ANALYSIS OF INJURY SEVERITY IN PEDESTRIAN CRASHES USING CLASSIFICATION REGRESSION TREES

By

Vichika Iragavarapu, P.E.
(Corresponding author)
Assistant Research Engineer
Texas A&M Transportation Institute, 3135 TAMU
College Station, TX 77843-3135
Phone: 979/845-5686, fax: 979/845-6006
Email: v-iragavarapu@ttimail.tamu.edu

Dominique Lord, Ph.D, P.Eng.
Associate Professor and Zachry Development Professor I
Zachry Department of Civil Engineering, 3136 TAMU
Texas A&M University
College Station, TX 77843-3136
Phone: 979/458-3949, fax: 979/845-6481
Email: d-lord@tamu.edu

And

Kay Fitzpatrick, Ph.D., P.E.
Senior Research Engineer
Texas A&M Transportation Institute, 3135 TAMU
College Station, TX 77843-3135
Phone: 979/845-7321, fax: 979/845-6006
Email: K-Fitzpatrick@tamu.edu

TOTAL WORDS: 6,164 [3664 Words, 9 Tables, 1 Figure (2500)]

Submitted to the Transportation Research Board 94th Annual Meeting
January 11-15, 2015, Washington D.C.
ABSTRACT

Texas is considered to be an “opportunity” state by the Federal Highway Administration (FHWA), due to the high number of pedestrian crashes. Data from the Fatality Analysis Reporting System (FARS) show that the number of pedestrian fatal crashes in Texas is the third highest in the U.S and is significantly higher than the national average. The research team explored the Texas Department of Transportation (TxDOT) Crash Record Information System (CRIS) database to identify characteristics of crashes involving pedestrians in Texas. A Classification and Regression Tree (CART) analysis of all pedestrian crashes was conducted to find the significant factors influencing the severity of crashes involving pedestrians in Texas. The classification tree identified that light condition, road class, traffic control, right shoulder width, involvement of a commercial vehicle, pedestrian age, and the collision manner, have the most influence on the severity of pedestrian crashes.
INTRODUCTION

Pedestrian crashes are a major safety issue; on average, a pedestrian is killed every two hours and injured every eight minutes in traffic crashes in the United States (1). For the years 2007 through 2011, Texas had the third highest number of pedestrian fatal crashes in the U.S., with about 400 pedestrian fatalities per year, which is equal to 15 percent of all traffic crash fatalities in the US (2, 3). To understand the characteristics and factors influencing pedestrian fatalities, the Texas Department of Transportation (TxDOT) requested an in-depth analysis of these crashes. Understanding of these relations will provide insight for the development of effective countermeasures focused on pedestrians. The objective of this analysis was, therefore, to find the significant factors and their interactions that influence the severity of crashes involving a pedestrian in Texas.

STUDY METHODOLOGY

A variety of methodologies have been used to understand factors influencing crash severities. A comprehensive review of these methodologies is provided in Savolainen et. al, 2011 (4). Data mining techniques are widely applied in the areas of business, medicine, industry and engineering and are gaining attention in the transportation safety area (5, 6, 7, 8, 9). The classification-regression tree (CRT) methodology is a popular data mining technique that does not need a specific functional form and is effective with large data sets containing a large number of explanatory variables (10).

The CRT framework is based on the algorithm first proposed in Breiman et al. (11). The “root node”, which is the node with all the data, is divided into two child nodes on the basis of an independent variable (splitter) that creates the best homogeneity. This process is repeated for each child node until all data in each node has the greatest possible homogeneity. This node is called a “leaf node”. The most famous index for splitting of nominal data is the Gini index. For each tree created, the “goodness of fit” index is calculated using the “misclassification error rate” or “misclassification cost.” “Pruning” is performed according to the cost-complexity algorithm to avoid over-fitting of the training data and to create an optimal tree. An optimal tree is the one that has the least misclassification cost for the test data. Misclassification cost allows inclusion of information about the relative penalty associated with incorrect classification and is inversely proportional to the accuracy of prediction. Correct classifications always have a misclassification cost of 0 (discussed further below). Importance of each variable is calculated using the variable importance index and is scaled such that its summation is one. (12)

