Addressing Transportation Equity by Comparing In-Service Performance of Roadside Safety Devices through Machine Learning Modeling

Hanzhen Wang Wang

ADDRESSING TRANSPORTATION EQUITY BY COMPARING IN-SERVICE PERFORMANCE OF ROADSIDE SAFETY DEVICES THROUGH MACHINE LEARNING MODELING

THESIS

Presented in Partial Fulfillment of the Requirements for

the Master of Science Degree in the Graduate School

of Texas Southern University

By

Hanzhen Wang, B.S.

Texas Southern University

2021

Approved By

Dr.Fengxiang Qiao

Chairperson, Thesis Committee

Dr. Gregory H. Maddox

Dean, The Graduate School

Approved By:

Dr.Fengxiang Qiao	06/25/2021
Chairperson, Thesis Committee	Date
Dr. Lei Yu	06/25/2021
Committee Member	Date
Dr. Mehdi Azimi	06/25/2021
Committee Member	Date
Dr. Yachi Wanyan	06/25/2021
Graduate School Representative	Date

ADDRESSING TRANSPORTATION EQUITY BY COMPARING IN-SERVICE PERFORMANCE OF ROADSIDE SAFETY DEVICES THROUGH MACHINE LEARNING MODELING

By

Hanzhen Wang, B.S.

Texas Southern University, 2021

Professor Dr. Fengxiang Qiao, Advisor

Transportation equity plays an important role in modern communities, and a fair distribution of transportation infrastructures is vital as an integral part of transportation planning process. The In-Service Performance Evaluation (ISPE) satisfies transportation safety requirements by identifying the problems of roadside safety devices during installation and maintenance process with proper solutions, and the performance results reveal the current statue of target devices in specific areas. Although several studies have been conducted to emphasize transportation equity, there is still a lack of equity research specifically focusing on the deploying of roadside safety devices associated with ISPE results. With proper comparison of in-service performance results in different areas, the importance of ensuring transportation equity of all communities and areas in the decision-making process is able to be demonstrated.

This thesis utilizes Machine Learning models to analyze linked crash and roadway data related to major roadside safety devices implemented in Texas. Three typical roadside safety devices are selected to be assessed, including: (1) guardrail, (2) median barrier, and (3) bridge rail. By comparing both statistical and Machine Learning based modeling analysis with rural and metropolitan areas in specific counties, it is demonstrated that distributions of crashes that end up causing heavy property damage or serious injuries is higher in rural communities regardless of its lower crash frequency. The data analysis result suggests that parameters related to roadway conditions and transportation infrastructures tend to have higher influence over the performances of rural safety devices. Additional one year of crash data analysis also addresses the importance of transportation equity under the COVID-19 pandemic period. Recommendations on improving overall equity and Environmental Justice (EJ) within all regions are conducted with stated findings.

Keywords: Environmental Justice, In-Service Performance Evaluation, Machine Learning, Roadside Safety Devices, Transportation Equity

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

VITA

2015-2019 .... B.E. in Electrical Engineering and Automation

Xuchang University, Henan, China

2019-2021 .... Graduate Research Assistant

Department of Transportation Studies,

Texas Southern University

Major Field....................................... Transportation Planning & Management

ACKNOWLEDGEMENT

I would like to give my gratitude to my advisor, Dr. Fengxiang Qiao, who has offered me huge help and guidance on the idea and revisions of this thesis. Dr. Qiao has been dedicated in mentoring me through my study of master's program at TSU. I am more than thankful for his continuous and professional support not only in my study but also in my life spent as a Graduate Research Assistant. By sharing Dr. Qiao's wisdom in academic studies, I have learned a lot and made my own achievement as receiving 2020 ITS Texas Scholarship. Thanks to the recommendation letters by Dr. Qiao and Dr. Yunpeng Zhang in UH.

I am truly grateful for Dr. Lei Yu who provided me opportunity to study at TSU, and I also appreciate other committee members of my thesis defense, Dr. Mehdi Azimi and Dr. Yachi Wanyan. I am sincerely thankful for Dr. Yi Qi and other professors in the Department of Transportation Studies as well as my colleagues for their help in my study and life.

This thesis study is partially associated with a research project sponsored by Texas Department of Transportation (TxDOT). I am more than grateful for the help in collecting and processing data from the teammates of this project.

CHAPTER 1

INTRODUCTION

1.1 Background

Transportation equity means distributing transportation resources, benefits, and services fairly in each community. The U.S. Office of Management and Budget (OMB) designated different communities into metropolitan communities and rural communities. The majority popularity with more than 50,000 population and consists of a core urban area categorized under metropolitan communities (HRSA, 2021). The actions to address Environmental Justice (EJ) in minority population which includes transportation users in rural communities are purposed in the 2012 Federal Highway Administration (FHWA) order 6640.23A (FHWA, 2012). With lower population in rural areas, funds for transportation maintenance are limited in these areas due to less tax distribution.

However, 68.13% of American total lane miles are located in rural areas till 2019 and fatality rate in rural areas is two times higher than metropolitan areas (FHWA, 2019; NHTSA, 2018). In addition, around 80% of railroad crossings in rural areas lack of suitable warning devices (U.S. DOT, 2021). According to FHWA’s rural transportation planning, 40% roads in rural areas do not meet the requirement for current travel, and around 50% of bridges that over 20 feet in rural areas have structural flaws now (FHWA, 2017). In a word, there is a lack of investment in the maintenance and preservation of rural transportation infrastructure which leads several safety issues in rural areas. The inequality of traffic risk needs to be addressed under current circumstances.

The roadside safety device is the transportation infrastructure designed and served as an engineering hardware, which has the general function of decreasing the risk of roadway crashes significantly and ensuring transportation safety (Cantisani, *et al.*, 2017). Several common types of roadside safety devices have been implemented on the roadways to redirect and protect vehicles, such as: (1) longitudinal barriers (guardrails); (2) barrier terminals (guardrail end treatments); (3) crash cushions; and (4) breakaway hardware (signs, luminaires, etc.) (FHWA, 2020). Among listed safety hardware, barriers are typical devices that can be found along highways and roadways. According to Roadside Design Guide conducted by American Association of State Highway Transportation Officials (AASHTO), the barriers refer to three major categorizes by their specific functions: (1) roadside barrier; (2) median barrier; and (3) bridge railing. Accordingly, to assess the performance of this type of safety transportation infrastructure, In-Service Performance Evaluation (ISPE) is an important procedure to process when determining whether the devices perform as they designed to in real world conditions. These considerable conditions include traffic conditions, site maintenance, and environmental conditions (AASHTO, 2011).

Previous research focused on transportation equity are mainly aiming at discussing burdens and benefits of transportation infrastructure among various areas and populations. These approaches include redefining equity rules on political side and identifying equalization standard for transportation planning process. While other factors such as measures of performance should be considered in a broader perspective (Karner, *et al.*, 2020). Another research also recommended improving analytical approaches to evaluate the features of inequity conditions in burdens and benefits of transportation infrastructure (Karner, *et al.*, 2016). For the analytical approach as performance measures for particular transportation safety infrastructure roadside safety devices, there are currently a lack of comprehensive ISPE results to demonstrate transportation equity among different areas. As a defined procedure in ISPE, outcome of crash data analysis illustrates the performance of safety hardware. As road crashes becoming the major death causes for people among 1 to 54 ages in the United States (ASIRT, 2020), the thorough analysis of crash data can significantly reveal the performance of safety transportation infrastructure. This thesis will fill the gap by addressing transportation equity and providing recommendations through comparing ISPE results between various areas.

1.2 Research Objectives

The major objective of this research is to perform proper comparison of in-service performance results in different areas through Machine Learning modeling on Texas crash data, and to demonstrate the importance of ensuring transportation equity of all communities and areas in the decision-making process. The specific objectives can be concluded as following points:

- Find crash distribution of main safety devices on crash severity and property damage in rural and metropolitan areas,
- Analyze the main safety devices related roadway and crash data in Texas using statistical analysis and Machine Learning modeling method,
- Analyze related crash data under COVID-19 pandemic period in Texas and compare the impact factors by Machine Learning modeling method, and
- Assess the performance of in-service roadside safety devices in Texas metropolitan and rural areas, while addressing the transportation equity conditions for safety devices in the decision-making process.
- Provide recommendations to improve transportation equity for roadside safety devices, identify future study needs in this direction of research.

1.3 Outline

This thesis is conprised of five chapters. Chapter one presents the introduction of the study by presenting the background, as well as research gaps, research objectives, and the thesis outline. Chapter two performs a thorough literature review on the studies related to transportation equity and environmental justice, roadside safety devices, ISPE practices, and transportation equity conditions under pandemic period. Chapter three demonstrates the design of the study, including the methodology, data collection, data processing, Machine Learning modeling method, and related algorithms. Chapter four illustrates the data analysis results and discussion, which is divided into statistical analysis and

Machine Learning modeling analysis, while crash data under pandemic period will also be analyzed separately. Chapter five presents the conclusion of this study with related recommendations based on the results.

