Crash Severity Analysis Through Nonparametric Machine Learning Methods

Vaibhav AGRAWAL\textsuperscript{a}, Sudipa CHATTERJEE\textsuperscript{b}, Sudeshna MITRA\textsuperscript{c}

\textsuperscript{a,b,c} Department of Civil Engineering, Indian Institute of Technology Kharagpur, 721302, India
\textsuperscript{a} E-mail: va8817@gmail.com
\textsuperscript{b} E-mail: sudipachatterjee13@outlook.com
\textsuperscript{c} E-mail: sudeshna@civil.iitkgp.ernet.in

Abstract: Various parametric models such as logistic regression or linear discriminant analysis have been most commonly used to explore factors contributing to the severity of crashes. These models assume a functional form and learn the coefficients for the function from the training data. If the assumptions are breached the models can lead to an incorrect prediction of the crash severity. In such cases, Non-parametric models seek to be the best fit for the training data in constructing the mapping function while maintaining some ability to generalize the unseen data. In this study, three non-parametric machine-learning methods viz. Classification and Regression Tree (CART), Extreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM) have been utilised to identify the critical factors affecting the severity of traffic crashes on the State Highways of India. Among the three studied methods, the XGBoost was found to be the best performing. The results show that the presence of pedestrian facility, type of collision, weather conditions, intersection proximity, type of traffic control and speed limits are the critical factors affecting crash severity on the State Highways.

Keywords: Crash severity, Data Mining, Traffic Crashes, XGBoost.

1. INTRODUCTION

Rise in per capita income in many developing countries have led to rapid motorization, while road safety management and regulations have not kept pace. As a result, traffic fatalities have disproportionately affected the low and middle-income countries, where 90% of the global road deaths occur. In India a total of 4,64,910 traffic crashes have been reported in the calendar year 2017, claiming 1,47,913 lives and causing injuries to 4,70,975 persons (MoRTH, 2017). Although only 3.1% of the total road lengths in India are State Highways, they contribute to 26.9% of the total road fatalities of the country. The rising number of human deaths and major injury due to the traffic crashes had become a major concern among the general public and government authorities in India and studies on identifying the causal factors for crash occurrence and their outcome on such highways is of immense importance.

When gone through literature related to crash studies, it was found that the regression method or multivariate logistic regression methods are the most commonly used techniques to identify the influence of factors on crash occurrence and severity. These parametric models presume a functional form between the important external factors and crash severity. By assuming the functional form parametric methods become highly constrained as on violation of the assumptions can lead to the incorrect prediction of the factors influencing to injury severity and can also result in designing of faulty countermeasure. However, in the non-parametric method, unlike a parametric model, the functional form of the model is derived from the training
data. This is one of the major advantages in non-parametric approaches, which has no prior assumptions and requires no prior knowledge of the functional dependency between the dependent and independent variables. Thus, many researchers are recurrently implementing non-parametric approaches for identifying the factors contributing to crash severity. In doing that several machine learning techniques could be useful to find the association of crash severity and external factors such as weather conditions, geographic characteristics, traffic state, and drivers’/pedestrians’ behaviour.

Therefore, the objective of this study is twofold –

• To examine the factors influencing injury severity in vehicular crashes on State Highways in India and identifying the key determinants.
• To evaluate the effectiveness of non-parametric models used to predict the injury severity of crashes and identify the best model for our dataset by comparing the accuracies of each model for a test dataset.

To address the study objectives, a total of three non-parametric approaches Classification Regression Tree (CART), Extreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM) are explored.

2. STUDY MOTIVATION

Several studies have been carried out in the world to determine the factors that affect the crash severity the most, and by the elimination or control of these factors had eventually prevented the occurrence of severe crashes, major injuries, and fatalities. From a functional standpoint, logit models are of the most common ones used in the analysis of injury severity. As an example, Al-Ghamdi (2002) showed that the location and the cause of the crash were essential factors for the crash severity. Wood and Simms (2002) identified that the size of the cars is the primary factor for crash severity.

Recently, many researchers have analysed crash data using non-parametric methods and other data mining techniques such as decision trees or any other machine learning approaches.

