities. Other crash configuration (such as head-on crashes, left/right turn crashes, and different direction sideswipe crashes) would increase the fatality risk of the passenger car driver. The traffic control device is the second most important predictor. Crashes occurred in places with some kind of traffic

TABLE 3. Classification performance by data imbalance treatment approaches for the passenger car driver crash severity analysis.

Models	Evaluation metric[a]	No[b]	Data preprocessing approach			Cost-sensitive learning
			US[c]	OS[d]	Hybrid	
XGBoost	Overall accuracy (%)	79.97	79.85	55.85	46.94	*60.37*
	class 1 accuracy (%)	0.37	0	45.92	54.01	47.35
	class 2 accuracy (%)	99.91	99.15	57.26	43.9	62.6
	class 3 accuracy (%)	0.27	12.94	65.77	77.36	66.6
	G-mean (%)	2.15	0	55.71	56.82	*58.23*
	AUC	0.72	0.64	0.68	0.69	*0.72*
Decision tree	Overall accuracy (%)	79.97	79.14	54.84	46.05	48.22
	class 1 accuracy (%)	0	0	40.28	51.93	48.97
	class 2 accuracy (%)	100	98.05	57.39	43.38	46.53
	class 3 accuracy (%)	0	16.98	60.38	74.12	77.66
	G-mean (%)	0	0	51.87	55.07	56.14
	AUC	0.68	0.64	0.68	0.68	0.68
Naive Bayes	Overall accuracy (%)	79.97	79.97	57.71	48.69	44.85
	class 1 accuracy (%)	0	0	42.06	50.89	74.37
	class 2 accuracy (%)	100	100	61.33	47.85	39.44
	class 3 accuracy (%)	0	0	47.71	56.33	37.79
	G-mean (%)	0	0	49.74	51.57	48.04
	AUC	0.68	0.67	0.64	0.66	0.68
Multinomial logistic regression	Overall accuracy (%)	79.89	79.75	53.96	44.17	43.86
	class 1 accuracy (%)	0.52	0	44.66	53.56	57.27
	class 2 accuracy (%)	99.46	99.1	55.14	40.74	39.62
	class 3 accuracy (%)	6.2	11.59	66.04	73.58	74.95
	G-mean (%)	6.84	0	54.58	54.35	55.4
	AUC	0.71	0.66	0.68	0.67	0.70
Support vector machine	Overall accuracy (%)	79.97	79.97	56.87	48.14	47.59
	class 1 accuracy (%)	0	0	42.06	51.36	53.26
	class 2 accuracy (%)	100	100	60.32	46.97	45.38
	class 3 accuracy (%)	0	0	46.56	58.46	68.68
	G-mean (%)	0	0	49.06	52.05	54.96
	AUC	0.68	0.67	0.66	0.66	0.69

[a] class 1 = No injury crash, class 2 = Injury crash, class 3 = Fatality. [b] No = No treatment. [c] US = Under-sampling. [d] OS = Over-sampling

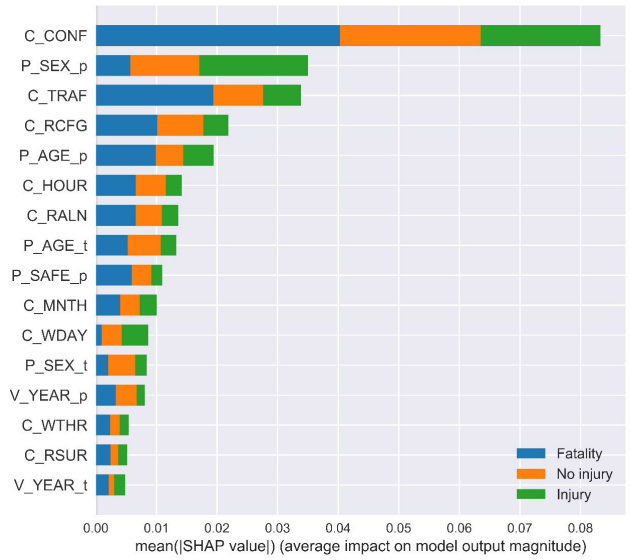

FIGURE 2. Importance of features on passenger car driver crash severity.