CHAPTER 2

LITERATURE REVIEW

This chapter summarizes findings and reviews of the existing studies related to the proposed research topic. A general idea about current update and perspectives from correlated literature is presented to address the significance of the study. To start with, studies and practices related to transportation equity will be reviewed and summarized. The literature based on several common types of roadside safety devices will be discussed to learn the applications and functions of typical safety hardware. In addition, ISPE definitions and procedures from existing reports and studies will be summarized. ISPE levels will be introduced, and performance measures will be emphasized based on the review of related literature.

2.1 Transportation Equity

2.1.1 Transportation Equity Policies

The concept of equity is derived from Environmental Justice (EJ) when related actions or proposals by authorities have been started in late 1900s. The Title VI of the Civil Rights Act was passed in 1964, it forbids any form of discrimination against race, color, and national origin (Civil Rights Act, 1964). After that, The National Environmental Justice Advisory Council (NEJAC) was established by Environmental Protection Agency (EPA) in 1993, which provides recommendations about issues related to EJ and consider EJ into the development of agency policies and activities (EPA, 1993). It is serving as the legal satisfaction of following the 1994 Executive Order (EO) 12898 as actions on federal level to address EJ (Twaddell & Zgoda, 2020). The EO orders that each Federal agency to conduct an agency wide EJ strategy. DOT issued its initial EJ Strategy in 1995 to meet the requirements. In the FHWA order 6640.23A, the EJ within minority and low-income populations are addressed. It is suggested that FHWA's continuing policy has the intention of identifying and eliminating discriminatory influences throughout the decision-making process (FHWA, 2012). DOT has been continuously considering the EJ strategy into the programs, activities, and policies, while the minority and low-income population have been given more notices when implementing the EJ strategy (U.S. DOT, 2017).

According to the Environmental Justice Reference Guide by FHWA, the fairly participation of all potentially affected communities and agencies is required when addressing EJ in the decision-making process. Thus, as the federal financial recipients, apart from U.S. DOT and FHWA, the Metropolitan Planning Organizations (MPOs) and Federal Transit Administration (FTA) are also required to follow the EJ strategy polices to further ensure transportation equity (FHWA, 2015). Following by the existing laws, the National Environmental Policy Act (NEPA) has also been involved relatively to ensure the equity of healthy environment.

Public engagement in the decision-making process is also required by the DOT to fulfil EJ strategy. It is intended to avoid, minimize, or mitigate the disproportionately influence and environmental effects from transportation infrastructure projects. The transportation infrastructure planning projects is also considered for the EJ influence (U.S. DOT, 2017).

2.1.2 Transportation Equity Practices

To satisfy the continuing transportation equity needs from EJ strategy policies, research and practices focused on developing transportation equity have been carried out in different aspects. Some studies implement various models to evaluate the transportation equity impacts on projects and practices. Methods of training models to simulate and present the equity effects on particular conditions are regarded as equity analysis. Bills *et al.* emphasized that the equity analysis consists of three major parts including assessment priority, models used, and equity indicators (Bills, *et al.*, 2012). Priority assessment refers to distinguishing the importance of involved transportation factors in considering of the major interests. The models used for equity analysis include transportation simulating and forecasting models, while the primary goal of applying models is to conduct comprehensive analysis that the model output represents the difference of settled scenarios to address equity or inequity situations. Equity indicators present the overall impacts focused on equity side and demonstrate the transportation distribution in comparison.

Accordingly, Rodier *et al.* implemented a spatial economic model to evaluate the influences of equity on the land use in the purpose of reducing gas emissions (Rodier, *et al.*, 2010). The distribution measures are conducted by proposed model in this study and the recommendation of smaller urban form around transit stations help developing transportation equity. More related studies intend to apply travel forecast and travel demand models to perform equity analysis on various transportation issues or projects (Castiglione, *et al.*, 2006). In these practices on transportation equity, some are reginal transportation practices, and the analysis targets ranging from household pricing to land use planning. The equity analysis in related studies tend to present the distribution of transportation sources, by ensuring the equity in transportation investments among all communities and areas, avoiding the negative or inequity in the decision-making process.

2.1.3 Transportation Equity Needs under Pandemic

Under the impact of COVID-19 pandemic since earlier 2020, social activities of various communities have been affected and the influences on transportation tend to be complicated (Du, *et al.*, 2020). During this period, transportation equity is also affected and the needs for satisfying EJ strategy are rising according to recent situations. Due to the decrease of overall revenue, the funding and budgets for equitable transportation projects have been changed in some areas. In addition, public transportation has been majorly affected by the pandemic and transportation equity on the transportation modes is also influenced in a way. It is suggested that not only the funding amount for public transit should be increased but also the funding approach should be developed for equitable public transit. The engagement with community residents is also recommended for better equitable results in the policy decision-making process (Davis & Stacy, 2021).

According to another research in the period of pandemic, urban EJ is also not relatively balanced. Some communities have situations of socio-environmental injustice due to the condition of the housing and popularity density under the pandemic (Cole, *et al.*, 2020). Abdoli and Hosseinzadeh have also studied the spatial equity among public transit under pandemic period. They concluded that the pandemic effect on various transportation transit systems is uneven. Thus, the inequity was addressed for public transport and social groups (Abdoli & Hosseinzadeh, 2021).

2.2 In-Service Performance Evaluation

2.2.1 ISPE Practices

As a vital process to evaluate performance of roadside safety devices, ISPE practices have been conducted in several states. Four levels of ISPE are included in ISPE implementations in Arizona and Texas (Mak & Sicking, 2002; Zhang, *et al.*, 2019). Level I of ISPE focused on developing a comprehensive database that collects data from various data sources including crash data, highway and roadway data, roadside safety devices inventory data, and maintenance data. Level II of ISPE based on evaluating specific individual road safety devices and obtain further information through field studies. Level III of ISPE collects more detailed data of target safety devices that involved serious crashes resulted in fatal or incapacity injuries or heavy property damage. Level IV of ISPE evaluates the improved or newly designed safety devices and fulfil the gaps of limited information on these types of roadside safety devices. It is crucial that all the results from each level of ISPE will eventually enrich the Level I ISPE database (Qiao, *et al.*, 2020).

Early ISPE practices implemented in Texas consists of two phases of evaluation process. The Phase I ISPE collects data mostly from existing database while Phase II requires more detailed investigation of specific safety device when the failure rate is too high. The crash data analysis mainly focused on statistical counting on the crash number, ranking crashes through counties, and comparing vehicle type related to safety features (Schalkwyk, *et al.*, 2004). Studies that focused on specific safety device like cable median barrier using performance evaluation was also carried out. This research by Schalkwyk, *et al.* (2004) analyzed data related to cost, maintenance and repair, safety, and field performance evaluation. By counting the fatal and incapacity injuries before and after the target cable median barrier were putting into service, the significance of the specific safety device was demonstrated. It also concluded that, weather conditions have important effect on the occurrence of crashes related to studied safety devices as shown in Figure 1 (Cooner, *et al.*, 2009).

Figure 1 Impact of weather condition on cable median barriers crash (Source: Cooner, *et al.*, 2009)

2.2.2 Roadside Safety Devices

As the early evaluation guideline for roadside safety device, National Cooperative Highway Research Program (NCHRP) report 350 presents uniform guidelines for evaluating highway safety devices and helps determine the criteria of the evaluation in the assessment of tests targeting at various types of roadside safety systems (Ross Jr, *et al.*, 1993). Followed by such criteria, the AASHTO's Manual for Assessing Safety Hardware (MASH) was published later in 2009 and 2016 to provide guidelines for crash testing temporary and permanent highway safety structures (AASHTO, 2009; AASHTO, 2016). Some listed types of roadside safety devices are included under the MASH 2016, which is implemented in 2020. Categorized safety devices in MASH 2016 include: (1) Longitudinal Barriers, (2) Terminals, (3) Crash Cushions, (4) Support Structures, (5) Work Zone Attenuation and Channelizers, (6) Other Devices, and (7) Drainage and Geometric Features (AASHTO, 2016).

Relatively, some specific roadside safety devices are mainly implemented on the roadways such as guardrail, median barrier, bridge railing, crash cushion, and roadside barrier. According to Roadside Safety Field Guide 2014, median barrier, bridge railing, and roadside barrier all longitudinal barriers with different functions. Among these devices, median barrier serves as barrier that preventing out-of-line vehicles from crossing the median area. Bridge railing has the function of preventing out-of-line vehicles from driving through the side of bridge. Roadside barrier works as a barricade to shield obstacles along roadside, which also has the function of protecting pedestrians from passing vehicles. Crash cushions function at decelerating or redirecting out-of-line vehicles to prevent serious crashes (TxDOT, 2014). Some typical safety devices on the Texas roadways are shown in Figure 2 as examples. The left one in Figure 2 is a terminal at the end of guardrail, while the right one presents a metal bridge railing.