Classification and Regression Tree (CART) is the most widely used non-parametric model for estimation of injury severity and identifying the key determinants of a crash (Chang and Chien, 2013; Abdel-Aty et al., 2005; Kashani and Mohaymany, 2011; Golias and Karlaftis, 2001; Pande et al., 2010; Chong et al., 2004). CART is an important data mining method that is used in various fields ranging from medicine, industries, engineering sector, and other services. It is a decision tree approach that has no predefined underlying relationship between dependent and independent variables. The number and the nature of the parameters are flexible and are determined from data. Various other non-parametric approaches can be used for this study.

Chang and Chien (2013) implemented the CART model to analyse driver injury severity in truck-involved crashes. They demonstrated a non-parametric classification tree model to be an effective technique. They identified drinking-drive, seatbelt used, vehicle type, collision type, driver/vehicle action and number of vehicles involved as the vital factors affecting the severity of injury in truck-involved crashes. Abellán et al. (2013) used decision rules extracted from decision trees to analyse traffic crash severity. They used traffic crash data from rural roads in Granada (Spain). They established decision rules that were useful in understanding the behavioural characteristic of the dataset. Decision trees are governed by the structure of the decision tree. In their study, they extracted knowledge in the form of decision rules by constructing multiple trees with varying root nodes. These decision rules were used to understand the key characteristics of crash cause and its effect on the severity of injuries. CART
is the most popular non-parametric method that is used for predicting injury severity in vehicular crashes. Kashani and Mohaymany (2011) have implemented CART in a similar study related to traffic safety on two-lane, two-way rural roads in Iran. They decomposed the three-class problem to a binary classification problem. They successfully increased the overall accuracy of their model. In addition, the prediction of fatalities was also improved, which was nearly 0% in previous studies. They identified improper overtaking and not using a seatbelt as the key determinants influencing the injury severity in crashes.

Support Vector Machines (SVM) is another non-parametric data mining technique that has been used in the field of transportation studies. Li et al. (2008) applied SVM models for analysing the crash injury. They compared the Ordered Probit (OP) model with an SVM model and found that SVM produced better prediction accuracy. They concluded that for the study of crash injury severity and safety studies SVM can be used to model the functional dependency between injury severity and crash causing factors.

However, Extreme Gradient Boosting (XGBoost) is an ensemble learning method that has been used in this study. It is a relatively new modelling approach that was first introduced in early 2000.

There are very few literatures on non-parametric modelling of crash severity specially for Indian context. Of the few, in the most recent study Kumar et al. (2017) have implemented data mining techniques to analyse the crash data from the state of Uttarakhand in India and also identify the key factors affecting the powered two-wheeler (PTW) road crashes. The authors used three classification approaches namely, CART, Naïve Bayes and SVM for their study. CART was the best performing model out of the three approaches used in the paper. They showed that each district in the state has different factors that influence the accident severity and thus targeted policies need to be adapted in each district in order to reduce the risks associated with the PTW crashes. In a similar study, Kumar et al. (2016) used k-means algorithms to identify the high and low frequency crashes locations in Dehradun district in the state of Uttarakhand. With these crashes, location clusters, authors identified the critical factors affecting the crashes associated with the corresponding cluster. They used association rule mining to identify the dependency among different factors corresponding to the road crashes. Although the method used in this study can reveal specific information related to the road crashes and the severity of crashes, due to the unavailability of specific information at the time of crashes (like speed of vehicle, weather condition, road surface condition etc.) the result found were at a general level. They identified that highways are more prone to traffic crashes, intersections have a high frequency of pedestrian crashes, vehicles are more susceptible to multi-vehicular crashes near the curves and intersection in the marketplaces are exposed to severe crashes.

In most of the works of literature based on road safety in India, the non-parametric methods are applied for urban road networks although the severity of crashes on the State Highways is comparatively larger. With the expansion of highway networks in India, it is important to identify the factors that result in severe crashes and treat the issues scientifically to enhance safety on these highways.

Although CART is the most widely used non-parametric approaches in the study of traffic safety, there is a need to explore different models and find the appropriate technique for a dataset. In this paper, ensemble learning approaches like boosting have been employed that generates multiple classifiers and combine them to obtain a better prediction performance than the traditional methods.