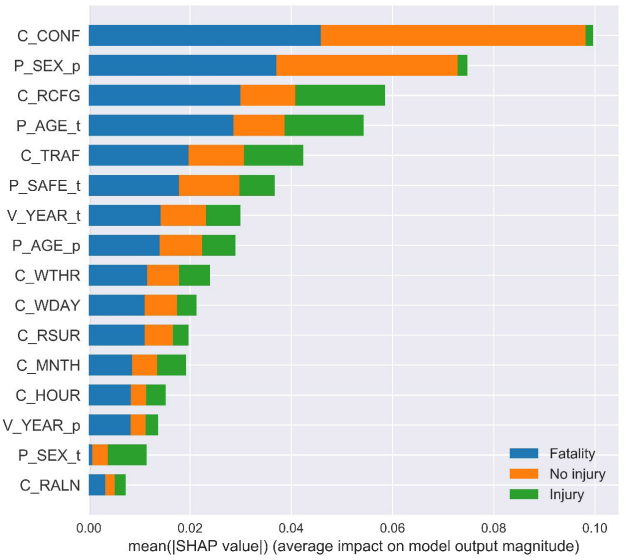

FIGURE 3. Importance of features on truck driver crash severity.

control devices (such as traffic signals, stop signs, and warning signs) are less likely to result in passenger car driver fatalities. Properly installing traffic control devices could be an economical and effective method to reduce fatality risks

TABLE 4. Classification performance by data imbalance treatment approaches for the truck driver crash severity analysis.

Models	Evaluation metric (%)[a]	No[b]	Data preprocessing approach			Cost-sensitive learning
			US[c]	OS[d]	Hybrid	
XGBoost	Overall accuracy (%)	84.75	84.63	61.16	50.03	*63.57*
	class 1 accuracy (%)	99.79	100	64.64	48.67	67.75
	class 2 accuracy (%)	1.94	0	42.03	57.62	40.49
	class 3 accuracy (%)	0	0	37.5	37.5	62.5
	G-mean (%)	0	0	46.71	47.2	*55.55*
	AUC	0.77	0.59	0.66	0.65	*0.70*
Decision tree	Overall accuracy (%)	84.63	84.63	48.21	48.77	49.87
	class 1 accuracy (%)	100	100	46.19	46.82	49.57
	class 2 accuracy (%)	0	0	59.34	59.51	51.51
	class 3 accuracy (%)	0	0	50	50	50
	G-mean (%)	0	0	51.56	51.84	50.34
	AUC	0.62	0.56	0.66	0.67	0.62
Naive Bayes	Overall accuracy (%)	84.63	84.63	63.25	55.65	65.06
	class 1 accuracy (%)	100	100	66.58	55.33	74.84
	class 2 accuracy (%)	0	0	45.06	57.57	10.91
	class 3 accuracy (%)	0	0	12.5	12.5	75
	G-mean (%)	0	0	33.47	34.15	39.41
	AUC	0.68	0.60	0.62	0.62	0.68
Multinomial logistic regression	Overall accuracy (%)	84.83	84.63	61.5	51.71	61.01
	class 1 accuracy (%)	99.49	100	65.41	51.36	64.92
	class 2 accuracy (%)	4.11	0	39.92	53.68	39.46
	class 3 accuracy (%)	0	0	50	50	50
	G-mean (%)	0	0	50.73	51.66	50.41
	AUC	0.70	0.69	0.69	0.69	0.69
Support vector machine	Overall accuracy (%)	84.63	84.63	63.27	55.72	62.11
	class 1 accuracy (%)	100	100	66.53	55.42	62.66
	class 2 accuracy (%)	0	0	45.46	57.57	59.34
	class 3 accuracy (%)	0	0	12.50	12.50	12.50
	G-mean (%)	0	0	33.56	34.164	36.95
	AUC	0.68	0.66	0.63	0.60	0.70

[a] class 1 = No injury crash, class 2 = Injury crash, class 3 = Fatality. [b] No = No treatment. [c] US = Under-sampling. [d] OS = Over-sampling

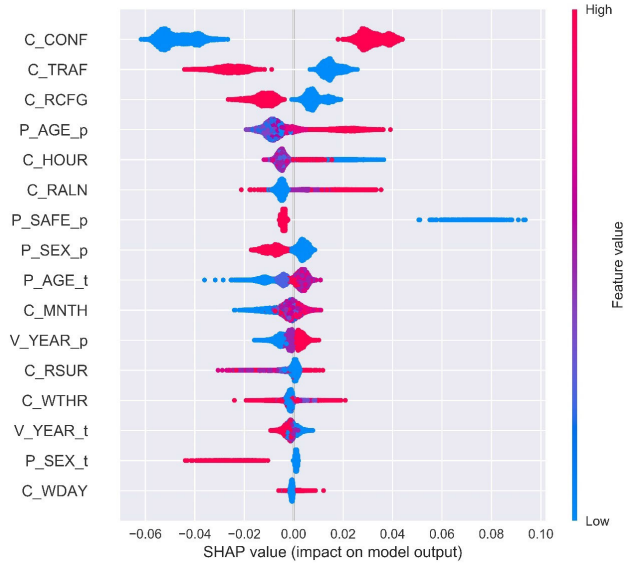

FIGURE 4. Impact of features on passenger car driver fatalities.

