Analytical Assessment of Highway Fatalities in United States: Frontier Approaches

Conference Paper · January 2014

CITATIONS
0

READS
38

3 authors, including:

Gary R Weckman
Ohio University
107 PUBLICATIONS **420** CITATIONS

SEE PROFILE

Proceedings of the 2014 Industrial and Systems Engineering Research Conference
Y. Guan and H. Liao, eds.

Analytical Assessment of Highway Fatalities in United States: Frontier Approaches

Naime Fulya Kayaalp, Can Celikbilek* and Gary Weckman

Industrial & Systems Engineering,
Russ College of Engineering, Ohio University
Athens, Ohio, USA, 45701
fk580711@ohio.edu, cc340609@ohio.edu*, weckmang@ohio.edu
Tel: (740)-593-1548, Fax: (740)-593-0778

Abstract

A 5.3% increase in motor vehicle traffic crashes in 2012 [1] brings up the discussion of related traffic safety parameters. This paper considers the analytical assessment and evaluation of various highway safety factors that will eventually trigger fatalities. The related safety parameters are mainly divided into four categories-- economical investment, system usage, road condition, and personal safety. Three data mining algorithms-- K-nearest Neighbors algorithm (KNN), Random Forest and Support Vector Machine (SVM), and also a probabilistic Artificial Neural Network (ANN)-- are used for the prediction of highway fatalities among the eight different safety indicators. According to the Bureau of Transportation Statistics' most recent available data, the analysis of this study covers the years from 2003 to 2011. The preliminary results indicated that out of the three, the proposed Random Forest data mining approach predicted the data with the highest percentage. The sensitivity analysis is conducted and, according to the results, the bad-road condition is found to be the most important factor that affects the highway fatalities. This research shows the significant relationship between safety indicators and highway fatalities, and also provides guidance for policy makers when it comes to preventing highway fatalities in United States.

Keywords

Transportation Safety, Data Mining Algorithms, Artificial Neural Networks, Highway Fatalities, Prediction Assessment

1. Introduction

For almost half a century, the U.S Department of Transportation has made tremendous progress and put a lot of effort into tackling the highway fatalities all around the United States. However, recently the U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA) released 2012 Fatality Analysis Reporting System (FARS) data indicating that highway deaths increased to 33,561 in 2012, which amounts to 1,082 more fatalities than the previous year. In other words, Americans drove the same amount of miles in 2012 compared to 2011 and statistics showed that fatalities increased about 3.3%. Overall, highway deaths claim more than 30,000 lives each year. Therefore, it is obvious and significant that a profound assessment is necessary for elaborating upon the causes of highway fatalities[1].

For many years, several statistical methods and techniques were conducted in order to analyze highway traffic safety and assist policy makers in their decision making processes in the transportation safety field. In the field of Artificial Neural Networks (ANN), Abdelwahab and Abdel [2] studied the traffic safety of toll plazas and the impact of electronic toll collection (ETC) systems on highway safety. Toll plaza traffic accident reports of the Central Florida expressway system for years 1999 and 2000 were studied and the analysis was conducted on accident location with respect to the plaza structure (before, at, after plaza) and driver injury severity (no injury, possible, evident, severe injuries). Two Artificial Neural Networks (ANN) paradigms were analyzed. These were the Multi-Layer Perceptron and Radial Basis Functions neural networks. The performance was compared with logic models. The modeling and analysis showed that especially medium/heavy-duty trucks had a higher risk of being involved in accidents at the toll plaza structure. Also, the analysis indicated that main-line toll plazas had a higher percentage of accidents occurring upstream of the toll plaza. Delen et al.,[3] conducted a series of artificial neural networks to model the potentially non-linear relationships between the injury severity levels and crash-related factors. Factors that could lead to more severe injuries in the event of an accident include demographics, behavioral characteristics

of people, environmental factors, and roadway conditions at the time of the accident occurrence. Abdel and Abdelwahab [4] studied the use of two well-known artificial neural network (ANN) paradigms: the multilayer perceptron (MLP) and fuzzy adaptive resonance theory (ART) neural networks in analyzing driver injury severity. The main emphasis of that study was to investigate the viability and potential benefits of using the ANN in predicting driver injury severity, conditioned on the premise that a crash had occurred. The performance of ANN model was compared with the calibrated ordered probit model. The results indicated that ANN (particularly MLP) provided a more accurate prediction capability over other traditional methods (73.5%). Also, results indicated that gender, vehicle speed, seat belt use, type of vehicle, point of impact, and area type affected the likelihood of injury severity levels.

