THE USE OF MACHINE LEARNING METHOD FOR MODELING AND ANALYZING PEDESTRIAN CRASH DATA AND COMPARISONS WITH TRADITIONAL DISCRETE CHOICE MODELING METHODS

by

Yang Li

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2020

Approved by:

_____

Dr. Wei Fan

_____

Dr. Martin Kane

_____

Dr. David Weggel

_____

Dr. Jay Wu

_____

Dr. Jing Yang

ABSTRACT

YANG LI. The use of machine learning method for modeling and analyzing pedestrian crash data and comparisons with traditional discrete choice modeling methods.
(Under the direction of DR. WEI FAN)

As one of the most vulnerable entity within the transportation system, pedestrians might face more dangers and sustain severer injuries in the traffic crashes than others. The safety of pedestrians is particularly critical within the context of continuous traffic safety improvements in US. Moreover, traffic crash data are inherently heterogeneous, and such data heterogeneity can cause one to draw incorrect conclusions in many ways. Therefore, developments and applications of proper modeling approaches are needed to identify causes of pedestrian-vehicle crashes to better ensure the safety of pedestrians.

On the other hand, with the development of artificial intelligence techniques, a variety of novel machine learning methods have been established. Compared to conventional discrete choice models (DCMs), machine learning models are more flexible with no or few prior assumptions about input variables and have higher adaptability to process outliers, missing and noisy data. Furthermore, the crash data has inherent patterns related to both space and time, crashes happened in locations with highly aggregated uptrend patterns should be worth exploring to examine the most recently deteriorative factors affecting the pedestrian injury severities in crashes.

The major goal of this dissertation is intended to build a framework for modeling and analyzing pedestrian injury severities in single-pedestrian-single-vehicle crashes with providing a higher resolution on identification of contributing factors and their associating effects on the injury severities of pedestrians, particularly on those most recently

deteriorative factors. Developments of both conventional DCMs and the selected machine learning model, i.e., XGBoost model, are established. Detailed comparisons among all developed models are conducted with a result showing that XGBoost model outperforms all other conventional DCMs in all selected measurements. In addition, an emerging hotspot analysis is further utilized to identify the most targeted hotspots, followed by a proposed XGBoost model that analyzes the most recently deteriorative factors affecting the pedestrian injury severities. By completions of all abovementioned tasks, the gaps between theory and practice could be bridged. Summary and conclusions of the whole research are provided, and further research directions are given at the end.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincerest gratitude to my advisor, Professor Wei Fan, whose kindest support was invaluable in shaping my Ph.D. research. It was his insightful guidance sharpening my thoughts and brought my work to such a high level.

I would like to give appreciations to the rest of my thesis committee: Professor Martin Kane, Professor David Weggel, Professor Jay Wu, and Professor Jing Yang. Without those thought-provoking questions and advisable comments from them, this research could hardly be accomplished solely by myself.

I would also like to thank my colleagues in Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE) lab. There is no doubt that my progresses are tightly related to their kindest help in both my life and work during the past three and a half years. In addition, I am always very grateful to Professor Burak Eksioglu, who has offered his generous help when I was in Clemson University and during my transition to UNCC.

Indeed, I shall have my biggest thanks and respects to my parents, my father Xuming Li and my mother Xiucheng Yang, for all their selfless support the whole time in my life, otherwise all my successes and achievements would be in vain. Same to my grandparents, who always held strong believes in their grandson.

And for my dearest fiancée, Jueminsi Wu, thanks for all her support, without which I would have stopped these studies a long time ago. All unconditional encouragements and care received from her have lighted me up through all these days. What more can I say? Love conquers all!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AADT	annual average daily traffic
AdaBoost	adaptive boosting
ANN	artificial neural network
DCM	discrete choice model
HSIS	Highway Safety Information System
IIA	independence of irrelevant alternatives
LightGBM	light gradient boosted machine
MAPRT	multiple additive Poisson regression trees
MNL	multinomial logit
ML	mixed logit
NCDOT	North Carolina Department of Transportation
NHTSA	National Highway Traffic Safety Administration
PO	proportional odds
PPO	partial proportional odds
SVM	support vector machine
XGBoost	extreme gradient boosting

CHAPTER 1: INTRODUCTION

1.1. Problem Statement and Motivation

Compared to other entities in the transportation system, pedestrians are among the most vulnerable ones. The injuries and deaths of pedestrians in traffic crashes causes huge impacts both socially and economically. Such issue is particularly critical within the context of continuous traffic safety improvements in US. According to the National Highway Traffic Safety Administration (NHTSA), compared to other entities, pedestrians are the most vulnerable on American roadways, and on average, a pedestrian was killed every 88 minutes in traffic crashes in 2017 (NHTSA, 2017). In recent reports from Governors Highway Safety Association (Retting and Schwartz 2019, 2020), there is a 53% increase in pedestrian fatalities in 2018 compared to 2009. Figure 1.1 shows the number of pedestrian fatalities from 2007 to 2018 along with its percentages in the total traffic fatalities in US (NHTSA, 2018).

Figure 1.1 shows an increasing trend of the pedestrian death in US after the Great Recession of 2009 not only in the number but also the percentage compared to other traffic fatalities (NHTSA, 2018). On the other hand, based on the available data from North Carolina Department of Transportation (NCDOT), there are more than 2000 pedestrians that are involved in crashes with vehicles each year during the past decades in North Carolina (NC). On average, a total of 150–200 pedestrians are annually killed on NC roads, and additional 200–300 pedestrians are severely injured (Thomas and Levitt, 2018). With that being said, the safety and risk issues of pedestrians within the transportation system cannot be neglected. As a result, a myriad of researches have been done to explore the factors to the abovementioned issues, such as the alcohol involvement, demographic

features, at-fault operations, collision types, environmental characteristics, roadway and locality features, time-of-day (Kim et al., 2008a, 2010; Kim et al., 2008b; Dai, 2012; Chen and Fan, 2019a,b; Li and Fan, 2019a,b; Mokhtarimousavi, 2019).

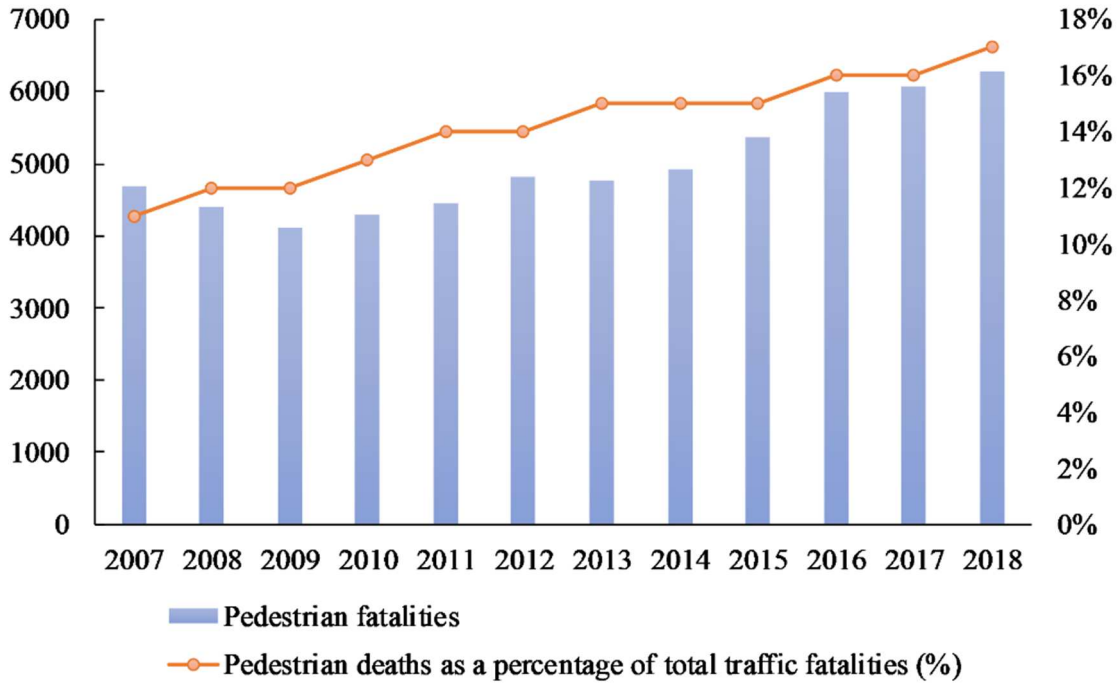

FIGURE 1.1: Number of pedestrian fatalities and its percentages in the total traffic fatalities in US (NHTSA, 2018)

Moreover, due to the heterogeneity inherent in the traffic crash data, which arises from unobservable factors that are not reported by law enforcement agencies and cannot be collected from state crash records, it is not easy to identify and evaluate factors that significantly affect injury severity of pedestrians in such crashes. Such heterogeneity might result in biased estimation of parameters and thus drawing potentially incorrect conclusions (Mannering and Bhat, 2014; Shaheed and Gkritza, 2014).

Generally speaking, conventional widely used discrete choice modeling (DCM) methods such as multinomial logit models (MNL), ordered logit/probit models, mixed logit models (ML), and partial proportional odds logit models (PPO) have been applied to analyze the crash data. However, almost all the beforementioned methods highly rely on prior assumptions. Compared with those statistical models, machine learning models (Tang et al., 2018) are more flexible with no or few prior assumptions for input variables and have higher adaptability to process outliers, missing and noisy data. Furthermore, machine learning techniques are also good examples of data driven methods which aim to increase efficiency and accuracy of the analysis and prediction of crash data. Recently, different machine learning approaches such as neural network, ensemble learning, and support machines have been employed by the researchers and their results indicate that such approaches are highly adaptable and can give better performances than traditional models. Therefore, the machine learning-based approach is selected for the analysis of pedestrian involved crash data in this study.

Furthermore, the crash data inherent has patterns in both space and time, and crashes happened in locations with highly aggregated uptrend patterns should be worth exploring to examine the most recently deteriorative factors that contribute to severer injuries (i.e., fatalities and incapacitating injuries) of pedestrian in the pedestrian-vehicle crashes. With such consideration, the emerging hotspot analysis tool developed by the ArcGIS could provide solid references to help identify the spatiotemporal patterns of the crash related data. Hence, by taking advantages of both the emerging hotspot analysis and the selected machine learning technique, a more targeted model is further developed to analyze the pedestrian injury severity in this study.

1.2. Study Objectives

The proposed work in this research is intended to fulfill the following objectives:

1. To select and develop traditional DCMs for modeling pedestrian-injury severities based on previous studies falling into the field of transportation crash data analysis, particularly those focusing on pedestrian involved crashes;
2. To model pedestrian-injury severities in pedestrian-vehicle crashes using advanced, appropriate and accuracy machine learning-based approach;
3. To use real-world police-reported pedestrian crash data to examine and validate the developed models so that the gaps between the theoretical research and the application of the developed pedestrian injury severity model can be bridged;
4. To compare the results between traditional DCMs and advanced machine learning-based approach and provide conclusions and recommendations;
5. To provide a framework for modeling pedestrian injury severities by combining emerging hotspots analysis and the selected machine learning method.

1.3. Expected Contributions

In order to better improve the environment for pedestrians, many research studies have been conducted; however, efforts are still needed to establish proper and accurate methods in data analysis and modeling to identify contributing factors affecting injury severities of pedestrian in pedestrian-vehicle crashes. This would highly guide traffic and safety engineers to make appropriate measures for creating a safer environment for pedestrians. Expected contributions of this research work to the state-of-the-art and state-of-the-practice include the following:

1. Ability to select and develop traditional DCMs for modeling pedestrian-injury severities and analyzing the data for improving the safety of pedestrians;
2. Ability to develop more advanced, appropriate and accurate pedestrian injury severity models;
3. Ability to accurately analyze real world police-reported crash data and identify key contributing factors to pedestrian injury severity by using the developed prediction models.
4. Ability to conduct comparisons between different modeling methods and provide recommendations on selecting appropriate approach to modeling pedestrian-injury severity and on countermeasures for improving pedestrian safety;
5. Ability to analyze the spatiotemporal patterns of pedestrian crashes and apply the corresponding machine learning technique to develop the best model to provide up-to-date analysis of pedestrian injury severities in pedestrian-vehicle crashes.

1.4. Research Overview

The research will be structured as shown in Figure 1.2. In Chapter 1, the significance and motivation of modeling pedestrian-injury severity in pedestrian-vehicle crashes has been discussed, followed by the description of study objectives and expected contributions.

Chapter 2 presents a comprehensive literature review of the current state-of-the-art and state-of-the-practice of modeling pedestrian injury severities in pedestrian-vehicle crashes, including both conventional DCMs and novel machine learning models. Then methodologies of the selected widely used conventional DCMs to model and analyze

pedestrian-injury severity in pedestrian-vehicle crashes are recognized in the dissertation as: (1) basic DCMs (i.e., MNL model); (2) advanced DCMs (i.e., ML model and PPO model). In addition, the selected machine learning-based methodology (i.e., XGBoost model) is introduced and summarized in this chapter as well.

Chapter 3 describes the basic information on police-reported pedestrian crash data collected between 2007 and 2018 in North Carolina utilized in this study. A wide range categorical factors with motorist, pedestrian, environmental, and roadway features of the dataset are inspected. The data processing steps are also described in this chapter.

Chapter 4 presents the applications of the selected widely used conventional DCMs (i.e., MNL, PPO, and ML models) for modeling pedestrian-injury severity in pedestrian-vehicle crashes. Firstly, the developments of each model are explained in detail, followed by the exhibits of results on both parameter estimations and marginal effects on each developed model. Then, the general guidance of interpreting the model results (i.e., mainly based on the results of marginal effects) of conventional DCMs is provided in this chapter.

Chapter 5 introduces the development of the selected advanced machine learning method, which is the XGBoost model for modeling pedestrian-injury severity in pedestrian-vehicle crashes. The detailed process of the selected method is described with parameter tuning and model training. Additionally, partial dependence of the contributing factor in machine learning methods introduced in Chapter 2 is calculated with the similar utility as the marginal effect of conventional DCMs. A general guidance on the interpretation of partial dependence is also presented for the purpose of demonstration of the result analysis for the developed machine learning model.

Chapter 6 provides an analysis of comparison between selected conventional DCMs and advanced machine learning method on modeling and analyzing pedestrian-injury severity in pedestrian-vehicle crashes. Evaluations on all models in this study are also provided with several most widely used criteria such as accuracy, precision (i.e., positive predictive value), recall (i.e., sensitivity), and F1 score. The discussions on the comparison results are provided for finding the most appropriate, advanced and accurate model on modeling and analyzing pedestrian-injury severity in pedestrian-vehicle crashes.

Chapter 7 illustrates a framework for modeling pedestrian injury severities in pedestrian-vehicle crashes by combining the emerging hotspot analysis and machine learning method. Associated results are also presented for exploring the contributing factors affecting the pedestrian injury severities with uptrend in the hotspots in North Carolina to further provide recommendation references to policymakers for improving the safety of pedestrian.

Chapter 8 concludes the whole study with a summary of the developed models, solution approaches, and research results. Suggestions for future research are provided.

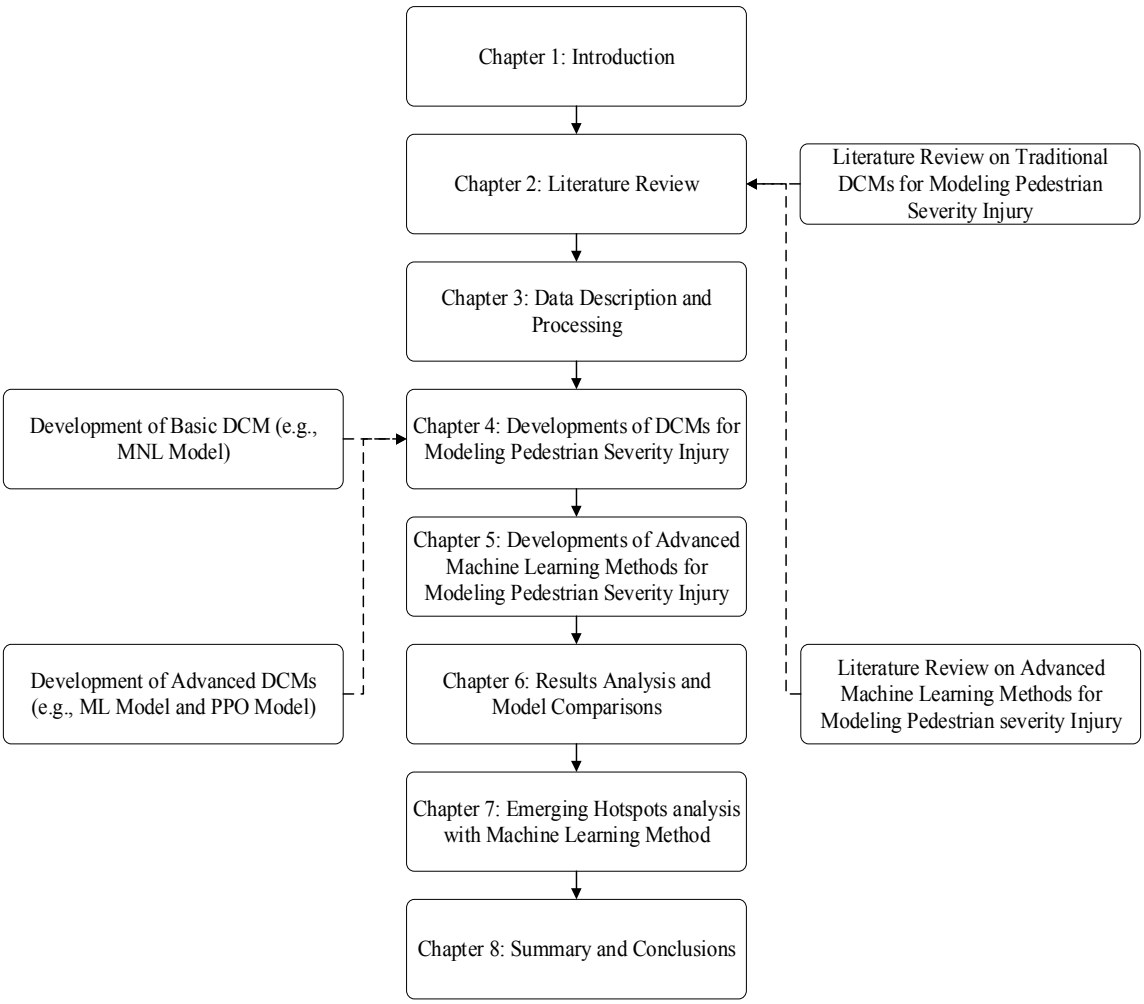

FIGURE 1.2: Research Structure

CHAPTER 2: LITERATURE REVIEW

2.1. Introduction

This chapter provides a comprehensive review of various aspects related to traffic crash data analysis, particularly those majorly focusing on pedestrian involved crashes. Contents include literatures on each approach and the associated introduction of methodology. This should give a clear picture of existing current efforts made towards the modeling of pedestrian injury severities in pedestrian-vehicle crashes.

The following sections are organized as follows. Section 2.2 presents the literatures on basic DCMs, followed by the presentation of methodology for the MNL model, which is the most popular DCM model used in modeling pedestrian-injury severity in pedestrian-vehicle crashes. Section 2.3 gives a comprehensive review of existing methods of advanced DCMs, and also the introduction of two widely used models, which are the ML model and the PPO model. Section 2.4 will show the literatures with the use of machine learning methods in crash data analysis and the associated methodology of the selected method (i.e., XGBoost method). Finally, section 2.5 concludes this chapter with a summary.

2.2. Basic Discrete Choice Models

2.2.1. Introduction of Basic Discrete Choice Models

Among all traditional DCMs, multinomial logit (MNL) models have been widely and popularly used for analyzing the crash injury severity relevant data. Many researchers had demonstrated that MNL models could be successfully applied to identify contributing factors for most traffic crash types, including pedestrian-vehicle crashes (Himanen and

Kulmala 1988; Shankar and Mannering 1996; Quddus et al., 2009; Zhou et al., 2013). However, MNL model is highly dependent on the assumption of independence of irrelevant alternatives (IIA) property. IIA considers the same effects of the independent variables across individual cases, which would be invalid when there are unobserved data heterogeneities. It is true because of the incompleteness of the data and it also implies various effects of the estimators across different observations. In addition to the MNL model, several other discrete choice models were also used by many scholars to model the severity of pedestrian-injury in pedestrian-vehicle crashes, including the binary logit model (Sze and Wong, 2007; Moudon et al., 2011; Sasidharan et al., 2015) and ordered logit/probit model (Zajac and Ivan, 2003; Yasmin et al., 2014; Chen et al., 2016).

However, most of the beforementioned models ignored the unobserved heterogeneity across individual injury observations, especially when the police-reported data were utilized in which only a limited number of the explanatory variables were included and analyzed. According to some studies (McFadden and Train, 2000; Train, 2009; Mannering et al., 2016), such ignorance may lead to biased analysis results.

2.2.2. Literature Review on Using Basic Discrete Choice Models to Model Pedestrian Crashes

2.2.2.1. Work of Zajac and John (2003)

Zajac and John (2003) applied an ordered probit model to investigated factors that have impacts on injury severities of motor vehicle-crossing pedestrian crashes in rural area of Connecticut. Key factors, such as clear roadway width, vehicle type, alcohol

involvement on both driver and pedestrian, and elder pedestrian (age ≥65), increase the chance of fatal injury for pedestrians. Based on the results of such key factors, policy related suggestions were also made for practical purpose.

2.2.2.2. Work of Rifaat and Chin (2007)

To better understand the relationship between injury severities and risk factors for mitigating severity of pedestrian injury in Singapore, Rifaat and Chin (2007) applied an ordered probit model to examine the factors affecting crash severities with broad considerations of driver characteristics, roadway features, vehicle types, pedestrian characteristics and crash characteristics and also explored how the interaction of these factors affect the injury severities. Data used in this study were divided into three groups, of which one portion with only pedestrian involved crashes was dedicated for modeling pedestrian injury severity. Factors of elder pedestrian and nighttime were identified to raise the risk of pedestrians being severely injured.

2.2.2.3. Work of Sze and Wong (2007)

Sze and Wong (2007) used historical crash data in Hong Kong to determine the risk of pedestrian being injured and killed in traffic crashes and to investigate the contributing factors to injury severity. The authors proposed a binary logit model and found that elder pedestrians (age ≥65), head injury, crash at crossing or within 15m of crosswalk, speed limit (≥50 mph), a signalized intersection, and number of lanes (≥2) obviously resulted in a higher risk of being killed and severely injured for pedestrians.

2.2.2.4. Work of Kim et al. (2008a)

Kim et al. (2008a) utilized police-reported crash data between 1997 and 2000 of North Carolina to explore the pedestrian injury severities in pedestrian-vehicle crashes. Two models including MNL and a heteroskedastic model were used. Factors, such as PM traffic peak, traffic signal control, driver age, curved roadways, inclement weather, crosswalk, and walking along roadway were identified to raise the chance of a fatal crash. Other than MNL model, the results showed that the heteroskedastic model performs better than the MNL model.

2.2.2.5. Work of Ulfarsson et al. (2010)

Ulfarsson et al. (2010) developed an MNL model to explore the assignment of fault in pedestrian-vehicle crashes for the purpose of improving the safety of pedestrians in the transportation system. In this study, different scenarios were selected, where observed factors are associated with pedestrian at fault, driver at fault, or both at fault. Results of this study showed the followings: 1) Pedestrian-at-fault factors: pedestrian crossing streets, pedestrian dash/dart, pedestrian (age ≤12), drunk pedestrian; 2) Driver-at-fault factors: turning/merging/backing up movement, speeding, driver backing up, drunk driver, and multiple pedestrians; and 3) Other important factor: darkness.

2.2.2.6. Work of Tay et al. (2011)

A study by Tay et al. (2011) was conducted to explored factors that have impacts on the injury severity of pedestrian in pedestrian-vehicle crashes in South Korea. The authors applied an MNL model, and several contributing factors were found to have impacts on pedestrian injury severity, which belong to traffic control methods, roadway characteristics, weathers, pedestrian characteristics, driver characteristics, and vehicle types. Several key factors, such as heavy vehicles, drunk drivers, male drivers, drivers with

age ≤65, elder pedestrians (age ≤65) or female pedestrians, midblock, high speed limit, inclement weather conditions, night, etc., were found to severely increase the risk of each severity level of injury to pedestrians. Related recommendations were also made to hold campaigns targeted at male drivers, drivers under the age of 65, female pedestrians and very old pedestrians for improving the safety of pedestrians.

2.2.2.7. Work of Kwigizile et al. (2013)

Kwigizile et al. (2013) examined the inconsistencies between ordered and unordered probability models for pedestrian injury severity. Two models (i.e., ordered probit model and MNL model) were developed using data on crashes between a single vehicle and a pedestrian recorded in Florida from 2004 to 2008. The results of the comparison analysis indicated that the effects of contributing factors are consistent on levels of the lowest and highest injury, but inconsistent for some factors on intermediate injury levels. Thus, cautions should be given when conducting the model selections.

2.2.2.8. Work of Obeng and Rokonuzzaman (2013)

Obeng and Rokonuzzaman (2013) deployed an ordered logit model to model injury severity for pedestrians injured from the pedestrian-vehicle crashes at signalized intersections in Greensboro, North Carolina. Through the results, the study showed that female drivers and presence of a sidewalk were identified as the contributing factors to increase the risk of pedestrian being severely injured. On the other hand, passenger cars, sport utility vehicles and pickups were found to contribute to minor injuries of pedestrians.

2.2.2.9. Work of Zhou et al. (2013)

Zhou et al. (2013) conducted a study on pedestrian crossing behaviors at signalized intersections in Nanjing, China. An MNL model with latent variables was developed to examine the impacts of contributing factors on pedestrians' behavior. The results showed that 1) arrival time, oncoming cars, and crosswalk length have the most effects on late starters; 2) for pedestrians with sneaking behavior, gender has the greatest impact; and 3) when pedestrians with partial sneaking behavior, age is the most significant contributing factor. The authors also provided recommendations on several aspects to improve the safety of pedestrians, including facility designs and safety educations.

2.2.2.10. Work of Yasmin et al. (2014)

An ordered logit model was proposed by Yasmin et al. (2014) to examine pedestrian injury severity in New York City. Two other enhanced ordered logit models (i.e., generalized ordered logit model and latent segmentation based ordered logit model) were also developed for an in-depth analysis. The key contributing factors affecting pedestrian injury severity levels were identified as weathers, lighting conditions, vehicle characteristics, pedestrian ages, and seasons. Elder pedestrian (age $\geq$ 65) was identified to raise the risk of pedestrian being killed in crash.

2.2.2.11. Work of Chen and Fan (2019a)

Chen and Fan (2019a) developed an MNL model to investigate the pedestrian-vehicle crash in North Carolina, by using the data obtained from Highway Safety Information System (HSIS) database from 2005 to 2012. From the results, factors that significantly increase the chance of severer injuries (i.e., fatal and incapacitating injuries) include: bad condition of driver, motorcycle and heavy truck, young and elder pedestrians (age between 26-65; $\geq$65), weekends, light condition (dawn, dusk and dark), curved

roadways, roadway surface with water, NC route, speed limit (35-50 mph; ≥50 mph). Relevant suggestions on improving the safety of pedestrian within the transportation system were also provided in this study.

Table 2.1 provides a summary of existing studies that utilized the basic DCM methods (i.e., binary logit model, ordered logit/probit model and MNL model) that majorly focus on pedestrian crash data analysis within transportation safety research in chronological order.

TABLE 2.1: Summary of Existing Studies Utilized the Basic DCM Methods Focusing on Pedestrian Crash Data Analysis

Authors	Year	Case study location	Methodology	Key findings
Zajac and John	2003	Rural Connecticut USA	Ordered probit model	Factors contributing to fatality of pedestrians: clear roadway width, vehicle type, alcohol involvement on both driver and pedestrian, and elder pedestrian (≥65).
Rifaat and Chin	2007	Singapore	Ordered probit model	Factors contributing to severe injury of pedestrians: elder pedestrian and nighttime.
Sze and Wong	2007	Hong Kong China	Binary logit model	Factors contributing to severe injury such as pedestrian age (≥65), head injury, pedestrian crossing, speed limit (≥31 mph), signalized intersection, and number of lanes (≥2).
Kim et al.	2008	North Carolina USA	MNL model	Factors contributing to fatality of pedestrians: PM traffic peak, traffic signal control, curved roadways, inclement weather, crosswalk, and walking along roadway.