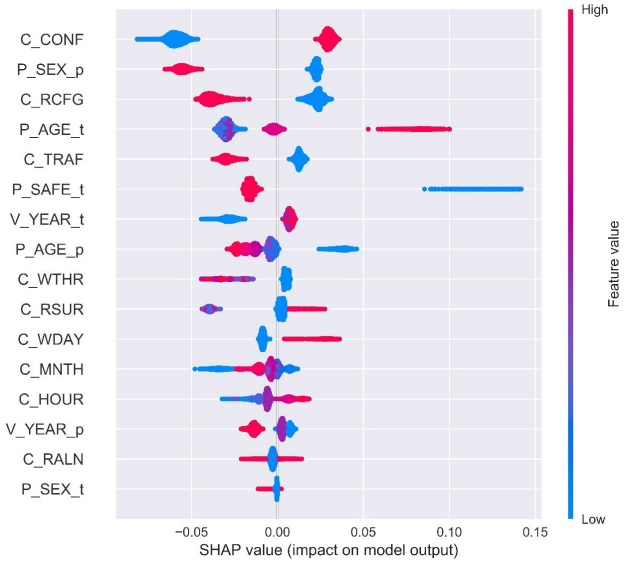

FIGURE 5. Impact of features on truck driver fatalities.

in passenger car-truck crashes. In general, older passenger car drivers are more likely to suffer from fatalities. Besides, compared to female drivers, male passenger car drivers are more prone to fatal crashes. As such, traffic safety education

campaign targeted at older male passenger car drivers should be promoted. It's worth noting that passenger car safety device usage is not the most important feature, but driving without safety belts could significantly increase the fatality risk, given the long tail in the plot. Although the gender of the truck driver is one of the least influential factors, it seems that the presence of female truck drivers could reduce the fatality risk of the passenger car driver.

Compared with Figure 4, most features in Figure 5 have similar effects on truck driver fatalities. For instance, truck drivers are also more likely to suffer from fatalities in crashes involving vehicles travelling in different directions. Crashes occurred at intersections are less likely to result in truck driver fatalities. Nevertheless, six features exhibit opposite effects on passenger car driver and truck driver fatalities, including the age of the passenger car driver, crash hour, the passenger car age, road surface condition, weather condition and the truck age. Compared to younger passenger car drivers, older drivers are more prone to fatalities. However, the presence of older passenger car drivers would reduce the fatality risk of truck drivers. This result further emphasizes the vulnerability of older passenger car drivers. Regarding the crash hour, crashes occurred earlier in the day (0:00∼6:59) would increase the fatality risk of passenger car drivers but decrease the fatality risk of truck drivers. This is probably due to passenger car drivers are more likely to be affected by fatigue compared to truck drivers on night shift, who have more experience in this condition. Newer passenger cars are less prone to passenger car driver fatalities, but more prone to truck driver fatalities. This may be due to a newer car has more safety features and better mechanical condition. When road surface is covered with ice, truck drivers are more likely to suffer from fatalities. With higher center of gravity, trucks are more prone to rollover on icy road, leading to fatal injuries. Likewise, icy road surface would increase the fatality risk of passenger car drivers in some cases. But in other cases, it might decrease the fatality risk. This is probably due to some passenger car drivers are more cautious in this situation. As such, they tend to reduce speed to avoid severe crashes. Regarding weather condition, snow would always decrease the fatality risk of truck drivers. But it might increase the fatality risk of passenger car drivers in some cases. Compared to older trucks, newer trucks (within 5 years) are less prone to truck driver fatalities as the small values of V_YEAR_t (the truck age) are associated with negative SHAP values in Figure 5. Nevertheless, Figure 4 suggests that newer trucks would slightly increase the passenger car drivers' fatality risks.