The equation used for the Gini index is (12):

\[(\text{Gini Index}) \quad \text{Gini}(m) = 1 - \sum_{j=1}^{J} p^2 (j|m)\]

\[P(j|m) = \frac{p(j|m)}{p(m)}, \quad p(j, m) = \frac{\pi(j)N_j(m)}{N_j}, \quad p(m) = \sum_{j=1}^{J} p(j, m)\]
Where, \( j \) is the number of target variables or classes, \( \pi(j) \) is the prior probability for class \( j \), \( p(j|m) \) is the conditional probability of a record being in class \( j \), provided that it is in the node \( m \), \( N_j(m) \) is the number of records in class \( j \) of node \( m \), \( N_j \) is the number of records of class \( j \) in the root node, and Gini (m) or the Gini index is the indication of impurity in node \( m \). The prior probability shows the proportion of observations in each class in the population. (12)

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

Where, \( p(m) \) is the proportion of existing observations in the terminal node or leaf \( m \) (from all observations) and \( M \) is the number of terminal nodes. (12)

\[
(\text{Variable Importance Index}) \ 
\text{VIM}(x_j) = \sum_{t=1}^{T} \frac{n_t}{N} \Delta \text{Gini}(S(x_j, t))
\]

Where, \( \Delta \text{Gini}(S(x_j, t)) \) is the reduction in the Gini index at node \( t \) that is achieved by splitting variable \( x_j \), \( \frac{n_t}{N} \) is the proportion of the observations in the dataset that belong to node \( t \), \( T \) is the total number of nodes and \( N \) is the total number of observations. (12)

\section*{DATA}

The TxDOT Crash Record Information System (CRIS) database has three subsets: crash, person, and unit. The crash dataset contains information on the characteristics of the crash (e.g., date, time, weather) and its location (e.g., intersection relation, surface condition, traffic control devices). The person dataset has information on the characteristics of the people involved (e.g., age, gender, blood alcohol content, etc.) and injuries sustained (e.g., fatal, no injury, etc.). The unit dataset describes the characteristics of the units (e.g., type of vehicle, contributing factor for unit) involved, along with the contributing factors for each unit involved. Each of these subsets has different codes that distinguish crashes involving pedestrians from other reported crashes. 34,620 pedestrian crashes were available between 2007 and 2011 for use in this analysis. Table 1 lists the variables used in this analysis.
**Table 1. Variables used in analysis.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rpt_Road_Part_ID</td>
<td>Reported roadway part on which the crash occurred</td>
</tr>
<tr>
<td>Wthr_Cond_Rev</td>
<td>Weather condition at the time of crash (Revised to merge some categories)</td>
</tr>
<tr>
<td>Light_Cond_ID</td>
<td>The type and level of light at the time of the crash</td>
</tr>
<tr>
<td>Road_Algn_ID</td>
<td>Roadway Alignment at the crash site</td>
</tr>
<tr>
<td>Surf_Cond_ID</td>
<td>The surface condition at the time and place of the crash</td>
</tr>
<tr>
<td>Traffic_Cntl_Rev</td>
<td>Type of traffic control at the scene of the crash (Revised to merge some categories)</td>
</tr>
<tr>
<td>FHE_Collsn_Rev</td>
<td>The manner in which the vehicle(s) were moving prior to the first harmful event (Revised to merge some categories)</td>
</tr>
<tr>
<td>Othr_Factr_Rev</td>
<td>Additional detail of events/circumstances concerning the crash (Revised to merge some categories)</td>
</tr>
<tr>
<td>Road_Cls_ID</td>
<td>The functional classification group of the priority road the motor vehicle(s) was traveling on before the crash</td>
</tr>
<tr>
<td>Road_Relat_ID</td>
<td>Where the crash occurred in relation to the roadway</td>
</tr>
<tr>
<td>Month</td>
<td>Month of year when the crash occurred</td>
</tr>
<tr>
<td>Hour</td>
<td>Time of day when the crash occurred</td>
</tr>
<tr>
<td>DriverAge</td>
<td>Age of one of the drivers involved in the crash</td>
</tr>
<tr>
<td>DriverGender</td>
<td>Gender of one of the drivers involved in the crash</td>
</tr>
<tr>
<td>PedestrianAge</td>
<td>Age of one of the pedestrians involved in the crash</td>
</tr>
<tr>
<td>PedestrianGender</td>
<td>Gender of one of the pedestrians involved in the crash</td>
</tr>
<tr>
<td>Hwy_Dsgn_Lane_ID</td>
<td>Lane design on the applicable section of highway for crashes located on the state highway system</td>
</tr>
<tr>
<td>Hwy_Dsgn_Hrt_ID</td>
<td>Part of the Highway Design Code indicating HOV (High Occupancy Vehicle), railroads, and toll roads (HRT) for crashes located on the state highway system</td>
</tr>
<tr>
<td>Hp_Shldr_Left</td>
<td>Width of inside shoulder on divided sections, or width of shoulder traveling in descending marker direction, measured in feet, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Hp_Shldr_Right</td>
<td>Width of outside shoulder on divided sections, or width of shoulder traveling in ascending marker direction, measured in feet, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Hp_Median_Width</td>
<td>Median width plus both inside shoulders, measured in feet, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Nbr_Of_Lane</td>
<td>Number of lanes, not including turning and climbing lanes, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Shldr_Type_Left_ID</td>
<td>Type of shoulder on the left side of the road, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Shldr_Type_Right_ID</td>
<td>Type of shoulder on the right side of the road, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Median_Type_ID</td>
<td>Median type description, for crashes located on the state highway system</td>
</tr>
<tr>
<td>Adt_Curnt_Amt</td>
<td>Average daily traffic amount for a given road segment and year for crashes located on the state highway system</td>
</tr>
<tr>
<td>Trk_Aadt_Pct</td>
<td>Adjusted average daily traffic percent for trucks for crashes located on the state highway system</td>
</tr>
</tbody>
</table>
ANALYSIS