Figure 2 Typical roadside safety devices in Texas

2.2.3 Crash Analysis on Safety Devices

With increased efficient methods in data analysis, the literature of data analysis on transportation infrastructure to evaluate their performance and give corresponding recommendations or advice are reviewed in this section. Du *et al.* conducted frequent pattern analysis on six specific types of roadside safety devices using crash data. Variables from crash database that may influence the crash results were analyzed such as weather condition, light condition, and speed limit (Du, *et al.*, 2021). In this study, The Apriori and FP-Growth frequent pattern mining algorithms were implemented to identify the variables that have higher impact on crashes. It is concluded by Du *et al.* (2021) that, although median barriers have the highest frequency on crashes, "side of bridge" can statistically result in more serious crashes as an important safety device. A conceptional notation about the methodology in this study is shown in Figure 3.

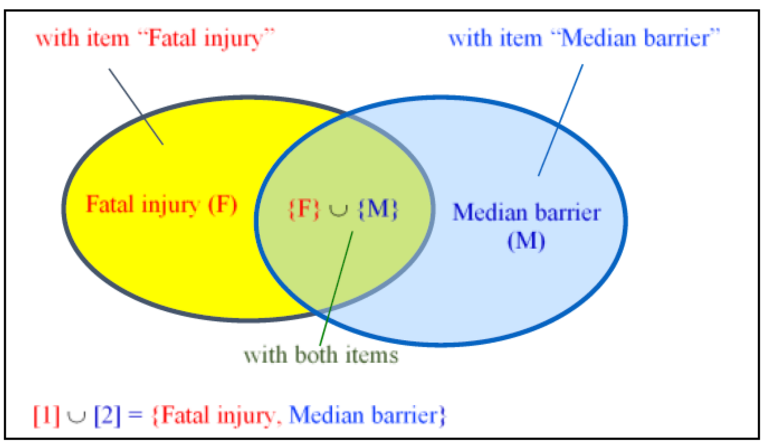


Figure 3 An Example of intersection between device item and crash item

(Source: Du, *et al.*, 2021)

Similarly, another study compared the performance of safety devices between rural and metro areas by analyzing crash data by different device types in two groups. It is recommended considering fair distribution of maintenance and installation of safety devices in rural and metro counties as the conclusion (Wang, *et al.*, 2021). Crash analysis and evaluation on W-beam guardrail was conducted by Gutowski *et al.* (2017) in recent years. The Finite Element (FE) model was applied in this study to simulate and evaluate crash scenarios due to the needs of testing whether the W-beam guardrail are functioning effectively. The result suggested a promising function of the implemented model and some specific size of W-beam guardrail meet the requirement compared to MASH criteria (Gutowski, *et al.*, 2017).

2.3 Machine Learning Modeling

With the fast development of data science technologies, the Machine Learning technique has become a vital part in processing data in not only transportation field but also various implementation sites. The Intelligent Transportation System (ITS) has utilized many computational technologies to allocate suitable approaches and address transportation needs. Different kinds of computer algorithms and models are capable of handling tasks such as regression, classification, pattern recognition, clustering, prediction, etc. A significant number of Machine Learning algorithms have been implemented especially for smart transportation applications (Zantalis, *et al.*, 2019).

One of the examples is to use Machine Learning classifier to predict people's choice for travel modes. Wang *et al.* (2020) analyzed the National Household Travel Survey (NHTS) 2017 database using more than 80 Machine Learning classifiers. In this study, Random Forest, Deep Neural Network (DNN), and Classical discrete choice models were performed to predict travel modes choice. As a conclusion, this research presents the advantages of implementing Machine Learning on public dataset.

Relatively, Rezapour *et al.* conducted Machine Learning based analysis on severity of injury types in crashes involved motorcycles. The Binary Logistic Regression and Classification Tree (CT) models were implemented together to predict injury severity for motorcycle related crashes (Rezapour, *et al.*, 2020). Zhang *et al.* applied Gradient Boosting (GB) and Decision Tree (DT) model to predict crash, thus addressing the grad crossing crashes happened on highway rails. The performances of two Machine Learning models were compared and it is concluded that GB has higher accuracy on the objectives. By effectively training GB model, Zhang *et al.* found several factors impacting the crashes, including travel speed of train, traffic volume of railway and highway, etc. (Zhang, *et al.*, 2020)

2.4 Summary and Research Needs

The important role of transportation equity stays vital throughout the development of transportation history. Along with the continuous policies and actions being proposed focused on equity and EJ, increasing number of research and practices are also highlighting the needs for equity requirements. It is distinctly concluded from the literature that, EJ needs should not be ignored even under the pandemic situation. Since the development of transportation infrastructure can be an indispensable part in transportation planning and operation, when evaluating those infrastructures, considerations related to equity side is essential by comparing the results among different communities.

As a method of assessing whether the particular transportation infrastructure as in roadside safety devices are functioning as expected, crash data analysis is included in multiple previous ISPE studies and practices. However, most analysis conducted through finished projects were statistical analysis with different inputs and outputs. With listed studies utilizing Machine Learning based approaches to process transportation data including crash data, it can be demonstrated that this analysis technology gives more specific results to evaluate the proposed goals effectively. Thus, the needs to fulfil the research gap in ISPE studies are presented. In addition to this, there are few studies associate the performance of roadside safety devices with transportation equity and EJ by dividing the analysis groups into different communities. Therefore, this thesis study analyzes crash data by Machine Learning Modeling and comparing the outcomes in both rural and metropolitan counties in Texas. With solid results presenting the impacts during crashes, the transportation equity can be evidently addressed, and related recommendations can be proposed reasonably.

CHAPTER 3

DESIGN OF THE STUDY

3.1 Methodology

After reviewed related literature on the practices and case studies related to transportation equity on the infrastructures, this study is designed to aim at presenting modeling results to address particular equity problems on roadside safety devices. In doing so, a design of this study is illustrated to present the research procedure and fulfil the study objectives. A designed flowchart as shown in Figure 4 describes the whole process conducted in this study.

From the flowchart of study methodology in Figure 4, the entire procedure mainly consists of data collection, data processing, Machine Learning model selection, and data analysis. After data preprocessing to validate original data for modeling, the additional data are linked with preprocessed crash data. By doing so, more attributes information related to additional database are included in the analysis for broader and more precise consideration. To better address the objective of this study, analysis results through statistical analysis and Machine Modeling analysis are compared. Additionally, a sample amount of data under specific time period was individually analyzed by Machine Learning analysis using selected model to emphasize the impact and difference under pandemic scenario.

