

APPLICATION OF MULTIPLE LOGISTIC REGRESSION, BAYESIAN LOGISTIC AND CLASSIFICATION TREE TO IDENTIFY THE SIGNIFICANT FACTORS INFLUENCING CRASH SEVERITY

MILAD TAZIK¹, IMAN AGHAYAN^{2,*}, MOHSEN SADEGHI²

¹Department of Civil Engineering, Shahrood Branch, Islamic Azad University, Shahrood, Iran

²Department of Civil Engineering, Shahrood University of Technology, Shahrood, Iran

*Corresponding Author: iman.aghayan@shahroodut.ac.ir

Abstract

Identifying cases in which road crashes result in fatality or injury of drivers may help improve their safety. In this study, datasets of crashes happened in Tehran-Qom freeway, Iran, were examined by three models (multiple logistic regression, Bayesian logistic and classification tree) to analyse the contribution of several variables to fatal accidents. For multiple logistic regression and Bayesian logistic models, the odds ratio was calculated for each variable. The model which best suited the identification of accident severity was determined based on AIC and DIC criteria. Based on the results of these two models, rollover crashes (OR = 14.58, %95 CI: 6.8-28.6), not using of seat belt (OR = 5.79, %95 CI: 3.1-9.9), exceeding speed limits (OR = 4.02, %95 CI: 1.8-7.9) and being female (OR = 2.91, %95 CI: 1.1-6.1) were the most important factors in fatalities of drivers. In addition, the results of the classification tree model have verified the findings of the other models.

Keywords: Bayesian logistic, Classification tree, Injury severity, Multiple logistic regression, Traffic crash.

1. Introduction

A growing trend in road traffic crashes is considered as a serious problem for public health across the world. Injuries caused by traffic crashes are the leading cause of death, disability, hospitalization as well as economic value in the world [1]. Road traffic crashes are associated with a variety of factors such as the large population of young individuals in the country which increases the exposure to these incidences, low price of fuel, decreased rates of using public transportation system, preferring

Nomenclatures

| | |
|-------|---|
| A | An attribute |
| e | Euler's number |
| g | Logit transformation |
| p_i | Probability of the outcome of i in training data |
| S | Set of examples |
| S_V | Subset of S for which attribute A has value V |
| var | Variance |
| x | Independent variable |
| Y | Response variable |

Greek Symbols

| | |
|-----------|-----------------|
| β_i | Model parameter |
|-----------|-----------------|

Abbreviations

| | |
|------|------------------------------------|
| ADT | Average Daily Traffic |
| AIC | Akaike Information Criterion |
| BL | Bayesian Logistic |
| CART | Classification and Regression Tree |
| CI | Confidence Interval |
| DIC | Deviance Information Criterion |
| EMA | Expectation Maximization Algorithm |
| F | Frequency |
| FARS | Fatal Accident Reporting System |
| MCMC | Markov Chain Monte Carlo |
| MLE | Maximum Likelihood Estimation |
| MLR | Multiple Logistic Regression |
| OR | Odds Ratio |
| SE | Standard Error |

personal vehicles and non-standard designs of cars in terms of safety. Road traffic crashes are the third cause of fatalities in Iran [2]. The rate of injury/death in the world is three people per 10,000 vehicles while it is 33 people per 10,000 vehicles in Iran [3]. The proper management of the main risk factors leads to the reduction of the frequency and severity of injuries resulting from traffic crashes.

Many statistical and mathematical methods have been employed to identify factors influencing the level of drivers' injuries in crashes. Some of these models have focused on identifying the most important factors influencing and predicting the injury severity level. Some simple methods such as cross-tabulation method and χ^2 -tests were applied to compare the distribution of different levels of injury and various groups (e.g., age and gender) or risk factors (e.g., crash location, time of crash, and type of collision) [4, 5].

Lui and McGee [6] used the logistic regression model to analyse the probability of death in crashes via datasets obtained from the Fatal Accident Reporting System (FARS). The death probability was considered as the output variable while the age and gender of driver, the directions of crash impacts, car deformation, use of seat belt and vehicle weight were input parameters of model. The results of this study revealed that occupants of lighter vehicles were at higher risks of injury. Moreover,

the risk of injury in men was lower than women and each 10-year increase in age enhanced the risk of death as much as 1.6 times.