Authors	Year	Case study location	Methodology	Key findings
Ulfarsson et al.	2010	North Carolina USA	MNL model	1) Pedestrian-at-fault factors: pedestrian crossing streets, pedestrian dash/dart, pedestrian age (≤12), drunk pedestrian; 2) Driver-at-fault factors: turning/merging/backing up movement, speeding, driver backing up, drunk driver, and multiple pedestrians. 3) Other important factor(s): darkness.
Tay et al.	2011	South Korea	MNL model	Factors contributing to severe injury of pedestrians: heavy vehicles, drunk drivers, driver gender (male), driver age (≤65), pedestrian age (≥65), pedestrian gender (female), pedestrians in roadway, high speed limit, inclement weather conditions, night, on road links, in tunnels, on bridges, on wider roads.
Obeng and Rokonuzzaman	2013	Greensboro, NC USA	Ordered logit model	Factors contributing to severe injury of crashes in signalized intersection: female drivers and presence of a sidewalk.
Kwigizile et al.	2011	Florida USA	Ordered probit model, MNL model	Comparisons on two model structures (ordered probit model vs. MNL model): effects of contributing factors are consistent on levels of lowest and highest injury levels, but inconsistent for some factors on intermediate injury levels.
Zhou et al.	2013	Nanjing China	MNL model	Crossing behaviors of pedestrian at signalized intersections were investigated and contributing factors for different behavior groups are identified: 1) late starters (arrival time, oncoming cars, and crosswalk length); 2) pedestrian with sneaking behavior (gender); and 3) pedestrian with partial sneaking behavior (age).

Authors	Year	Case study location	Methodology	Key findings
Yasmin et al.	2014	New York City USA	Ordered logit model	Factor contributing to fatality: elder pedestrian (≥65).
Chen and Fan	2019 a	North Carolina USA	MNL model	Factors contributing to fatalities and disabling injuries: impaired driver, motorcycle and heavy truck, pedestrians age (26-65; ≥65), weekends, light condition (dawn, dusk and dark), curved roadways, roadway surface with water, NC route, speed limit (35-50 mph; ≥50 mph).

Compared to all other basic DCMs, the MNL model is the most popular one and widely used by the majority of scholars whose interest is in safety research, due to the ease of its use in both model development and interpretation. Thus, the following subsection gives a brief introduction to the methodology on constructing the MNL model.

2.2.3. Multinomial Logit Model

The utility function U_{ij} of MNL model is a linear function, which denotes the relationship between injury severities ($j = 0, 1, 2,..., J$) and contributing factors, as presented in Equation 2.2.3.1:

$$U_{ij} = \beta_j X_{ij} + \varepsilon_{ij} \qquad (2.2.3.1)$$

where X_{ij} represents the vector of observable factors (variables) for ith individual with jth injury severity level, β_j denotes the vector of estimated coefficients, and ε_{ij} is the error component, which captures the unobserved factors and is assumed to be independently and identically distributed (i.e., independence of irrelevant alternatives, IIA property). If the error component follows the generalized extreme-value distribution, the MNL model could

be presented in Equation 2.2.3.2 (Manski and McFadden, 1981):

$$P_{ij} = \frac{exp(\alpha_j - X_{ij}^{'}\beta_j)}{\sum_{j \in J} exp(\alpha_j - X_{ij}^{'}\beta_j)} \quad (2.2.3.2)$$

where P_{ij} is the probability of *i*th pedestrian-vehicle crash with *j*th pedestrian injury severity level outcome.

The marginal effect analysis could help evaluate how the significant variables estimated in the MNL model impact the pedestrian injury outcome probabilities (Scott-Long, 1997). Since binary indicator variables (with the value of 0 or 1) are used in this study, the marginal effect can be computed as:

$$\frac{\partial P_{ij}}{\partial X_{ijk}} = P_{ij}\big(given\ X_{ijk} = 1\big) - P_{ij}\big(given\ X_{ijk} = 0\big) \quad (2.2.3.3)$$

Combined the direct interpretation of the coefficients with the marginal effects of the MNL model, marginal effect is more appropriate for use to explain the results to provide proper recommendations on improving the safety of pedestrians in the transportation system.

2.3. Advanced Discrete Choice Models

2.3.1. Introduction of Advanced Discrete Choice Models

As mentioned in Subsection 2.2.1, MNL model has the assumption of the same effects of independent variables across individual cases which could be violated if there are inherent unobserved data heterogeneities. It is true due to the incompleteness of the traffic accidents data, which implies that effects could vary across different cases. Therefore, in order to overcome the limitation caused by such IIA property, the ML model

is developed by setting the parameters to be randomly distributed across individual observations. In the transportation safety research domain, many researchers had applied the ML model for crash injury severity relevant data analysis (Milton et al., 2008; Malyshkina and Mannering 2010; Chen and Chen 2011; Yasmin and Eluru, 2013; Gong and Fan, 2017; Chen and Fan 2019a, b). By comparing results with the MNL model, they also found that ML model is more appropriate for dealing data with unobserved heterogeneities.

However, in crash injury severity analyses, despite the improvement of ML model that overcomes the limitation of MNL model, both models consider all injury severities as non-ordered. Hence, both models ignore the inherent hierarchical nature of injury severities. Meanwhile, data utilized in ordered logit/probit models needs to be strictly subjected to the proportional odds (PO)/parallel lines assumption. With that being said, ordered logit/probit models treats the parameter estimates the same and constant across severity levels (Savolainen et al., 2011). Such assumption would be unreasonable, which requires relaxation in the modeling of crash injury severity.

By considering those limitations of conventional discrete choice models, there is an emerging need for establishing and utilizing more elaborate models. Basically, such models need to consider the inherent ordered nature of the crash injury severity. Then, they should allow some of the parameter estimates to have different effects on different injury severity levels. Peterson and Harrell (1990) firstly proposed the partial proportional odds (PPO) model by relaxing the PO assumption. And by applying the model to several well-examined datasets, the model by allowing non-proportional odds for a subset of the independent variables has been proved its effectiveness. Since then, the PPO model has

shown its attractiveness to scholars and been successfully deployed to handle a variety of research problems (Wang and Abdel-Aty, 2008; Rifaat et al., 2012; Gong et al., 2016; Pour-Rouholamin and Jalayer, 2016; Pour et al., 2016; Li and Fan, 2019a,b, 2020). Sasidharan and Menéndez (2014) had done a literature review on applications and utilized PPO model to model the pedestrian crash injury severities using a dataset of national pedestrian safety from 2008 to 2012 in Switzerland. Compared with results from MNL model and ordered logit/probit model, the authors found PPO model performs better in modeling crash injury severity.

Therefore, with the consideration of the unobserved heterogeneities, the inherent hierarchical nature of crash data, and the popularity of models within the field of traffic crash data analysis, the ML model and the PPO model are selected in this study to represent the advanced DCMs and the their methodologies are also included in the following subsections.

2.3.2. Literature Review on Using Advanced Discrete Choice Models to Model Pedestrian Crashes

2.3.2.1. Work of Kim et al. (2010)

In order to address the unobserved heterogeneity within the crash data, Kim et al. (2010) applied an ML model to analyze pedestrian-injury severity in pedestrian-vehicle crashes in North Carolina. Several contributing factors were identified as significant factors with raising the risk of being killed for pedestrians in crashes, which are light condition of darkness without streetlights, truck freeway, speeding involvement, and drunk driver. It

was also found that heterogeneity for factors of freeway and pedestrian-at-fault collisions in the mean of the random parameters was highly associated with the gender of pedestrians, and heterogeneity for traffic control (sign) and backing vehicle in the mean of the random parameters was associated with the age of pedestrians.

2.3.2.2. Work of Aziz et al. (2013)

Aziz et al. (2013) developed an ML model to investigate pedestrian injury severity levels in New York City by accounting for unobserved heterogeneity in the population and across the boroughs. Several key factors, such as number of lanes, grade, light condition, road surface, presence of signal control, type of vehicle, parking facilities, commercial land use, and industrial land use were identified as significant factors in the developed model. Besides, the results of the loglikelihood ratio test indicated the necessity of developing segmented models for each borough.

2.3.2.3. Work of Islam and Jones (2014)

Islam and Jones (2014) deployed an ML model to examine the contributing factors affecting the injury severities of pedestrians at-fault crashes in both rural and urban locations in Alabama by considering unobserved heterogeneity across individuals. Through the results, obvious differences exist between the impacts of a set of variables on the injury severities of pedestrian of urban versus rural pedestrian at-fault crashes. Different statistically significant variable sets were identified in both locations (i.e., urban and rural). In addition, several variables with random effects were also detected by allowing unobserved heterogeneity across individuals.

2.3.2.4. Work of Haleem et al. (2015)

A study focused on analyzing pedestrian crash injury severity at signalized and non-signalized intersections in Florida was conducted by Haleem et al. (2015), in which an ML model was employed. The results showed that for signalized intersections, factors towards severe injury of pedestrian were identified, which are higher AADT, higher speed limit, and higher percentage of trucks, elder pedestrians, pedestrian-at-fault, rain, and darkness. And for unsignalized intersections, factors contributing to severe injury of pedestrian were also detected including higher speed limits, pedestrian walking along roadway, mid-age and elder pedestrians, pedestrian-at-fault, vans, and darkness.

2.3.2.5. Work of Tulu et al. (2017)

Tulu et al. (2017) used police-reported pedestrian crashes in Addis Ababa, Canada from 2009 to 2012 and applied an ML model with accounting for the unobserved heterogeneity in the crash data that were potentially not reported by law enforcement agencies and/or could not be collected from crash records to explore contributing factors affecting the pedestrian injury severities in pedestrian-vehicle crashes. Results revealed several factors having negative impacts towards severe and fatal injuries of pedestrians, including high speed limit, intersections, darkness, and heavy vehicle. Moreover, drivers with less education were more likely contributing to fatal injury to pedestrians in the crashes.

2.3.2.6. Work of Kim and Ulfarsson (2019)

Kim and Ulfarsson (2018) used an ML model to examine a wide range of variables of driver, vehicle type, vehicle movement, location, and environment for pedestrian crashes. A comparison study was also conducted to explore differences between older adult pedestrians (age ≥65) and younger adult pedestrians (age between18-59) by using the

2012-2013 crash data retrieved from U.S. National Automotive Sampling System (NASS) General Estimate Systems (GES) database. The study identified several factors contributing to severe injury of elder pedestrians, such as pedestrian crossing roadway, left/right turning movement of driver in parking areas, minivans, and SUVs.

2.3.2.7. Work of Chen and Fan (2019b)

Build upon the previous study (Chen and Fan, 2019a), Chen and Fan (2019b) deployed an ML model to investigate and identify significant contributing factors affecting the pedestrian injury severities in pedestrian-vehicle crashes by segmenting data into rural and urban areas in North Carolina, United States. The results show that impaired driver, heavy trucks, darkness, speed limit (35-50 mph; ≥50 mph) were detected as statistically significant factors towards severe and fatal injury severities in both rural and urban areas. Differences were also identified for some variables with different effects in urban and rural areas. The findings indicate that in order to better understand the pedestrian crashes, it is necessary to model separate models with segmentation of data as urban and rural areas.

Unlike ML model, only few research works have applied PPO models to analyze crash injury severity and it is also true for those studies mainly focusing on pedestrian injury severity analysis.

2.3.2.8. Work of Rifaat et al. (2012)

A study by Rifaat et al. (2012) was conducted to explore the effects of different urban street patterns on pedestrian injury severity in pedestrian-vehicle crashes. Police-reported pedestrian crash data collected by the City of Calgary from 2003-2005 were utilized to develop the PPO model. The results of the model implied that compared to the

traditional gridiron pattern, loops and lollipops design were identified to be related to severe pedestrian crash injury.

2.3.2.9. Work of Sasidharan and Menéndez (2014)

Sasidharan and Menéndez (2014) developed a PPO model as an alternative to model the pedestrian injury severity in crashes. A national pedestrian safety data ranging from 2008 to 2012 of Switzerland was collected and used. Variables found to be statistically significant in the model were identified, such as peak hour, familiarity of route, traffic signals and signs. In addition, factors contributing to severe injury of pedestrians in the crashes were also detected as elder pedestrians (≥75), male pedestrians, dark unlighted roadways, and midblock of pedestrians.

2.3.2.10. Work of Pour et al. (2016)

Pour et al. (2016) deployed a PPO model to analyze pedestrian crashes with midblocks behavior in Melbourne metropolitan area, Australia. And for the first time, factors such as distance of crashes to public transport stops, average road slope and some social characteristics were taken into consideration on developing the PPO model. The results of the study identified several statistically significant factors affecting the injury severities of pedestrians in pedestrian-vehicle crashes with midblock behavior, such as speed limit, light condition, pedestrian age and gender, and vehicle type.

2.3.2.11. Work of Li and Fan (2019a)

Li and Fan (2019a) used a PPO model associated with a segmentation of data based on pedestrian ages to examine the contributing factors affecting the injury severities of pedestrians in pedestrian-vehicle crashes in North Carolina. Results indicated the necessity

with accounting for different age groups when modeling the pedestrian injury severity. And differences were found among all three age groups (i.e., young [age ≤24], middle-aged [age 25-55], and older pedestrians [age ≥55]).

2.3.2.12. Work of Li and Fan (2019b)

In this study, authors applied a sequential analysis framework by combining latent class clustering and PPO model to explore potential unobserved heterogeneity across observations in the crash data. Six sub-models with different representing variables were developed and heterogeneities did exist among different classes. Some general major factors contributing to severe injury were identified as: heavy vehicle, pedestrian crossing and dash/dart-out, and pedestrian age (≥55).

Table 2.2 displays a summary of existing studies that utilized the advanced DCM methods (i.e., ML model and PPO model) focusing on pedestrian crash data analysis within transportation safety research in chronological order.

TABLE 2.2: Summary of Existing Studies Utilized the Advanced DCM Methods Focusing on Pedestrian Crash Data Analysis

Authors	Year	Case study location	Methodology	Key findings
Kim et al.	2010	North Carolina USA	ML model	1) Factors contributing to fatality of pedestrians: dark unlighted roadway, truck, freeway, speeding, and drunk driver; 2) Heterogeneity in the mean of the random parameters: pedestrian gender (freeway and pedestrian-at-fault collision), and pedestrian age (traffic control [sign] and backing vehicle).
Aziz et al.	2013	New York City USA	ML model	1) Significant factors: number of lanes, grade, light condition, road surface, presence of signal control, type of vehicle, and parking facilities, commercial and industrial land use; 2) LR test indicated necessity of separate models for different areas (i.e., boroughs).
Rifaat et al.	2012	City of Calgary Canada	PPO model	Compared to the traditional gridiron pattern, currently popular urban street patterns, such as loops and lollipops design, were identified to relate to severe pedestrian crash injury.
Sasidharan and Menéndez	2014	South Korea	PPO model	Factor contributing to fatality: elder pedestrian (≥75), pedestrian gender (male), dark unlighted roadways, and mid-block crossing behavior of pedestrians.
Islam and Jones	2014	Alabama USA	ML model	Pedestrian-at-fault accidents were analyzed and obvious different effects on some variables were identified between urban and rural areas.
Haleem et al.	2015	Florida USA	ML model	Factors contributing to severe injury:

Authors	Year	Case study location	Methodology	Key findings
				1) At signalized intersections: higher AADT, higher speed limit, percentage of trucks, elder pedestrians, at-fault pedestrians, rain, and darkness; 2) At unsignalized intersections: pedestrian walking along roadway, mid-age and elder pedestrians, at-fault pedestrians, vans, darkness, and higher speed limit.
Pour et al.	2016	Melbourne metropolitan area Australia	PPO model	Mid-block crashes of pedestrian were investigated, and factors contributing to severe injury of pedestrian were identified: higher speed limit, darkness, male pedestrian.
Tulu et al.	2017	Addis Ababa Ethiopia	ML model	Factor contributing to fatality: higher speed limit, intersections, heavy vehicle, and less educated drivers.
Kim and Ulfarsson	2019	USA	ML model	Factors contributing to severe injury of elder pedestrian: crossing street, left/right turning movement, parking lot, minivan, and SUV.
Chen and Fan	2019b	North Carolina USA	ML model	Factors contributing to severe injuries of pedestrian (on both urban and rural areas): impaired driver, heavy trucks, darkness, speed limit (35-50 mph; ≥50 mph). Differences were also identified for some variables.
Li and Fan	2019a	North Carolina USA	PPO model	Three sub-models with consideration of age differences (i.e., young [age ≤24], middle-aged [age 25-55], and older pedestrians [age ≥55]) were developed and obvious differences of effects of variables were identified to denote the necessity of accounting for different age segmentations.

Authors	Year	Case study location	Methodology	Key findings
Li and Fan	2019b	North Carolina USA	LCC and PPO model	Six sub-models with different representing variables were developed and heterogeneities did exist among different classes. Some general major factors contributing to severe injury were identified: heavy vehicle, pedestrian crossing and dash/dart-out, and pedestrian age (≥55).

Based on the ML model and PPO model introduced in this subsection, there are also some variant models that have been derived from these two models, such as generalized ordered logit model, mixed generalized ordered logit model and mixed PPO model. However, most of these variants have the difficulty of computation when handling the large dataset in the real world. Thus, the following subsection gives a brief introduction to the methodologies on constructing both the ML model and the PPO model.

2.3.3. Mixed Logit Model

This subsection gives a brief introduction to the mixed logit (ML) model that has been used to model pedestrian crash severity data in this dissertation. According to the previous section, unobserved heterogeneities could exist in the dataset and such unobserved factors can highly affect the crash outcome. Thus, considering only fixed effects of the variables (e.g., in the MNL model) may underestimate the unobserved heterogeneities. Compared to MNL models, the ML models use a similar linear utility function U_{ij} to represent the relationship between injury severity levels ($j = 0, 1, 2 \ldots J$) and explanatory variables, as shown below in Equation 2.3.2.1:

$$U_{ij} = \beta_j X_{ij} + \varepsilon_{ij} \quad (2.3.2.1)$$

where X_{ij} represents the vector of independent variables for *i*th individual with *j*th injury severity level, β_j is the vector of estimated coefficients for X_{ij}, and ε_{ij} denotes the error term representing the unobserved factors. Different from the MNL models, β_j is a vector of estimated coefficients with probabilistic distributions, and some elements within it could be different across cases of each pedestrian-vehicle crash with the consideration of unobserved data heterogeneities. If ε_{ij} follows a Gumbel type I distribution, then the probability of individual *i* suffering injury severity *j* can be expressed in Equation 2.2.3.2:

$$P_{ij}|\beta_j = \frac{exp(X_{ij}^{'}\beta_j)}{\sum_{j \in J} exp(X_{ij}^{'}\beta_j)} \quad (2.3.2.2)$$

By taking account of the randomly distributed parameters across individual observations, a mixing distribution can be further written in Equation 2.3.2.3:

$$P_{ij} = \int (P_{ij}|\beta_j) f(\beta_j|\varphi) d\beta_j \quad (2.3.2.3)$$

where $f(\beta_j|\varphi)$ is the probability density function (PDF) of random vector β_j and φ denotes a vector of parameters that describe the PDF, which are the mean and variance of the normal distribution, correspondingly.

This study follows Gong and Fan (2017)'s work. Prior to developing the ML model, a standard MNL model is developed for preselecting the significant variables with a backward stepwise process by eliminating variable(s) that at each step has(have) p-value(s) less than 0.05. Since there is no initial knowledge of implying the randomly distributed parameters (Moore et al. 2011), all effects of variables are set to be randomly distributed initially. Then a backward stepwise process is applied to determine which parameters should remain fixed or be treated as randomized. After all steps, parameters that are found

significantly different across observations are set to be randomly distributed and if not, they are constrained to be equal.

According to (Milton et al. 2008; Gkritza and Mannering 2008; Train 2009; Gong et al. 2016), normal distribution has been found to be the more suitable one for the ML models, which therefore is also used in this study for the model parameters. Normally, 200 to 1000 Halton draws are deployed to compute the approximation of the integral in Equation 3. Based on the research of (Koppelman et al., 2003; Behnood and Mannering, 2016), 500 Halton draws are enough to obtain a relatively accurate maximum likelihood estimate. Thus, in this study, 500 Halton draws is employed in the simulation.

After the ML model is fitted, marginal effects of all variables are also calculated to evaluate the impacts of the associated variables on the probabilities of injury severity levels. The formulation for calculating the marginal effect is the same as what was presented in Subsection 2.3.2. Since the ML model is developed by using a simulation-based method, the marginal effects are calculated via average simulation-based marginal effects over all observations. More details regarding this evaluation can be referred to (Moore et al. 2011; Kim et al. 2013; Gong and Fan 2017).

2.3.4. Partial Proportional Odds Logit Model

As mentioned previously, parallel-lines assumption could be invalid in many cases if there are unobserved inherent data heterogeneities. Based on the ordered logit/probit models, the generalized ordered logit model with a full relaxation of the PO assumption to all variables could overcome such issue, as shown below in Equation 2.3.3.1:

$$P(Y_i \geq j) = \frac{exp(\alpha_j - X_i^{'}\beta_j)}{1+exp(\alpha_j - X_i^{'}\beta_j)} \qquad (2.3.3.1)$$

where Y_i is the recorded crash injury for crash i, X_i presents a $p \times 1$ vector including the values of crash i on the full set of p independent variables, β_j denotes a $p \times 1$ vector of estimated coefficients, and α_j is the cut-point for jth cumulative logit. The only difference between the ordered logit model and this model is that β is not fixed across equations. However, in most cases, not all the variables violate the PO assumption. Hence, the PPO might be more realistic than the generalized ordered logit model. PPO model relaxes the PO assumption by having particular variables to violate the PO assumption, when the PO assumption can be applied to the rest variables. By partitioning the variables related to crash i into two groups associated with/without violating the PO assumption: X_i and T_i, the partial proportional odds model with logit function can be shown as (Peterson and Harrell, 1990), as presented below in Equation 2.3.3.2:

$$P(Y_i \geq j) = \frac{exp[\alpha_j - (X_i'\beta_j + T_i'\gamma_j)]}{1 + exp[\alpha_j - (X_i'\beta_j + T_i'\gamma_j)]} \qquad (2.3.3.2)$$

Variables in either vector *Xi* or *Ti* could be decided by deploying a series of Wald Chi-square tests. The test can help decide whether the PO assumption is violated or not for each independent variable in the generalized ordered logit model (Wang and Abdel-Aty, 2008; Sasidharan and Menéndez 2014).

It should be cautious when examining the results of the PPO model. The sign of β does not always denotes the direction of the effect of the intermediate outcomes (Wooldridge, 2002; Washington et al., 2020). Thus, this dissertation further uses the marginal effects to conduct the analysis. In this study, all variables are dummy variables. Therefore, a difference of probability changes rather than the derivative is computed as the marginal effect for each variable.

2.4. Machine Learning Approaches

2.4.1. Introduction

Sections 2.2 and 2.3 summarize the selected conventional statistical DCMs, which have been widely applied to model pedestrian injury severities. However, by accounting for effectiveness and accuracy, these models might become substitutable and outdated. In addition, most of the conventional DCMs are regression-based models, which have limitations with pre-assuming linear or nonlinear relationships between the exploratory variables and the response variable. When violating such conditions, the models might inevitably result in improper inferences (Chang and Chen, 2005). With the rapid development of machine learning (or artificial intelligent) techniques and the increase of data accumulations, though few efforts have been made, it becomes popular with applying machine learning to handle transportation related problems. And compared to conventional statistical and econometric DCMs, fewer requirements on the pre-defined assumptions about the relationships between outcomes of injury severity and contributing factors is an important advantage of machine learning methods as a non-parametric method (Gong et al., 2019; Rahman et al., 2019).

The machine learning methods contain a set of techniques, such as support vector machine regression, neural network approaches (e.g., deep neural network, convolutional neural network), random forest, and gradient boosting (e.g., CatBoost, LightGBM, AdaBoost, XGBoost), etc. It should be pointed out that the traffic safety related studies with the identifications of contributing factors in traffic accidents are essentially multiclass classification problems. And according to existing literatures in the field of transportation

safety research, among all the machine learning methods, decision/binary tree-based models would be the most popular and appropriate techniques (Pande et al., 2010; Chang and Chien, 2013; Rahman et al., 2019). On the other hand, despite the popularity of machine learning methods, there are really few applications focusing on applying machine learning methods in exploring the issues of pedestrian safety within the transportation system.

Thus, Subsection 2.4.2 focuses on reviewing and summarizing the existing researches in the field of transportation safety with a focus on pedestrian safety related studies by applying machine learning methods. Research studies that used machine learning methods to model and analyze pedestrian injury severities in pedestrian-vehicle crashes are reviewed and summarized in this section.

2.4.2. Literature Review on Using Machine Learning Methods to Model Pedestrian Crashes

2.4.2.1. Work of Ding et al. (2018)

By adopting the Multiple Additive Poisson Regression Trees (MAPRT), Ding et al. (2018) examined non-linear effects of the built environment on pedestrian injury severities in pedestrian-vehicle crashes. Factors of density and the mixed land level were identified to have the most impacts on pedestrian injury severity with a 66% of the total effects. Results indicated some strong non-linear relationships between the contributing factors and the responded injury severities of pedestrians in the crashes, which disagreed with the widely applied conventional statistical models with linearity assumptions.

2.4.2.2. Work of Mokhtarimousavi (2019)

By utilizing the HSIS data of pedestrian crash in Los Angeles, Mokhtarimousavi (2019) applied MNL models and support vector machine (SVM) models to analyze the pedestrian injury severity respectively. Segmentation on the data based on time-of-day was also considered. Different contributing factor sets were identified between day and night, and comparisons between MNL and SVM models were also conducted. It was found that SVM models outperformed MNL models in terms of the prediction performance. However, despite the prediction performance, this study only provided result interpretations of the contributing factors' effects on pedestrian injury severities using MNL and no explanations about the results of SVM models were elaborated.

2.4.2.3. Work of Mokhtarimousavi (2020)

In this study, authors applied an Artificial Neural Network (ANN) model and a random parameter ordered response model to examine the factors affecting pedestrian injury severities in vehicle-pedestrian crashes while accounting for possible day-of-week effects. A variety of variables were explored by two models with a further comparison of two proposed models. Results showed superiority of the optimized ANN model to the conventional statistical model in terms of the prediction performance. Additional impacts analysis of the significant variables was also derived in order to examine the effects of the variables on pedestrian injury severities in pedestrian-vehicle crashes. Strong instabilities between weekdays and weekends of the factors' effects were also found.

2.4.3. Selected Machine Learning Method (XGBoost)

This subsection provides a brief introduction to the selected machine learning method, which is the XGBoost approach. Associated inherent analysis methods for analyzing results by XGBoost are also presented as: 1) variable importance; and 2) partial dependence of variables used in the XGBoost model.

2.4.3.1. XGBoost Method

XGboost, also known as "eXtreme Gradient Boosting", is an improved gradient boosting algorithm developed by Chen and Guestrin (2016). It is one of the most popular machine learning algorithms that has been successfully and widely used by many winners in many machine learning competitions and various domains with a significant popularity.

And compared to other regular machine learning approaches, it is well recognized by its parallel capability of processing large amount of data in a much faster speed. In addition, due to its tree-based characteristics, the XGBoost algorithm has its own way to handle missing values, which provides a relatively low sensitivity to the missing value. Furthermore, as mentioned by its creators, XGBoost could produce better performance by applying a more regularized model formalization to deal with the issue of data over-fitting (Chen and Guestrin, 2016). Tests on other machine learning approaches were also conducted with algorithms such as AdaBoost, CatBoost, Light Gradient Boosting method, and Deep Neural Network, but XGBoost does provide a much better results in terms of the accuracy. Thus, XGBoost is selected to model and analyze pedestrian injury severities in pedestrian-vehicle crashes in this study. The information on this algorithm is briefly introduced in the rest of this subsection.

As mentioned previously, one of the particular improvements made by Chen and Guestrin to the basic gradient boosting method with XGBoost is the regularized objective to the loss function, which is shown as below in Equation 2.4.3.1 for the k^{th} iteration:

$$L_k = \sum_{i=1}^{m} l\left(y^i, \hat{y}_k^i\right) + \sum_{j=1}^{k} \Omega\left(f_j\right) \quad (2.4.3.1)$$

where m is the number of samples, y^i represents the actual value of sample i, which is the observed injury severity of individual i in this study, $\hat{y}_k^i$ is the prediction of the sample i at iteration k, $l(\cdot)$ denotes the original loss function without regularized term. Ω is the regularization term, which can be calculated in Equation 2.4.3.2 as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{k} w_j^2 \quad (2.4.3.2)$$

where T is the number of leaf nodes. γ and λ are the penalty coefficient of the number of leaves, and the penalty coefficient of regularization, respectively, which are used to control the degree of regularization in the algorithm. w_j is the score of the leaf j.