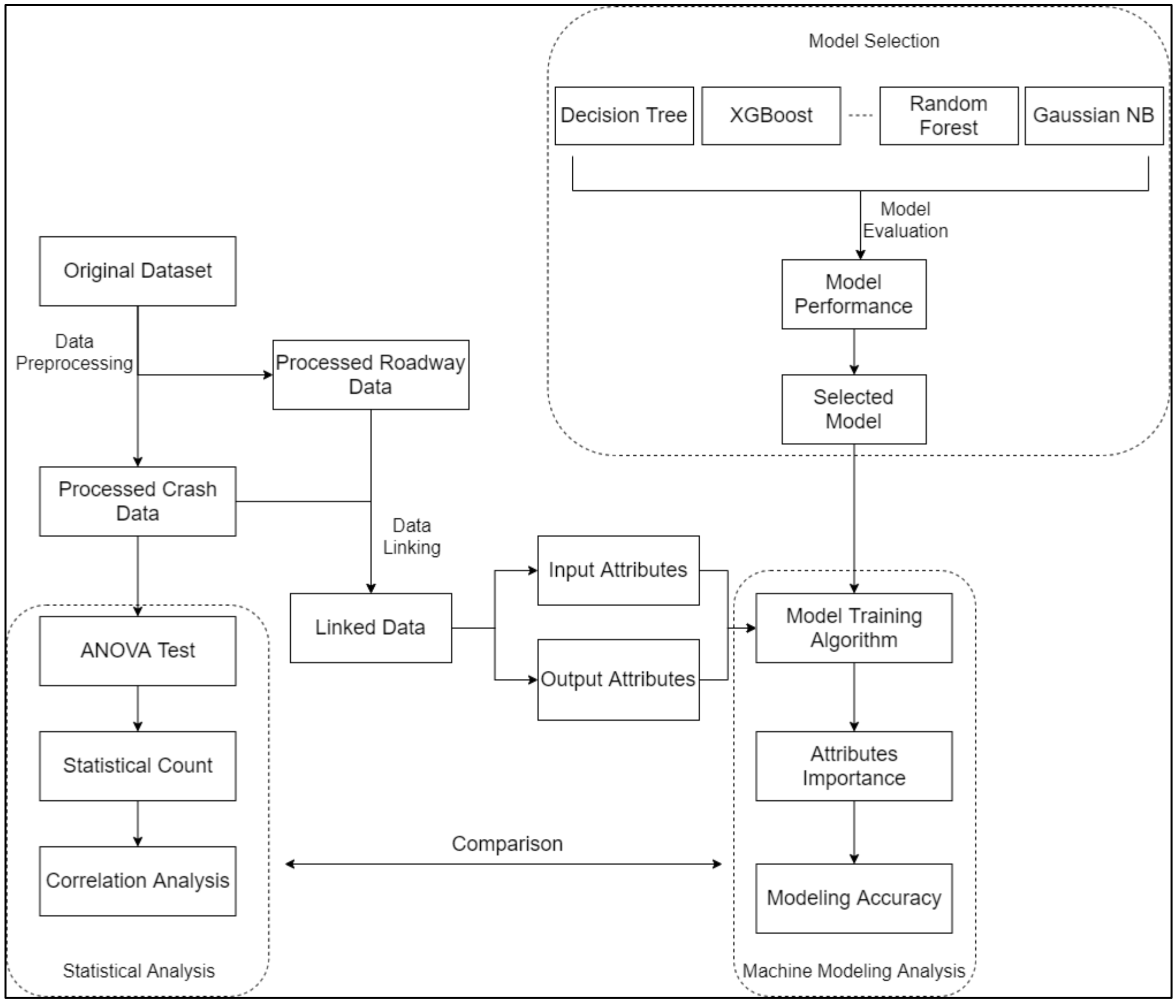


Figure 4 Study methodology flowchart

3.2 Data Collection

The collected data of this study including crash data, county designation data, and roadway data. To begin with, a total of eleven years statewide crash data from January 1st, 2010 to December 31st, 2020 with 5,704,523 crash records and 172 attributes information are gathered from the Crash Record Information System (CRIS). The CRIS is a statewide database containing information of documented traffic crashes related to motor vehicle collected from Texas Peace Officer's Crash

Reports (CR-3) and handled by the TxDOT. It contains all the crash related information. By picking attribute that indicates the hit object during the crash, crashes that involved main types of roadside safety devices were filtered for the study.

According to the OMB, the community groups are divided into metropolitan communities and rural communities. The counties that meet the requirements of majority popularity exceeding 50,000 population and consisting of a core urban area are categorized under metropolitan communities (HRSA, 2021). The designations of Texas rural and metropolitan areas are collected to divide study areas based on 254 counties in Texas. There are 77 counties are designated as metropolitan areas, while the rest 177 are categorized into rural areas (Texas DSHS, 2020). The detailed distribution of rural and metropolitan counties is shown in Figure 5.

By statistically comparing the percentage of crash severity and property damage by crash amount, total population, and area between rural and metropolitan areas, it has been concluded that the total crash number during 2010 to 2019 in rural counties is a lot lower than that in metropolitan counties. However, the percentages of crashes resulted in serious injuries including incapacity injury and fatal injury are higher in rural counties both by amount and by population. Relatively, percentage comparison results of crashes that caused heavier property damage are similar as rural area are higher by crash amount and population (Wang, *et al.*, 2021). With bigger area and lower popularity density in rural areas, the statistical result from the study emphasizes crashes in rural areas may result in serious injury or property damage, regardless of lower total crash number.

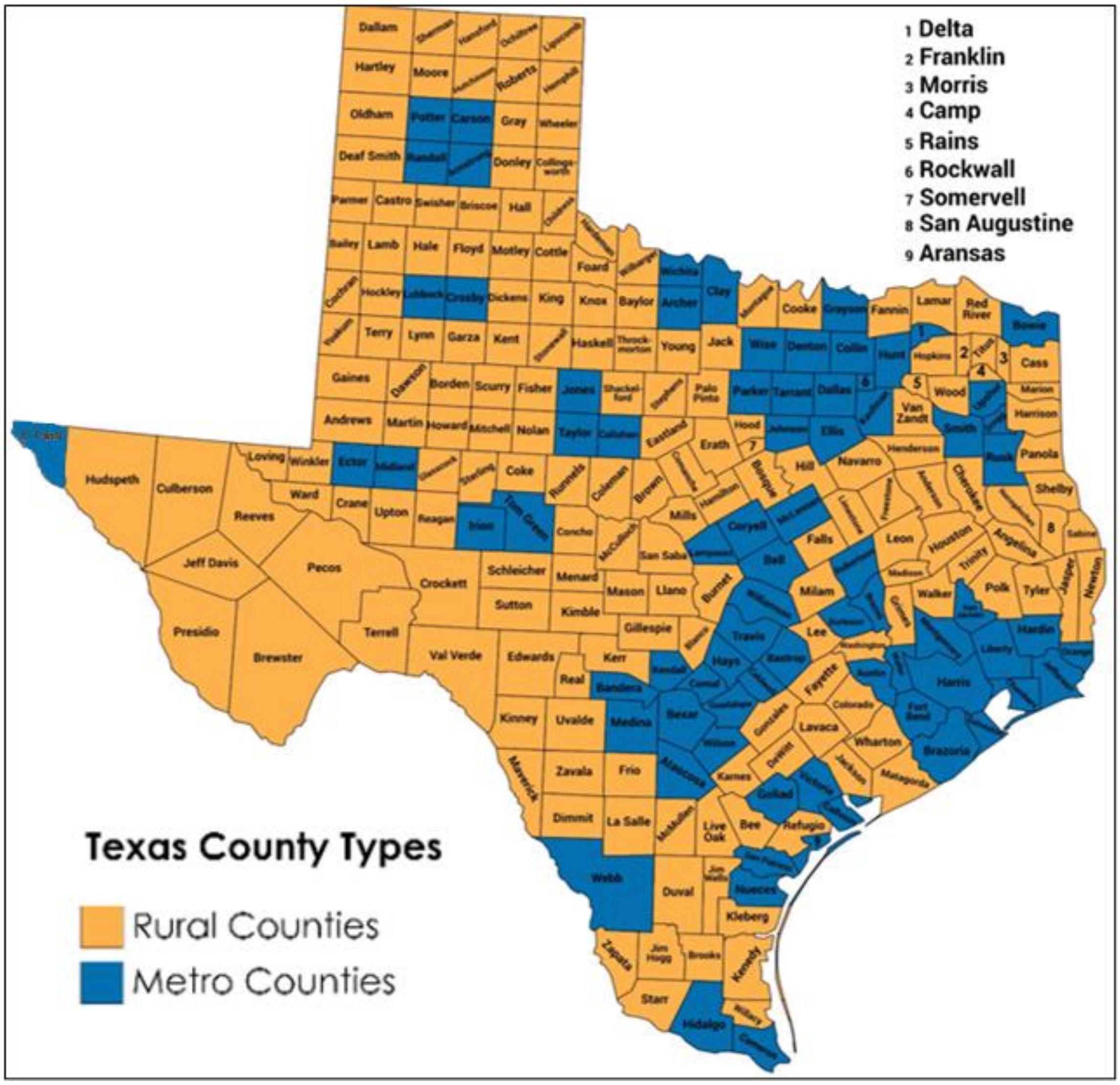


Figure 5 Texas rural and metropolitan counties (Sources: SORH, 2012; Wang, *et al.*, 2021)

Apart from state crash data, in order to include more attributes for analysis in this study, another dataset that have impacts on roadside safety devices related crashes are collected. The geospatial data retrieved from Highway Performance Monitoring System (HPMS) database present documented roadway information annually. Due to the accessibility of current released public HPMS data, seven years of HPMS data from 2011 to 2017 are collected for this study. Several important roadway attributes are included in the original data, including: International Roughness Index (IRI), Annual Average Daily Traffic (AADT), High-Occupancy Vehicle (HOV) types, etc.

3.3 Data Processing

3.3.1 Attributes Filtering

While the latest collected 2020 crash data in CRIS system has been updated with specific description of attributes in each column, older crash data collected from 2010 to 2019 are still in previous format with IDs and abbreviated headers. In addition, there are columns appearing to be with no data in both crash data and roadway data sets, which also need to be replaced with certain values for further Machine Learning modeling. Thus, a step of data preprocessing is required before proceeding.

Some attributes related to roadside safety devices that may impact the performance evaluation process have been selected and filtered in this section. Thirteen attributes including their IDs are listed in Table 1 and Table 2 (except for Crash ID, Crash Speed Limit, and county ID), where three types of struck object with IDs 23, 39, 41 are listed as selected roadside safety devices in this study to divide crashes involved different safety devices. Attribute “$1,000 Damage to Any

One Person's Property" and "Crash Severity" present the results of crashes as property damage and personal safety side.

Table 1 Selected Attributes from CRIS Crash Data

$1,000 Damage to Any One Person's Property	**Roadway Part**	**Weather Condition**	**Light Condition**	**Bridge Detail**
0- no 1- yes	1- main/proper lane 2- service/frontage road 3- entrance/on ramp 4- exit/off ramp 5-connector/flyover 7- other	0- unknown 2- rain 3- sleet/hail 4- snow 5- fog 6- blowing sand/snow 7- severe crosswinds 8- other 11- clear 12- cloudy	0- unknown 1- daylight 2- dawn 3- dark, not lighted 4- dark, lighted 5- dusk 6- dark, unknown lighting 8- other	1- vehicle retained on bridge or overpass 2- vehicle went through rail 3- vehicle went over rail 4- crash involved underpass 5- vehicle went between parallel structures 6- structure not hit 8- not applicable

Table 2 Selected Attributes from CRIS Crash Data (Cont.)