Results of another study using a multivariate logistic regression model indicated that the OR of a fatal outcome of a crash increased with age. This ratio in drivers aged over 80 years old reached 4.98 compared to drivers aged between 40 and 49. Moreover, another finding of this study was the increase of death OR in female drivers and vehicles which had side impacts compared to vehicles which had frontal impacts [7]. The study of the factors influencing crashes of Riyadh city, in Saudi Arabia using the logistic regression model that its output specified two levels of fatal and non-fatal crashes showed two variables significant among input variables. The location of crash: for example, the OR of death in crashes at non-intersection location was 2.64 higher than those at intersection. Another significant variable was speeding [8]. The modelling results of crash data in one of the provinces in Turkey using multinomial logit model revealed that factors such as age over 65, drivers with primary school education and single-vehicle crashes increased the probability of fatality in a crash [9].

Bayesian network models can be used as a classifier for crashes based on the level of injuries to determine the relationship between injuries resulting from crashes and variables related to driver, vehicle and road conditions [10]. Mujalli and de Oña [11] presented a method based on Bayesian networks and variable selection algorithm to predict the level of crash injury severity. 1536 crash data in suburban roads of Spain were used which contained eighteen input variables of each crash occurrence. After examining the different subsets of variables, it was concluded that the number of variables used in a Bayesian network can be reduced for modeling injury severity level, without reducing performance of the model. Seven variables including crash type, driver's age, driver's gender, atmospheric factors, lighting conditions, occupants involved and the number of injuries were recognized as the most important factors influencing crashes. In 2015, Bayes classifiers were used to analyze imbalanced traffic crashes datasets in Jordan and the influencing parameters were identified. The important points were using different methods of balancing data and comparing three different classification methods [12].

Kwon et al. [13] applied two methods of classification, the decision tree classifiers and Naive Bayes for ranking the risk factors. Applying non-parametric classification tree techniques on the crash data of Taipei, the capital of Taiwan, indicated that the vehicle type (e.g., cars and trucks) was the most important factor in the injury severity level (i.e., fatality, injury and non-injury) [14]. Kashani and Mohaymany [15] in a study on the level of crashes injury severity on two-lane, two-way rural roads in Iran using classification and regression tree (CART) concluded that improper overtaking and not using a seat belt were the most influential factors among input variables.

Akgüngör and Yıldız [16] analysed the sensitivity of a crash prediction model by the fractional factorial method. The reason for choosing this model was both traffic and road geometry parameters in addition to the terrain characteristics among the data. The evaluation of the sensitivity analysis revealed that the average daily traffic (ADT), lane width, width of paved shoulder, median and interaction (e.g., ADT-W, ADT-PA and ADT-H) had a considerable effect on the number of

crashes. Based on the absolute value of parameter effects at the three- and two-standard deviation thresholds, ADT was found to be of primary importance.

This study primarily intends to compare three models of multiple logistic regression, Bayesian logistic and classification tree to identify the most important factors influencing crash injury severity. The input variables of these three models include driver's gender, driver's age, lighting conditions, road surface, collision type, crash type, use of seat belt, sleepiness and tiredness and exceeding the speed limits. The output variables include injury in three levels of no injury, injury and fatality. For the multiple logistic regression and Bayesian logistic models, the OR of each variable was calculated and the probability model of fatal crashes was determined. A CART model was employed to determine the relationship between fatality of drivers and driver/vehicle characteristics, highway/environmental variables and crash variables. The classification accuracy of CART model in determining fatal crashes was also specified.

2. Data Description

The dataset used in this study consisted of 912 traffic crashes derived from traffic crash reports between 2013 and 2014 in Tehran-Qom freeway, Iran. The datasets related to the injury severity levels have been divided into the three levels of no injury, injury and fatality. Table 1 presents the independent variables used in the models with their frequency distribution. This variable was normalized in the range of [0,1]. In addition, for ease of interpretations, one reference class was considered for each variable in multiple logistic regression and Bayesian logistics. In CART model, normalization procedure has been applied on all variables.

