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

ARTICLE INFO

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

ABSTRACT

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.

Multinomial logistic regression and its derivative models (i.e., ordered-logit or ordered-probit) have been the most commonly used techniques in developing injury severity analysis models. In a recent study, a multivariate ordered-response probit (MORP) model system was developed by Abay et al. (2013) to explain the injury severity in two-vehicle crashes primarily focusing on the use of seat belts. Another study was conducted within the state of Ohio that focused on the risk factors contributing to injury severity at freeway merging and diverging locations (Mergia et al., 2013). The data set was extracted from the police-reported crash data for the years 2006–2009 and the modelling technique used was the generalized ordinal logit model. The levels of the variables related to road contour were adopted from the state agency's crash database. They split the mode estimation in two parts: one for converging roads and the other for diverging roads. The fit for the model was given in terms of R-square, which was 0.10 and 0.09 for the converging and diverging roads respectively. In an earlier study, Lui and McGee (1988) used statistical methods (specifically logistic regression models) to analyze the likelihood of fatal outcomes at traffic accidents. The data for this study was procured from FARS (the Fatal Accident Reporting System) database, which contained crashes where at least one fatality was observed. A similar statistical study was carried out by Wood and Simms (2002) for the goal of identifying risk factors that lead to fatality of incapacitating injuries.

As opposed to using only one statistical technique, some of the previous studies preferred to use two or more techniques in an either comparative or a complementary manner. For instance, Park et al (2012) used three statistical techniques—ordered probit, ordered logit, and multinomial logit—to analyze influential factors on the severity of injuries sustained in traffic crashes based on the crash data obtained from the entire network of Korean expressways in 2008. The results of the three methods combined to provide a more reliable estimate of the most influential factors. In another comparative study O'Donnell and Connor (1996) developed and compared two standard statistical modeling techniques: ordered probit and ordered logit models. In this study they wanted to investigate the likelihood of sustaining varying levels of injury severity using variety of descriptive variables including the driver characteristics. The findings of this study suggested that the level of injury severity increases with increasing driving speed, the age of the vehicle, the age of the driver and other occupants, along with high blood alcohol level (over 0.08 percent), not using a restraining system (i.e., seatbelt), female gender, and manner in which the collision happened (i.e., head-on collations). Many of the postulated outcomes and findings of this study were partially confirmed by earlier studies (Hauer, 2009; Delen et al., 2006) and also supported by the machine learning models developed in this study.

Our study can be differentiated from those of the previous studies in two respects. First, we used a number of most-respected machine learning methods (including artificial neural networks, support vector machines and decision trees) as a collective means to capture the potentially non-linear highly-complex relationships between the levels of injury severity and crash related risk factors. The prediction accuracy on the holdout/test sample is used as the measure of evaluation for different prediction model types and their overall contribution to the predicted outcome. Second, we developed and used an information-fusion based sensitivity analysis on those trained prediction models to identify the prioritized importance of crash-related factors as they relate to different injury severity levels. Using an ensemble of prediction methods for sensitivity analysis (as opposed to relying on only one model type) provided us with a coherent list of risk factors that are ranked based on their level of contribution/importance to the predicted phenomenon.

3. Research methodology

To lay the foundation upon which our work is carried out, in this section we describe the major steps involved in carrying out our research methodology. Specifically, we start the section with a description of the particular techniques that we used for data acquisition and data preparation. Then, we described the specifics about the prediction methods that we developed for the study; followed by a short section on description of the metrics used to assess the predictive accuracy of the classification models. Lastly, we briefly explain the details about the information-fusion based sensitivity analysis method that we used to identify the ranked-order importance/significance of the risk factors.

To effectively and efficiently perform the individual tasks in the proposed methodology, we employed several statistical and data mining software tools. Specifically, we used JMP (a statistical and data mining software tool developed by SAS Institute), Microsoft Excel and Tableau for data inspection, data understanding and data preprocessing; IBM SPSS Modeler and KNIME for data merging, predictive model building and sensitivity analysis.

3.1. Data acquisition and preparation

The crash data used in this study was originally acquired from NASS GES (National Automotive Sampling System General Estimates System) database. This database contains roughly one percent of all US automobile crashes as they are reported by law enforcement agencies (GES, 2014). The database is specifically designed to have a sufficiently large nationwide sample of all crash reports created by local police departments. The original dataset contains information related to property damage, driver characteristics, crash circumstances, environmental conditions and severity of the injury specifications for all occupants.

The GES dataset used for the study covered all incidents for the most recent two years (2011 and 2012). The complete data set was obtained in the form of three separate flat/text files—accident, vehicle, and person. The *accident* files contained specific characteristics about road conditions, environmental conditions and the crash related settings. The *vehicle* files

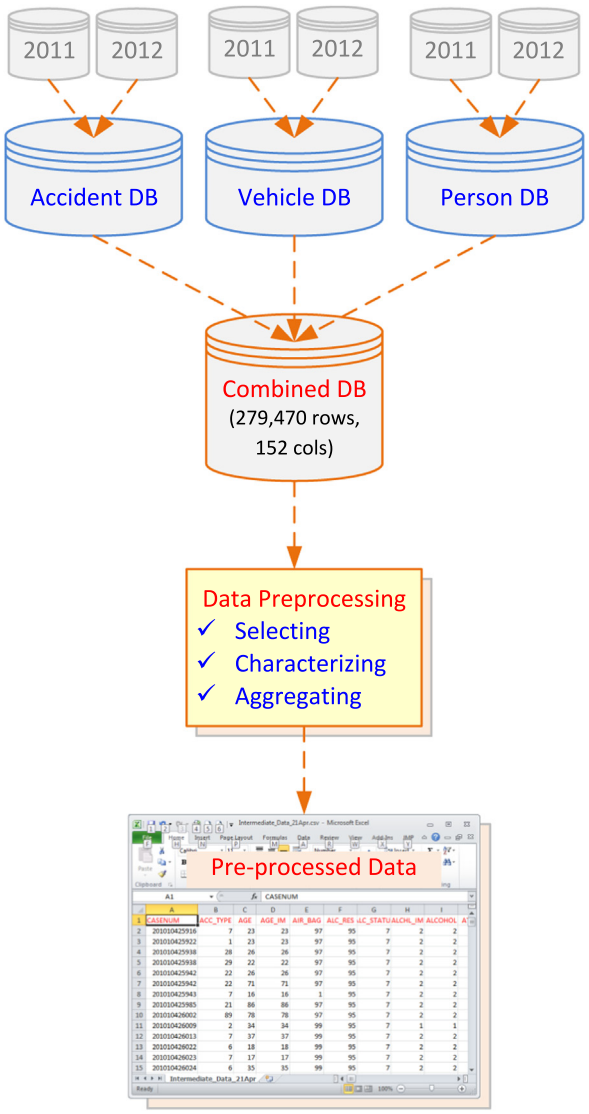

Fig. 1. The data acquisition/merging/preparation process.