Surface Condition	**Object Struck**	**Crash Severity**	**Base Type**	**Median Type**
0- unknown 1- dry 2- wet 3- standing water 5- slush 6- ice 8- other 9- snow	23- guardrail 39- median barrier (concrete or cable) 41- side of bridge (bridge rail)	-1- unknown 1- incapacitating injury/suspected serious injury 2- non-incapacitating injury 3- possible injury	-1- no data 1- roadbed soil 2- flex base (granular) 3- stabilized earth or flex (granular) 8- asphalt base (hot mix, asphalt concrete)	-1- no data 0- no median 1- curbed 2- positive barrier 3- unprotected 4- one-way pair

10- sand, mud, dirt		4- killed 5- not injured	9- concrete	

As is shown in Table 1 and Table 2, values that represent no data under selected attributes are replaced with “-1”. “Yes” and “No” binary value under attribute “$1,000 Damage to Any One Person’s Property” are replaced with “1” and “0” to indicate whether there is over or less $1,000 Damage to any One Person’s Property in a crash. IDs for crash severity were replaced with the Equivalent Property Damage Only (EPDO) weights in the reference of crash costs analysis conducted by FHWA (Harmon, *et al.*, 2018). The assigned EPDO weights are decided by the societal costs of crashes under different severity levels, the crash costs under target crash severity are divided by Property Damage Only (PDO) costs to calculate the EPDO weights for the target crash severity (AASHTO, 2010; Wemple, *et al.*, 2014). There is a total of six types of crash severity replaced as calculated EPDO weights: (1) “-1” for unknown records; (2) “1” for no injury; (3) “6” for possible injury; (4) “11” for non-incapacitating injury; (5) “30” for incapacitating injury; and (6) “568” for fatal injury.

Attributes from HPMS roadway database contain the information of a part of roadway and contribute to crashes are selected. Ten attributes including their IDs in the HPMS data are selected and listed in Table 3 and Table 4 by looking up the HPMS Field Manual (FHWA, 2016). Values that represent no data under selected attributes are replaced with “-1”.

Table 3 Selected Attributes from HPMS Roadway Data

AADT	Functional System	Facility Type	HOV Lanes	HOV Type
-1- no data 0 20 50 60 … 326,677	-1- no data 0- unknown 1- interstate 2- principal arterial – other freeways and expressways 3- principal arterial – other 4- minor arterial 5- major collector 6- minor collector	-1- no data 0- unknown 1- one-way roadway 2- two-way roadway 4- ramp 5- non mainline 6- non inventory direction 7- planned/unbuilt	-1- no data 0 1 2	-1- no data 0- unknown 1- full-time managed lanes 2- part-time managed lanes 3- part-time managed lanes

Table 4 Selected Attributes from HPMS Roadway Data (Cont.)

International Roughness Index	National Highway System	Strategic Highway Network Type	Toll Charged	Toll Type
-1- no data 0 1 2 16 … 554	-1- no data 0- unknown 1- non collector NHS 2- major airport 3- major port facility 4- major amtrack station 5- major rail/truck terminal 6- major intercity bus terminal 7- major public transportation or multi-modal passenger terminal 8- major pipeline terminal	-1- no data 0- unknown 1- regular strategic highway network 2- connector	-1- no data 0- unknown 1- toll charged in one direction only 2- toll charged in both directions 3- no toll charged	-1- no data 0- unknown 1- has toll lanes but no special tolls (e.g., HOT lanes) 2- has HOT lanes 3- has other special tolls

	9- major ferry terminal			

After selecting attributes that may impact the crashes related to chosen roadside safety devices from both databases, abbreviations of each attribute are presented for the convenience of plotting in the following data analysis. The listed abbreviations are shown in Table 5.

Table 5 List Abbreviations for Selected Attributes

Selected Features	**Abbreviation**
Roadway Part	RRP
Crash Speed Limit	CSL
Weather Condition	WC
Light Condition	LC
Surface Condition	SC
Bridge Detail	BD
Road Base Type	BT
Median Type	MT
Functional System	FS
Facility Type	FT
HOV Lanes	HL
HOV Type	HT
International Roughness Index	IRI
National Highway System	NHS
Strategic Highway Network Type	ST
Toll Charged	TC

Toll Type	TT

3.3.2 Data Linking

To include both crash and roadway attributes in the Machine Learning modeling analysis, seven years of filtered CRIS crash data and HPMS roadway data are linked accordingly by year. Since both data have geospatial attributes, Distance from Origin (DFO) attribute in crash data, Begin and End Point attributes in roadway data both represent the reference marker. It used locations of existing milepost and a listing equivalent to a mile post as a control document during the identification and installation process (Randall, 2005).

By implementing Spatial Join through closeness factor in ArcGIS, the original database is uploaded to the ArcGIS platform to create a feature layer for loading data and avoiding conflicting coordinate systems issues. A search radius is set to perform spatial join with acceptable error range. For every target point on the map, a circle is drawn and the nearest joinable object within that radius is selected and attached to the target dataset. In this case, the HPMS roadway data is directly imported into the platform to link with CRIS crash data. Each year of linked data from 2011 to 2017 contain selected attributes from both original datasets.

3.4 Model Selection

Machine learning modeling is conducted by choosing candidate Machine Learning models, such as: Random Forest, Decision Tree, Multi-layer Perceptron

(MLP) Classifier, KNeighbors Classifier, Gaussian Naive Bayes, and XGBoost (Ahmed, *et al.*, 2010). The *k*-Fold Cross-Validation is implemented to evaluate accuracy of each candidate models under different output (Burman, 1989). Machine Learning models with the highest accuracy under different main safety devices are selected for modeling. The importance ranking of chosen input attributes is presented after training the selected Machine Learning model, while the results are discussed for each input/output and safety devices set. The last part of Machine Learning modeling is model evaluation, which assesses the modeling results accordingly.

3.4.1 Candidate Machine Learning Models

After collected related and properly processed related data, six Machine Learning based models are selected as candidate Machine Learning models for modeling data analysis. The description of the six models and their fundamental principles are introduced in this section.

Decision Tree Model. When considering candidate models for Machine Learning based analysis, Decision Tree Model is always been widely used to evaluate output features through inputs (Drummond *et al.*, 2005). The model refers to a tree-like procedure with multiple nodes and paths as binary trees, which conducts evaluation based on the decision rules that learned from original data. It implements regression and classification to pick important features. For example, the classifier of this model has the function of recognizing patterns of features and is applicable of selecting features (Safavian & Landgrebe, 1991).

Random Forest Model. Based on a single Decision Tree model, the Random Forest model is developed relatively. It selects features with the results of a set of Decision Tree models (Belgiu & Drăguţ, 2016). By collecting and summing results from multiple Decision Trees, the outcome of Random Forest is finalized. It also rises the shortage that, it is hard to ensure the performance by concluding various Decision Trees results, and a model compression is needed sometimes (Madeh Piryonesi & El-Diraby, 2021). Random Forest model can also be used to decide the feature importance, which is calculated through Gini index shown in Equation 1.

$$G = 1 - \sum_{i=1}^{n} f^2 \tag{1}$$

where, G is the Gini index,

n is the number of candidate features, and

f is the feature frequency.

Gradient Boosting Model. Gradient Boosting model also has the function of regression and classification as one of the Machine Learning models, which is an ensemble of prediction models with weak outcome such as Decision Trees (Hastie, *et al.*, 2009). As an extreme Gradient Boosting model, the XGBoost model is developed by Chen and Guestrin, while the basic algorithm is the Decision Tree featured as scalable tree boosting system. (Chen & Guestrin, 2016). It tends to have higher running speed and usually has better performance than traditional Machine Learning models. By minimizing the value followed by Equation 2 and Equation 3, the XGBoost model builds trees through a loss function.

$$L(\emptyset) = \sum_i l(\hat{y_i}, y_i) + \sum_k \Omega(f_k) \tag{2}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3}$$

where, y_i is the target feature in the given dataset,

$\hat{y_i}$ is the prediction through the model,

l is the convex loss function,

L is the pseudo residuals of predicted feature value,

f is the space of built regression trees,

Ω is the regularization term used to reduce complexity of prediction model,

γ is the user-definable penalty,

T is the number of leaves in the built trees, and

w is the weights of leaf in the built trees.