Table 1. The frequency of the studied independent and dependent variables.

| Crash conditioning variables | Type | Description | F | Coding (CART) | Coding (MLR and BL) |
|------------------------------|-------------|------------------------|-----|---------------|---------------------|
| Driver's gender | Qualitative | Female | 53 | 1 | 1 |
| | | Male | 859 | 0 | 0 |
| Light conditions | Qualitative | Night | 295 | 1 | 1 |
| | | Day | 617 | 0 | 0 |
| Road surface | Qualitative | Wet | 66 | 1 | 1 |
| | | Dry | 846 | 0 | 0 |
| Use of seat belt | Qualitative | Not in use | 199 | 1 | 1 |
| | | In use | 713 | 0 | 0 |
| Type of vehicle | Qualitative | Passenger cars | 714 | 0 | (1,0,0) |
| | | Pickup | 133 | 1 | (0,1,0) |
| | | Bus/large trucks | 65 | 2 | (0,0,1) |
| Exceeding speed limits | Qualitative | Yes | 119 | 1 | 1 |
| | | No | 793 | 0 | 0 |
| Sleepiness and tiredness | Qualitative | Yes | 390 | 1 | 1 |
| | | No | 522 | 0 | 0 |
| Crash type | Qualitative | Rollover | 164 | 0 | (1,0,0,0) |
| | | With vehicle | 282 | 1 | (0,1,0,0) |
| | | With multiple vehicles | 378 | 2 | (0,0,1,0) |
| | | With fixed | 88 | 3 | (0,0,0,1) |

| | | object | | | |
|------------------------|-------------|---------------------------------------|-----|---|-----------|
| Collision type | Qualitative | Sideswipe (with fixed object) | 152 | 0 | (1,0,0,0) |
| | | Rear-end | 557 | 1 | (0,1,0,0) |
| | | Left-turn head-on/ right-turn head on | 134 | 2 | (0,0,1,0) |
| | | Sideswipe | 69 | 3 | (0,0,0,1) |
| Driver injury severity | Qualitative | Fatality | 63 | 0 | (1,0,0) |
| | | Injury | 168 | 1 | (0,1,0) |
| | | No-injury | 681 | 2 | (0,0,1) |

3. Methods

3.1. Multiple logistic regression model

The aim of an analysis using logistic regression is the same as that of any model-building technique used in statistics: to find the best fitting and parsimonious. What distinguishes a logistic regression model from a linear regression model is the response variable. In the logistic regression model, the response variable is binary or dichotomous. The expected value of Y , given the value x is defined as Eq. (1):

$$E(Y | x) = \beta_0 + \beta_1 x \quad (1)$$

where Y and x are the response variable and the independent variable, respectively and β_i values denote the model parameters. Many distribution functions have been proposed to be used in the analysis of a dichotomous response variable. The specific form of the logistic regression model is as follow:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (2)$$

Logistic pattern can be directly generalized to situations where there are several predictor variables.

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}} \quad (3)$$

where to simplify the notation $\pi(x) = E(Y | x)$. The transformation of the $\pi(x)$ logistic function is known as the logit transformation.

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (4)$$

In logistic regression model, variables can be selected through three ways (i.e., forward selection, backward elimination and stepwise entry) to enter the model.

Logistic slope coefficients can be interpreted as the change in the logit related to a change of one unit in the independent variable. The exponent of difference between two logits gives the odds ratio. The OR is the ratio of odds that an independent variable will be present to the odds that it will not be present. This illustrates how the logit serves as a link function between the probability and the

linear regression expression. For a continuous independent variable, the OR can be defined as:

$$OR = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1x}} = e^{\beta_1} \quad (5)$$

In the multiple logistic regression model, the model fit has been performed based on the stepwise procedures. In addition, the OR for each of the risk factors have been identified, in this study.