Additionally, rather than using the stochastic gradient descent method, XGBoost applies an additive learning strategy to complement the associating optimization process by adding the best tree model into the current model to provide prediction result for subsequent iteration. Thus, when a new tree is added, the Equation 2.4.3.1 could be rewritten as:

$$L_k = \sum_{i=1}^{m} l(y^i, \hat{y}_{k-1}^i + f_k(x^i) + \Omega(f_k) + \sum_{j=1}^{k-1} \Omega\left(f_j\right) \quad (2.4.3.3)$$

By adopting the Taylor Expansion to the objective function, Equation 2.4.3.3 could be further rewritten as in Equation 2.4.3.4:

$$L_k = \sum_{i=1}^{m} [l\left(y^i, \hat{y}_{k-1}^i\right) + g_i \cdot f_k\left(x^i\right) + \frac{1}{2} h_i \cdot f_k\left(x^i\right)] + \Omega(f_k) + C \quad (2.4.3.4)$$

where $g_i = \partial_{\hat{y}_{k-1}^i} l(y^i, \hat{y}_{k-1}^i)$ is the first order derivative of the loss function, $h_i = \partial^2_{\hat{y}_{k-1}^i} l(y^i, \hat{y}_{k-1}^i)$ is the second order derivative of the loss function, and *C* is a constant.

The objective can be condensed as:

$$L_k = \sum_{j=1}^{k} \left[G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T \qquad (2.4.3.5)$$

where $G_j = \sum_{I_j} g_i$ and $H_j = \sum_{I_j} h_i$. I_j is the set of instance to the j^{th} leaf. Then the best w_j can be obtained for the objective function is:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \qquad (2.4.3.6)$$

which transforms the final objective function to be:

$$L_k = -\frac{1}{2} \sum_{j=1}^{k} \frac{G_j}{H_j + \lambda} + \gamma T \qquad (2.4.3.7)$$

After splitting the data across all regression trees, the loss reduction could be illustrated as:

$$Gain = \frac{1}{2} \left[\frac{G_{Right}^2}{H_{Right} + \lambda} + \frac{G_{Left}^2}{H_{Left} + \lambda} + \frac{(G_{Right} + G_{Left})^2}{H_{Right} + H_{Left} + \lambda} \right] - \gamma \qquad (2.4.3.8)$$

where $\frac{G_{Right}^2}{H_{Right} + \lambda}$ and $\frac{G_{Left}^2}{H_{Left} + \lambda}$ are the scores of right and left nodes after the cuts respectively, and $\frac{(G_{Right} + G_{Left})^2}{H_{Right} + H_{Left} + \lambda}$ denote the score of combination without the cut. By minimizing the objective with the enumeration of various tree structures (or maximizing the total gains for the trees generated in the model), the best model structure could be ultimately retrieved.

2.4.3.2. Variable Importance and Partial Dependence

After obtaining the best model structure from the XGBoost, the algorithm also has the capability of ranking the important variables by using total leaf gains in the path of the

branch for each feature (i.e., contributing factors in this study), which denotes the importance for an input feature. The interpretation of *gain* of an input feature could be expressed as the relative contribution of the associating feature to the model computed by having contribution of each feature for each tree generated in the model. A feature with higher value of this metric means that it is more important in generating a prediction in the model than another. The feature importance is used to see the most influencing factor contributing to the outcomes. However, other than this functionality, it is truly helpless to interpret and analyze the results, and particularly it has little to do with examining how a factor impacts the injury severities.

Despite the feature importance, this study also utilizes the partial dependence to show the marginal effect of features affecting the predicted outcome of a machine learning model (Friedman 2001). Since all input variables in this study are dummy coded with 0 and 1 values with meaning of presence or not of the corresponding contributing factor, the partial dependence of each factor has the same meaning to the marginal effect used in the conventional DCMs. This could further help and provide convenience for comparing models between both groups.

The partial dependence function for XGBoost is defined as:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) dP(x_C) \qquad (2.4.3.9)$$

where x_S is(are) the input variable(s) for which the partial dependence function with corresponding set of S, x_C are the other variables used in the machine learning model $\hat{f}$ with corresponding set of C. In another word, the variable(s) in S is(are) the variable(s) that one wants to examine associated effects towards predicted outcomes. Partial dependence is

computed by marginalizing the machine learning model output over the distribution of the variables in set *C* with result of the function showing the relationship between the x_S and the predicted outcome of the model.

By utilizing the Monte Carlo method, the partial dependence can be estimated by calculating averages in the fitted dataset, which is shown as:

$$\hat{f}_{x_s}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{x_s}\left(x_S, x_C^i\right) \quad (2.4.3.10)$$

Therefore, the partial dependence of the selected important contributing factors identified in the XGBoost model will be used for the purpose of interpretation and comparison with conventional statistical models. Furthermore, the average change of the partial dependence, which is also the probability change (or marginal effect) of a specific factor from 0 to 1 (i.e., from the status of "not presented in the crash" to "presented in the crash") is used and main focuses are given to those factors whose changes show impacts to the severer injury levels (i.e., "K" of fatality and "A" of incapacitating injury).

2.5. Summary

A comprehensive review and synthesis of the current and historical research efforts related to modeling and analyzing pedestrian injury severities in pedestrian-vehicle crashes have been presented, associating with the introductions to methodologies of all selected approaches. This is intended to provide a solid reference and assistance in developing models for future tasks.

CHAPTER 3: DATA DESCRIPTION AND PROCESSING

3.1. Introduction

This chapter provides the basic information about the data used to analyze pedestrian-injury severities in pedestrian-vehicle crashes, which is the police-reported crash data obtained from the Division of Bicycle and Pedestrian Transportation in North Carolina Department of Transportation (NCDOT) from 2007-2018. The following sections are organized as follows. Section 3.2 presents detailed descriptive analysis of the data and section 3.3 concludes this chapter with a summary.

3.2. Descriptive Analysis of the Collected Data

In this research, police-reported pedestrian crash data of North Carolina between 2007-2018 are were acquired from the Division of Bicycle and Pedestrian Transportation of North Carolina Department of Transportation (NCDOT). The data consist of much categorical information about pedestrian characteristics, driver characteristics, crash characteristics, locality and roadway characteristics, time and environment characteristics, and traffic control characteristics and work zone. During the data cleaning process, incomplete and clearly improper observations are excluded. Additionally, only cases involving single pedestrian and single vehicle are kept. A total of 17,480 observations of pedestrian-vehicle crashes are eventually selected and used. The percentages at each injury severity level are: K: Killed: 1,154 (6.6%), A: Incapacitating Injury: 1,292 (7.39%), B: - Non-incapacitating Injury: 6,571 (37.59%), C: Possible Injury: 7,553 (43.21%), and O: No Injury: 910 (5.21%). Figure 3.1 show the distributions of each injury severity level of

pedestrians in pedestrian-vehicle crashes of the collected data, and Figure 3.2 displays the crash frequency of the five categories in each year.

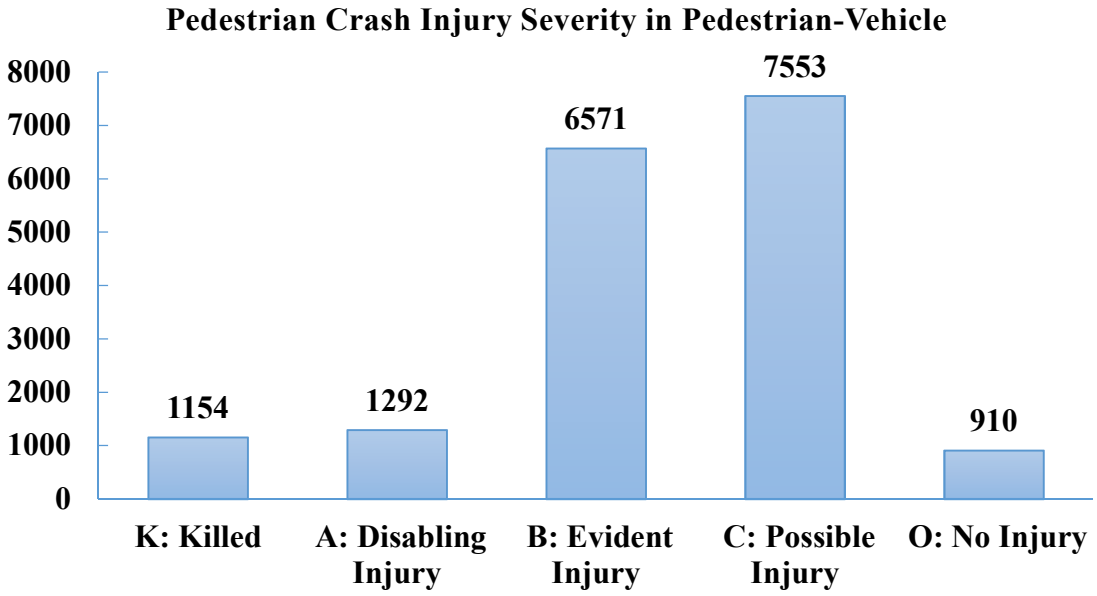

FIGURE 3.1: Distributions of Each Injury Severity Level of Pedestrians in Pedestrian-Vehicle Crashes of the Collected Data

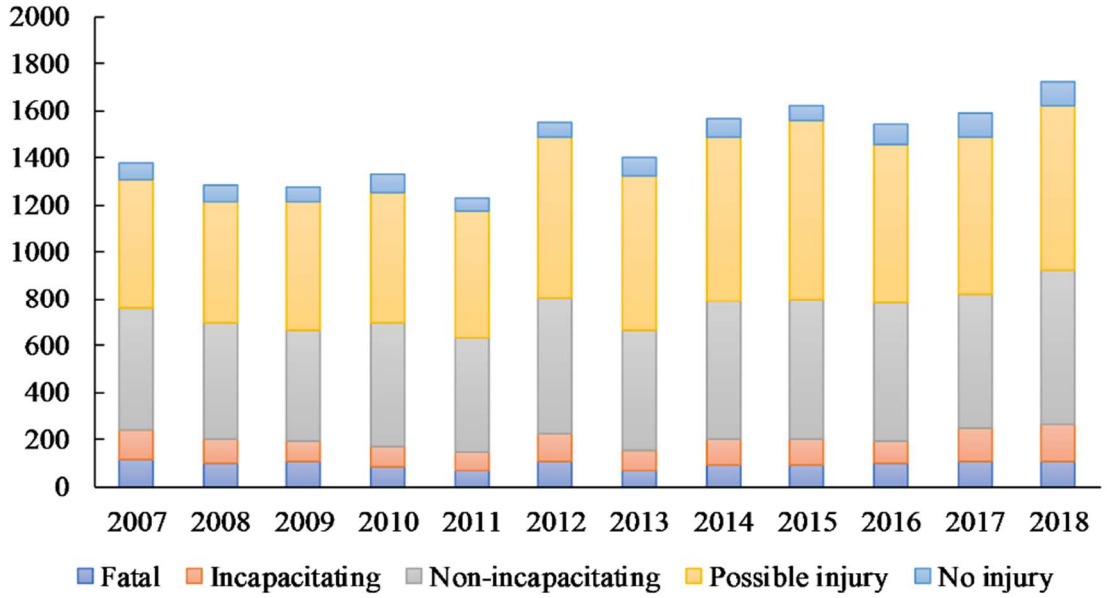

FIGURE 3.2: Crash Frequency Distribution of Injury Severity Category by Year

The data used in this research for various pedestrian injury severity levels in the pedestrian-vehicle crashes are summarized and displayed in Table 3.1. Categories of the variable and descriptions, as well as the percentages of observed crash frequency at each severity level are also included for each individual crash. Dummy variables are coded for each variable and reference variables in each category are marked with asterisks in the table.

TABLE 3.1: Descriptive Statistics of Explanatory Variable

Variable	**Total**	**K**[a]	**A**[b]	**B**[c]	**C**[d]	**O**[e]
Pedestrian–vehicle Crashes	17480 (100%)	1154 (6.60%)	1292 (7.39%)	6571 (37.59%)	7553 (43.21%)	910 (5.21%)
Pedestrian Characteristics						
Pedestrian age: 25 - 44 (1 if pedestrian is younger than 45 years old and older than 24 years old; 0 otherwise) *	5183 (29.65%)	355 (2.03%)	366 (2.09%)	1842 (10.54%)	2326 (13.31%)	294 (1.68%)
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	5574 (31.89%)	238 (1.36%)	418 (2.39%)	2369 (13.55%)	2263 (12.95%)	286 (1.64%)
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	4928 (28.19%)	400 (2.29%)	391 (2.24%)	1642 (9.39%)	2257 (12.91%)	238 (1.36%)
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	1795 (10.27%)	161 (0.92%)	117 (0.67%)	718 (4.11%)	707 (4.04%)	92 (0.53%)
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	2364 (13.52%)	468 (2.68%)	332 (1.90%)	945 (5.41%)	536 (3.07%)	83 (0.47%)
Male pedestrian (1 if pedestrian is male; 0 otherwise)	10150 (58.07%)	820 (4.69%)	866 (4.95%)	3946 (22.57%)	3966 (22.69%)	552 (3.16%)
Driver Characteristics						
Driver age: 25 - 44 (1 if driver is younger than 45 years old and older than 24 years old; 0 otherwise) *	6413 (36.69%)	462 (2.64%)	512 (2.93%)	2459 (14.07%)	2654 (15.18%)	326 (1.86%)
Driver age: ≤ 24 (1 if driver is younger than 25 years; 0 otherwise)	3276 (18.74%)	232 (1.33%)	256 (1.46%)	1292 (7.39%)	1317 (7.53%)	179 (1.02%)
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	5337 (30.53%)	325 (1.86%)	371 (2.12%)	1916 (10.96%)	2440 (13.96%)	285 (1.63%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Driver age: ≥ 65 (1 if driver is older than 64 years old; 0 otherwise)	2454 (14.04%)	135 (0.77%)	153 (0.88%)	904 (5.17%)	1142 (6.53%)	120 (0.69%)
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	460 (2.63%)	80 (0.46%)	60 (0.34%)	185 (1.06%)	117 (0.67%)	18 (0.10%)
Male driver (1 if driver is male; 0 otherwise)	9662 (55.27%)	776 (4.44%)	790 (4.52%)	3637 (20.81%)	3990 (22.83%)	469 (2.68%)
Crash characteristics						
Ambulance rescue (1 if service presents; 0 otherwise)	13367 (76.47%)	1040 (5.95%)	1212 (6.93%)	5607 (32.08%)	5219 (29.86%)	289 (1.65%)
Hit and run (1 if crash is hit-and-run; 0 otherwise)	344 (1.97%)	39 (0.22%)	36 (0.21%)	103 (0.59%)	140 (0.80%)	26 (0.15%)
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	2049 (11.72%)	28 (0.16%)	76 (0.43%)	610 (3.49%)	1189 (6.80%)	146 (0.84%)
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	6702 (38.34%)	517 (2.96%)	533 (3.05%)	2550 (14.59%)	2792 (15.97%)	310 (1.77%)
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	2054 (11.75%)	124 (0.71%)	219 (1.25%)	1058 (6.05%)	585 (3.35%)	68 (0.39%)
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	134 (.77%)	15 (0.09%)	11 (0.06%)	55 (0.31%)	46 (0.26%)	7 (0.04%)
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	271 (1.55%)	5 (0.03%)	16 (0.09%)	132 (0.76%)	102 (0.58%)	16 (0.09%)
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	2769 (15.84%)	32 (0.18%)	97 (0.55%)	878 (5.02%)	1571 (8.99%)	191 (1.09%)
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	1488 (8.51%)	244 (1.40%)	164 (0.94%)	505 (2.89%)	493 (2.82%)	82 (0.47%)
Waiting to cross (1 if crash occurred when pedestrian is waiting to cross the roadway; 0 otherwise) *	15 (.09%)	1 (0.01%)	1 (0.01%)	6 (0.03%)	6 (0.03%)	1 (0.01%)
Walking along roadway (1 if crash occurred when pedestrian is walking along roadway; 0 otherwise)	1998 (11.43%)	188 (1.08%)	175 (1.00%)	777 (4.45%)	769 (4.40%)	89 (0.51%)
Locality and roadway Characteristics						
Mixed (1 if crash occurs in mixed roadway; 0 otherwise) *	2470 (14.13%)	240 (1.37%)	221 (1.26%)	928 (5.31%)	946 (5.41%)	135 (0.77%)
Rural (1 if crash occurs in rural roadway; 0 otherwise)	2205 (12.61%)	386 (2.21%)	262 (1.50%)	814 (4.66%)	658 (3.76%)	85 (0.49%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Urban (1 if crash occurs in urban roadway; 0 otherwise)	12805 (73.26%)	528 (3.02%)	809 (4.63%)	4829 (27.63%)	5949 (34.03%)	690 (3.95%)
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	824 (4.71%)	112 (0.64%)	107 (0.61%)	304 (1.74%)	264 (1.51%)	37 (0.21%)
One-way, not divided (1 if the road configuration is one-way not divided; 0 otherwise) *	1452 (8.31%)	30 (0.17%)	59 (0.34%)	492 (2.81%)	767 (4.39%)	104 (0.59%)
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	3322 (19.00%)	377 (2.16%)	305 (1.74%)	1365 (7.81%)	1136 (6.50%)	139 (0.80%)
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	12706 (72.69%)	747 (4.27%)	928 (5.31%)	4714 (26.97%)	5650 (32.32%)	667 (3.82%)
Commercial (1 if crash occurred in commercial area; 0 otherwise)	9475 (54.2%)	452 (2.59%)	606 (3.47%)	3375 (19.31%)	4500 (25.74%)	542 (3.10%)
Farms, Woods, Pastures (1 if crash occurred in areas of farms, woods, or pastures; 0 otherwise)	1641 (9.39%)	334 (1.91%)	211 (1.21%)	595 (3.40%)	440 (2.52%)	61 (0.35%)
Industrial (1 if crash occurred in industrial area; 0 otherwise)	98 (.56%)	4 (0.02%)	8 (0.05%)	42 (0.24%)	41 (0.23%)	3 (0.02%)
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	670 (3.83%)	13 (0.07%)	21 (0.12%)	244 (1.40%)	334 (1.91%)	58 (0.33%)
Residential (1 if crash occurred in Residential area; 0 otherwise) *	5596 (32.01%)	351 (2.01%)	446 (2.55%)	2315 (13.24%)	2238 (12.80%)	246 (1.41%)
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	122 (.7%)	14 (0.08%)	9 (0.05%)	63 (0.36%)	32 (0.18%)	4 (0.02%)
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	2263 (12.95%)	218 (1.25%)	216 (1.24%)	859 (4.91%)	867 (4.96%)	103 (0.59%)
Hillcrest (1 if crash occurred at the hillcrest of the roadway; 0 otherwise)	619 (3.54%)	44 (0.25%)	58 (0.33%)	234 (1.34%)	248 (1.42%)	35 (0.20%)
Level (1 if crash occurred at level roadway; 0 otherwise) *	14476 (82.81%)	878 (5.02%)	1009 (5.77%)	5415 (30.98%)	6406 (36.65%)	768 (4.39%)
Interstate (1 if crash occurred on interstate; 0 otherwise)	194 (1.11%)	73 (0.42%)	27 (0.15%)	58 (0.33%)	33 (0.19%)	3 (0.02%)
Local street (1 if crash occurred on local street; 0 otherwise)	9005 (51.52%)	379 (2.17%)	632 (3.62%)	3701 (21.17%)	3840 (21.97%)	453 (2.59%)
NC route (1 if crash occurred on NC route; 0 otherwise)	956 (5.47%)	149 (0.85%)	142 (0.81%)	344 (1.97%)	286 (1.64%)	35 (0.20%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	348 (1.99%)	17 (0.10%)	47 (0.27%)	154 (0.88%)	121 (0.69%)	9 (0.05%)
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	4104 (23.48%)	34 (0.19%)	113 (0.65%)	1185 (6.78%)	2476 (14.16%)	296 (1.69%)
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	1803 (10.31%)	259 (1.48%)	200 (1.14%)	739 (4.23%)	533 (3.05%)	72 (0.41%)
US route (1 if crash occurred on US route; 0 otherwise) *	1070 (6.12%)	243 (1.39%)	131 (0.75%)	390 (2.23%)	264 (1.51%)	42 (0.24%)
Time and Environment characteristics						
Weekday (1 if crash occurred during weekday; 0 otherwise)	13517 (77.33%)	803 (4.59%)	963 (5.51%)	5039 (28.83%)	6001 (34.33%)	711 (4.07%)
Morning (1 if crash occurred during morning; 0 otherwise)	10380 (59.38%)	370 (2.12%)	595 (3.40%)	3781 (21.63%)	5064 (28.97%)	570 (3.26%)
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	3561 (20.37%)	264 (1.51%)	315 (1.80%)	1426 (8.16%)	1383 (7.91%)	173 (0.99%)
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	3053 (17.47%)	588 (3.36%)	399 (2.28%)	1122 (6.42%)	827 (4.73%)	117 (0.67%)
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	747 (4.27%)	44 (0.25%)	50 (0.29%)	271 (1.55%)	348 (1.99%)	34 (0.19%)
Daylight (1 if light condition is daylight; 0 otherwise) *	10119 (57.89%)	258 (1.48%)	528 (3.02%)	3752 (21.46%)	4995 (28.58%)	586 (3.35%)
Clear (1 if the weather is clear; 0 otherwise) *	13433 (76.85%)	888 (5.08%)	990 (5.66%)	5079 (29.06%)	5736 (32.81%)	740 (4.23%)
Cloudy (1 if the weather is cloudy; 0 otherwise)	2445 (13.99%)	176 (1.01%)	186 (1.06%)	894 (5.11%)	1090 (6.24%)	99 (0.57%)
Fog, Smog, Smoke (1 if the weather is fog, smog, or smoke; 0 otherwise)	79 (.45%)	17 (0.10%)	4 (0.02%)	27 (0.15%)	28 (0.16%)	3 (0.02%)
Rain (1 if the weather is raining; 0 otherwise)	1457 (8.34%)	71 (0.41%)	111 (0.64%)	544 (3.11%)	664 (3.80%)	67 (0.38%)
Snow, Sleet, Hail, Freezing Rain/Drizzle (1 if the weather is snow, sleet, hail, freezing rain, or drizzle; 0 otherwise)	66 (.38%)	2 (0.01%)	1 (0.01%)	27 (0.15%)	35 (0.20%)	1 (0.01%)
Traffic control characteristics and workzone						
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	1809 (10.35%)	272 (1.56%)	234 (1.34%)	707 (4.04%)	539 (3.08%)	57 (0.33%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Workzone (1 if crash on work-zone related road segment; 0 otherwise)	172 (.98%)	11 (0.06%)	14 (0.08%)	73 (0.42%)	68 (0.39%)	6 (0.03%)
Human control (1 if the type of traffic control is human control; 0 otherwise)	220 (1.26%)	4 (0.02%)	8 (0.05%)	62 (0.35%)	126 (0.72%)	20 (0.11%)
No control present (1 if there is no control present; 0 otherwise) *	11267 (64.46%)	737 (4.22%)	823 (4.71%)	4208 (24.07%)	4891 (27.98%)	608 (3.48%)
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	1305 (7.47%)	44 (0.25%)	52 (0.30%)	451 (2.58%)	672 (3.84%)	86 (0.49%)
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	2879 (16.47%)	97 (0.55%)	175 (1.00%)	1143 (6.54%)	1325 (7.58%)	139 (0.80%)

K[a] - Fatal Injury

A[b] - Incapacitating Injury

B[c] - Non-incapacitating Injury

C[d] - Possible Injury

O[e] - No Injury

Figure 3.2-3.8 display examples of several distributions of some important features in the dataset associating with the number of crashes under each injury severity level. Such features include the "alcohol-impaired pedestrians", "alcohol-impaired drivers", "pedestrian age groups", "driver age groups", "pedestrian gender", "driver gender", "rural, mixed, or urban area", and "roadway geometry". This gives some examples on how to statistically interpret the data via descriptive analysis and provides a clear picture on what kind of the contributing factors are included when developing the model. Moreover, this gives a more intuitive and visual impression of each feature in the dataset. For instance, in Figure 3.4, even though the total number of crashes involving alcohol-impaired pedestrians is much larger than the ones without alcohol-impaired pedestrians being involved, the difference in the number of fatalities happened under both situations is relatively small.