Multilayer Perceptron Model. As one of the most common Neural Network model in Machine Learning modeling, Multilayer Perceptron Model (MLP) is one type of Neural Network that used to solve simple regression problems. Multiple layers are components or nodes that coordinated with each other and form the MLP model (Pal & Mitra, 1992). Except from input and output layers, one or more hidden layers are also included in the MLP model. Due to its non-linear and couple of layers features, the MLP model can be implemented in non-linear data. It has relatively lower training time compared with other complex models (Car, *et al.*, 2020). The MLP model also serves as both classifier and regressor. According to the concluded equation by Nicholson, the relationship between inputs and outputs are shown in Equation 4 (Nicholson, 2020):

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(w^T x + b) \tag{4}$$

where, x is the inputs vector,

w is the weights vector,

b is the bias, and

φ is the non-linear activation function.

Naive Bayes Model. By applying the Bayes' theorem with higher independence assumptions among features, a set of probabilistic classifiers called Naïve Bayes (NB) Model can achieve higher accuracy (Hastie, *et al.*, 2009). the probability calculation for each assumption through NB modeling are simplified to make their calculations easy to follow. The independence is too high as strong or naïve state that is not usual in actual data (Brownlee, 2016). One of the extended NB models is the Gaussian NB model that estimates the mean and the standard deviation with real-value inputs. The probability density is calculated through the training of Gaussian NB model in Equation 5, where the mean and variance of input attributes are computed.

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}} \tag{5}$$

where, x is the attributes in the given dataset,

C_k is the assigned class for the attributes,

μ_k is the mean of the values,

σ_k^2 is the Bessel corrected variance of the values, and

p is the possibility density.

K-Nearest Neighbors Model. The K-Nearest Neighbors (KNN) model can be used both in classification and regression side. It was developed as a non-parametric classification method in early 19s and extended for regression also in 1992 (Altman, 1992). The KNN model focuses on similarity measure of training data, so it needs to calculate the similarity of all training data to select nearest k neighbors. Thus, the disadvantage of KNN model is longer running time when processing large amount of data, and the cost is relatively high when implementing KNN for big data (Deng, *et al.*, 2016). The distance measure through KNN model is valid for continuous variables only and can be computed in three ways of distance measuring, including: (1) Euclidean distance measuring, (2) Manhattan distance measuring, and (3) Minkowski distance measuring (Sayad, 2010). The calculations are shown in Equation 6, Equation 7, and Equation 8, respectively.

$$d_E = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{6}$$

$$d_{Ma} = \sum_{i-1}^{k}|x_i - y_i| \tag{7}$$

$$d_{Mi} = \left(\sum_{i-1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \tag{8}$$

where, d_E is the Euclidean distance,

d_{Ma} is the Manhattan distance,

d_{Mi} is the Minkowski distance,

k is the number of nearest neighbors,

x and y are the attributes in the given data set, and

q is an integer as the Minkowski distance of order.

3.4.2 Model Performance Measure

To finalize which candidate Machine Learning model is the most suitable one for feature selection in Machine Learning modeling, the performance measure for selected models is needed in the data analysis procedure. The *k*-Fold Cross-Validation has the function of assessing the accuracy of a group of Machine Learning models and provides understandable results for model selection. Cross-validation has the convenience of only requiring mild distributional assumptions for the modeling data and it does not need the detailed features of each evaluated model such as their model dimensions (Yang, 2007). In this statistical evaluation method, *k* often remains a fix number but 10-Fold Cross-Validation is the most widely used one in related situations (McLachlan, *et al.*, 2005).

For the 10-Fold Cross-Validation when *k* is selected as 10, given data set is randomly spitted into 10 sampled sets with the same size. A single sample set is picked up to test the selected model, while the remaining 9 datasets are assigned as training data. The evaluating model was fitted to the other 9 parts of sampled data, and the prediction error of the fitted model is calculated. The procedure of cross-validation is repeated by 10 times and the results of these prediction errors are combined to give final performance result (Hastie, *et al.*, 2009). The 10-Fold Cross-Validation prediction error estimation is computed through Equation 9.

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-10(i)}(x_i)) \tag{9}$$

where, CV is the estimate prediction error by cross-validation,

$\hat{f}$ is the predicted evaluating model,

N is the partition of allocated observations,

i is the number of fitted observations, and

L is the loss function.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Statistical Analysis

4.1.1 ANOVA Test on Statistical Significance of Crash Data

Before detailed analysis on various groups of crash data, the input data need to be verified comparable and have significant differences. Analysis of Variance (ANOVA) test was performed to check the statistical significance of differences on crash data so that the comparison between different groups makes sense. By comparing groups of data under variable, the significant probability (P-value) is conducted and the invalid assumptions as null hypothesis will be returened when the P-value is lower than the identified level of significance (Kim, 2017). In addition, statistic F value is also a result of ANOVA test that shows the ratio of between group variances and within group variances as a reference.

Thus, the significant differences between selected groups will be justified and the crash data can be proceeded in the following analysis. In this case, different groups based on types of safety devices involves only one independent attribute and One-way ANOVA test (Howell, 2012). This particular testing method was performed in this section, while attributes in crash data are coordinated with selected attributes in Table 1 and Table 2.

A significance level threshold was set as 0.05 to test P-value under different attributes. According to the results shown in Table 6 and Table 7, most P-values were very close to zero, and the highest P-value (7.7E-62<<0.05) was under attribute “Light Condition”. With most light conditions for crashes related to each safety device are daylight, the comparison of light condition in crash data appears to be the least significant different for the six roadside safety devices. However, since all the P-value are far less than 0.05, comparisons of crashes related to different groups of safety devices under selected ten attributes show significant difference and are comparable.

Table 6 ANOVA Test Result for Crash Data Attributes

Attribute	$1,000 Damage to Any One Person's Property	Roadway Part	Weather Condition	Light Condition	Bridge Detail
P-value	8.3E-142	0	0	7.7E-62	0

Table 7 ANOVA Test Result for Crash Data Attributes (Cont.)

Attribute	Surface Condition	Crash Speed Limit	Crash Severity	Base Type	Median Type
P-value	0	0	7.4E-122	0	0

Another ANOVA test was conducted based on the severity of crashes, the crash data was divided into two groups with higher or lower property damage according to attribute “$1,000 Damage to Any One Person's Property”. The values in each group are “Object ID” from attribute “Object Struck” to represent the types of selected roadside safety devices. To verify the analysis results when considering property damage and crash severity as outputs and safety device type as input are comparable, the F statistic and P-values of each crash severity type are shown in Table 8.

Table 8 ANOVA Test for Safety Devices by Crash Severity Type

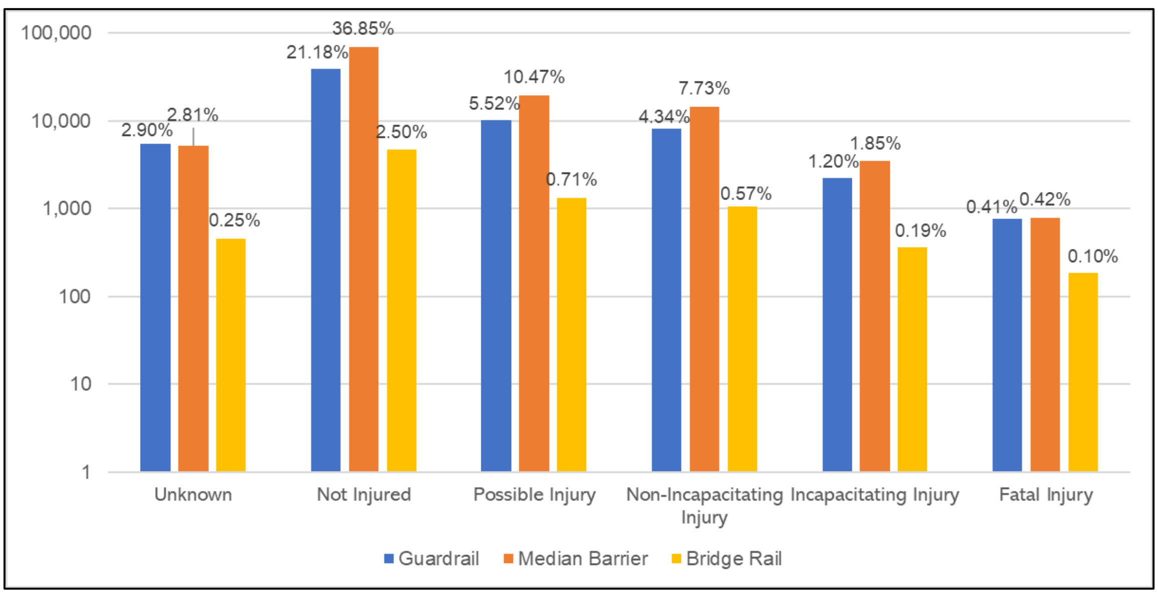
Crash Severity	No Injury	Possible Injury	Non-incapacity	Incapacity	Fatal	Unknown
P-value	4.3E-15	0.00019	0.022	0.011	0.22	0.23
F	61.5	13.9	5.3	6.5	1.5	1.4

The F statistic was included in this ANOVA test to present the significant difference of safety device types with different property damage under each crash severity type. Higher F statistic corresponding to lower P-value and higher significant difference. As the result shown in Table 8, the P-value under fatal and unknown crashes are higher than significance level and the F statistic are lower than 2. Since fatal and unknown type of crashes only take up 6.4% of whole crash records, the significant of safety devices are limited. In these two types of crashes, crashes related to guardrail and median barriers all have the highest portion, and significances of difference are lower due to the data size limitation.