3. STUDY METHODOLOGY
Non-parametric methods perform reasonably well in the case when the data is subjected to outliers. In parametric approaches like regression analysis, outliers can lead to serious issues as they change the value of the parameter coefficient. In decision tree models like CART or XGBoost, the outliers are pruned from the tree in the tree-pruning step of the algorithm. The advantage is that these methods perform exceptionally well when the data size is large. They can produce useful results from only a few important variables by effectively dealing with large data sets containing a large number of independent variables (Chang and Wang, 2006).

However, these methods also have certain limitation, which includes slow training. Due to more training parameters, these methods are often slow to train. In XGBoost the decision trees are built iteratively which results in higher learning time. The non-parametric are also highly depended on the learning procedure. For example, the SVM model required functional mapping and parameter selection. This study used radial basis function (RBF) as the kernel function for SVM, but there are various other kernels that can be employed for the classification task. The three methods used in the study is described in brief in this section.

### 3.1 Classification and Regression Tree (CART)

Machine learning involves two different methods viz. classification task and regression task. Classification is a process of grouping the data into distinct classes with similar characteristic. Evaluation of injury severity in vehicular crashes is a classification problem and thus non-parametric decision trees models can be implemented to extract relevant information from the traffic data. Along with modelling the function form of the traffic data, key determinants of the injury severity can be identified. These key parameters can provide necessary insight into the decision-making bodies and help them to make suitable traffic policies for reducing the casualties on roads and making roadways a safer mode of transport.

CART is the most widely applied classification approach in the study of injury severity. A procedure of repeated binary splitting of each node is adopted in the CART model (Chang and Chien, 2013). The splitting starts at the root node. The nodes generated are termed as child nodes. The nodes that are split to form child nodes are termed as parent nodes. This method involves recurrent splitting of the parent node in the tree till all the child nodes are homogeneous. At every successive interval, the child node acts as a parent node and the splitting is continued. The process is terminated when no further nodes split can be made. Every splitting is governed by a decision rule. The terminal nodes of the tree are often termed as leaf nodes. A maximal tree is generated on termination of the splitting procedure. The maximal tree has a large number of terminal nodes and hence it is not the optimal tree. CART method involves further steps; tree growing, tree pruning, and optimal tree-selection.

The tree-growing step involves repeated sectioning of the target variable to purify the terminal nodes. The objective of this step is to minimize the impurity in the leaf nodes. Gini index is a measure of impurity in a particular node of a tree. A node corresponding to the zero value of the Gini index has the least impurity. For a node \( m \), Gini Index is given as (Hastie et al., 2009):

\[
Gini(m) = \sum_{j=1}^{n} \hat{p}_{mj}(1 - \hat{p}_{mj})
\]

where,

\[
\hat{p}_{mj}: \text{proportions of class } j \text{ observations in node } m \\
\text{n: number of target variables/classes.}
\]

In CART models tree growth is continued until homogeneous observation are achieved. Thus, tree pruning is needed to avoid highly complex maximal trees. The branches of little importance to the classifier are removed in this step. The tree pruning method involves the
removal of branches from maximal trees leading to smaller and simpler trees. These smaller trees and fewer leaf nodes that in turn leads to a greater misclassification error rate. If the increase in misclassification error is comparatively less than the decrease in complexity of the pruned tree, the branch is eliminated and a new tree is formed (Kashani and Mohaymany, 2011). Tree pruning is implemented according to the cost-complexity algorithm.

\[
\text{Misclassification Error Rate} = \sum_{m=1}^{M} p(m) Gini(m) \]

where,

\[ p(m): \text{proportions of existing observations in the leaf node } m \]
\[ M: \text{number of terminal nodes} \]

In Figure 1., a hypothetical plot is presented to show the variation of error rate with the complexity of the maximal tree formed in tree growing step. It can be noted that with an increase in the complexity of the maximal tree the training error rate is continuously minimised and we see an exponentially decreasing function for the training error. The high complexity of the tree leads to develop a bias in the classifier for the training dataset. This bias causes the overfitting of the model for training dataset. It can be noted that after a particular complexity, the test error starts to rise again. This verifies the overfitting issue, which is a very common problem encountered in the machine learning algorithms. Tree pruning is an essential step to avoid overfitting. If we choose to compromise on the complexity of the model, it can also lead to under fitting. Thus, an optimal tree needs to be generated, for which the test error rate is at its minimum.