FIGURE 3.3: Distributions of Crashes with Alcohol-impaired Drivers under Each Injury Severity Level

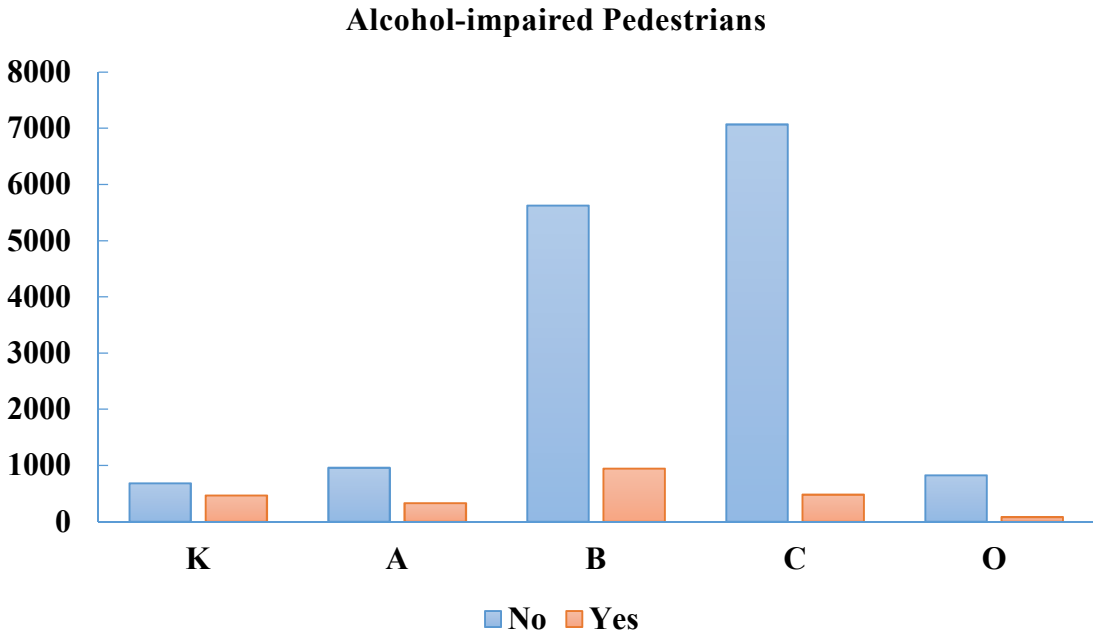

FIGURE 3.4: Distributions of Crashes with Alcohol-impaired Pedestrians under Each Injury Severity Level

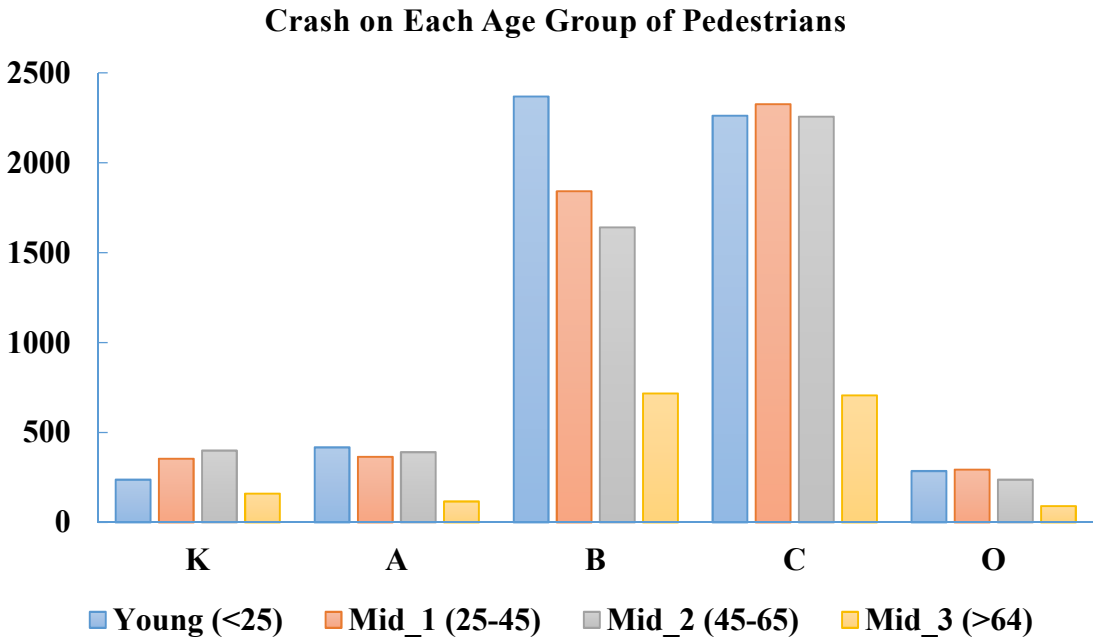

FIGURE 3.5: Distributions of Crashes on Each Age Group of Pedestrians under Each Injury Severity Level

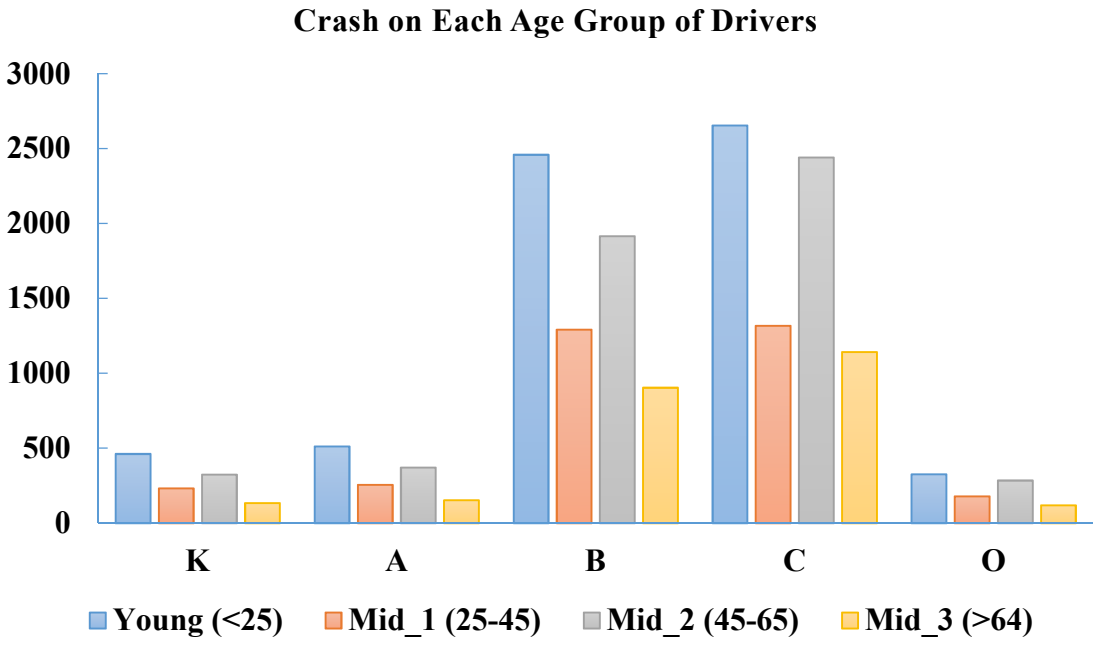

FIGURE 3.6: Distributions of Crashes on Each Age Group of Drivers under Each Injury Severity Level

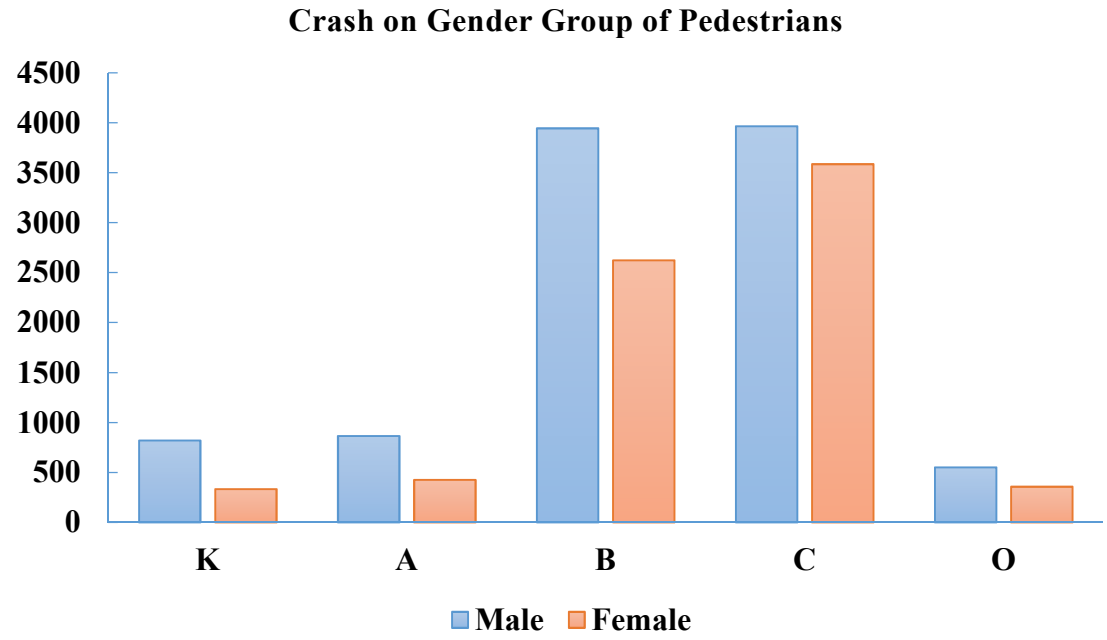

FIGURE 3.7: Distributions of Crashes on Each Gender of Pedestrians under Each Injury Severity Level

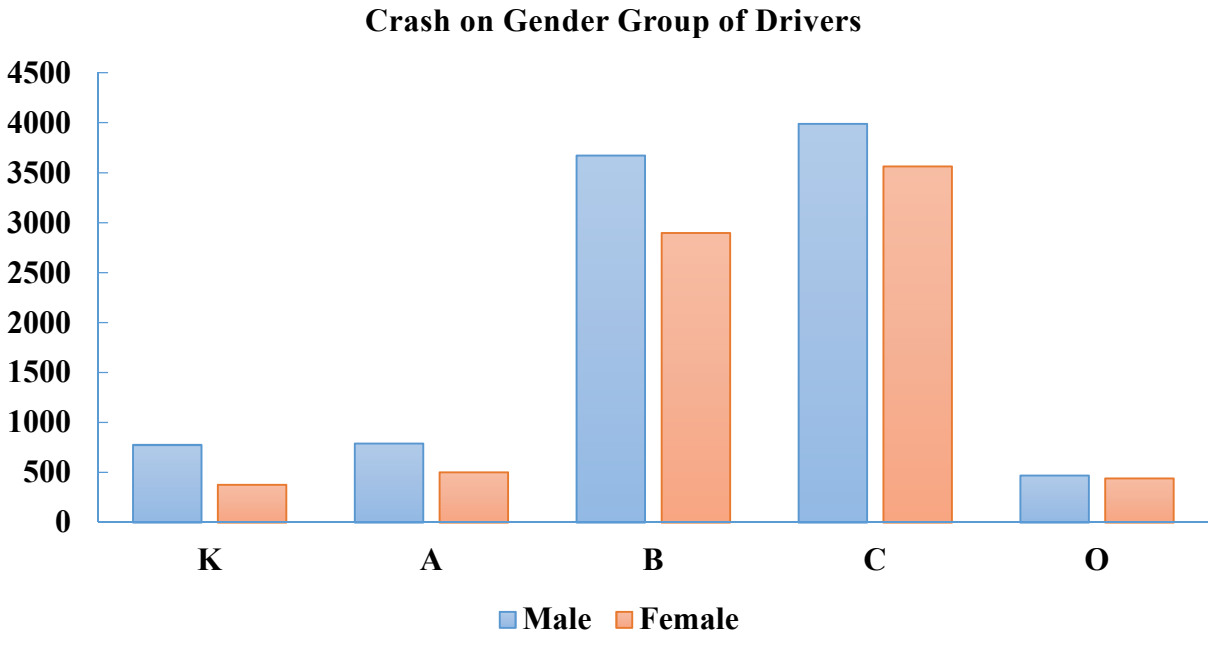

FIGURE: 3.8 Distributions of Crashes on Each Gender of Drivers under Each Injury Severity Level

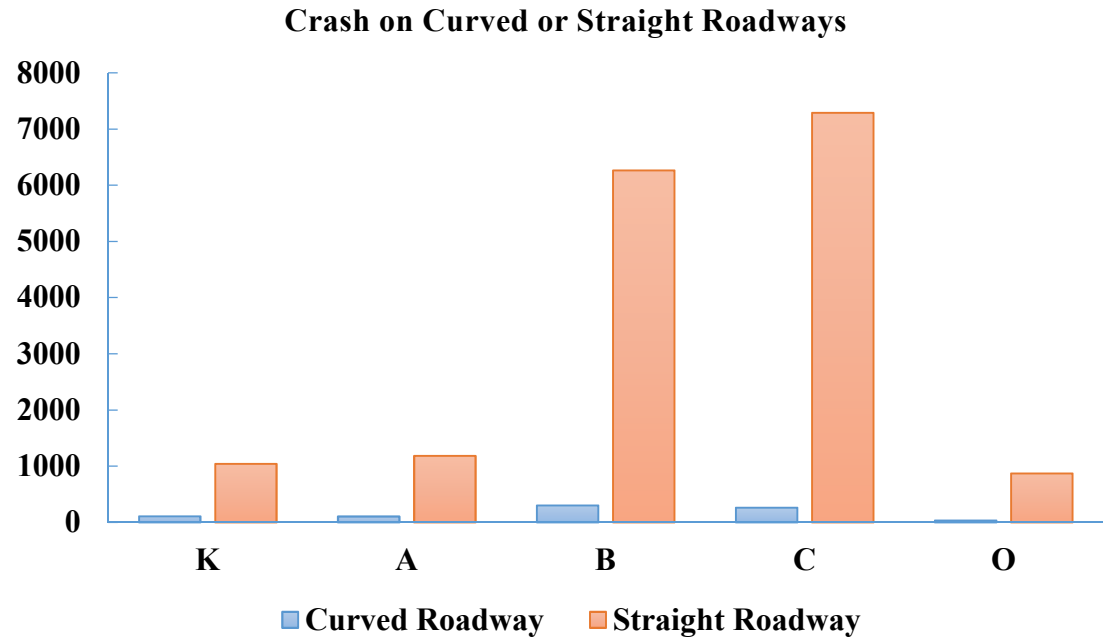

FIGURE 3.9: Distributions of Crashes on Curved or Straight Roadways under Each Injury Severity Level

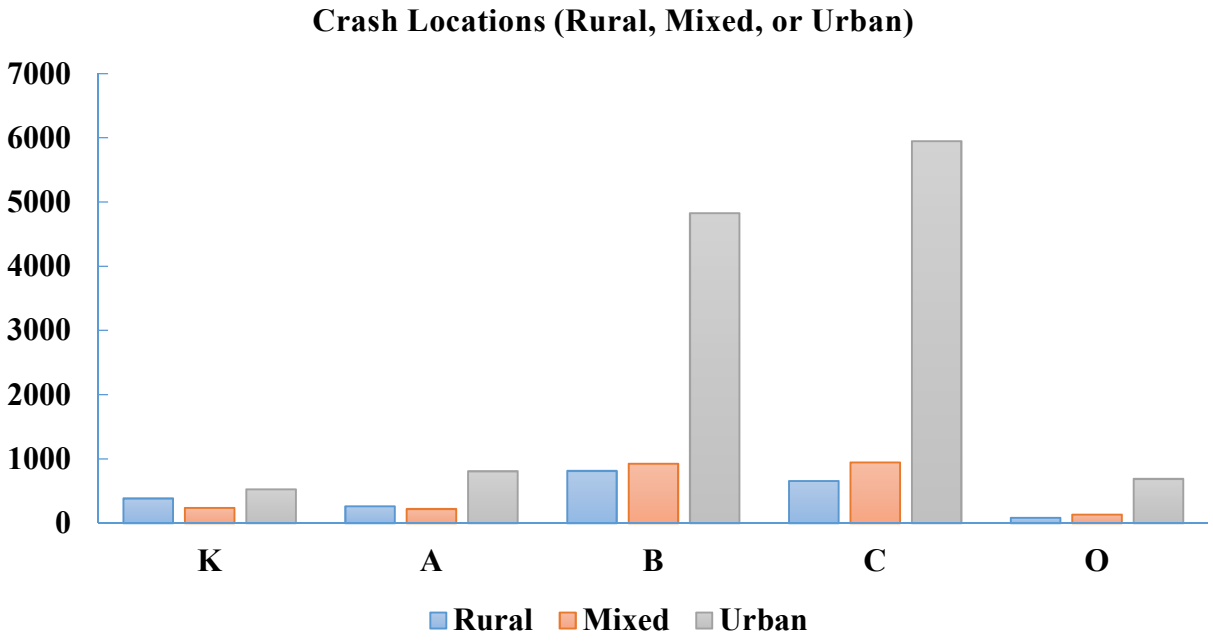

Figure 3.10: Distributions of Crashes in Rural or Urban Areas under Each Injury Severity Level

3.3. Summary

This chapter presents the detailed information on the data source, data structure, and processing methodology. This is intended to provide a solid reference and assistance for future tasks.

CHAPTER 4: DEVELOPMENTS OF DISCRETE CHOICE MODELS

4.1. Introduction

The chapter presents the model developments of conventional DCMs for analyzing and modeling pedestrian injury severities in pedestrian-vehicle crashes. The following sections are organized as follows. Section 4.2 shows development of basic DCM (i.e., MNL), including the model results and the associating interpretation of the results. Section 4.3 presents model developments of the selected advanced DCMs (i.e., ML model and PPO model), and the model results and associating result explanations are also provided. Section 4.4 shows some simple comparisons between basic DCM and the advanced DCMs. Finally, section 4.5 concludes this chapter with a summary.

4.2. Development of Basic Discrete Choice Model (MNL Model)

This section describes the results of the selected basic DCM for analyzing and modeling pedestrian-injury severity in pedestrian-vehicle crashes, which is the MNL model. Then the interpretations of the model results are briefly given as a general guide in the Subsection 4.2.2 to demonstrate the use of MNL model in analyzing and modeling pedestrian injury severities in pedestrian-vehicle crashes, especially its use in identifying the key contributors to the injury severity levels of pedestrians. The associated marginal effects of all individual contributing factors that remain significant in the final MNL model are also computed for the purpose of using them as supplements to the results from the direct interpretations of the developed MNL model in this study.

4.2.1. Multinomial Logit Model Results

Table 4.1 represents the model results of MNL model for modeling the pedestrian injury severity in pedestrian-vehicle crashes. It shows the coefficients and standard errors of each contributing factor in the developed MNL model.

Table 4.2 displays the results of the marginal effects for each contributing factor in the developed MNL model. And as mentioned in Section 2.2, the marginal effect analysis could help evaluate how the significant variables estimated in the MNL model impact the pedestrian injury outcome probabilities.

TABLE 4.1: MNL Model Results for Modeling the Pedestrian Injury Severity in Pedestrian-Vehicle Crashes

Variables	**K**[a]		**A**[b]		**B**[c]		**O**[c]	
	Coef.	**S.E.**	**Coef.**	**S.E.**	**Coef.**	**S.E.**	**Coef.**	**S.E.**
Intercept	-3.7295	0.2621	-3.9766	0.2209	-0.6062	0.0703	-1.0775	0.1317
Pedestrian Characteristics								
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	-	-	0.2049	0.0827	0.1869	0.0397	-	-
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	0.5320	0.0773	0.2935	0.0794	-	-	-	-
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	1.7530	0.1175	0.7824	0.1229	0.5618	0.0609	0.2985	0.1224
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	1.0846	0.0878	0.7643	0.0898	0.5094	0.0628	0.2935	0.1338
Male pedestrian (1 if pedestrian is male; 0 otherwise)	-	-	0.2092	0.0667	0.1433	0.0352	0.2657	0.0747
Driver Characteristics	-	-	-	-	-	-	-	-
Driver age: 45 - 64 (1 if driver is younger than 65 years old and	-0.2734	0.0780	-0.1906	0.0693	-0.1426	0.0377	-	-

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
older than 44 years old; 0 otherwise)								
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	1.1525	0.1700	0.7721	0.1727	0.4633	0.1213	-	-
Male driver (1 if driver is male; 0 otherwise)	0.3176	0.0717	0.1594	0.0617	-	-	-	-
Crash characteristics	-	-	-	-	-	-	-	-
Ambulance rescue (1 if service presents; 0 otherwise)	1.0512	0.1110	1.6819	0.1202	0.8386	0.0441	-1.6003	0.0764
Hit and run (1 if crash is hit-and-run; 0 otherwise)	0.9921	0.2122	0.7267	0.1940	-	-	-	-
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	-0.4461	0.2196	-	-	-	-	-	-
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	0.8344	0.1047	0.6201	0.0964	0.1904	0.0508	-	-
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	1.1227	0.1380	1.1466	0.1177	0.6982	0.0658	-	-
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	0.9581	0.3131	-	-	-	-	-	-
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	-	-	0.5942	0.2873	0.4889	0.1374	-	-
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	-	-	-	-	0.2069	0.0627	-	-
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	0.9115	0.1122	0.5243	0.1138	-	-	-	-
Locality and roadway Characteristics	-	-	-	-	-	-	-	-
Urban (1 if crash occurs in urban roadway; 0 otherwise)	-0.4697	0.0898	-0.1912	0.0805	-	-	-	-
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	0.4763	0.1258	0.5397	0.1169	-	-	-	-
One-way, not divided (1 if the road configuration is one-way not divided; 0 otherwise)	-	-	-	-	-	-	-	-
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	1.0373	0.2136	0.6390	0.1616	0.2210	0.0477	-	-

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	0.4351	0.2087	0.3547	0.1481	-	-	-	-
Commercial (1 if crash occurred in commercial area; 0 otherwise)	-	-	-	-	-0.1195	0.0378	-	-
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	-	-	-	-	-	-	0.3610	0.1498
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	0.7963	0.3466	-	-	0.7033	0.2040	-	-
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	0.3432	0.0903	0.1895	0.0827	-	-	-	-
Interstate (1 if crash occurred on interstate; 0 otherwise)	0.5603	0.1860	-	-	-	-	-	-
Local street (1 if crash occurred on local street; 0 otherwise)	-1.4281	0.1103	-0.6721	0.0885	-0.3599	0.0524	-0.2676	0.1157
NC route (1 if crash occurred on NC route; 0 otherwise)	-0.4099	0.1304	-	-	-0.2364	0.0823	-	-
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	-	-	0.8013	0.1887	-	-	-0.7935	0.3607
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	-2.4229	0.2118	-1.2418	0.1396	-0.8460	0.0704	-0.3690	0.1261
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	-0.3165	0.1144	-	-	-	-	-	-
Time and Environment characteristics	-	-	-	-	-	-	-	-
Morning (1 if crash occurred during morning; 0 otherwise)	-	-	-	-	-0.1632	0.0373	-0.2025	0.0790
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	0.9339	0.1043	0.4299	0.0824	-	-	-	-
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	1.1793	0.0996	0.5925	0.0875	-	-	-	-
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	0.6919	0.1786	-	-	-	-	-	-
Cloudy (1 if the weather is cloudy; 0 otherwise)	-	-	-	-	-	-	-0.2688	0.1120
Rain (1 if the weather is raining; 0 otherwise)	-0.4227	0.1359	-	-	-	-	-	-

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Traffic control characteristics and workzone	-	-	-	-	-	-	-	-
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	-	-	-	-	-	-	-0.3575	0.1652
Human control (1 if the type of traffic control is human control; 0 otherwise)	-1.5304	0.5297	-0.8822	0.3820	-0.3714	0.1609	-	-
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.7215	0.1733	-0.8445	0.1544	-0.3254	0.0685	-	-
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.8302	0.1256	-0.5103	0.0997	-0.1857	0.0539	-	-
K[a] - Fatal Injury	No. of observations: 17,480.							
A[b] - Incapacitating Injury	-2×Log-likelihood at convergence: 38,652.							
B[c] - Non-incapacitating Injury	-2×Log-likelihood (constant only): 56,266.							
C[d] - Possible Injury	AIC: 38,839.							
O[e] - No Injury	BIC: 39,658.							

TABLE 4.2: Average Marginal Effects for Each Contributing Factors in the MNL Model

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Pedestrian Characteristics					
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	-0.0063	0.0077	0.0353	-0.0334	-0.0032
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	0.0263	0.0149	-0.0229	-0.0167	-0.0016
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	0.0951	0.0132	0.0325	-0.1391	-0.0016
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	0.0408	0.0260	0.0527	-0.1206	0.0012
Male pedestrian (1 if pedestrian is male; 0 otherwise)	-0.0058	0.0087	0.0215	-0.0341	0.0096
Driver Characteristics					
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	-0.0087	-0.0057	-0.0193	0.0307	0.0030

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	0.0506	0.0272	0.0397	-0.1067	-0.0108
Male driver (1 if driver is male; 0 otherwise)	0.0146	0.0077	-0.0124	-0.0090	-0.0009
Crash characteristics					
Ambulance rescue (1 if service presents; 0 otherwise)	0.0269	0.0611	0.1611	-0.1247	-0.1245
Hit and run (1 if crash is hit-and-run; 0 otherwise)	0.0559	0.0452	-0.0547	-0.0423	-0.0040
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	-0.0201	0.0035	0.0095	0.0064	0.0007
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	0.0337	0.0280	0.0016	-0.0577	-0.0056
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	0.0324	0.0496	0.0822	-0.1492	-0.0151
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	0.0652	-0.0102	-0.0306	-0.0219	-0.0024
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	-0.0165	0.0252	0.0889	-0.0888	-0.0088
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	-0.0051	-0.0064	0.0462	-0.0316	-0.0032
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	0.0511	0.0284	-0.0435	-0.0328	-0.0032
Locality and roadway Characteristics					
Urban (1 if crash occurs in urban roadway; 0 otherwise)	-0.0233	-0.0087	0.0178	0.0129	0.0013
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	0.0211	0.0362	-0.0312	-0.0239	-0.0022
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	0.0488	0.0272	-0.0011	-0.0681	-0.0067
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	0.0183	0.0185	-0.0205	-0.0149	-0.0014
Commercial (1 if crash occurred in commercial area; 0 otherwise)	0.0029	0.0036	-0.0265	0.0182	0.0018
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	-0.0008	-0.0008	-0.0062	-0.0111	0.0189
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	0.0237	-0.0264	0.1361	-0.1209	-0.0126
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	0.0170	0.0096	-0.0147	-0.0108	-0.0010
Interstate (1 if crash occurred on interstate; 0 otherwise)	0.0340	-0.0056	-0.0160	-0.0112	-0.0012
Local street (1 if crash occurred on local street; 0 otherwise)	-0.0619	-0.0205	-0.0166	0.1017	-0.0027
NC route (1 if crash occurred on NC route; 0 otherwise)	-0.0144	0.0109	-0.0423	0.0417	0.0042
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	-0.0081	0.0719	-0.0245	-0.0110	-0.0283

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	-0.0614	-0.0365	-0.1120	0.2097	0.0002
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	-0.0153	0.0027	0.0072	0.0049	0.0005
Time and Environment characteristics					
Morning (1 if crash occurred during morning; 0 otherwise)	0.0043	0.0054	-0.0331	0.0304	-0.0070
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	0.0507	0.0197	-0.0386	-0.0290	-0.0028
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	0.0653	0.0314	-0.0528	-0.0401	-0.0039
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	0.0430	-0.0070	-0.0202	-0.0143	-0.0015
Cloudy (1 if the weather is cloudy; 0 otherwise)	0.0004	0.0005	0.0037	0.0068	-0.0115
Rain (1 if the weather is raining; 0 otherwise)	-0.0193	0.0034	0.0091	0.0062	0.0007
Traffic control characteristics and workzone					
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	0.0006	0.0006	0.0048	0.0087	-0.0147
Human control (1 if the type of traffic control is human control; 0 otherwise)	-0.0426	-0.0296	-0.0341	0.0970	0.0092
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0202	-0.0325	-0.0351	0.0802	0.0076
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0293	-0.0198	-0.0091	0.0530	0.0051

K[a] - Fatal Injury

A[b] - Incapacitating Injury

B[c] - Non-incapacitating Injury

C[d] - Possible Injury

O[e] - No Injury

4.2.2. Results Interpretations of Multinomial Logit Model

Since "Possible injury" is set to be the reference injury severity level, there is no model for this group. Positive (negative) sign of the coefficient indicates that the corresponding variable will increase (decrease) the probability of occurrence of crashes with the injury severity level versus the base level and the base variable within the same

group. For instance, variable “male driver” has the positive coefficient (i.e., 0. 3176) for the “Fatal Injury” level, which indicates that male driver involved pedestrian-vehicle crashes will increase the probability of fatal level of pedestrian injury severity compared to “Possible injury” and “female drivers”. On the contrary, crashes under the condition of traffic control with human control (i.e., variable “Human control”) will result in a lower chance of being fatally injured than possible injured for pedestrians.

On the other hand, as a supplement to the direct interpretation of the estimated parameter, the marginal effects are also very useful and could help extend the interpretation to the reference injury severity level (i.e., “Possible Injury” in this study) as well. And as presented in Subsection 2.3.2, the marginal effect denotes the probability change when the corresponding variable changes one unit (i.e., from 0 to 1). Since all variables are dummy coded in the crash data, the associated marginal effect of a variable indicates the impacts of the presence of the variable in the crash on the injury severities of pedestrians in the pedestrian-vehicle crashes. For instance, the marginal effect of “alcohol-impaired driver” for “Possible Injury” is -0.1067, which means that when crash occurred involves a drunk driver, the probability of pedestrian sustaining possible injury decreases by 10.67%. Due to the ease of such straightforward interpretations, explanations of other variables and their associated parameters are not repeated here.

4.3. Development of Advanced Discrete Choice Models

This section presents the results of the selected advanced DCMs for analyzing and modeling pedestrian injury severities in pedestrian-vehicle crashes, which are the ML model and the PPO model. The interpretations of the model results are briefly given in the Subsections 4.2.4 and 4.2.5 to demonstrate the uses of ML model and PPO model,

respectively in analyzing and modeling pedestrian injury severities in pedestrian-vehicle crashes, especially its use in identifying the key contributors to the injury severity levels of pedestrians. Unlike the MNL model, in ML model, due to allowance of the random distributed setting of some variables, and in the PPO model, due to the ordered property of the injury severity levels, the signs of the estimated parameters could not always denote the directions of the associated effect changes of contributing factors, which do require the computations of the marginal effects for both models. Therefore, the marginal effects of all individual contributing factors that remain statistically significant in the final ML model and PPO model are computed for the purpose of interpretations.

4.3.1. Mixed Logit Model

4.3.1.1. Model Results of Mixed Logit Model

Table 4.3 represents the model results of ML model for modeling the pedestrian injury severity in pedestrian-vehicle crashes. It shows the coefficients and standard errors of each contributing factor in the developed ML model.

Table 4.4 displays the results of the marginal effects for each contributing factor in the developed ML model. And as mentioned in Section 2.3, the marginal effect analysis could help evaluate how the significant variables estimated in the ML model impact the pedestrian injury outcome probabilities.

