CAIT-UTC-NC62

Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling

FINAL REPORT
June 2020

Submitted by:

Branislav Dimitrijevic, Ph.D.
Assistant Professor

Sina Darban Khales
Research Assistant

Roksana Asadi
Research Assistant

Joyoung Lee, Ph.D.
Associate Professor

Kitae Kim, Ph.D.
Senior Research Associate

John A. Reif, Jr. Department of Civil and Environmental Engineering
New Jersey Institute of Technology
University Heights
Newark, NJ

External Project Manager
Joseph Weiss
New Jersey Division of Highway Traffic Safety

In cooperation with
Rutgers, The State University of New Jersey And
State of New Jersey
Division of Highway Traffic Safety
And
U.S. Department of Transportation Federal
Highway Administration

DISCLAIMER STATEMENT

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

The Center for Advanced Infrastructure and Transportation (CAIT) is a National UTC Consortium led by Rutgers, The State University. Members of the consortium are the University of Delaware, Utah State University, Columbia University, New Jersey Institute of Technology, Princeton University, University of Texas at El Paso, Virginia Polytechnic Institute, and University of South Florida. The Center is funded by the U.S. Department of Transportation.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
CAIT-UTC-NC62		

4. Title and Subtitle	5. Report Date
Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling	June 2020
	6. Performing Organization Code CAIT/New Jersey Institute of Technology

7. Author(s)	8. Performing Organization Report No.
Branislav Dimitrijevic, Ph.D. https://orcid.org/0000-0002-1259-2037, Sina Darban Khales https://orcid.org/0000-0003-4577-5101, Roksana Asadi, https://orcid.org/0000-0003-1286-0318, Joyoung Lee, Ph.D. https://orcid.org/0000-0003-4888-0679, Kitae Kim, Ph.D., https://orcid.org/0000-0003-1027-5625	CAIT-UTC-NC62

9. Performing Organization Name and Address	10. Work Unit No.
John A. Reif, Jr. Department of Civil and Environmental Engineering New Jersey Institute of Technology University Heights Newark, NJ	11. Contract or Grant No. DTRT13-G-UTC28

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Center for Advanced Infrastructure and Transportation Rutgers, The State University of New Jersey 100 Brett Road	Final Report 01/01/2020 – 05/31/2020
	14. Sponsoring Agency Code

15. Supplementary Notes

U.S. Department of Transportation/OST-R
1200 New Jersey Avenue, SE
Washington, DC 20590-0001

16. Abstract

Highway crashes are the most significant challenge to the goal of providing a safe and efficient highway transportation system. They result in significant societal toll reflected in numerous fatalities, personal injuries, property damage, and traffic congestion. To that end, much attention has been given to developing models to study and predict crash occurrence and severity. Most of these models are reactive: they aim to identify the significant crash factors, crash hot-spots and crash-prone roadway locations, analyze and select the most effective countermeasures for reducing the number and severity of crashes. More recently advancements have been made in developing proactive crash risk models, aiming to assess crash risks in a short term, and inform traffic management strategies to prevent and mitigate the negative effects of crashes. This study developed and tested several models for segment-level crash risk and severity assessment considering the data available to most transportation agencies in real time on a regional network scale. The data included roadway geometry characteristics, traffic flow characteristics, and weather condition data. The models included Bayesian Logistics Regression, Decision Tree, Random Forest, Gradient Boosting Machine, K-Nearest Neighbor, and Gaussian Naïve Bayes (GNB). The models were trained and tested using a dataset containing records of 10,155 crashes that occurred on two interstate highways in New Jersey over a period of two years. It was found that for the given dataset the models provided limited predictive value.

17. Key Words	18. Distribution Statement
Crash analysis, crash severity, Crash risk forecasting, machine learning	

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	43	

Form DOT F 1700.7 (8-69)

Acknowledgments

The authors would like to express gratitude to the National University Transportation Center Consortium led by CAIT, as well as the John A. Reif, Jr. Department of Civil and Environmental Engineering at the New Jersey Institute of Technology for providing funding for this research. The authors also thank the external Project Manager, Mr. Joseph Weiss for supporting the proposed research, and the New Jersey Department of Transportation for providing the traffic data used in the study.

Table of Contents

List of Figures

List of Tables

DESCRIPTION OF THE PROBLEM

Research Need

The primary goal and purpose of highway transportation agencies is to provide a safe and efficient highway transportation system. Highway crashes are the most significant challenge to this goal. They result in significant societal toll reflected in numerous fatalities, personal injuries, and property damage. According to the National Highway Traffic Safety Administration, over 8.7 million people were involved in reported highway crashes in the United States in 2018. Among these, there were 33,654 fatalities, and over 1.5 million people were injured, some sustaining incapacitating injuries. Highway crashes are also a major cause of traffic congestion, accounting for about 25% of non-recurring delays. More severe crashes, especially those occurring during peak commuting hours or in adverse weather conditions, may result in prolonged roadway closures and excessive traffic backups, thus affecting the ability of highway operating agencies to efficiently respond to and manage the clearance of crashes. The key for improving highway safety and reducing the number and severity of crashes is in better understanding of how, why, when, and where the highway crashes occur. With this knowledge one can ascertain the necessary actions and strategies for reducing the probability of crash occurrence and their severity. This has been a subject of numerous research studies resulting in a variety of crash risk assessment and cash prediction models. Most of these efforts and models are reactive: they aim to help identify the significant crash factors, identify crash hot-spots and crash-prone roadway locations, analyze and select the most effective countermeasures for reducing the number and severity of crashes.

More recently there has been a great level of interest in proactive crash risk modeling, aiming to assess crash risks in a short term, and use traffic management strategies to prevent the occurrence of highway crashes and mitigate their negative effects on the overall traffic safety and mobility. The analytics resulting from such models can help the highway agencies to strategically plan the deployment of assets dedicated to traffic incident management and take preemptive traffic management actions targeting the locations with elevated crash risk. The underlying assumption of these models is that real-time traffic, geometric and weather conditions can characterize ‘crash-prone’ conditions. These models are focused on identifying crash precursors that are likely to lead to crash occurrence in dynamic traffic environment using high-resolution traffic data (such as traffic monitoring data for 5–10 min intervals), weather characteristics and road geometry. The data analysis methods and techniques employed in developing dynamic crash risk models include different regression analysis models, Bayesian network models, data envelop analysis, and more recently the modeling approaches such as supervised and deep learning model. Different modeling techniques have different advantages and shortcomings. The impetus and motivation for the proposed research is the interest in developing and evaluating effectiveness of a crash risk prediction model for New Jersey highways. Such model would be useful to different transportation agencies in the State by providing means for a proactive decision making related to traffic incident management and law enforcement, especially at the outset of specific conditions with adverse effects on highway traffic safety, such as adverse weather conditions during peak commute ours.

Research Goals and Objectives

The main objective of the proposed research is to develop a modeling framework for segment-level crash risk assessment considering roadway geometry characteristics and dynamic parameters affecting the crash risk, including temporal characteristic (e.g., season, day of the week, time of day), traffic flow characteristics (e.g., vehicle volume, average speed or travel time), and weather conditions (e.g., precipitation and visibility). In developing the model framework, the historical crash data from the selected roadways in the State of New Jersey were analyzed to identify important patterns and statistical significance of various contributing factors. The data considered in the analysis was limited to information currently available to transportation agencies in real time, on the roadway segment level, and providing network-wide coverage for major roadways in New Jersey. This was done purposely, aiming to only include the data that could be used for dynamic short-term crash prediction. Based on the results of this analysis, different modeling techniques will be considered to select the one or a combination of techniques that would yield the best crash risk assessment results. Ultimately, the aim of this research is to utilize the findings in advancing the development of analytical models and tools to predict relative crash risk and their severity for a given roadway segment under the given traffic and weather conditions, or provide a ranking of roadway segments by relative crash risk under a given set of conditions. The crash risk ranking, or other safety performance measures, could then be used to select and prioritize crash and crash-related congestion mitigation strategies and actions by the highway operations agencies.

BACKGROUND

Literature Review

Conceptually, the crash risk and severity are influenced by a set of factors related to driver performance, roadway characteristics, vehicle characteristics, and environmental factors. The data related to these factors at the time of crash is collected after the crash occurrence by the responding law enforcement officers as part of crash investigation and reporting. However, most of the factors, especially related to driver performance and vehicle characteristics, are not known, or rather cannot be ascertained in real-time for a specific roadway segment. Advances in Intelligent Transportation Systems (ITS) and data collection technologies have vastly improved the ability of transportation agencies to collect and analyze traffic and road performance data in real time, such as segment-level travel time, speed, volume, occupancy, and road-weather data. Nevertheless, the challenges in this respect remain as the data collection is often times focused on specific roadway segments, limiting the coverage of the regnal transportation network.

At the same time, numerous studies have already been conducted with the goal of utilizing the data collected in real-time and advanced data analysis methods to assess the likelihood of crashes and their severity. Yu and Abdel-Aty (2014b) used four different models to classify and compare the non-severe crashes and severe crashes on two high-speed facilities: I-70 freeway in Colorado and State Road 408 (SR-408) in Orlando, Florida. Four datasets were utilized to study the severity of crashes on I-70: (1) crash data for I-70 provided by the Colorado Department of Transportation (CDOT), (2) roadway segment geometry data from the roadway characteristics inventory, (3) real-time weather data from six weather stations located along the study area, and (4) real-time traffic data collected by automatic vehicle identification (AVI) detectors. The real-time traffic data was aggregated into 6-min intervals and the mean, standard deviation, and coefficient of variation of the speeds for 6-12 minutes prior to each crash were calculated to represent the traffic conditions before the crashes happened. The visibility condition from the closest weather station prior to the time of crash was also assigned to each crash to investigate the impact of weather on the severity of crashes. Two binary indicator variables (snow season vs. dry season and longitudinal grade ≥ 4% vs. longitudinal grade < 4%), one real-time traffic variable (standard deviation of speed), and two joint variables (visibility * snow season and visibility * dry season) were used as inputs for the I-70 models. To analyze the severity of crashes on SR-408, crash data from the crash analysis reporting (CAR) system, and real-time AVI data from the Orange County Expressway Authority (OOCEA) were used. The same approach as in the I-70 model was also implemented to aggregate and assign the traffic data for each crash. Three binary indicator variables (passenger car vs. non-passenger car, daytime vs. nighttime, and whether the impact point is the driver side), one roadway geometry variable (shoulder width), and one real-time traffic variable (standard deviation of speed) were investigated in the crash severity models for SR-408. Four different models were used to analyze the crash injury severities for the two studied roadways: regular binary probit (BP) with maximum likelihood estimation, Bayesian BP, segment level random-effect hierarchical Bayesian BP, and crash-level random-effect Bayesian BP. First, the results of the BP model were compared to the Bayesian BP, showing that for both roadways the Bayesian BP model outperformed the regular BP model in terms of the number of significant

variables. Second, the Bayesian BP model was compared with the segment level random-effect hierarchical Bayesian BP model. The Bayesian models were compared based on the deviance information criterion (DIC): the lower value of DIC in random-effect Bayesian models indicated that they were superior as they better accounted for the unobserved heterogeneity in the data that was not captured in the Bayesian BP model. Finally, the comparison between the two hierarchical Bayesian BP models showed that the model performance can be improved by the crash level random effect model as it allowed for a more flexible error term.