3.2 Extreme Gradient Boosting (XGBoost)

Boosting is a powerful learning method, which combines the multiple weak classifiers into a single powerful learner. Boosting was initially designed for classification problems but it can be also extended to regression. Multiple classifiers are iteratively generated from the weighted versions of the training dataset. Each weight is continuously updated at each step to give higher weight to the cases that had a higher misclassification error in the previous step. Boosting is a process of combining the classifiers generated from this iterative process. In CART we discussed the tree growing method through Gini Index. The boosted tree model is the combination of such trees. On weighing the result of the iteratively produced classifiers we produce our final predictions. The prediction function is described as (Friedman, 2001):

\[
\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i) \]
where,
\( t \): step count
\( \hat{y}_t \): predicted value at step \( t \)
\( f_t(x_i) \): learning function

Unlike the classical gradient descent algorithm, which in every iteration reduces the output error, gradient boosting predicts the optimal gradient for the additive model. After iterative computation of the gradient of the loss function, gradient boosting algorithms involves fitting the learning function \( f_t(x_i) \) on the gradient of the loss function. Usually, mean square error is used as the loss function for this algorithm. The mean square error loss function is described as:
\[
L(\theta) = \sum (y_i - \hat{y}_i)^2
\]
where,
\( \theta \): set of input parameters
\( y_i \): output label
\( \hat{y}_i \): predicted value

The optimization is done for the objective function, which is the sum of the regularization term and the loss function. The objective function in the XGBoost method at a given time step \( t \) is given as:
\[
ob_j^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \varphi(f_t) + c
\]
where,
\( l(y_i, \hat{y}_i) \): loss function
\( \varphi(f_t) \): regularization term
\( c \): constant

The regularization term used in XGBoost method helps to control the complexity of the tree, which in turn prevents the model from over-fitting. This gives XGBoost better performance over other decision tree approaches.

### 3.3 Support Vector Machines (SVM)

SVM is a supervised machine learning technique that can be used for classification and regression task. The SVM algorithm classifies the data into distinct classes by identifying a hyperplane in \( N \)-dimensional space (\( N \)-the number of features). The hyperplane is found by maximizing the margin between the data points of the classes on either side of the hyperplane. The hyperplane with maximum margin is termed as an optimal hyperplane.

Our training dataset is a set of \( n \)-dimensional vectors \( x_i \in \mathbb{R}^n \), where \( n \) is the number of features or crash related parameters in our data set. The optimal hyperplane is obtained by maximizing the margin to every nearest data point of each target class. The hyperplane that separates the data into several groups is described as:
\[
W \cdot X - b = 0
\]
where,
\( W \): normal vector to the hyperplane
\( X \): input vector
\( B \): bias

The SVM models need to solve the following optimization problem (Li et al. 2011):
\[
\min_{w,b,\xi} \left( \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \right) \quad \text{Subject to} \quad \xi_i \geq 0, \ y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i \ \forall i \quad (7)
\]

where,
- \(\xi_i\): slack variable calculating misclassification errors
- \(C\): cost parameter/penalty factor to the error
- \(y_i\): injury severity, the output of the training data

We describe a quadratic programming solution using Lagrange multipliers. Lagrange multipliers are used to further solve the minimizing function and also determine the \(C\) coefficient. The Lagrange function is defined as:

\[
L_p = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{n} \alpha_i [y_i (w^T \phi(x_i) + b) - (1 - \xi_i)] - \sum_{i=1}^{n} \beta_i \xi_i \quad (8)
\]

where,
- \(\alpha_i, \beta_i > 0\): Lagrange multipliers.