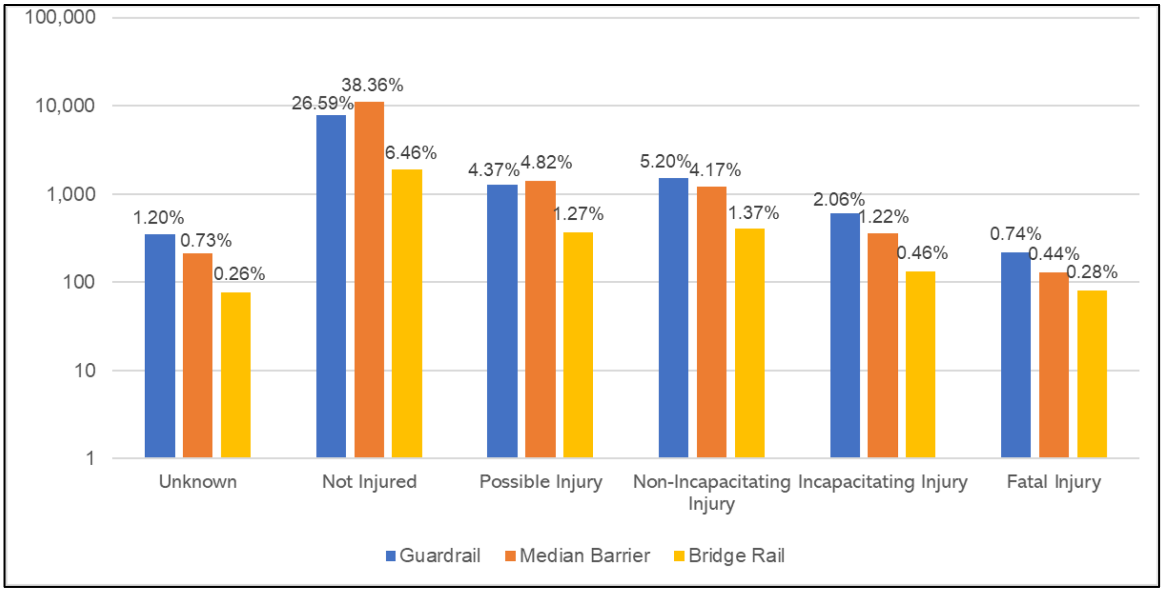
Apart from fatal and unknown crashes, the P-values under other crash severity type are lower than the significance levels and the data groups are comparable. Larger size of crash data under specific crash severity tends to be more significantly different except for non-incapacity crashes. Compared to incapacity crashes, although the number of crashes resulted in non-incapacity injuries is larger, the P-value is higher and F statistic is lower than incapacity crashes. The safety devices type under incapacity crashes have higher significant difference regardless of its number of crashes. In conclusion, selected grouped crash data under selected attributes are generally comparable and suitable for further analysis.

4.1.2 Statistical Count of Crash Data Attributes

After ensuring the statistical significance of attributes in CRIS crash data satisfies the analysis requirement as statistical comparable, basic count of crash number by crash severity and property damage type are conducted to demonstrate overall statistical differences. The counting results are divided into rural and metropolitan groups for each roadside safety device. Distribution of crash amount by crash severity for three selected safety devices is shown in Figure 6.



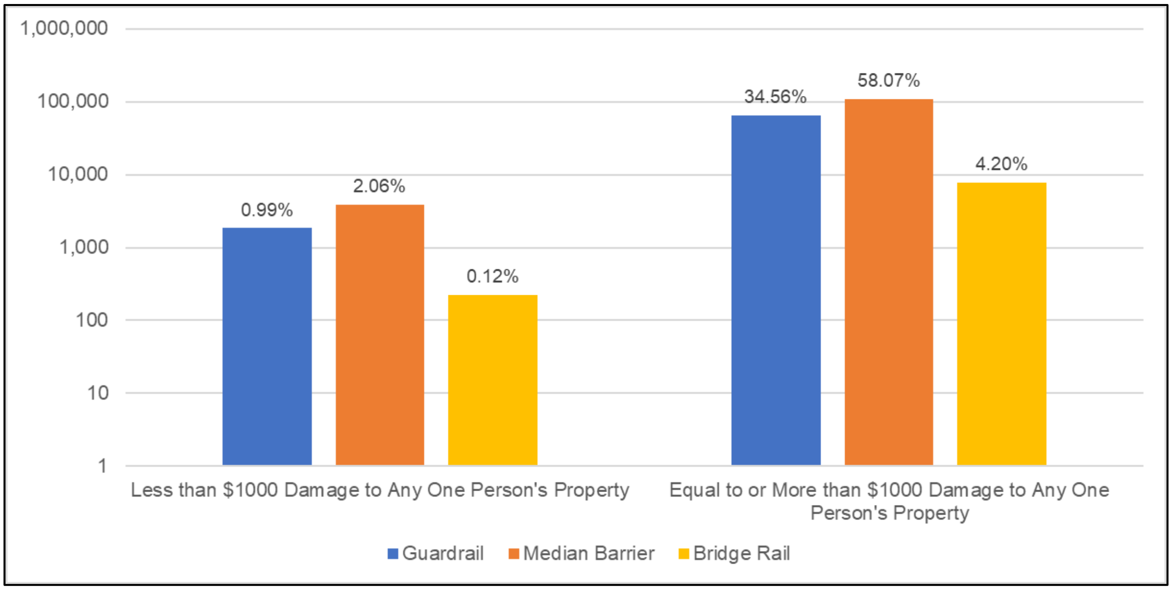
(a) Metropolitan Areas



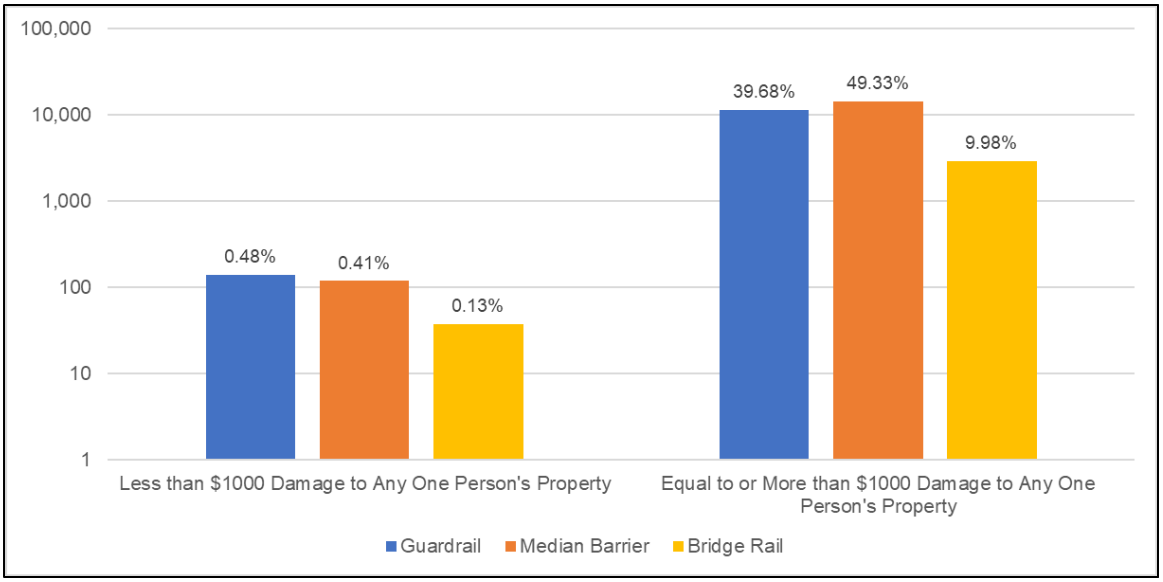
(b) Rural Areas

Figure 6 Statistical distribution of crash severity by roadside safety devices in Texas from 2010 to 2019

As is shown in Figure 6, the crash distribution in Metropolitan and rural counties are presented. The labels show the distribution percentage out of total crash number related to three selected safety devices. It can be concluded from Figure 6 that, median barriers have the highest crash amount among all the areas in ten years of Texas crash records. The percentage of crashes ended up causing fatal injuries demonstrates the proportion of each roadside safety device compared to the total crash number. It clearly suggests from the statistics that, although crash frequency in rural areas is much lower compared to metropolitan areas, the percentage of fatal crashes related to each safety device is higher in rural counties. In addition, guardrail involved more fatal crashes in rural areas. while in metropolitan areas, median barriers tend to result in more fatal injuries. Similarly, the distribution for property damage is shown in Figure 7.