TABLE 4.3: ML Model Results for Modeling the Pedestrian Injury Severity in Pedestrian-Vehicle Crashes

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Intercept	-4.3381	0.3443	-4.3167	0.3018	-0.6887	0.0800	-1.0845	0.1335
Pedestrian Characteristics								
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	-	-	0.2243	0.0953	0.1915	0.0439	-	-
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	0.7028	0.1157	0.3411	0.0926	-	-	-	-
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	2.1832	0.2192	0.8229	0.1463	0.3678	0.1457	0.3235	0.1243
Standard deviation of "Pedestrian age: ≥ 65"	-	-	-	-	-	-	2.1140	0.6255
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	1.3267	0.1429	0.8020	0.1036	0.4087	0.1002	0.3160	0.1377
Standard deviation of "Alcohol-impaired pedestrian"	-	-	-	-	-	-	1.3724	0.4913
Male pedestrian (1 if pedestrian is male; 0 otherwise)	-	-	0.2458	0.0772	0.1650	0.0402	0.2639	0.0755
Driver Characteristics								
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	-0.3453	0.1026	-0.2097	0.0793	-0.1528	0.0426	-	-
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	1.4109	0.2355	0.8476	0.1937	0.5241	0.1364	-	-
Male driver (1 if driver is male; 0 otherwise)	0.3890	0.0985	-0.2991	0.2924	-	-	-	-
Standard deviation of "Male driver"	-	-	-1.2280	0.4062	-	-	-	-
Crash characteristics								
Ambulance rescue (1 if service presents; 0 otherwise)	0.9253	0.2275	1.8366	0.1581	0.9408	0.0539	-1.6040	0.0767
Standard deviation of "Ambulance rescue"	1.0628	0.4036	-	-	-	-	-	-
Hit and run (1 if crash is hit-and-run; 0 otherwise)	1.2023	0.3140	0.7935	0.2365	-	-	-	-

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	-0.5673	0.2809	-	-	-	-	-	-
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	1.0630	0.1479	0.7222	0.1126	0.2085	0.0561	-	-
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	1.3852	0.1902	1.2665	0.1400	0.7667	0.0870	-	-
Standard deviation of "Dash/dart out"	-	-	-	-	-1.5212	0.5708	-	-
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	1.2834	0.3982	-	-	-	-	-	-
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	-	-	0.6799	0.3080	0.5592	0.1400	-	-
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	-	-	-	-	0.0769	0.1256	-	-
Standard deviation of "Off roadway"	-	-	-	-	-1.2309	0.3575	-	-
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	0.9911	0.2487	0.5974	0.1331	-	-	-	-
Standard deviation of "Pedestrian in roadway"	1.0666	0.5460	-	-	-	-	-	-
Locality and roadway Characteristics								
Urban (1 if crash occurs in urban roadway; 0 otherwise)	-1.0127	0.2610	-0.2277	0.0942	-	-	-	-
Standard deviation of "Urban"	1.3273	0.3427	-	-	-	-	-	-
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	0.6664	0.1747	0.6092	0.1365	-	-	-	-
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	1.2091	0.2609	0.7015	0.1935	0.2193	0.0527	-	-
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	0.4703	0.2503	0.4289	0.1782	-	-	-	-
Commercial (1 if crash occurred in commercial area; 0 otherwise)	-	-	-	-	-0.1575	0.0435	-	-
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	-	-	-	-	-	-	0.3619	0.1509
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	0.8790	0.3987	-	-	0.8191	0.2309	-	-

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	-0.3896	0.3164	0.2064	0.0944	-	-	-	-
Standard deviation of "Grade-road"	2.0681	0.3978	-	-	-	-	-	-
Interstate (1 if crash occurred on interstate; 0 otherwise)	0.8206	0.2650	-	-	-	-	-	-
Local street (1 if crash occurred on local street; 0 otherwise)	-1.8356	0.1995	-0.7113	0.1011	-0.3515	0.0582	-0.2614	0.1160
NC route (1 if crash occurred on NC route; 0 otherwise)	-0.5594	0.1759	-	-	-0.2778	0.0932	-	-
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	-	-	0.9424	0.2226	-	-	-0.8047	0.3676
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	-2.8319	0.2874	-3.0266	1.0631	-0.9051	0.0806	-0.3517	0.1262
Standard deviation of "Public vehicular area"	-	-	-2.2511	0.7803	-	-	-	-
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	-0.4382	0.1548	-	-	-	-	-	-
Time and Environment characteristics								
Morning (1 if crash occurred during morning; 0 otherwise)	-	-	-	-	-0.1774	0.0419	-0.2073	0.0798
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	1.1656	0.1649	0.4826	0.0976	-	-	-	-
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	1.4778	0.1654	0.6590	0.0980	-	-	-	-
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	0.9110	0.2271	-	-	-	-	-	-
Cloudy (1 if the weather is cloudy; 0 otherwise)	-	-	-	-	-	-	-0.2707	0.1135
Rain (1 if the weather is raining; 0 otherwise)	-0.5744	0.1910	-	-	-	-	-	-
Traffic control characteristics and workzone								
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	-	-	-	-	-	-	-0.3527	0.1668

Variables	K[a]		A[b]		B[c]		O[e]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Human control (1 if the type of traffic control is human control; 0 otherwise)	-1.8655	0.6432	-0.9143	0.4172	-0.4403	0.1817	-	-
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.8935	0.2325	-0.8885	0.1682	-0.3434	0.0754	-	-
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	-1.0337	0.1754	-0.5628	0.1107	-0.1660	0.0593	-	-
K[a] - Fatal Injury	No. of observations: 17,480.							
A[b] - Incapacitating Injury	-2×Log-likelihood at convergence: 38,596.							
B[c] - Non-incapacitating Injury	-2×Log-likelihood (constant only): 56,266.							
C[d] - Possible Injury	AIC: 38,803.							
O[e] - No Injury	BIC: 39,602.							

TABLE 4.4: Average Marginal Effects for Each Contributing Factors in the ML Model

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Pedestrian Characteristics					
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	-0.0043	0.0066	0.0361	-0.0351	-0.0033
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	0.0229	0.0140	-0.0211	-0.0143	-0.0015
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	0.0935	0.0162	-0.0026	-0.1112	0.0041
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	0.0374	0.0251	0.0368	-0.1048	0.0055
Male pedestrian (1 if pedestrian is male; 0 otherwise)	-0.0043	0.0078	0.0258	-0.0392	0.0098
Driver Characteristics					
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	-0.0076	-0.0049	-0.0226	0.0319	0.0031
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	0.0427	0.0227	0.0595	-0.1135	-0.0114
Male driver (1 if driver is male; 0 otherwise)	0.0141	-0.0177	0.0019	0.0018	0.0000
Crash characteristics					
Ambulance rescue (1 if service presents; 0 otherwise)	0.0129	0.0496	0.1925	-0.1237	-0.1313
Hit and run (1 if crash is hit-and-run; 0 otherwise)	0.0459	0.0406	-0.0483	-0.0347	-0.0035

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	-0.0160	0.0027	0.0077	0.0049	0.0006
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	0.0292	0.0257	0.0085	-0.0577	-0.0056
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	0.0294	0.0420	0.1038	-0.1592	-0.0159
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	0.0598	-0.0089	-0.0288	-0.0197	-0.0023
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	-0.0123	0.0214	0.1036	-0.1027	-0.0100
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	-0.0012	-0.0020	0.0168	-0.0124	-0.0012
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	0.0352	0.0283	-0.0358	-0.0251	-0.0025
Locality and roadway Characteristics					
Urban (1 if crash occurs in urban roadway; 0 otherwise)	-0.0331	-0.0065	0.0227	0.0153	0.0017
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	0.0208	0.0324	-0.0299	-0.0212	-0.0020
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	0.0381	0.0249	0.0058	-0.0625	-0.0062
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	0.0125	0.0182	-0.0176	-0.0120	-0.0012
Commercial (1 if crash occurred in commercial area; 0 otherwise)	0.0025	0.0040	-0.0344	0.0254	0.0024
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	-0.0005	-0.0006	-0.0063	-0.0124	0.0199
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	0.0150	-0.0234	0.1655	-0.1425	-0.0145
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	-0.0129	0.0133	-0.0001	-0.0005	0.0001
Interstate (1 if crash occurred on interstate; 0 otherwise)	0.0339	-0.0053	-0.0163	-0.0109	-0.0013
Local street (1 if crash occurred on local street; 0 otherwise)	-0.0514	-0.0187	-0.0224	0.0961	-0.0036
NC route (1 if crash occurred on NC route; 0 otherwise)	-0.0129	0.0102	-0.0512	0.0491	0.0048
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	-0.0063	0.0690	-0.0254	-0.0073	-0.0300
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	-0.0428	-0.0610	-0.1255	0.2283	0.0009
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	-0.0136	0.0023	0.0066	0.0042	0.0005
Time and Environment characteristics					
Morning (1 if crash occurred during morning; 0 otherwise)	0.0030	0.0048	-0.0355	0.0350	-0.0074

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	0.0423	0.0179	-0.0339	-0.0238	-0.0025
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	0.0521	0.0290	-0.0456	-0.0322	-0.0033
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	0.0380	-0.0059	-0.0183	-0.0123	-0.0015
Cloudy (1 if the weather is cloudy; 0 otherwise)	0.0003	0.0004	0.0038	0.0076	-0.0121
Rain (1 if the weather is raining; 0 otherwise)	-0.0164	0.0028	0.0079	0.0050	0.0006
Traffic control characteristics and workzone					
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	0.0004	0.0005	0.0048	0.0095	-0.0152
Human control (1 if the type of traffic control is human control; 0 otherwise)	-0.0320	-0.0230	-0.0559	0.1013	0.0096
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0169	-0.0262	-0.0435	0.0792	0.0074
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0234	-0.0180	-0.0091	0.0461	0.0045

K[a] - Fatal Injury

A[b] - Incapacitating Injury

B[c] - Non-incapacitating Injury

C[d] - Possible Injury

O[e] - No Injury

4.3.1.2. Results Interpretations of Mixed Logit Model

As defined in Section 2.3.3, there are two subsets in the coefficient vector β, which consist of fixed ones and randomly distributed ones respectively. For fixed coefficients, they follow the same explanation as that for the MNL model. For demonstration purpose, the explanatory variable "ambulance rescue" specific to "Fatal Injury (K)" is selected to show the interpretations for variables with random effects in the ML model. The fitted normal distribution of the coefficient of "ambulance rescue" specific to "Killed (K)" has a mean and standard deviation of 0.9253 and 1.0628 respectively. Thus, the probability of such distribution to be below zero is 25.70% and 74.30% to be above zero. Such

phenomenon indicates that the likelihood of being killed increases for 74.3% of pedestrians involved in pedestrian-vehicle crashes when they walk along the roadway, while for the minority (25.70%) of pedestrians, this likelihood decreases. Thus, this results in a more severe outcome of pedestrian injury severity level under such condition in most cases. Meanwhile, there are still a certain proportion of pedestrians not being killed. This kind of interpretation would be more plausible than the result from the MNL model in which the constant positive effect of "ambulance rescue" in MNL is assumed, which implies that the presence of the ambulance rescue service would always be a positive contributing factor to severer injury levels in pedestrian-vehicle crashes for pedestrians. Thus, the ML model can reveal heterogeneities across individual observations.

Additionally, the average marginal effects across individuals of all statistically significant contributing factors have been computed as well. Like MNL model, the marginal effects can be explained in the same manner and therefore are not repeated. However, there are some interesting differences between the MNL and ML models due to the consideration of unobserved heterogeneity or without. For example, the average marginal effect of "grade-road" under level "K" are with opposite signs in MNL and ML models (i.e., 0.0170 in MNL, and -0.0129 in ML), which implies that this factor would increase the risk of pedestrian being killed in pedestrian-vehicle crashes in MNL model but decrease such probability in ML model, on average. Other than such difference, most effects of the other contributing factors in ML model tend to have smaller absolute average marginal effects towards "K" level than they are in the MNL model.

4.3.2. Partial Proportional Odds Model

4.3.2.1. Model Results of Partial Proportional Odds Model

Table 4.5 represents the model results of PPO model for modeling the pedestrian injury severity in pedestrian-vehicle crashes. It shows the coefficients and standard errors of each contributing factor in the developed PPO model.

Table 4.6 displays the results of the marginal effects for each contributing factor in the developed PPO model. And as mentioned in Section 2.3, the marginal effect analysis could help evaluate how the significant variables estimated in the PPO model impact the pedestrian injury outcome probabilities.

TABLE 4.5: PPO Model Results for Modeling the Pedestrian Injury Severity in Pedestrian-Vehicle Crashes

Variable	All Level		K[d]		A[c]		B[b]		C[d]	
	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.	Coef.	S.E.
Pedestrian Characteristics										
Pedestrian age: ≤ 24	-	-	-0.3692	0.0815	-0.1102	0.0579	0.211	0.0424	0.1291	0.0805
Pedestrian age: 45 - 64	0.1687	0.0384	-	-	-	-	-	-	-	-
Pedestrian age: ≥ 65	0.7003	0.055	-	-	-	-	-	-	-	-
Alcohol-impaired pedestrian	0.5972	0.0474	-	-	-	-	-	-	-	-
Male pedestrian	0.0766	0.0306	-	-	-	-	-	-	-	-
Driver Characteristics	-	-	-	-	-	-	-	-	-	-
Driver age: 45 - 64	-0.149	0.0317	-	-	-	-	-	-	-	-
Alcohol-impaired driver	0.6294	0.0922	-	-	-	-	-	-	-	-
Male driver	-	-	0.2725	0.0668	0.2161	0.0482	0.0823	0.0332	0.0844	0.0705
Crash characteristics	-	-	-	-	-	-	-	-	-	-
Ambulance rescue	-	-	0.5336	0.1008	1.0418	0.0795	1.0979	0.0408	1.9644	0.0749
Hit and run	-	-	0.8367	0.1896	0.8364	0.1464	0.1924	0.1195	-0.2469	0.2148
Backing Vehicle	-	-	-1.0992	0.2153	-0.7388	0.1289	-0.4029	0.0741	-0.184	0.1408
Dash/dart out	-	-	-0.1284	0.1045	0.1186	0.0719	0.5032	0.0574	0.1945	0.1403

Variable	All Level		K[d]		A[c]		B[b]		C[d]	
	Coef.	**S.E.**	**Coef.**	**S.E.**	**Coef.**	**S.E.**	**Coef.**	**S.E.**	**Coef.**	**S.E.**
Off roadway	-	-	-0.8279	0.2427	-0.4592	0.1427	-0.1296	0.078	-0.0911	0.1464
Walking along roadway	-	-	-0.676	0.0923	-0.6567	0.073	-0.4094	0.0598	-0.0277	0.1348
Locality and roadway Characteristics	-	-	-	-	-	-	-	-	-	-
Urban	-	-	-0.5349	0.0775	-0.3417	0.0604	-0.0777	0.0466	0.1866	0.0913
Curved roadway	-	-	0.4433	0.1145	0.5031	0.0904	0.2483	0.0814	-0.0473	0.1779
Two-way, divided	0.378	0.0649	-	-	-	-	-	-	-	-
Two-way, not divided	0.1198	0.0566	-	-	-	-	-	-	-	-
Commercial	-0.0947	0.0354	-	-	-	-	-	-	-	-
Farms, Woods, Pastures	0.1578	0.0621	-	-	-	-	-	-	-	-
Bottom-road	0.4923	0.1698	-	-	-	-	-	-	-	-
Grade-road	0.1305	0.0439	-	-	-	-	-	-	-	-
Interstate	0.4832	0.1495	-	-	-	-	-	-	-	-
Local street	-0.6718	0.0605	-	-	-	-	-	-	-	-
NC route	-0.2036	0.0812	-	-	-	-	-	-	-	-
Public vehicular area	-	-	-1.548	0.2428	-1.3172	0.1421	-1.1049	0.0827	-0.5981	0.14
State secondary route	-0.2613	0.0719	-	-	-	-	-	-	-	-
Time and Environment characteristics	-	-	-	-	-	-	-	-	-	-
Dark - lighted roadway	-	-	0.6796	0.0915	0.5024	0.0631	0.2504	0.0434	-0.0767	0.0918
Dark - roadway not lighted	-	-	1.003	0.084	0.7773	0.0634	0.2998	0.0548	-0.3125	0.1215
Traffic control characteristics and workzone	-	-	-	-	-	-	-	-	-	-
Human control	-0.5537	0.132	-	-	-	-	-	-	-	-
Traffic sign	-0.4276	0.0585	-	-	-	-	-	-	-	-
Traffic signal	-	-	-0.679	0.1123	-0.4904	0.0752	-0.2652	0.0501	-0.0906	0.1094

K[a] - Fatal Injury

A[b] - Incapacitating Injury

B[c] - Non-incapacitating Injury

C[d] - Possible Injury

O[e] - No Injury

No. of observations: 17,480.

-2×Log-likelihood at convergence: 38,920.

-2×Log-likelihood (constant only): 43,916.

AIC: 39,076.

BIC: 39,682.

TABLE 4.6: Average Marginal Effects for Each Contributing Factors in the PPO Model

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Pedestrian Characteristics					
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	-0.0187	0.0076	0.0554	-0.0385	-0.0058
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	0.0092	0.0083	0.0178	-0.0278	-0.0075
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	0.0451	0.0370	0.0607	-0.1172	-0.0257
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	0.0356	0.0328	0.0571	-0.1029	-0.0225
Male pedestrian (1 if pedestrian is male; 0 otherwise)	0.0040	0.0037	0.0084	-0.0126	-0.0035
Driver Characteristics					
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	-0.0078	-0.0071	-0.0165	0.0244	0.0069
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	0.0406	0.0339	0.0544	-0.1062	-0.0227
Male driver (1 if driver is male; 0 otherwise)	0.0142	0.0075	-0.0045	-0.0135	-0.0039
Crash characteristics					
Ambulance rescue (1 if service presents; 0 otherwise)	0.0252	0.0623	0.1507	-0.1228	-0.1154
Hit and run (1 if crash is hit-and-run; 0 otherwise)	0.0578	0.0461	-0.0637	-0.0526	0.0123
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	-0.0423	-0.0209	-0.0222	0.0765	0.0088
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	-0.0066	0.0189	0.0931	-0.0970	-0.0083
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	-0.0346	-0.0077	0.0148	0.0231	0.0042
Walking along roadway (1 if crash occurred when pedestrian is walking along roadway; 0 otherwise)	-0.0316	-0.0271	-0.0266	0.0840	0.0013
Locality and roadway Characteristics					
Urban (1 if crash occurs in urban roadway; 0 otherwise)	-0.0299	-0.0064	0.0199	0.0252	-0.0088
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	0.0268	0.0307	-0.0057	-0.0541	0.0022
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	0.0216	0.0192	0.0386	-0.0638	-0.0156
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	0.0063	0.0057	0.0131	-0.0195	-0.0056
Commercial (1 if crash occurred in commercial area; 0 otherwise)	-0.0050	-0.0046	-0.0103	0.0157	0.0043

Variables	K[a]	A[b]	B[c]	C[d]	O[e]
Farms, Woods, Pastures (1 if crash occurred in areas of farms, woods, or pastures; 0 otherwise)	0.0087	0.0079	0.0166	-0.0263	-0.0068
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	0.0307	0.0259	0.0448	-0.0828	-0.0186
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	0.0072	0.0064	0.0138	-0.0217	-0.0057
Interstate (1 if crash occurred on interstate; 0 otherwise)	0.0301	0.0255	0.0441	-0.0814	-0.0183
Local street (1 if crash occurred on local street; 0 otherwise)	-0.0369	-0.0344	-0.0659	0.1061	0.0312
NC route (1 if crash occurred on NC route; 0 otherwise)	-0.0103	-0.0094	-0.0230	0.0327	0.0100
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	-0.0546	-0.0503	-0.1319	0.2064	0.0304
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	-0.0132	-0.0120	-0.0294	0.0416	0.0130
Time and Environment characteristics					
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	0.0408	0.0139	-0.0021	-0.0561	0.0036
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	0.0625	0.0283	-0.0274	-0.0789	0.0156
Traffic control characteristics and workzone					
Human control (1 if the type of traffic control is human control; 0 otherwise)	-0.0246	-0.0238	-0.0673	0.0847	0.0310
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0201	-0.0192	-0.0507	0.0676	0.0223
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0309	-0.0149	-0.0100	0.0515	0.0042

K[a] - Fatal Injury

A[b] - Incapacitating Injury

B[c] - Non-incapacitating Injury

C[d] - Possible Injury

O[e] - No Injury

4.3.2.2. Results Interpretations of Partial Proportional Odds Model

As mentioned in Subsection 2.3.4, the sign(s) of the estimation(s) does (do) not always represent the direction(s) of the effect(s) on the intermediate outcomes. Hence marginal effects of each variable are computed for further interpretation and used to show how key factors affect the injury severity levels in pedestrian–vehicle crashes. With

categorical independent variables, marginal effects measure discrete change. It should be pointed out that the interpretations of the marginal effects are the same due to the same calculation process compared with both MNL and ML models. Thus, the interpretation of the corresponding factors with their effects could refer to MNL and ML models.

By examining all results from three models, some inconsistencies could be observed in term of the contributing factors identified by each model and also the associated marginal effects towards each injury severity level. Subsection 4.4.2 provides more detailed comparisons in these two phenomena.

4.4. Brief Comparisons Between Basic and Advanced Discrete Choice Models

4.4.1. Model Comparison Criteria of Discrete Choice Models

Since model structures of the three conventional DCMs in this dissertatiom are different, the Akaike's information criterion (*AIC*) and Bayesian information criterion (*BIC*) values of each model are also computed for comparisons. Both *AIC* and *BIC* take goodness of fit, prediction accuracy, and the number of significant variables into consideration. Many researches have shown that both *AIC* and *BIC* work well as a measure of goodness of fit (Abdel-Aty and Radwan, 2000; Sasidharan and Menendez, 2014; Cafiso et al., 2010; Ma et al., 2016). The *AIC* is computed as shown in Equation 4.4.1.1:

$$AIC = 2k - 2\,ln(L) \qquad (4.4.1.1)$$

where k is the number of parameters in the model, and L means the maximum likelihood value of the fitted model. The *BIC* differs from the *AIC* in the penalty term, as presented below in Equation 4.4.1.2:

$$BIC = k\,ln(O) - 2\,ln(L) \qquad (4.4.1.2)$$

where O is the number of the observations. A model with smaller *AIC* and *BIC* values performs better than others.

4.4.2. Comparison Results

TABLE 4.7: Indicator for Model Comparison

Model	No. of Obs (*O*)	No. of Vars (*k*)	-2ln(*L*)	*AIC*	*BIC*
MNL	17,480	93	38,652	38,839	39,658
ML	17,480	103	38,596	38,803	39,602
PPO	17,480	55	38,983	39,093	39,520

Table 4.7 provides the summaries of indicators (-2×log-likelihood, *AIC* and *BIC*,) for all three models. Though the likelihood value might not be good to compare different model structures, this value of ML is still the largest. Additionally, the *AIC* value of the ML model is the smallest among all models. The result of *BIC* values is different in which PPO ranks the first, followed by the ML model and then the MNL model. These three measures indicate that, given the same data set, the ML model yields the highest likelihood, and the smallest *AIC* value, when PPO model has the smallest *BIC* value. In summary, the ML model performs better than the PPO and MNL models for modeling the pedestrian injury severity in pedestrian-vehicle crashes in this dissertation.

Since the ML model has been built based on the MNL model, the number of statistically significant contributing factors identified in both models might be the same in most cases. However, as mentioned in Subsection 4.3.2, inconsistencies in terms of the identified contributing factors do exist while comparing PPO model with both MNL and ML models. For examples, the factors of “pedestrian walking along roadway” and “land

development: farms, woods, pastures" are found to be significant in PPO model, but not in MNL and ML models. Table 4.8 provides a summary of different significant contributing factors sets with associating marginal effects towards levels "K" and "A" in each model.

TABLE 4.8 Different Factors and Marginal Effects to K and A Levels

Variables	MNL Model		ML Model		PPO Model	
	K[a]	A[b]	K[a]	A[b]	K[a]	A[b]
Pedestrian Characteristics						
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	-0.0063	0.0077	-0.0043	0.0066	-0.0187	0.0076
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	0.0263	0.0149	0.0229	0.0140	0.0092	0.0083
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	0.0951	0.0132	0.0935	0.0162	0.0451	0.0370
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	0.0408	0.0260	0.0374	0.0251	0.0356	0.0328
Male pedestrian (1 if pedestrian is male; 0 otherwise)	-0.0058	0.0087	-0.0043	0.0078	0.0040	0.0037
Driver Characteristics						
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	-0.0087	-0.0057	-0.0076	-0.0049	-0.0078	-0.0071
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	0.0506	0.0272	0.0427	0.0227	0.0406	0.0339
Male driver (1 if driver is male; 0 otherwise)	0.0146	0.0077	0.0141	-0.0177	0.0142	0.0075
Crash characteristics						
Ambulance rescue (1 if service presents; 0 otherwise)	0.0269	0.0611	0.0129	0.0496	0.0252	0.0623
Hit and run (1 if crash is hit-and-run; 0 otherwise)	0.0559	0.0452	0.0459	0.0406	0.0578	0.0461
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	-0.0201	0.0035	-0.0160	0.0027	-0.0423	-0.0209
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	0.0337	0.0280	0.0292	0.0257	-	-

Variables	MNL Model		ML Model		PPO Model	
	K[a]	A[b]	K[a]	A[b]	K[a]	A[b]
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	0.0324	0.0496	0.0294	0.0420	-0.0066	0.0189
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	0.0652	-0.0102	0.0598	-0.0089	-	-
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	-0.0165	0.0252	-0.0123	0.0214	-	-
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	-0.0051	-0.0064	-0.0012	-0.0020	-0.0346	-0.0077
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	0.0511	0.0284	0.0352	0.0283	-	-
Walking along roadway (1 if crash occurred when pedestrian is walking along roadway; 0 otherwise)	-	-	-	-	-0.0316	-0.0271
Locality and roadway Characteristics						
Urban (1 if crash occurs in urban roadway; 0 otherwise)	-0.0233	-0.0087	-0.0331	-0.0065	-0.0299	-0.0064
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	0.0211	0.0362	0.0208	0.0324	0.0268	0.0307
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	0.0488	0.0272	0.0381	0.0249	0.0216	0.0192
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	0.0183	0.0185	0.0125	0.0182	0.0063	0.0057
Commercial (1 if crash occurred in commercial area; 0 otherwise)	0.0029	0.0036	0.0025	0.0040	-0.0050	-0.0046
Farms, Woods, Pastures (1 if crash occurred in areas of farms, woods, or pastures; 0 otherwise)	-	-	-	-	0.0087	0.0079
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	-0.0008	-0.0008	-0.0005	-0.0006		
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	0.0237	-0.0264	0.0150	-0.0234	0.0307	0.0259
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	0.0170	0.0096	-0.0129	0.0133	0.0072	0.0064

Variables	MNL Model		ML Model		PPO Model	
	K[a]	A[b]	K[a]	A[b]	K[a]	A[b]
Interstate (1 if crash occurred on interstate; 0 otherwise)	0.0340	-0.0056	0.0339	-0.0053	0.0301	0.0255
Local street (1 if crash occurred on local street; 0 otherwise)	-0.0619	-0.0205	-0.0514	-0.0187	-0.0369	-0.0344
NC route (1 if crash occurred on NC route; 0 otherwise)	-0.0144	0.0109	-0.0129	0.0102	-0.0103	-0.0094
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	-0.0081	0.0719	-0.0063	0.0690	-	-
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	-0.0614	-0.0365	-0.0428	-0.0610	-0.0546	-0.0503
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	-0.0153	0.0027	-0.0136	0.0023	-0.0132	-0.0120
Time and Environment characteristics						
Morning (1 if crash occurred during morning; 0 otherwise)	0.0043	0.0054	0.0030	0.0048	-	-
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	0.0507	0.0197	0.0423	0.0179	0.0408	0.0139
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	0.0653	0.0314	0.0521	0.0290	0.0625	0.0283
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	0.0430	-0.0070	0.0380	-0.0059	-	-
Cloudy (1 if the weather is cloudy; 0 otherwise)	0.0004	0.0005	0.0003	0.0004	-	-
Rain (1 if the weather is raining; 0 otherwise)	-0.0193	0.0034	-0.0164	0.0028	-	-
Traffic control characteristics and workzone						
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	0.0006	0.0006	0.0004	0.0005	-	-
Human control (1 if the type of traffic control is human control; 0 otherwise)	-0.0426	-0.0296	-0.0320	-0.0230	-0.0246	-0.0238
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0202	-0.0325	-0.0169	-0.0262	-0.0201	-0.0192
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	-0.0293	-0.0198	-0.0234	-0.0180	-0.0309	-0.0149

K[a] - Fatal Injury

A[b] - Incapacitating Injury

Variables	MNL Model		ML Model		PPO Model	
	K[a]	A[b]	K[a]	A[b]	K[a]	A[b]
B[c] - Non-incapacitating Injury						
C[d] - Possible Injury						
O[e] - No Injury						

Despite the goodness-of-fit and the identified contributing factors, from Table 4.8 one can see that there are also some similarities and differences in the effects of the contributing factors, particularly towards fatality and incapacitating injuries of pedestrians in the pedestrian-vehicle crashes according to Tables 4.2, 4.4, and 4.6. For instances, the effects of "alcohol-impaired driver" towards fatality in all three models are quite closed to each other (marginal effects: 0.0506 in MNL, 0.427 in ML, and 0.406 in PPO), which denotes that about 5% of the probability increase for pedestrians being killed in the crashes. On the other hand, the factor of "male pedestrian" tends to mitigate the risk of pedestrians being killed in the MNL and ML models (marginal effects of -0.0058 and -0.0043, respectively), but tends to increase such risk in the PPO model (marginal effect of 0.0040). In other words, such differences might result from the ignorance of the unobserved heterogeneity in the crash data by the PPO model, when compared to ML model which allows further random effects of some factors to capture the unobserved heterogeneity.