included a large number of variables about the specific features of the vehicle involved in the crash. The *person* files provided detailed demographics, injury and situational information about the occupants (i.e., driver and the passengers) impacted from the automobile crash. In order to consolidate the data into a single database, the two years of data is merged within each file types (i.e., accident, person, vehicle), and the resulting files are combined using unique accident, vehicle and person identifiers to create a single dataset. After the data consolidation/aggregation the resulting dataset included person-level records—one record per person involved in a reported automobile crash. At this point in the process (before the data cleaning, preprocessing and slicing/dicing), the complete dataset included 279,470 unique records (i.e., persons/occupants involved in crashes) and more than 150 variables (a combination of accident, person and vehicle related characteristics). Fig. 1 graphically illustrates the individual steps involved in the process of data preparation (i.e., acquiring→merging→filtering→reformatting).

In the final dataset, all of the samples/records used in this study reported some level of injury sustained by the driver as a result of a car crash. In order to obtain a consistent data set, some of the most recent studies have also focused only on driver specific injury characteristics (Zeng et al., 2016; Chen et al., 2016b). In this study, the levels of injury severity were encoded as *low*-level-of-injury (i.e., a combination of possible injury and minor non-incapacitating injury) or *high*-level-of-injury (a combination of incapacitating injury and fatality). In order to have a representative sample with common characteristics, any crash that did not result in any injury or any damage to the vehicle were removed from the dataset by filtering the GES data based on the variable VEH_SEV, which specifies the extend of the damage sustained by the vehicle at the time of crash. This filtration process did not eliminate many records, because the crash that caused injuries to occupants but had very

Table 1
List of variables included in the study.

Variable	Description	Data Type	Descriptive Statistics*	% Miss/Unk.
AIR_BAG	Airbag is deployed	Binary	Yes: 52, No: 26	5.2
ALC_RES	Alcohol test results	Numeric	12.68 (15.05)	0.4
BDYTYP_IMN	Vehicle body type	Nominal	Sedan: 34, Sm-SUV: 13	3.2
DEFORMED	Extend of damage	Nominal	Major: 43, Minor: 22	3.7
DRINKING	Alcohol involvement	Binary	Yes: 4, No: 67	28.8
AGE	Age of the person	Numeric	36.45 (18.49)	6.9
DRUGRES1	Drug test results	Binary	Yes: 2, No: 72	25.5
EJECT_IM	Ejection	Binary	Yes: 2, No: 93	4.9
FIRE_EXP	Fire occurred	Binary	Yes: 3, No: 97	0.0
GVWR	Vehicle weight category	Nominal	Small: 92, Large: 5	2.9
HAZ_INV	Hazmat involved	Binary	Yes: 1, No: 99	0.0
HOUR_IMN	Hour of the day	Nominal	Eve: 39, Noon: 32	1.2
INT_HWY	Interstate highway	Binary	Yes: 13, No: 86	0.7
J_KNIFE	Jackknife	Binary	Yes: 4, No: 95	0.2
LGTCON_IM	Light conditions	Nominal	Daylight: 70, Dark: 25	0.3
MANCOL_IM	Manner of collision	Nominal	Front: 34, Angle: 28	0.0
MONTH	Month of the year	Nominal	Oct: 10, Dec: 9	0.0
NUMINJ_IM	Number of injured	Numeric	1.23 (4.13)	0.0
PCRASH1_IMN	Pre-crash movement	Nominal	Going Str.52: , Stopped: 14	1.3
REGION	Geographic region	Nominal	South: 42, Midwest: 24	0.0
REL_ROAD	Relation to traffic way	Nominal	Roadway: 85, Median: 9	0.1
RELJCT1_IM	At a junction	Binary	Yes: 4, No: 96	0.0
REST_USE_N	Restraint system used	Nominal	Yes: 76, No: 4	7.4
SEX_IMN	Gender of the driver	Binary	Male: 54, Female: 43	3.1
TOWED_N	The car was towed	Binary	Yes: 49, No: 51	0.0
VEH_AGE	Age of the vehicle	Numeric	8.96 (4.18)	0.0
WEATHR_IM	Weather condition	Nominal	Clear: 73, Cloudy: 14	0.0
WKDY_IM	Weekday	Nominal	Friday: 17, Thursday 15	0.0
WRK_ZONE	Work zone	Binary	Yes: 2, No: 98	0.0
INJ_SEV	Injury severity (**DV**)	Binary	Low: 79, High: 21	0.0

* For numeric variables: mean (St. Dev.); for binary or nominal variables: % frequency of the top two classes.

minor or no damage to the vehicle were very sporadic in existence. As a result, the data set used in this study contained automobile crashes that exhibited some level of injury and small, medium or severe damage to the car. Since this study is focused on discovering injury severity risk factors for drivers involved in automobile crashes, the dataset also excluded all records involving injury to only pedestrians and/or animals. Exclusions were also made for the type of vehicle: since the study focuses on car crashes the injury records that were associated with bicycles, motorcycles, tractors and other farm equipment we also filtered out of the final data set.

As is the case in any data-driven analytical studies, the quality and reliability of results depends on the level of efforts put forth during the data preparation process. Constituting more than 80% of the total time spent on the project, data pre-processing is by far the most critical and time demanding activity. While merging the multiple years of data, we had to consolidate the discrepancies in the way the variables were represented over time. We also needed to use a multi-source approach (including published literature, statistical analysis and common sense) in identifying the variables to include in the model building. Of all the variables—directly obtained from the GES databases and the ones that were derived/re-calculated using the existing GES variables—29 were selected as relevant and potentially influential in determining the varying levels of injury severity involved in car crashes. This extant of variables expected to provide a rich description of the people and the vehicle involved in the accident, the specifics about the environmental conditions at the time of the crash, the settings surrounding the crash itself, and specifics about time and place of the crash. Table 1 lists and briefly describes the variables created and used for this study.

After the preprocessing the data, we noticed that in the representation of the dependent variable we've had significantly less number of reported cases for high-level of injury severity than the number of cases for low-level of injury severity. Indeed, we had an underreported crash data problem (Abay, 2015; Ye and Lord, 2011), which is also commonly called as data imbalance problem in data mining. This problem is quite common in machine learning and data mining applications as it appears in many real-world data sets. It is a problem because if not handled, the prediction results would be skewed significantly in favor of the majority class (Thammasiri et al., 2014). That is, for significantly skewed datasets, the results may have very high accuracy for the majority class and the overall prediction problem while having very low accuracy for the minority class. For instance, a dataset having 95% representation of the majority class and 5% of the minority class would have 95% accuracy without correctly classifying any of cases for the minority class. In data mining, this is often referred to as "fool's gold." There is not a best way to handle imbalanced data problem. Therefore, in this study, we experimented with the most common procedures—under sampling, over sampling and a hybrid of over and under sampling—and settled on using the under sampling method for balancing the dataset.