The equation 8 is minimised with respect to \(\phi(x_i), b\) and \(\xi_i\). The function \(\phi(x_i)\) maps the input vector \(x_i\) to a higher dimension space. SVM algorithm uses a set of mathematical function termed as a kernel function. A kernel is given by \(K(x_i, x_j) = \phi(x_i)^T \phi(x_j)\). The objective of the kernel is to transform the input data into the desired form. There are several kernels that are used by researchers, but the most widely used kernel for injury severity and safety study is the radial basis function (RBF) kernel. The RBF is described as:

\[
K(x_i, x_j) = e^{-\gamma ||x_i - x_j||^2}, \gamma > 0 \quad (9)
\]

where,
- \(\gamma\): kernel parameter

A globally optimal solution to \(W\) and \(b\) can always be calculated for input values of the parameter \((C, \gamma)\). Thus, the optimal hyperplane can be found that will distinctly classify into separate groups.

The discussed SVM approach is for a binary class classification problem and the classifier can be easily extended for multi-class classification problem. In this study, the SVM is implemented for 4-class injury severity and also for the binary classification problem. The test data set is used to estimate the accuracy of the SVM classifier. The accuracy is the proportion of correctly classified severity level.

4. CRASH DATABASE

Extracting information about the key factors influencing the severity of crashes can be very useful in designing the policies related to traffic management and safety. For the past few decades, the major gap in the studies related to crash severity in India was the unavailability of detailed crash records. The sheer size of India made it a very laborious and difficult task to keep records of traffic crashes but with the advent of technologies, there has been an increase in the concern for traffic safety and crash data collection. In this study, key factors influencing the severity of traffic crashes have been identified from a detailed accident dataset recorded by West Bengal police department.

4.1 Description of the Crash Data

With an increasing awareness of road safety, traffic police departments of many states in India are now recording details of the road traffic crashes, and subsequently, storing the information
in a dataset. The crash data was collected from the West Bengal Police Department on the five State Highways (SH-1, SH-2, SH-5, SH-6, and SH-11) covering over 1400 KM in the state of West Bengal in eastern India. The data were extracted for the period from January 2017 to December 2018. The dataset includes detailed information about each crash reported. The total dataset has information for different labels related to –

- Law – FIR number, Indian Penal Code (IPC) section, motor vehicle (MV) act section.
- Crash type – number of motorised vehicles involved, number of non-motorised vehicle involved, type of collision.
- Crash severity – number of crashes, number of fatalities, number of persons with major injury and minor injury.
- Crash location – district, police station, area type, road name, road number, landmark, latitude and longitude of the crash location
- Road Infrastructure at the crash location - road type, number of lanes, pedestrian facility, speed limit, type of traffic control.
- Crash time – Time of the day, month of the year, weather.

A summary of crash statistics on these five highways is presented in Table 1. When compared it was observed that SH-6 had experienced the highest share of fatal crashes followed by SH-5 and SH-11. However, SH-1 experienced a greater number of fatalities in person although it had a share of 32% fatal crash only.

<table>
<thead>
<tr>
<th>Road</th>
<th>Length (in KM)</th>
<th>Total Crashes</th>
<th>% of Fatal Crashes</th>
<th>Number of Fatalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>SH-1</td>
<td>118</td>
<td>526</td>
<td>32</td>
<td>181</td>
</tr>
<tr>
<td>SH-2</td>
<td>358</td>
<td>275</td>
<td>17</td>
<td>48</td>
</tr>
<tr>
<td>SH-5</td>
<td>411</td>
<td>211</td>
<td>52</td>
<td>126</td>
</tr>
<tr>
<td>SH-6</td>
<td>269</td>
<td>217</td>
<td>55</td>
<td>125</td>
</tr>
<tr>
<td>SH-11</td>
<td>272</td>
<td>353</td>
<td>51</td>
<td>255</td>
</tr>
</tbody>
</table>

### 4.2 Data Preparation

As mentioned earlier, the crash data had detailed information and was classified into several labels. However, a few labels such as reference number, FIR number, Indian Penal Code (IPC) section, motor vehicle (MV) act section, district, police station, road name, road number, landmark etc. were removed from the study dataset as they represent unique characteristic for a crash. Since all the data considered is for State Highways thus road type label was also removed from the dataset.

However, a new label of “Intersection proximity” was generated from the latitude and longitude of the crash location that can affect the severity of crashes. From the data of longitude and latitude, the distance of the crash location from the nearest intersection on the highway was manually extracted from Google Maps for each state highway. The intersection proximity is a binary class parameter. If a crash location is less than or equal to 100 m away from its nearest intersection, it is given a class of one and zero otherwise.