The analysis was performed with IBM SPSS Statistics 21 Decision Tree tool, using the tree growing method of CRT (Classification-regression Tree). The response variable assessed in this analysis was crash severity, which is defined as the level of injury sustained by the most severely injured person involved in the crash. The crash severity variable used for developing the classification tree for this analysis was categorized as either fatal or non-fatal. The tree depth was restricted to five levels and impurity was measured with the Gini index. Minimum change in impurity improvement was set at 0.0001. Seventy percent of the data was randomly assigned to train the model and the remaining thirty percent was allocated to the test. The tree was pruned to avoid over-fitting.

In the dataset used for this analysis, the number of non-fatal crashes was almost fifteen times than that of fatal crashes (32,388 vs. 2,232) and the overall prediction accuracy for test sample was 93.5%, whereas the prediction accuracy for fatal crashes was only 10.9%. To ensure the same prediction accuracy for both severity levels, the prior probabilities are set equal so that the target variable level that has a lower proportion is also taken into consideration in predictions. Although this decreases the overall accuracy of the model, the prediction accuracy of the data with the least proportion increases (12). Table 2 shows that with using equal prior probabilities across all categories, the overall prediction accuracy of the model (for the training and the test data) decreased slightly, but the prediction accuracy for fatal crashes improved tremendously (74.3% vs. 10.9%).

Table 3 shows the misclassification costs and Table 4 shows the risk estimate for the model, which is the proportion of cases incorrectly classified after adjustment for prior probabilities and misclassification costs. As discussed above, correct classifications represented on the diagonal in Table 3 have a misclassification cost of 0.
The classification tree generated (Figure 1) shows that the initial split at node 0 is based on the variable of light condition, which implies that light condition is the best variable to classify and predict pedestrian crash severity (fatal versus non-fatal). More (64%) pedestrian crashes are predicted to occur in daylight, whereas a higher proportion of fatal crashes are predicted to occur in dark conditions (13% vs. 3%, Node 1 vs. Node 2). Traffic control, road class, and pedestrian age are selected to be the splitters more than once, implying that these variables have multiple effects on the crash severity outcome.
Figure 1. Classification Tree.
DISCUSSION

The classification tree developed in this analysis (shown in Figure 1) indicates that the following variables are critical in classifying the injury severity of pedestrian crashes:

- Light condition
- Road class
- Traffic control
- Right shoulder width
- Involvement of a commercial vehicle
- Pedestrian age
- Manner in which the vehicle(s) were moving prior to the first harmful event

Daylight conditions are associated with more pedestrian; however, the severity of the crash is higher in dark conditions. When a pedestrian is struck at night, he or she is four times more likely to be killed when compared to daylight conditions (13% vs. 3%, Node 2 vs. Node 1). This result is in agreement with a study on pedestrian crashes in North Carolina that found that dark conditions (with and without streetlights) significantly increase the probability of fatal injury for pedestrians (13). This could be a reflection of higher speeds at night, along with greater difficulty in detecting pedestrians in dark conditions; hence, not being able to reduce the speed in time.

Under all light conditions, the probability of pedestrian crashes is higher on city streets, county roads, and other lower speed roads that have segment-related traffic control device (e.g. warning sign); whereas the severity of the crash is more on higher speed roads, i.e. Interstates, US & State Highways, FM roads, and Tollways (25% vs. 6%, Node 4 vs. Node 3, and 13% vs. 3%, Node 14 vs. Node 13). This result was further investigated and is discussed in greater detail in Iragavarapu et al. (14). Kim et al. (9) also found that freeway, U.S. route, and state route increased the probability of fatal injury in pedestrian crashes, compared with local city streets (13).