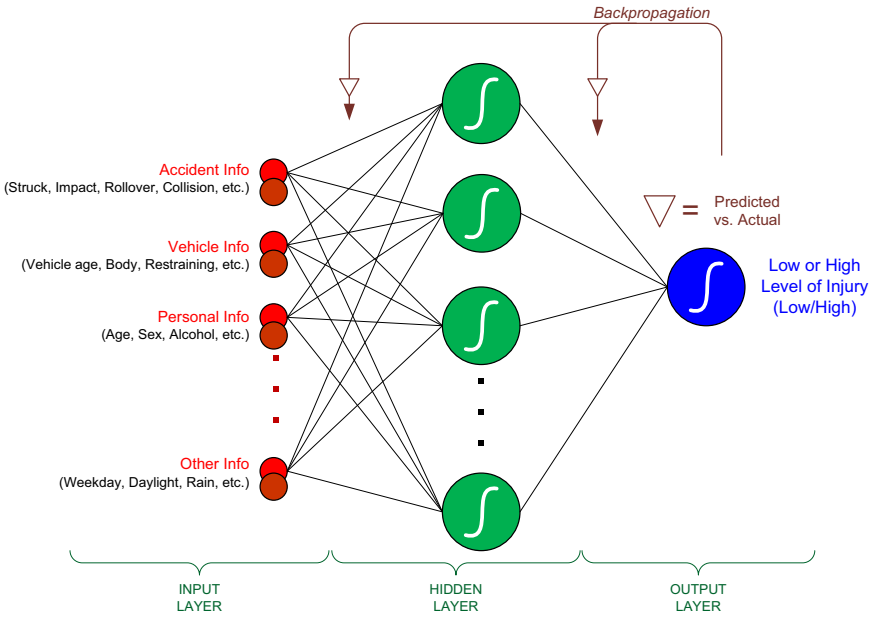

Fig. 2. Graphical representation of our multi-layered-perceptron-type neural network model.

3.2. Prediction methods

In this study, we represented the dependent/response variable (injury severity) as a binary variable with two possible outcomes (low- versus high-level of injury severity). Accordingly, the problem is characterized as a binomial classification problem. There are a number of machine learning and statistical techniques that can handle this type of classification problems. Based on the extant literature and our preliminary experimentation, we settled on employing three prevalent machine learning techniques (neural networks, support vector machines and decision trees) along with the most common statistical technique for this type of classification problems (i.e., logistic regression). The following sections very briefly describes these prediction (i.e., classification) techniques and their unique specification used in this research study.

3.2.1. Neural networks

Since their inception in early 1960s, neural networks have become increasingly more ubiquitous machine learning technique to unravel variety of complex analytics problems. They are recognized as biologically motivated (brain metaphor), greatly adept machine learning tools that are capable of capturing and representing highly complex non-linear relationships exist in real-world data sets. Neural networks are modeled to mimic the learning process of the human brain (i.e., neurological functions and cognitive system of the brain). The predictive models developed with neural networks can estimate the outcome of new observations/cases based on the patterns captured from other observations/cases after executing a process called "learning" from historic data (Haykin, 2008). After experimenting with several different neural network architectures, for this classification problem, we choose to use feed-forward multi-layer perceptron (MLP) with back-propagation supervised learning algorithm—arguably the most common neural network architecture used in data mining studies. Fig. 2 shows the graphical representation of the single-middle layer MLP architecture used in this study. After experimenting with more than one hidden-layer MLP architectures, for this dataset, we observed that the prediction accuracy on hold-out sample is not improving, but the model is getting more complex. Therefore, we settled on using a single hidden later MLP as our neural network architecture.

An MLP type neural network is basically a collection of processing elements (nonlinear neurons or perceptrons) organized in layers and connected to other layers in a feed-forward multi-layer structure, where the input layer obtains the data/signal and passes it on to the next hidden layer, then potentially to another hidden layer, and ultimately to the output layer. The end signal at the output layer is then compared to actual observation and the error/difference is fed back to the network (in some form) for gradual adjustment of it parameters/weights (i.e., the learning process). Neural networks in general and MLP in specific are known to be very capable complex function approximators for prediction (both regression as and classification) as well as for clustering/segmentation type analytic problems. In an empirical study, Hornik et al. (1990) showed that an MLP neural network model (given that it is designed optimally with proper model parameter values) can learn highly complex non-linear relationships to optimal accuracy level. Accordingly, Alkheder et al (2016) used ANN to successfully predict severity of traffic accidents, while Karlaftis and Vlahogianni (2011) showed the differences and similarities of ANN to its statistical counterparts within the context of transportation research.

3.2.2. Support vector machines

In addition to neural networks, support vector machines (SVMs) have become another widely used machine learning techniques of the recent years, mostly because of their superior predictive performance and their sound theoretical foundation. SVMs are often characterized as an extension of neural networks (especially when RBF kernel-type is employed). Simply put, SVMs are part of the supervised learning methods that produce input-output functions from a set of labeled training data, somewhat the same way that neural networks do. The relationship between the input and output vectors can be either a classification type function (used to assign cases into predefined output class labels) or a regression type function (used to estimate the continuous numerical value of the desired output). That is, similar to neural networks, SVMs can also handle both types of prediction model types (i.e., classification and regression).

For classification type prediction problems, as is the case in this study, SVMs constructs hyperplanes to optimally separate the output classes from each other using the patterns in the training data set. For this type of classification problems, generally speaking, many linear classifiers (two or more dimensional planed—hyperplanes) can separate the data into multiple sub-sections each representing one of the classes. However, only one hyperplane achieves the maximum separation between these classes. Data used in SVMs may have more than two dimensions (i.e., two distinct classes). In that case, SVMs separated the data using n–1 dimensional hyperplane, where n is the number of dimensions (i.e., class labels). This may be seen as a typical form of linear classifier, where we are interested in finding n–1 the hyperplane so that the distance from the hyperplanes to the nearest data points are maximized. The assumption is that the larger the margin or distance between these parallel hyperplanes, the better the generalization power of the classifier (i.e., prediction power of the SVM model). If such hyperplanes exist; they can be mathematically represented using quadratic optimization modeling. These hyperplanes are known as the maximum-margin hyperplane and such a linear classifier is known as a maximum margin classifier Cristianini and Shawe-Taylor (2000).