Out of all available parameters, a total of 10 predictor variables were used in this study with the qualitative target variable of crash severity level to identify the crucial patterns we wish to understand. These 10 predictor variables included - the number of the motorized vehicle
involved, number of the non-motorised vehicle involved, pedestrian facility, area type, weather condition, type of collision, number of lanes, speed limit, type of traffic control, and intersection proximity (Table 2).

### Table 2. Description of predictor variables for the cumulative dataset

<table>
<thead>
<tr>
<th>Sl.</th>
<th>Variable Type</th>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Discrete</td>
<td>Number of Motorised Vehicle Involved</td>
<td>Maximum: 7, Minimum: 0, Mean: 1.3, Standard Deviation: 0.51</td>
</tr>
<tr>
<td>2.</td>
<td>Discrete</td>
<td>Number of Non-motorised Vehicle Involved</td>
<td>Maximum: 2, Minimum: 0, Mean: 0.12, Standard Deviation: 0.33</td>
</tr>
<tr>
<td>4.</td>
<td>Nominal</td>
<td>Area Type</td>
<td>1. Urban, 2. Rural</td>
</tr>
<tr>
<td>7.</td>
<td>Nominal</td>
<td>Number of Lanes</td>
<td>1. 2 lanes or less, 2. More than 2 lanes</td>
</tr>
<tr>
<td>8.</td>
<td>Nominal</td>
<td>Speed Limit (in Kmph)</td>
<td>1. Less than 40, 2. 40 to 60, 3. 60 to 80, 4. More than 80, 5. No speed sign</td>
</tr>
<tr>
<td>10.</td>
<td>Nominal</td>
<td>Intersection Proximity</td>
<td>1. 100m or less, 2. More than 100 m</td>
</tr>
</tbody>
</table>

The crash severity analysis was conducted by utilising the 10 predictor variables for different sets of the target variable (Table 3). The first set of target variable includes four classes of injury severity: [Fatal, Major Injury, Minor Injury, and No Injury]. As per the police surveillance system a fatal crash is a crash that causes someone to die and a major injury results in amputation, skeletal injuries, burns, injuries to internal organs etc. However, minor injuries are mostly scratches, scrapes or cuts that in general do not need urgent hospitalisation.

Based on the police data, four class of injury severity was formed: Fatal (if number of fatalities > 0), Major Injury (if number of fatalities = 0 and number of major injury >0), Minor Injury (if number of fatalities = 0, number of major injury = 0 and number of minor injury > 0) and No Injury (if number of fatalities = 0, number of major injury = 0 and number of minor
injury = 0). The second classification output is a binary set of injury severity (Figure 2.) The first class comprises of crashes resulting in urgent hospitalization (i.e. the combined data set of Fatal and Grievous Injury [1] and the second class comprises of crashes not needing any hospitalized in an urgent need (i.e., the combined set off Minor Injury and Non- Injury [0]). The main motive to conduct the binary set classification is mainly to address the underreporting of the minor and no injury data.

<table>
<thead>
<tr>
<th>Road</th>
<th>Four-Class Model</th>
<th>Two-Class Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fatal</td>
<td>Major Injury</td>
</tr>
<tr>
<td>SH-1</td>
<td>169</td>
<td>245</td>
</tr>
<tr>
<td>SH-2</td>
<td>46</td>
<td>172</td>
</tr>
<tr>
<td>SH-5</td>
<td>109</td>
<td>88</td>
</tr>
<tr>
<td>SH-6</td>
<td>120</td>
<td>79</td>
</tr>
<tr>
<td>SH-11</td>
<td>179</td>
<td>149</td>
</tr>
</tbody>
</table>

Figure 2. Representation of (a) four-class and (b) two-class (binary) model configurations
5. RESULTS AND DISCUSSION

Three non-parametric models viz. CART, XGBoost and SVM were used to analyse the crash data. A sample CART result is shown in Figure 3. At the initial stage, all the three non-parametric approaches were applied on the separate dataset from each of the five state highways. Thereafter all the five datasets were combined and each approach was trained and tested for this combined dataset. The aim to combine the data set was to identify the key important safety deficiencies which were common in all the State Highways.