VI. CONCLUSION

This study aims to explore factors affecting the passenger car driver and truck driver crash severity. For the modeling purpose, crashes involving one passenger car and one truck from 2007 to 2017 in Canada are collected and processed. To compare the classification performance of different classifiers, this study uses five different classification models: multinomial logistic regression, Naive Bayes, Classification and Regression Tree (CART), support vector machine, and eXtreme Gradient Boosting (XGBoost). In view of the imbalanced crash severity distribution, four data imbalance treatment approaches are separately applied, including over-sampling, under-sampling, a hybrid method combining under-sampling and over-sampling, and a cost-sensitive learning method. Each classification model is combined with one data imbalance treatment approach to generate the classification result. In light of the data imbalance issue, this study selects geometric mean (G-mean), overall accuracy, and AUC to evaluate the classification performance. To improve the interpretability of classification model results, a recently proposed approach called SHAP (Shapley Additive Explanations) is applied.

For the passenger car driver crash severity analysis, XGBoost combined with cost-sensitive learning generates the best result (G-mean = 58.23%, overall accuracy = 60.37%, AUC = 0.72). Likewise, regarding the truck driver crash severity analysis, the best result is also achieved by combining XGBoost with cost-sensitive learning (G-mean = 55.55%, overall accuracy = 63.57%, AUC = 0.70). G-mean in the current study outperforms the best results in previous literature regarding the 3-class classification [6], [7].

To make the XGBoost model results more informative, this study adopts a recently proposed approach called SHAP. Based on game theory and local explanation, SHAP is designed to explain the result of any machine learning model in a consistent and accurate way. Impacts of features on passenger car driver and truck driver injury severity are separately reported. Additionally, to explore factors affecting driver fatalities, impacts of features on passenger car and truck driver fatal crashes are presented. Among all these features, most of them have similar effects on passenger car driver and truck driver fatalities. However, six features exhibit opposite effects, including the age of the passenger car driver, crash hour, the passenger car age, road surface condition, weather condition and the truck age. Results of the current study could provide some valuable insights to alleviate severity of passenger car-truck crashes. For instance, both passenger car and truck drivers should be aware that wearing safety belts could significantly reduce their fatality risks in crashes. Besides, properly installing traffic control devices (such as traffic signals, stop sign or warning sign) could be an economical and effective method to reduce fatality risks in passenger car-truck crashes. Passenger car drivers should be extremely cautious when driving between midnight to 6 am, especially on roads with heavy truck traffic. Besides, traffic safety education campaign targeted at elderly male passenger car drivers should be promoted. They should be more careful when driving on truck corridors. Traffic administration department should set up more safety warning signs on curved roads and mountain roads. For a snowy

country like Canada, timely clearing of snow and icing on the road is essential to improve traffic safety, especially on truck corridors.

This study demonstrates that XGBoost combined with cost-sensitive learning is a decent method to identify factors affecting injury severity when data is highly imbalanced. Besides, SHAP could be a valuable tool for interpreting results of crash severity analysis models.

VII. STUDY LIMITATIONS

Meanwhile, the study limitations should be noted. Firstly, variables used in this study could be expanded to increase the reliability of model results. Examples include traffic flow information prior to the crash, speeding, driver fatigue, and working experience of the truck driver, etc. Secondly, although this study applies four different approaches to handle the data imbalance issue, other methods in imbalanced learning field are also worth exploring. For instance, Generative Adversarial Networks (GAN) seem promising in improving the performance of imbalanced learning models. Future studies should test the potential of GAN in traffic safety analysis. Thirdly, this study uses only police-reported crash data for modeling. Future studies should consider building models with data generated from traffic simulators. Then, the performance of different models could be compared. Lastly, the quality of model results is only as good as the input data. All crashes are recorded and coded by police officers based on the best available information. As such, some errors in the data are inevitable, which might affect the quality of model results.