In addition to their solid mathematical foundation in learning theory, SVMs have also demonstrated highly competitive performance in numerous real-world prediction problems, such as medical diagnosis, bioinformatics, face/voice recognition, demand forecasting, image processing and text mining, which has established SVMs as one of the most common analytics tools for knowledge discovery and data mining. Similar to artificial neural networks, SVMs possess the well-known ability of being universal approximators of any multivariate function to any desired degree of accuracy. Therefore, they are of particular interest to modeling highly nonlinear, complex problems, systems and processes. SVM have already become a standard analytics tool in transportation research (Chen et al., 2016a; Li et al., 2008).

3.2.3. Decision trees

Decision trees are not new. Some form of decision trees has been used for a variety of decision problems since early 1930s. Early implementation of decision trees was based on expert knowledge, and not so much on data (i.e., deductive as opposed to inductive). With the widespread use of data mining and the new algorithmic advances that allowed them to be developed inductively from historic data, decision trees have become common complement to existing analytics tools again. They are not only viable but also increasingly more popular alternative to neural networks and support vector machines in wide range of analytics problems.

The biggest advantage that decision trees has over other machine learning methods (i.e., neural networks and support vector machines) is the fact that decision trees are capable of explain the inner structure of “how they do what they do” (Delen, 2015). They can do that in the form of an inverse tree like structure, a list of indented statements or a series of production rules. Therefore, they are known to produce relatively transparent model structures (i.e., trees with nodes edges), as opposed to being called black-boxes, which has been the biggest complaint in the analytics community about neural networks and support vector machines. This advantage has made easily understandable and easily deployable modeling techniques, increasing their popularity and use as an analytical tool for real world problems, even when they are not the best performer. There are a large number of decision tree algorithms. The most commonly cited ones include Quinlan's ID3, C4.5, C5 (Quinlan, 1986, 1993) and Breiman et al.'s (1984) classification and regression trees (CART). After experimenting with a number of decision tree algorithms, based on the performance results that we have obtained, in this classification problem we choose to use C5 (an improved version of C4.5 and ID3) algorithm as the decision tree prediction method.

3.2.4. Logistic regression

As mentioned in the earlier sections of the paper, historically, logistic regression and its variants (i.e., ordered logit and probit models) have been the most commonly used statistical method for studying injury severity risk factors. Generally speaking, they are amongst the most common statistical techniques for binomial and multinomial classification type prediction problems. Logistic regression builds probability-based classification models by employing a supervised learning-based expectation maximization algorithm. It was developed in the 1940s as a complement to linear regression and linear discriminant analysis methods. It is primarily used for predicting binary dependent variables, but its extended version can also handle multi-class classification problems. Even though it is a part of the family of regression techniques, logistic regression is not meant to model linear functions using least squared method, instead it models discrete outputs using a heuristic technique. Rather than predicting a point estimate of the event itself, logistic regression builds a model to estimate the odd ratio of its occurrence. Although logistic regression has been a widespread statistical algorithm for classification type analytic problems, its restrictive assumptions on independence, normality and multi-collinearity made it a less desired

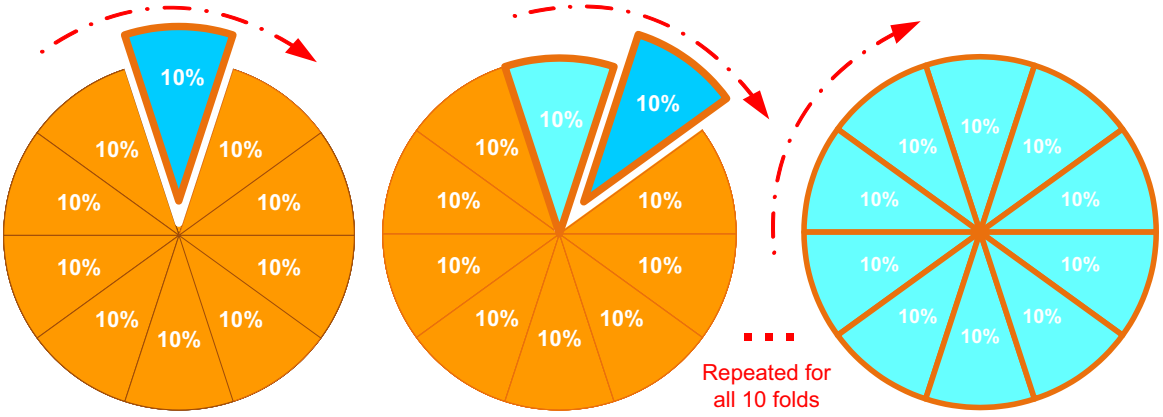

Fig. 3. A graphical illustration of the *10*-fold cross validation methodology.

alternative against increasingly more capable machine learning techniques for real-world prediction problems.

3.3. Experimental framework (k-fold cross-validation)

Estimating the true accuracy of a classification model induced by a supervised learning algorithm is important for the following two reasons: first, it can be used to estimate its future prediction accuracy on real cases, which could indicate the level of confidence one should have in the classifier's output in the prediction system. Second, it can be used for choosing a classifier from a given set (identifying the "best" classification model out of the trained ones). The most common estimation methodologies used for classification-type data mining models is to split the data into training and testing sets. A conventional implementation of this methodology is often called *k*-fold cross validation. In *k*-fold (or *v*-fold, depending on which literature you look at) cross-validation methodology, the complete dataset (all of the samples/rows) is randomized and then split into *k* distinct subsets of near equal number of samples/rows. To operationalize the experimentation, a classifier is trained on the *k*-1 number of records and tested on the remaining one subset. This experimentation process (i.e., training followed by testing) is repeated for *k* times, each time a different fold is used as the test dataset and the remainder of the samples is used as the training dataset. Then the overall accuracy of the classifier is calculated using a simply average of the *k* individual test sample accuracy measures. The formula for this aggregation is shown Eq. (1).

$$CVA\ (CrossValidationAccuracy) = \frac{1}{k}\sum_{i=1}^{k} Accuracy_i \tag{1}$$

In this study, we used 10 as the value of *k* (i.e., 10-fold cross validation). A number of experimental studies showed that 10 seems to be a good (or perhaps an "optimal") value for the number of folds to use. It seems to have produced an optimal balance between the time it takes to complete the experimentation and minimizing the bias and variance associated with the validation process (Breiman et al., 1984, Kohavi, 1995). Fig. 3 graphically illustrates the 10-fold cross validation methodology we used for this investigative study.

