INVESTIGATING THE IMPACT OF ROADWAY GEOMETRY, SPEED DISTRIBUTION, AND WEATHER CONDITION ON ROADWAY DAILY CRASH OCCURRENCE AND SEVERITY BY USING MACHINE LEARNING METHODS

A Thesis

by

ZIHANG WEI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Yunlong Zhang
Co-Chair of Committee,	Xiubin B. Wang
Committee Members,	David E. Jones
Head of Department,	Robin Autenrieth

August 2021

Major Subject: Civil Engineering

ABSTRACT

Conventional traffic crash analysis methods often use highly aggregated data, making it difficult to understand the effects of many time-varying factors on crash occurrence. Although studies have used data with small aggregation intervals, they typically analyze the effect of a single factor on crash occurrence. In this study, the collaborative effect of roadway geometry, speed distribution, and weather conditions on crash occurrence and severity is investigated using an interpretable or explainable machine learning method XGBoost (eXtreme Gradient Boosting) on daily level crash data. The data are collected from four different sources on roadways in Texas. Three roadway facility types are considered in this study: (1) Rural Interstate; (2) Rural Two-Lane; (3) Rural Multilane. In the feature selection process, the Pearson correlation coefficient is applied to remove highly correlated variables. The study then uses the synthetic minority over-sampling technique (SMOTE) method to mitigate the data imbalance issue. The XGBoost model is trained twice: first on data with all crash severity levels, and then only on data with fatal and severe injury crash levels. Finally, the SHAP (SHapley Additive exPlanation) method is applied to investigate the contribution of all variables on the model's output. The results show that on different roadways facility types the contributions of variables tend to be different, and moreover, the variables also contribute differently on crashes with different severity levels.

ACKNOWLEDGEMENTS

First of all, I would like to thank my committee chair, Dr. Yunlong Zhang, my committee co-chair, Dr. Bruce Wang, and my committee members, Dr. David Jones and Dr. Subasish Das for giving me their guidance and support on this master thesis. I would also like to appreciate all the transportation engineering faculty members at the Zachary Department of Civil and Environment Engineering for helping me throughout my two-year study.

My special thanks also go to Dr. Subasish Das at Texas A&M Transportation Institute (TTI) where I have been spending one year as a graduate research assistant. Dr. Das provides me lots of valuable research experience besides coursework which will tremendously benefit my research career in the future.

I would also like to thank my parents for supporting me during my study at Texas A&M University. Your supports help me concur so many difficulties and help me go through the darkest time. Thanks again for your encouragement, patience, and love.

Finally, I would like to thank all the friends I meet here at Texas A&M University. Thanks for your companies which make my life at Aggieland unforgettable.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Yunlong Zhang [committee chair] and Professor Xiubin B. Wang [committee co-chair] of the Department of Civil and Environmental Engineering and Professor David E. Jones of the Department of Statistics and Dr. Subasish Das of Texas A&M Transportation Institute.

The raw data in this study were from a project funded by the Texas Department of Transportation at Texas A&M Transportation Institute. I was part of this project as a graduate research assistant.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

To the student's best acknowledgment, there is no outside funding source for this thesis.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

1. INTRODUCTION

Traditional roadway crash analysis methods usually use highly aggregated data. Roadway and crash-related variables are often aggregated over a large time interval. Thus, for some explanatory variables that may change significantly during this interval, these variations are usually not considered because of the lack of detailed information (*1*). However, for many time-varying explanatory variables such as speed and weather information, the variations inside these variables are very important for crash analysis as well. Many studies have found that weather conditions, especially precipitation and visibility (*2*)(*3*)(*4*), and speed distribution, especially average speed and speed variation (*5*)(*6*)(*7*), are closely related to crash occurrence. It is unreasonable to aggregate these variables over a long time interval. For example, if two specific locations have identical monthly average precipitation values, it is problematic to say these two locations will have the same crash occurrence level because precipitation varies from day to day. Thus, ignoring the potential within-period variation in these explanatory variables may result in the loss of valuable information. The best way to avoid this problem is to aggregate crash data into smaller time intervals. In particular, data can be aggregated into daily, hourly, or even minute-by-minute levels.

However, on the other hand, many other explanatory variables such as road geometry data (i.e., curve, lane width, and shoulder width) are relatively static. For the same roadway segment, road geometric data rarely change over any time interval. By aggregating roadway geometry data into smaller intervals, it means that for the same

roadway segment, more identical observations will be generated. Moreover, roadway segments that are close to each other always share similar geometry features which result in correlation over space. As a result, the temporal and spatial correlation will negatively affect the analysis of these relatively static explanatory variables (*8*)(*9*).

Many studies have studied the effect of roadway geometric data on crash occurrence and some have studied the effect of speed distribution and weather conditions. However, few of them are conclusive enough, and even fewer researchers have investigated the collaborative effect of these three factors on crash occurrence.

To address the problem of time-varying variables and the temporal and spatial correlation in crash frequency analysis, data should not only be aggregated into small time intervals, but also include lots of roadway segments with various geometry features. There are three main aggregation intervals previously studied by researchers. The first one is the yearly aggregation interval under which the effect of roadway geometric can be properly analyzed. However, the effect of speed distribution and weather conditions cannot be analyzed under a yearly aggregation interval. The second one is the daily aggregation interval under which the effect of roadway geometric and weather conditions can be properly analyzed. However, it is not ideal to study speed distribution under this interval. The third one is hourly or minute-by-minute intervals under which the effect of speed distribution can be properly analyzed. However, it is not ideal for studying the effect of roadway geometric and weather conditions. Considering the above-mentioned facts, this study chooses the daily aggregation interval to prepare the dataset. Moreover, since speed distribution tends to vary differently throughout different time periods within a day, this

study introduces speed measurement variables of different periods during a day to address this problem. For example, average speed and speed variation are calculated for daytime and nighttime separately.

2. RESEARCH OBJECTIVE

This study develops a comprehensive roadway crash dataset that contains segments of three different facility types: Rural Interstate, Rural Two-Lane, and Rural Multilane in the state of Texas. The goal of this study is to analyze the hidden relationship between daily crash occurrences, roadway geometric, speed distribution, and weather conditions on these three different roadway facility types and to compare the variables' relative contributions on affecting roadway daily crashes occurrence by using machine learning techniques. Moreover, this study will also investigate various explanatory variables' contributions on crashes with different severity levels (i.e., all crashes and severe crashes) and see whether the same explanatory variable contributes differently to crash occurrences of different severity levels.

3. LITERATURE REVIEW

Many previous researchers have studied the effects of roadway geometry, weather conditions, and speed distribution factors on crash occurrence separately. Some researchers have also studied the collaborative effects of two of the three factors. For example, Shankar et al. studied highway crash frequency by analyzing the effect of geometric elements and weather conditions. The study was conducted by applying a negative binomial model, and the data were aggregated into a monthly interval (*2*). Dutta and Fontaine introduced crash prediction modeling on a freeway segment using disaggregated speed data and roadway geometric data. The results indicated that by including hourly averaged speed data and selected roadway geometric data, the crash prediction performance improved compared with the one using annual data without speed information (*10*).

Many studies have analyzed the relationship between roadway geometric features and crash occurrence or severity. Miaou and Lum utilized two linear regression models and two Poisson regression models to study the relationship between roadway geometric factors and crash frequency (*11*). Anderson et al. applied Poisson, negative binomial, and log-normal regression analysis to study the relationship between rural two-lane highway crash frequency and roadway geometric design consistency (*12*). Haghighi et al. investigated the effect of roadway geometric factors on crash severity with data collected from rural two-lane highways. They developed a multilevel ordered logit model to deal

with the hierarchical structure of the crash data. They found that the introduction of crash type as a variable can better explain the crash severity levels variation (*13*).

Speed distribution is another important contributor to crash occurrence. Garber and Gadiraju investigated the impact of mean speed and speed variation on crash rate. They concluded that a higher mean speed does not necessarily increase the crash rate, but a higher speed variation can lead to a higher crash rate (*14*). Lee et al. used real-time traffic flow data from loop detectors to predict crash occurrence. They applied an aggregated log-linear model to estimate the crash occurrence and found that speed variation and traffic density are strong indicators of crash frequency (*15*). Pei et al. analyzed the effect of mean speed on crash occurrence using disaggregated speed and crash data within a 4-hour interval from different periods of a day. They found that the mean speed and crash occurrence are positively related when distance exposure is considered, however, mean speed and crash occurrence are negatively related when time exposure is considered (*16*). Wang et al. studied the relationship between mean speed, speed variation, and crash frequency on arterials in urban areas. A hierarchical Poisson log-normal (HPLN) model was applied to model the crash frequency. Since speed distribution tends to vary significantly during different periods, this study aggregated crash data into three study periods (morning, midday, and evening), each period being three hours long. The results revealed that a higher average speed and higher speed variation will lead to higher crash frequencies on urban arterial roads (*7*).

Weather condition factors can significantly affect crash occurrence as well. Scott included temperature and rainfall as explanatory variables to model the time-series crash

data. A regression model was applied to model single-vehicle crashes, and a Box-Jenkins model was applied to model two-vehicle crashes (*17*). Eisenberg analyzed the effects of precipitation on traffic crashes by applying the negative binomial regression method. Two data aggregation intervals (monthly and daily) were studied in this study. Its results revealed a significant negative relationship between monthly fatal crash frequency and precipitation. However, the results indicated a significant positive relationship between daily fatal crashes and precipitation (*18*). Brijs et al. applied an integer autoregressive model on daily crash data to model the time dependency nature of crash occurrences. Their results showed that the intensity of the rainfall is significantly related to the daily crash count (*19*). Jaroszweski and McNamara analyzed the influence of precipitation on crashes by utilizing weather radar images. This novel approach offers improvements to the analysis of weather-related accidents by giving a more representative rainfall measure in urban areas (*20*). Yu and Abdel-Aty analyzed the relationship between weather conditions and crash severity on mountainous freeways. The results indicate: (1) snow weather is less likely to cause severe crashes and (2) lower temperature increases the likelihood of severe crashes (*21*).

Although previous studies have investigated these three factors separately, fewer studies have considered these three factors together and analyzed their collaborative effects on crash occurrence. One major problem of studying the collaborative effects on these three factors is which data aggregation level should be used. Previous studies tended to analyze the relationship between crash and roadway geometric data based on yearly aggregation intervals (*11*) (*12*), the relationship between crash and weather condition

variables based on daily aggregation intervals or monthly aggregation intervals (*17*) (*18*) (*19*), and the relationship between crashes and speed distribution based on real-time data with hourly intervals or minute by minute aggregation (*15*) (*16*) (*7*).

By summarizing previous research, it is found that daily aggregation intervals seem ideal for collectively analyzing the effects of roadway geometry and weather conditions on crash occurrence. Although it is not ideal for analyzing the effect of speed distribution on crash occurrence using daily aggregation interval because speed distribution tends to differ significantly throughout a day, variables such as average daytime speed and average nighttime speed can be included in the daily model to alleviate this problem.

4. DATA DESCRIPTION

Data are collected from roadways in the state of Texas. Three roadway types are considered in this study: (1) Rural Interstate, (2) Rural Two-Lane, and (3) Rural Multilane. First, a comprehensive dataset is developed by using the data conflation method. The dataset analyzed in this study contains four parts: (1) roadway geometry feature and traffic information, (2) weather condition data, (3) speed measurements data, and (4) crash data. These data are collected from four different sources respectively:

- *Texas Department of Transportation Road-Highway Inventory Network Offload (RHiNO 2018)*
- *Automated Surface Observing System (ASOS)*
- *National Performance Management Research Dataset (NPMRDS)*
- *Crash Record Information System (CRIS)*

Figure 1 demonstrates the flow chart of the data preparation process. Figure 2 shows the base roadway network used in this study.

4.1. Base Roadway Network and Geometric Data

The base roadway network is collected from the Texas Department of Transportation (TxDOT) Road-Highway Inventory Network Offload (RHiNO 2018), which contains roadway GIS linework and roadway inventory attributes, including geometric features and traffic information. TxDOT submitted this dataset to Federal Highway Administration (FHWA) as part of the Highway Performance Monitoring System (HPMS) program. This study only selects rural interstates from the base network.