Many research projects have been conducted in the area of data mining as well. Tseng et al., (2005) applied data mining techniques to discover the relationship between driver inattention and motor vehicle accidents. The research was focused on the Washington, D.C., and Maryland area from the years 2000 to 2003. The results showed that when inattention and physical/mental conditions take place at the same time, the driver has a higher tendency of being involved in a crash. Additionally, this research showed that the relative importance of colliding into a moving vehicle is two times higher than into a fixed object. That research was one of the few research papers that utilized data mining techniques to analyze the possible relationships between driver inattention and motor vehicle crashes. Bayam et al.,[6] applied data mining techniques such as neural networks and decision trees to examine the relationships between senior driver characteristics such as age, gender, driving cessation, alcohol usage, fragility, medical condition, the perception of the road signings, and accidents. The provided data mining techniques performed well in terms of prediction of the accidents. Similarly, Chan and Chen [7] applied Classification and Regression Tree (CART) data mining techniques to analyze one of the national freeways in Taiwan. The study's scope mainly covered the years 2001-2002. Briefly, the CART model and a negative binomial regression model were developed to assess the empirical relationship between traffic accidents and highway geometric variables, traffic characteristics, and environmental factors. It was found that CART demonstrated a good alternative method for analyzing freeway accident frequencies, and also that traffic volume and precipitation were the most important indicators for traffic accidents. Chong et al.,[8] discussed the performance of four machine learning paradigms, which is applied to the modeling of injury severity that occurred in traffic accidents. They considered neural networks trained using hybrid learning approaches, support vector machines, decision trees, and a concurrent hybrid model involving decision trees and neural networks. Experiment results indicated that the hybrid decision tree-neural network approach outperformed the others. Moreover, Chong et al., [9] studied the severity of injury resulting from traffic accidents using artificial neural networks and decision trees. The data set was obtained from the National Automotive Sampling System (NASS) General Estimates System (GES). Different variables such as drivers' age, gender, alcohol usage, restraint system, vehicle body type, vehicle age, road surface conditions, and lighting conditions were considered in relation to the different impact types (different collision types). Experiment results indicated that in all the cases, the decision tree outperforms the neural network approach. The research analysis also revealed that the three most important factors in fatal injuries were a driver's seat belt usage, the light condition of the roadway, and the driver's alcohol usage.

All in all, the literature review that has been done so far indicates that there were no studies covering three data mining algorithms-- K-nearest Neighbors algorithm (KNN), Random Forests and Support Vector Machines (SVM) and a probabilistic Artificial Neural Network (ANN) – that were used for the prediction of highway fatalities while also considering the variety of safety indicators. Our research mainly considers four categories-- economical investment, system usage, road condition, and personal safety-- for the prediction of highway fatalities. The flow of the paper is as follows; section 2 discusses background information of the aforementioned three different data mining techniques; section 3 briefly indicates the research aims; and section4 discusses the problem in general. Section 5 presents the analysis results, and section 6 relates the conclusions and relevant discussions.

2. Background

This section explains all of the data mining techniques that are used in this research.

2.1. K-Nearest Neighbor (KNN) Classification

KNN is an easy classification technique to understand and implement in data mining literature. Basically, the technique tries to find a neighbor in the training set that is closest to the test object, and identifies the class of that

test object based on the dominance of a particular class within its nearest neighbor. At first, the number of neighbors denoted as k should be identified and then the algorithm runs accordingly.