3.4. Evaluation metrics

Methodological advancement (including recent applications of advanced analytics and ensemble models) has substantially improved our understanding of the factors that affect crash-frequencies and crash severities. It is perhaps the combination of evolving methodologies and assessment techniques that holds the greatest promise in advancing the analytics studies in this application domain (Lord and Mannering, 2010). Therefore, to have a good comparison of the analytics models, in this study we employed several performance assessment measures (Pal et al. 2016):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (true\ classification\ of\ all\ cases) \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (true\ classification\ rate\ of\ the\ positive\ cases) \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \quad (true\ classification\ rate\ of\ the\ negative\ cases) \tag{4}$$

$$AUC\ (area\ under\ the\ ROC\ curve) = [x = \text{Sensitivity},\ y = 1 - \text{Specificity}] \tag{5}$$

In these formulas, *TP, TN, FP, FN* represents the number of cases that fall under true positive, true negative, false positive, and false negative counts, respectively. The overall accuracy, shown by Eq. (2), estimates the proportion of correctly

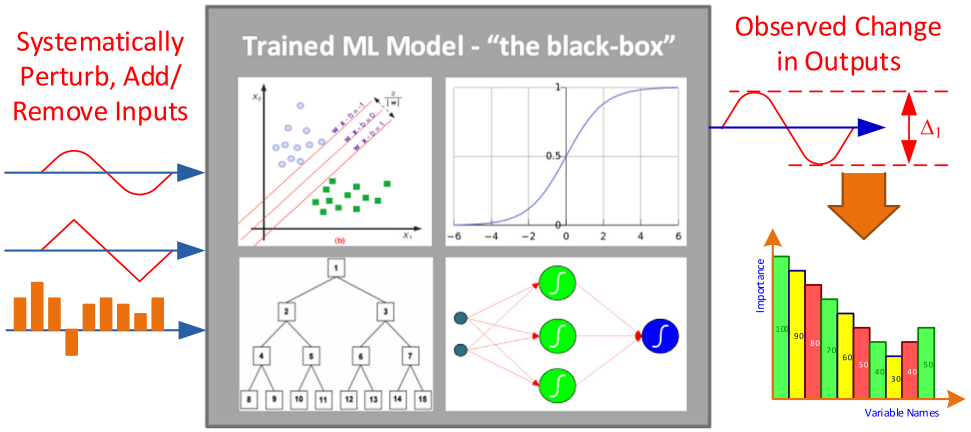

Fig. 4. A graphical depiction of the sensitivity analysis process.

classified test examples (sum of all correctly classified samples divided by all of the samples), and therefore providing the overall ratio of the correct classifications. Sensitivity, specificity, and AUC (**A**rea **U**nder the Receiver Operating Characteristics [ROC] **C**urve) shown by Eqs. (3)–(5) respectively, measure the model's ability to predict the individual class labels within itself (i.e., how accurately model predicts the positives and negatives).

3.5. *Sensitivity analysis*

Machine learning algorithms are really good at capturing complex relationships between input and output variables (producing very accurate prediction models), but are not nearly as good at explaining how they do what they do (i.e., model transparency). In order to mitigate this deficiency (also called the "black-box syndrome"), machine learning community developed several sensitivity analysis methods. In the context of predictive modeling, sensitivity analysis refers to an exclusive experimentation process aimed at discovering the cause and effect relationship between the input and output variables (Davis, 1989).

The sensitivity analysis method used in this study is based on the experimental process of systematically removing input variables, once at a time, from the model and observing the impact of the absence of this variable on the predictive performance of the machine learning model. The model is trained and tested for each input variable (i.e., its absence in the input variable collection) to measure its contribution/importance to the model. A graphical depiction of the process is shown in Fig. 4.

This method is often used for support vector machines, decision trees, logistic regression as well as for artificial neural networks. Saltelli (2002), in his sensitivity analysis book, formalized the algebraic representation of this measurement process (see Eq. (6)).

$$S_i = \frac{V_i}{V(F_t)} = \frac{V(E(F_t|X_i))}{V(F_t)} \tag{6}$$

In the denominator of the equation, $V(F_t)$ refers to the variance in the output variable. In the numerator, $V(E(F_t|X_i))$, E is the expectation operator to calls for an integral over parameter X_i; that is, inclusive of all input variables except X_i, the V, the variance operator applies a further integral over X_i. The variable contribution (i.e., importance), represented as S_i, for the i^{th} variable, is calculated as the normalized sensitivity measure. In a later study, Saltelli et al. (2004) proved that Eq. (6) is the most probable measure of model sensitivity that is capable of ranking input variables (i.e., the predictors) in the order of importance for any combination of interactions including the non-orthogonal relationships amongst the input variables. In order to properly combine the sensitivity analysis results for several prediction methods, we used an information fusion based methodology. Particularly, we modified the Eq. (6) in such a way that the sensitivity measure of an input variable n based on the information obtained (i.e., fused) from m number of prediction models can be shown as in Eq. (7).

$$S_{n(\,fused)} = \sum_{i=1}^{m} \omega_i S_{in} = \omega_1 S_{1n} + \omega_2 S_{2n} + \ldots + \omega_m S_{mn} \tag{7}$$

In this equation ω_i represents the normalized contribution/weight for each prediction model where the level of contribution/weight of a model is calculated as a function of its relative predictive power. In this study, we used four prediction model types (i.e., $m=4$) and calculated the sensitivity/importance measure of a variable (i.e., the nth variable) for a specific model type (i.e., the ith model type) using S_{in}.