The model’s performance between the two datasets was compared by the overall prediction accuracy of the classifier. In general, accuracy is a metric for evaluating the
performance of the classification model and is described as the ratio of the number of correct injury severity prediction and the total number of predictions. In order to evaluate the efficiency of data classification, the performance of the proposed three machine learning methods, a logistic regression model was developed as the baseline model. Table 4 lists down the prediction accuracies of each classifier for both 4-class labels and binary classification on each dataset. The accuracy obtained against the 4-class classifier was lower in comparison to the binary classifier for all the datasets. This could be because of the erratic reporting, as the police do not always enquire about the hospital stay, so often there is a misclassification of fatal, major and minor injury. This signifies that in case of low-quality crash data instead of 4-class severity classification a binary classification can provide better predictive models.

Table 4. Accuracy check for different learning paradigms

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CART Accuracy (%)</th>
<th>XGBoost Accuracy (%)</th>
<th>SVM Accuracy (%)</th>
<th>Logistic Regression Model Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4-class</td>
<td>Binary</td>
<td>4-class</td>
<td>Binary</td>
</tr>
<tr>
<td>SH-1</td>
<td>50.01</td>
<td>75.31</td>
<td>53.44</td>
<td>83.62</td>
</tr>
<tr>
<td>SH-2</td>
<td>67.86</td>
<td>89.39</td>
<td>48.45</td>
<td>89.69</td>
</tr>
<tr>
<td>SH-5</td>
<td>57.14</td>
<td>93.7</td>
<td>44.68</td>
<td>97.87</td>
</tr>
<tr>
<td>SH-6</td>
<td>58.33</td>
<td>90.76</td>
<td>58.46</td>
<td>95.83</td>
</tr>
<tr>
<td>SH-11</td>
<td>50.04</td>
<td>90.47</td>
<td>56.41</td>
<td>92.33</td>
</tr>
<tr>
<td>All data</td>
<td>43.66</td>
<td>83.08</td>
<td>51.83</td>
<td>89.97</td>
</tr>
</tbody>
</table>

To determine the models’ prediction performance in forecasting the injury severity, the five-fold cross validation was used (Meng et al. 2018) for binary classifier dataset (Table 5). In five-fold cross validation, the sample was divided into five parts and for each run, one part was selected as test set, and the rest for the training set. The model was trained and tested for a total of five times. The average accuracy of five results was recorded as the performance of the model. When compared across the highways it was found that XGBoost performed better than the other models.

Table 5 Prediction Performance for different learning paradigms

<table>
<thead>
<tr>
<th>Model</th>
<th>SH 1</th>
<th>SH 2</th>
<th>SH 5</th>
<th>SH 6</th>
<th>SH 11</th>
<th>All data</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>0.76</td>
<td>0.89</td>
<td>0.92</td>
<td>0.88</td>
<td>0.88</td>
<td>0.82</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.84</td>
<td>0.88</td>
<td>0.96</td>
<td>0.95</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>SVM</td>
<td>0.72</td>
<td>0.85</td>
<td>0.89</td>
<td>0.91</td>
<td>0.95</td>
<td>0.88</td>
</tr>
</tbody>
</table>

So, the results from the XGBoost were only considered for the identifying critical safety factors. The results obtained from the XGBoost classification on the cumulative dataset for binary labels identified *intersection proximity, pedestrian facility, type of traffic control, type of collision, weather conditions and speed limit* as the key factors influencing the severity of injury in crashes in decreasing order of relevance. Whereas for the same model, the key factors identified for the 4-way classification on the cumulative dataset are *intersection proximity, number of lanes, pedestrian facility, type of traffic control, speed limit, type of collision and weather conditions* in decreasing order of relevance (Table 6). Although the identified factors
Table 6. Relative importance of variables and their feature importance scores (XGBoost)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Feature importance Score*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4-class classifier</td>
</tr>
<tr>
<td>Intersection proximity</td>
<td>0.077</td>
</tr>
<tr>
<td>Pedestrian facility</td>
<td>0.030</td>
</tr>
<tr>
<td>Type of traffic control</td>
<td>0.025</td>
</tr>
<tr>
<td>Type of Collision</td>
<td>0.012</td>
</tr>
<tr>
<td>Weather Condition</td>
<td>0.009</td>
</tr>
<tr>
<td>Speed limit</td>
<td>0.013</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>0.034</td>
</tr>
</tbody>
</table>

*For each variable, the feature importance score is the average score over its every value.