3.2. Bayesian logistic

In classical statistics, the analysis of the logistic regression model is based on estimating parameters through Maximum Likelihood Estimation (MLE) and calculating an estimate using Expectation Maximization Algorithm (EMA). In such method, it is possible to evaluate local maximum instead of general MLE without calculating the convergence of the model. Furthermore, given the asymptotic properties, MLE in small samples encounters with serious inferential problems. Therefore, in Bayesian method, inference about the model parameters is conducted on the basis of their posterior distribution which is the combination of observed data and information from previous studies or personal experiences that are known as prior distribution. In the case of uncertainty of posterior distribution, Markov Chain Monte Carlo (MCMC) simulation methods are used. This method allows a detailed inference from parameters and can be obtained for any arbitrary sample size.

The Bayes model formula is almost the same as Eq. (2). However, the difference between them is due to the absence of a particular idea about prior distributions. This distribution can be considered non-informative for betas. Moreover, the prior distribution parameters of these parameters were considered normal distribution with mean value 0 and three variances of 100, 10 and 1, in this study. The reason for selecting these three variances is because in Bayesian logistic, datasets should be examined at different modes of backgrounds. The large variance was applied for the first two modes (i.e., variances of 100 and 10), in fact, the model was run by considering datasets without background. In the third mode (variance 1), for prior distributions, more information about the parameters was presented to the model.

3.3. Classification and regression tree (CART)

Regression models have their own assumptions and predefined underlying relationships between independent and dependent variables. If these assumptions are violated, the model can lead to erroneous estimations of the injury severity. The CART models are non-parametric models without any pre-defined underlying relationship between the response variable (dependent variable) and predictors (independent variables). CART models are useful means for prediction and classification issues. The classification tree is applied for issues which are discrete response variable. Regression tree is used when the response variable is

continuous. Since the purpose of this study was to model the driver injury severity levels in traffic crashes regarding to the discrete output (i.e., fatality and non-fatality). Thus, the classification tree was employed.

The concept called Information Gain is used to choose the best attribute for being located at the root of the tree. An attribute with the Information Gain (information which is more useful for classification) is selected. To introduce the Information Gain, first the idea of entropy is presented.

$$Entropy(Training\ table) = - \sum_{i \in outcomes} p_i \log_2 p_i \quad (6)$$

p_i =Probability of the outcome of i in training data (The number of items in the training data with outcome= i)/ (The number of all items in the training data).

Entropy makes decision about the degree of randomness in the data and measures the amount of information available in an attribute. The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches). In other words, the Information Gain (S, A) for an attribute such as A is relative to a set of examples S defined as follows:

$$Gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (7)$$

where values (A) is a set of all possible values for attribute A and S_v is a subset of S for which attribute A has value V . In the above definition, the first term is the entropy value of data and the second is the expected entropy value after splitting data.

4. Modelling and Results

4.1. Multiple logistic regression and Bayesian logistic

To achieve the best model which identifies the most important variable in associated with fatal crashes, firstly, the two multiple logistic regression models with forward and backward stepwise methods were fitted to the data. Secondly, the 95% confidence interval (CI) has been used to investigate the significance of variables in each method. In other words, the significant level of 0.05 was considered. The results obtained from OR and CIs of fatal crash based on logistic regression model and Bayesian logistic are presented in Table 2.

With regard to the sparse data and large maximum likelihood estimates in the two mentioned models, in Bayesian logistic model with the previous normal functions for the model parameters were fitted to the data with the presence of such variables. In this model, the variance of prior distributions was considered 100, 10 and 1. In the fit of each three Bayesian logistics, different initial values were given to model to evaluate the convergence of the algorithms. The results were very similar which represents the convergence of the algorithm. A reference class was considered for each variable to measure the variables in terms of the OR and/or the amount of risk taking. For example, not using seat belt could increase the risk of death.

