IATSS Research 45 (2021) 277–285

Contents lists available at ScienceDirect

IATSS Research

Research Article

Using machine leaning techniques for evaluation of motorcycle injury severity

Mahdi Rezapour [a,b,c,*], Ahmed Farid [b], Sahima Nazneen [b], Khaled Ksaibati [c]

[a] Department of Civil and Architectural Engineering, University of Wyoming, 1000 E. University Avenue, Department 3295, Laramie, WY 82071, United States
[b] Wyoming Technology Transfer Center, Department of Civil and Architectural Engineering, University of Wyoming, 1000 E. University Avenue, Department 3295, Laramie, WY 82071, United States
[c] Wyoming Technology Transfer Center, Department of Civil and Architectural Engineering, 1000 E. University Avenue, Department 3295, Laramie, WY 82071, United States

ARTICLE INFO

Article history:
Received 22 August 2019
Received in revised form 6 February 2020
Accepted 27 July 2020
Available online 21 August 2020

Keywords:
Motorcycle crashes
Injury severities
Machine learning methods
random forest

ABSTRACT

There is a growing interest in the application of the machine learning techniques in predicting the motorcycle crash severity. This is partly due to a progress in autonomous vehicles technology, and machine learning technique, which as a main component of autonomous vehicle could be implemented for traffic safety enhancement. Wyoming's motorcycle crash fatalities constitute a concern since the count of riders being killed in motorcycle crashes in 2014 was 11% of the total road fatalities in the state. The first step of crash reduction could be achieved through identification of contributory factors to crashes. This could be accomplished by using a right model with high accuracy in predicting crashes. Thus, this study adopted random forest, support vector machine, multivariate adaptive regression splines and binary logistic regression techniques to predict the injury severity outcomes of motorcycle crashes. Even though researchers applied all the aforementioned techniques to model motorcycle injury severities, a comparative analysis to assess the predictive power of such modeling frameworks is limited. Hence, this study contributes to the road safety literature by comparing the performance of the discussed techniques. In this study, Wyoming's motorcycle crash injury severities are modeled as functions of the characteristics that give rise to crashes. Before conducting any analyses, feature reduction was used to identify a best number of predictors to be included in the model. Also to have an unbiased estimation of the performance of different machine learning techniques, 5-fold cross-validation was used for model performance evaluation. Two measure, Area under the curve (AUC), and confusion matrix were used to compare different models' performance. The machine learning results indicate that random forest model outperformed the other models with the least misclassification and higher AUC. It was also revealed that a dichotomous response variable, with fatality and incapacitation injury in one category, along with all other categories in another group would result in a lower misclassification rate than a polychotomous response variable. This might result from the nature of motorcycle crashes, lacking a protection compared with passenger cars, preventing machine learning technique to get trained properly. Moreover, the most important variables identified by the random forest model are those related to the operating speed, resentful other party, traffic volume, truck traffic volume, riding under the influence, horizontal curvature, wide roadway with more than two lanes and rider's age.

© 2020 International Association of Traffic and Safety Sciences. Production and hosting by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Traffic crashes result in significant number of fatalities and injuries which, in turn, lead to tremendous social and economic losses. Motorcycle crashes pose serious threats considering the fact that they are subject to higher risks of being severe or fatal crashes. Understanding the relationship between motorcycle crash injury severity outcomes and the crash contributing factors is crucial in order to address motorcycle safety. That provides insights about the appropriate safety countermeasures to implement and the policies to enforce.

Motorcycle safety is one of the critical concerns in transportation engineering. The National Highway Traffic Safety Administration (NHTSA) stated that according to the per vehicle miles traveled (VMT) in 2016, the count of motorcyclist deaths was 28 times higher than other vehicle occupants while the risk of being injured was nearly 5 times higher [1]. The NHTSA also reported that, in the same year, fatal motorcycle crashes increased by about 5.1% from 2015 [2]. Mountainous topography and adverse weather conditions contribute to higher fatal motorcycle

* Corresponding author.
E-mail addresses: mrezapou@uwyo.edu (M. Rezapour), afarid@uwyo.edu (A. Farid), snazneen@uwyo.edu (S. Nazneen), khlaed@uwyo.edu (K. Ksaibati).
Peer review under responsibility of International Association of Traffic and Safety Sciences.

https://doi.org/10.1016/j.iatssr.2020.07.004
0386-1112/© 2020 International Association of Traffic and Safety Sciences. Production and hosting by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

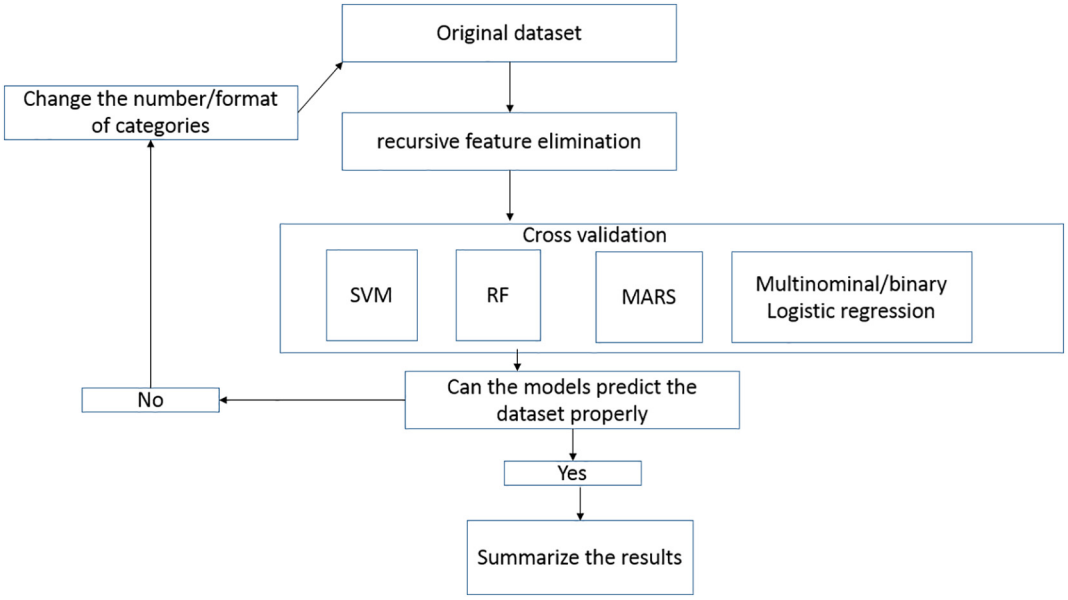

Fig. 1. Flow chart of the methodology.

crash counts in Wyoming. Based on a study conducted in Wyoming in 2014, motorcyclist fatalities accounted for 11% of the total traffic fatalities in the state [3].