To classify an unlabeled data point, the closest neighbor to that data point is identified by calculating the Euclidean distance between the neighbors and said data point. Then the class labels of these neighbors are used to identify the class label of the point in question. Many issues might affect the performance of the KNN algorithm. Firstly, the performance of this algorithm is heavily depending on the *k* value. It is realized that when *k* is too small, then the results become very sensitive. On the other hand, if *k* is too large, then the neighborhood may encompass too many points from the other classes. The grouping of neighbors is another factor that affects the performance. A smaller distance between data points implies a greater likelihood of having the same class. Therefore, if data points that are close to each other form a group as a neighbor, classification accuracy is expected to be higher. During this process, conflicting neighbors should be prevented. [10]. All in all, KNN algorithms are not quick learners; modeling is easy, but the classification of unknown objects is relatively challenging, due to the computation of k-nearest neighbors of the labeled objects. The KNN classification technique requires some computation requirements and time, but the classification accuracy can be improved and the speed can be greatly enhanced.

2.2 Support Vector Machines (SVM)

SVM offers one of the most accurate and robust algorithms in data mining literature. The aim of SVM is to find the best classification function to differentiate between different classes in the data set. The metric for the best classification is defined geometrically. In the SVM approach, each data point is defined as a vector in the space. For a linearly separable data set, a linear classification function (hyperplane) separates these vectors such that the distance between each class represented by the vectors is maximized. However, sometimes it is not possible to have a linear classification function to separate the vectors into the classes they represent. Therefore, SVMs may make use of different types of functions including linear, polynomial, sigmoid, and exponential functions, which are all called kernels. The choice of kernel is important; it provides better data performance when chosen correctly.

2.3 Random Forest Classification

Random forest is an algorithm for classification that uses an ensemble of classification trees. In the first step of random forests, pre-defined numbers of classification trees are constructed randomly with the training data. When a data instance needs to be classified, all of the trees in the forest classify the same data instance. The dominant class is chosen as the resulting class of the aimed data instance. Random forests can be used for two-class and multi-class problems, and also can be used when there are many more variables than observations. Random forest is a good predictive algorithm, although noise is involved in the dataset; hence, no pre-selection process can be made. Moreover, the categorical and continuous predictors can be handled and interactions among predictor variables can be incorporated [11].

2.4 Probabilistic Artificial Neural Networks (PNN)

Neural networks are currently applied to classify the patterns based on learning from examples. Different neural network algorithms utilize different learning rules, but in general they identify the pattern statistics from a set of training samples. Currently, back propagation uses heuristic approaches, which require long computation times for training, and each sigmoid activation function has its own unique characteristics. However, these features that are under certain conditions are tackled by probabilistic neural networks (PNN) by considering Bayesian decision strategy and non-parametric estimators of probability density function [12]. In PNN, instead of using sigmoid activation functions which are commonly used for back-propagation, the non-linear operation is used to accomplish the interconnections for neuron activation. It was found out that PNN has higher prediction accuracy with a higher speed advantage relative to back-propagation. All in all, PNN has various advantages such as; the shape of the decision surfaces can be complex; Bayesian optimal decision surfaces are used; erroneous samples are tolerated; and higher network performance is obtained [12].

Given the promising features of these different classification algorithms, it is important to understand the comparisons among each other and observe which one provides the better prediction accuracy for highway fatalities.

3. Research Aims

The aim of this research is to apply different data mining algorithms which will be used for the prediction of highway fatalities among the eight different safety indicators. The primary emphasis of this study is to mitigate the

number of highway fatalities in United States. Moreover, important factors will be mentioned and indicated for the guidance of decision makers when it comes to preventing highway fatalities in United States.

4. Problem Description

Identification of safety performance indicators are always an issue. Therefore, this paper uses safety parameters based on those given by the National Highway Traffic Safety Administration (NHTSA). There are many indicators that eventually affect the highway fatalities. However, the scope of the study needs to be defined further. Therefore, Hermans et al.,[11]'s study is considered as the reference for the identification of parameters. Hermans et al.,[11] considered alcohol and drug usage, speed, protective systems, vehicle, and infrastructure as safety performance indicators for affecting the fatalities in transportation. Inspired by that research, eight safety performance indicators in highways were determined in this paper for assessing the highway fatalities. These safety performance indicators are; safety belt usage, fatalities involving high blood alcohol, licensed drivers per registered vehicles, vehicle miles traveled, highway safety expenditure, total road length, and road condition for the states in question. The outcome was considered as highway fatalities. The analysis years were chosen from 2003-2011. The dataset is gathered from RITA's website, where each data entry corresponds to a state in the United States. However, state information was not used during the analysis. Thus, each data entry has been considered as an individual measurement.