4.2. Speed Distribution Data

Speed data are collected from FHWA's National Performance Management Research Dataset (NPMRDS). The NPMRDS contains travel time and speed data collected from a fleet of probe vehicles (cars and trucks). The NPMRDS can generate speed and travel time data by using probe vehicle location information. The data are aggregated in 3-time intervals (5-minute, 15-minute, and 1-hour), and this study uses data with 5-minute intervals because more detailed information can be kept when calculating speed variation. Speed data are available across the National Highway System (NHS), and the spatial resolution is set by different Traffic Message Channel (TMC) location codes (*22*). Daily speed distribution variables (i.e., average speed, speed standard deviation, 85th percentile speed, etc.) are calculated based on 5-minute speed data at each TMC. Each TMC is conflated with its corresponding roadway segment by using Geographic Information system (GIS) software.

4.3. Weather Condition Data

Weather condition data are collected from the Automated Surface Observing System (ASOS) of the National Oceanic and Atmospheric Administration (NOAA). Each roadway segment is matched with an ASOS station closest to it.

4.4. Crash Data

Crash data are collected within the state of Texas from 2017 to 2019 through Crash Record Information System (CRIS). Each crash record includes location and date information. Through GIS software, all crashes are assigned to the roadway segments on which they occurred. Then, the daily crash count of each roadway segment can be

summarized. The dataset used in this study contains 2,601,106 non-crash observations and 26,210 crash observations, including 1,428 severe crash observations. The definition of a severe crash in this study is a crash that resulted in severe injuries or fatalities. In this process, crash severity is classified into 5 different levels: (1) K: Killed; (2) A: Incapacitating Injury; (3) B: Non-Incapacitating Injury; (4) C: Possible Injury; (5) O: Not Injured or Unknown.

4.5. Data Conflation

The data from the four parts above are conflated by using ArcGIS software. All data are aggregated into a daily interval. The final dataset is made up of 26 variables from the above-mentioned four parts. The total number of segments, total segment length, and the number of crashes at different severity levels (KABCO) are summarized in Table 1. In this study, all crashes include severity levels KABCO and severe crashes only include severity levels KA. The detailed definitions of all variables are listed in Table 2.

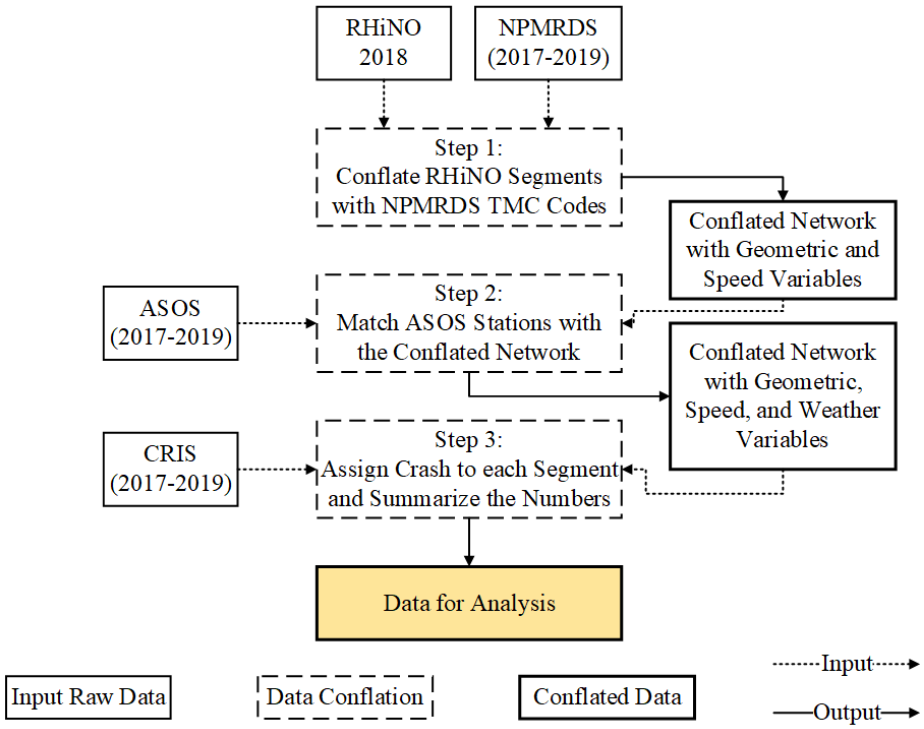


Figure 1. Flowchart of the data preparation process

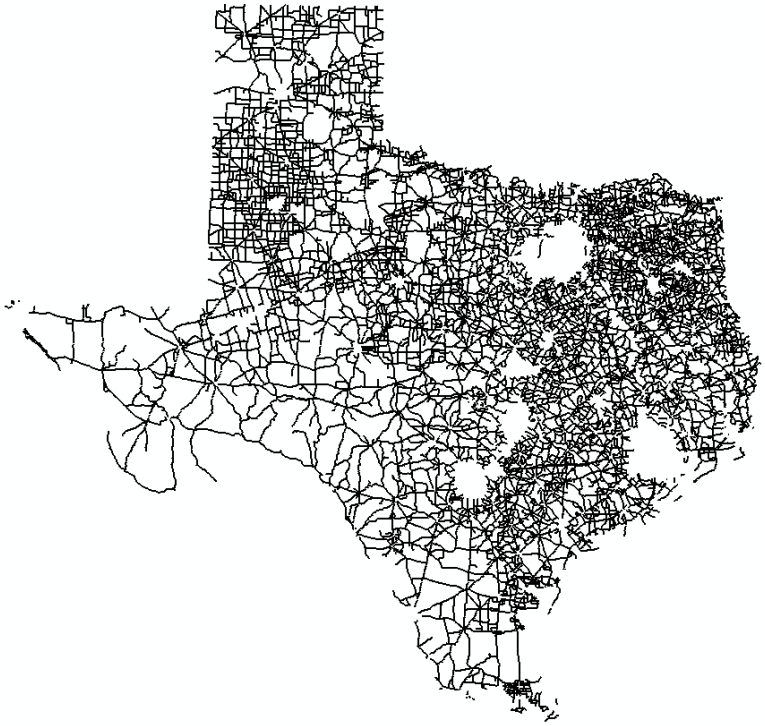


Figure 2. Roadways analyzed in this study

Table 1. Summary of segments and number of crashes at different severity levels

Roadway Facility Type	Number of Segments	Total Mileage (miles)	Number of Crashes (2017-2019)				
			K	A	B	C	O
Rural Interstate	2,594	1,743	398	1043	2863	3212	20616
Rural Two-Lane	8,247	3,055	245	527	1258	1214	6473
Rural Multilane	8,999	3,516	556	1391	3383	3425	17702

Table 2. Variables' names and definitions

Variable Names	Definition
Weather Condition	
DailyPrecip	Daily Precipitation
VsbyAve	Average Visibility
VsbyStd	Visibility Standard Deviation
Speed Distribution	
SpdAve	Average of Daily Speed
SpdStd	Standard Deviation of Daily Speed
SpdCV	Coefficient of Variation (CV) of Daily Speed
Spd85	85th Percentile of Daily Speed
RefSpd	Reference Speed
SpdAveDay	Average of Daily Speed Using Data during Daytime
SpdStdDay	Standard Deviation of Daily Speed Using Data during Daytime
SpdCVDay	Coefficient of Variation (CV) of Daily Speed Using Data during Daytime
SpdAveNight	Average of Daily Speed Using Data during Nighttime
SpdCVNight	Coefficient of Variation (CV) of Daily Speed Using Data during Nighttime
SpdStdNight	Standard Deviation of Daily Speed Using Data during Nighttime
SpdFFAve	Average of Daily Speed Larger than Reference Speed
SpdFF85	85th Percentile of Daily Speed Larger than Reference Speed
Roadway Geometry and Traffic	
SpdMax	Maximum Speed Limit
MedWid	Median Width

	NumLanes	Number of Through Lanes
	LaneWidth	Lane Width
	SWid_I	Inside Shoulder Width
	SWid_O	Outside Shoulder Width
	SrfType	Surface Type (Categorical)
	AADT	AADT
	TrkAADTP	Truck AADT Percentage
	Length	Roadway Segment Length

5. METHODOLOGY

5.1. Feature Selection

In the prepared dataset, some explanatory variables might be highly correlated with others. When machine learning models are trained on the dataset, these variables do not have extra benefits in distinguishing the target variables. Thus, to improve modeling efficiency and accuracy, these highly correlated explanatory variables need to be removed.

In this section, the feature selection and oversampling processes are presented by using the rural interstate dataset. The processes for the datasets of all three roadway facility types are the same. Firstly, the Pearson correlation coefficient is applied to evaluate feature correlation. The Pearson correlation coefficient of two variables can be calculated by Equation 1. Figure 3 shows the Pearson correlation coefficient heatmap of the rural interstate dataset. Brighter cells represent a higher correlation between two explanatory variables. Two explanatory variables are considered highly correlated if their Pearson correlation coefficient is higher than 0.7. There are 23 pairs of highly correlated explanatory variables identified from the rural interstate dataset (see Table 3).

$$c_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

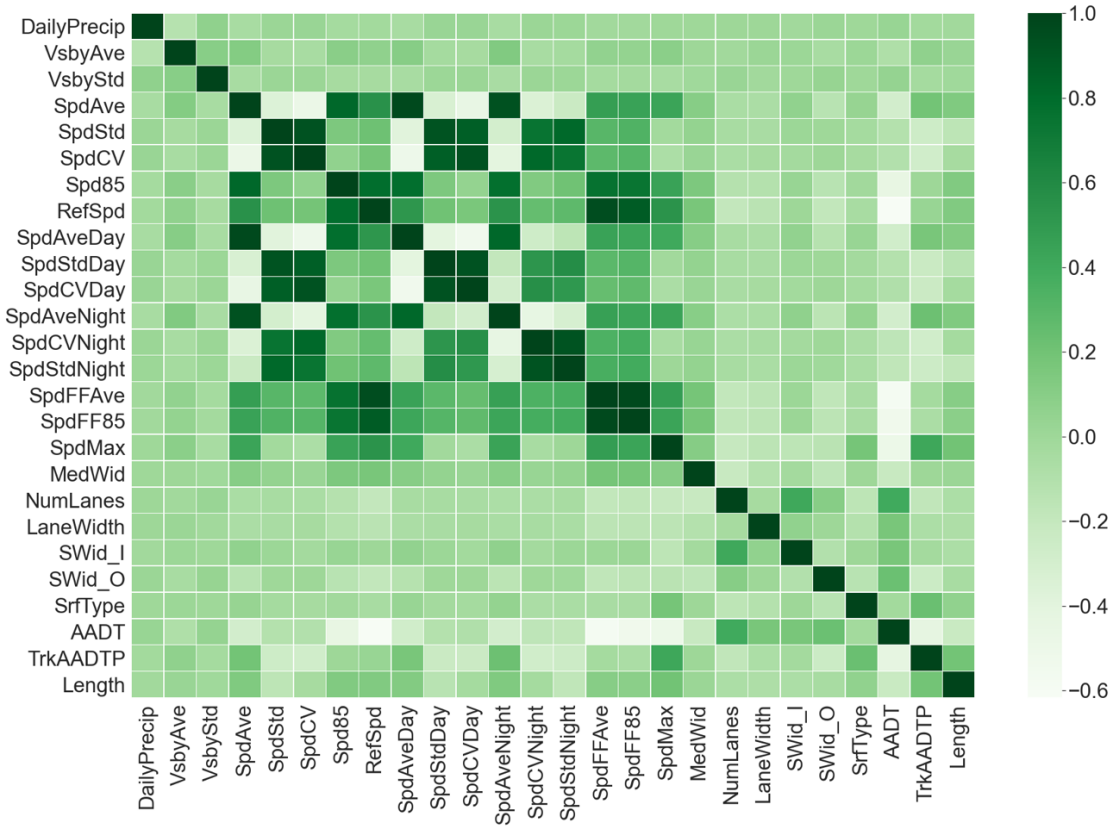


Figure 3. Pearson correlation coefficient heatmap

After Pearson correlation coefficients are calculated for all explanatory variable pairs, a Random Forest (RF) model is trained using all explanatory variables on the dataset and the feature importance values of all explanatory variables are available. All variables are ranked based on their feature importance values. The feature selection criterion is that for each highly correlated explanatory variable pair, the one with a lower feature importance value is removed. Finally, 9 explanatory variables (SpdCV; SpdStdDay; SpdStd; SpdCVNight; SpdAveNight; SpdAve; SpdFF85; Spd85; RefSpd) are removed from the rural interstate dataset.