In this current study, crash injury severity analyses were performed on Wyoming's motorcycle crashes using various analytical techniques. Non-parametric machine learning methods implemented were random forest regression, multivariate adaptive regression splines (MARS) and support vector machine (SVM) modeling. Furthermore the multinomial logistic regression method, also known as the multinomial logit regression method, was employed. In order to improve the accuracy of the models, different types of categories with a change in category number were implemented. All the modeling techniques are discussed and the misclassification rates are used to compare the performances of the models to determine the best fit model. The general framework of the analyses is presented in Fig. 1.

Every machine learning technique consists of few main components: optimizer, cost function and iteration processes. Thus the performance of various machine learning techniques would vary based on their capabilities in training themselves, and implementing that performance when there are exposed to a new dataset.

Having said that, it is important for the dataset to be well-prepared for the algorithms. Thus one of the objectives of this study was to determine whether modeling the response as three injury severity categories or two categories would result in a better prediction accuracy. The findings of the superior model, which are the interpretations of the influences of the factors that contribute to motorcycle crash injury severity, would aid traffic engineers in making better informed decisions when choosing appropriate mitigation measures to improve motorcycle safety. Studies, in which the aforementioned modeling techniques are employed, are discussed in the background section followed by the methodology section in which the modeling techniques being employed for this research are elaborated. Subsequently, the results are presented and discussed in the empirical analysis section. Finally, the conclusion of this study are discussed.

2. Background

Non-parametric machine learning techniques and parametric techniques are popular in severity analyses [4–6]. As previously mentioned, random forest regression, MARS, SVM and logistic regression are implemented. That is to predict the association between motorcycle crash severity and the crash contributing factors. A concise literature review of these methods is discussed in the following paragraphs.

2.1. Random forest regression

The random forest model, a powerful machine learning algorithm is widely used in injury severity analyses. Harb et al. [7] employed the random forest modeling technique to interpret drivers', vehicles' and environmental characteristics associated with drivers' crash avoidance maneuvers. Such characteristics pertain to angle, head-on and rear-end crashes.

Saha et al. [8] applied the random forest model to gauge the importance of the Highway Safety Manual's (HSM) calibration parameters in crash frequency predictions. Siddiqui et al. [9] adopted the random forest approach to understand the impacts of the variables associated with total crashes, variables that contribute to severe crashes and planning variables so as to incorporate proactive safety measures in transportation planning. Hossain and Muromachi [10] employed the random forest model to rank the influential factors associated with traffic crashes at basic freeway segments and ramps. Hassan and Abdel-Aty [11] applied the random forest method to select significant variables affecting poor visibility related crashes by using real-time traffic flow data of freeways. In another study, Xu et al. [12] employed the random forest model to evaluate the risk of crashes as a function of the traffic flow operation for each level of service A through F. The random forest model is demonstrated to be a promising data mining approach in the aforementioned studies. The model's advantage is that no specific relationship between the outcome and the predictors ought to be assumed.

2.2. Support vector machine

The SVM method is a discriminative supervised machine learning algorithm based on the structural learning theory and the structural risk minimization principal. A supervised method is one in which an outcome variable is modeled as opposed to an analysis in which associations among variables are interpreted without having an outcome being modeled.

SVM models were used in various transportation research studies. Li et al. [4] investigated the feasibility of using SVM models in crash injury

severity analysis. The authors compared the performances of the SVM model and the ordered probit (OP) model for predicting the injury severity of crashes. The results indicated that the SVM model provides more accurate predictions than the OP model. Yu and Abdel-Aty [13] employed the SVM technique with the radial basis kernel function to develop crash injury severity models for a mountainous freeway section. Two other techniques, namely the fixed parameter logit model and the random parameter logit model were also implemented. The comparison of the three models' performances demonstrated that the SVM and the random parameter models exhibited better fits than the logit model. Li et al. [14] also compared the performances of the SVM model and the traditional negative binomial (NB) regression model for predicting motor vehicle crashes. It was concluded that the SVM model demonstrated a higher prediction accuracy than the NB model. Gu et al. [15] developed a traffic fatalities prediction model using the SVM technique. The authors generated particle swarm with mutation optimization for selecting parameters to improve prediction accuracy. The authors also developed other classification algorithms. Compared to the other developed algorithms, the prediction model of the traffic fatalities based on particle swarm with mutation optimization SVM was the most accurate algorithm.

2.3. Multivariate adaptive regression splines

The MARS technique is a nonparametric regression technique developed by Friedman [16]. This technique does not assume any particular relationship between the dependent variable and the independent variables. Instead, it explores the nonlinear relationship between the response and the predictors by fitting the data into spline linear regression functions.

Although the application of the MARS in transportation is limited, this technique shows promising predictive power in crash injury analysis.

Chang et al. [17] employed the MARS technique to assess the effects of geometric design, traffic and environmental factors on freeway crashes. The research team found that horizontal alignment, vertical alignment, average daily traffic (ADT) volume, heavy vehicle ADT volume and annual precipitation vary with crash frequencies in a nonlinear fashion. Abdel-Aty and Haleem [18] developed NB regression and MARS models to explore the effects of geometric and traffic variables on angle crashes at unsignalized intersections. The mean square prediction error (MSPE) results demonstrated that the MARS models outperformed the NB models. Haleem et al. [19] also employed the MARS method and NB technique to predict rear-end crashes at unsignalized intersections.