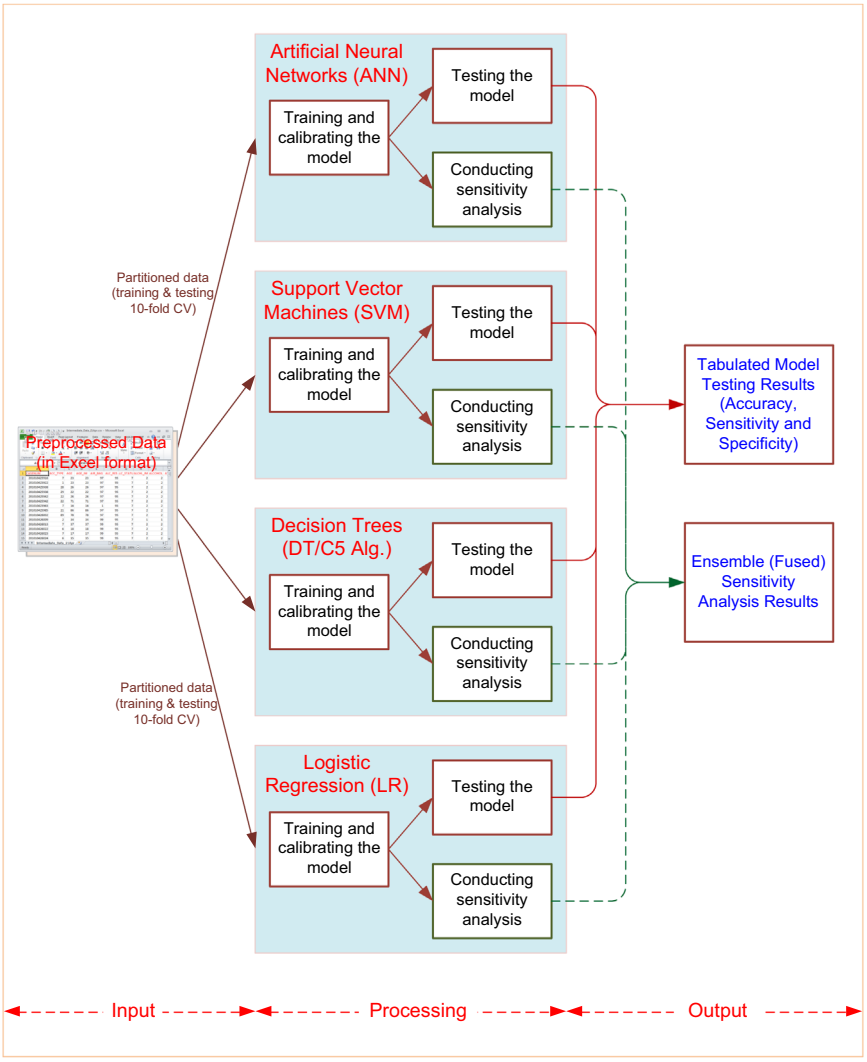

Fig. 5. The process of model building, testing and validation.

4. Results and discussion

Fig. 5 pictorial represents the process of model training (i.e., building), testing (i.e., validating) and importance measuring (i.e., assessing variable importance with information-fusion based sensitivity analysis). As shown on the left hand-side of the figure, the input data is pre-processed and converted to a flat file in Excel format. As part of the cross validation procedure, the dataset is then randomly split into 10 mutually exclusive partitions—to be used as training and testing sets recursively for each model types—that resulted in 40 prediction models. As the right hand-side of the figure shows, the prediction accuracy and sensitivity analysis results of all models are accumulated and represented using aforementioned performance metrics.

Table 2 shows the predictive accuracies of all four model types. Specifically, it shows the confusion matrices, overall accuracy, sensitivity, specificity and area under the ROC curve measures obtained using *10*-fold cross validation for all four mode types. As the results indicate, SVM was the most accurate classification techniques with better than 90% overall accuracy, comparably high sensitivity and specificity, and an AUC value of 0.928 (out of maximum 1.000). The next best model type was C5 decision tree algorithms with slightly better accuracy measured than ANN. The last in the accuracy ranking was LR, also with decent accuracy measures but not as good as the machine learning methods.

Even though the accuracy measures obtained from all four model types are high enough to validate our methodology, the main goal of this study was to identify and prioritize the significant risk factors influencing the level of injury severity sustained by drivers during an automobile crash. To achieve this goal, we conducted sensitivity analysis on all of the developed prediction models. Employing the procedure described in the previous section, focusing on each mode type individually, we calculated the variable importance measures for each fold, and then summed the 10 vectors for each model

Table 2
Tabulation of all prediction results based on *10*-fold cross validation.

Model Type	Confusion Matrices			Accuracy	Sensitivity	Specificity	AUC
		Low	High				
Artificial Neural Networks (ANN)	Low	12864	1464	85.77%	81.31%	89.78%	0.865
	High	2409	10477				
Support Vector Machines (SVM)	Low	13192	1136	**90.41%**	88.55%	92.07%	**0.928**
	High	1475	11411				
Decision Trees (DT/C5)	Low	12675	1653	86.61%	84.55%	88.46%	0.879
	High	1991	10895				
Logistic Regression (LR)	Low	8961	2742	76.97%	77.27%	76.57%	0.827
	High	3525	11986				

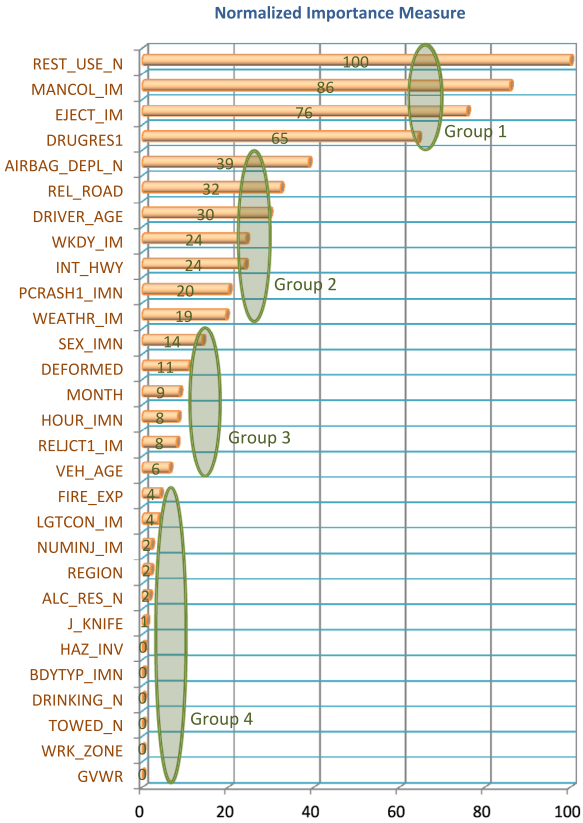

Fig. 6. Variable importance values.

types. In order to properly fuse (i.e., ensemble) the sensitivity analysis results for all four model types we used Eq. (7). That way, we wanted to make sure that the models contributed to the fused/combined variable importance values based on their cross validation accuracy. That is, the best performing model type had the largest weight/contribution while the least performing model type had the smallest weight/contribution. The fused variable importance values are tabulated, normalized and then graphically presented in Fig. 6.