Table 3. Correlated Variable Pairs

Identified Correlated Variable Pairs		Abs Correlation Coefficient
SpdFF85	Spd85	0.74271541
SpdCV	SpdStdNight	0.74397319
SpdCVNight	SpdStd	0.75024058
SpdFFAve	Spd85	0.76033368
Spd85	SpdAveNight	0.7782837
SpdAveDay	Spd85	0.78406856
Spd85	RefSpd	0.78770835
SpdStd	SpdStdNight	0.8128073
SpdCV	SpdCVNight	0.81324744
Spd85	SpdAve	0.8222168
SpdAveDay	SpdAveNight	0.82246082
SpdStdDay	SpdCV	0.86242186
SpdCVDay	SpdStd	0.86389696
RefSpd	SpdFF85	0.87994823
SpdStdDay	SpdStd	0.92355007
SpdCVNight	SpdStdNight	0.92385636
SpdCVDay	SpdCV	0.92874229
SpdCVDay	SpdStdDay	0.92976377
SpdCV	SpdStd	0.92985596
SpdAveNight	SpdAve	0.93383347
SpdFFAve	RefSpd	0.95036766
SpdAveDay	SpdAve	0.96857392
SpdFFAve	SpdFF85	0.97319593

5.2. Resampling Imbalanced Dataset

This study aggregates crash occurrence into a daily interval. Because of the rareness nature of crashes, for any roadway segment, crashes only happen on a small portion of the total observations. Thus, the number of non-crash observations is significantly larger than crash observations. This nature results in an imbalanced dataset. To address this problem, this study applies a resampling method to rebalance the original dataset. Many previous studies have applied resampling methods. Abdel-Aty et al. applied a matched case-control method that manually matches crash samples with non-crash samples (*23*). Chawala et al. proposed the synthetic minority over-sampling technique

(SMOTE) to address the problem of imbalanced datasets (*24*). SMOTE is an oversampling method that only oversamples the minority class, and it is applied to the training dataset only. The minority is oversampled by creating "synthetic" samples along the line segments joining the k minority class nearest neighbors. The number of neighbors is randomly chosen based on the amount of over-sampling required. Since SMOTE is not applied to the testing dataset, the testing result can still be considered to reflect reality. Many previous studies have applied SMOTE to address imbalanced datasets (*25*) (*26*) (*27*). Figure 4 presents the data points before and after oversampling process.

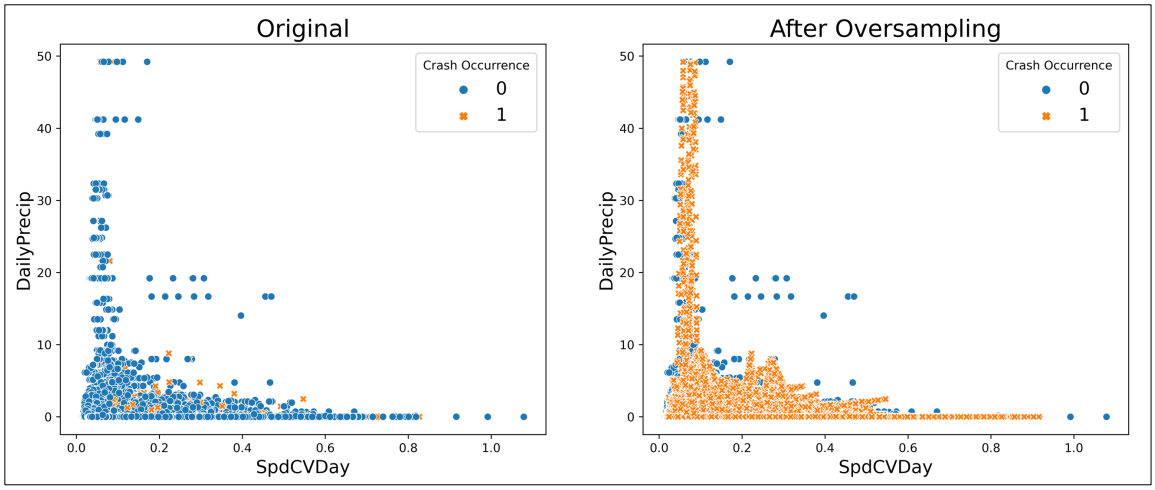


Figure 4. SMOTE oversampling method

5.3. XGBoost (eXtreme Gradient Boosting)

XGBoost (eXtreme Gradient Boosting) is a scalable end-to-end tree boosting system. It implements machine learning algorithms under the Gradient Boosting framework (*28*). XGBoost is an additive boosting tree package that is built by *k* essential tree functions implemented with regularization, missing value imputation, shrinkage and column subsampling, sparsity-aware split finding, and column block for parallel learning.

Comparing with Gradient Boosting, XGBoost can deliver more accurate approximations by using the strengths of the second-order derivative of the loss function, L1 and L2 regularization, and parallel computing. It can run more than ten times faster than existing popular solutions on a single machine, and it scales to billions of examples in distributed or memory-limited settings. The scalability of XGBoost is due to several important systems and algorithmic optimizations. These innovations include: a novel tree learning algorithm introduced for handling sparse data, and a theoretically justified weighted quantile sketch procedure enables handling instance weights in approximate tree learning. Parallel and distributed computing make learning faster which enables quicker model exploration. XGBoost can solve real-world scale problems by using a minimal number of resources. It is currently one of the fastest and best open-source boosting tree tools for modeling and prediction analyses.

Given a dataset with n observations, each observation has multiple features x_i, and a corresponding response variable y_i. $\hat{y}_i^{(t)}$ is the predicted response value after t^{th} iterations by adding one tree function $f(x_i)$ to the predicted value of the $(t-1)^{th}$ iteration corresponding to the i^{th} observation. The boosting process is shown in Equation 1.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{1}$$

The objective of this process is to minimize Equation 2. $l(y_i, \hat{y}_i)$ is a loss function and $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$ represents the penalty for the complexity of the model where T is the number of leaves and w_j^2 is the L2 norm of j^{th} leaf scores. This term is used to avoid over-fitting.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{t} \Omega\,(f_k) \tag{2}$$

By solving equations (1) to (2), the optimal value of w_j is:

$$w_j{}^* = -\frac{\sum_i \partial_{\hat{y}_i{}^{(t-1)}} l(y_i, \hat{y}_i{}^{(t-1)})}{\sum_i \partial^2_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) + \lambda} \tag{3}$$

And the corresponding minimum object value is:

$$Obj^{min} = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_i \partial_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right)\right)^2}{\sum_i \partial^2_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) + \lambda} + \gamma T \tag{4}$$

6. RESULT – MODEL COMPARISON

In order to show the advantage of the XGBoost model, the performance of the XGBoost model is compared with Random Forest - a machine learning method that has been widely used for classification in this section. These two models are both trained on the rural interstate dataset with all crash occurrences as the target variable. Table 4 summarizes the performance measures of these two models.

Table 4. Comparison between XGBoost and Random Forest

Performance Measures	XGBoost	Random Forest
Accuracy	78.7%	97.4%
Sensitivity	64.2%	14.2%
Specificity	78.8%	98.2%
Weighted Accuracy	71.5%	56.2%

Although the Random Forest model seems to have higher accuracy and specificity values, the sensitivity and weighted accuracy values are very low. Since the SMOTE oversample method is only applied to the training set, the testing set is still highly imbalanced. The Random Forest model tends to classify most of the records in the testing set as the majority class (non-crash), giving the model higher accuracy and specificity values (see Table 5). However, the sensitivity and weighted accuracy of the Random Forest is very low. This indicates that Random Forest's performance on the rural interstate dataset is not ideal.

On the other hand, the difference between sensitivity and specificity scores of the XGBoost model are small. This indicates that the XGBoost model performs better than the Random Forest model on imbalanced datasets. In the following sections, the XGBoost model will be trained on all three roadway facility types and the results will be demonstrated and discussed.

Table 5. Confusion matrix (Comparison between Random Forest and XGBoost)

Predicted Labels		**XGBoost**		**Random Forest**	
		Non-Crash	**Crash**	**Non-Crash**	**Crash**
True Labels	**Non-Crash**	512,731	137,546	638,668	11,609
	Crash	2,343	4,209	5,622	930

7. RESULT - RURAL INTERSTATE

7.1. All Crash Model

Firstly, the XGBoost model is trained on the rural interstate dataset with all crash occurrences as the target variable. The dataset is split into a training dataset and a testing dataset. The SMOTE oversampling method is applied to the training dataset to balance the minority group and majority group. The training dataset contains 1,950,829 non-crash observations and 19,658 crash observations. After the oversampling process, the numbers of both non-crash and crash observations are 1,950,829. The model evaluation is made with the testing dataset. Table 6 presents the confusion matrix of the XGBoost model. Several measurements are selected to evaluate the model performance.

Table 6. Confusion matrix (Rural Interstate)

Predicted Labels		**All Crash Model**		**Severe Crash Model**	
		Non-Crash	**Crash**	**Non-Crash**	**Crash**
True Labels	**Non-Crash**	512,731	137,546	552,184	104,288
	Crash	2,343	4,209	114	243

Performance measures for all crash model:

(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 78.7\%$

(2) $Sensitivity = \frac{TP}{TP+FN} = 64.2\%$

(3) $Specificity = \frac{TN}{TN+FP} = 78.8\%$

(4) $Weighted\ Accuracy = \frac{Sensitivity+Specificity}{2} = 71.5\%$

Performance measures for severe crash model:

(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 84.1\%$

(2) $Sensitivity = \frac{TP}{TP+FN} = 68.1\%$

(3) $Specificity = \frac{TN}{TN+FP} = 84.1\%$

(4) $Weighted\ Accuracy = \frac{Sensitivity+Specificity}{2} = 76.1\%$

Where: TP: True positive; TN: True Negative; FP: False Positive; FN: False Negative

SHAP (SHapley Additive exPlanation) is selected to interpret the feature importance in the XGBoost model. Figure 5 is the SHAP summary plot. It ranks all explanatory variables based on their impact on the model output. A variable with a higher feature importance indicates that the variable has a more determining weight on classifying observations as crash or non-crash.

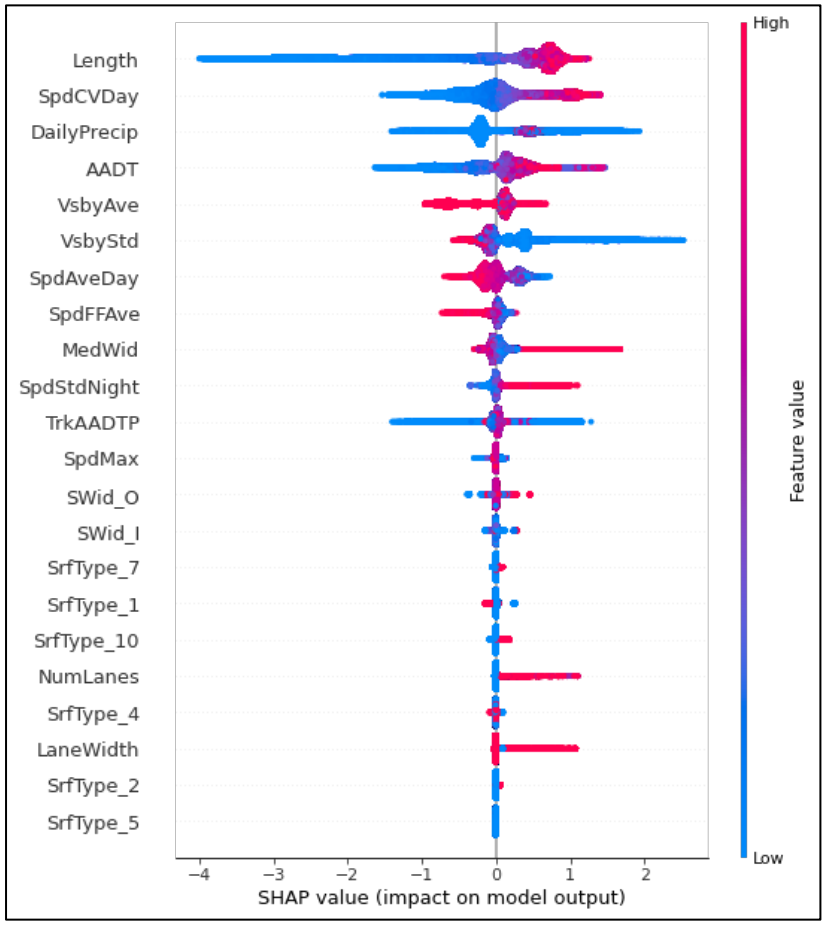


Figure 5. SHAP summary plot (Rural Interstate All Crash Model)

The most important feature identified by SHAP is segment length. This is no surprise because longer roadway segments have a higher chance of capturing daily crash occurrences. This study applies machine learning methods to address a binary classification problem (crash or non-crash) so normalizing crash occurrence numbers as crashes per mile does not make a difference. This is one of the limitations of this study and thus in future studies; it would be better to prepare all segments with equal length to address this problem. The second most important feature is the daily speed CV during daytime. This is reasonable because, as proved by previous studies (*5*)(*6*)(*7*), speed

variation is closely related to crash occurrence. The third most important feature is daily precipitation. A higher precipitation level has a positive impact on the model's outcome. This indicates that, as for daily crash occurrence, precipitation is a major factor and it is more important than roadway geometric features and most of the speed measurement features. Other top important features are AADT, Average Visibility, Standard Deviation of Visibility, and Daily speed average during daytime. The results indicate that weather condition variables, especially daily precipitation, have significant impacts on daily crash occurrence.