2.4. Logistic regression model

In the multinomial logit model, as modified version of binary logit, the response variable is nominal with multiple categories. This model is a common models applied in crash injury severity analyses. Shankar and Mannering [20] implemented the multinomial logit model to analyze motorcycle rider injury severity in single-vehicle crashes. The authors considered five injury severity levels which are property damage only (PDO), possible injury, evident injury, disabling injury and fatality. Schneider and Savolainen [21] modeled motorcycle injury severities by crash type using the multinomial logit structure. The crash types were single- and multi-vehicle motorcycle crashes. As per the study's findings, the impacts of crash contributing factors varied by crash type and crash location. The study also revealed that collision with fixed objects and crashes of which the impact forces were large increased the risk of severe injuries. Geedipally et al. [22] assessed the factors that contributed to the severity of motorcycle crashes in urban and rural areas using the multinomial logit model. Alcohol consumption, gender, lighting, presence of horizontal curves and presence of vertical curves were identified as influential factors in motorcycle injury crashes in urban areas. In rural areas, besides the aforementioned factors, elderly riders, single-motorcycle crashes, angle crashes and divided highways contributed to severe injuries. It should be noted that this method is considered for the primarily categories, which was three category response. If it was decided to change the number of category to two, the model would be changed into binary logistic regression.

In general, there is an interest in the use of the mentioned nonparametric and parametric modeling techniques to examine the factors that contribute to crash injury severity. Yet, it should be noted that there is no in-depth study conducted to assess the predictive power of each modeling method to recommend the most accurate method. For this study, the discussed techniques are implemented to model motorcyclist injury severity. The prediction accuracy of the models are compared via misclassification rate and the area under the curve (AUC) results and the best fit model for the data is recommended. Interpretations of the results of the best fit model are discussed as well.

3. Methodology

Although the previous section also talks about the methodological aspects briefly, this section would detail the above descriptions. The random forest, SVM, MARS and logistic regression modeling methods would be discussed. Note that for each model, *k*-fold cross-validation (CV) is implemented to assess the misclassification rate [23]. The *k*-fold CV method is used to fragment the data into *k* non-overlapping sets or folds. The entire data excluding one fold is used for training the model while the remaining fold would be used as the validation set to test the model. Hence, a misclassification rate would be computed. Then, the entire data excluding a fold would be used for model training while that remained fold would be used for model testing. Thus, another misclassification rate is computed. The process is repeated until all *k* folds are used as validation sets and *k* misclassification rates are obtained. The average of the *k* rates is the CV misclassification rate. For this research, *k* is chosen as five.

3.1. Random forest model

The random forest model consists of a large number of individual decision trees as ensemble. The core concept of random forest lie behind the fact that when a large number of individual, low correlated decision trees come together would outperform any of those single decision trees. The trees subset the datasets into further sets termed nodes. The method is a combination of Breiman's "bagging" idea and Ho's "random subspace method" [24]. At each node, a best split is chosen from a random subset of the predictors instead of all of them to determine the splitting decision. It should be noted while assigning observations to various decision tree, due to sampling with replacement, it is possible some records would be used by few decision trees. It should also be noted that sampling does not occur on just observations, but also features to give a generalization power to random forest, resulting in a lower variance.

Some of the main characteristics of decision trees are low bias and high variance. Low bias indicates that if a decision tree is trained to its complete depth, it would be probably trained for the training dataset. So the training error would be minimal. High variance means while implementing a trained decision trees on the test data, the decision trees are prone to give a larger amount of errors, high variance. This could be called as overfitting. This overfitting could be addressed in random forest by using multiple decision trees. Thus random forest would tend to have both low bias and low variance.

The splitting decision is made such that the misclassification error rate would be minimized. When a particular tree is developed from a bootstrap sample, one third of the observations would not be used in the development of the tree. These observations are termed out-of-bag (OOB) data points which comprise data used to compute an unbiased prediction error and estimate the measures of variable importance. The variable importance is a metric that gauges the influence of the

independent variables on the response. The plot of the OOB error rate against various tree counts is developed in this study to determine the optimal number of trees to achieve reliable results. The best number of trees is the least one which produces the lowest misclassification rate [21,25–27]. It should be noted that each individual decision tree would make a decision on the dataset and final decision would be made based on majority votes, voting.

The mean decrease Gini "IncNodePurity" diagram obtained from the "randomForest" function in R [28]. This measure was used to determine the important predictors identified by the random forest method. The diagram provides the node purity value for every predictor of a tree by means of the Gini index [27].

In summary, the condition that choose the optimal condition is called impurity. For classification it could be Gini impurity, or information gain/entropy, and for regression tree it is variance. Thus, while training, it could be computed how much each features decrease the weighted impurity. For random forest, the decrease impurity could be computed by averaging, and the features could be ranked. A larger node purity value represents a higher variable importance. For more details about the random forest technique, the reader is referred to Breiman [24] and Harb et al. [7].

Variable importance of random forest could be presented based on mean decrease accuracy, or mean decrease in gini. The mean decrease in accuracy is equivalent to the number of observations that are misclassified by removing that feature/predictors. On the other hand, gini importance, or mean decreasegini, is defined as the average purity gain by splitting a given predictor. Permuting an important predictor would results in a large decrease in mean decrease in gini.