In another study by Yu and Abdel-Aty (2014a), similar data sources were used to develop crash injury severity models for the I-70 freeway. First, Random Forest (RF) algorithm was used to rank the variables: the steep grade indicator, speed standard deviation, temperature, and snow season indicators were found to be the most important factors. Second, a Bayesian fixed-parameter binary logit model was developed to model the injury severity (severe vs. non-severe). The results of the model showed that the temperature was not statistically significant. To account for the potential non-linearity between the injury severity levels and independent explanatory variables, a Support Vector Machine (SVM) model with radial basic function (RBF) kernel was performed. The effect of the explanatory variables was also quantified through the sensitivity analyses. Next, a random parameter logit model with an unrestricted variance-covariance matrix was used to model the injury severities by considering the unobserved heterogeneities and correlation between the input variables. Finally, the three models were compared based on the area under the Receiver Operating Characteristic (ROC) curve values. The results indicated that SVM model and logit model with random parameters provided better results than the binary logit model with fixed parameters.

Xu, Tarko, Wang, and Liu (2013) developed a model to predict the crash likelihood at three different severity levels. The study area covered a 29-mile segment on the I-88 freeway in San Francisco. The model inputs included 22 traffic flow variables derived from the vehicle count, occupancy, and speed data for the upstream and downstream stations, obtained from the Highway Performance Measurement System (PeMS) and based on 30-second raw detector readings. The traffic data was aggregated into 5-minute intervals, and traffic data for the period 5-10 minutes prior to crash at the upstream and downstream detectors used to represent the traffic condition at the time of the crash. In addition, the data for nine roadway-geometry variables obtained from PeMS were also included in the dataset, such as width of the roadway, number of lanes, and geometric type of the roadway. The weather condition data (clear vs. adverse), was obtained from the National Climate Data Center (NCDC). For each crash case, 20 non-crash cases were randomly selected. Traffic data, geometric data, and weather data were assigned to all crash cases and non-crash cases for model development. A three-stage sequential binary logit model was used to assess the likelihood of crashes at each severity level. The 20-fold cross-validation was also performed to evaluate the model's performance. The findings of the study showed that the traffic flow characteristics contributing to crash likelihood were substantially different at each severity level.

Theofilatos (2017) investigated accident likelihood and severity by incorporating real-time traffic and weather data for urban arterials in Athens, Greece. To build the dataset, traffic data from the

nearest upstream loop detector and weather data from the closest weather station were matched to each crash. The traffic and weather data were aggregated into 1-hour intervals and used for the analysis. For every crash case, two non-crash cases were collected for the same location and same time, one week before and one week after the crash occurrence. Traffic and weather data were assigned to non-crash cases using a similar method as the crash cases. For the crash likelihood model, a random forest (RF) approach was used to select the significant variables. Five parameters were found to be significant and therefore, selected to be implemented in the final model: 1-hour coefficient of variation of flow upstream, 1-hour standard deviation of occupancy up-stream, 1-hour standard deviation of speed upstream, 1-hour coefficient of variation of speed upstream, and 1-hour coefficient of variation of occupancy upstream. Next, a correlation matrix was built to assess the correlation between the significant variables to avoid multicollinearity problem. Finally, a Bayesian logistic regression was used to model the likelihood of crashes. The model outputs showed that the standard deviation of occupancy and the coefficient of variation of flow, impact the likelihood of crashes. A similar approach was undertaken for the crash severity, where RF model was used to identify the following important variables: 1-hour average flow upstream, accident type, 1-hour coefficient of variation of flow upstream, 1-hor average speed upstream as well as 1-hour coefficient of variation of speed upstream. The correlation matrix was also generated to find the possible correlations between these variables. Two different approaches were utilized to model the crash severity in the next step. A finite mixture logit (latent class) model and a mixed effect logit model. The results of the study revealed that the finite mixture model showed a better fit and proved to be superior as the latent classes are optimally chosen by the model based on the Bayesian Information Criterion (BIC).

Yu and Abdel-Aty (2013) studied the real-time crash risk by analyzing a 15-mile mountainous freeway section of I-70 in Colorado. The datasets used in the study, included: (1) crash data provided by CDOT, and (2) real-time traffic data from the Remote Traffic Microwave Sensor (RTMS) radars. The RTMS radars collect data on speed, volume, and occupancy at 30-second intervals. This data were further aggregated into 5-minute intervals and assigned to each crash from the nearest downstream detector. Similar to the Xu et al. (2013), the data aggregated for the period 5-10 minutes prior to the time of crash time was selected to represent the traffic condition at the time of the crash. In addition, the upstream and downstream speed, volume, and occupancy were also used in the analysis. For each crash, the average and standard deviation of the three traffic flow parameters were calculated for three detectors (downstream, crash location, and upstream), which makes the total number of 18 traffic-related explanatory variables associated with each observation. Furthermore, for each crash case, four non-crash cases were identified and matched for the same location, day of the week, and time of day, two weeks before and two weeks after the crash occurrence. For the modeling part, firstly, a classification and regression tree (CART) was incorporated to estimate the significant variables to be used as inputs for the crash likelihood models. The selected variables included: downstream average speed, crash location average speed, crash location standard deviation of occupancy, and crash location standard deviation of volume. The correlation matrix was also calculated to find potential correlations between the identified variables. In the next step, the dataset was split into a training set (70%), and three testing sets with varying sample sizes (30%, 20%, and 10%). Three Bayesian logistic regression models were applied using the training set: (1) Bayesian fixed-parameter

logistic regression, (2) Bayesian random-parameter logistic regression accounting for the seasonal variation, and (3) Bayesian random-effect logistic regression considering the segment level heterogeneity. Comparing the DIC values for the three models demonstrated that the Bayesian fixed-parameter model showed better performance than the other two models. Next, two SVM models, one with linear kernel and one with RBF kernel, were employed and tested using different testing sets. The results were compared to the results produced by the Bayesian logistic regression, using the Area under the ROC curve (AUC). The findings of the study showed that the SVM with RBF kernel models was superior, and therefore, concluded that some non-linear relationships existed between the dependent variable and independent variables in the real-time crash risk model.

Wang, Shi, and Abdel-Aty (2015) conducted a study to predict crashes on expressway ramps. Three expressways in Central Florida were chosen as the study area: SR-408 (14.2 mi), SR-417 (26.9 mi), and SR-528 (7.6 mi). To reduce the noise, traffic data was aggregated into 5-minute intervals, and the period 5-10 minutes prior to the time of crash was selected to represent the traffic condition. Compared with the traffic data 0-5 minutes before the crash, it was discovered that the period 5-10 minutes prior to the time of crash provides better model performance and is also sufficient enough to disseminate warning information to the drivers. The non-crash cases were generated through a random process in which 0.05% of the 11,270,808 5-minute intervals (12 intervals * 24 hours * 141 ramps) were selected in SAS. The data used for the study included: (1) crash data from the Florida DOT statewide crash database, (2) traffic flow data provided by the Central Florida Expressway Authority, (3) roadway geometry data derived from the roadway Geographic Information System (GIS), and (4) weather data from the National Climate Data Center. The final dataset was further divided into two parts based on the crash type (single vehicle vs. multi-vehicle). The dataset for each crash type was also split into training and validation datasets with a ratio of 70:30. The Pearson correlation test was performed before the model development to detect potential correlations between the explanatory variables. The Bayesian logistic regression was used to establish the prediction models for a single vehicle (SV) and multi-vehicle (MV) crashes. Five variables were found to be significant in the SV crash prediction model: logarithm of the vehicle count in 5-min intervals, speed, ramp configuration, road surface condition, and visibility. The AUC for the training and validation were also found to be 0.9346 and 0.9710, respectively. In addition, the overall accuracy was 0.89 for the training set and 0.904 for the validation set. All the significant variables in the SV model, except the speed, were found to be significant in the MV model as well. The AUCs for the training and validation were 0.7644 and 0.76, respectively, and the overall accuracy was obtained as 0.643 for the training set and 0.764 for the validation set.

Theofilatos, Chen, and Antoniou (2019) compared the performance of machine learning (ML) and deep learning (DL) methods in predicting crash occurrence. To achieve this, the Attica Tollway, an urban motorway in Greece was selected as the study location. For the analysis purposes, real-time traffic data and weather data were obtained and matched to the crash and non-crash cases. A 1:2 ratio of crash cases to non-crash cases was selected for this study. In addition, the raw data were aggregated to obtain the average, standard deviation, and coefficient of variations of traffic-related parameters. To develop the models, data was first split into a training set (75%), and a

validation set (25%), and various ML methods were employed to predict the crash likelihood using the training set. The ML models considered in the study included: k-nearest neighbor, Naïve Bayes, decision tree (DT), RF, SVM, and shallow neural network. These models were generated and compared based on their performance metrics (accuracy, sensitivity, specificity, and AUC). An RF model was first applied to identify important variables. Afterward, a binary logistic model was generated with the selected variables to check and confirm the degree of significance for each of them. The result of the binary logistic model indicated that the standard deviation of speed 0-15 minutes before the crash time, and the total amount of rainfall were the only significant variables, and they were used as inputs for the ML and DL models. The results of the study showed that the DL model outperformed the ML techniques as it provided a relatively balanced performance among all metrics.

Analysis Methods Applied in this Study

Based on the lit review and initial analysis using the input dataset, a number of modeling methods was considered, including regression models and machine learning models. Considering the scope and time frame of the study, the following methods were selected for evaluation in the analysis of crash likelihood and severity: Bayesian Logistics Regression, Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), K-Nearest Neighbor (KNN), and Gaussian Naïve Bayes (GNB). Each method is briefly explained in the following subsections.

Bayesian Logistic Regression

This study applied the Bayesian logistic regression model to predict the likelihood of crash occurrence. Unlike the classical logistic regression that treats the parameters of the independent variables as fixed, the coefficients in Bayesian logistic regression are assumed to follow a distribution, such as Gaussian, Bernoulli, or multinomial. In the study of crashes in this study, the binary outcomes are $y_i = 1$ and $y_i = 0$: in the crash likelihood model, “1” represents a crash and “0” represents a non-crash case; in the crash severity model, “1” represents an injury/fatal crash, and “0” represents a property-damage-only (PDO) crash. The probabilities associated with the binary events are p_i and $1 - p_i$, respectively. Thus, applying the Bayes theorem, the Bayesian logistic regression is built as follows:

$$y_i \sim Bernoulli(p_i) \quad (1)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha_0 + \sum_{m=1}^{M} \alpha_m x_{mi} \quad (2)$$

where y_i is assumed to follow a Bernoulli distribution, α_0 presents the intercept of the model, α_m the parameter of m^{th} explanatory variable, and x_{mi} the value of m^{th} explanatory variable for i^{th} observation, e.g., volume, average speed, hourly precipitation, etc. The parameters including α_0 and α_m are assumed to follow the normal distributions. The likelihood of an event is then calculated as:

$$likelihood_i = \pi(x_i)^{y_i}\left(1 - \pi(x_i)\right)^{(1-y_i)} \quad (3)$$

where $\pi(x_i)$ denotes the probability of an event for the i^{th} observation which has a vector of independent variables x_i.

Decision Tree (DT)

The description of the principles of DT is explained in various road safety studies (Kwon et al., 2015; Theofilatos et al., 2019). A DT algorithm employs a tree framework that builds the data set from the root node and splits it to the leaf nodes, which introduces a class value in the dataset. A DT algorithm aims to recursively form the child nodes that consist of a high proportion of data points from a single class. Therefore, the recurrently constructed DT maximizes the "purity" in the child nodes that represent one class. To measure the purity of data split, the Gini impurity factor is introduced as a measure of diversity of a predictor. The Gini impurity is formulated as follows:

$$\text{Gini impurity} = 1 - \sum_1^n p_j^2 \quad (4)$$

where j represents the class of targets including crash or non-crash, n defines the number of targets (which is equal to two in this study), and p represents a probability of picking a datapoint with crash or non-crash cases.