All abovementioned comparisons tend to provide a brief view on the conventional DCMs. Chapter 7 will further provide a detailed comparison between all developed models in this study, including XGBoost model, with several other more interpretable, widely used and accepted measures in the field of multiclass classification problems, which are accuracy, precision (i.e., positive predictive value), recall (i.e., sensitivity), and F1 score.

4.5. Summary

In summary, this chapter provides the developments of one basic discrete choice model (i.e., MNL model) and two advanced discrete choice models (i.e., ML model and PPO model). Detailed estimation results with the corresponding marginal effects are also computed. Typical interpretations of the model results are illustrated to examine the effects of the identified contributing factors to pedestrian injury severity in the pedestrian-vehicle crashes. This tends to provide a general guidance on interpreting the results produced by conventional DCMs. Regarding comparisons between discrete choice models, a set of criteria have been introduced and the results show that ML outperforms PPO and MNL models in short. This is intended to provide a solid reference for future tasks in comparisons with developed machine learning approach.

CHAPTER 5: DEVELOPMENTS OF MACHINE LEARNING MODEL(S)

5.1. Introduction

Chapter 5 provides the developments of the selected advanced machine learning method (i.e., XGBoost method) for modeling the pedestrian injury severities in pedestrian-vehicle crashes based on the literatures presented in Section 2.4 by using the collected data. Detailed modeling results in the developed machine learning model and detailed analysis of results from the selected model is also presented in this chapter.

5.2. Modeling and Parameter Tuning

Basically, to model pedestrian injury severities in pedestrian-vehicle crashes is a multiclass classification problem in this study with much categorical information (variables) and multiple discrete outcomes (injury severities). Thus, the objective in the XGBoost model used in this study is “multi:softmax” or “multi:softprob”, which is designed to accomplish the multiclass classification problem.

As mentioned in Section 2.4, parameter optimization (also known as parameter tuning) of the XGBoost method is applied by adjusting the algorithm parameters to achieve better results in the pedestrian injury severity modeling. By reviewing several literatures (Cheng and Ma, 2015; Jun and Cheng, 2017; Zhang and Cheng, 2017), the following hyperparameters in the XGBoost model have been selected to be tuned to optimize the performance, associated with the searching ranges:

- Learning rate: a parameter relates to the learning steps at which the model learns the patterns of input data. Smaller value of this parameter leads to

slower calculation, while larger value results in non-convergence. [0.1, 0.2, 0.3]

- Max_depth: it denotes the maximum depth of a generated tree. Higher value of this parameter means more complexity of the model but more likely to overfit. [6, 10, 14, 18]
- N_estimators: this parameter presents the number of boosting trees, or the number of training iterations on the data. Too few trees result in under fitting, while too many trees lead to overfitting. [1, 50, 200, 500]
- Reg_lambda: this parameter could help to deal with the regularization part in the XGBoost model. [0.1, 0.2, 0.3, 0.4]
- Colsample_bytree: it is the percentage of features that are used per tree. [0.6, 0.7, 0.8, 0.9]

Though fully tuning of all hyperparameter works in theory, it might be impossible in practices. The selected parameters and their associated searching ranges have been pre-tested to be narrowed down in which the performance of the model is most sensitive to. In addition, by considering both time-efficiency and accuracy of the model performance, the randomized search method is utilized in this study with ten sets of random selected parameters by the algorithm. 5-fold cross-validation is used in the optimization process for a more stable result. Table 5.1 shows the tuning results by randomized search and the best set is ranked first in the table.

TABLE 5.1: Randomized Search Results for Hyperparameter Tuning in XGBoost Model

Reg_lambda	N_estimators	Max_depth	Learning_rate	Colsample_bytree	Mean accuracy	Rank
0.2	200	18	0.1	0.7	0.7444	1
0.4	200	10	0.3	0.9	0.7336	2
0.3	500	6	0.3	0.7	0.7210	3
0.4	500	6	0.3	0.7	0.7208	4
0.2	500	6	0.2	0.9	0.7126	5
0.4	50	10	0.2	0.7	0.6994	6
0.3	200	6	0.1	0.7	0.6393	7
0.3	1	14	0.3	0.9	0.6077	8
0.4	1	10	0.3	0.7	0.5004	9
0.4	1	6	0.2	0.8	0.4760	10

It should be noted that the accuracy is applied here to examine the performance of the model, which can be calculated as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \tag{5.2.1}$$

Accuracy gives an overall measurement of the model. Despite this straight forward measure, there are also other three measures at each outcome level (i.e., injury severity in this study) computed to examine the performance of the developed models (they are also applied to conventional DCMs and introduced in Chapter 6 with model comparisons), which are precisions, recall, and the F1 score and they can be computed as follows:

$$Precision_i = \frac{Number\ of\ correct\ predictions\ under\ injury\ severity\ i}{Total\ number\ of\ predictions\ under\ injury\ severity\ i} \tag{5.2.2}$$

$$Recall_i = \frac{Number\ of\ correct\ predictions\ under\ injury\ severity\ i}{Total\ number\ of\ actual\ cases\ injury\ severity\ i} \tag{5.3.3}$$

$$F1_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{5.2.3}$$

After the tuning process, the best selected parameter set is identified as: Reg_lambda = 0.2; N_estimators = 200; Max_depth = 18; Learning_rate = 0.1; Colsample_bytree = 0.7. In addition, the associated test accuracy is 74.44%. With the best

model by utilizing the parameters randomized search, the whole dataset is refitted. The overall accuracy of the best model on the whole dataset is 93.23%, which shows a relatively high performance. The associated precisions, recalls, and F1 scores are further presented in Chapter 6 for comparison with the conventional DCMs.

5.3. Variable Importance and Partial Dependence of Top 15 Contributing Factors

As stated in Subsection 2.4.3, the total gains of each contributing factor are calculated as the variable importance. Figure 5.1 shows the importance of all the contributing factors in the final XGBoost model. This provides an intuitive image to the important factors impacting the model structure and their contributions to the predicted outcomes of the model.

Additionally, as described in Section 2.4, the importance of a factor does not have the direct interpretation on its effect towards pedestrian injury severities and partial dependences should be used to examine how factors affect the pedestrian injury severities. Figure 5.2 shows the average partial dependence changes of top 15 contributing factors.

Despite the top 15 factors, the changes of the partial dependences of factors which show their impacts on increasing the risk of pedestrian sustaining severer injuries (i.e., "K" of fatality and "A" of incapacitating injury) are further computed for more in-depth analysis. Nine factors with such effects are identified and their associated average partial dependence changes are illustrated in the Figure 5.3. Factors such as alcohol involvements for both drivers and pedestrians are in line with results from most of the existing studies by applying the conventional DCMs. Similar to the interpretation of marginal effect

utilized in the conventional DCMs, the average partial dependence change of a factor has the same way to be examined.

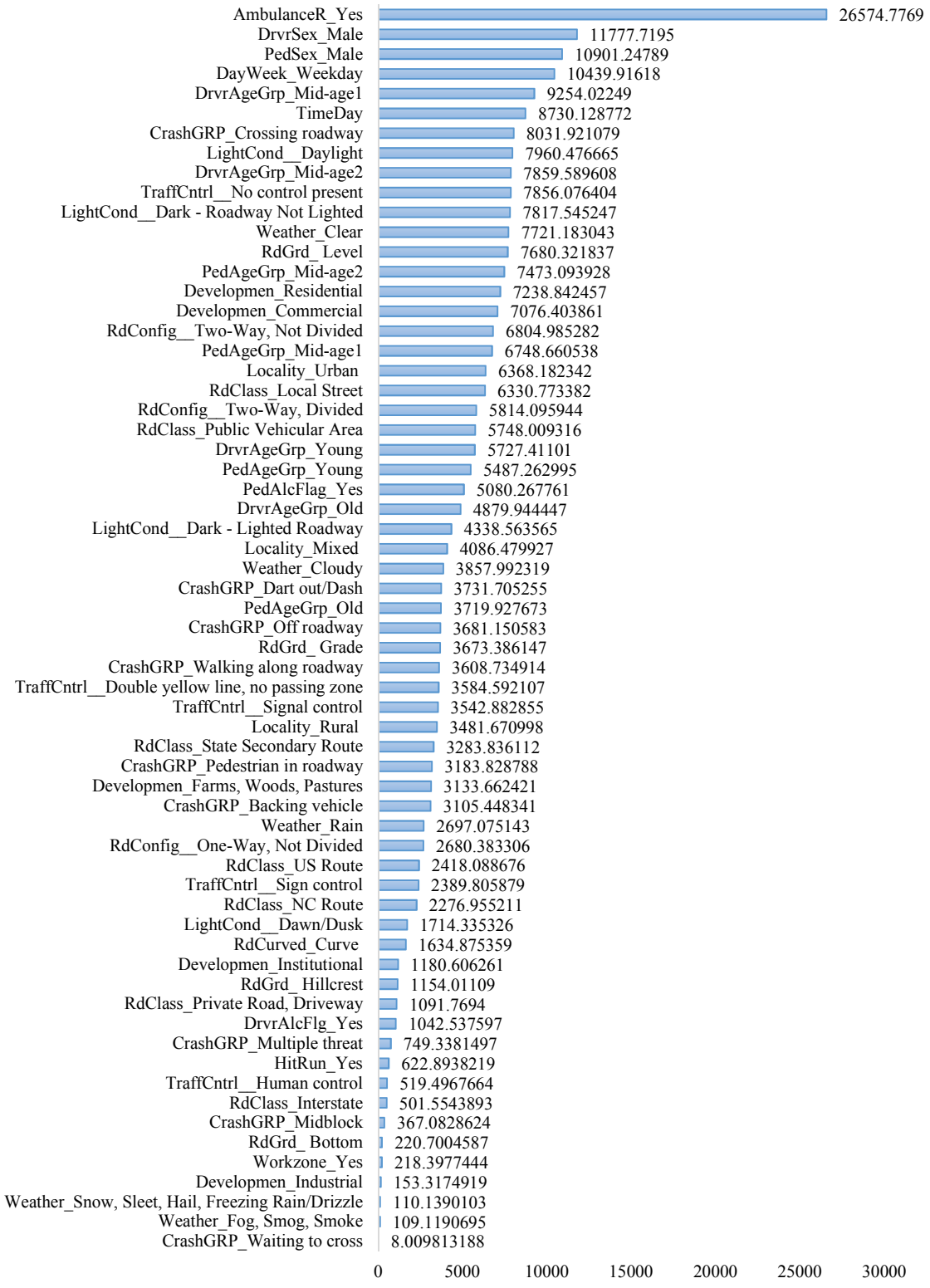

FIGURE 5.1: Importance of All Contributing Factors

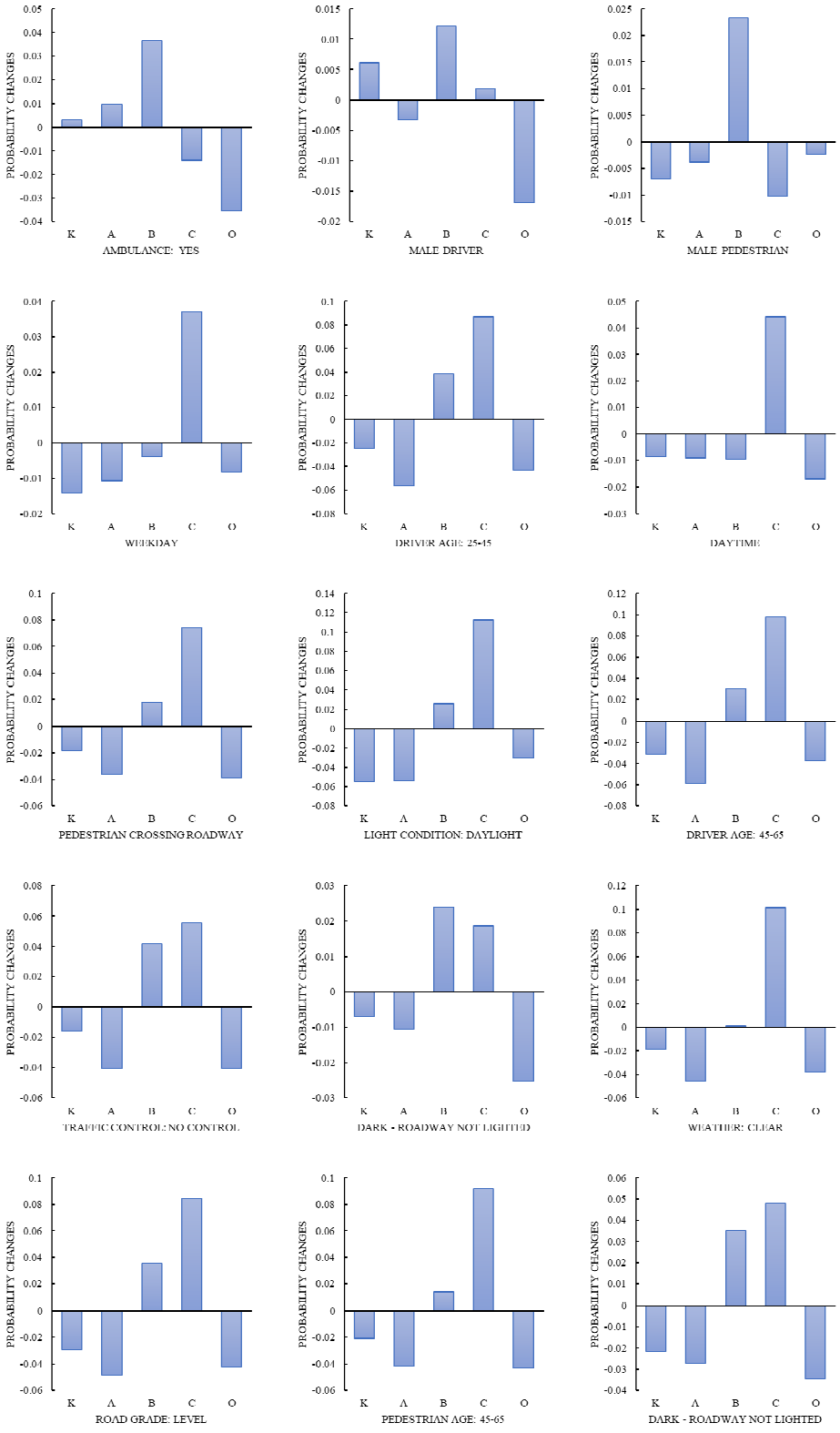

FIGURE 5.2: Average Partial Dependence Changes of Top 15 Contributing Factors

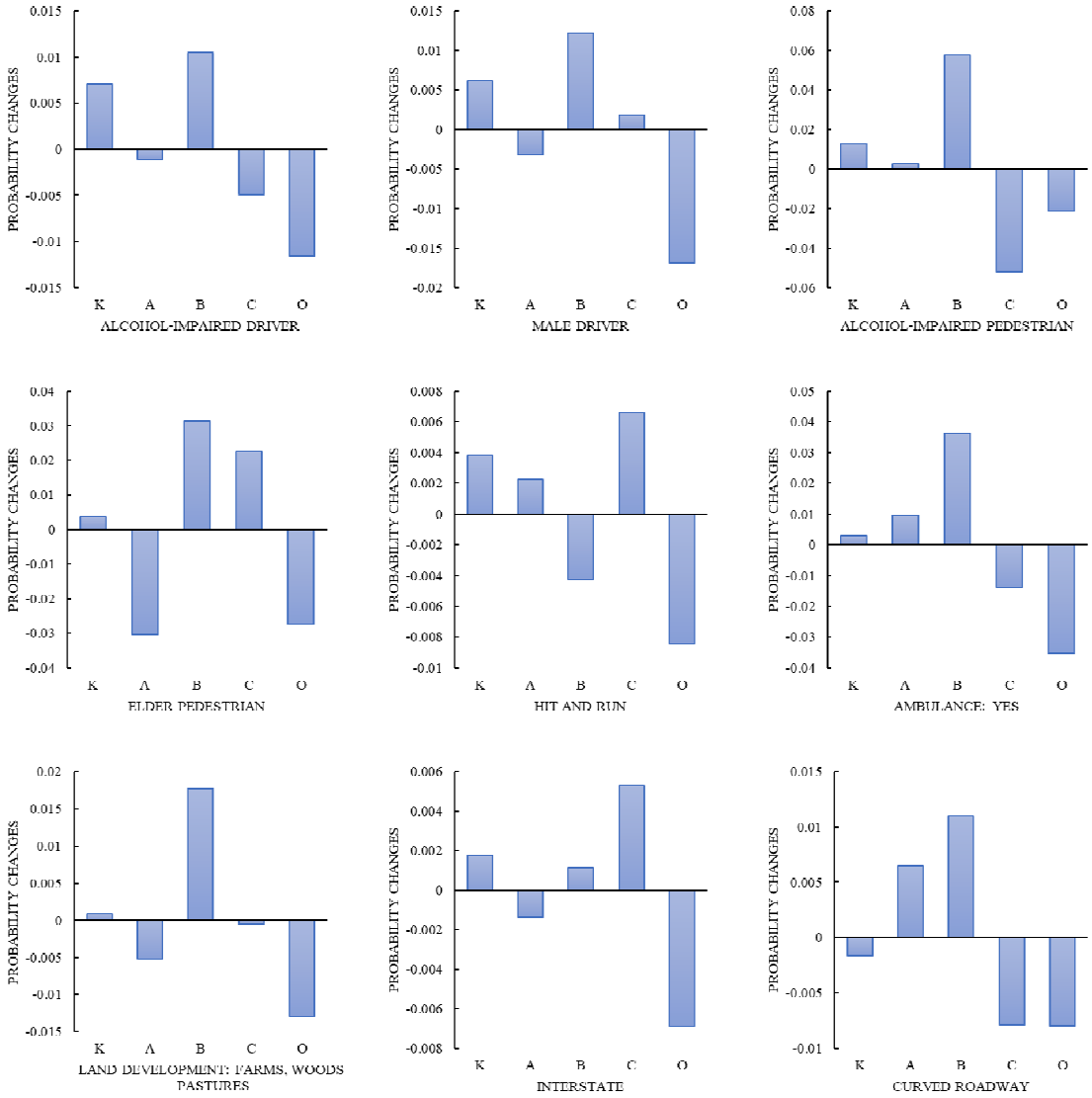

FIGURE 5.3: Average Partial Dependence Changes of Contributing Factors Increasing the Risk of Pedestrians Sustaining Severer Injuries (i.e., Fatality and Incapacitating Injury)

5.4. Summary

In summary, this chapter provides the developments of the selected machine learning method, which is the XGBoost model. Designed optimization process with hyperparameter tuning to the XGBoost model is provided to obtain the best model structure in order to achieve a relatively high performance of the proposed XGBoost model. Detailed numerical results with the corresponding feature importance and partial dependences of the contributing factors are also computed. Typical interpretations of the model results are presented to examine the effects of the identified contributing factors to pedestrian injury severities in pedestrian-vehicle crashes.

CHAPTER 6: MODEL COMPARISONS

6.1. Introduction

Chapter 6 provides the comprehensive evolutions and comparison of all proposed models. Section 6.2 presents the performance comparison between conventional DCMs and the XGBoost model based on several performance measures (i.e., accuracy, precision, recall, and F1 score). Section 6.3 concludes this chapter with a summary

6.2. Model Comparisons

Despite the traditional model statistics used in the conventional DCMs, the comparisons with machine learning models could be very different. As introduced in Section 5.2, overall model accuracy, model precision, model recall, and model F1 score on each predicted outcome with more straightforward and interpretable meaning in showing the performance of a proposed model are utilized in this study.

TABLE 6.1: Predicted Results of XGBoost Model with Accuracy, Precisions, Recalls, and F1 Scores

	Fatality	**Incapacitating Injury**	**Non-incapacitating Injury**	**Possible Injury**	**No Injury**	**Actual Total**	**Recall**	**F1**
Fatality	1113	3	20	16	2	1154	96.45 %	96.36 %
Incapacitating Injury	7	1196	33	52	4	1292	92.57 %	92.14 %
Non-incapacitating Injury	21	41	6090	375	44	6571	92.68 %	93.33 %
Possible Injury	15	61	310	7070	97	7553	93.61 %	93.52 %
No Injury	0	3	26	53	828	910	90.99 %	87.85 %

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Predicted Total	1156	1304	6479	7566	975	17480	**Accuracy**	
Precision	96.28%	91.72%	94.00%	93.44%	84.92 %		93.23%	

TABLE 6.2: Predicted Results of MNL Model with Accuracy, Precisions, Recalls, and F1 Scores

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Fatality	306	1	669	178	0	1154	26.52%	32.83%
Incapacitating Injury	106	1	846	339	0	1292	0.08%	0.15%
Non-incapacitating Injury	196	3	3471	2901	0	6571	52.82%	49.84%
Possible Injury	85	1	2216	5251	0	7553	69.52%	61.93%
No Injury	17	0	156	737	0	910	0.00%	0.00%
Predicted Total	710	6	7358	9406	0	17480	**Accuracy**	
Precision	43.10%	16.67%	47.17%	55.83%	0.00%		51.65%	

TABLE 6.3: Predicted Results of ML Model with Accuracy, Precisions, Recalls, and F1 Scores

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Fatality	268	4	676	206	0	1154	23.22%	30.77%
Incapacitating Injury	84	5	838	365	0	1292	0.39%	0.76%
Non-incapacitating Injury	149	6	3362	3054	0	6571	51.16%	49.24%

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Possible Injury	71	3	2075	5404	0	7553	71.55%	62.32%
No Injury	16	0	134	760	0	910	0.00%	0.00%
Predicted Total	588	18	7085	9789	0	17480	**Accuracy**	
Precision	45.58%	27.78%	47.45%	55.20%	0.00%		51.71%	

TABLE 6.4: Predicted Results of PPO Model with Accuracy, Precisions, Recalls, and F1 Scores

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Fatality	262	0	747	145	0	1154	22.70%	30.27%
Incapacitating Injury	84	0	897	311	0	1292	0.00%	0.00%
Non-incapacitating Injury	151	0	3522	2898	0	6571	53.60%	49.63%
Possible Injury	66	0	2299	5188	0	7553	68.69%	61.64%
No Injury	14	0	158	738	0	910	0.00%	0.00%
Predicted Total	577	0	7623	9280	0	17480	**Accuracy**	
Precision	45.41%	0.00%	46.20%	55.91%	0.00%		51.33%	

Tables 6.1-6.4 present the predicted results for all proposed models in this study. It is obvious that the XGBoost model outperforms the other three conventional DCMs with more than 40% higher in the overall accuracy. It should be noted that the overall accuracy denotes total actual cases divided by the total correctly predicted cases and this value being about 50% in all three conventional DCMs means that these models fit the data with randomly predicted outcomes. Recalls and F1 scores of incapacitating injury and no injury

in all three conventional DCMs almost equal to 0.00%, implying that the conventional DCMs can hardly identify the corresponding patterns in these two injury levels. Such results show extremely bad fitting and predicting capabilities of the conventional DCMs with tasks of modeling and analyzing the pedestrian injury severities. From the other hand, results of all three measurements (i.e., precisions, recalls, and F1 scores) in XGBoost model do provide a much better performance in fitting the pedestrian-vehicle crash data.

Furthermore, unlike conventional DCMs, since no variables have been excluded in the model development, with the best XGBoost model structure (more than 90% accuracy), one is able to examine the effects of all existing contributing factors towards each injury severity level for pedestrians in the pedestrian-vehicle crashes. More information could be extracted from the results to further support policymakers with more accurate and targeted safety improvement plans to help pedestrians in the transportation system.

6.3. Summary

This chapter illustrates detailed comparisons between conventional DCMs and the proposed XGBoost model with several interpretable, widely used and accepted measures in the field of multiclass classification problems, which are accuracy, precision (i.e., positive predictive value), recall (i.e., sensitivity), and F1 score. Results show that the XGBoost model outperforms the developed conventional DCMs in all measurements. And with more comprehensive results in terms of the coverage on the contributing factors, XGBoost model would be able to provide more information in guiding the plans of safety improvements to pedestrians in the transportation system.

CHAPTER 7: EMERGING HOTSPOTS ANALYSIS AND XGBOOST FOR MODELING PEDESTRIAN INJURY

7.1. Introduction

Chapter 6 has already proved the superiority of the proposed XGBoost model in modeling and analyzing the pedestrian injury severities in pedestrian-vehicle crashes. In order to better understand the spatiotemporal distributions of the pedestrian-vehicle crashes, this chapter provides a framework to combine the emerging hotspot analysis with XGBoost model for more explorations. Section 7.2 briefly introduces the emerging hotspot analysis technique and presents the associated results of the used crash data in this study. Then Section 7.3 provides the development and numerical results of XGBoost model by using the data retrieved from the emerging hotspot analysis. Section 7.4 discusses the corresponding model results in a manner similar to what has been presented in Chapter 5. Finally, Section 7.5 summarizes the whole chapter.

7.2. Emerging Hotspot Analysis

This study utilizes the emerging hotspot analysis tool in ArcGIS Pro to examine the spatiotemporal distribution patterns of single-pedestrian-single-vehicle crashes across the whole State of North Carolina based on aggregated crash data with grid size of 5000ft×5000ft. Figure 7.1 shows such aggregated crash density spatial distribution with a grid size of 5000ft×5000ft. In emerging hotspot analysis, a three-dimensional analysis integrating both temporal and spatial clusters could be achieved.

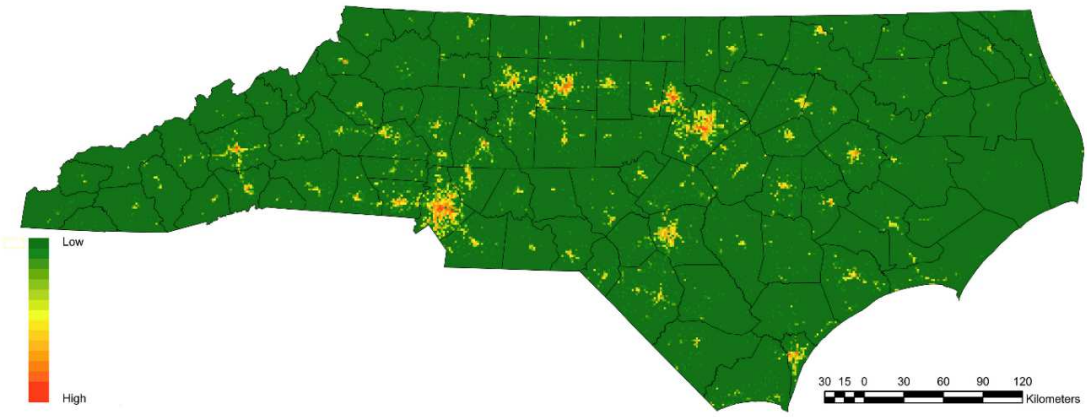

FIGURE 7.1: Crash Density Spatial Distribution

The first step is to create a space-time cube from the defined location. As seen in Figure 7.2, the tool generates a three-dimensional cube as the 12-year (2007–2018) temporal trend of the crash frequency is determined when the location of each grid is of fixed size.