5. Analysis of Results

For the experiments, popular data mining algorithms such as k-nearest neighbors (KNN), support vector machines (SVM), and Random Forests were used. Experiments with these algorithms were performed in Orange Data Mining Software. Additionally, a probabilistic neural network (PNN) model was developed through the NeuroSolutions Software.

For each model, appropriate parameters were identified as follows. In the KNN model, the number of neighbors is data dependent and needs to be tuned appropriately. During the experiments, it is observed that having a number of neighbors less than 5 or more than 5 lowers the accuracy. Therefore, the number of neighbors was chosen as 5. On the other hand, for the SVM model, a kernel was chosen as an exponential function, which is also called a Radial Basis Function. The equation for the kernel function is shown below:

$$K(x, y) = \exp(-g|x - y|^2) \tag{1}$$

For the above equation, the optimal g is found to be approximately 0.03.

In the random forests model, the only parameter-- which is the number of trees-- was chosen as 10. Lastly, for the PNN model, no additional parameters are specified other than the ones assigned automatically by the NeuroSolutions Software.

Four approaches and parameters that are mentioned above were tested on the same dataset. The dataset has been divided into two-- as training and test data sets. Predictions have been done over the test data sets. According to the results, KNN predicted 86% and SVM predicted 90%, while Random Forests predicted 95% accurately. Meanwhile, the PNN algorithm's prediction performance was observed as 83%. Overall, the best model is achieved in Random Forest in terms of the highest prediction percentage. After assessing the best algorithm for predicting the highway fatalities, the importance analysis was done for the considered parameters by using information gain scoring; parameters are listed below from most effective to least effective.

1. Bad Road Condition
2. Total Road Length
3. High Blood Alcohol
4. Miles Traveled
5. Highway Safety Expenditure
6. Good Road Condition
7. Safety Belt Usage
8. Licensed Drivers per Vehicle

According to the performed analysis, the bad road condition played the most important role in relation to highway fatalities. Road length came in second and blood alcohol percentage came in third, as expected. The miles traveled,

safety expenditure, good road condition, and safety belt usage played increasingly less critical roles in accidents with fatalities. The least important factor was found to be licensed drivers per vehicle.

An analysis was performed again without the input of licensed drivers per vehicle. The results for each of the models are given below: SVM's parameters have been changed, while the others were kept the same. The optimal value is achieved by g=0.50 of RBF function. From the obtained results, the prediction performance of KNN is 86%, the SVM prediction performance is 91%, the Random Forest is obtained as the highest of 97%, and finally, the PNN algorithm predicts the highway fatalities with a percentage of 78. Here again, the best results are obtained by Random Forest algorithm. It can be concluded that having the number of licensed drivers per vehicle is a negligible parameter.

The correlation between some of the inputs to our analysis provided some interesting results, which are interpreted and shown visually below. Figure 1 analyzes the expenditures for different road conditions.