3.2. Support vector machine model

The SVM method is used to develop an optimal separating hyperplane to categorize the observations into several groups while maximizing the margin between the decision boundaries. The predictors are defined as the vectors $x_p \in R^p$ for $p = 1, 2, 3, \ldots, p$ representing the full set of crash related variables, and the outcome is defined as $y_k \in R^k$ which represents the injury severity levels of the crashes. For a case with two predictors and two outcome categories, the hyperplane is a line that separates the data points that belong to one outcome category and those that belong to the other. Hence, the plane constitutes the decision boundaries. For cases with multiple predictors, the hyperplane is a $p - 1$ dimensional plane. The decision boundaries may or may not be linear depending on the pre-set kernel function. The goal of maximizing the distance between the hyperplane and the nearest data points is also known as the support vectors. When it comes to modeling multiple outcome categories, two SVM approaches are attempted. They are the one versus one, and one versus all approaches. The reader may consult James et al. [23] and Li et al. [4] for more detailed information about the SVM model. There are linear and non-linear SVM models. In linear SVM, y_i could be either 1 or −1, indication various classes that $\vec{x}$ belongs to. In summary, the objective is to find a maximum margin hyperplane that could divide the groups of observation with a highest distance between the hyperplane.

3.3. Multivariate adaptive regression splines model

The core component of this method is Hinge function. Which has a form of.

$$\max(0, x-c) \text{ or } \max(0, c-x) \tag{1}$$

where c is a constant, knot. It is clear from the above equation that the function would be zero for part of its range resulting in partitioning the data into disjoint regions.

The MARS method is a local regression method that uses a series of functions to model complex nonlinear relationships [16]. The global MARS model is defined as shown in Eq. (2) as per Chang et al. [17] and Friedman [16].

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{2}$$

Where:
$\hat{y}$ = predicted response;
a_0 = coefficient of the constant basis function;
a_m = coefficient of the m^{th} basis function;
$B_m(x)$ = m^{th} basis function which may be a single spline function or an interaction of two or more spline functions and
M = number of basis functions included in the MARS model.

The MARS model involves three steps [16,17]. The first step is a constructive phase in which basis functions are developed in several regions of the predictors, and are combined in a weighted sum to define the global MARS model as shown in Eq. (2). This global model typically includes many basis functions which may cause overfitting. The second step is the pruning phase in which some basis functions of the overfitted MARS model are deleted. In the third step, the optimal MARS model is selected from a sequence of smaller models.

In the first step, basis functions are continually added to the model. Basis functions in MARS include either a single spline function or a product (interaction) of two or more spline functions for different predictors. Those basis functions are added in a "two at-a-time" forward stepwise procedure, which selects the best pairs of spline functions to improve the model fit. Each pair consists of one left-sided and one right-sided truncated function defined by a given
knot location as shown in Eqs. (3a) and (3b) respectively.

$$\left[-(x-t)_+^q\right] = \begin{cases} (t-x)^q & x < 1, \\ 0 & \text{otherwise} \end{cases} \tag{3a}$$

$$\left[+(x-t)_+^q\right] = \begin{cases} (x-t)^q & x > 1, \\ 0 & \text{otherwise} \end{cases} \tag{3b}$$

Note that t pertains to the knot location while q is the degree of the polynomial term. The search for the best predictor and knot location is conducted in an iterative process. The predictor that contributes the most to explaining the variance in the data, and the knot location are selected first. Whether an interaction term can be included would be checked at the end of each iteration to improve the model fit. The order of any fitted MARS model denotes the maximum number of basis functions that interact. For example, in a second-order MARS model, the interaction order of the splines is not greater than two. The iterative building procedure would proceed until a maximum number of basis functions, M_{max}, is included.

The second step is the pruning step. This procedure is mainly based on the generalized cross-validation (GCV) criterion [16,17]. The GCV criterion is used to find the overall best fit model from a sequence of fitted models while avoiding overfitting. The GCV criterion incurs a penalty for the model's complexity. A larger GCV value tends to produce a smaller model and vice versa. The GCV criterion is computed by using Eq. (4).

$$GCV(M) = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}{\left(\frac{1-C(M)}{N}\right)^2} \tag{4}$$

Where:
N = number of observations;
y_i = response for observation i;
$\hat{y}$ = predicted response for observation i and
$C(M)$ = complexity penalty function equal to $M + d \times M$.

The term, M, is the number of non-constant basis functions and d is the user-defined cost for each basis function. The higher the cost of d the more basis functions are excluded. Typically, d is magnified during the pruning step to obtain smaller models.

The third step is to select the optimal model. The selection is based on an evaluation of the predictive properties of the different fitted MARS models. For further details about the MARS formulation, see Friedman [16], Put et al. [29] and Sekulic and Kowalski [30].

3.4. Multinomial logit regression model

The multinomial logit regression model [6] is structured such that the injury severity of motorcyclist, i, is defined as $\boldsymbol{\beta_k X_i} + \varepsilon_{ki}$ in which the predictors' vector is denoted by $\boldsymbol{X_i}$ and the coefficients' vector is $\boldsymbol{\beta}$. The parameter ε_{ki} is an error term, which accounts for latent effects that impact the outcome. The index, k, represents the outcome category. According to Manski and McFadden [31], if ε_{ki} follows the generalized extreme value (GEV) distribution, the probability, P_{ik}, of the motorcycle crash of being classified as injury severity outcome k is the following.

$$P_{ik} = \frac{\exp(\boldsymbol{\beta_k X_i})}{\sum_{k=1}^{K} \exp(\boldsymbol{\beta_k X_i})} \tag{5}$$

The coefficients, β's, obtained are those that maximize the log-likelihood function and hence are known as the maximum likelihood estimates (MLE). In this study different number of response categories are considered and based on the results, multinomial or binary logistic would be selected. As a multinomial logistic is an extension of binary logistic regression the equation related to this model is not presented here.

In summary, logistic regression is a type of machine learning technique that can be used for predictive problems. Its predictive analysis is based on the concept of probability. It uses a sigmoid function as the cost function. For a binary classification, the cost function would be limited to 1 and 0. The sigmoid function is used to map the predicted values to probabilities. This function could be written as:

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{6}$$

For a binary classification, for instance, a threshold would be set as 0.5, and response would be rounded to give a binary outcomes. For instance if the prediction function returned 0.6, this would classify the response as 1, as severe crashes in our case.