Random Forest (RF)

Random Forest (RF) is an ensemble learning method that can be defined as the combination of Breiman's bagging idea, (Breiman, Friedman, Stone, & Olshen, 1984) and random feature selection. The basic idea behind RF is to build a collection of decision trees by bootstrapping the sample and use a random subset of input factors for splitting at each node. Thus, an RF consists of multiple decision trees where each of them presents a model (e.g., classification) with a subset of features. RF outputs are generated as the averages of all decision trees in the forest, which is referred to as voting. The RF models often outperform the traditional classification and regression trees (CART) in terms of accuracy and capability of providing unbiased error. The other advantage of RF over CART is that it obviates the need for a separate cross-validation dataset. RF is a common method used in different crash likelihood studies (Theofilatos, 2017; Theofilatos et al., 2019).

During the training procedure, about one-third of the training data is held out and is not used in model development. These cases are referred to as the out-of-bag (OOB) data (Breiman, 2000). The main objective of RF is to tune the primary model by selecting the optimal number for hyperparameters to minimize the OOB error. For example. Reducing the number of randomly sampled variables available for splitting at each tree node (*mtry)*, reduces both the correlation and the strength. Therefore, an important step in model development is to find the optimal number of *mtry*s. OOB error is a function of the correlation between each pair of trees in the forest and the strength of each individual tree. There is a positive relationship between the inter-tree

correlation and OOB error, while the relationship between the strength of the individual trees and OOB error is negative.

The OOB data can further be used to quantify the variable importance. The importance of a variable can be explained by examining the change in the prediction error when that variable is permutated or excluded in the OOB data, while all the other variables remained unchanged. After obtaining the new OOB error, the variable importance can be determined by calculating Mean Decrease Accuracy (MDA) as an average difference in the new error and the initial error over all trees in the random forest (Nicodemus, 2011). Higher values of MDA indicate greater relative importance of a variable. Another variable importance measure is Mean Decrease Gini, which is defined as the average across the forest of the decrease in Gini impurity indicator for a factor (Nicodemus, 2011). While both methods have been used in the literature, MDA was chosen for variable ranking in this study.

Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is a powerful ML technique, proposed by Friedman (2001). Like RF, GBM is also an ensemble technique, using decision trees as the base model. However, unlike RF, which creates large trees, GBM grows a sequence of small trees such that each tree tries to capture those parts of the training set which were missed in the preceding tree (Hastie, Tibshirani, & Friedman, 2009). To this end, GBM identifies the missing parts by using the gradient of some differentiable loss functions, using a random subsample of the training set with different sizes. In our problem, the multinomial deviance is used as the loss function.

K-Nearest Neighbor (KNN)

KNN is a machine learning approach in which the classification of observations of interest is based on the labels of its k-th nearest neighbors, identified based on some measure of multi-dimensional distance. As all the *K* neighboring observations do not normally belong to the same class, the class label of the majority of them is selected as the class label of the unclassified observation (Bishop, 2006). Two decisions need to be made with regards to KNN: the value of *K* and the distance function. Normally, the best value of *K* is achieved through an iterative process in which, different values are examined and the one that results in the best model performance in terms of the selected performance metric is chosen. Small values of *K* may create weak models unable to classify features in the model, while large values of *K* can lead to overfitting. I addition, as a rule of thumb, where there are only two classes, which is the case in our study, *K* should be odd in integer to avoid ties (Cigdem & Ozden, 2018). With respect to the distance function, Euclidean distance, weighted Euclidean distance, and cosine method are the most commonly used in KNN models. In this study, the Euclidean distance was used as the distance function. Euclidean distance can be formulated as:

$$dist(x_i, x_j) = \sqrt{(\sum_{k=1}^{p}(x_{ik} - x_{jk})^2)} \quad (5)$$

where $dist(x_i, x_j)$ denotes the distance between observation *i* and *j*, and x_{ik} and x_{jk} are the value of the *K*th factor for *i* and *j*, respectively.

Gaussian Naïve Bayes (GNB)

The Naïve Bayes (NB) algorithm is one of the probabilistic classification techniques based on Bayes' theorem, which assumes that the features are strongly independent of each other. This method has been used in various road safety studies (Shanthi & Ramani, 2011; Theofilatos et al., 2019). Using the Bayes theorem, the posterior probability of a class target y occurs given the attribute vector X, $x_i \in X$, $1 \leq i \leq n$, which is calculated as follows:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(X)} \quad (6)$$

where $P(y|X)$ denotes the posterior probability that class y occurs given feature x, $P(X|y)$ and denotes the likelihood probability of x given class y. The $P(y)$ and $P(X)$ represent the prior probabilities of class y and X respectively, which occur independently. In this study, the Gaussian Naïve Bayes (GNB) method is applied, which uses the Gaussian likelihood function:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\pi\sigma_y^2}\right) \quad (7)$$

where the parameters σ_y, and μ_y are estimated using maximum likelihood.

METHODOLOGICAL APPROACH

The methodological approach applied in the study consisted from the flowing steps:

1) Identify data sources and collect data to be used in model development.
2) Identify the roadways to be included as the study location.
3) Prepare data for the analysis, i.e., input to the models identified in the previous chapter.
4) Define the model performance criteria.
5) Apply the models (including model tuning) and analyze the results.

The steps 1-4 are explained in the following subsections. The model application and results are summarized in the next chapter titled *Results*.

Data Sources

In identifying the data sources and datasets to be collected and used in the analysis, the following types of data were of interest:

- Historical crash records – needed to obtain a record of crashes and their severity as the outcomes to be predicted by the crash likelihood and crash severity models.
- Roadway characteristics dataset – providing roadway geometry data.
- Traffic condition datasets – providing real-time data on speeds, travel times, and vehicle volume at a roadway segment level.
- Weather datasets – providing real-time weather information, such as temperature, precipitation, visibility, wind, etc.

The criteria for identifying the data sources and datasets for the analysis included: availability of data in real-time, availability of data at a roadway segment level, and availability of data for all sections of the major roadways in the State of New Jersey. Following the detailed search and review of the datasets available to New Jersey Department of Transportation, the following were selected as the data source for the analysis in this study:

1) NJDOT Crash Records Database – the database contains records of all crashes reported by the Police Departments in the State of New Jersey using the NJTR-1 Accident Report Form. The data provides detailed information about the crash characteristics, roadway condition, environmental (ambient) conditions, vehicle characteristics, as well as the condition and characteristics of all participants in a crash. The crash data for the period January 2017 through December 2018 were acquired from the NJDOT website and used in the analysis.

2) NJDOT Congestion Management System (NJCMS) – this dataset provides estimated, synthesized hourly volume and congestion levels (expressed in terms of average speed and volume-to-capacity ratio) at a roadway segment level for all highways in NJDOT jurisdiction. This dataset also provides the basic roadway geometry data, such as number

of lanes, median types, and shoulder, which were also acquired and used in developing the analysis dataset for this study. The datasets with 2012 and 2016 vehicle volume data were used as the baseline for calculating 2017 and 2018 hourly volumes for all roadway segments in this study. Moreover, the seasonal traffic factors were applied to calculate vehicle volumes specific to each month of the year.

3) Probe-vehicle speeds at roadway segment level – this dataset provides the actual prevailing vehicle speeds and travel times aggregated from the probe vehicles and recorded in 1-minute increments. The data was obtained from the RITIS system for the sample of roadway segments and the time period analyzed in the study. In spatial terms the speeds and travel times are aggregated and reported for traffic management channel (TMC) links. The limits of TMC links do not coincide with the roadway segments defined in the NJCMS dataset, and therefore it was necessary to match and conflate the speed records from the RITIS dataset to the roadway segments defined in the NJCMS dataset for the roadways included in the analysis.

4) Historical weather data from the National Centers for Environmental Information (NCEI) dataset – the historical weather observation data was obtained from the dataset sourced from weather stations (AWOS and ASOS) managed by the National Weather Service (NWS) and the Federal Aviation Administration (FAA). The weather observation data was matched to the locations of reported crashes and time intervals prior to the reported crash time (e.g. 15-30-minute interval). This data provides additional insight into ambient conditions at the time of crash and non-crash cases included in the model dataset. The Local Climatological Data (LCD) was identified as the most complete and reliable dataset that provides local weather information from permanent weather stations in 15-minute increments. The data record for each location and time stamp contains the ambient temperature, air pressure, visibility, hourly precipitation, hourly visibility, and average wind speed. The LCD data was obtained from the National Centers for Environmental Information (NCEI) of the National Oceanic and Atmospheric Administration (NOAA) (National Centers for Environmental Information, 2019). Hourly visibility and hourly precipitation are considered as one of the prominent variables affecting the crash likelihood and severity.

In the next step the data available from the above listed data sources was reviewed and key explanatory variables were identified for inclusion in the crash likelihood and crash severity models.

Explanatory Variables

The explanatory variables that were identified as the most critical and informative for crash likelihood and crash severity analysis are listed in Table 1. The selection of explanatory variables was largely informed by the previous studies identified in the literature review. The data for each variable was collected for each crash event analyzed in the study and obtained from the new Jersey crash records database.

In preparation for the analysis the categorical variables LANES and MEDIAN were converted to binary variables, which are shown in Table 2 with the values corresponding to the number of lanes and type of median, respectively.

Table 1. Definition of Explanatory Variables Used in the Study

Variable	Type	Description
LANES	Categorical	Number of lanes (the values are: 2, 3, 4, or 5)
MEDIAN	Categorical	Median type (can be curbed, positive, on unprotected)
CAPLINK	Continuous	Capacity of the highway section [vehicles/hour]
VOL	Continuous	Estimated hourly vehicle volume at the highway section during a given hour of the day and month [vehicles/hour]
VC_RATIO	Continuous	Volume-to-capacity ratio at the highway section during a given hour of the day and month [unitless]
HourlyPrecipitation	Continuous	Hourly precipitation at the highway section during the hour of the crash or non-crash event obtained from the weather records for the closest weather station [inches/hour]
HourlyVisibility	Continuous	Hourly visibility at the highway section during the hour of the crash or non-crash event obtained from the weather records for the closest weather station [miles]
speed_avg	Continuous	Average speed on the highway section [miles/hour]. It is calculated for each crash and non-crash event as an average of 1-minute prevailing speeds for the pertinent highway section over a 15-minute period preceding the crash or non-crash event.
speed_sd	Continuous	Standard deviation of speed on the highway section [miles/hour]. It is calculated as a standard deviation of 1-minute prevailing speeds for the pertinent highway section over a 15-minute period preceding the crash or non-crash event.
speed_cv	Continuous	Coefficient of variation of speed [unitless]. Calculated as the ratio of average speed (speed_avg) and standard deviation (speed_sd) for the 15-minute period preceding the crash or non-crash event.
speed_ex	Continuous	Speed deviation from the speed limit [miles/hour]. Calculated as the difference between the average speed (speed_avg) and the speed limit (obtained for each roadway segment from the NJCMS dataset) for each crash and non-crash event at the given highway section.

Table 2. Conversion of Categorical to Binary Variables (LANES and MEDIAN)

Binary Variable	LANES			
	2	3	4	5
2 LANES	1	0	0	0
3 LANES	0	1	0	0
4 LANES	0	0	1	0
5 LANES	0	0	0	1

	MEDIAN		
Binary Variable	CURBED	POSITIVE	UNPROTECTED
MEDIAN_CURBED	1	0	0
MEDIAN_POSITIVE	0	1	0
MEDIAN_UNPROTECTED	0	0	1

Study Area

The study location was focused on two interstate highways in New Jersey: I-80 and I-287. The interstate I-80 has a west-to-east alignment and the New Jersey section is 68.5 miles long. The interstate I-287 has a south-to-north alignment and the New Jersey section is 67.5 miles long. Both roadways are located in the northern part of the State and had the highest number of crashes among the interstate highways in the State. The location of I-80 and I-287 on the map of New Jersey is shown in Figure 1.