7.2. Severe Crash Model

In the previous section, the XGBoost model is trained to make binary classification on whether crashes happen on a particular roadway segment during a particular day. In this section, another XGBoost model is trained to make binary classification on whether severe crashes happen. For severe crashes, this study includes crashes that lead to death or severe injuries (K and A as defined in section 4). In Figure 6, most of the top important features are the same except for daily precipitation. In this model, daily precipitation ranks at 14th place. This means that daily precipitation makes only a little contribution to the model's output.

Figure 7 (b) presents the details of daily precipitation's impact on the model's output. There is no obvious pattern that can be seen in this plot. At different daily precipitation levels, the impacts on the model's output seem to be evenly distributed along the y axis. The results reveal that although higher daily precipitation levels make a significant impact on the occurrence of total daily crashes, in terms of the occurrence of

severe crashes, the level of daily precipitation has little impact. This is because when people drive during rainy weather, they tend to drive more carefully. Even though crashes are more likely to happen during rainy weather, it does not necessarily cause severe crashes. However, visibility (both average and standard deviation) still seems to play an important role in distinguishing the occurrence of severe crashes.

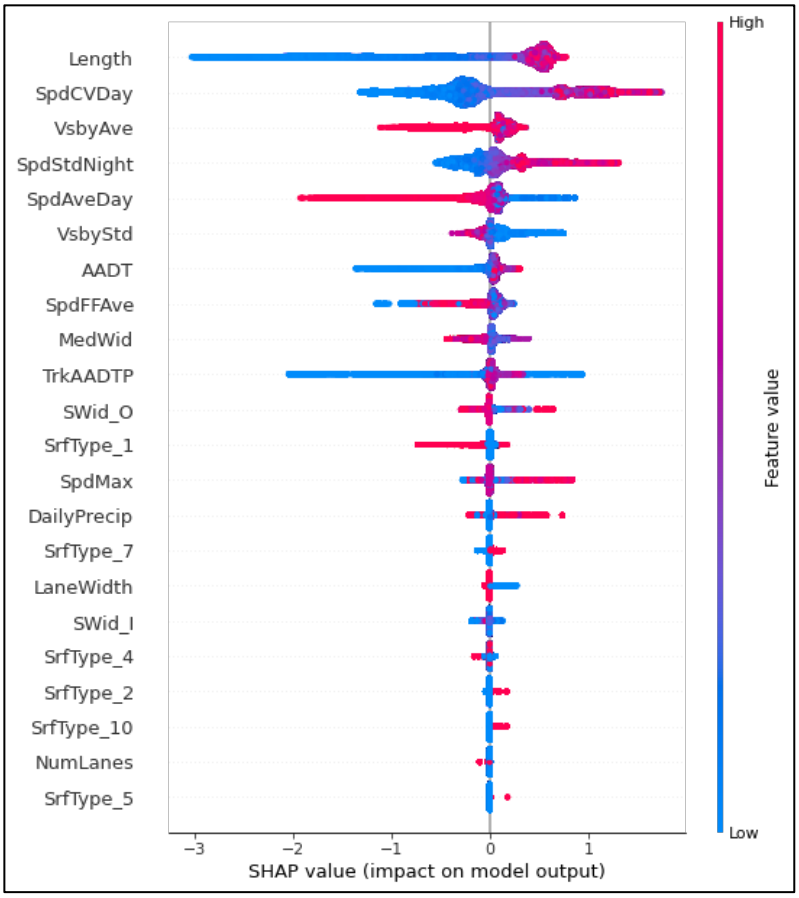
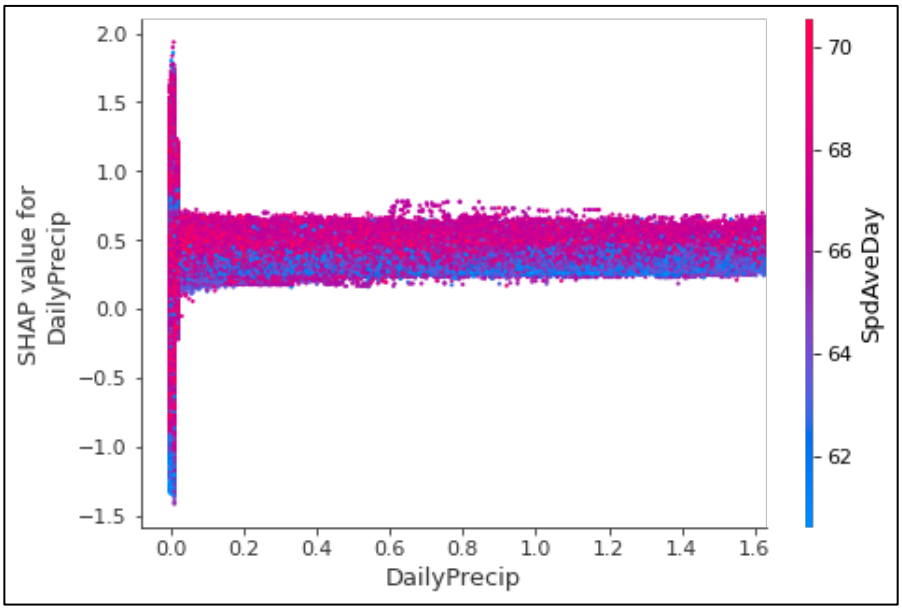


Figure 6. SHAP summary plot (Rural Interstate Severe Crash Model)

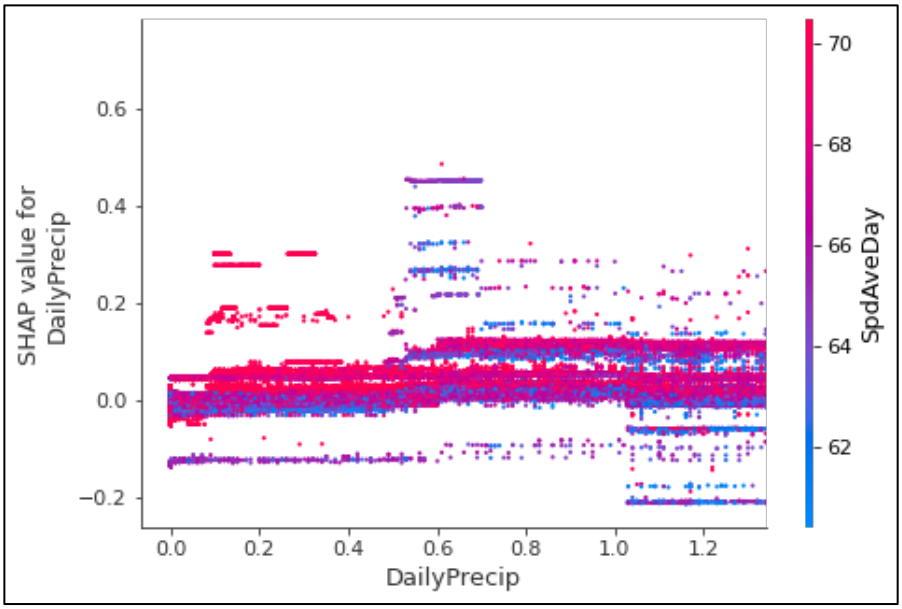
7.3. SHAP Dependence Plot

Below are the dependence plots between several key variables to show the variables' collaborative effects on crash occurrence probability. Figure 7 (a) presents the

SHAP dependency plot between daily precipitation and daytime average speed of the all crash model. When the daily precipitation level is near 0, it makes little contribution to distinguishing between crash observation and non-crash observation because crashes happened on the non-raining day as well. However, as the value of daily precipitation increases, the impact of this explanatory variable tends to stay positive. This indicates that larger precipitation tends to cause daily crash occurrences. As for average speed during the daytime, it is obvious in Figure 7 (a) that when the precipitation level is high, larger daytime average speeds are more likely to cause daily crash occurrence. This is different from the general contribution of this explanatory variable to the model output which can be found in Figure 5- that larger daytime average speed tends to decrease all crash occurrence probability.



(a) Rural Interstate All Crash Model



(b) Rural Interstate Severe Crash Model

Figure 7. Dependence plot between daily precipitation and daytime average speed

Figure 8 is the dependency plot between average visibility and visibility standard deviation of the all crash model. Similar to daily precipitation, when the value of visibility is near 10 (the maximum value), it makes little contribution to distinguishing between crash observation and non-crash observation because there are crashes that happened on clear days as well. When average visibility starts to decrease, it tends to have positive impacts on the model's output. It is noteworthy that on the left side of Figure 8, lower visibility standard deviation tends to make a positive contribution to the model's output when average visibility is low. This is because if average visibility is low and the standard deviation is also low, the adverse visibility condition barely changes throughout the day which is a hazardous condition for drivers.

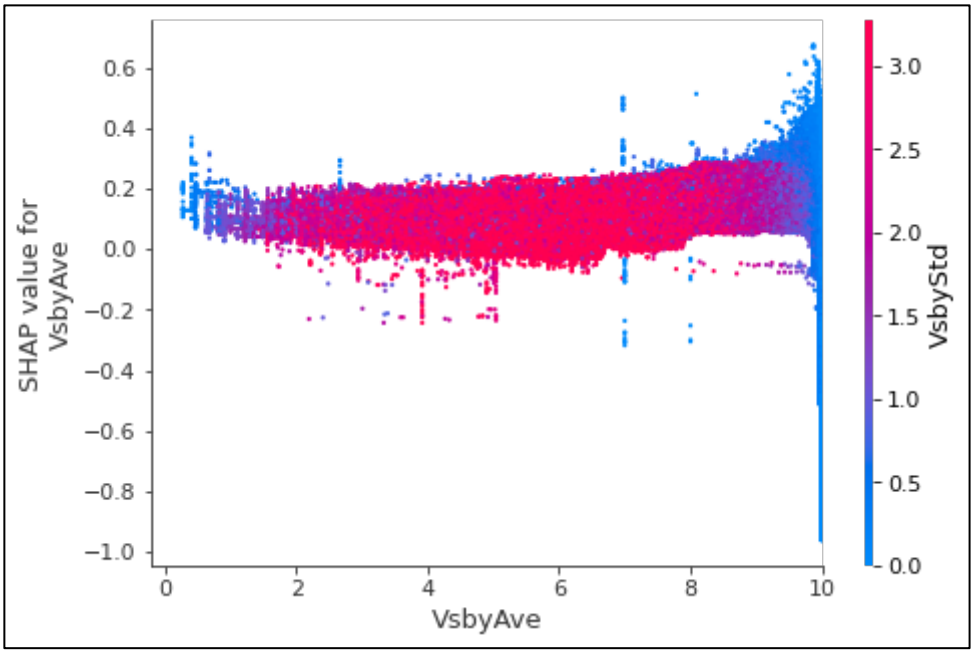


Figure 8. Dependence plot between average visibility and visibility standard deviation (All Crash Model)

Figure 9 presents the collaborative impact of speed CV and daily precipitation on the probability of all crash occurrences. Larger daytime speed CV values tend to push the model's output toward positive. Interestingly, the red dots in the dependency plot indicate that when daily precipitation levels are high, larger daytime speed CV values are more likely to cause daily crash occurrence. This shows that the effect of speed CV on daily crash occurrence becomes more significant under rainy weather conditions.