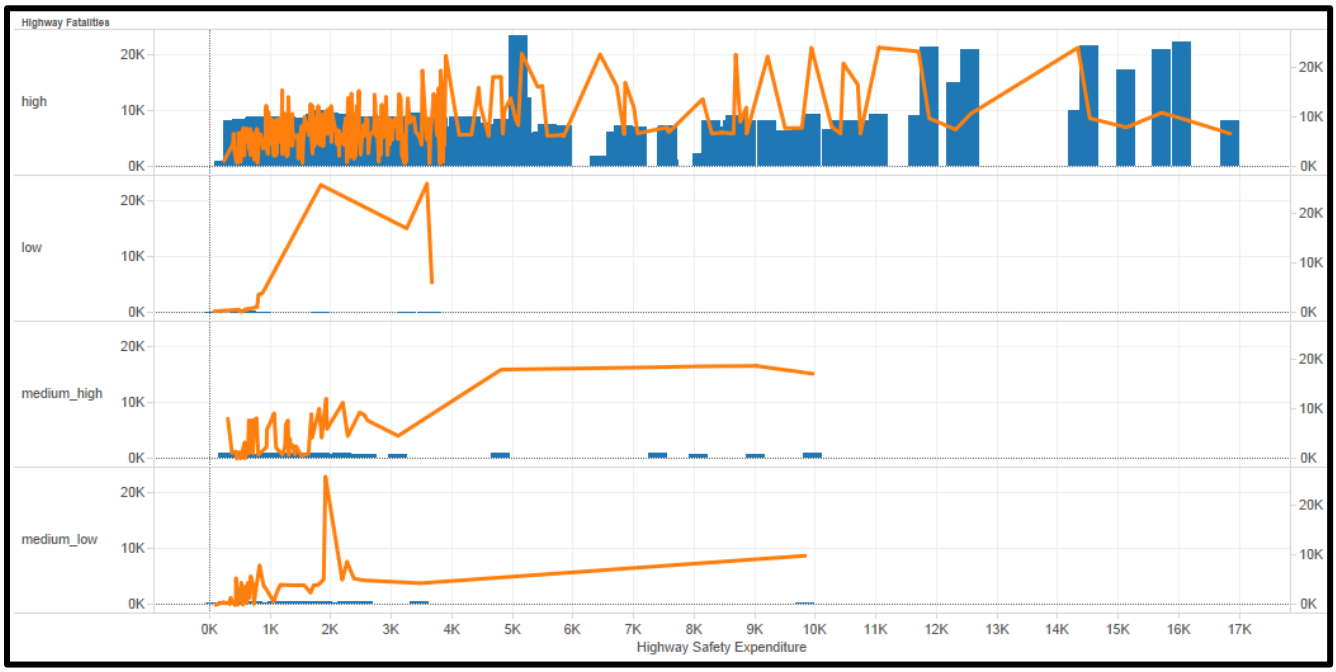

Figure 1: Assessment of Highway Safety Expenditure for Different Road Conditions

The orange line indicates good road conditions (in miles), and the blue bars indicate bad road conditions (in miles). The results obtained from Figure 1 tell us that, in the lowest fatality section (2nd row), the highest levels of good road conditions and lowest levels of bad road conditions were observed, which is not surprising. Additionally, in a high fatality section (1st row), fluctuations are observed for both good roads and bad roads. However, as shown for the other sections, bad roads are more likely to be stable and observed in low levels. This explains the lower levels of highway fatalities in the 2nd, 3rd, and 4th rows compared to the 1st row. It could also be concluded that once the bad road condition levels are the same, the highway fatality rates have an inverse relationship with good road conditions.

According to the analysis results, safety belt usage was not found to be a good indicator of highway fatalities. Figure 2 depicts the relationship between the highway fatalities and safety belt usage.