3.5. Data collection and preparation

The data are collected from the Wyoming Department of Transportation's Critical Analysis Reporting Environment (CARE) package. Specifically, there were 2484 motorcycle crash records belonging to the years 2008 through 2017 of which data are collected. The advantage of the CARE package is that it contains data of a wide spectrum of crash contributing factors mainly collected from police reports. The outcome modeled is the motorcycle crash injury severity. The severities are defined by the KABCO scale which are fatal (K), incapacitating injury (A), non-incapacitating injury (B), possible injury (C) and property damage only (PDO or O) as per the American Association of State Highway and Transportation Officials (AASHTO) [32]. Note that when multiple crashes of different severity levels are referred, their designation are abbreviated. For instance, C and O crashes altogether are referred to as CO crashes. Once the outcome categories modeled are KA, BC and O crashes while in another trial the outcome categories modeled are KA and BCO crashes.

4. Analysis

4.1. Feature selection

As a preliminary step, the recursive feature elimination (RFE) algorithm is implemented to select the independent variables which are strongly associated with the response, namely the motorcycle injury

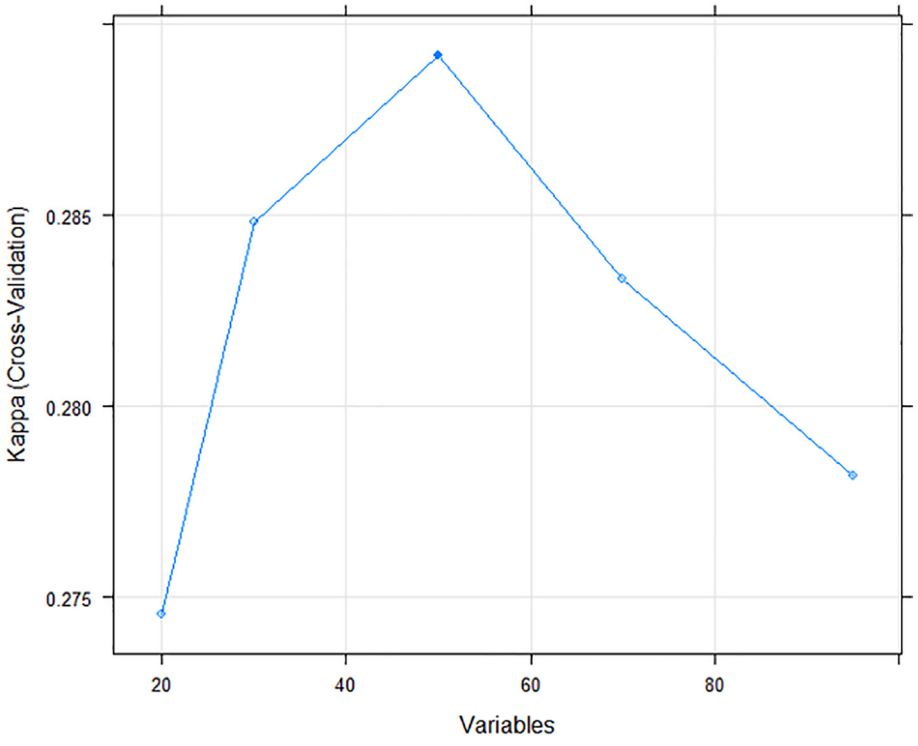

Fig. 2. Results of the recursive feature elimination algorithm.

severity. The package "caret" was used to implement the RFE algorithm [33]. The RFE algorithm utilizes the random forest model to evaluate and rank the variables based on their importance. Multiple counts of variables, 20, 30, 50, 70, and 100 are input (see Fig. 2). As per the CV kappa measure results, 50 variables produced the least error. Hence, only the 50 identified variables are selected for further analysis (see Fig. 2). The reader is referred to Twaddle and Busch [34] for more information about the RFE algorithm.

The method is based on the decrease of permuted OOB kappa coefficients. In other words, kappa value is performance estimate used in this study. This value could be written as:

$$\boldsymbol{k} = 1 - \frac{1-p_0}{1-p_e}$$

Where p_0 is the relative observed agreement among the raters, accuracy, and p_e is the hypothetical probability of chance agreement, expected accuracy. It should be noted that when there is a complete agreement, $k = 1$, while with no agreement $k = 0$.

For this method, first each predictor would be ranked using its importance in the model. So the ordered features would be selected based on the number of predictors to retain in the syntax of the model. After retaining the top *n* features, the model would be reassessed and the kappa would be reported. The simple flowchart of the process in depicted in Fig. 3. About 45 predictors were selected by RFE such as traffic, restrain conditions, gender.

4.1.1. Model evaluation

Two methods namely confusion matrix and AUC were used to compare different models. The error rate can be calculated based on Eq. (6).

$$\text{Error } rate = \frac{\sum_{i=1}^{s} p_{ij}}{\sum_{i=1}^{s} N_i} \tag{7}$$

where p_{ij} is the number of crashes with severity level of i that is predicted as severity level j, and N_i is the total number of crashes.

4.2. Random forest model

The random forest model is developed using the statistical software R. The function "randomForest" belonging to the package of the same name [28] is used. The number of trees considered is 2100 to comprise the forest because it provides stable estimation of variable importance. Also, three variables are selected to be randomly sampled as candidates at each split of each tree. The variable importance measures are estimated from the developed model.

4.3. Support vector machine

The SVM model is developed using the R function "svm" belonging to the package "e1071" [35]. The training dataset is mapped into a higher dimensional space by the kernel function to obtain the optimal hyperplane. The linear kernel function is adopted in this study because of its accelerated performance. For determining the optimal hyperplane, the SVM method employs an iterative training algorithm to minimize the error function.

4.4. Multivariate adaptive regression splines model

The R software is also utilized to estimate the MARS model. The function is "earth" which belongs to the package of the same name [36]. In the forward pass step, the maximum degree of interaction is selected as three.