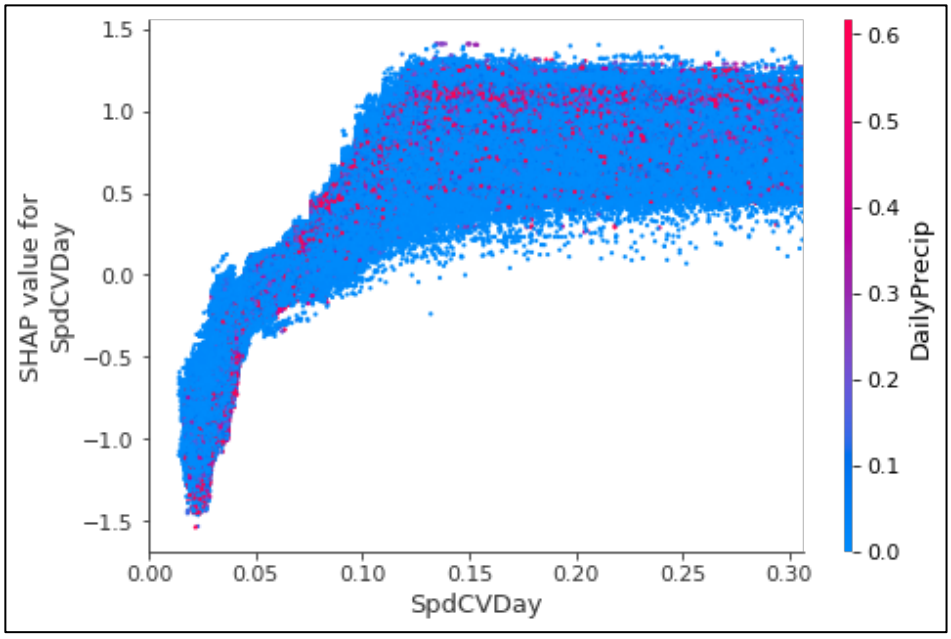


Figure 9. Dependence plot between daytime speed CV and daily precipitation (All Crash Model)

7.4. Findings on Rural Interstate Roadway

- Weather factors (precipitation and visibility) and speed variation are the main influential factors of roadway daily all crash occurrence.
- Daily precipitation is highly ranked in the all-crash occurrence model. However, its rank falls significantly in the severe crash occurrence model. This indicates that higher precipitation level is more likely to cause roadway crashes while it does not necessarily lead to extremely severe crashes.
- Generally, higher daytime average speed tends to decrease crash occurrence probability. However, with higher daily precipitation levels, higher daytime average speed is more likely to cause crash occurrences.

- When daily visibility is low, and it keeps low throughout the day, it is likely to lead to crash occurrences.
- Night-time speed standard deviation is a strong contributor to extremely severe crash occurrences on rural interstate roadways. Higher night-time speed standard deviation is more likely to cause extremely severe crashes.

8. RESULT - RURAL TWO-LANE

The same feature selection process is performed on the rural two-lane dataset. 9 highly correlated variables are removed from the dataset. (Spd85, RefSpd, SpdCV, SpdStd, SpdStdDay, SpdAveNight, SpdStdNight, SpdFFAve, SpdAveDay). Two XGBoost models are trained on the all crash occurrence dataset and the severe crash dataset. Table 7 presents the confusion matrixes of these two models.

Table 7. Confusion matrix (Rural Two-Lane)

		All Crash Model		Severe Crash Model	
Predicted Labels		**Non-Crash**	**Crash**	**Non-Crash**	**Crash**
True Labels	**Non-Crash**	1,447,028	413,749	1,590,621	272,352
	Crash	841	1,548	93	100

Performance measures for all crash model:

(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 77.7\%$

(2) $Sensitivity = \frac{TP}{TP+FN} = 64.8\%$

(3) $Specificity = \frac{TN}{TN+FP} = 77.8\%$

(4) $Weighted\ Accuracy = \frac{Sensitivity+Specificity}{2} = 71.3\%$

Performance measures for severe crash model:

(5) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 85.4\%$

(6) $Sensitivity = \frac{TP}{TP+FN} = 51.8\%$

(7) $Specificity = \frac{TN}{TN+FP} = 85.4\%$

(8) $Weighted\ Accuracy = \frac{Sensitivity+Specificity}{2} = 68.6\%$

Where: TP: True positive; TN: True Negative; FP: False Positive; FN: False Negative

8.1. All Crash Model

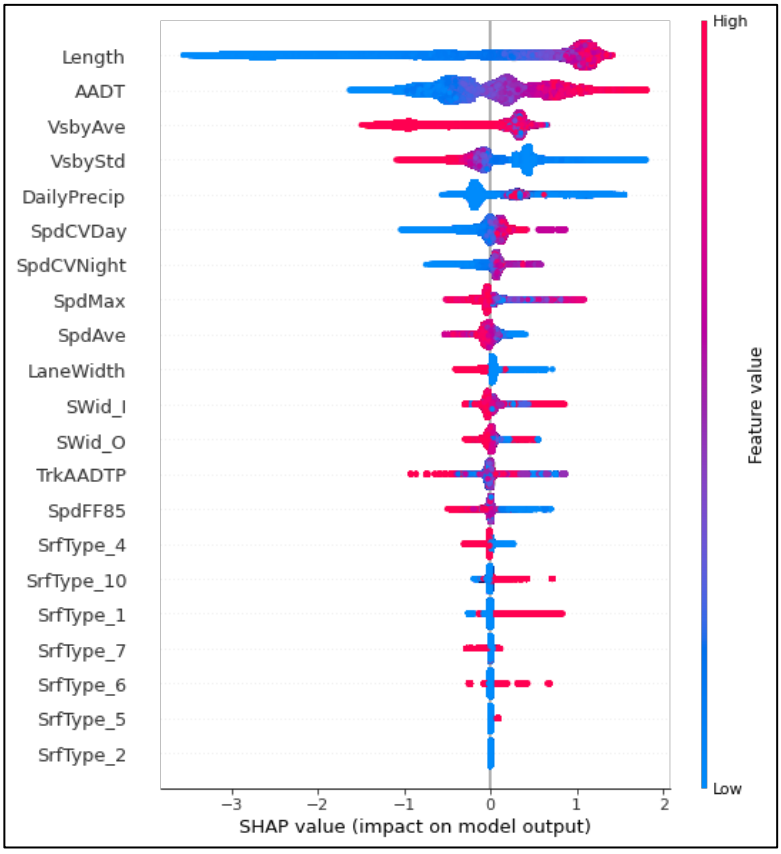


Figure 10. SHAP summary plot (Rural Two-Lane All Crash Model)

For rural two-lane roadways, AADT becomes the most important variable besides roadway segment length, followed by average daily visibility, visibility standard deviation, and daily precipitation. Higher AADT increases the probability of crash occurrence. Visibility is more important than precipitation on rural two-lane roadways.

This is different from rural Interstate roadways. Moreover, the importance of speed variation also decreases on rural two-lane roadways. Additionally, the importance of lane width significantly increases on rural two-lane roadways. Narrower lane width increases the probability of crash occurrences.

8.2. Severe Crash Model

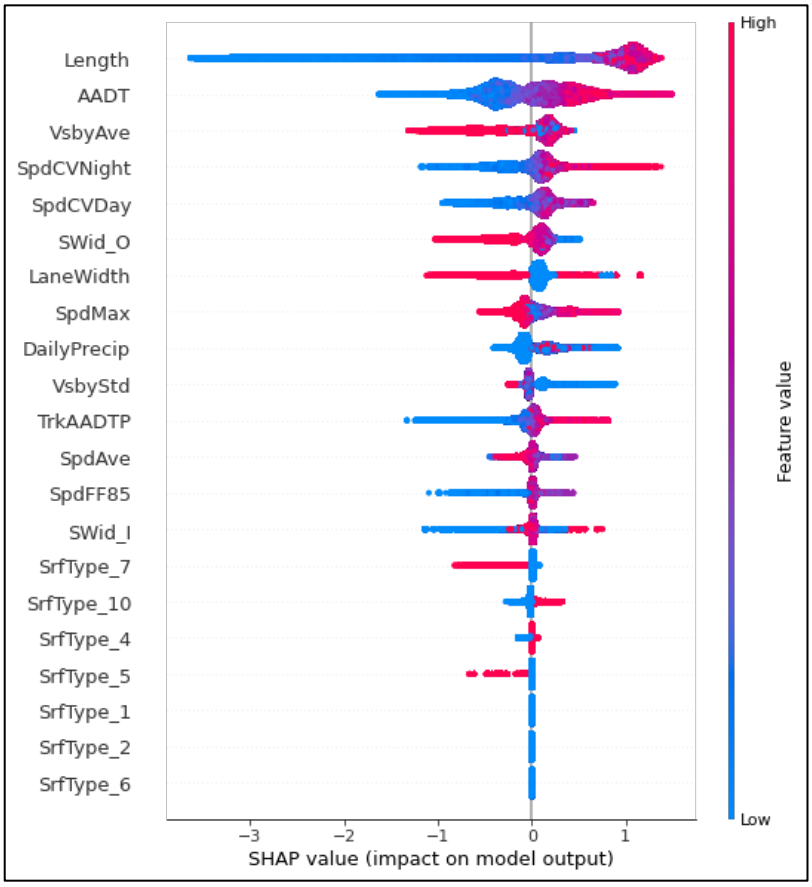
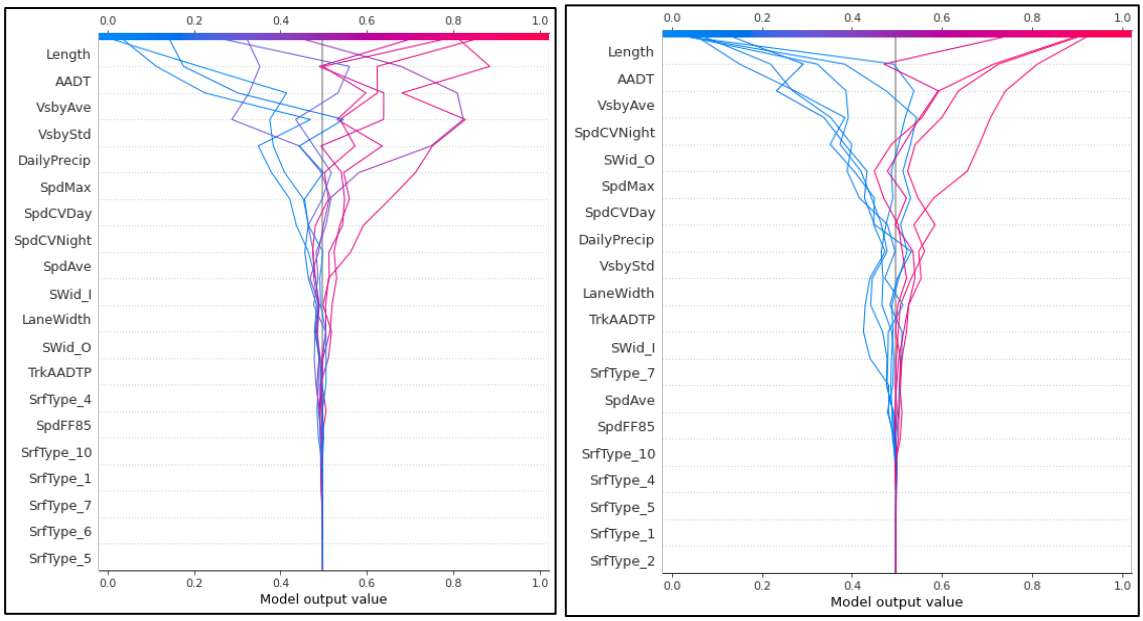


Figure 11. SHAP summary plot (Rural Two-Lane Severe Crash Model)

In the severe crash model for rural two-lane roadways, the contribution of speed variation becomes more important. On rural two-lane roadways, nighttime speed CV contributes more than daytime speed CV on severe crash occurrence. Similar to rural

Interstate roadways, the contribution of daily precipitation also decreases. However, the contribution pattern of daily precipitation is still observable on rural two-lane roadways. Higher precipitation still somehow tends to increase severe crash occurrence likelihood. Another noteworthy point is that the contribution of lane width significantly increases in the severe crash model. For truck AADT percentage, in the all crash occurrence model, the effect of this variable on the model's output is not clear. While in the severe crash model, it is obvious that higher truck AADT percentage increases the likelihood of severe crash occurrence. This indicates that truck AADT percentage is an important contributor to the probability of severe crash occurrence on rural two-lane roadways. Another explanatory variable that becomes more important in the severe crash model is outside shoulder width (ranked at 6th place). Its importance increases compared with the all crash occurrence model. Lower outside shoulder width increased the probability of severe crash occurrence on rural two-lane roadways.