The results obtained depicts that, there is a significant lack of infrastructural facilities leading to severe crashes on the State Highways. Other than the road infrastructural parameters only weather and type of collision were found to have a significant impact on the crash severity. The study shows that the pedestrian facility is one of the most critical factors that influence the severity of injury in traffic crashes. Although most of such highways pass through rural areas often, they connect district capitals and important villages. However, inadequate infrastructures at such locations expose the pedestrians to high-speed motorized traffic resulting in severe crashes. Another important observation is that intersection proximity is also a major contributor to the severity level. This is mainly because in most of the cases major intersections on these highways have restricted sight distance due to encroachments and even due to vegetation. Moreover, it is very common to find vehicles overtaking, taking wrong turns and over speeding near the intersection. Lack of proper traffic control at such locations also aggregates the risk of a severe crash. Similarly, the lack of appropriate speed limits near built-up areas, schools, and even horizontal curves can result in a significant number of fatal and injury crashes.

6. CONCLUSION

The aim of the study was to model a functional relationship between the injury severity classes and the different label that govern the cause of these crashes. In addition, different classification models were explored to accurately predict the injury severity class and further identify the key parameters to understand the dependency of the severity of injury with factors such as environmental condition, roadway, vehicular character and interaction of motorized traffic with pedestrians and non-motor vehicles. Various useful information was obtained from the datasets using data mining techniques. The findings from this study can guide decision-making authorities to make better policies for preventing severe crashes on the State Highways in India. In this study, three different non-parametric approaches were implemented to model the severity level of injuries taking place in crashes on Indian State Highways. The performance of each model was estimated and from the study and it can be deduced that XGBoost was the overall better performer for identifying factors influencing severity in the crash data. The results showed that intersection proximity, pedestrian facility, type of traffic control, weather conditions, type of collisions, number of lanes and speed limit are significant contributors to severe crashes. All three approaches used in the study have provided a reasonable estimate of the severity level in traffic crashes. These approaches capture the underlying relationship
between the injury severity and the factors affecting severity. The decision-making authorities can use the knowledge of these factors to make relevant traffic policies and safety plan to reduce the fatalities and injuries on the roads.

From the study, it emerged that, India being a developing country, lacks many basic infrastructural facilities. From the result, it is evident that there is a serious need for pedestrian facilities like subway, FOB, sidewalks, pedestrian traffic lights, and others. Moreover, lack of proper traffic control had led to an increase in wrong turns, overtaking and over speeding near the intersections. Another interesting observation was that the absence of appropriate speed limit was also one of the key determinants that influenced the severity of the crashes. This highlights that there are several stretches on these highways where there is a need to control speed.

The insights from the study can help the concerned authorities in decision-making policies and establishing safety plans. This study can help in prioritizing the critical issues common to the crash prone locations on the State highways and can help in designing specific countermeasures based on the distinctive safety deficiencies. As for example, from the results it can be suggested that, law enforcement needs to be intensified for the adherence to traffic rules near the intersections. In this context, CCTV cameras can be installed at the major intersections to detect over speeding, unsafe overtaking and lane changing. Similarly, providing adequate pedestrian facilities near the built-up areas can also help to cut down the severe crash counts. Even low-cost intelligent transportation system (ITS) solutions can be designed to warn vehicles of the presence of pedestrians on the crosswalks on highways. This study also highlights that over speeding should be controlled by providing traffic calming measure or posted speed limit sign, which can actually reduce the proportion of severe accident on such highways.

In terms of future work, further exploration by other non-parametric approaches such as non-parametric Bayesian network and artificial neural network can provide an unknown insight into the dependency relation between injury severity and factors affecting the severity level. From this, it can be concluded that the non-parametric approaches are good techniques, which can be engaged in the studies in the field of transportation and safety.

REFERENCES