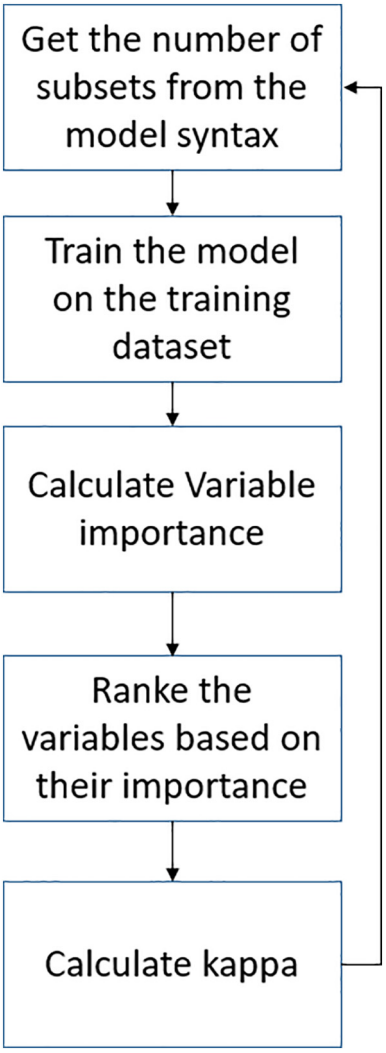

Fig. 3. RFE flow chart.

4.5. Multinomial logit regression model

The R function "multinom" of the package "nnet" is used to model the injury severities of the motorcycle crashes [37]. The function is used without altering its default settings. Later this model is changed to binary logistic regression due changes in categories.

5. Results

Initially, the outcome's categories were set as KA, BC and O crashes. However, the random forest, SVM, MARS and multinomial logit regression models were unable to be trained to predict the severities in the test data accurately. Hence, only two outcome categories are considered namely KA and BCO crashes. Savolainen and Mannering [6] combined C and O crashes as one category capable of being nested in a nested logit structure. The model comparison results considering the two outcome categories, KA and BCO, are presented in Tables 1 and 2. The training error results are presented in Table 1 while the CV error results are presented in Table 2. Note that the multinomial logit regression structure reduces to the binary logit regression structure when modeling two outcome categories.

As shown in Table 2, although different models resulted in a very close results, the random forest model produces the best fit results as per the overall CV misclassification rate of 24% and AUC of 0.7. The

Table 1
Models' training confusion matrices considering two outcome categories.

Model	Confusion Matrix					
Random Forest			Predicted Category			AUC
			KA	BCO	Overall Misclassification Rate	
	Actual Category	KA	490	18	10%	0.86
		BCO	17	1182		
Multivariate Adaptive Regression Splines			Predicted Category		Overall Misclassification Rate	
			KA	BCO		
	Actual Category	KA	446	402	25%	0.70
		BCO	237	1245		
Support Vector Machine			Predicted Category		Overall Misclassification Rate	
			KA	BCO		
	Actual Category	KA	362	128	23%	0.71
		BCO	302	1072		
Binary Logistic Regression			Predicted Category		Overall Misclassification Rate	
			KA	BCO		
	Actual Category	KA	362	128	23%	0.73
		BCO	302	1072		

training errors presented in Table 1 are considerably lower than their counterpart CV errors as expected. The results of the most important variables identified by the random forest model in terms of mean decrease in model accuracy and mean decrease in the Gini index are presented in Figs. 4 and 5. Note that the AADT, AADTT and RUI are the average annual daily traffic, average annual daily truck traffic and riding under the influence variables respectively.

The results of Figs. 3 and 4 are based on a finalist model with an optimal number of predictors in that model. These figures just presented the top highest important variables. As shown in Figs. 3 and 4, each measure identifies a different ranking of the important variables. The AADT, AADTT and speed related variables were found to influence motorcycle injury severity. That is possibly because during heavy traffic conditions, the speed differentials between motorcycles' speeds and other vehicles' speeds pose hazards. According to Savolainen and

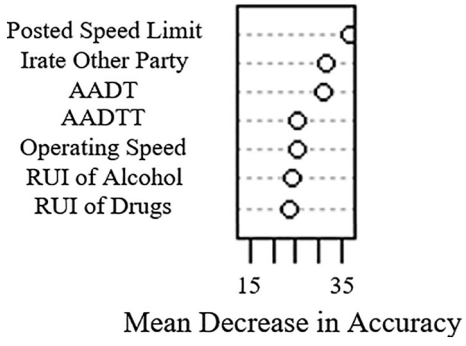

Fig. 4. Variable importance measured by mean decrease in model accuracy.

Table 2
Models' cross validation confusion matrices considering two outcome categories.

Model	Confusion Matrix					
			Predicted Category			
RandomForest			KA	BCO	Overall Misclassification Rate	AUC
	Actual Category	KA	417	431	24%	0.70
		BCO	142	1340		
			Predicted Category		Overall Misclassification Rate	
Multivariate Adaptive Regression Splines			KA	BCO		
	Actual Category	KA	446	402	27%	0.68
		BCO	237	1245		
			Predicted Category		Overall Misclassification Rate	
Support Vector Machine			KA	BCO		
	Actual Category	KA	438	410	26%	0.69
		BCO	194	1288		
			Predicted Category		Overall Misclassification Rate	
Binary Logit Regression			KA	BCO		
	Actual Category	KA	461	387	26%	0.69
		BCO	209	1273		

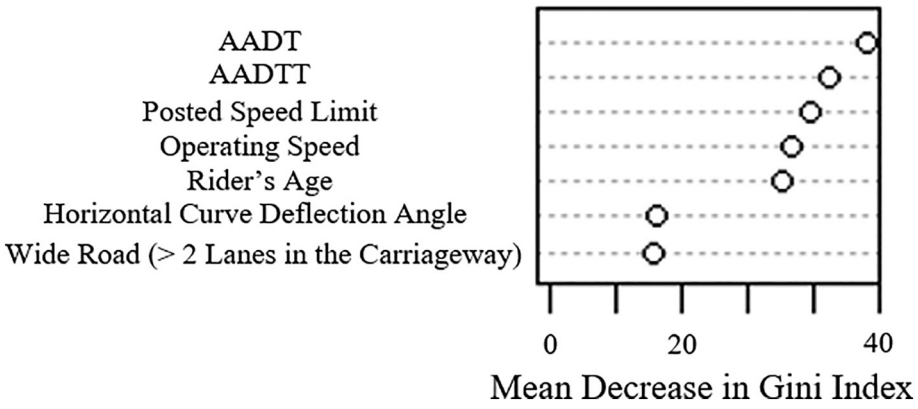

Fig. 5. Variable importance measured by mean decrease in Gini index.