Figure 12 is the SHAP decision plots of rural two-lane roadways all crash and severe crash models. The SHAP decision plot helps to visualize how the model reaches its decision based on all explanatory variables. For both decision plots, 10 observations are randomly selected and plotted. Red lines indicate the probability of crash occurrence is higher and blue lines indicate the probability of crash occurrence is lower.



(a) All Crash Model (b) Severe Crash Model

Figure 12. SHAP Decision Plot

8.3. Findings on Rural Two-Lane Roadway

- AADT is the most important contributing factor except segment length in both all crash and severe crash models of rural two-lane roadways. Higher AADT significantly increases crash occurrence probability.
- On rural two-lane roadways, nighttime speed variation makes more contributions than daytime speed variation on severe crash occurrences. This pattern cannot be seen on rural multilane roadways (introduced in the next section).
- On rural two-lane roadways, the importance of lane width is more significant compared with rural interstate roadways. Especially for the severe crash model, lower lane width increases the probability of severe crash occurrence.

- The importance of outside shoulder width also increases in the severe crash model. Narrower shoulder width tends to increase severe crash probability.
- On rural two-lane roadways, higher truck percentage tends to increase severe crash occurrence probability while this pattern is not obvious on the all crash model.

9. RESULT - RURAL MULTILANE

For the rural multilane dataset, 10 highly correlated variables are removed from the dataset including SpdCVNight; SpdStdDay; SpdFFAve; Spd85; RefSpd; SpdStd; SpdAveNight; SpdCV; SpdFF85; and SpdAve. Two XGBoost models are trained on the all crash occurrence dataset and the severe crash dataset. Table 8 presents the confusion matrixes of these two models.

Table 8. Confusion matrix (Rural Multilane)

		All Crash Model		Severe Crash Model	
Predicted Labels		**Non-Crash**	**Crash**	**Non-Crash**	**Crash**
True Labels	**Non-Crash**	1,629,323	539,110	1,888,631	285,743
	Crash	2,174	4,251	237	247

Performance measures for all crash model:

(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 75.1\%$

(2) $Sensitivity = \frac{TP}{TP+FN} = 66.2\%$

(3) $Specificity = \frac{TN}{TN+FP} = 75.1\%$

(4) $Weighted\ Accuracy = \frac{Sensitivity+Specificity}{2} = 70.7\%$

Performance measures for severe crash model:

(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = 86.9\%$

(2) $Sensitivity = \frac{TP}{TP+FN} = 51.0\%$

(3) $Specificity = \frac{TN}{TN+FP} = 86.9\%$

(4) $Weighted\ Accuracy = \frac{Sensitivity+Specificity}{2} = 69.0\%$

Where: TP: True positive; TN: True Negative; FP: False Positive; FN: False Negative

9.1. All Crash Model

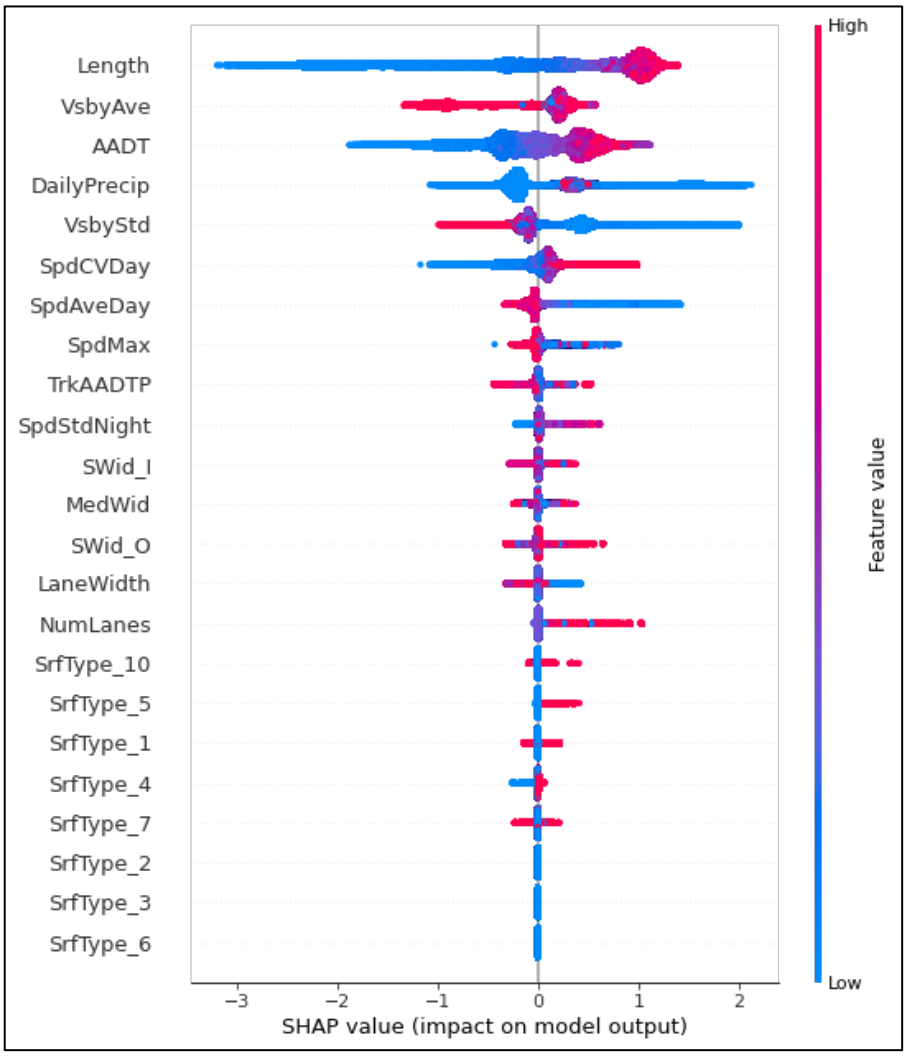


Figure 13. SHAP summary plot (Rural Multilane All Crash Model)

Figure 13 presents the SHAP summary plot of the rural multilane all crash occurrence model. In this model, as for rural multilane roadways, average daily visibility becomes the most important variable besides segment length to distinguish crash and non-crash observation, followed by AADT, daily precipitation, visibility standard deviation, and daytime speed CV. On both rural multilane roadway and rural two-lane roadways, average daily visibility tends to have a more critical effect on all crash occurrences than daily precipitation. Similar to rural two-lane roadways, the importance of roadway geometric variables becomes more important compared with those in the rural interstate model. For rural two-lane roadway, the importance of lane width is higher than the importance of shoulder width, whereas for rural multilane roadway, the importance of lane width is lower than that of shoulder width.

9.2. Severe Crash Model

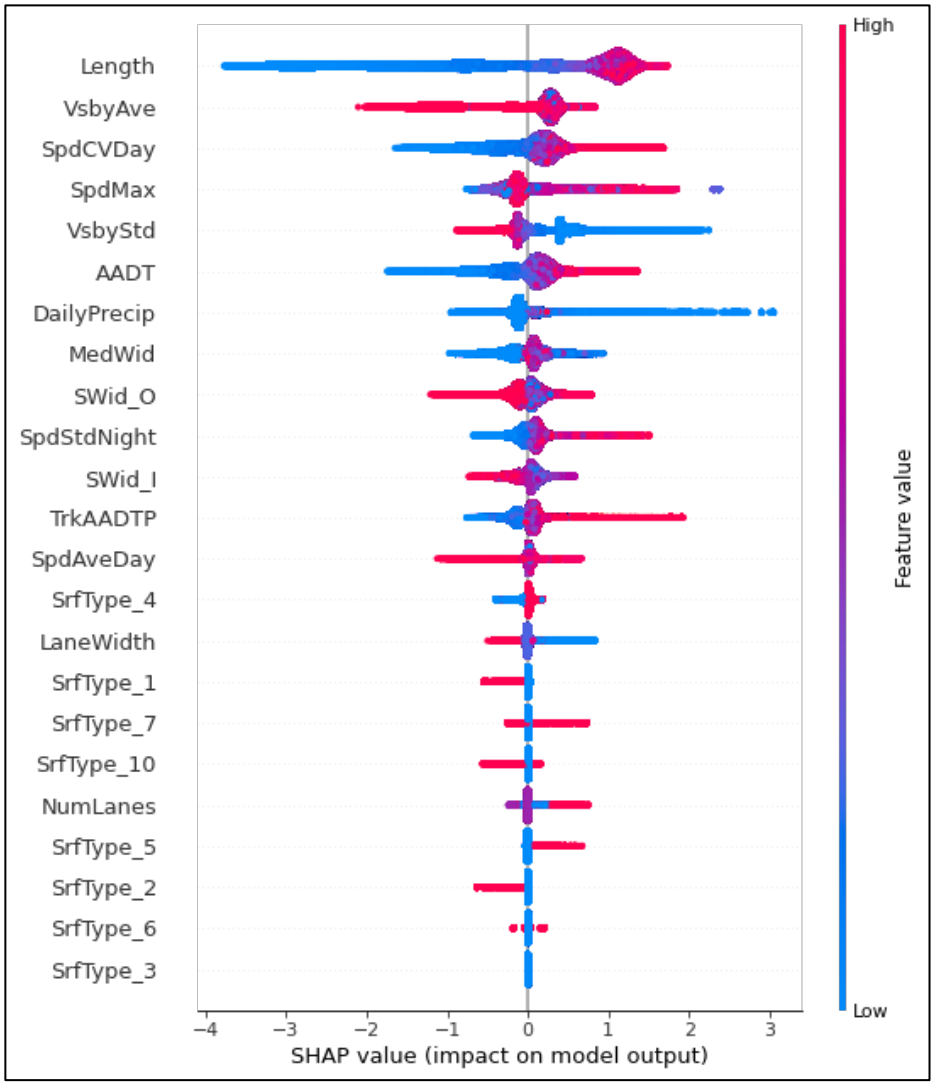


Figure 14. SHAP summary plot (Rural Multilane Severe Crash Model)

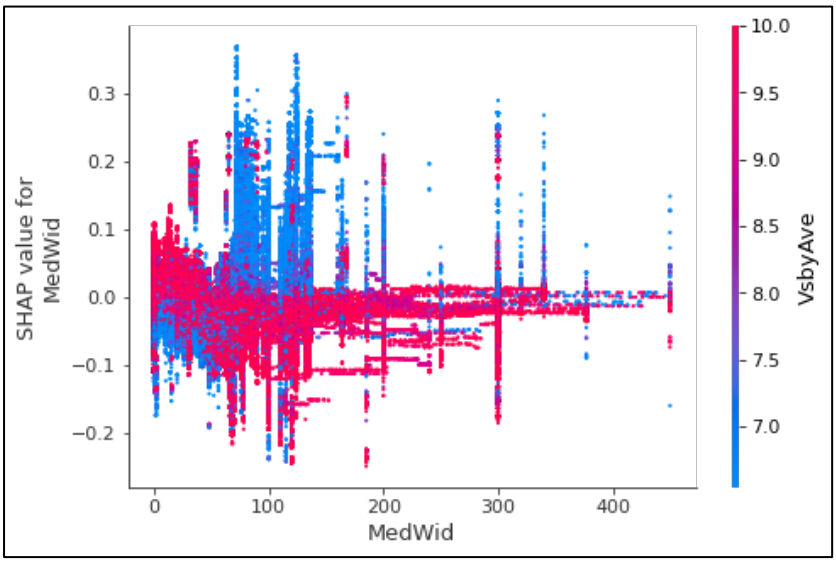
Figure 14 presents the rural multilane severe crash model, in which the importance rank of daytime speed CV increases to 3rd place. The importance rank of the maximum speed limit significantly increases to 4th place with higher maximum speed limit being more likely to cause severe crash occurrence. As for median width, its importance rank also increases in the severe crash model and it seems that smaller median width tends to

increase the probability of severe crash occurrence. Moreover, similar to rural two-lane roadways, outside shoulder width also gain more importance in the rural multilane severe crash model. Higher outside shoulder width tends to decrease the likelihood of severe crash occurrence and the impact of outside shoulder width is more significant on rural two-lane roadways than rural multilane roadways.