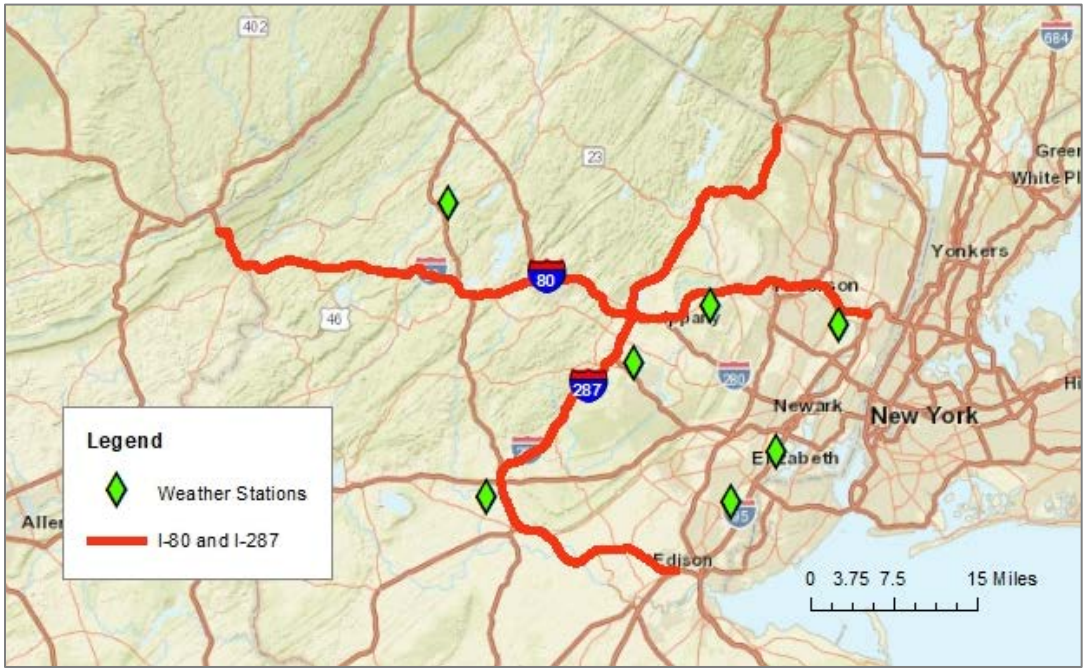

Figure 1. The study area with the location of I-80, I-287, and weather stations

The weather data was obtained from the LCD database from seven weather stations located in the proximity of I-80 and I-287. For each roadway segment the closest LCD station was identified based on the Euclidian distance. The locations of LCD weather stations that provided data for the study area are shown in Figure 1. All stations are located at the regional airports.

Data Preparation

Summary of Data Inputs

The summary of the dataset used in the study is provided in Table 3 - Table 6. The dataset included the total of 10,155 crashes that recorded along interstate I-80 and interstate I-287 during the period January 2017 – December 2018. Each crash was matched to a corresponding NJCMS record based on the unique road identifier (standard road identifier, or SRI) and milepost. The matching NJCMS record provided the segment-level roadway data, such as speed limit, hourly vehicle volume, v/c ratio, number of lanes and type of median.

The traffic speed data at the crash location prior to the time of crash was obtained from the RITIS dataset. The RITIS data was also matched to the NJCMS segment based on route name and milepost and added to the record of each crash. As previously indicated, the average speed for each segment in RITIS dataset is reported at a 1-minute interval. Nevertheless, to reduce the noise and the impact of human error in reporting the exact time of the crash, the speed data was extracted for a period of 15 minutes prior to the crash occurrence, and then aggregated to calculate the average speed, the standard deviation of speed, the coefficient of variation of speed, and the deviation from the speed limit over the same 15-minute period. For each crash these speed indicators were used as model inputs.

Lastly, the weather data was extracted from the LCD data for the date and time of crash, and the weather station closest to the crash location, i.e., closest to the NJCMS segment associated with the crash record. The weather data extracted from the LCD dataset included hourly precipitation and hourly visibility observed during the hour of the crash.

Table 3. Summary of the Roadway Segment Characteristics (Including Crash Statistics)

Characteristic	I-287	I-80	Total
Number of crashes (total)	1,267	8,888	10,155
Number of injury/fatal crashes	236	1,903	2,139
Number of PDO crashes	1,031	6,985	8,016
Roadway length (in miles)	67.5	68.5	136
Number of roadway segments (both ways)	116	164	280
Minimum length of a roadway segment (in miles)	0.020	0.100	0.020
Maximum length of a roadway segment (in miles)	5.140	4.020	5.140
Average length of a roadway segment (in miles)	1.218	0.936	1.053

Table 4. Summary of Basic Statistics for the Explanatory Variables

Variable	Description	Min	Max	Mean	Median
CAPLINK	Road capacity	3,268	8,857	5951	5,314
VOL	Vehicle volume	125	8928	3621	3482
VC_RATIO	v/c ratio	0.024	1.614	0.608	0.585
HourlyPrecipitation	Hourly precipitation	0.0	1.54	0.003	0.0
HourlyVisibility	Hourly visibility	0.0	74.0	8.917	10.0
speed_avg	Average speed	2	80.875	60.958	64.562
speed_sd	St. deviation of speed	0	32.482	2.588	2.048
speed_cv	Coef. of variation of speed	0	1.122	0.049	0.033
speed_ex	Average deviation form speed limit	-63.0	28.0	0.391	3.0

Table 5. Percentage of Roadway Segments by Number of LANES

Number of LANES	Percent of Road Segments
2	7.61
3	43.96
4	46.51
5	1.91

Table 6. Percentage of Roadway Segments by Type of MEDIAN

MEDIAN Type	Percent of Road Segments
Curbed	3.74
Positive	90.54
Unprotected	5.72

Generating Non-crash Cases for the Crash Likelihood Modeling

In order to evaluate crash likelihood, this study employed a matched case–control methodology, which involved introduction of non-crash cases to match the crash cases in terms of crash characteristics such as location and time. To that end, for every crash case four non-crash cases were generated for the same location, day of the week and time, including one each in the week before, two weeks before, a week after, and two weeks after the crash occurrence. The 1:4 ratio of crash cases to non-crash cases was recommended by Ahmed and Abdel-Aty (2011) who found this value to provide slightly better results when compared to other crash to non-crash case ratios. In addition, according to the finding of another study by S. Kuhn, Egert, Neumann, and Steinbeck (2008), negligible improvement can be achieved by adding non-crash cases beyond 1:3 ratio. It

should be noted that the matched case–control methodology employed in this study only accounted for the location (roadway) and time as the crash factors; the other factors, such as vehicle, driver, and environmental characteristics were not considered in case-control matching.

After identifying the non-crash cases, the same procedure that was applied to crashes was used to match the traffic flow, speed, and weather data to each non-crash case. After completing this step, the study dataset for the crash likelihood model had additional 40,620 records representing non-crash cases (in addition to the 10,155 crash records).

It should be noted that the crash severity dataset remained unchanged – it only contained the 10,155 records pertaining to crashes and their severity.

Determination of Significant Variables in the Crash Likelihood Model

In this study, Random Forest (RF) model was used to determine relative importance of variables to be used in the crash likelihood and crash severity models. This allows to only include the significant variables in models such as KNN, which can easily produce misleading results in high-dimensional space. For both datasets, the Mean Decrease in Accuracy (MDA) was used as the criterion in determining the relative variable importance. The mean decrease in accuracy for a variable is calculated based on the out of bag (OOB) error. The importance of a variable can be explained by examining the change in the prediction error when that variable is permutated or excluded in the OOB data, while all the other variables remained unchanged. After obtaining the new OOB error, the variable importance can be determined by calculating MDA as an average difference in the new error and the initial error over all trees in the random forest (Nicodemus, 2011). Higher values of MDA indicate greater relative importance of a variable.

Before identifying the significant variables, it is also important to check for correlation between the decision variables in the analysis dataset. To that end, the correlation matrix was created using Pearson correlation coefficient to identify the correlated variables, as shown in Figure 2.

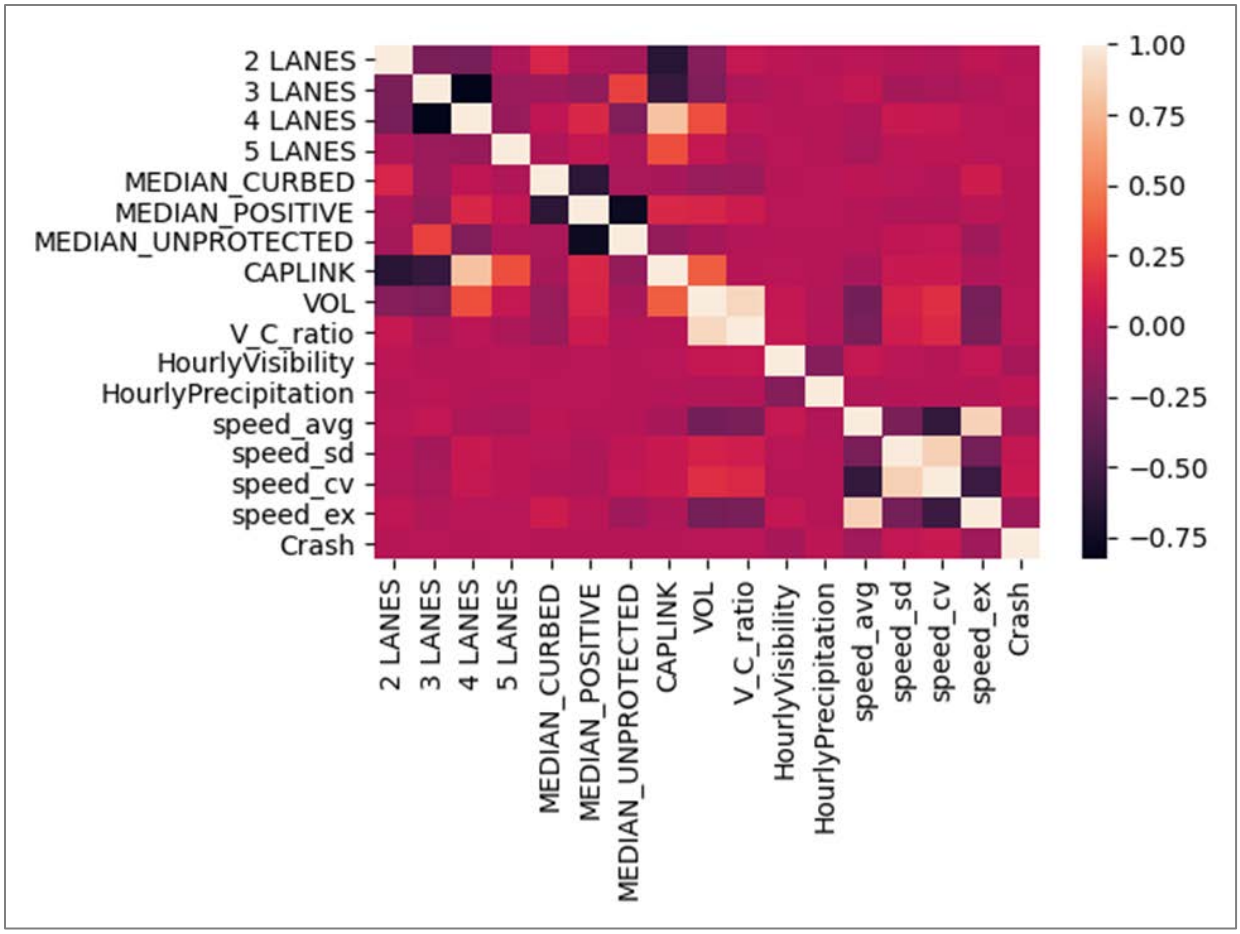

Figure 2. Correlation matrix for the crash likelihood analysis dataset

Based on the correlation matrix, it was decided to exclude from further consideration all of the LANES variables as they are correlated with highway capacity (CAPLINK), as well as MEDIAN-POSITIVE since it was correlated with the other two MEDIAN variables. When it comes to variables related to speed, based on the correlation matrix it was decided to exclude average speed (speed_avg) as it was highly correlated to speed_cv, speed_sd and speed_ex; it was also decided to exclude speed coefficient of variance (speed_cv) for the same reason. The standard deviation of speed (speed_sd) and deviation of speed from speed limit (speed_ex) are kept for evaluation of variable significance in the Random Forest (RF) model. It can also be observed that V_C_ratio and VOL are highly correlated, so they should not be used in the models together.