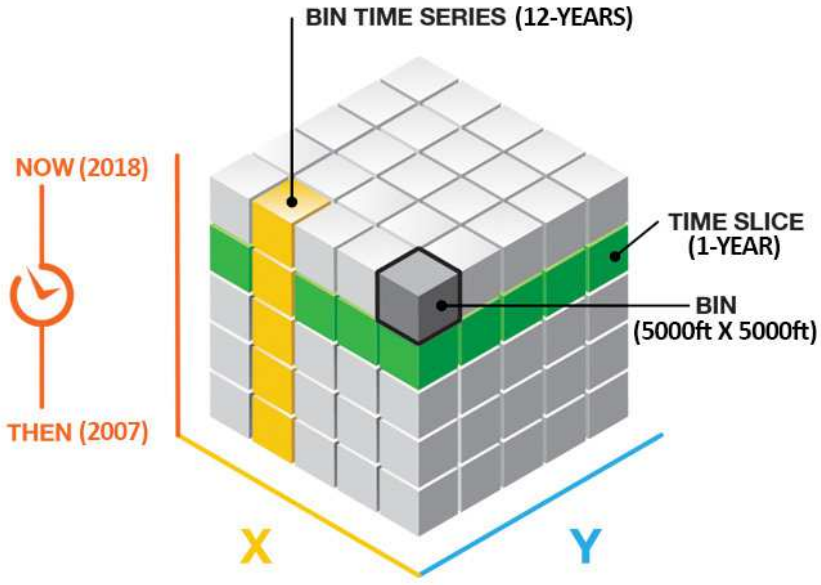

FIGURE 7.2: Space-Time Cube Used in This Study (Revised from Esri, ArcGIS.com)

With the cube generated, the Getis-Ord Gi* statistic (Getis and Ord, 2010) is then calculated to categorize those values (i.e., z-scores and p-values) in high or low values based on the given bin relative to its neighbor bins within the clustered cubes (Betty et al., 2020). The equation of Getis-Ord Gi* index can be expressed as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{X}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}} \quad (7.2.1)$$

where x_j is the attribute value for j^{th} bin, W_{ij} is the spatial weight between bins i and j (equals to 1 if j^{th} bin is within the spatiotemporal neighborhood distance of the i^{th} bin; equals to 0 otherwise); n is the number of total bins; and G_i^* is a z-score. It should be pointed out that both the z-scores and p-values indicate the time-series trends of each cube. After obtaining those values, the Mann–Kendall trend test has been utilized to identify the spatial and temporal patterns of each grid with the associating evaluation results of the trends.

Figure 7.3 displays the result of the emerging hotspot analysis and a zoom-in of the Mecklenburg County areas. Table 7.1 displays the descriptions and statistics for the spatiotemporal patterns of the associated single-pedestrian-single-vehicle crashes identified by the emerging hotspot analysis technique.

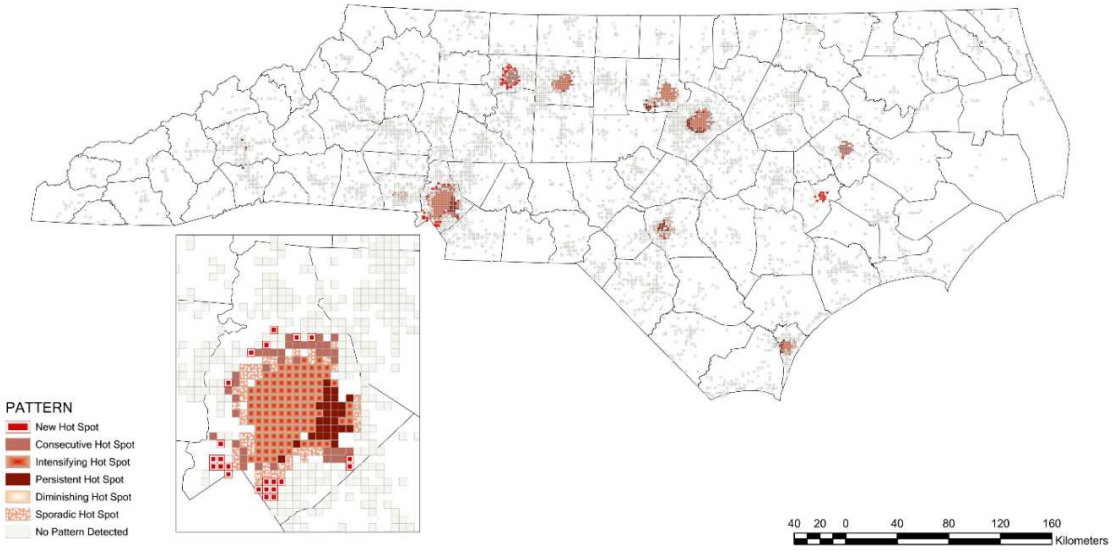

FIGURE 7.3: Spatiotemporal Patterns of Single-pedestrian-single Vehicle Crash Locations in North Carolina

TABLE 7.1: Descriptions and Statistics for Spatiotemporal Patterns of Single-pedestrian-single vehicle Crash Locations in North Carolina

Spatiotemporal patterns	**Description**	**Total bins**	**Total crashes**	**K[a]**	**A[b]**	**B[c]**	**C[d]**	**O[e]**
New Hot Spot	A location that is identified as a statistically significant hot spot only for the final year.	66	287	12	16	113	131	15
Intensifying Hot Spot	A location that has been identified as a statistically significant hot spot for 90% of all years, including the final year, with a statistically significant increase in the intensity of clustering of high counts over time.	301	4728	142	253	1759	2350	224
Persistent Hot Spot	A location that has been identified as a statistically significant hot spot for 90% of all years with no recognizable tendency showing an increase or decrease in the intensity of clustering over time.	95	900	34	55	347	414	50
Consecutive Hot Spot	A location with a single uninterrupted run of statistically significant hot spot bins in the final year. The location has never been a statistically significant hot spot prior to the final hot spot run and less than 90% of	154	854	32	57	353	375	37

Spatiotemporal patterns	Description	Total bins	Total crashes	K[a]	A[b]	B[c]	C[d]	O[e]
	all bins are statistically significant hot spots.							
Diminishing Hot Spot	A location that has been identified as a statistically significant hot spot for 90% of all years, including the final year with a statistically significant decrease in the intensity of clustering over time.	5	39	1	6	16	14	2
Sporadic Hot Spot	A location that is an on-again then off-again hot spot. Less than 90% of the years have been statistically significant hot spots and none of the years have been statistically significant cold spots.	208	1139	66	80	423	513	57
No Pattern Detected	Does not fall into any of the hot defined above	4490	9533	867	825	3560	3756	525

K[a] - Fatal Injury; A[b] - Incapacitating Injury; B[c] - Non-incapacitating Injury; C[d] - Possible Injury; O[e] - No Injury

From the descriptions of Table 7.1, locations (or bins) within categories of "New Hot Spot", "Intensifying Hot Spot", "Persistent Hot Spot", and "Consecutive Hot Spot" are the most targeted hotspots that need to be further focused and explored. Then the cases within these targeted hotspot areas have been filtered out and aggregated for modeling and analyzing. Table 7.2 shows the descriptive statistics of the explanatory variables used for the hotspot analysis.

TABLE 7.2: Descriptive Statistics of Explanatory Variable for Hotspots Dataset

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Pedestrian–vehicle Crashes	6769 (100%)	220 (3.25%)	381 (5.63%)	2572 (38.00%)	3270 (48.31%)	326 (4.82%)
Pedestrian Characteristics						
Pedestrian age: 25 - 44 (1 if pedestrian is younger than 45 years old and older than 24 years old; 0 otherwise) *	2076 (30.67%)	56 (0.83%)	110 (1.63%)	758 (11.20%)	1053 (15.56%)	99 (1.46%)
Pedestrian age: ≤ 24 (1 if pedestrian is younger than 25 years; 0 otherwise)	2127 (31.42%)	47 (0.69%)	117 (1.73%)	916 (13.53%)	942 (13.92%)	105 (1.55%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Pedestrian age: 45 - 64 (1 if pedestrian is younger than 65 years old and older than 44 years old; 0 otherwise)	2008 (29.66%)	95 (1.40%)	120 (1.77%)	672 (9.93%)	1024 (15.13%)	97 (1.43%)
Pedestrian age: ≥ 65 (1 if pedestrian is older than 64 years old; 0 otherwise)	558 (8.24%)	22 (0.33%)	34 (0.50%)	226 (3.34%)	251 (3.71%)	25 (0.37%)
Alcohol-impaired pedestrian (1 if pedestrian is alcohol-impaired; 0 otherwise)	751 (11.09%)	94 (1.39%)	97 (1.43%)	353 (5.21%)	179 (2.64%)	28 (0.41%)
Male pedestrian (1 if pedestrian is male; 0 otherwise)	3746 (55.34%)	155 (2.29%)	255 (3.77%)	1485 (21.94%)	1673 (24.72%)	178 (2.63%)
Driver Characteristics						
Driver age: 25 - 44 (1 if driver is younger than 45 years old and older than 24 years old; 0 otherwise) *	2632 (38.88%)	95 (1.40%)	161 (2.38%)	991 (14.64%)	1267 (18.72%)	118 (1.74%)
Driver age: ≤ 24 (1 if driver is younger than 25 years; 0 otherwise)	1247 (18.42%)	37 (0.55%)	81 (1.20%)	514 (7.59%)	550 (8.13%)	65 (0.96%)
Driver age: 45 - 64 (1 if driver is younger than 65 years old and older than 44 years old; 0 otherwise)	2078 (30.70%)	69 (1.02%)	103 (1.52%)	747 (11.04%)	1053 (15.56%)	106 (1.57%)
Driver age: ≥ 65 (1 if driver is older than 64 years old; 0 otherwise)	812 (12.00%)	19 (0.28%)	36 (0.53%)	320 (4.73%)	400 (5.91%)	37 (0.55%)
Alcohol-impaired driver (1 if driver is alcohol-impaired; 0 otherwise)	144 (2.13%)	22 (0.33%)	19 (0.28%)	54 (0.80%)	44 (0.65%)	5 (0.07%)
Male driver (1 if driver is male; 0 otherwise)	3684 (54.42%)	150 (2.22%)	224 (3.31%)	1392 (20.56%)	1750 (25.85%)	168 (2.48%)
Crash characteristics						
Ambulance rescue (1 if service presents; 0 otherwise)	5158 (76.20%)	187 (2.76%)	353 (5.21%)	2197 (32.46%)	2317 (34.23%)	104 (1.54%)
Hit and run (1 if crash is hit-and-run; 0 otherwise)	145 (2.14%)	10 (0.15%)	16 (0.24%)	40 (0.59%)	67 (0.99%)	12 (0.18%)
Backing Vehicle (1 if crash occurred when driver is backing vehicle; 0 otherwise)	715 (10.56%)	4 (0.06%)	19 (0.28%)	185 (2.73%)	455 (6.72%)	52 (0.77%)
Crossing roadway (1 if crash happened when pedestrian is crossing roadway; 0 otherwise)	3451 (50.98%)	134 (1.98%)	209 (3.09%)	1326 (19.59%)	1634 (24.14%)	148 (2.19%)
Dash/dart out (1 if pedestrian movement preceding crash is dashing/darting out; 0 otherwise)	892 (13.18%)	33 (0.49%)	85 (1.26%)	477 (7.05%)	274 (4.05%)	23 (0.34%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Midblock (1 if crash happened when pedestrian is crossing at mid-block location; 0 otherwise)	42 (0.62%)	0 (0.00%)	0 (0.00%)	14 (0.21%)	26 (0.38%)	2 (0.03%)
Multiple-threat (1 if crash is a multiple-threat crash; 0 otherwise)	174 (2.57%)	2 (0.03%)	9 (0.13%)	89 (1.31%)	65 (0.96%)	9 (0.13%)
Off roadway (1 if pedestrian move off the roadway when vehicle approach; 0 otherwise)	915 (13.52%)	5 (0.07%)	25 (0.37%)	249 (3.68%)	569 (8.41%)	67 (0.99%)
Pedestrian in roadway (1 if pedestrian is in the roadway; 0 otherwise)	332 (4.90%)	31 (0.46%)	22 (0.33%)	122 (1.80%)	141 (2.08%)	16 (0.24%)
Waiting to cross (1 if crash occurred when pedestrian is waiting to cross the roadway; 0 otherwise) *	6 (0.09%)	1 (0.01%)	1 (0.01%)	3 (0.04%)	1 (0.01%)	0 (0.00%)
Walking along roadway (1 if crash occurred when pedestrian is walking along roadway; 0 otherwise)	242 (3.58%)	10 (0.15%)	11 (0.16%)	107 (1.58%)	105 (1.55%)	9 (0.13%)
Locality and roadway Characteristics						
Mixed (1 if crash occurs in mixed roadway; 0 otherwise) *	431 (6.37%)	9 (0.13%)	24 (0.35%)	179 (2.64%)	195 (2.88%)	24 (0.35%)
Rural (1 if crash occurs in rural roadway; 0 otherwise)	128 (1.89%)	12 (0.18%)	3 (0.04%)	59 (0.87%)	48 (0.71%)	6 (0.09%)
Urban (1 if crash occurs in urban roadway; 0 otherwise)	6210 (91.74%)	199 (2.94%)	354 (5.23%)	2334 (34.48%)	3027 (44.72%)	296 (4.37%)
Curved roadway (1 if road geometry is curved roadway; 0 otherwise)	206 (3.04%)	12 (0.18%)	28 (0.41%)	78 (1.15%)	78 (1.15%)	10 (0.15%)
One-way, not divided (1 if the road configuration is one-way not divided; 0 otherwise) *	671 (9.91%)	7 (0.10%)	18 (0.27%)	206 (3.04%)	395 (5.84%)	45 (0.66%)
Two-way, divided (1 if the road configuration is two-way divided; 0 otherwise)	1753 (25.90%)	111 (1.64%)	153 (2.26%)	756 (11.17%)	669 (9.88%)	64 (0.95%)
Two-way, not divided (1 if the road configuration is two-way not divided; 0 otherwise)	4345 (64.19%)	102 (1.51%)	210 (3.10%)	1610 (23.78%)	2206 (32.59%)	217 (3.21%)
Commercial (1 if crash occurred in commercial area; 0 otherwise)	4339 (64.10%)	142 (2.10%)	246 (3.63%)	1563 (23.09%)	2181 (32.22%)	207 (3.06%)
Farms, Woods, Pastures (1 if crash occurred in areas of farms, woods, or pastures; 0 otherwise)	50 (0.74%)	11 (0.16%)	1 (0.01%)	25 (0.37%)	10 (0.15%)	3 (0.04%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Industrial (1 if crash occurred in industrial area; 0 otherwise)	36 (0.53%)	0 (0.00%)	1 (0.01%)	15 (0.22%)	18 (0.27%)	2 (0.03%)
Institutional (1 if crash occurred in Institutional area; 0 otherwise)	334 (4.93%)	2 (0.03%)	11 (0.16%)	134 (1.98%)	153 (2.26%)	34 (0.50%)
Residential (1 if crash occurred in Residential area; 0 otherwise) *	2010 (29.69%)	65 (0.96%)	122 (1.80%)	835 (12.34%)	908 (13.41%)	80 (1.18%)
Bottom-road (1 if crash occurred at the bottom of the roadway; 0 otherwise)	39 (0.58%)	5 (0.07%)	1 (0.01%)	23 (0.34%)	10 (0.15%)	0 (0.00%)
Grade-road (1 if crash occurred on grade-road; 0 otherwise)	728 (10.75%)	36 (0.53%)	67 (0.99%)	272 (4.02%)	316 (4.67%)	37 (0.55%)
Hillcrest (1 if crash occurred at the hillcrest of the roadway; 0 otherwise)	272 (4.02%)	14 (0.21%)	17 (0.25%)	104 (1.54%)	121 (1.79%)	16 (0.24%)
Level (1 if crash occurred at level roadway; 0 otherwise) *	5730 (84.65%)	165 (2.44%)	296 (4.37%)	2173 (32.10%)	2823 (41.70%)	273 (4.03%)
Interstate (1 if crash occurred on interstate; 0 otherwise)	57 (0.84%)	17 (0.25%)	10 (0.15%)	14 (0.21%)	16 (0.24%)	0 (0.00%)
Local street (1 if crash occurred on local street; 0 otherwise)	4918 (72.65%)	171 (2.53%)	308 (4.55%)	2047 (30.24%)	2181 (32.22%)	211 (3.12%)
NC route (1 if crash occurred on NC route; 0 otherwise)	77 (1.14%)	7 (0.10%)	10 (0.15%)	37 (0.55%)	21 (0.31%)	2 (0.03%)
Private road, driveway (1 if crash occurred on driveway of private road; 0 otherwise)	95 (1.40%)	3 (0.04%)	3 (0.04%)	40 (0.59%)	47 (0.69%)	2 (0.03%)
Public vehicular area (1 if crash occurred on public vehicular area; 0 otherwise)	1477 (21.82%)	6 (0.09%)	39 (0.58%)	363 (5.36%)	961 (14.20%)	108 (1.60%)
State secondary route (1 if crash occurred on State secondary route; 0 otherwise)	45 (0.66%)	4 (0.06%)	5 (0.07%)	20 (0.30%)	14 (0.21%)	2 (0.03%)
US route (1 if crash occurred on US route; 0 otherwise) *	100 (1.48%)	12 (0.18%)	6 (0.09%)	51 (0.75%)	30 (0.44%)	1 (0.01%)
Time and Environment characteristics						
Weekday (1 if crash occurred during weekday; 0 otherwise)	5312 (78.48%)	153 (2.26%)	276 (4.08%)	1997 (29.50%)	2632 (38.88%)	254 (3.75%)
Morning (1 if crash occurred during morning; 0 otherwise)	4880 (72.09%)	87 (1.29%)	214 (3.16%)	1778 (26.27%)	2560 (37.82%)	241 (3.56%)
Dark - lighted roadway (1 if light condition is lighted roadway; 0 otherwise)	1754 (25.91%)	104 (1.54%)	154 (2.28%)	733 (10.83%)	692 (10.22%)	71 (1.05%)
Dark - roadway not lighted (1 if light condition is dark - roadway not lighted; 0 otherwise)	442 (6.53%)	52 (0.77%)	44 (0.65%)	188 (2.78%)	140 (2.07%)	18 (0.27%)

Variable	Total	K[a]	A[b]	B[c]	C[d]	O[e]
Dawn/dusk light (1 if light condition is dawn/dusk light; 0 otherwise)	314 (4.64%)	11 (0.16%)	13 (0.19%)	107 (1.58%)	169 (2.50%)	14 (0.21%)
Daylight (1 if light condition is daylight; 0 otherwise) *	4259 (62.92%)	53 (0.78%)	170 (2.51%)	1544 (22.81%)	2269 (33.52%)	223 (3.29%)
Clear (1 if the weather is clear; 0 otherwise) *	5142 (75.96%)	167 (2.47%)	292 (4.31%)	1985 (29.32%)	2451 (36.21%)	247 (3.65%)
Cloudy (1 if the weather is cloudy; 0 otherwise)	967 (14.29%)	40 (0.59%)	50 (0.74%)	344 (5.08%)	483 (7.14%)	50 (0.74%)
Fog, Smog, Smoke (1 if the weather is fog, smog, or smoke; 0 otherwise)	11 (0.16%)	0 (0.00%)	0 (0.00%)	1 (0.01%)	9 (0.13%)	1 (0.01%)
Rain (1 if the weather is raining; 0 otherwise)	630 (9.31%)	13 (0.19%)	38 (0.56%)	232 (3.43%)	319 (4.71%)	28 (0.41%)
Snow, Sleet, Hail, Freezing Rain/Drizzle (1 if the weather is snow, sleet, hail, freezing rain, or drizzle; 0 otherwise)	19 (0.28%)	0 (0.00%)	1 (0.01%)	10 (0.15%)	8 (0.12%)	0 (0.00%)
Traffic control characteristics and workzone						
Double yellow line, no passing zone (1 if crash occurs within no passing zone with double yellow line; 0 otherwise)	72 (1.06%)	7 (0.10%)	6 (0.09%)	36 (0.53%)	19 (0.28%)	4 (0.06%)
Workzone (1 if crash on work-zone related road segment; 0 otherwise)	60 (0.89%)	0 (0.00%)	1 (0.01%)	18 (0.27%)	34 (0.50%)	7 (0.10%)
Human control (1 if the type of traffic control is human control; 0 otherwise)	4225 (62.42%)	159 (2.35%)	265 (3.91%)	1599 (23.62%)	1999 (29.53%)	203 (3.00%)
No control present (1 if there is no control present; 0 otherwise) *	596 (8.80%)	10 (0.15%)	20 (0.30%)	183 (2.70%)	346 (5.11%)	37 (0.55%)
Traffic sign (1 if the type of traffic control is traffic sign; 0 otherwise)	1816 (26.83%)	44 (0.65%)	89 (1.31%)	736 (10.87%)	872 (12.88%)	75 (1.11%)
Traffic signal (1 if the type of traffic control is traffic sign; 0 otherwise)	71 (1.05%)	2 (0.03%)	4 (0.06%)	26 (0.38%)	37 (0.55%)	2 (0.03%)

K[a] - Fatal Injury

A[b] - Incapacitating Injury

B[c] - Non-incapacitating Injury

C[d] - Possible Injury

O[e] - No Injury

7.3. Modeling and Parameter Tuning for Hotspot Data

The same hyperparameter set, tuning criteria, and procedures as what has been presented in Section 5.2 are used in the hotspot dataset. Table 7.3 shows the tuning results by randomized search and the best set is ranked first in the table.

TABLE 7.3: Randomized Search Results for Hyperparameter Tuning in XGBoost Model for Hotspot Data

Reg_lambda	N_estimators	Max_depth	Learning_rate	Colsample_bytree	Mean accuracy	Rank
0.2	200	18	0.1	0.7	0.7968	1
0.4	200	10	0.3	0.9	0.7844	2
0.3	500	6	0.3	0.7	0.7830	3
0.4	50	10	0.2	0.7	0.7810	4
0.2	500	6	0.2	0.9	0.7795	5
0.4	500	6	0.3	0.7	0.7788	6
0.3	200	6	0.1	0.7	0.7330	7
0.3	1	14	0.3	0.9	0.6924	8
0.4	1	10	0.3	0.7	0.5541	9
0.4	1	6	0.2	0.8	0.4555	10

The associated test accuracy is 79.68%. With the best model by utilizing the parameters randomized search, the hotspot dataset is refitted. The overall accuracy of the best model on the hotspot dataset is 94.49%, which shows a relatively high performance. The associated precisions, recalls, and F1 scores are presented in Table 7.4.

TABLE 7.4: Predicted Results of XGBoost Model for Hotspot Data with Accuracy, Precisions, Recalls, and F1 Scores

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Fatality	209	0	6	5	0	220	95.00 %	96.09 %

	Fatality	Incapacitating Injury	Non-incapacitating Injury	Possible Injury	No Injury	Actual Total	Recall	F1
Incapacitating Injury	2	346	12	21	0	381	90.81 %	91.29 %
Non-incapacitating Injury	2	12	2426	127	5	2572	94.32 %	94.32 %
Possible Injury	2	19	119	3120	10	3270	95.41 %	95.05 %
No Injury	0	0	9	22	295	326	90.49 %	92.77 %
Predicted Total	215	377	2572	3295	310	6769	**Accuracy**	
Precision	97.21%	91.78%	94.32%	94.69%	95.16 %		94.49%	

7.4. Variable Importance and Partial Dependence of Top 15 Contributing Factors for Hotspot Data

Same as Section 5.4, the total gains of each contributing factor are calculated as the variable importance. Figure 7.4 shows the relative importance of all the contributing factors in the final XGBoost model developed for the hotspot data. The partial dependences are also further used to examine how factors affect the pedestrian injury severities for the hotspot data. Figure 7.5 shows the average partial dependence changes of top 15 contributing factors. It could be observed that the ranking of factor importance in the XGBoost model that uses the hotspot dataset is different from the one in the XGBoost model which utilizes the whole dataset. However, most of the contributing factors are relatively stable in terms of total gains and their rankings in both models with different datasets.

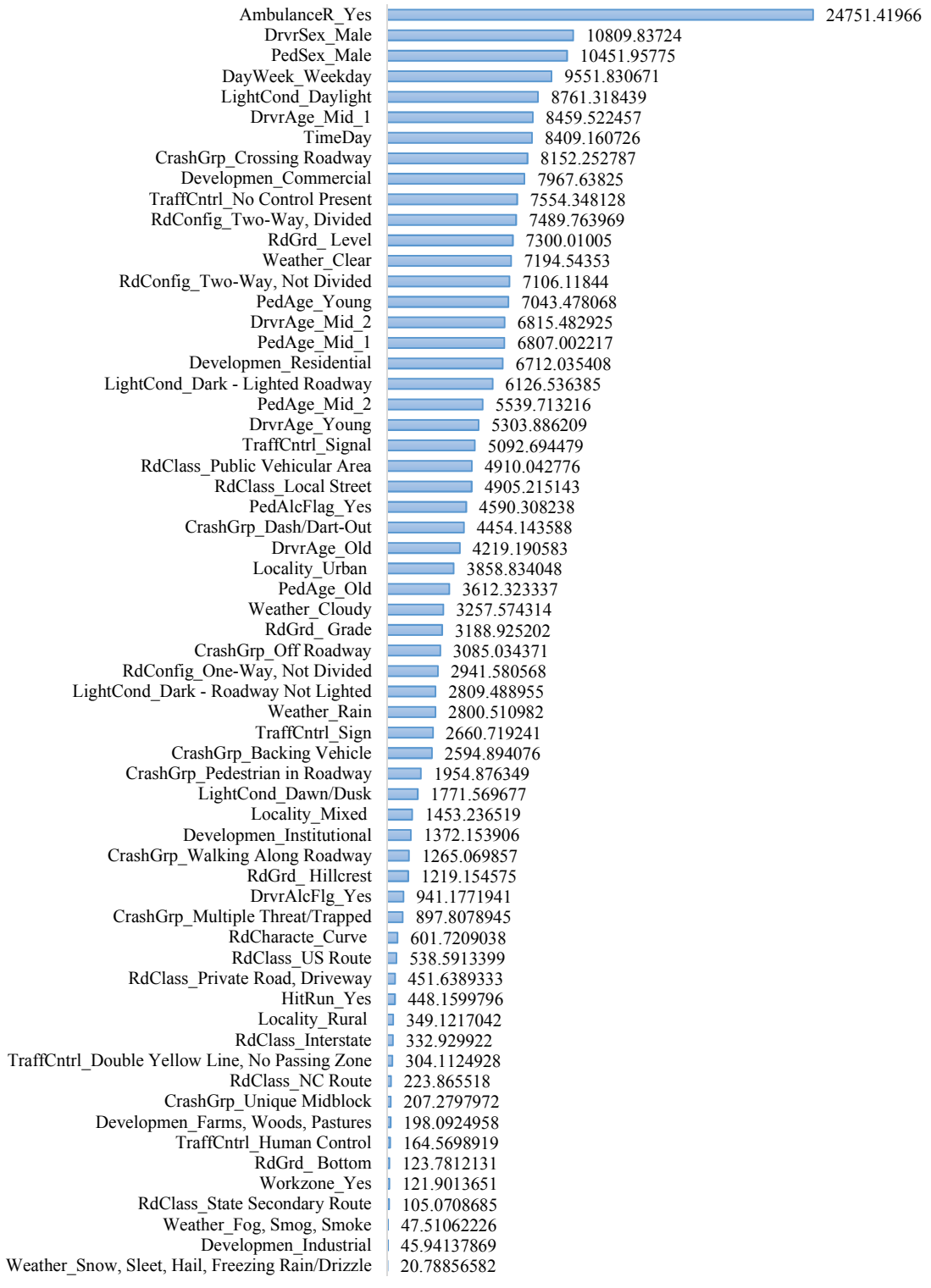

FIGURE 7.4: Importance of All Contributing Factors in the Hotspot Data

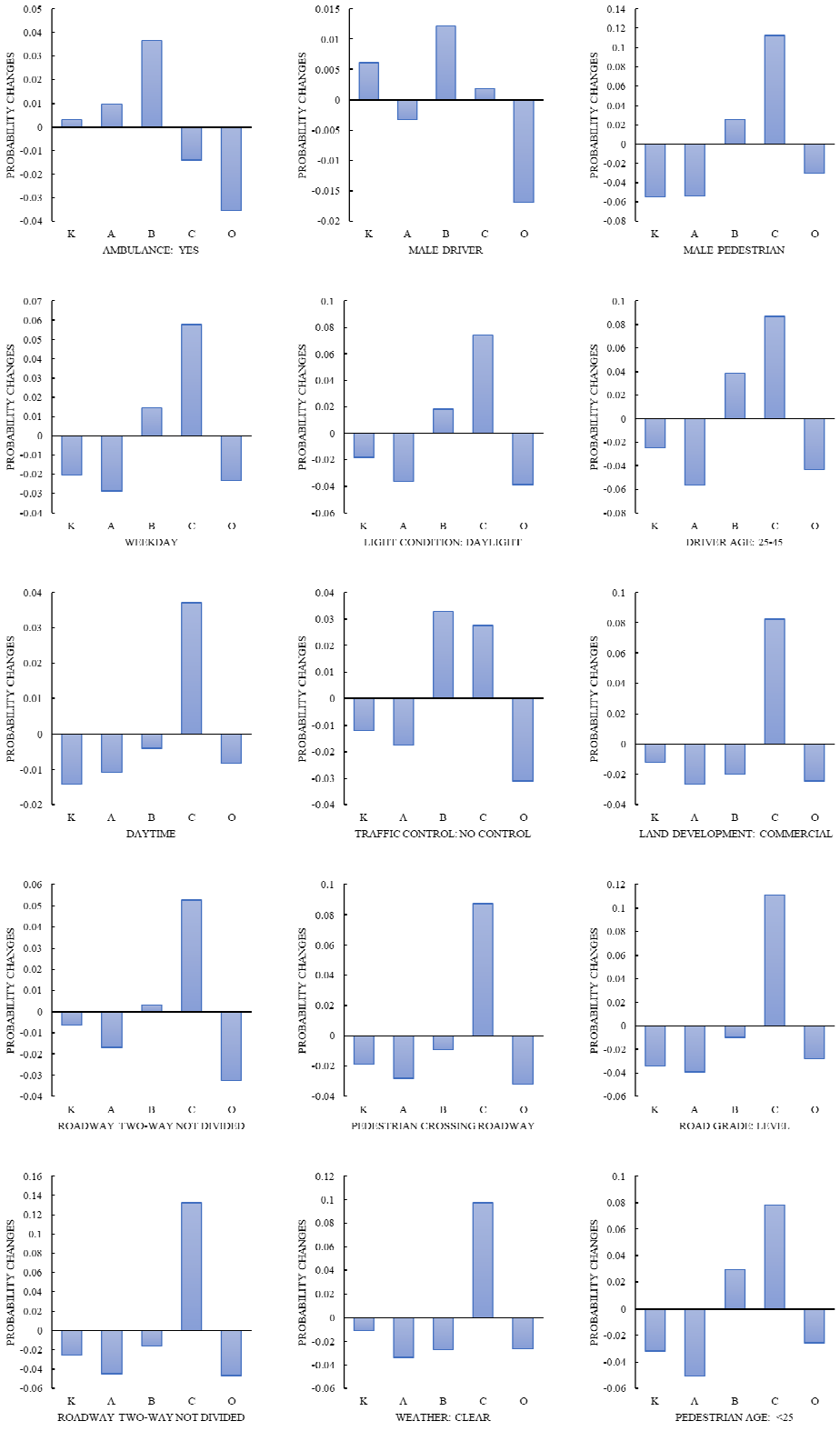

FIGURE 7.5: Average Partial Dependence Changes of Top 15 Contributing Factors in the Hotspot Data