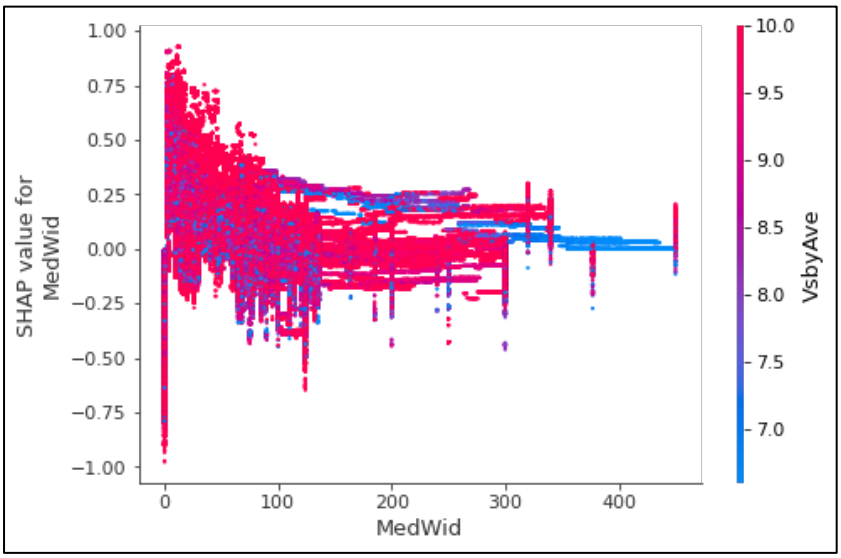
9.3. SHAP Dependence Plot

Figure 15 presents the dependence plots between median width and average visibility. For all crash occurrences, lower visibility significantly increases all crash occurrence probability. This is likely because lower visibility condition affects drivers' abilities to observe the traffic condition in the opposite direction. The exception is that when the median width is very small, lower visibility tends to decrease all crash occurrence probability. However, as for the severe crash model, larger median width significantly decreases severe crash occurrence probability, and when the median width is larger, lower visibility tends to increase the severe crash likelihood. When the median width is lower, average visibility levels do not have an obvious contribution to the model's output. This pattern indicates that lower average visibility may increase severe crash occurrence likelihood when median width is larger, while visibility level does not affect severe crash occurrence likelihood when the median is relatively narrower. This is probably because during bad visibility conditions, drivers may have difficulty observing the traffic condition of the opposite direction and thus drivers will tend to drive more carefully if the median is narrow. In this scenario, even if crashes happen, the severity will usually not be too high. However, if the median becomes wider, drivers will tend to be

less vigilant even when the visibility condition is not ideal. Thus, the crash severities are likely to be high under this situation.

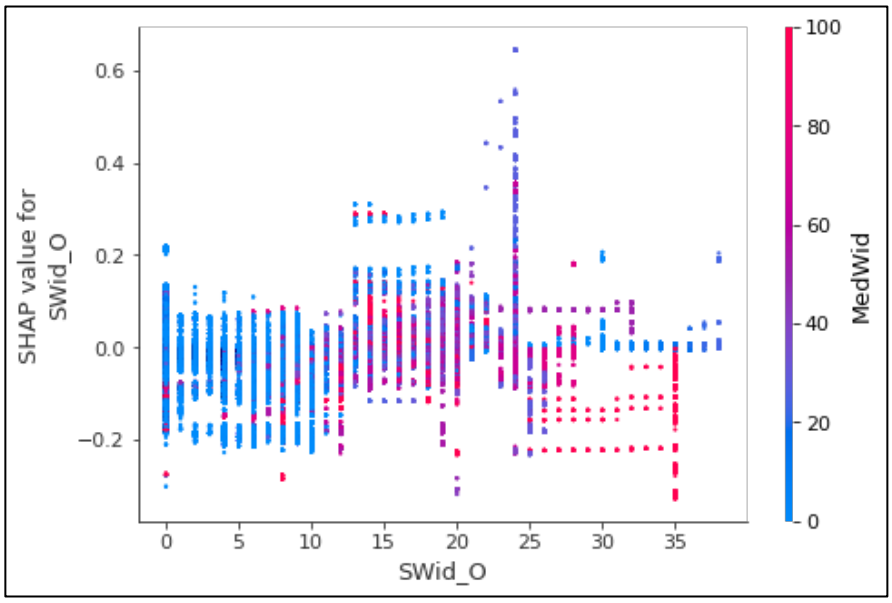


(a) Rural Multilane All Crash Model

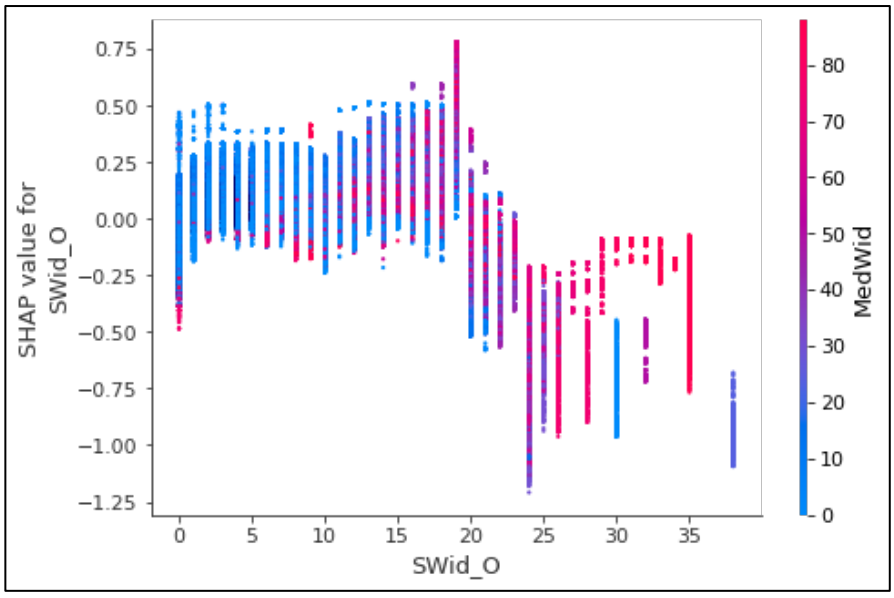


(b) Rural Multilane Severe Crash Model

Figure 15. Dependence plot between median width and average visibility



(a) Rural Multilane All Crash Model

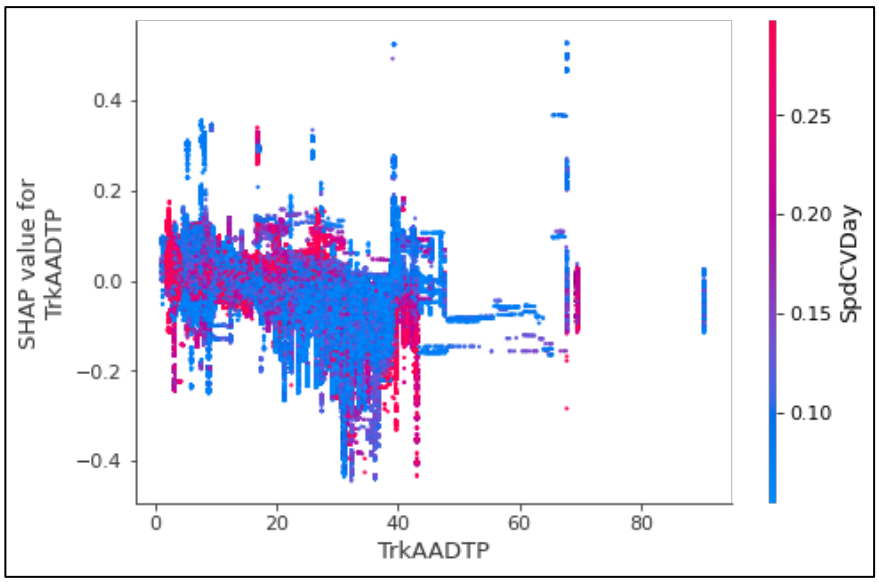


(b) Rural Multilane Severe Crash Model

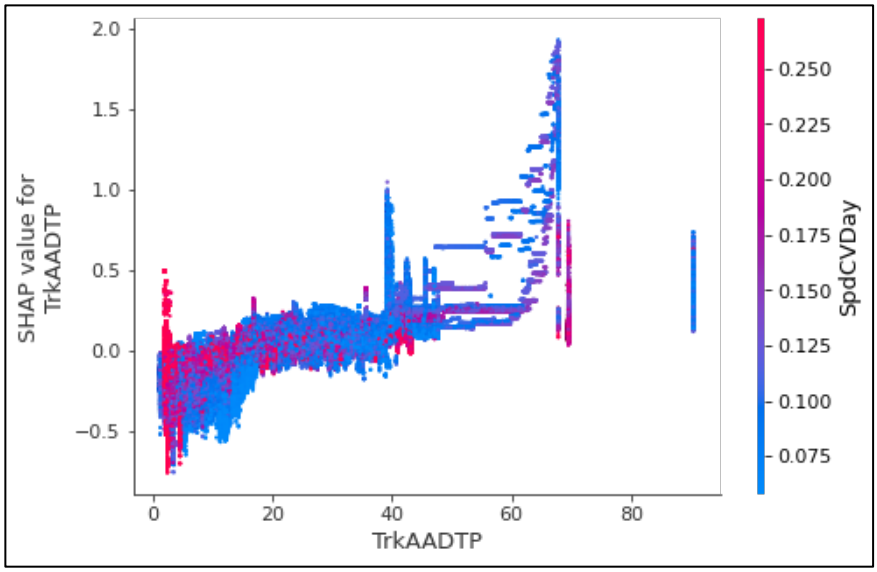
Figure 16. Dependence plot between outside shoulder width and median width

As the importance ranks of outside shoulder width and median width increase in the severe crash model, Figure 16 presents the dependence plot between these two

variables. The comparison between Figure 16 (a) and Figure 16 (b) indicates that the change in outside shoulder width does not have an obvious effect on all crash occurrences. However, for severe crash occurrences, when the outside shoulder width is higher, the probability of severe crash occurrence decreases significantly. Moreover, when outside shoulder width is narrower, larger median width tends to decrease the likelihood of severe crash occurrence, whereas when the outside shoulder width becomes larger, the effect of median width becomes less obvious.