Mannering [6], motorcycle PDO crashes are unlikely to occur at roads with high posted speed limits. Such conditions are instead associated with severe crashes.

It was found an irate other party involved in the motorcycle crash influences injury severity. Such crashes may be those in which the other party is at-fault. The other party may be likely to fail to yield to the motorcycle possibly due to the motorcycle's poor conspicuity [38] coupled with exasperation. Sideswipe crashes, rear-end crashes in which the other party abruptly cuts off the traveled way of the motorcycle, head-on crashes and angle crashes among other types of crashes are also critical [21]. In such crashes, riders are ejected from the motorcycles and are prone to severe injury.

The rider's age is identified as one of the factors that contributes to injury motorcycle crashes. Elderly riders are characterized by deteriorated health conditions and are therefore more prone to be heavily injured than younger riders. Savolainen and Mannering [6] interpreted that motorcycle crashes involving elderly riders are likely to be severe.

Riding under the influence parameters influence the injury severities of motorcycle crashes since they cause brain damage. As per Schneider and Savolainen [21], alcohol limits the capacity to analyze the roadway conditions, elongates perception-reaction time (PRT) and renders the rider to be likely to execute unsafe maneuvers. The authors also found that alcohol consumption considerably increased the risk of motorcycle fatalities. Turner and Georggi [39] recommended awareness campaigns to target riders riding under the influence.

The deflection angle of a horizontal curve experiencing motorcycle crashes is one of the key variables associated with injury severity. Perhaps the riding speed of riders at tight curves is large for the conditions posing hazards. Savolainen and Mannering [6] found that motorcycle crashes at horizontal curves are likely to be severe crashes. The interaction of navigating curves and wet surfaces is another hazard but was not modeled for this study.

Wide roads having more than two lanes in both travel directions combined is another important factor identified by the random forest model. Having more lanes translates to larger AADT and hence larger exposure. That in turn increases the risk of crashes in general regardless of severity.

6. Concluding remarks

Reduction in crashes, especially severe crashes, is a main goal of traffic safety policy makers. A correct understanding of causal factors contributing to severe crashes could help to minimize the severe crashes in a most efficient way. Although numerous statistical methods have been utilized in identification of contributory factors to severe crashes such as motorcycle crashes, a comparative analysis of the predictive power of the models is not extensively investigated.

This study, thus, conducted to present a comprehensive study by considering 4 analyses to model motorcycle crashes, and evaluate the accuracy of the models. Before running any model, feature reduction was conducted to come up with an optimum number of predictors in the model. Also to have an unbiased results cross validation technique was employed. Originally, the outcome categories modeled were KA, BC and O crashes. However, model was not able to train itself properly. Therefore, the outcomes were further grouped into KA and BCO crashes.

To have a better understanding of the prediction power, in addition to AUC, confusion matrix was presented. Although all the models perform similarly in terms of prediction of motorcycle crashes, the best fit model according to the misclassification rate and AUC measure was the random forest model.

After identification of a model outperforming the prediction of motorcycle crashes, the random forest model identified the contributing factors that influence motorcycle crash injury severity the most as travel speed, discontented other party involved in the crash, AADT, AADTT, RUI, horizontal curvature, wide roadway with multiple lanes and rider's age.

Reliable and accurate methods to predict and model crash severity are critical for improving traffic safety. Although the majority of previous studies presented different models to identify the severity of motorcycle crashes, they did not evaluate the prediction power of their models. This paper first evaluates the predictive power of different modeling techniques, and then it presents results of a best performed technique. Machine learning techniques could help artificial intelligence to act as tools for human and policy makers, such as highway patrol, to make better decision for improvement of safety.

Undoubtedly, this study is not without limitations. This research may be extended by including interaction terms involving multiple variables. Also, random parameters may be incorporated in the logit model [40]. The advantage of random parameters is that they capture the effects of the crash contributing factors that are not incorporated in the model, also known as unobserved heterogeneity effects. Other approaches worth resorting to are the uses of neural network and deep learning algorithms [5] to model motorcycle injury severities.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The funding of this study was supported by WYT2.