Table 2. The OR values of fatal crashes factors based on the multiple logistic regression and Bayesian logistic models.

| Variables | Type | BL (var=1) OR (95%CI) | BL (var=10) OR (95%CI) | BL (var=100) OR (95%CI) | MLR OR (95%CI) |
|--------------------------|--|--------------------------------|---------------------------------|----------------------------------|----------------------|
| Driver's gender | Female | 2.91 (1.1,6.1) | 3.96 (1.2,9.1) | 4.13 (1.3,9.5) | 3.68 (1.4,9.4) |
| | Male | Reference | Reference | Reference | Reference |
| Use of seat belt | Not in use | 5.79 (3.1,9.9) | 7.08 (3.6,12.7) | 7.19 (3.7,12.8) | 6.60 (3.5,12.2) |
| | In use | Reference | Reference | Reference | Reference |
| Exceeding speed limits | Yes | 4.02 (1.8,7.9) | 5.42 (2.1,11.7) | 5.71 (2.3,12.1) | 5.01 (2.2,11.4) |
| | No | Reference | Reference | Reference | Reference |
| Sleepiness and tiredness | Yes | 0.98 (0.5,1.7) | 1.08 (0.5,2.0) | 1.09 (0.5,2.1) | 1.03 (0.5,2.0) |
| | No | Reference | Reference | Reference | Reference |
| Crash type | Rollover | 14.58 (6.8,28.6) | 26.6 (9.7,61.7) | 30.1 (10.7,75.4) | 23.65 (9.3,59.9) |
| | With fixed object | 1.63 (0.4,4.3) | 3.06 (0.4,9.8) | 3.52 (0.4,11.7) | 2.77 (0.6,11.9) |
| | With vehicle | 1.3 (0.5,3) | 2.29 (0.6,6.0) | 2.51 (0.7,6.4) | 2.06 (0.6,6.1) |
| | With multiple vehicles | Reference | Reference | Reference | Reference |
| Light conditions | Night | 0.58 (0.5,1.1) | 0.65 (0.3,1.0) | 0.34 (0.6,1.0) | 0.96 (0.1,2.0) |
| | Day | Reference | Reference | Reference | Reference |
| Road surface | Wet | 1.89 (0.4,2) | 2.24 (0.4,4.6) | 2.26 (0.3,2.9) | 2.78 (0.6,2.5) |
| | Dry | Reference | Reference | Reference | Reference |
| Type of vehicle | Bus/ large trucks | 0.61 (0.3,4.6) | 1.62 (0.1,5.4) | 1.45 (0.4,5.7) | 1.54 (0.7,7.9) |
| | Pickup | 1.78 (0.4,4.2) | 2.96 (0.4,9.8) | 2.81 (0.4,6.7) | 2.84 (0.6,8.9) |
| | Passenger cars | Reference | Reference | Reference | Reference |
| Collision type | Rear end | 0.89 (0.5,1.6) | 1.67 (0.4,1.0) | 1.41 (0.6,2.0) | 1.45 (0.7,3.0) |
| | Left-turn head-on/ right-turn head-on | 1.32 (0.4,3) | 1.92 (0.7,6.0) | 2.24 (0.3,6.6) | 2.65 (0.6,5.6) |
| | Sideswipe | 1.89 (0.6,3) | 2.24 (0.6,6.6) | 2.26 (0.7,3.5) | 2.78 (0.6,4.9) |
| | Sideswipe (with fixed object) | Reference | Reference | Reference | Reference |
| Driver's age | Year | 0.45 (0.1,1.5) | 1.15 (0.3,1.1) | 1.95 (0.4,2.2) | 1.58 (0.7,3.9) |

As can be seen from Table 2, the significance and prioritization of factors were similar in all models. The Akaike information criterion (AIC) and deviance information criterion (DIC) were calculated to find the best-fitted model. According to the results, the multiple logistic regression model based on stepwise method had the lowest AIC and was very close to the calculated values in Bayesian logistics (see Table 3). With regard to these criteria, it can be concluded that Bayesian logistic with variance 1 was more accurate than the other three models.