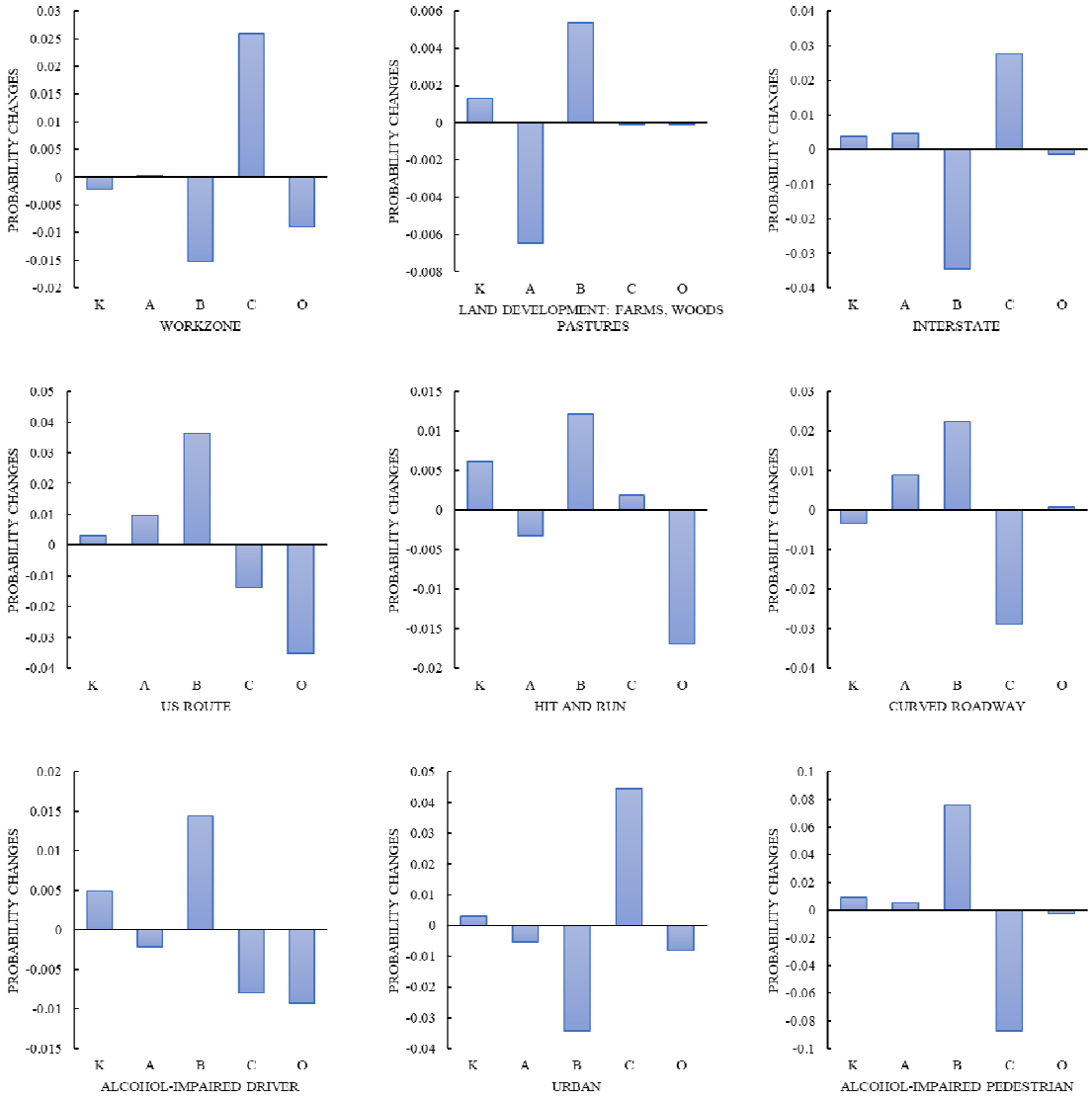

FIGURE 7.6: Average Partial Dependence Changes of Contributing Factors Increasing the Risk of Pedestrians Sustaining Severer Injuries (i.e., Fatality and Incapacitating Injury) in the Hotspot Data

Despite the top 15 factors, the changes of the partial dependences of factors which show their impacts on increasing the risk of pedestrians sustaining severer injuries (i.e., "K" of fatality and "A" of incapacitating injury) are further computed for more in-depth analysis. Nine factors with such effects are identified and their associated average partial

dependence changes are illustrated in the Figure 7.6. Again, similar to the interpretation of the marginal effect utilized in the conventional DCMs, the average partial dependence change has the same way to be examined.

However, some factors that are found to increase the risk of pedestrians sustaining severer injuries in the hotspot XGBoost model which appear to have mitigating effects on severer injuries for pedestrians in the XGBoost model with the whole crash dataset. The factors are “workezon”, “US route”, and “urban area”. On the other hand, factors of “male driver”, “elder pedestrian”, and “ambulance: yes” are found to decrease the risk of pedestrians being severely injured in the hotspot XGBoost model rather than their deteriorative effects in the XGBoost model with the whole dataset.

7.5. Summary

In summary, this chapter conducts the emerging hotspot analysis of the single-pedestrian-single-vehicle crashes data. Cases within the hotspot areas have been filtered out to be further fitted with a developed XGBoost model. Same optimization process with hyperparameter tuning to the XGBoost model is provided to obtain the best model structure in order to achieve a relatively high performance of the proposed XGBoost model on the hotspot dataset. Detailed numerical results with the corresponding feature importance and partial dependences of the contributing factors are also computed. Typical interpretations of the model results are presented to examine the effects of the identified contributing factors to pedestrian injury severities in the pedestrian-vehicle crashes.

CHAPTER 8: SUMMARY AND CONCLUSIONS

8.1. Introduction

As one of the most vulnerable entity within the transportation system, a pedestrian might face more dangers and sustain severer injuries in the traffic crashes than others. The safety of pedestrians is becoming more and more critical with an increasing trend of pedestrian fatality when comparing to other fatalities in traffic crashes over the past decades (NHTSA, 2017). The relationship between pedestrian injury severities and a variety of contributing factors (i.e., *pedestrian characteristics*, *driver characteristics*, *crash characteristics*, *locality and roadway characteristics*, *time and environment characteristics*, and *traffic control characteristics and workzone*) is highly complex. Moreover, it should be noted that in the police-reported crash data, some unobservable factors are not reported by law enforcement agencies and cannot be collected from state crash records, which may induce unobserved heterogeneity and have impacts on injury severities. Neglecting such unobserved heterogeneity might lead to biased estimation of parameters and therefore having possibly improper inferences (Mannering and Bhat, 2014; Shaheed and Gkritza, 2014). Therefore, applications and developments of proper modeling approaches are needed to identify causations in pedestrian-vehicle crashes to better ensure the safety of pedestrians.

On the other hand, with the development of artificial intelligence techniques, a variety of novel machine learning methods have been established (Jordan and Mitchell, 2015). And compared to conventional DCMs, machine learning models (Tang et al., 2018) are more flexible with no or few prior assumptions about input variables and have higher adaptability to process outliers, missing and noisy data. As one of the highly representative

among all those techniques, XGBoost algorithm has been successfully and widely used by many winners in many machine learning competitions and various domains with a significant popularity. Thus, the XGBoost algorithm has the potential to be deployed in the field of crash data and traffic safety related analysis to better promote safety to particularly vulnerable entities such as pedestrians within the transportation system.

Furthermore, the crash data inherent has patterns related to both space and time, crashes happened in locations with highly aggregated uptrend patterns should be worth exploring to examine the most recently deteriorative factors which contribute to severer injuries (i.e., fatalities and incapacitating injuries) of pedestrians in the pedestrian-vehicle crashes. With such consideration, emerging hotspot analysis tool developed by the ArcGIS could provide solid references to identify the spatiotemporal patterns of the crash related data. Hence, the combination of both emerging hotspot analysis and XGBoost model could offer an opportunity to extract the most accurate and up-to-date information on deteriorative contributing factors affecting the pedestrian injury severities in pedestrian-vehicle crashes to achieve a better understanding for researchers and policymakers.

The major goal of this study is to develop a framework for modeling and analyzing pedestrian injury severities in single-pedestrian-single-vehicle crashes with providing a higher resolution on identifications of contributing factors and their associating effects affecting the injury severities of pedestrians, particularly on those most recently deteriorative factors. The pedestrian crash data in North Carolina ranging from 2007 to 2018 is used and several different categories of variables (i.e., pedestrian characteristics, driver characteristics, crash characteristics, locality and roadway characteristics, time and environment characteristics, and traffic control characteristics and workzone) are

considered. Developments of both conventional DCMs and the selected machine learning model, i.e., XGBoost, are established. Detailed comparisons among all developed models are conducted with a result showing that XGBoost outperforms all other conventional DCMs in all selected measurements. In addition, an emerging hotspot analysis is further utilized to identify the most targeted hotspots (i.e., "New Hot Spot", "Intensifying Hot Spot", "Persistent Hot Spot", and "Consecutive Hot Spot"), followed by a proposed XGBoost model that analyzes the most recently deteriorative factors affecting the pedestrian injury severities. By completions of all abovementioned tasks, the gaps between theory and practice could be bridged.

The rest of this chapter is presented as: Section 8.2 provides a summary and conclusion of the comparisons between conventional DCMs and XGBoost model. Section 8.3 gives a summary and conclusion of the developments of pedestrian injury severity modeling and contributing factors analysis by applying both emerging hotspot analysis and the proposed XGBoost method. Section 8.4 presents a brief discussion of the limitations of the current framework and approaches and gives future research directions.

8.2. Summary and Conclusions of Comparisons between Conventional DCMs and XGBoost Model for Modeling and Analyzing Pedestrian Injury Severities

In this study, three conventional DCMs (i.e., MNL, PPO, and ML models) and one XGBoost model have been developed by using the whole retrieved single-pedestrian-single-vehicle crash dataset in North Carolina. After developments, several widely used measurements denoting goodness-of-fit of model, which are accuracy, precision, recall,

and F1 score, are utilized to evaluate and compare the performances of all proposed model structures.

Results show that the proposed XGBoost model outperforms the other three conventional DCMs with more than 40% higher in the overall accuracy (i.e., 93.23% vs 51.65% [MNL], 51.71% [ML], and 51.33% [PPO]). Recalls and F1 scores of incapacitating injury and no injury of all three conventional DCMs almost equal to zero, which means that the conventional DCMs can rarely correctly identify outcomes for observations with these two injury severity patterns. While these three measurements of the proposed XGBoost model for these two categories imply a far better performance, according to Tables 6.1-6.4.

Referring to the power for identifying contributing factors, as mentioned in Section 2.4, unlike conventional statistical and econometric DCMs, there are fewer or no requirements on the pre-defined assumptions about the relationships between outcomes of injury severity and contributing factors in the XGBoost model so that no variables have been excluded in the model development. And with the best XGBoost model structure (more than 90% accuracy), one would be able to examine the effects of all existing contributing factors towards each injury severity level for pedestrians in the pedestrian-vehicle crashes, more information could be extracted from the results for further supporting policymakers with more accurate and targeted safety improvement plans to help pedestrians in the transportation system.

8.3. Summary and Conclusions of XGBoost Model on Emerging Hotspot Crash Data for Modeling and Analyzing Pedestrian Injury Severities

By taking advantage of the emerging hotspot analysis, this study further develops a XGBoost model based on the identified emerging hotspot crash dataset in which crashes happened in locations with highly aggregated uptrend patterns (i.e., “New Hot Spot”, “Intensifying Hot Spot”, “Persistent Hot Spot”, and “Consecutive Hot Spot”). The ranking of factor importance in the developed hotspot XGBoost model seems to be similar as that in the XGBoost model with whole crash dataset. However, some factors that are found to increase the risk of pedestrians sustaining severer injuries in the hotspot XGBoost model which appear to have mitigating effects on severer injuries for pedestrian in the XGBoost model with whole crash dataset. The factors are “workezon”, “US route”, and “urban area”. On the other hand, factors of “male driver”, “elder pedestrian”, and “ambulance: yes” are found to decrease the risk of pedestrians being severely injured in the hotspot XGBoost model instead of their deteriorative effects in the XGBoost model with whole dataset. Additionally, different magnitudes of the effects for same factors can also be observed according to the calculated partial dependence changes.

8.4. Future Research Directions

The framework and results for modeling and analyzing pedestrian injury severities in pedestrian-vehicle crashes, at particularly most recent hotspots with uptrend of crash occurrences in this research could be a solid reference for the identifications of contributing factors affecting the pedestrian injury severities to further promote safety to pedestrians within the transportation system. Though relatively higher resolutions could be achieved

when compared to conventional statistical models, it is helpful to identify the most recently deteriorative factors, and limitations still exist in its current form.

Firstly, other than the variables in the retrieved police-reported crash data, variables related to traffic characteristics such as traffic volumes (e.g., pedestrian volumes and vehicle volumes) are not considered in this study. By examining the current pedestrian crash data, obvious "two-peak" pattern could be observed, which is quite similar to the vehicle traffic during weekdays. And this should be further considered, since the pedestrian injury severities in pedestrian-vehicle crashes are also highly affected by such traffic characteristics.

Secondly, the distance interval used in the emerging hotspot analysis is fixed as 5000ft × 5000ft and the distance of nearest neighbors is set as 20000ft, which is manually set. Theory-based methods might be deployed with a solid reference towards the achievement of more accurate results. Additionally, possible effects on the detailed built in environment factors related to roadway segments within the ranges of hotspots could be included for an in-depth analysis in the future.

Finally, since differences among all patterns identified by the emerging hotspot analysis do exist according to the criteria used in the analysis, future research could focus on modeling and analyzing segmented dataset based on different spatiotemporal patterns.

REFERENCES

Abdel-Aty, M. A., Radwan, A. E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention, 32(5), 633-642.

Behnood, A., Mannering, F. L., 2016. An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. Analytic methods in accident research, 12, 1-17.

Betty, E. L., Bollard, B., Murphy, S., Ogle, M., Hendriks, H., Orams, M. B., Stockin, K. A., 2020. Using emerging hot spot analysis of stranding records to inform conservation management of a data-poor cetacean species. Biodiversity and Conservation, 29(2), 643-665.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B., 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accident Analysis and Prevention, 42(4), 1072-1079.

Chang, L. Y., Chen, W. C., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36(4), 365–375.

Chen, C., Zhang, G., Huang, H., Wang, J., Tarefder, R. A., 2016. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Accident Analysis and Prevention, 96, 79-87.

Chen, F., S. Chen., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis and Prevention, 43(5), pp. 1677-1688. doi: 10.1016/j.aap.2011.03.026.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, Z., Fan, W., 2019a. A multinomial logit model of pedestrian-vehicle crash severity in North Carolina. International journal of transportation science and technology, 8(1), 43-52.

Chen, Z., Fan, W., 2019b. Modeling pedestrian injury severity in pedestrian-vehicle crashes in rural and urban areas: mixed logit model approach. Transportation Research Record, 2673 (4), 1023-34.

Cheng, J. C., Ma, L. J., 2015. A non-linear case-based reasoning approach for retrieval of similar cases and selection of target credits in LEED projects. Building and Environment, 93, 349-361.

Dai, D., 2012. Identifying clusters and risk factors of injuries in pedestrian–vehicle crashes in a GIS environment. Journal of Transport Geography, 24, 206-214.

Ding, C., Chen, P., Jiao, J., 2018. Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach. Accident Analysis and Prevention, 112, 116-126.

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

Getis, A., Ord, J. K., 2010. The analysis of spatial association by use of distance statistics. In Perspectives on spatial data analysis (pp. 127-145). Springer, Berlin, Heidelberg.

Gkritza, K., Mannering, F. L., 2008. Mixed logit analysis of safety-belt use in single-and multi-occupant vehicles. Accident Analysis and Prevention, 40(2), 443-451.

Gong, L., Fan, W. D., 2017. Modeling single-vehicle run-off-road crash severity in rural areas: Accounting for unobserved heterogeneity and age difference. Accident Analysis and Prevention, 101, 124-134.

Gong, L., Fan, W., Washing, E. M., 2016. Modeling severity of single vehicle run-off-road crashes in rural areas: model comparison and selection. Canadian journal of civil engineering, 43(6), 493-503.

Gong, Y., Abdel-Aty, M., Cai, Q., Rahman, M. S., 2019. A decentralized network level adaptive signal control algorithm by deep reinforcement learning. Transportation Research Board 98th Annual Meeting.

Haleem, K., Alluri, P., Gan, A., 2015. Analyzing pedestrian crash injury severity at signalized and non-signalized locations. Accident Analysis and Prevention, 81, 14-23.

Himanen, V., Kulmala, R.,1988. An application of logit models in analysing the behaviour of pedestrians and car drivers on pedestrian crossings. Accident Analysis and Prevention, 20(3), 187-197.

Islam, S., Jones, S. L., Dye, D., 2014. Comprehensive analysis of single-and multi-vehicle large truck at-fault crashes on rural and urban roadways in Alabama. Accident Analysis and Prevention, 67, 148-158.

Jalayer, M., Pour-Rouholamin, M., Zhou, H., 2018. Wrong-way driving crashes: A multiple correspondence approach to identify contributing factors. Traffic injury prevention, 19(1), 35-41.

Jordan, M. I., Mitchell, T. M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Jun, M. A., Cheng, J. C., 2017. Selection of target LEED credits based on project information and climatic factors using data mining techniques. Advanced Engineering Informatics, 32, 224-236.

Kim, J. K., Ulfarsson, G. F., Kim, S., Shankar, V. N., 2013. Driver-injury severity in single-vehicle crashes in California: a mixed logit analysis of heterogeneity due to age and gender. Accident Analysis and Prevention, 50, 1073-1081.

Kim, J.-K., Ulfarsson, G., Shankar, V., Kim, S., 2008a. Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. Accident Analysis and Prevention, 40 (5), 1695-1702.

Kim, J.-K., Ulfarsson, G., Shankar, V., Mannering, F., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accident Analysis and Prevention, 42 (6), 1751-1758.

Kim, K., Brunner, I., Yamashita, E., 2008b. Modeling fault among accident—Involved pedestrians and motorists in Hawaii. Accident Analysis and Prevention, 40 (6), 2043-2049.

Kim, S., Ulfarsson, G. F., 2019. Traffic safety in an aging society: Analysis of older pedestrian crashes. Journal of Transportation Safety and Security, 11(3), 323-332.

Koppelman, F. S., Bhat, C., 2006. A self instructing course in mode choice modeling: multinomial and nested logit models.

Kwigizile, V., Sando, T., Chimba, D., 2011. Inconsistencies of ordered and unordered probability models for pedestrian injury severity. Transportation Research Record, 2264 (1), 110-118.

Li, Y., Fan, W., 2019. Modelling severity of pedestrian-injury in pedestrian-vehicle crashes with latent class clustering and partial proportional odds model: a case study of North Carolina. Accident Analysis and Prevention, 131, 284-296.

Li, Y., Fan, W., 2019. Pedestrian injury severities in pedestrian-vehicle crashes and the partial proportional odds logit model: accounting for age difference. Transportation Research Record, 2673 (5), 731-746.

Li, Y., Fan, W., 2020. Modelling the severity of pedestrian injury in pedestrian—vehicle crashes in North Carolina: A partial proportional odds logit model approach. Journal of Transportation Safety and Security, 12(3), 358-379.

Ma, L., Wang, G., Yan, X., Weng, J., 2016. A hybrid finite mixture model for exploring heterogeneous ordering patterns of driver injury severity. Accident Analysis and Prevention, 89: 62-73.

Malyshkina, N. V., Mannering, F. L., 2010. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis and Prevention, 42(1), 131-139.

Mannering, F. L., Bhat, C. R., 2014. Analytic Methods in Accident Research: Methodological frontier and future directions. Analytic methods in accident research, 1, 1-22.

Mannering, F., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.

Manski, C. F., McFadden, D. (Eds.), 1981. Structural analysis of discrete data with econometric applications (pp. 2-50). Cambridge, MA: MIT press.

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of applied Econometrics, 15(5), 447-470.

Milton, J. C., Shankar, V. N., Mannering, F. L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention, 40 (1), 260-266.

Mokhtarimousavi, S., 2019. A time of day analysis of pedestrian-involved crashes in California: Investigation of injury severity, a logistic regression and machine learning approach using HSIS data. Institute of Transportation Engineers. ITE Journal, 89 (10), 25-33.

Mokhtarimousavi, S., Anderson, J. C., Azizinamini, A., Hadi, M., 2020. Factors affecting injury severity in vehicle-pedestrian crashes: A day-of-week analysis using random parameter ordered response models and Artificial Neural Networks. International Journal of Transportation Science and Technology.

Moudon, A. V., Lin, L., Jiao, J., Hurvitz, P., Reeves, P., 2011. The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accident Analysis and Prevention, 43(1), 11-24.

NHTSA, 2018. Fatality Analysis Reporting System (FARS). National Highway Traffic Safety Administration, Washington (DC). www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars. Accessed July 31, 2020.

NHTSA, 2017. Traffic safety facts: Pedestrians. National Highway Traffic Safety Administration, Washington (DC). Report No.: DOT HS 812 681.

Obeng, K., Rokonuzzaman, M., 2013. Pedestrian injury severity in automobile crashes. Open Journal of Safety Science and Technology, 3 (02), 9.

Peterson, B., Harrell Jr, F. E., 1990. Partial proportional odds models for ordinal response variables. Journal of the Royal Statistical Society: Series C (Applied Statistics), 39(2), 205-217.

Pour, A. T., Moridpour, S., Tay, R., Rajabifard, A., 2016. A partial proportional odds model for pedestrian crashes at mid-blocks in Melbourne metropolitan area. In MATEC web of conferences (Vol. 81, p. 02020). EDP Sciences.

Quddus, M. A., Wang, C., Ison, S. G., 2009. Road traffic congestion and crash severity: econometric analysis using ordered response models. Journal of Transportation Engineering, 136(5), 424-435.

Rahman, M. S., Abdel-Aty, M., Hasan, S., Cai, Q., 2019. Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. Journal of safety research, 70, 275-288.

Retting, R., Schwartz S., 2019. Pedestrian Traffic Fatalities by State, 2020 Preliminary Data. Washington, DC: Governors Highway Safety Association.

Retting, R., Schwartz, S., 2018. Pedestrian Traffic Fatalities by State: 2019 Preliminary Data. Washington, DC: Governors Highway Safety Association.

Rifaat, S. M., Chin, H. C., 2007. Accident severity analysis using ordered probit model. Journal of Advanced Transportation, 41 (1), 91-114.

Rifaat, S. M., Tay, R., de Barros, A., 2012. Urban street pattern and pedestrian traffic safety. Journal of urban design, 17(3), 337-352.

Sasidharan, L., Menéndez, M., 2014. Partial proportional odds model—An alternate choice for analyzing pedestrian crash injury severities. Accident Analysis and Prevention, 72, 330-340.

Sasidharan, L., Wu, K. F., Menendez, M., 2015. Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. Accident Analysis and Prevention, 85, 219-228.

Savolainen, P. T., Mannering, F. L., Lord, D., Quddus, M. A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention, 43(5), 1666-1676.

Scott Long, J., 1997. Regression models for categorical and limited dependent variables. Advanced quantitative techniques in the social sciences, 7.

Shaheed, M. S., Gkritza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Analytic Methods in Accident Research, 2, 30-38.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of safety research, 27(3), 183-194.

Sze, N. N., Wong, S. C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis and Prevention, 39 (6), 1267-1278.

Tang, J., Liu, F., Zhang, W., Ke, R., Zou, Y., 2018. Lane-changes prediction based on adaptive fuzzy neural network. Expert Systems with Applications, 91, 452-463.

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian–vehicle crash severity. International Journal of Sustainable Transportation, 5 (4), 233-249.

Thomas, L., Vann, M., Levitt, D., 2018. North Carolina Pedestrian Crash Trends and Facts 2011 - 2015. Project RP 2017 - 42, The North Carolina Department of

Transportation, Division of Bicycle and Pedestrian Transportation, Raleigh, North Carolina.

Train, K., 2003. Discrete choice methods with simulation, Cambridge University Press, Cambridge, UK.

Tulu, G. S., Washington, S., Haque, M. M., King, M. J., 2017. Injury severity of pedestrians involved in road traffic crashes in Addis Ababa, Ethiopia. Journal of Transportation Safety and Security, 9(sup1), 47-66.

Ulfarsson, G. F., Kim, S., Booth, K. M., 2010. Analyzing fault in pedestrian–motor vehicle crashes in North Carolina. Accident Analysis and Prevention, 42 (6), 1805-1813.

Wang, X., Abdel-Aty, M., 2008. Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models. Accident Analysis and Prevention, 40(5), 1674-1682.

Washington, S., Karlaftis, M. G., Mannering, F., Anastasopoulos, P., 2020. Statistical and econometric methods for transportation data analysis. CRC press.

Wooldridge, J. M., 2002. Econometric analysis of cross section and panel data MIT Press. Cambridge, MA, 108.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention, 59, 506-521.

Yasmin, S., Eluru, N., Ukkusuri, S. V., 2014. Alternative ordered response frameworks for examining pedestrian injury severity in New York City. Journal of Transportation Safety and Security, 6 (4), 275-300.

Zajac, S. S., Ivan, J. N., 2003. Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut. Accident Analysis and Prevention, 35 (3), 369-379.

Zhang, L., Zhan, C., 2017. Machine learning in rock facies classification: an application of XGBoost. In International Geophysical Conference, Qingdao, China, 17-20 April 2017 (pp. 1371-1374). Society of Exploration Geophysicists and Chinese Petroleum Society.

Zhou, Z. P., Liu, Y. S., Wang, W., Zhang, Y., 2013. Multinomial logit model of pedestrian crossing behaviors at signalized intersections. Discrete Dynamics in Nature and Society, 2013.