References

[1] NHTSA, Motorcycle Safety, National Highway Traffic Safety Administration, U.S. Department of Transportation, 2018 https://www.nhtsa.gov/road-safety/motorcycle-safety (Accessed January 7, 2019).
[2] NHTSA, Traffic Safety Facts: Motorcycles, National Highway Traffic Safety Administration, U.S. Department of Transportation, 2018 https://crashstats.nhtsa.dot.gov/Api/Public/Publication/812492 Accessed Jan. 7, 2019.
[3] R. Retting, H. Rothenberg, Spotlight on Highway Safety: Motorcyclist Traffic Fatalities by State 2015 Preliminary Data, Governors Highway Safety Association, https://www.ghsa.org/sites/default/files/2016-12/motorcycles_2015.pdf 2016 Accessed Jan. 7, 2019.
[4] Z. Li, P. Liu, W. Wang, C. Xu, Using support vector machine models for crash injury severity analysis, Accid. Anal. Prev. 45 (2012) 478–486.
[5] S. Das, A. Dutta, K. Dixon, L. Minjares-Kyle, G. Gillette, Using deep learning in severity analysis of at-fault motorcycle rider crashes, Transp. Res. Rec. 2672 (34) (2018) https://doi.org/10.1177/0361198118797212.
[6] P. Savolainen, F. Mannering, Probabilistic models of Motorcyclists' injury severities in single- and multi-vehicle crashes, Accid. Anal. Prev. 39 (5) (2007) 955–963.
[7] R. Harb, X. Yan, E. Radwan, X. Su, Exploring Precrash Maneuvers using classification trees and random forests, Accid. Anal. Prev. 41 (1) (2009) 98–107.
[8] D. Saha, P. Alluri, A. Gan, A random forests approach to prioritize highway safety manual (HSM) variables for data collection, J. Adv. Transp. 50 (4) (2015) 522–540.
[9] C. Siddiqui, M. Abdel-Aty, H. Huang, Aggregate nonparametric safety analysis of traffic zones, Accid. Anal. Prev. 45 (2012) 317–325.
[10] M. Hossain, Y. Muromachi, Understanding crash mechanism on urban expressways using high-resolution traffic data, Accid. Anal. Prev. 57 (2013) 17–29.
[11] H. Hassan, M. Abdel-Aty, Predicting reduced visibility related crashes on freeways using real-time traffic flow data, J. Saf. Res. 45 (2013) 29–36.
[12] C. Xu, P. Liu, W. Wang, Z. Li, Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service, Transp. Res. A Policy Pract. 69 (2014) 58–70.
[13] R. Yu, M. Abdel-Aty, Utilizing support vector machine in real-time crash risk evaluation, Accid. Anal. Prev. 51 (2013) 252–259.
[14] X. Li, D. Lord, Y. Zhang, Y. Xie, Predicting motor vehicle crashes using support vector machine models, Accid. Anal. Prev. 40 (4) (2008) 1611–1618.
[15] X. Gu, T. Li, Y. Wang, L. Zhang, Y. Wang, J. Yao, Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization, J. Algor. Comput. Technol. 12 (1) (2017) 20–29.
[16] J. Friedman, Multivariate adaptive regression splines, Ann. Stat. 19 (1) (1991) 1–67.
[17] L. Chang, H. Chu, D. Lin, P. Lui, Analysis of freeway accident frequency using multivariate adaptive regression splines, Proc. Engineering. 45 (2012) 824–829.
[18] M. Abdel-Aty, K. Haleem, Analyzing angle crashes at Unsignalized intersections using machine learning techniques, Accid. Anal. Prev. 43 (1) (2011) 461–470.
[19] K. Haleem, M. Abdel-Aty, J. Santos, Multiple applications of multivariate adaptive regression splines technique to predict rear-end crashes at Unsignalized intersections, Transp. Res. Rec. 2165 (1) (2010) 33–41.
[20] V. Shankar, F. Mannering, An exploratory multinomial Logit analysis of single-vehicle motorcycle accident severity, J. Saf. Res. 27 (3) (1996) 183–194.
[21] W. Schneider, P. Savolainen, Comparison of severity of motorcyclist injury by crash types, Transp. Res. Rec. 2265 (1) (2011) 70–80.
[22] S. Geedipally, P. Turner, S. Patil, Analysis of motorcycle crashes in Texas with multinomial Logit model, Transp. Res. Rec. 2265 (1) (2011) 62–69.
[23] G. James, D. Witten, T. Hastie, R. Tibshirani, Support Vector Machines. In an Introduction to Statistical Learning with Applications in R, Springer, New York City, New York, 2013 341–372.
[24] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[25] H. Pang, A. Lin, M. Holford, B. Enerson, B. Lu, M. Lawton, E. Floyd, H. Zhao, Pathway analysis using random forests classification and regression, Bioinformatics 22 (16) (2006) 2028–2036.
[26] R. Grimm, T. Behrens, M. Märker, H. Elsenbeer, Soil organic carbon concentrations and stocks on Barro Colorado Island – digital soil mapping using random forests analysis, Geoderma 146 (1–2) (2008) 102–113.
[27] S. Kuhn, B. Egert, S. Neumann, C. Steinbeck, Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction, BMC Bioinform. 9 (1) (2008) 400.
[28] L. Breiman, A. Cutler, A. Liaw, M. Wiener, Breiman and Cutler's Random Forests for Classification and Regression, https://cran.r-project.org/web/packages/randomForest/randomForest.pdf 2018.
[29] R. Put, Q. Xu, D. Massart, Y. Vander-Heyden, Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies, J. Chromatogr. A 1055 (1–2) (2004) 11–19.
[30] S. Sekulic, B. Kowalski, MARS: A Tutorial, J. Chemom. 6 (4) (1992) 199–216.
[31] C. Manski, D. McFadden, Econometric models of probabilistic choice, Structural Analysis of Discrete Data with Econometric Applications, M.I.T. Press, Cambridge, Massachusetts 1981, pp. 198–272.
[32] American Association of State Highway and Transportation Officials, Highway Safety Manual, American Association of State Highway and Transportation Officials, Washington D.C., 2010
[33] M. Kuhn, Variable Selection Using the Caret Package, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.1655&rep=rep1&type=pdf 2010 (Accessed February 18, 2019).
[34] H. Twaddle, F. Busch, Binomial and multinomial regression models for predicting the tactical choices of bicyclists at signalised intersections, Transport. Res. F: Traffic Psychol. Behav. 60 (2019) 47–57.
[35] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, 2019 https://cran.r-project.org/web/packages/e1071/e1071.pdf (Accessed February 19, 2019).
[36] S. Milborrow, Multivariate Adaptive Regression Splines, https://cran.r-project.org/web/packages/earth/earth.pdf 2019.
[37] B. Ripley, W. Venables, Feed-Forward Neural Networks and Multinomial Log-Linear Models, https://cran.r-project.org/web/packages/nnet/nnet.pdf 2016.
[38] M. Haque, H. Chin, H. Huang, Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections, Accid. Anal. Prev. 42 (1) (2010) 203–212.
[39] P. Turner, N. Georggi, Analysis of alcohol-related motorcycle crashes in Florida and recommended countermeasures, Transp. Res. Rec. 1779 (1) (2001) 189–196.
[40] F. Mannering, V. Shankar, C. Bhat, Unobserved heterogeneity and the statistical analysis of highway accident data, Anal. Methods Accident Res. 11 (2016) 1–16.