Table 3. The results of the fitted models based on two criteria (AIC and DIC).

| Criteria | Bayesian logistic (var=1) | Bayesian logistic (var=10) | Bayesian logistic (var=100) | Multiple logistic regression |
|-----------------|----------------------------------|-----------------------------------|------------------------------------|-------------------------------------|
| AIC | 325.7 | 324.5 | 324.5 | 316.2 |
| DIC | 316.5 | 316.3 | 316.4 | N/A |

Based on the obtained results in multiple logistic regression and Bayesian logistic models, rollover crash (OR= 14.58, CI: 6.8-28.6), not using seat belt (OR= 5.79, CI: 3.1-9.9), exceeding of speed limits (OR= 4.02, CI: 1.8-7.9) and being female (OR= 2.91, CI: 1.1-6.1) were the most significant risk factors for traffic crashes, respectively. So that, rollover crash had the highest impact on deaths caused by crashes based on the Bayesian logistic with variance 1 (OR= 14.58).

It should be noted that the sleepiness and tiredness variable had not become significant in none of the models. In other words, this variable had no significant effect on driver's death. Overall, because of the effect of rollover on the probability of fatality, there is also a policy implication that the roadway and roadside design community should continue to build forgiving roadsides that do not produce rollover (e.g., clear roadsides, less severe side slopes).

According to the outcomes of these two models, not using seat belt was one of the main risk factors contributing to the increase in crash injuries. This result was also pointed out in some previous studies [17, 18]. Regarding to the important role of human risk factors in traffic crashes, struggling efforts are required to control dangerous driving behaviors such as exceeding speed limit, which has been identified in relation to traffic crashes and injuries [19]. In addition, being female was a significant factor for fatal outcome due to car crashes, which is consistent with previous reports [20].

After determining the OR values and identifying the most important risk factors, the probability model of fatal crashes has been determined. As shown in Table 4, the coefficient (β) and standard error (SE) of each of the significant factors were estimated. Thus, the Bayesian logistic model developed in this study is as Eq. (8). The validity of the model was checked by considering the statistical level of significant for its coefficients.

$$\pi_i = \frac{\exp(4.72 + 1.715 \text{seatbelt}_i + 1.325 \text{exceeding speed limits}_i + 0.975 \text{gender}_i + 2.612 \text{rollover}_i)}{1 + \exp(4.72 + 1.715 \text{seatbelt}_i + 1.325 \text{exceeding speed limits}_i + 0.975 \text{gender}_i + 2.612 \text{rollover}_i)} \quad (8)$$

Table 4. Estimated coefficients and estimated standard errors for the model variables.

| Crash conditioning variables | Type | BL (var=1) | BL (var=10) | BL (var=100) | MLR |
|------------------------------|------------------------|----------------|----------------|-----------------|----------------|
| | | β (SE) | β (SE) | β (SE) | β (SE) |
| Driver's gender | Female | 0.97 (0.43) | 1.25 (0.49) | 1.29 (0.49) | 1.30 (0.48) |
| | Male | Reference | Reference | Reference | Reference |
| Use of seat belt | Not in use | 1.71 (0.29) | 1.90 (0.32) | 1.92 (0.31) | 1.88 (0.31) |
| | In use | Reference | Reference | Reference | Reference |
| Exceeding speed limits | Yes | 1.32 (0.36) | 1.60 (0.42) | 1.65 (0.41) | 1.61 (0.41) |
| | No | Reference | Reference | Reference | Reference |
| Crash type | Rollover | 2.61 (0.36) | 3.16 (0.47) | 3.28 (0.48) | 3.16 (0.47) |
| | With fixed object | 0.31 (0.59) | 0.83 (0.77) | 0.95 (0.80) | 1.02 (0.75) |
| | With vehicle | 0.23 (0.44) | 0.67 (0.55) | 0.75 (0.56) | 0.72 (0.55) |
| | With multiple vehicles | Reference | Reference | Reference | Reference |

4.2. Classification tree

One of the useful properties of classification tree in comparison to other modelling methods is that it gives decision makers some rules to answer “if-then” questions efficiently. A decision tree of fatal crashes displays in Fig. 1. The tree has ten terminal nodes and can be easily seen that crash type, using of seat belt were the basic splitters in the classification tree. The initial split at node 1 was based on the variable of crash type. Classification tree directed the rollover of car to the left forming node 2; the rest of crash types to the right, forming node 5. Classification tree further split node 2 based on the variable of using seat belt and directs whether the driver used seat belt or not. Not using of seat belt to the right, forming terminal node 1; the tree predicts the driver involved in this condition was more likely to death (56%).

