Journal of Transport & Health 4 (2017) 118–131

Contents lists available at ScienceDirect

Journal of Transport & Health

journal homepage: www.elsevier.com/locate/jth

Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods

Dursun Delen [a,*], Leman Tomak [b], Kazim Topuz [c], Enes Eryarsoy [d]

[a] Center for Health Systems Innovation, Spears School of Business, Oklahoma State University, Stillwater, OK, USA
[b] Biostatistics and Public Health, Ondokuz Mayis University, Samsun, Turkey
[c] Center for Health Systems Innovation, Oklahoma State University, Stillwater, OK, USA
[d] Management Information Systems, Istanbul Sehir University, Istanbul, Turkey

A R T I C L E I N F O

Article history:
Received 4 August 2016
Received in revised form
22 January 2017
Accepted 23 January 2017
Available online 21 February 2017

Keywords:
Automobile crashes
Predictive analytics
Risk factors
Injury severity
Machine learning
Sensitivity analysis

A B S T R A C T

Investigation of the risk factors that contribute to the injury severity in motor vehicle crashes has proved to be a thought-provoking and challenging problem. The results of such investigation can help better understand and potentially mitigate the severe injury risks involved in automobile crashes and thereby advance the well-being of people involved in these traffic accidents. Many factors were found to have an impact on the severity of injury sustained by occupants in the event of an automobile accident. In this analytics study we used a large and feature-rich crash dataset along with a number of predictive analytics algorithms to model the complex relationships between varying levels of injury severity and the crash related risk factors. Applying a systematic series of information fusion-based sensitivity analysis on the trained predictive models we identified the relative importance of the crash related risk factors. The results provided invaluable insights for the use of predictive analytics in this domain and exposed the relative importance of crash related risk factors with the changing levels of injury severity.

1. Introduction

Big Data has become a dominant term in describing the exponential growth, accessibility, availability, and widespread use of information—in structured, semi-structured and unstructured format—in a variety of business context (Delen, 2015). Big Data by itself, regardless of the level of volume, variety, or velocity, is worthless unless analysts do something with it that delivers value. That's where "big" analytics comes into the picture. Although organizations have long run reports and dashboards against data warehouses, most have not opened these repositories to in-depth on-demand exploration. This is partly because analytics tools were too complex for the average user and partly also because the repositories often did not contain all the data needed for the power users. But this is about to change (and had already changed for some) in a dramatic fashion, thanks to the new Big Data Analytics paradigm. One area where Big Data Analytics has greatest potential to make a significant impact is in critical analysis of traffic accidents, especially automobile crashes and resultant injures.

As the technology keeps advancing, new and improved safety measures are being developed and incorporated into vehicles and roads to prevent crashes from happening and/or reduce the impact of the injury sustained by passengers

* Correspondence to: Spears School of Business, Oklahoma State University, 700 North Greenwood Avenue, Suite North Hall 341, Tulsa, OK 74106, USA.
E-mail addresses: dursun.delen@okstate.edu (D. Delen), lemant@omu.edu.tr (L. Tomak), ktopuz@okstate.edu (K. Topuz), eneseryarsoy@sehir.edu.tr (E. Eryarsoy).

http://dx.doi.org/10.1016/j.jth.2017.01.009
2214-1405/

caused by such incidents. Despite the extent of these efforts the number of car crashes and the resulting injuries are increasing worldwide. For instance, according to NHTSA (the National Highway Traffic Safety Administration), only in the US more than six million traffic accidents claim over 30,000 lives and injure more than 2 million people each year (NHTSA, 2014). The latest NHTSA report presented to the US Congress on April 2014 states that in 2012 highway fatalities in the United States reached 33,561, which is an increase of 1,082 over the previous year (Friedman, 2014). In the same year, an estimated 2.36 million people were injured in motor vehicle traffic crashes, compared to 2.22 million in 2011. As a result, an average of nearly 4 lives were lost and nearly 270 people were injured on America's roadways every hour in 2012. In addition to the staggering number of fatalities and injuries, these traffic accidents also cost the tax payers more than $230 billion. Hence, road safety is a major problem in the US, and around the world.