Examination of the sensitivity analysis results reveals four somewhat distinct risk groups, each comprising four to eight variables. The top group, in an order from most to least importance, included REST_USE_N (whether the seat belt of any other restraining system was used), MANCOL_IM (manner of collision), EJECT_IM (whether the driver was ejected from the car), and DRUGRES1 (results of the drug test). According to the combined sensitivity analysis results of all prediction models, these four risk factors seem to be significantly more important than the rest. Of these four, use of restraining system

(REST_USE_N) is the outright frontrunner. As suggested in some of the previous studies (Delen et al., 2006; Abay et al., 2013), the use of seat belt is one of the most important factors to determine the level of injury severity. As indicated in Delen et al. (2006), compared to minor automobile crashes, use of seat belt (or any other straining system) becomes significantly more important predictor of injury severity for more serious crash that involves multi-cars, high speed and adverse driving conditions. The second most important predictor in the group one came out as MAN_COL_IM (manner of collision), which specifies the orientation of the motor vehicle while involved in the crash that includes front-to-front, rear-end, head-on, angle, sideswipe, and so on. The importance of this variable is consistent with some of the previous studies including Kononen et al. (2011), Jung et al. (2011) and Abaya et al. (2013). The next most important variable, EJECT_IM (whether the driver was ejected from the car) is an obvious and at the same time an inquisitive factor. It is inquisitive because it is highly "correlated" with the use of restraining system. Usually, in situations where similar information is provided to a machine learning algorithm (using two or more variables), algorithm picks the most informative one and ignores the other. In this case, all of the prediction models found these two variables collectively predictive of the injury severity level. The last variable in the first group, DRUGRES1 (results of the drug test), signifies the contribution of drug use in injury severity in traffic accidents. Represented with a continuous numerical value, where larger values indicate high dosage of drug, this variable also finds validation from the existing literature, although not as prevalent as it is found in this study.

The second group of predictors included AIRBAG_DEPL_N (whether the airbag was deployed), REL_ROAD (relation to the road—entering, exiting, lane-changing or driving), DRIVER_AGE (the age of the driver), WKDY_IM (the day of the week designation), INT_HWY (whether it was an interstate highway), PCRASH1_IMN (movement/activity prior to crash), WEATHR_IM (weather conditions). In this second group of risk factors, the age of the driver and the weather conditions at the time of accident are perhaps the most widely investigated variables in the previous studies. Ordinary reasoning may suggest that when the weather conditions are not favorable, the chance of severe crashes and hence possibility of severe injury is more possible. Similarly, if the numerical value for the age of the driver is small (new inexperienced drivers) or large (elderly driver) the chances of injury severity increase (Dissanayake and Lu, 2002, Abay et al. 2013, Zeng and Huang, 2014). In our study these two variables found to be moderately significant, but not as significant as they were covered in the previous literature. This can partially be attributed to us using a more inclusive, non-linear modeling methodology that is capable of identifying relative importance of risk factors in a collective and holistic manner. Besides, rational thinking also suggests that, if the age is in the riskier range (very young or elderly) or if the weather conditions are not favorable, the driver compensated for the deficiencies by paying extra attention to the traffic and driving slower.

The third group of predictors included SEX_IMN (the gender of the driver), DEFORMED (the level of deformation of the car), MONTH (month of the year), HOUR_IMN (hour of the day), RELJCT1_IM (related to a junction), VEH_AGE (the age of the vehicle). Even though these factors are also influencing the prediction of injury severity, their contribution is rather small. Similarly, some of the previous regression based, small scale focused studies have also found gender of the driver and the hour of the day among the moderately significant predictors. For instance, Delen et al (2006) found that for the gender of the driver is significant predictor for less severe injuries while it is not a significant factor for more severe injury crashes. Another stimulating study suggests that while females are more likely to experience a more severe injury, males are more likely to be involved in fatal accidents Abdelwahab and Abdel-Aty (2001). Common knowledge suggests that the time of the day (early morning hours, daylight hours, evening hours and mid-night hours) ought to be a significant predictor of traffic crashes and resulting injury severities, especially if the other related variables ignored. A counter argument suggests that the cause of crashes and resulting injuries may be better explained by other, more specific, factors such as driver characteristics, involvement of drug or alcohol, road and environmental conditions, and therefore are not heavily explained by *time of the day* risk factor.

The last group of predictors included FIRE_EXP (the involvement of fire)LGTCON_IM (light conditions), NUMINJ_IM (number of injured in the accident), REGION (the region—Northeast, Midwest, South, and West), ALC_RES_N (results of alcohol test), J_KNIFE (whether jackknifed), HAZ_INV (involvement of hazardous materials, BDYTYP_IMN (body type of the car), TOWED_N (whether the car was towed). DRINKING (police reported alcohol involvement), GVWR (gross weight of the vehicle), WRK_ZONE (whether the crash was at a work zone). The significance of these variables (i.e., level at which they are contributing to the prediction of injury severity of the driver) have been found to be rather marginal. Even though some of the previous studies found some of these variables (i.e., light conditions, involvement of alcohol, body type and the weight of the vehicle) as significant risk factors, because of the rich representation of all of the other variables, our study did not find them very significant.

5. Summary and conclusion

In this paper we presented the results of our investigative study which was conducted for the purposes of identifying person, vehicle, and accident related risk factors that are influential in making a difference in the level of injury severity sustained by a driver in a car crash. After numerous experimentation with possible model types and modeling parameters, the best performing four model types—neural networks, support vector machines, decision trees and logistic regression—are employed to develop and test forty prediction models (4^* *10*-fold $= 40$). In order to capture the complex relationships a large and feature rich crash data set is obtained and meticulously preprocessed for the model building efforts. According to the cross validation results, support vector machines are the most accurate predictor of the injury severity levels followed by

decision trees and neural networks. Although logistic regression models also showed reasonably good prediction performance, among the four model types used in the study, they were the least accurate.

Prediction of the injury severity sustained by a passenger in an automobile crash, although thought-provoking and rather challenging, all by itself may not have much practical value. What is more valuable is the true identification and understanding of the crash related factors that can increase (or decrease) the risk of injury severity. So that, based on the finding, further technological and behavioral enhancements can be implemented to improve the overall traffic safety situation. Initial identification and subsequent assessment of the traffic safety related factors can be accomplished by a critical analysis of the prediction models. Since these models captures the mathematical relationship between the crash-related risk factors and the injury severity levels, a methodical analysis of these relationships can reveal the relative importance of these factors. Sensitivity analysis is the method that is commonly used for this purpose. In this study, as opposed to conducting sensitivity analysis on one model (perhaps on the best performing model), we choose to use an all-inclusive ensemble approach. That is, we fused the sensitivity analysis results of all prediction models by assigning weights to each model type based on its predictive power, where the better performers contributing the final results more heavily. We believe and hope that employing an ensemble approach produces more reliable and improved variable importance results. Sensitivity analysis results showed that use of a restraining system (i.e., seat belt), manner of collision and drug involvement are the top predictors of the injury severity. Contrary to most previous studies, some of the most commonly pronounced person, crash and vehicle related risk factors such as age or gender of the driver, weather or light conditions, type and weight of the of vehicle were not found to be among the top predictors of injury severity levels. It is worth mentioning that although sensitivity analysis performed on predictive models provide an invaluable insight on the ranked importance of independent variables (i.e., factors), it does not capture and/or explain the variables' directional contribution to the target variable (injury severity level), which can be mentioned as a limitation of this type of analytics studies.