In addition to the risk factors identified by multiple logistic regression and Bayesian logistic, the CART showed that age was a main risk factor in fatality. In Fig. 1, the right side of the CART shows that not using seat belt, exceeding speed limits, driver's gender and driver's age were correlated with each other. Considering the age variable, the CART basically categorized the data into three age groups: 19

to 33 years, 34 to 36 and 37 to 79. The female drivers who were between 34 and 79 years in Terminal Node 9 had the largest probability of fatality (Fatality=1), while the probability of fatality was 0.05 for men in the same age. This result indicated, in the same conditions, women were more vulnerable to death than men.

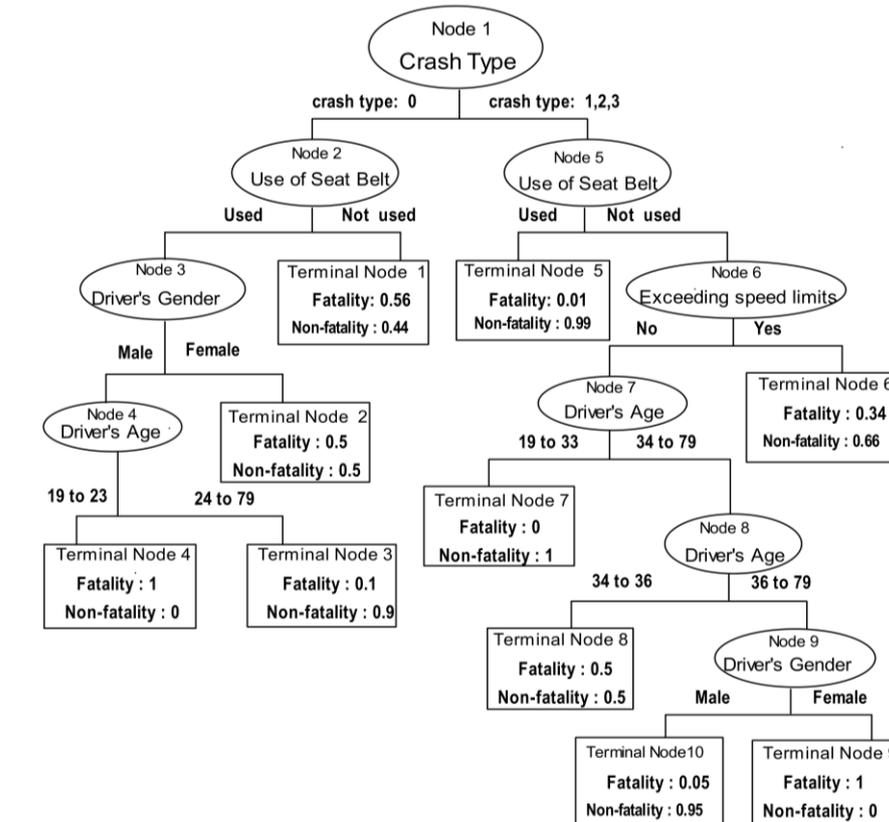


Fig. 1. The output of CART.

This pattern could be used for other conditions to obtain the probability of fatality in the crash from the tree. In summary, the obtained results revealed that crash type, using seat belt, exceeding speed limits and driver's gender were the main causes of the fatalities. In classification modelling, 70% of the data were used for training. Testing of the model for 30% of all observations showed that the model has had the accuracy of 88%.

5. Conclusions

This study compared classification tree with multiple logistic regression and Bayesian logistic to identify the most important factors contributing to death of drivers caused by crashes in Tehran-Qom freeway, Iran.