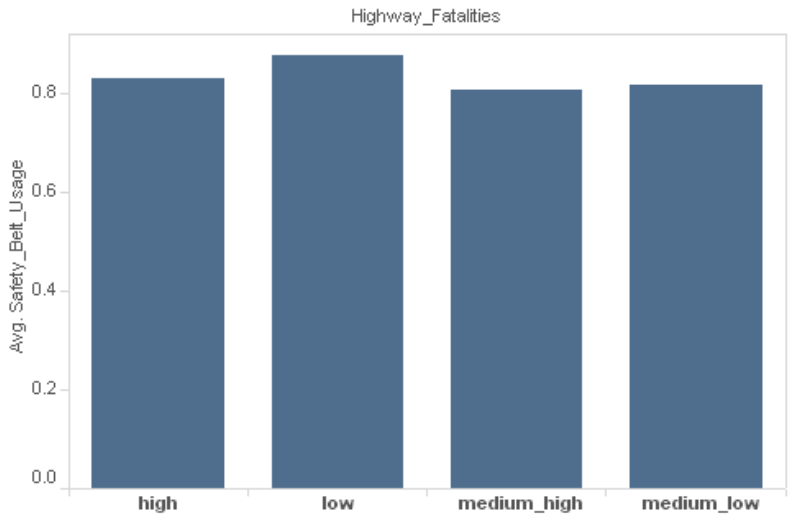

Figure 2: Safety Belt usage

According to Figure 2, safety belt usage is almost stable, and does not affect the prediction results critically. It could be concluded that because of the current rules, regulations signs, and warning technology within the new-generation cars, the drivers usually fasten their seat belts. Hence, safety belt usage is not significantly affecting highway fatalities. In Figure 3, the relationship of different road conditions and highway safety expenditures are reviewed.

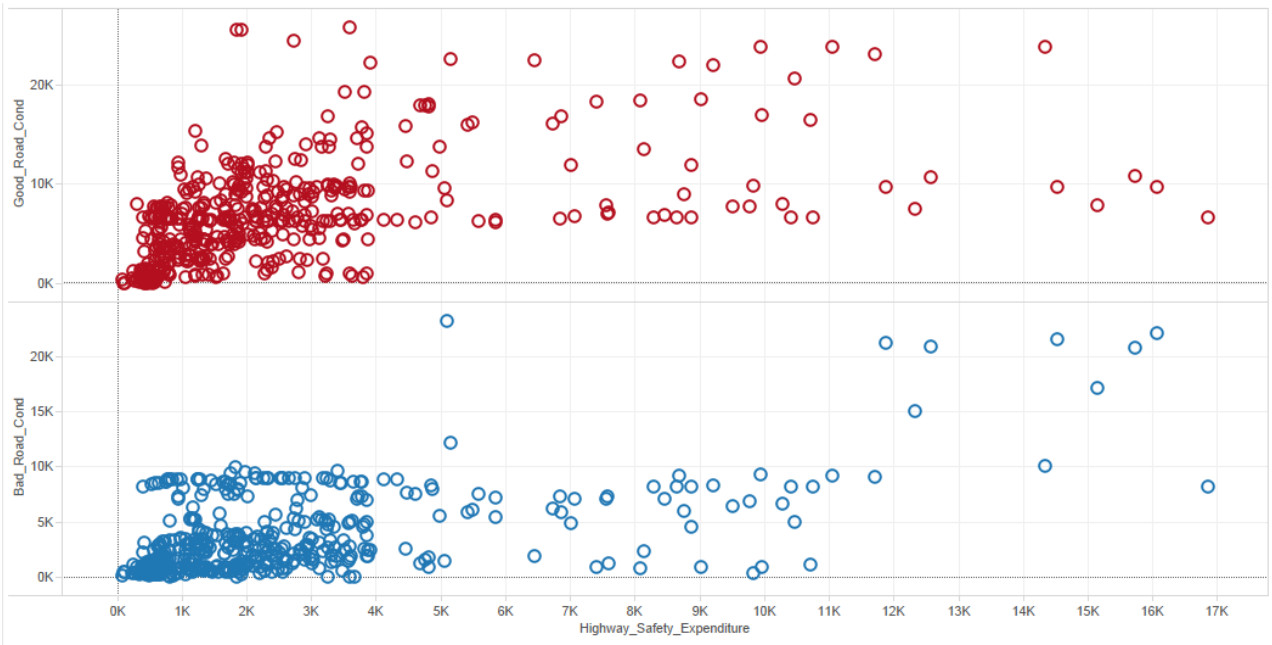

Figure 3: Road Condition vs Highway Safety Expenditure

According to Figure 3, money spent to highways does not often go over 4K. More data points are observed below 10K for road conditions, and below 4K in expenditures for both good and bad road conditions. This means that safety expenditure is not closely related with good or bad road conditions. However, it is interesting to observe that bad road conditions are higher when more money is spent on that section of highway. In reality, a lower percentage of bad roads should be observed when we spend more money on them. This might indicate that money has been spent on the constructing new roads, rather than repairing the existing bad roads. Therefore, it might be beneficial to rearrange policies towards fixing the bad roads and increasing the overall safety budget.