Root causes of traffic accidents and crash related injury severity are of special concern to general public, but especially to researchers (in academia, government and industry) since such investigation would be aimed not only at prevention of crashes but also at reduction of their severe outcomes, potentially saving many lives and money. In addition to laboratory- and experimentation-based engineering research methods, another way to address the issue is to identify the most probable factors that affect injury severity by mining the historical data on car crashes. Intimate understanding of the complex circumstances where drivers and/or passengers are more likely to sustain severe injuries or even be killed in a car crash has a great potential to mitigate the risks involved in automobile crashes and thereby advance the well-being of people involved in these automobile crashes. Many factors were found to have an impact on the severity of injury sustained by occupants in the event of an automobile accident. These factors include behavioral or demographic features of the occupants (e.g., drugs and/or alcohol levels, seatbelt or other restraining system usage, gender and age of the driver, etc.), crash related situational characteristics (e.g., road type/situation, direction of impact, strike versus struck, number of cars and/or other objects involved, etc.), environmental factors and related roadway conditions at the time of the accident (road surface condition, weather conditions, visibility and/or light conditions, time of the day, etc.), and the technical characteristics of the vehicle itself (the age of the vehicle, weight of the vehicle, body type of the vehicle, etc.).

The main goal of this analytic study is to determine the most prevailing risk factors and their relative importance/significance in influencing the likelihood of increasing severity of injury on automobile crashes. The car crashes examined in this study included a collection of geographically well-represented sample. In order to have a consistent sample, the data set comprised of only collations of specific types: single or multi-vehicle head-on collisions, single or multi-vehicle angled collisions, and single vehicle fixed-object collisions. To obtain reliable and accurate results, in this investigative study we employed the most prevalent machine learning techniques to identify the significance of crash related factors as they relate to the changing levels of injury severity in automobile crashes. Although some of the machine learning techniques that we have developed were also investigated by other researchers in this application domain, our approach relies on methodological innovations in collective use of multiple prediction models and to develop information fusion based sensitivity analysis from the developed machine learning models.

The rest of the paper is organized as follows. The next section, Section 2, summarizes some of the most relevant research where data-driven analytical methods are used to study injuries in automobile crashes. Section 3 describes our methodology which includes brief descriptions of the methods used to obtain and preprocess the data, machine learning techniques used to develop the models, and the assessment techniques used to evaluate the findings. Section 4 presents the predictive model building results and summarizes and discusses the sensitivity analysis findings. The last section, Section 5, recaps the research outcomes and provides the concluding remarks.

2. Literature review

There is a wealth of literature published in reputable journals and conference proceeding on analysis of traffic accidents and resulting outcomes. A vast majority of these studies dealt primarily with the analysis of vehicle related physical properties and roadway related environmental factors that provokes the crash involvement so that the road, traffic and vehicle related features would be reengineered (designed and developed) to prevent the car crashes from happening in the first place. On the other hand, especially in the recent years, there seem to be an equally strong emphasis on data and analytics-based studies that focus on crash related injuries. Since the research explained herein deals with the analysis of crash related injuries and its root causes (underlying risk factors), the literature in this section will primarily be specific to the most relevant and rigorous work in this specific area.

A number of previous studies in this area have developed injury severity models using crash related data sets (Savolainen et al., 2011; Huang et al., 2008; Hauer, 2006). Although, most of them concentrated on traffic accident records limited to a small/specific geographic region, a particular crash type, or a specific road or environmental conditions, they have paved the road for more comprehensive analytics studies. The main reason for such limiting characteristics of these studies was perhaps to make the application domain as narrowly defined as possible so that a somewhat homogenous dataset can be obtained and used to derive more accurate prediction and explanatory models. Most of these early studies used traditional statistical techniques to evaluate a set of well-defined hypotheses using a purposefully samples data set. The following section provides a summary of a number of representative sample of these studies that developed and tested analytic models to discover and assess the factors that are influential to increasing or decreasing the level of injury severity experienced by occupants (drivers and/or passengers) during motor vehicle crashes.