The multiple logistic regression and Bayesian logistics showed that rollover crash, not using seat belt and exceeding speed limits were the most significant risk factors for traffic crashes in Iran. Thus, rollover crash had the highest impact on

deaths from traffic crashes based on the best-fitted model that was the Bayesian logistic with variance 1 (OR= 14.58). This ratio for not using seat belt, exceeding speed limits and being female was 5.79 (CI: 3.1-9.9), 4.02 (CI: 1.8-7.9) and 2.91 (CI:1.1, 6.1), respectively. The classification tree model identified that crash type, use of seat belt, driver's gender and exceeding speed limits influenced the fatality of a driver in the crashes. This finding was consistent with the results obtained from multiple logistic regression and Bayesian logistic models. Therefore, classification tree also provided a good overall classification for testing data.

The results obtained from the three models indicated that by controlling human error such as exceeding speed limits and not using of seat belt decreases the probability of driver's fatality in crashes.

References

1. Peden, M.; Scurfield, R.; Sleet, D.; Mohan, D.; Hyder, A.A.; Jarawan, E.; and Mathers, C. (2004). *The world report on road traffic injury prevention*. Geneva: World Health Organization.
2. Bhalla, K.; Naghavi, M.; Shahrzad S.; Bartles, D.; and Murray, C.J. (2009). Building national estimates of the burden of road traffic injuries in developing countries from all available data sources: Iran. *Injury Prevention*, 15(3), 150-156.
3. Nikzad, F.; (2006). *First book in road traffic injury and its damages, causes and suggestions for prevention of outcomes (in Persian)*. Tehran: Asas Publiser.
4. Al-Ghamdi, A.S. (2003). Analysis of traffic accidents at urban intersections in Riyadh. *Accident Analysis and Prevention*, 35(5), 717-724.
5. Holubowycz, O.T.; Klodeden, C.N.; and McLean, A.J. (1994). Age, sex and blood alcohol concentration of killed and injured drivers, riders and passengers. *Accident Analysis and Prevention*, 26(4), 483-492.
6. Lui, K.J.; McGee, D.; Rohdes, P.; and Pollock, D. (1989). An application of conditional logistic regression to study the effects of safety belts, the principal impact points, and car weights on drivers' fatalities. *Journal of Safety Research*, 19(4), 197-203.
7. Bédard, M.; Guyatt, G.H.; Stones, M.J.; and Hirdes, J.P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis and Prevention*, 34(6), 717-727.
8. Al-Ghamdi, A.S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis and Prevention*, 34(6), 729-741.
9. Celik, A.K.; and Oktay, E. (2014). A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. *Accident Analysis and Prevention*, 72, 66-77.
10. De Oña, J.; Mujalli, R.O.; and Calvo, F.J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis and Prevention*, 43(1), 402-411.
11. Mujalli, R.O.; and De Oña, J. (2011). A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. *Journal of Safety Research*, 42(5), 317-326.

12. Mujalli, R.O.; López, G.; and Garach, L. (2016). Bayes classifiers for imbalanced traffic accidents datasets. *Accident Analysis and Prevention*, 88, 37-51.
13. Kwon, O.H.; Rhee, W.; and Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis and Prevention*, 75, 1-15.
14. Chang, L.Y.; and Wang, H.W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, 38(5), 1019-1027.
15. Kashani, A.T.; and Mohaymani, A.S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. *Safety Science*, 49(10), 1314-1320.
16. Akgüngör, A.P.; and Yildiz, O. (2007). Sensitivity analysis of an accident prediction model by the fractional factorial method. *Accident Analysis and Prevention*, 39(1), 63-68.
17. Kashani, A.T.; Mohaymani, A.S.; and Ranjbari, A. (2010). A data mining approach to identify key factors of traffic injury severity. *Promet - Traffic&Transportation*, 23(1), 11-17.
18. Bedrad, M.; Guyatt, G.H.; Stone, J.M.; and Hirdes, P.J. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis and Prevention*, 34, 717-727.
19. Afukaar, F.K. (2003). Speed control in developing countries: issues, challenges and opportunities in reducing road traffic injuries. *Injury Control and Safety Promotion*, 10(1-2), 77-81.
20. Majdzadeh, R.; Khalagi, K.; Naraghi, K.; Motevalian, A.; and Eshraghian, M. R. (2008). Determinants of traffic injuries in drivers and motorcyclists involved in an accident. *Accident Analysis & Prevention*, 40(1), 17-23.