REFERENCES

[1] USDOT/BTS. (2018). *Commodity Flow Survey Overview 2017*. Accessed: Oct. 28, 2019. [Online]. Available: https://www.bts.gov/cfs
[2] (2019). *National Highway Traffic Safety Administration 2017 Data: Large Trucks*. Accessed: Oct. 29. 2019. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812663
[3] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.
[4] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.
[5] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases," *Neurocomputing*, vol. 175, pp. 935–947, Jan. 2016.
[6] H. Jeong, Y. Jang, P. J. Bowman, and N. Masoud, "Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data," *Accident Anal. Prevention*, vol. 120, pp. 250–261, Nov. 2018.
[7] L.-Y. Chang and J.-T. Chien, "Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model," *Saf. Sci.*, vol. 51, no. 1, pp. 17–22, Jan. 2013.
[8] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Anal. Prevention*, vol. 45, pp. 478–486, Mar. 2012.
[9] Z. Zheng, P. Lu, and B. Lantz, "Commercial truck crash injury severity analysis using gradient boosting data mining model," *J. Saf. Res.*, vol. 65, pp. 115–124, Jun. 2018.
[10] F. Mannering, "Temporal instability and the analysis of highway accident data," *Analytic Methods Accident Res.*, vol. 17, pp. 1–13, Mar. 2018.
[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
[12] A. J. Khattak and F. Targa, "Injury severity and total harm in truck-involved work zone crashes," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1877, no. 1, pp. 106–116, Jan. 2004.
[13] J. D. Lemp, K. M. Kockelman, and A. Unnikrishnan, "Analysis of large truck crash severity using heteroskedastic ordered probit models," *Accident Anal. Prevention*, vol. 43, no. 1, pp. 370–380, Jan. 2011.
[14] H.-C. Chu, "An investigation of the risk factors causing severe injuries in crashes involving gravel trucks," *Traffic Injury Prevention*, vol. 13, no. 4, pp. 355–363, Jul. 2012.
[15] S. Choi, C. Oh, and M. Kim, "Risk factors related to fatal truck crashes on korean freeways," *Traffic Injury Prevention*, vol. 15, no. 1, pp. 73–80, Jan. 2014.
[16] C. Dong, Q. Dong, B. Huang, W. Hu, and S. S. Nambisan, "Estimating factors contributing to frequency and severity of large truck-involved crashes," *J. Transp. Eng. Part A Syst.*, vol. 143, no. 8, pp. 1–9, 2017.
[17] M. Uddin and N. Huynh, "Truck-involved crashes injury severity analysis for different lighting conditions on rural and urban roadways," *Accident Anal. Prevention*, vol. 108, pp. 44–55, Nov. 2017.
[18] S. Newnam, D. Blower, L. Molnar, D. Eby, and S. Koppel, "Exploring crash characteristics and injury outcomes among older truck drivers: An analysis of truck-involved crash data in the united states," *Saf. Sci.*, vol. 106, pp. 140–145, Jul. 2018.
[19] M. Moomen, M. Rezapour, and K. Ksaibati, "An investigation of influential factors of downgrade truck crashes: A logistic regression approach," *J. Traffic Transp. Eng. (English Ed.)*, vol. 6, no. 2, pp. 185–195, Apr. 2019.
[20] M. Osman, R. Paleti, S. Mishra, and M. M. Golias, "Analysis of injury severity of large truck crashes in work zones," *Accident Anal. Prevention*, vol. 97, pp. 261–273, Dec. 2016.
[21] S. A. Useche, B. Cendales, F. Alonso, and L. Montoro, "Multidimensional prediction of work traffic crashes among spanish professional drivers in cargo and passenger transportation," *Int. J. Occupational Saf. Ergonom.*, vol. 26, pp. 1–8, Apr. 2020.
[22] Y. Wang and C. G. Prato, "Determinants of injury severity for truck crashes on mountain expressways in China: A case-study with a partial proportional odds model," *Saf. Sci.*, vol. 117, pp. 100–107, Aug. 2019.
[23] E. Rahimi, A. Shamshiripour, A. Samimi, and A. Mohammadian, "Investigating the injury severity of single-vehicle truck crashes in a developing country," *Accident Anal. Prevention*, vol. 137, Mar. 2020, Art. no. 105444.
[24] A. Behnood and F. Mannering, "Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes," *Analytic Methods Accident Res.*, vol. 23, Sep. 2019, Art. no. 100102.
[25] A. Behnood and N. S. S. Al-Bdairi, "Determinant of injury severities in large truck crashes: A weekly instability analysis," *Saf. Sci.*, vol. 131, Nov. 2020, Art. no. 104911.
[26] M. T. Haq, M. Zlatkovic, and K. Ksaibati, "Investigating occupant injury severity of truck-involved crashes based on vehicle types on a mountainous freeway: A hierarchical Bayesian random intercept approach," *Accident Anal. Prevention*, vol. 144, Sep. 2020, Art. no. 105654.
[27] R. O. Mujalli, G. López, and L. Garach, "Bayes classifiers for imbalanced traffic accidents datasets," *Accident Anal. Prevention*, vol. 88, pp. 37–51, Mar. 2016.
[28] Y. M. Goh, C. U. Ubeynarayana, K. L. X. Wong, and B. H. W. Guo, "Factors influencing unsafe behaviors: A supervised learning approach," *Accident Anal. Prevention*, vol. 118, pp. 77–85, Sep. 2018.
[29] Transport Canada. (2019). *National Collision Database—Open Government Portal*. Accessed: Nov. 26, 2019. [Online]. Available: https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a?tdsourcetag=s_pctim_aiomsg
[30] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[31] B. Zhou, Z. Li, S. Zhang, X. Zhang, X. Liu, and Q. Ma, "Analysis of factors affecting hit-and-run and non-hit-and-run in vehicle-bicycle crashes: A non-parametric approach incorporating data imbalance treatment," *Sustainability*, vol. 11, no. 5, p. 1327, Mar. 2019.
[32] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972.
[33] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.
[35] P. T. Savolainen, F. L. Mannering, D. Lord, and M. A. Quddus, "The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident Anal. Prevention*, vol. 43, no. 5, pp. 1666–1676, Sep. 2011.
[36] A. T. Kashani and A. S. Mohaymany, "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models," *Saf. Sci.*, vol. 49, no. 10, pp. 1314–1320, Dec. 2011.
[37] D. Li, Y. Zhao, Q. Bai, B. Zhou, and H. Ling, "Analyzing injury severity of bus passengers with different movements," *Traffic Injury Prevention*, vol. 18, no. 5, pp. 528–532, Jul. 2017.
[38] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[39] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[40] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems* (Lecture Notes in Computer Science), vol. 1857. Berlin, Germany: Springer, 2000, pp. 1–15.
[41] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *Int. J. Knowl. Eng. Soft Data Paradigms*, vol. 3, no. 1, p. 4, 2011.
[42] S. Lim and S. Chi, "Xgboost application on bridge management systems for proactive damage estimation," *Adv. Eng. Informat.*, vol. 41, Aug. 2019, Art. no. 100922.
[43] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "Explainable AI for trees: From local explanations to global understanding," 2019, *arXiv:1905.04610*. [Online]. Available: https://arxiv.org/abs/1905.04610