An RF model for the crash likelihood analysis dataset was then used to determine the relative importance of the variables. The RF model had mtry = 5 (number of factors randomly sampled at each split), number of trees = 500, split rule = Extra trees, and node size = 1 (minimum number of observations in each terminal node). The ranking of the relative variable importance in the crash likelihood model based on the RF model is illustrated in Figure 3. The ranking is provided for two alternate cases: (a) using the hourly v/c ratio (v_c_ratio) as the decision variable, and (b) using the hourly vehicle volume (VOL) as the decision variable. The vertical red lines denote a

cordon between the significant variables that should be considered (on the right-hand side) and variables that should be excluded as insignificant (on the left-hand side of the cordon line). The lines were placed where the gap between variables was relatively large in terms of MDA.

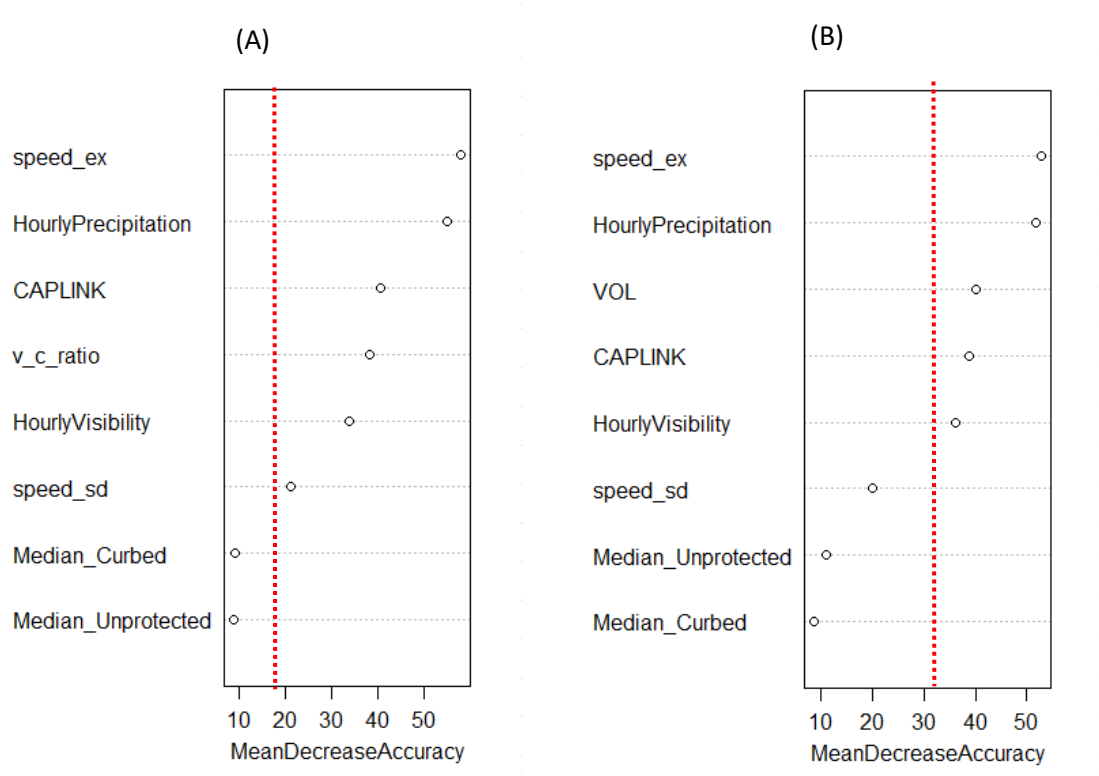

Figure 3. RF variable importance plot for the crash likelihood model: (A) with v_c_ratio, (B) with VOL as the decision variable

As it can be observed, in both cases the variables Median_Curbed and Median_Unprotected were not significant. Furthermore, the plot for the case with VOL as the decision variable suggested that standard deviation of speed (speed_sd) should be omitted due to low relative importance. It was decided to keep the combination of variables in case (A) (i.e., include V_C_ratio and speed_sd and omit VOL from further consideration) as the V_C_ratio captured the level of congestion, while VOL only provided the absolute vehicle volume on a highway link.

Thus, the final list of decision variables to be used in modeling the crash likelihood included:

- Speed_ex
- HourlyPrecipitation
- CAPLINK
- V_C_ratio
- HourlyVisibility
- Speed_sd

Determination of Significant Variables in the Crash Severity Model

Similar approach to the one used for the crash likelihood model was used to select the significant variables in the crash severity model. First, the correlation matrix was used to identify correlated variables (see Figure 4).

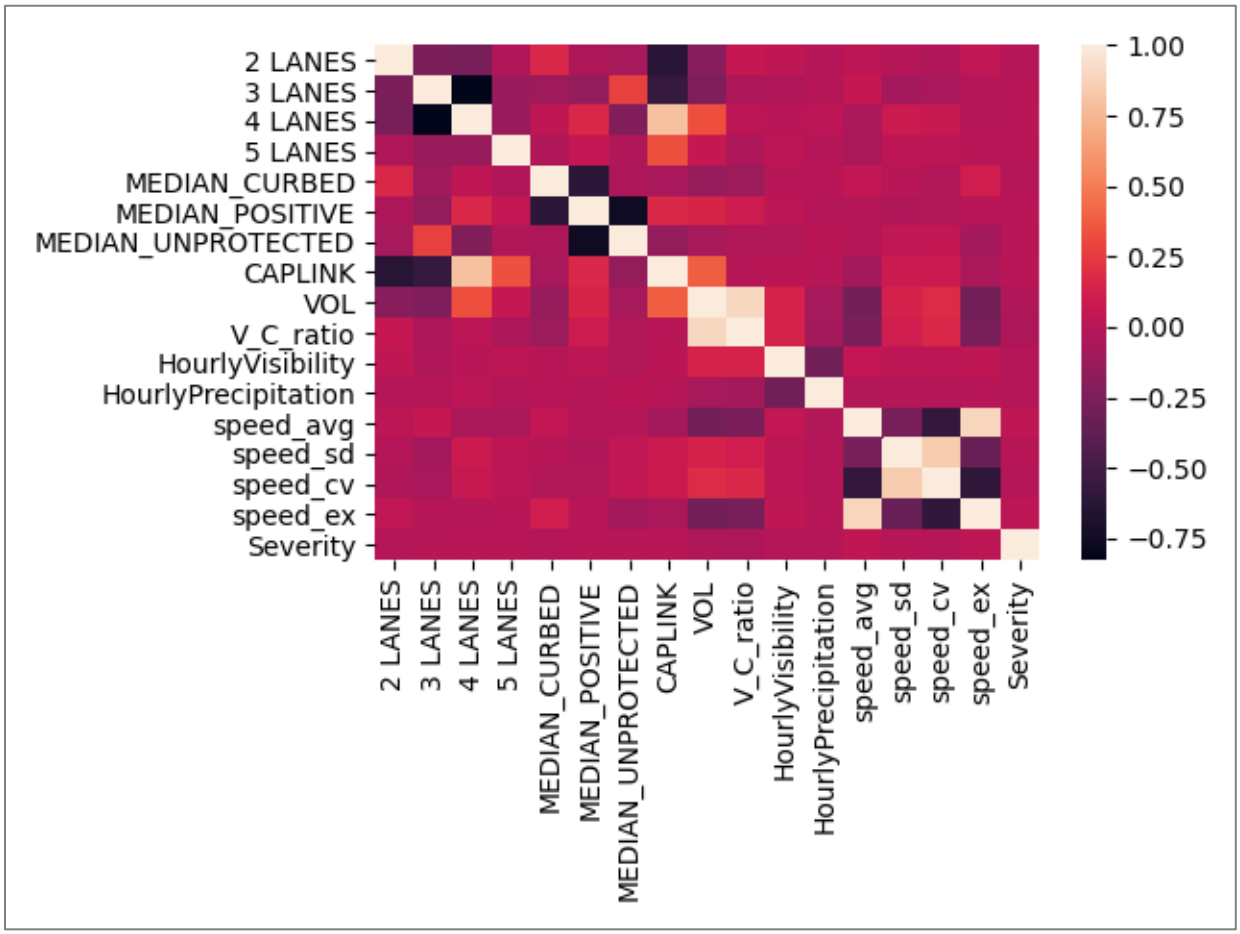

Figure 4. Correlation matrix for the crash severity analysis dataset

The correlation was very similar to that observed in the crash likelihood model. All of the LANES variables were excluded from further consideration as they were correlated with highway capacity – CAPLINK. The MEDIAN-POSITIVE was also omitted as it was correlated with the other two MEDIAN variables. It can be observed that VC_ratio and VOL are highly correlated, and thus should not be used in the models together. There is also high correlation between the average speed (speed_avg) and all other speed-related variables (including speed_cv, speed_sd, and speed_ex). However, correlation between the deviation of speed from speed limit (speed_ex) and standard deviation of speed is not significant, so these two variables can be considered in a model together.

The RF model was used to determine the relative importance of the variables for the crash severity analysis dataset with mtry = 2 (number of factors randomly sampled at each split), number of trees = 500, split rule = Extra trees, and node size = 1 (minimum number of observations in each terminal node). Considering the MDA criteria in the RF model, vehicle volume (VOL) was selected over v/c ratio (V_C_ratio) as the decision variable to enter the model. The ranking of the relative variable importance in the crash severity dataset is illustrated in Figure 5. The red line denotes the separation between the important and non-important variables.