Data mining (also known as knowledge discovery in databases, or predictive analytics) is meant to "discover" new and useful patterns (knowledge nuggets) in large databases (often secondary data that is collected for transactional and/or reporting purposes). The main keyword here is the "usefulness/practicality" of the discovered patterns. In this study, we designed and developed several predictive analytics models, and also discovered the variable importance measures. From the practicality standpoint, these variable importance measures can potentially be used to (1) identify features to improve in vehicles (e.g., better design and adherence of restraining systems), (2) improve environment and/or road related characteristics (e.g., lighting, surface, etc.), (3) build awareness towards high-risk personal and behavioral factors that contribute to the severity of injuries (e.g., importance of wearing seatbelts, effects of being under the influence, etc.), and (4) better planning and deployment of first responders to an accident (using predictive models to assess the potential severity of injuries).

There is no doubt that we are living in the era of Big Data and Analytics. The unprecedented popularity and the widespread use of analytics can be attributed to tree main factors: (1) desperate need for knowledge to do the best with constraint resources, (2) availability of data and information infrastructure (both hardware and software) and, (3) the affordability and increased capability of these evince-based decision support methods and tools. Data and analytics driven risk assessment and management is at the center of this analytics movement. Timely and proper use of analytics to derive actionable insight require a methodical and holistic approach. With this study we wanted show the viability, and to some extent the superiority, of data analytics approach in critical investigation of the risk factors that are related to different levels of injury severities in car crashes.

References

Abay, K.A., 2015. Investigating the nature and impact of reporting bias in road crash data. Transp. Res. Part A: policy Pract. 71, 31–45.

Abay, K.A., Paleti, R., Bhat, C.R., 2013. The joint analysis of injury severity of drivers in two-vehicle crashes accommodating seat belt use endogeneity. Transp. Res. Part B: Methodol. 50, 74–89.

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalizes intersection. Transp. Res. Rec. 1746, 6–13.

Alkheder, S., Taamneh, M., Taamneh, S., 2017. Severity prediction of traffic accident using an artificial neural network. J. Forecast. In press.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Chen, C., Zhang, G., Qian, Z., Tarefder, R.A., Tian, Z., 2016a. Investigating driver injury severity patterns in rollover crashes using support vector machine models. Accid. Anal. Prev. 90, 128–139.

Chen, C., Zhang, G., Huang, H., Wang, J., Tarefder, R.A., 2016b. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Accid. Anal. Prev. 96, 79–87.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, London.

Davis, G., 1989. Sensitivity analysis in neural net solutions. IEEE Trans. Syst., Man, Cybern. 19, 1078–1082.

Delen, D., 2015. Real-World Data Mining: Applied Business Analytics and Decision Making. Financial Times Press (a Pearson Publishing Co.), Upper Saddle River: New Jersey.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38 (3), 434–444.

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object–passenger car crashes. Accid. Anal. Prev. 34 (5), 609–618.

Friedman, D. (2014). Oral Testimony before the House Committee on Energy and Commerce, by the Subcommittee on Oversight and Investigations, April 1, 2014, ⟨www.nhtsa.gov/Testimony⟩ (accessed October 2014).

GES, 2014. National Automotive Sampling System General Estimates System. NASS GES.
Hauer, E., 2006. The frequency–severity indeterminacy. Accid. Anal. Prev. 38 (1), 78–83.
Hauer, E., 2009. Speed and safety. Transp. Res. Rec.: J. Transp. Res. Board 2103, 10–17.
Hornik, K., Stinchcombe, M., White, H., 1990. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Netw 3 (5), 551–560.
Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accid. Anal. Prev. 40 (1), 45–54.
Jung, S., Qin, X., Noyce, D.A., 2011. Injury severity of multivehicle crash in rainy weather. J. Transp. Eng. 138 (1), 50–59.
Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. Transp. Res. Part C: Emerg. Technol. 19 (3), 387–399.
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In S. Wermter, E. Riloff and G. Scheler (Eds.). In: Proceeding of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI), Montreal, Quebec, Canada. San Francisco, CA. Morgan Kaufman Publishing; pp. 1137–1145.
Kononen, D.W., Flannagan, C.A., Wang, S.C., 2011. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accid. Anal. Prev. 43 (1), 112–122.
Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using support vector machine models. Accid. Anal. Prev. 40 (4), 1611–1618.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A: Policy Pract. 44 (5), 291–305.
Lui, K.J., McGee, D., 1988. An application of conditional logistic regression to study the effects of safety belts, the principal impact points, and car weights on drivers' fatalities. J. Saf. Res. 19 (4), 197–203.
Mergia, W.Y., Eustace, D., Chimba, D., Qumsiyeh, M., 2013. Exploring factors contributing to injury severity at freeway merging and diverging locations in Ohio. Accid. Anal. Prev. 55, 202–210.
NHTSA (2014) National Highway Traffic Safety Administration (NHTSA's) General Estimate System (GES), ⟨www.nhtsa.gov⟩ (accessed January 20, 2017).
O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accid. Anal. Prev. 28 (6), 739–753.
Pal, C., Okabe, T., Kulothungan, V., Sangolla, N., Manoharan, J., Stewart, W., Combest, J., 2016. Factors influencing specificity and sensitivity of Injury Severity Prediction (ISP) algorithm for AACN. Int. J. Automot. Eng. 7 (1), 15–22.
Park, S., Jang, K., Park, S.H., Kim, D.-K., Chon, K.S., 2012. Analysis of injury severity in traffic crashes: a case study of Korean Expressways. J. Civil. Eng. 16 (7), 1280–1288.
Quinlan, J., 1986. Induction of decision trees. Mach. Learn. 1, 81–106.
Quinlan, J., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.
Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput. Phys. Commun. 145, 280–297.
Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M., 2004. Sensitivity Analysis in Practice – A Guide to Assessing Scientific Models. John Wiley and Sons.
Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.
Thammasiri, D., Delen, D., Meesad, P., Kasap, N., 2014. A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition. Exp. Syst. Appl. 41 (2), 321–330.
Wood, D.P., Simms, C.K., 2002. Car size and injury risk: a model for injury risk in frontal collisions. Accid. Anal. Prev. 34, 93–99.
Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. Transp. Res. Rec.: J. Transp. Res. Board 2241, 51–58.
Zeng, Q., Huang, H., 2014. A stable and optimized neural network model for crash injury severity prediction. Accid. Anal. Prev. 73, 351–358.
Zeng, Q., Wen, H., Huang, H., 2016. The interactive effect on injury severity of driver-vehicle units in two-vehicle crashes. J. Saf. Res. 59, 105–111.