Younger (≤ 60 years) pedestrians are predicted to be involved in more crashes, whereas older (> 60 years) pedestrians are more likely to be killed when struck by a vehicle (5% vs. 4%, Node 16 vs. Node 15, and 12% vs. 2%, Node 20 vs. Node 19). This finding is expected because in general older pedestrians have lesser physical strength to cope with the injuries and are less agile to escape to a safer location just before the crash. Kim et al. also found that older pedestrians are more likely to sustain greater injury than younger pedestrians (13,15). Holubowycz found the greatest fatality rates in pedestrians 75 years or older (16). This results supports the concept of a “pedestrian airbag” technology which would improve the chances of surviving for pedestrians hit by vehicles, more so for elderly pedestrians. Crandall et. al discuss this approach for pedestrian safety in their 2002 paper (17).

Commercial vehicle involvement in a pedestrian crash is associated with a greater probability of pedestrian fatality (10% vs. 2%, Node 12 vs. Node 11). This is obviously attributed to their larger weight, longer stopping distances, higher bumper height, and blunt geometry, which has also been documented in previous studies (13, 15, 18, 19).
Locations with no traffic control device or intersection-related traffic control devices (e.g., signal) are found to be associated with more number of pedestrian crashes; however, locations with segment traffic control devices (e.g. warning sign or flagger) or both traffic control devices (i.e., officer, flagman) are associated with higher proportion of fatal pedestrian crashes (15% vs. 4%, Node 8 vs. Node 7, and 6% vs. 2%, Node 6 vs. Node 5). This result is intuitive because more pedestrians are expected to be present at intersections and the vehicle speeds are relatively lower, when compared to mid-segment.

The results also show that on high speed roads, more crashes are expected at locations with right shoulder width less than 8.5 feet (Node 9), when one of the vehicles involved is either going straight or backing (Node 17). However, a higher proportion of fatal crashes on high speed roads are at locations with right shoulder width more than 8.5 feet.

**SUMMARY AND CONCLUSIONS**

This study has documented a regression tree analysis for examining factors that influence crash severity in crashes involving a pedestrian. A total of 34,620 pedestrian crashes that occurred between 2007 and 2011 were analyzed. The regression tree provided satisfactory results. The results (summarized in Table 5) were intuitive and consistent with the results of previous studies on pedestrian crashes that used other analytical techniques, such as probabilistic models of crash injury severity. The information provided in this study should help TxDOT and other transportation agencies to better target their efforts for reducing the number and severity of pedestrian collisions.
### Table 5. Summary of Classification Tree Result.

<table>
<thead>
<tr>
<th>Significant Factor</th>
<th>Classification Tree Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light condition</td>
<td>When a pedestrian is struck at night, they are four times more likely to be killed when compared to daylight conditions.</td>
</tr>
<tr>
<td>Road class</td>
<td>When a pedestrian is struck on higher speed roads, they are four times more likely to be killed when compared to lower speed roads.</td>
</tr>
<tr>
<td>Pedestrian age</td>
<td>Older (60+) pedestrians are six times more likely to be killed when compared to younger pedestrians when struck in daylight; at nighttime there is not much difference.</td>
</tr>
<tr>
<td>Traffic control</td>
<td>When a pedestrian is struck at a location with segment traffic control devices (e.g., warning sign or flagger) or generic traffic control devices (e.g., officer), they are four times more likely in dark conditions and three times more likely in daylight to be killed when compared to location with no traffic control or intersection traffic control (e.g., stop sign).</td>
</tr>
<tr>
<td>Right shoulder width</td>
<td>When a pedestrian is struck on high speed roads in dark conditions, they are twice as likely to be killed if the right shoulder width is more than 8.5 feet.</td>
</tr>
<tr>
<td>Involvement of a commercial vehicle</td>
<td>When a pedestrian in daylight, they are five times as likely to be killed if the crash involves a commercial vehicle.</td>
</tr>
<tr>
<td>Collision manner</td>
<td>On high speed roads with right shoulder width less than 8.5 feet, pedestrian crashes in dark conditions are six times more likely when one of the vehicles involved is either going straight or backing when compared to vehicles turning left or right.</td>
</tr>
</tbody>
</table>


ACKNOWLEDGMENTS

This paper is based on research sponsored by the Texas Department of Transportation (TxDOT) and the U.S. Department of Transportation, Federal Highway Administration (FHWA). The project was under the direction of Cary Choate of TxDOT. The research was performed at the Texas A&M Transportation Institute. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of TxDOT.

REFERENCES