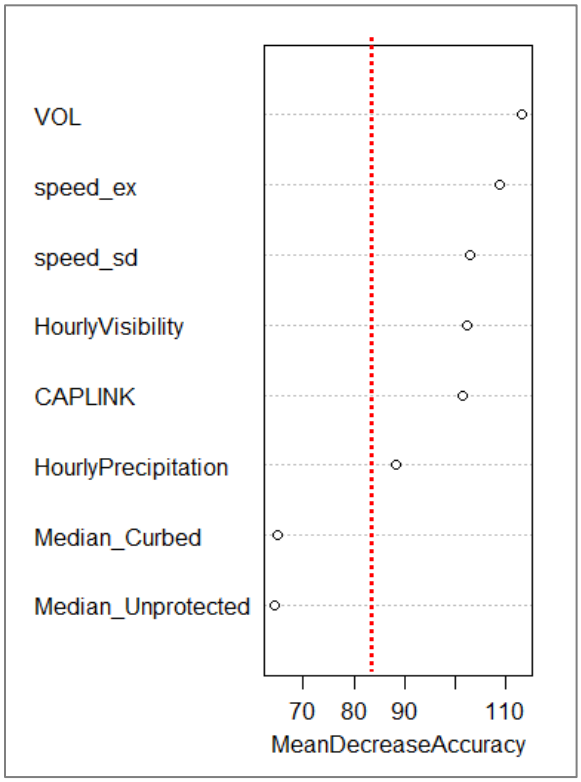

Figure 5. RF variable importance plot for the crash severity model

Thus, the final list of decision variables to be used in modeling the crash severity included:

- VOL
- Speed_ex
- Speed_sd
- HourlyVisibility
- CAPLINK
- HourlyPrecipitation

Dealing with the Data Imbalance Problem

To overcome the problem of a low frequency of fatal crashes, the fatality class was initially combined with the instances in the injury class. However, even after undertaking this action, 79%

of the cases were still non-injury crashes (8,016 PDO crashes out of the total of 10,155 crashes in the dataset) and only 21% of crashes (total of 2,139) with an injury or a fatal outcome. In the case of training the model with a skewed distribution of classes, the traditional accuracy maximizer techniques are not adequate and normally tend to perform better in favor of the prevalent class. Therefore, it is advantageous to transform the dataset so as to achieve a more balanced training dataset.

Random oversampling examples (ROSE) is a random bootstrapped-based technique, introduced by Menardi and Torelli (2014), which can alleviate the data imbalance issue in the binary classification problems. ROSE combines random oversampling and random undersampling by generating new artificial instances from the original classes based on a smoothed bootstrapped approach (Tibshirani & Efron, 1993).

Consider a training set of size *n*, consisting of a binary response variable y, with class labels Y_j and a set of input data for each class, $x_{ij}, i = 1, \dots, n_j$, where $n_j < n$ is the number of cases in class *j*. For each *x* belonging to the class Y_j, ROSE generates samples from a multivariate kernel density estimate of $f(x \mid y = Y_j)$ as follows:

$$\widehat{f}(x \mid y = Y_j) = \sum_{i=1}^{n_j} p_i \Pr(x \mid x_{ij}) = \sum_{i=1}^{n_j} \frac{1}{n_j} K_{H_j}(x - x_{ij}) \quad (8)$$

where K_{H_j} denotes an estimated kernel function and its smoothing matrix H_j is:

$$H_j = diag(h_1^{(j)}, \dots, h_d^{(j)}) \quad (9)$$

where d is the number of explanatory variables and

$$h_q^{(j)} = \left(\frac{4}{(d+2)n}\right)^{1/(d+4)} \hat{\sigma}_q^{(j)}, q = 1, \dots, d \quad (10)$$

where $\hat{\sigma}_q^{(j)}$ is the estimated standard deviation of the qth variable. According to Bowman and Azzalini (1997), the smoothing matrix minimizes the Asymptotic Mean Integrated Squared Error under the assumption that the true conditional densities underlying the data follow a Normal distribution.

The practical implementation of ROSE encompasses the following steps:

1) select $y^* = Y_j$ with probability π_j;
2) select *x* such that $y_k = y^*, k = 1, \dots, n$ with probability $\frac{1}{n_j}$;
3) sample x^* from the estimated kernel function.

Repeating steps 1 to 3 yields a newly generated training set of size *m*, with the probability of each class to be π_j.

Implementing the newly created dataset based on the ROSE approach is expected to provide better results than using the original imbalanced dataset. In addition, the findings of a study by Menardi and Torelli (2014) showed that ROSE outperformed other well-known oversampling methods, such as synthetic minority oversampling technique (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), by providing higher values of the area under ROC curve (AUC) in the logistic regression and classification tree models. In this study, the ROSE technique was applied to training sets for both the crash likelihood and the crash severity models to generate synthetic training sets.

Final Preparation of the Training and the Testing Datasets

As noted, the ROSE transformation is applied to the training datasets only. For that purpose, it was first necessary to split both the crash likelihood and the crash severity datasets into two subsets each: (a) training dataset, containing 80% of features (data records), and (b) testing dataset, containing 20% of features. A stratified sampling technique was used for splitting the datasets to ensure that there is the same proportion of output class labels in both the training set and testing set, as in the original data. Then, the ROSE transformation was applied to each training dataset.

Following the ROSE methodology, different probability values for the minority classes in each dataset were evaluated (e.g., 0.2, 0.3, 0.4, 0.5, 0.6). The evaluation showed that the probability of 0.5 yielded best results in terms of sensitivity and F1-scores in both the crash severity and crash likelihood model training datasets. A visual representation of the dataset before and after applying ROSE is shown in Figure 6, displaying the example of the data reflecting the deviation of speed vs. volume as the independent variables, and the crash severity (PDO vs. injury/fatality) as the dependent variable.

The number of crash records (features) in the training datasets for each class before and after the ROSE transformation, as well as the size of each class in the testing datasets are summarized in Table 7.

Table 7. Size of Input Datasets for the Crash Likelihood and Crash Severity Models

Models / Corresponding Classes	Training Dataset		Testing Dataset
	Before ROSE	After ROSE	
Crash Likelihood Dataset	40106	60159	10026
Crash Cases	8054	30052	2101
Non-crash Cases	32052	30107	7925
Crash Severity Dataset	8125	12187	2030
Fatal/Injury Crashes	1720	5917	419
PDO Crashes	6405	6270	1611

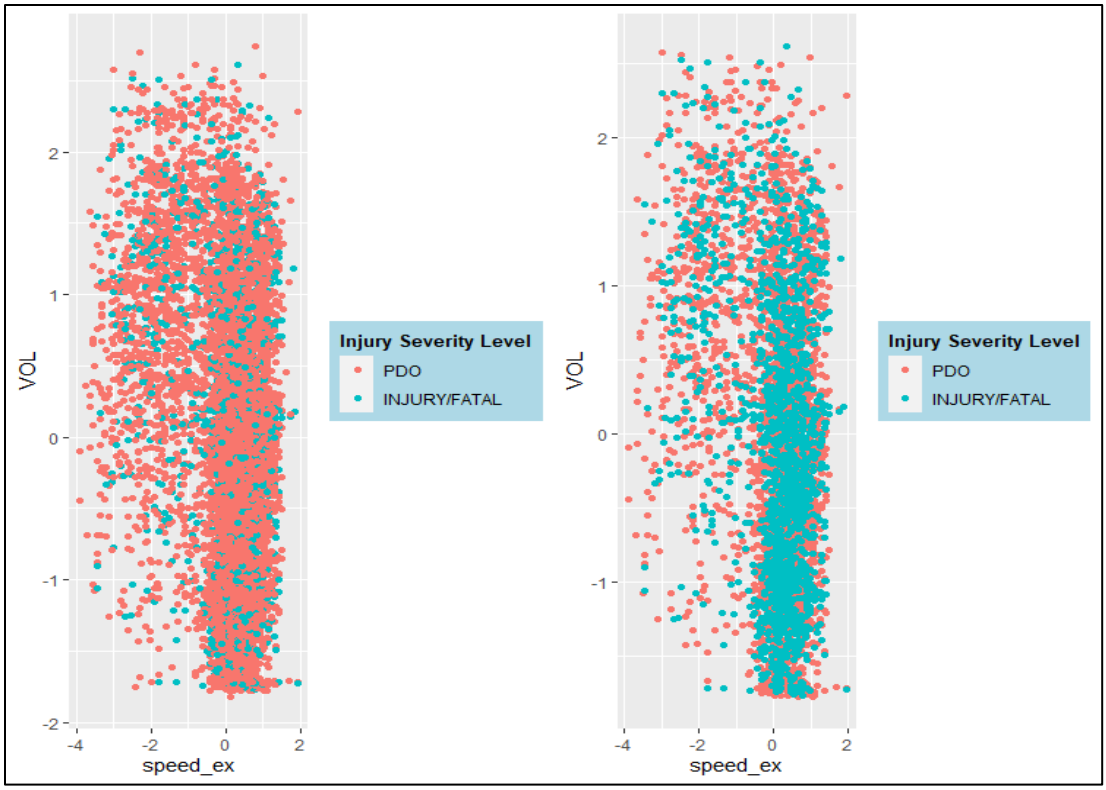

Figure 6. Deviation of speed vs. volume in the crash injury severity dataset: before ROSE (right) and after ROSE (left)

Model Performance Criteria

The quality of the predictions provided by different models considered in this study was evaluated based on the confusion metrics and its related performance measures: overall accuracy, sensitivity, specificity, and F1-score, as well as the AUC value. Calculating these metrics requires obtaining the True positive (TP), the True negative (TN), the False positive (FP), and the False negative (FN) predictions first. The definition of these values is provided as follows:

- *TP*: True positive value is defined as the number of crash cases (injury/fatality cases in the injury severity model) that are correctly predicted as crash cases (injury/fatality cases).
- *TN*: True negative value is defined as the number of non-crash cases (PDO cases in the injury severity models) that are correctly predicted as non-crash cases (PDO cases).
- *FP*: False positive value is defined as the number non-crash cases (PDO cases) that are falsely predicted as crash cases (injury/fatal cases).
- *FN*: False negative value is defined as the number of crash cases (injury/fatality cases) that are falsely predicted as non-crash cases (PDO cases).

Having TP, TN, FP, and FN, the performance measures can be formulated as:

$$\text{Overall accuracy} = \frac{TP+TN}{Total\ crashes} \quad (11)$$

$$\text{Sensitivity (True Positive Rate, Recall)} = \frac{TP}{TP+FN} \quad (12)$$

$$\text{Specificity (True Negative Rate)} = \frac{TN}{TN+FP} \quad (13)$$

$$\text{F1-score} = \frac{2\times TP}{2\times TP+FP+FN} \quad (14)$$

The closer the values of each of these measures is to 1, the better the prediction. However, very often the prediction models would provide better performance relative to one of these measures, and comparably worse performance relative to the other measure(s). Understanding the implications of the balance (or rather imbalance) of these measures in the model output is one of the critical aspects of interpreting the modeling results.

RESULTS

Model Application

In the Bayesian Logistic Regression (BLR) models applied in this study, the parameters are specified to be uninformative normally distributed priors, i.e., Normal (0, 10^{-6}) (Xu et al., 2014). The STATA data analysis software is used to calibrate the Bayesian logistic regression model (StataCorpLLC, 2015). The Bayesian model is applied utilizing a Markov Chain Monte Carlo (MCMC) algorithm (Gilks, 2005). Two chains of 12,500 iterations are set up based on the size of data and convergence speed and the first 2,500 samples are considered as burn-in. To consider the explanatory variable as significant, 95% Bayesian Credible Interval (BCI) should be reached (Gelman, 2003). The explanatory variable is statistically significant if zero is not included in the range of 95% confidence interval of the coefficient (Lunn et al., 2012). To evaluate the Bayesian models, deviance information criteria (DIC) are one of the factors utilized for model complexity and fit. DIC measures the goodness-of-fit in the model corresponding to the negative likelihood of the model as well as a penalty term corresponding to the number of coefficients. DIC's penalty term is measured by the deviation between the expected log-likelihood and the log-likelihood at the posterior mean point. The Bayesian logistic model with smaller values of DIC is preferable (StataCorpLLC, 2015). In this project, the Bayesian logistic regression models of crash severity and crash likelihood are estimated separately. The models are fitted on the training datasets treated with the oversampling and were then evaluated on the test dataset to derive the performance metrics.

The rest of the models were implemented in R statistical software using CARET package version 6.0-86 (M. Kuhn et al., 2020). A 5-fold cross validation was performed for all models to evaluate their performance. In addition, the preprocessing step included centering and scaling of all the continuous variables used in the models.