6. Conclusions & Discussions

This paper discusses various highway safety factors that will result in fewer highway fatalities. The related safety parameters are generally divided into four categories-- economical investment, system usage, road condition, and personal safety. In terms of economical investment, highway safety expenditure is considered. In system usage, miles traveled, total road length, and the number of licensed drivers per vehicle are evaluated. In the road condition category, bad and good road conditions (below and above a threshold value) are considered, respectively. In the personal safety category, safety belt usage and high blood alcohol percentage are considered. Three data mining algorithms-- K-nearest Neighbor algorithm (KNN), Random Forest and Support Vector Machine (SVM), and a probabilistic Artificial Neural Network (ANN) -- are used for the prediction of highway fatalities among the eight different safety indicators. The Bureau of Transportation Statistics' most recent available data from 2003 to 2011 for each state is considered. The preliminary results indicated that out of the three, the proposed Random Forest data mining approach predicted highway fatalities with the highest percentage. Table 1 shows the results of each data mining technique.

Table 1: The performance of considered data mining techniques

Data Mining Techniques	Prediction % (w/licensed drivers/ vehicle)	Prediction % (w/o licensed drivers/ vehicle)
KNN	86%	86%
Random Forest	90%	91%
SVM	95%	97%
Probabilistic ANN	83%	78%

The importance analysis is performed and, according to the results, the bad-road condition is found to be the most significant factor that affects highway fatalities. Highway safety expenditures should be more focused on fixing or changing the bad roads, which in turn decreases the fatality rate on highways. This research indicates the relationship between safety indicators and highway fatalities, and also provides guidance for policy makers when it comes to lessening the number of highway fatalities in the United States.

References

[1] "NHTSA Data Confirms Traffic Fatalities Increased In 2012", National Highway Traffic Safety Administration (NHTSA),[Online]November2013,http://www.nhtsa.gov/About+NHTSA/Press+Releases/NHTSA+Data+Confirms+Traffic+Fatalities+Increased+In+2012 (Accessed : 22 December 2013).

[2] Abdelwahab, H. T., & Abdel-Aty, M. A. "Artificial neural networks and logit models for traffic safety analysis of toll plazas" Journal of the Transportation Research Board, 1784(1), 115-125, 2002.

[3] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks" Accident Analysis & Prevention, vol. 38, no. 3, pp. 434–444, 2006.

[4] Abdel-Aty, M. A., & Abdelwahab, H. T. "Predicting injury severity levels in traffic crashes: a modeling comparison" Journal of Transportation Engineering, 130(2), 204-210, 2004.

[5] W.-S. Tseng , H. Nguyen , J. Liebowitz , and W. Agresti , "Distractions and motor vehicle accidents: Data mining application on fatality analysis reporting system (FARS) data files," Industrial Management & Data Systems, vol. 105, no. 9. pp. 1188–1205, 2005.

[6] E. Bayam, J. Liebowitz, and W. Agresti, "Older drivers and accidents: A meta-analysis and data mining application on traffic accident data," Expert Systems with Applications, vol. 29, no. 3. pp. 598–629, 2005.

[7] L.-Y. Chang and W.-C. Chen, "Data mining of tree-based models to analyze freeway accident frequency.," J. Safety Res., vol. 36, no. 4, pp. 365–375, 2005.

[8] M. Chong, A. Abraham, and M. Paprzycki, "Traffic accident analysis using machine learning paradigms," Inform., vol. 29, no. 1, pp. 89–98, 2005.

[9] Chong, M., Abraham A., Paprzycki M., "Traffic Accident Analysis Using Decision Trees and neural Networks" IADIS International Conference on Applied Computing, Portugal, IADIS Press, Pedro, 2004.

[10] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge of Information Systems, vol. 14, pp. 1–37, 2008.

[11]Hermans, E., Brijs, T., Wets, G., Vanhoof, K.,"Benchmarking road safety: lessons to learn from a data envelopment analysis" Accident; Analysis and Prevention 41 (1), 174–182, 2009.