BEI ZHOU received the B.S. degree in transportation engineering from Southeast University, China, in 2007, and the Ph.D. degree in transportation engineering from the Illinois Institute of Technology, Chicago, IL, USA, in 2012. He is currently a Lecturer with the College of Transportation Engineering, Chang'an University, China. His research interests include traffic safety analysis, driving behavior analysis, traffic simulation, and machine learning.

XIQING WANG received the M.S. degree in civil engineering from the University of Nottingham, U.K., in 2012. She is currently pursuing the Ph.D. degree in civil engineering with Chang'an University, Xi'an, China. Her research interests include safety evaluations and structure safeties.

SHENGRUI ZHANG received the M.S. and Ph.D. degrees in transportation engineering from Chang'an University, China, in 1997 and 2002, respectively. He was a Visiting Scholar with the Illinois Institute of Technology, Chicago, IL, USA, in 2013. He is currently a Professor with the College of Transportation Engineering, Chang'an University. His research interests include traffic flow theory, traffic simulation, transportation planning, and traffic safety.

ZONGZHI LI received the M.S. degree in transportation and infrastructure systems engineering, the M.S. degree in operations research, and the Ph.D. degree in transportation engineering from Purdue University, in 2000, 2002, and 2003, respectively. He is currently a Professor with the Illinois Institute of Technology (IIT), Chicago. He also coordinates the IIT Transportation Engineering Program and the Infrastructure Engineering and Management Program, and also serves as the Director of the IIT Sustainable Transportation and Infrastructure Research (STAIR) Center and the IIT Transportation Engineering Laboratory. His research interests include multimodal transportation infrastructure and dynamic traffic network performance modeling, big data driven transportation asset management, and transportation network economics.

SHAOFENG SUN received the B.E. degree in civil engineering from Chang'an University, Xi'an, China, in 2018, where he is currently pursuing the M.S. degree in transportation engineering. His research interests include traffic safety, traffic control and management, and sustainable transportation.

KUN SHU received the B.E. degree in road and river-crossing engineering from Chang'an University, Xi'an, China, in 2018, where he is currently pursuing the M.S. degree in transportation engineering. His research interests include traffic safety and traffic flow.

QING SUN received the B.E. degree in transportation engineering from Chang'an University, Xi'an, China, in 2018, where she is currently pursuing the M.S. degree in transportation engineering. Her research interests include traffic safety and traffic flow theory.