In developing and tuning the machine learning models, several parameters (referred to as hyperparameters) are considered and calibrated for the RF, GBM and KNN models. The set of tuning parameters that were found to yield the highest AUC value for the RF, GBM, and KNN models are summarized in Table 8.

Table 8. Summary of the Hyperparameters for the RF, GBM, and KNN Models

Model	Hyperparameters for the crash likelihood analysis	Hyperparameters for the crash injury severity analysis
RF	mtry = 4, split rule = extra tree, node size= 1, sample size = full training set	mtry = 2, split rule = extra tree, node size= 1, sample size = full training set
GBM	ntree = 50, interaction.depth = 3, shrinkage = 0.1, n.minobsinnode = 10	ntree = 250, interaction.depth = 5, shrinkage = 0.1, n.minobsinnode = 10
KNN	K = 5	K = 5

In the RF model, after preparing the training data, the OOB sample and 5-fold cross-validation based experimental design were used separately, to determine the optimal hyperparameters for the RF. Similar results were achieved through OOB error minimization and cross-validation. For the crash likelihood analysis, both approaches found that the combination of *mtry* = 4, *split rule* = extra trees, *node size* = 1, and *sample size* = full training set, to create the model with the lowest OOB error and highest AUC value. Using a similar approach for the injury severity analysis, the parameters *mtry* = 2, *split rule* = extra trees, *node size* = 1, and *sample size* = full training set, were found to yield the best result in terms of the AUC value.

In the GBM model, an important factor is the selection of the number of trees. Finding the optimal number of trees (*n.trees*) is a challenging task: larger number of trees contributes to good learning, while it might also increase the risk of overfitting (Opitz & Maclin, 1999). The size of the trees is another parameter which is indicated by *interaction.depth* in the R model and accounts for the order of predictor-to-predictor interaction captured in the model (Hastie et al., 2009). The learning rate or *shrinkage* is another hyperparameter pertaining to GBM, which determines the effect of each tree on the output result and takes values between 0 and 1. Overall, lower learning rates provide better results by adding more trees to the iteration (Friedman, 2001). Finally, the parameter *n.minobsinnode* defines the minimum number of observations allowed per node. In general, larger values of *n.minobsinnode* generate smaller trees that are less impacted by noise. Using a 5-fold cross-validation, the set of parameters *n.trees* = 250, *interaction.depth* = 5, *shrinkage* = 0.1, and *n.minobsinnode* = 10 was found to yield the result with the highest AUC value for the crash likelihood analysis. For the crash injury severity analysis, the set of parameters *n.trees* = 250, *interaction.depth* = 5, *shrinkage* = 0.1, and *n.minobsinnode* = 10 was found to return the best model in terms of the AUC value.

To tune the KNN model, one should find the optimal number of neighbors (*K*). The 5-fold cross-validation results showed that *K*= 5 produced the model with the highest AUC value in both the crash likelihood and the crash injury severity analysis.

Summary of Results – Crash Likelihood Model

The estimation of the BLR model is summarized in Table 9. As shown in the table, hourly precipitation, and speed deviation have positive impacts on crash occurrence, while v/c ratio, hourly visibility, and speed deviation from the speed limit have negative relationships with the crash occurrence. All explanatory variables except link capacity (CAPLINK) are significant at the 95% Bayesian credible intervals (BCI). To evaluate the Bayesian model, deviance information criteria (DIC) and AUC values are achieved as 82053.82 and 0.58, respectively. The DIC value is lower than the null model indicating that explanatory variables improve the model fit. The Odds Ratios suggest that based on the sample the odds of a crash increase by 6.4% with one unit increase of hourly precipitation, or by 5.7% with one unit increase in standard deviation of speed, while holding all other variables constant. Similarly, the odds of a crash decrease by 19.6% with one unit increase in deviation of speed from the speed limit, or by 13.1% with one unit increase in hourly visibility, while holding all the other independent variables constant.

Table 9. Summary of the Bayesian logistic regression model for crash likelihood

Variables	Mean	Std.	Odds Ratio	95% BCI
CAPLINK	-0.003	0.008	0.997	(-0.019, 0.013)
VC_RATIO	-0.044	0.008	0.957	(-0.061, -0.027)
HourlyPrecipitation	0.062	0.012	1.064	(0.039, 0.086)
HourlyVisibility	-0.140	0.008	0.869	(-0.156, -0.124)
speed_ex	-0.217	0.009	0.804	(-0.235, -0.201)
speed_sd	0.056	0.008	1.057	(0.039, 0.072)
Constant	-0.015	0.007	-	(-0.029, 0.002)

The performance statistics for the BLR, DT, RF, GBM, NB, and KNN models in terms of the overall accuracy, sensitivity, specificity, F1-score, and the AUC value is summarized in Table 10 and Figure 7. It should be noted that lager values for all metrics indicate better performance of the models. Looking at the AUC values, all models are very close, with GBM slightly outperforming the other models with the AUC value equal to 0.59.

Table 10. Crash likelihood models' performance summary

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC
BLR	0.64	0.27	0.82	0.34	0.58
DT	0.69	0.32	0.79	0.30	0.56
RF	0.70	0.30	0.80	0.29	0.56
GBM	0.66	0.43	0.72	0.35	0.59
GNB	0.63	0.45	0.68	0.34	0.58
KNN	0.54	0.56	0.50	0.32	0.54

It can be observed that RF provides the highest overall accuracy (0.70), but has very low sensitivity of 0.32 and the lowest F1-score among all the investigated models. These results for the RF model are closest to those obtained from the DT model. In fact, the performance of the RF and DT models with respect to all performance criteria is very close.

In relation to the sensitivity values, which indicated the capability of the models to correctly predict the crash cases, KNN has the highest sensitivity value (0.56), followed by GNB (0.45), and GBM (0.43). At the same time, the BLR model has the lowest sensitivity by only correctly predicting 27% of the crash cases.

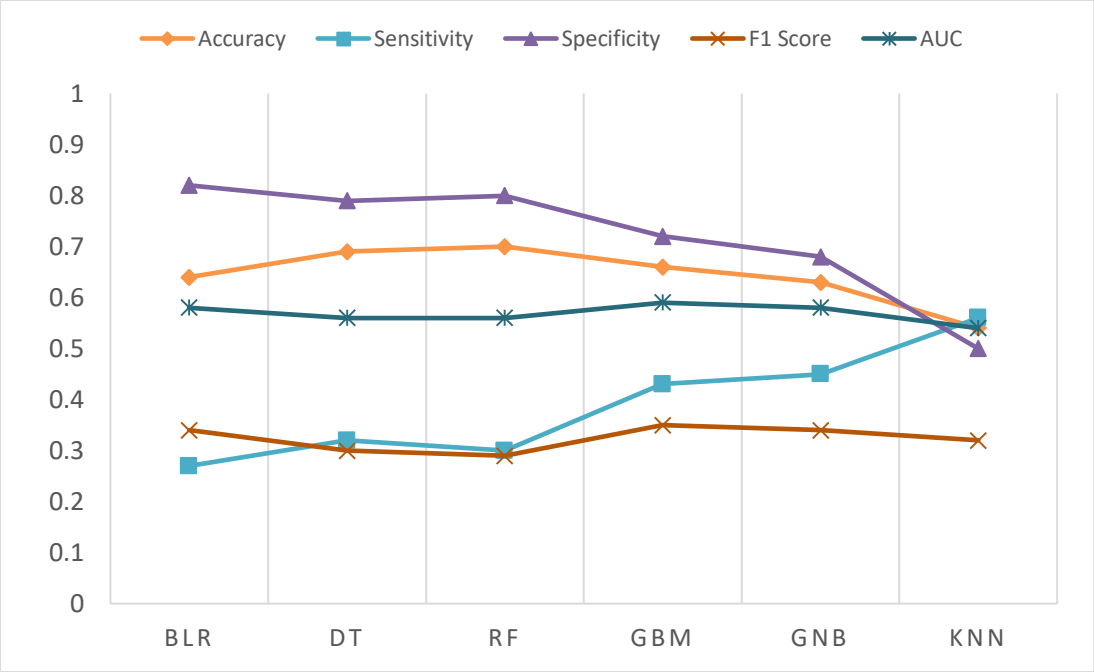

Figure 7. Crash likelihood models' performance summary (graph)

In terms of specificity, which reflects the ability of the models to correctly predict non-crash cases, the BLR has the highest value of 0.82, while the values for DT and RF are just slightly lower. The lowest specificity has the KNN model (0.50), which balances the sensitivity value (0.56). Therefore, despite the oversampling treatment of the training dataset, most of the models tend to favor majority class (non-crash cases) over the minority class (crash cases).

Overall, GBM appears to demonstrate the best performance of all tested models. It has the highest AUC value and F1-score, and the overall accuracy is comparable to slightly higher values achieved by the RF and DT. Nevertheless, even that performance cannot lead to a recommendation for a practical application of this model as it correctly predicted only 43% of crash occurrences in the testing.

Summary of Results – Crash Severity Model

The summary of findings in the BLR model for crash severity is provided in Table 11. The results show that an increase in link capacity, speed deviation from the speed limit, and standard deviation of speed result in an increase in crash severity, while the vehicle volume, hourly precipitation, and hourly visibility have negative relationships with crash severity. All the parameters except for the constant of the model are statistically significant at the 95% credible intervals. The DIC and AUC values of this model are equal to 16,850.74 and 0.54, respectively. Compared to the null model, the DIC value of this model is lower, which means that the explanatory variables help the model fit. The Odds Ratios indicate that an increase by one unit in the standardized values of link capacity, or standard deviation of speed, or deviation of speed from the speed limit, result in an increased odds of an injury of fatal crash by 8.2%, 6.4%, and

6.2% respectively, while holding all other independent variables constant. Similarly, an increase by one unit in the standardized values of vehicle volume, or hourly precipitation, or hourly visibility, reduce the odds of an injury or fatal crash outcome by 8.1%, 5.5%, and 4.2%, respectively, while holding all other independent variables constant.

Table 11. Summary of the Bayesian logistic regression model for crash severity

Variables	Mean	Std.	Odds Ratio	95% BCI
CAPLINK	0.079	0.018	1.082	(0.043, 0.114)
VOL	-0.084	0.019	0.919	(-0.124, -0.045)
HourlyPrecipitation	-0.057	0.024	0.945	(-0.106, -0.102)
HourlyVisibility	-0.043	0.018	0.958	(-0.080, -0.005)
speed_ex	0.059	0.021	1.062	(0.019, 0.101)
speed_sd	0.062	0.019	1.064	(0.026, 0.099)
Constant	-0.029	0.018	-	(-0.065, 0.008)

The performance metrics for the BLR, DT, RF, GBM, NB, and KNN models for the crash injury severity analysis is summarized in Table 12 and Figure 8. According to the results, the models have very similar AUC: for RF and KN the AUC = 0.51, and for all the other models AUC = 0.5. Among the models with AUC = 0.54, DT has the highest overall accuracy (0.69), but just like in the case of the crash likelihood model it has a very low sensitivity (0.28) and F1 score (0.25). In fact, the maximum sensitivity among all crash severity models is 0.5, achieved by the GBM, GNB, and KNN. All these three models have a similar specificity (between 0.49 and 0.54), which means that they are random – they correctly classify about 50% of events, either PDO or injury/fatal crashes. It can be observed that GBM and GNB have almost identical performance across all performance metrics. Overall, it can be concluded that none of the models is adequate in terms of predicting the crash severity.