(a) Rural Multilane All Crash Model



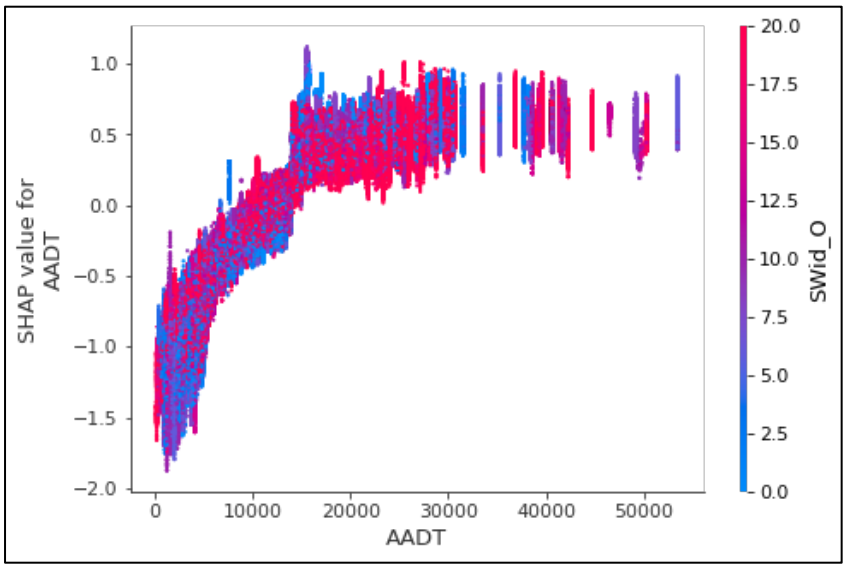
(b) Rural Multilane Severe Crash Model

Figure 17. Dependence plot between truck AADT percentage and daytime speed CV

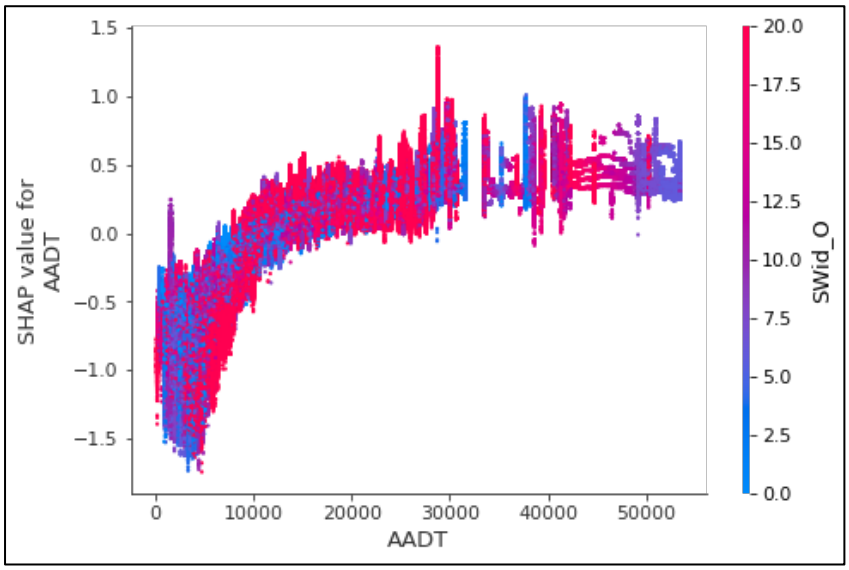
Similar to rural two-lane roadways, compared with the all crash model, truck AADT percentage has a more significant effect on severe crash occurrence probability. Figure 17 presents the dependence plot between truck AADT percentage and daytime speed CV. Figure 17 (a) shows no obvious effect of truck AADT percentage on all crash occurrence probability, while in Figure 17 (b) it is obvious that higher truck AADT percentage increases severe crash occurrence probability, especially when truck percentage is greater than 60% and when truck AADT percentage is lower than around 20%. Higher speed CV tends to have a positive effect on the likelihood of severe crash occurrence, but this pattern can't be seen when truck AADT percentage is higher.

Figure 18 is the dependence plot between AADT and outside shoulder width. For both all crash and severe crash models, higher AADT significantly increases crash occurrence probability. However, for the severe crash model, when the AADT level is

low, larger outside shoulder width may decrease the probability of severe crash occurrence according to Figure 18 (b).



(a) Rural multilane all crash model



(b) Rural Multilane Severe Crash Model

Figure 18. Dependence plot between AADT and outside shoulder width

9.4. Findings on Rural Multilane Roadway

- Daily average visibility is the most important contributing factor to crash occurrences on rural multilane roadways besides segment length. It is more important than daily precipitation.
- On rural multilane roadways, the contribution of median width is more obvious in the severe crash model compared with that in the all crash model. In general, lower median width tends to increase severe crash occurrence probability. When the median is wide, lower visibility is more likely to cause severe crash occurrences. However, when the median is narrow, lower visibility seems to slightly decrease severe crash occurrence probability.
- On rural multilane roadways, outside shoulder width makes a strong contribution to the probability of severe crash occurrence. When the outside shoulder width becomes wider, the probability of severe crash occurrence significantly decreases. Moreover, when the outside shoulder is narrow, increasing the median width may decrease the probability of severe crash occurrence. For the all crash model, the effect of outside shoulder width is not observable.
- Rural multilane roadways with larger truck percentages are more likely to observe severe crash occurrence, especially when the truck percentage is greater than 60%- the likelihood of severe crash occurrence is skyrocketing. Moreover, when the truck percentage is relatively low, larger speed CV can make severe crash occurrence probability increase, whereas when the truck percentage is relatively

larger, the effect of speed CV becomes less obvious. For the all crash occurrence model, the effect of truck percentage is not observable.

- AADT impacts both all crash and severe crash occurrence probabilities on rural multilane roadways. This effect is more obvious on the all crash occurrence model. For the severe crash model, when AADT is at a relatively lower level (less than 10,000), increasing outside shoulder width can decrease severe crash occurrence probability.

10. CONCLUSION

This study analyzes the impacts of roadway geometry, speed distribution, and weather condition on roadway daily crash occurrence with different severity levels by using a machine learning method named XGBoost (eXtreme Gradient Boosting). The dataset consists of information from four sources: (1) RHiNO 2018 (Roadway geometric and traffic information); (2) NPMRDS (Speed distribution information); (3) ASOS (Weather condition information); and (4) CRIS (Crash information). In this study, roadway segments are classified into three different facility types (Rural Interstate, Rural Two-Lane, and Rural Multilane) and to study the variables contribution on crash severity level, for each facility type, one dataset is prepared for all crash occurrence and another is prepared for severe crash occurrence. To address the rareness nature of crash observations, the study applied an oversampling method called SMOTE to oversample crash observations, and this study also applied a feature selection process to remove highly correlated variables in the dataset.

This study successfully analyzes the contribution patterns of geometric, speed distribution, and weather conditions on daily crash occurrence. Moreover, how these contribution patterns vary across different roadway facility types and crash severity levels is also investigated. Based on the findings of this study, it can be concluded that for rural interstate roadways, precipitation and speed variation play an important role in determining all crash occurrences. For severe crash occurrences, speed distribution is still very important whereas precipitation does not contribute too much. On rural interstate roadways, the contribution of roadway geometric variables is less significant for both all

crash and severe crash models. As for rural two-lane roadways, AADT plays a significant role in all crash and severe crash occurrences. Average visibility and visibility standard deviation make more contributions to crash occurrence than daily precipitation. The importance of some roadway geometric variables increases; for example, the impact of lane width significantly increases, especially in the severe crash model, and outside shoulder width also has an obvious contribution in the severe crash model. The result from rural multilane roadways offers a novel insight. Average visibility is the most important contributing variable besides segment length for both all crash and severe crash occurrences. The contribution of median width is apparent on rural multilane roadways. For all crash occurrences, low visibility increases all crash occurrence probability expect for when median width is very small, wherein low visibility tends to decrease all crash occurrence probability. For severe crash occurrences, only when median width is at a higher level, lower visibility can increase severe crash occurrence probability. Higher truck percentage increases severe crash occurrence probability on rural multilane roadways, especially when the truck percentage is greater than 60%. However, the truck percentage level seems does not affect all crash occurrence probability too much.

REFERENCE

1. Washington, S., M. G. Karlaftis, F. Mannering, P. Anastasopoulos, M. G. Karlaftis, F. Mannering, and P. Anastasopoulos. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC, 2020.
2. Shankar, V., F. Mannering, and W. Barfield. Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis & Prevention*, Vol. 27, No. 3, 1995, pp. 371–389.
3. Das, S., A. Dutta, and X. Sun. Patterns of Rainy Weather Crashes: Applying Rules Mining. *Journal of Transportation Safety & Security*, Vol. 12, No. 9, 2020, pp. 1083–1105.
4. Theofilatos, A., and G. Yannis. A Review of the Effect of Traffic and Weather Characteristics on Road Safety. *Accident Analysis & Prevention*, Vol. 72, 2014, pp. 244–256.
5. Choudhary, P., M. Imprialou, N. R. Velaga, and A. Choudhary. Impacts of Speed Variations on Freeway Crashes by Severity and Vehicle Type. *Accident Analysis & Prevention*, Vol. 121, 2018, pp. 213–222.
6. Quddus, M. Exploring the Relationship Between Average Speed, Speed Variation, and Accident Rates Using Spatial Statistical Models and GIS. *Journal of Transportation Safety & Security*, Vol. 5, No. 1, 2013, pp. 27–45.
7. Wang, X., Q. Zhou, M. Quddus, T. Fan, and S. Fang. Speed, Speed Variation and Crash Relationships for Urban Arterials. *Accident Analysis & Prevention*, Vol. 113, 2018, pp. 236–243.
8. Lord, D., and B. N. Persaud. Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. *Transportation Research Record*, Vol. 1717, No. 1, 2000, pp. 102–108.
9. Mountain, L., M. Maher, and B. Fawaz. The Influence of Trend on Estimates of Accidents at Junctions. *Accident Analysis & Prevention*, Vol. 30, No. 5, 1998, pp. 641–649.
10. Dutta, N., and M. D. Fontaine. Improving Freeway Segment Crash Prediction Models by Including Disaggregate Speed Data from Different Sources. *Accident Analysis & Prevention*, Vol. 132, 2019, p. 105253.
11. Miaou, S.-P., and H. Lum. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis & Prevention*, Vol. 25, No. 6, 1993, pp. 689–709.
12. Anderson, I. B., K. M. Bauer, D. W. Harwood, and K. Fitzpatrick. Relationship to Safety of Geometric Design Consistency Measures for Rural Two-Lane Highways. *Transportation Research Record*, Vol. 1658, No. 1, 1999, pp. 43–51.
13. Haghighi, N., X. C. Liu, G. Zhang, and R. J. Porter. Impact of Roadway Geometric Features on Crash Severity on Rural Two-Lane Highways. *Accident Analysis & Prevention*, Vol. 111, 2018, pp. 34–42.
14. Garber, N. J., and R. Gadiraju. Factors Affecting Speed Variance and Its Influence on Accidents. *Transportation research record*, Vol. 1213, 1989, pp. 64–71.

15. Lee, C. K., F. Saccomanno, B. Hellinga. Analysis of Crash Precursors on Instrumented Freeways. *Transportation Research Record*, No. 1784, 2002, pp. 1–8.
16. Pei, X., S. C. Wong, and N. N. Sze. The Roles of Exposure and Speed in Road Safety Analysis. *Accident Analysis & Prevention*, Vol. 48, 2012, p. pp 464-471.
17. Scott, P. P. Modelling Time-Series of British Road Accident Data. *Accident Analysis & Prevention*, Vol. 18, No. 2, 1986, pp. 109–117.
18. Eisenberg, D. The Mixed Effects of Precipitation on Traffic Crashes. *Accident Analysis & Prevention*, Vol. 36, No. 4, 2004, pp. 637–647.
19. Brijs, T., D. Karlis, and G. Wets. Studying the Effect of Weather Conditions on Daily Crash Counts Using a Discrete Time-Series Model. *Accident Analysis & Prevention*, Vol. 40, No. 3, 2008, pp. 1180–1190.
20. Jaroszweski, D., and T. McNamara. The Influence of Rainfall on Road Accidents in Urban Areas: A Weather Radar Approach. *Travel Behaviour and Society*, Vol. 1, No. 1, 2014, p. pp 15-21.
21. Yu, R., and M. Abdel-Aty. Analyzing Crash Injury Severity for a Mountainous Freeway Incorporating Real-Time Traffic and Weather Data. *Safety Science*, Vol. 63, 2014, p. pp 50-56.
22. Federal Highway Administration. The National Performance Management Research Data Set (NPMRDS) and Application for Work Zone Performance Measurement. https://ops.fhwa.dot.gov/publications/fhwahop20028/index.htm. Accessed Feb. 23, 2021.
23. Abdel-Aty, M., N. Uddin, A. Pande, M. F. Abdalla, and L. Hsia. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. *Transportation Research Record*, Vol. 1897, No. 1, 2004, pp. 88–95.
24. Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, Vol. 16, 2002, pp. 321–357.
25. Li, P., M. Abdel-Aty, and J. Yuan. Real-Time Crash Risk Prediction on Arterials Based on LSTM-CNN. *Accident Analysis & Prevention*, Vol. 135, 2020, p. 105371.
26. Yuan, J., M. Abdel-Aty, Y. Gong, and Q. Cai. Real-Time Crash Risk Prediction Using Long Short-Term Memory Recurrent Neural Network. *Transportation Research Record*, Vol. 2673, No. 4, 2019, pp. 314–326.
27. Parsa, A. B., H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian. Real-Time Accident Detection: Coping with Imbalanced Data. *Accident Analysis & Prevention*, Vol. 129, 2019, pp. 202–210.
28. Chen, T., and C. Guestrin. XGBoost: A Scalable Tree Boosting System. New York, NY, USA, 2016.