Table 12. Crash severity models' performance summary

Model	Accuracy	Sensitivity	Specificity	F1 Score	AUC
BLR	0.57	0.22	0.82	0.30	0.54
DT	0.69	0.28	0.78	0.25	0.54
RF	0.66	0.25	0.77	0.23	0.51
GBM	0.53	0.50	0.54	0.30	0.54
GNB	0.53	0.50	0.54	0.30	0.54
KNN	0.51	0.50	0.49	0.28	0.51

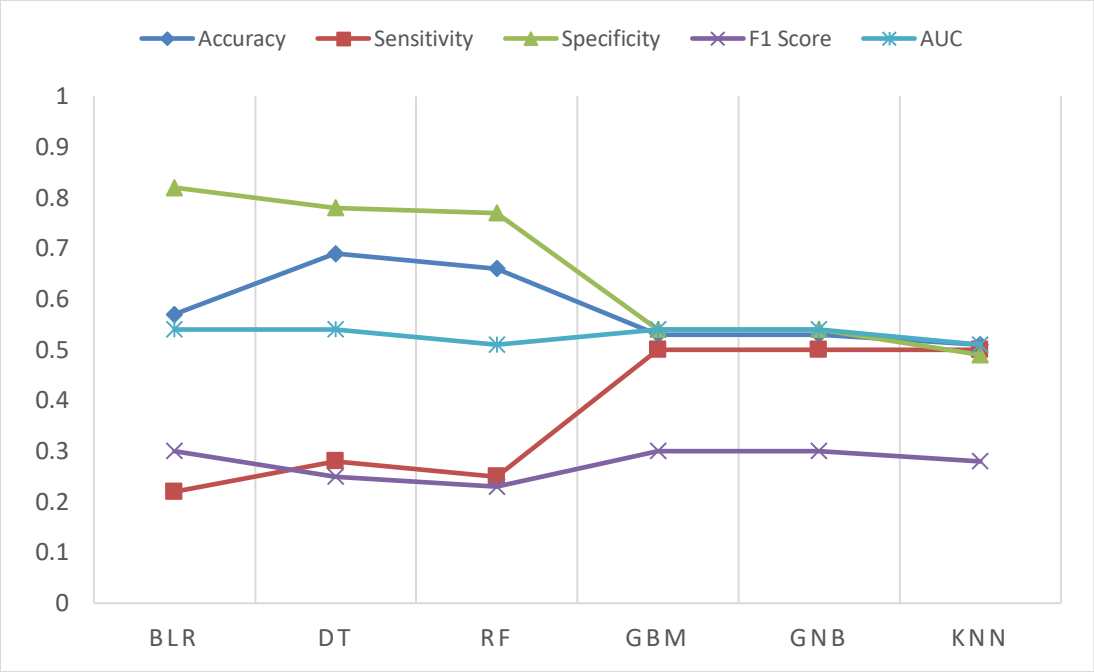

Figure 8. Crash likelihood models' performance summary (graph)

CONCLUSIONS

This main goal of this study was to apply advanced data analytics methods to develop and evaluate crash severity and crash likelihood prediction models that can be used in near-real time. For this purpose, the models were built using the data that is available to regional transportation agencies in real time and provides coverage of all major highway facilities on a reginal or statewide scale. The dataset applied in the study consisted of data collected for two interstate highways in New Jersey – I-80 and I-287, and included detailed crash data from the New Jersey State DOT crash records database, basic roadway geometry data, synthetic vehicle volume and capacity data, probe-vehicle traffic speed data, and weather data from the National Weather Service. All data is available in real time and is provided on a roadway segment level, which range in length between 0.02 miles and 5.14 miles.

The crash records dataset consisted of 10,155 crashes, including 2,139 crashes with an injury or fatal outcome, and 8,016 PDO crashes. For the crash likelihood model additional records were created to represent non-crash cases following the matched case–control methodology. To deal with the data imbalance between the crash cases and non-crash cases in the crash likelihood model, as well as between PDO and injury/fatality crashes in the crash severity model the study employed the random oversampling examples (ROSE) method. The relative importance of explanatory variables was evaluated using RF model and they were ranked based on mean decrease accuracy. The crash likelihood models had six significant explanatory variables, including the average deviation of speed from the speed limit, hourly precipitation, highway segment capacity, v/c ratio, hourly visibility, and standard deviation of speed over the 15 minutes preceding the crash. The crash severity models also had six significant explanatory variables, five of which were the same as in the crash likelihood model, with only synthetic hourly vehicle volume replacing the v/c ratio as the significant variable.

The BLR model further revealed (or rather confirmed) the significance of each explanatory variable in both crash likelihood and crash injury severity analyses. The Odds Ratios were calculated for all explanatory variables, and, for instance, showed that hourly precipitation and standard deviation of speed increased the odds of crash occurrence. Also, standard deviation of speed and deviation of speed from the speed limit were found to increase the odds of crash severity model. In addition to the BLR model, five additional machine learning (ML) methods were implemented for crash likelihood and crash severity prediction. A 5-fold cross-validation method was applied for tuning all ML models, which produced optimal combination of the hyperparameters for each model, as applicable.

The prediction accuracy of all models was evaluated using the performance metrics including the overall accuracy, sensitivity, specificity, F1-score, and the AUC value. The crash likelihood model estimation results revealed that the GBM model outperformed all the other investigated models in terms of AUC value (0.59) and F1-score (0.35). The RF provided the highest overall accuracy (70%), however it was only able to correctly identify 30% of the crash cases in the testing set. In conclusion, even the best performing model of crash likelihood could be characterized as having limited predictive value based on the performance metrics.

The results of the crash injury severity models were similar. Similar to crash likelihood, the GBM model was found to be the best performing model in terms of the AUC values and F1-score. DT was found to provide the highest overall accuracy (69%), while correctly predicting only 28% of severe (injury or fatal) crashes. Therefore, similar conclusion could be drawn regarding the crash severity models as in the case of crash likelihood models in this study – they provide limited predictive value, at best.

At the outset of the study, the aim was at developing models that would allow the transportation agencies and decision makers to assess the crash likelihood and anticipated severity of crashes in near-real-time, using the data already available to them. That in turn would allow them to make more effective operational decisions and implement operational countermeasures and tactics to reduce the likelihood and severity of crashes. Some examples would include proactive activation of advanced warnings on variable message sign (VMS), adjustments of variable speed limits (VSL) and ramp metering (RM), as well as deployment of highway safety patrols and other traffic operations and management assets.

The results of the analysis hint that the data used in this study is not sufficient or sufficiently informative to enable satisfactory separation of crash outcome and severity classes in the crash dataset. In that sense, and considering results of numerous previous studies and literature, it can be suggested that the impact of driver characteristics (e.g., driver age and gender, alcohol/drug usage, etc.), vehicle characteristics (e.g., vehicle type and age), and roadway condition characteristics may be more significant, even more critical, than the variables considered in this study. It should also be noted that most of the reviewed studies dealing with the real-time crash risk prediction are based on the real-time traffic counts and density collected from Automatic Vehicle Identification (AVI) and real-time weather data collected from weather stations, both with greater spatial and temporal resolution than the data used in this study. However, this kind of data is mostly available at specific, well-instrumented roadway segments, without providing a coverage of a larger regional scope. Application of models on a limited local scale where such data is available, even if they were highly accurate, would present a challenge in making regional operations decisions. And that precisely was the subject of analysis in this study.

RECOMMENDATIONS

The analysis conducted in the study presented in this report presents a solid basis for the research team's future work on crash prediction and related operational decision making. Considering the limited scope of the study, given that the analysis was based on the data from only two interstate highways, the future research would certainly benefit from including a larger sample from multiple roadways, including different roadway types.

Another aspect that may lead to an improved performance of crash risk and crash severity models is increased resolution of underlying data, specifically traffic and road-weather data. With advancements in ITS, increasing roadway instrumentation, and data analytics capabilities, the real-time or near-real time data on vehicle volume and occupancy will become more available and accessible for an increasing number if roadways and roadway networks. Similar can be expected with road-weather data, as the increasing number of agencies is pursuing expansion of road weather information systems (RWIS) to include both stationary and mobile sensors. Combined, stationary and mobile RWIS data can greatly increase the coverage and accuracy of road-weather data that can be used in traffic safety research.

Lastly, future research should also include analysis of the model sensitivity to different factors represented by the explanatory variables used in this study. While the Bayesian regression model did provide some analysis in this respect, the sensitivity of the ML models was not analyzed. This would provide additional understanding of the shortcomings of these models and would shed more light on potential pros and cons of including certain variables in the ML models. Furthermore, addition of the variables that have not been considered in this study, especially those that could serve as surrogates for driver characteristics and behavior, or vehicle characteristics, may also be beneficial towards the improved accuracy and predictive capability of the ML models.

REFERENCES

[1] Ahmed, M. M., & Abdel-Aty, M. A. (2011). The viability of using automatic vehicle identification data for real-time crash prediction. IEEE Transactions on Intelligent Transportation Systems, 13(2), 459-468.

[2] Bishop, C. M. (2006). Pattern recognition and machine learning: springer.

[3] Bowman, A. W., & Azzalini, A. (1997). Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations (Vol. 18): OUP Oxford.

[4] Breiman, L. (2000). Some infinity theory for predictor ensembles. Technical Report 579, Statistics Dept. UCB.

[5] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees: CRC press.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[7] Cigdem, A., & Ozden, C. (2018). Predicting the severity of motor vehicle accident injuries in Adana-turkey using machine learning methods and detailed meteorological data. International Journal of Intelligent Systems and Applications in Engineering, 6(1), 72-79.

[8] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[9] Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. International Statistical Review, 71(2), 369-382.

[10] Gilks, W. R. (2005). Markov chain monte carlo. Encyclopedia of Biostatistics, 4.

[11] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction: Springer Science & Business Media.

[12] Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., . . . Team, R. C. (2020). Package 'caret'. The R Journal.

[13] Kuhn, S., Egert, B., Neumann, S., & Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. BMC bioinformatics, 9(1), 400.

[14] Kwon, O. H., Rhee, W., & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. Accident Analysis & Prevention, 75, 1-15.

[15] Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). The BUGS book: A practical introduction to Bayesian analysis: CRC press.

[16] Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28(1), 92-122.

[17] National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce (2019). Local Climatological Data (LCD) Dataset Documentation. Downloaded from https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/LCD_documentation.pdf.

[18] Nicodemus, K. K. (2011). Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. Briefings in bioinformatics, 12(4), 369-373.

[19] Shanthi, S., & Ramani, R. G. (2011). Classification of vehicle collision patterns in road accidents using data mining algorithms. International Journal of Computer Applications, 35(12), 30-37.

[20] StataCorp, L. (2017). Stata Bayesian analysis reference manual.

[21] Theofilatos, A. (2017). Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. Journal of safety research, 61, 9-21.

[22] Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing machine learning and deep learning methods for real-time crash prediction. Transportation Research Record, 2673(8), 169-178.

[23] Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. Monographs on statistics and applied probability, 57, 1-436.

[24] Wang, L., Shi, Q., & Abdel-Aty, M. (2015). Predicting crashes on expressway ramps with real-time traffic and weather data. Transportation Research Record, 2514(1), 32-38.

[25] Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Predicting crash likelihood and severity on freeways with real-time loop detector data. Accident Analysis & Prevention, 57, 30-39.

[26] Xu, C., Wang, W., Liu, P., Guo, R., & Li, Z. (2014). Using the Bayesian updating approach to improve the spatial and temporal transferability of real-time crash risk prediction models. Transportation research part C: emerging technologies, 38, 167-176.

[27] Yu, R., & Abdel-Aty, M. (2013). Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention, 51, 252-259.

[28] Yu, R., & Abdel-Aty, M. (2014a). Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety science, 63, 50-56.

[29] Yu, R., & Abdel-Aty, M. (2014b). Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accident Analysis & Prevention, 62, 161-167.