(a) Metropolitan Areas



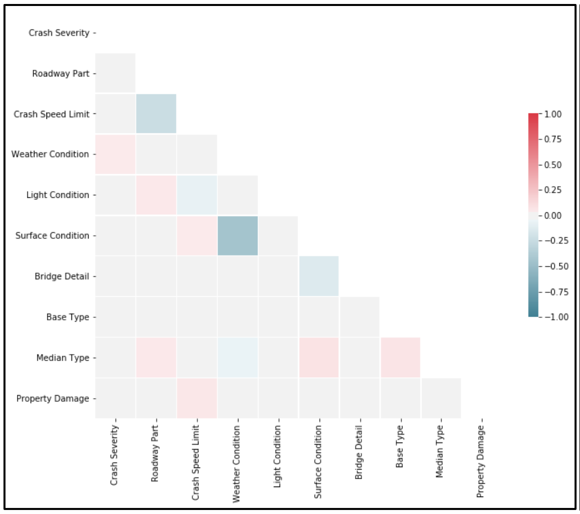
(b) Rural Areas

Figure 7 Statistical distribution of property damage by roadside safety devices in Texas from 2010 to 2019

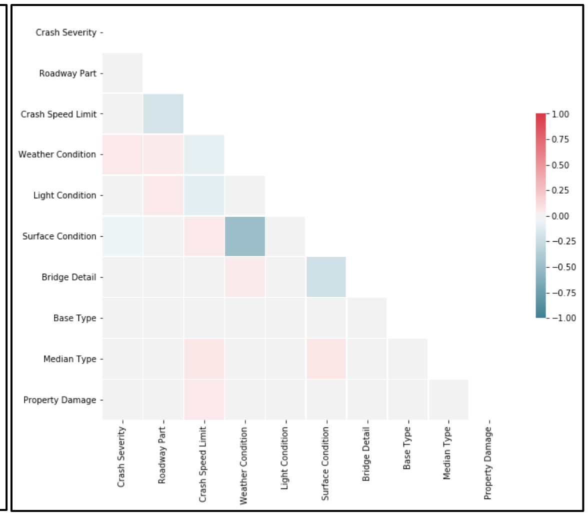
When focusing on another attribute that reflects the property damage in crashes, crash distribution of whether over $1,000 damage was resulted in a crash by safety device is shown in Figure 7. The labels show the distribution percentage out of total crash number related to three selected safety devices. According to the percentage of crashes that caused heavier property damage, median barriers have the highest distribution in both areas. However, the percentage of median barrier crashes that caused over or equal to $1,000 property damage is reduced in rural counties, while other three safety devices all have higher proportion of related serious crashes in damaging property.

4.1.3 Correlation Analysis on Crash Data Output Attributes

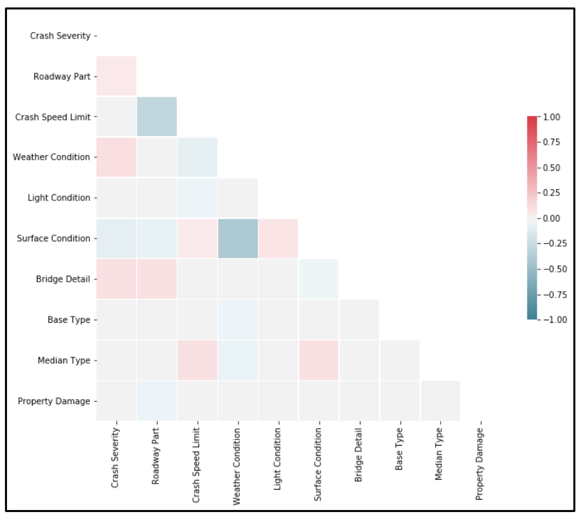
Apart from identifying statistical significance and counting crash distributions, a correlation analysis is conducted to present the impact of selected attributes on the injury type and property damage in crashes. Each selected safety device is individually analyzed with a corresponding diagonal correlation matrix. Ten years of Texas crash data from 2010 to 2019 are correlated in groups. “Crash severity” and “$1,000 Damage to Any One Person's Property” attributes are regarded as output variables to demonstrate the relationship between input attributes and output attributes in statistical analysis side. Three diagonal correlation matrixes in corporation with three studied safety devices are presented in Figure 8.



(a) Guardrail



(b) Median Barrier



(c) Bridge Rail

Figure 8 Correlation analysis results between attributes by safety devices

From the correlation results, the correlation between input attributes and output attributes varies with safety devices. The crash speed limit has relatively higher correlation with crash severity attribute for median barrier and bridge rail. The weather condition however, is an important parameter when considering property damage outcome for each safety device. Some attributes have impacts over the attributes for specific safety devices according to the matrix. For instance, the roadway part and bridge detail have relatively higher correlation relationship between crash severity in bridge rail related crashes. While other attributes except from weather condition do not show major relationship between severity of crashes in other two safety devices.

4.2 Machine Learning Modeling Analysis

The Machine Learning modeling analysis is conducted by comparing the performance of candidate Machine Learning models, while the candidate model with highest accuracy score through 10-Fold Cross-Validation is selected to perform Machine Learning based data analysis. The modeling analysis is conprised of two parts where one part focused on linked data ranged from 2011 to 2017 acquired by data processing part. The attributes from both roadway data and crash data are included. Another analysis procedure assessed the impact of attributes under COVID-19 pandemic period using crash data in year 2020, highlighting the transportation equity under pandemic situation.

4.2.1 Modeling Analysis on Linked Data

Model Performance. In order to choose the most suitable model, the repeated 10-fold cross validation was applied to assess the performance of selected six candidate models with the crash severity and property damage outcome. The linked data with roadway and crash information are divided into three groups for different specific safety device in rural crashes and metropolitan crashes. The results of model performance evaluation based on different output attributes are shown in Table 9 and Table 10.

Table 9 Accuracy Scores for Candidate Models on Crash Severity in Linked Data Analysis

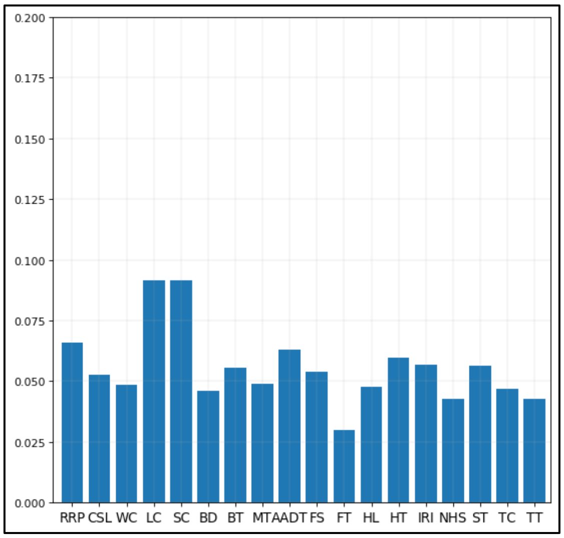
Area	Decision Tree	XGBoost	Random Forest	MLP Classifier	Gaussian NB	K-Nearest Neighbors
Rural	0.432	0.567	0.532	0.450	0.582	0.535
Metropolitan	0.457	0.607	0.558	0.547	0.602	0.560

Table 10 Accuracy Scores for Candidate Models on Property Damage in Linked Data Analysis

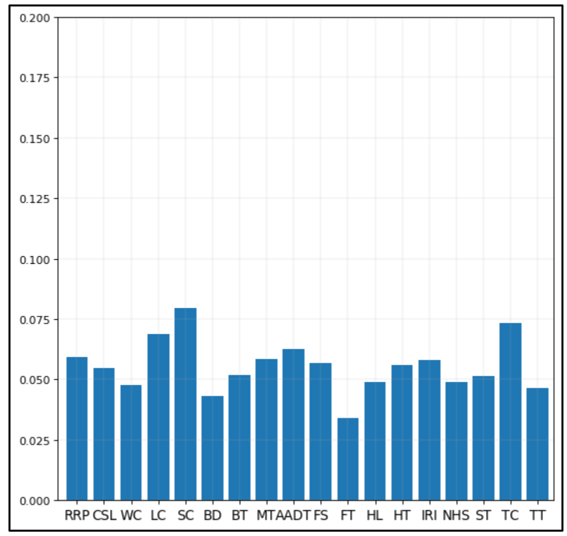
Area	Decision Tree	XGBoost	Random Forest	MLP Classifier	Gaussian NB	K-Nearest Neighbors
Rural	0.928	0.963	0.962	0.918	0.928	0.963
Metropolitan	0.954	0.978	0.976	0.975	0.954	0.978

As the results shown in Table 9 and Table 10, the highest accuracy scores are shown in red colors for both output attributes. XGBoost has highest accuracy scores on crash severity and property damage output. The performance of candidate Machine Learning Models are relatively better on property damage and in metropolitan data. Thus, XGBoost is selected as the Machine Learning model in this section.

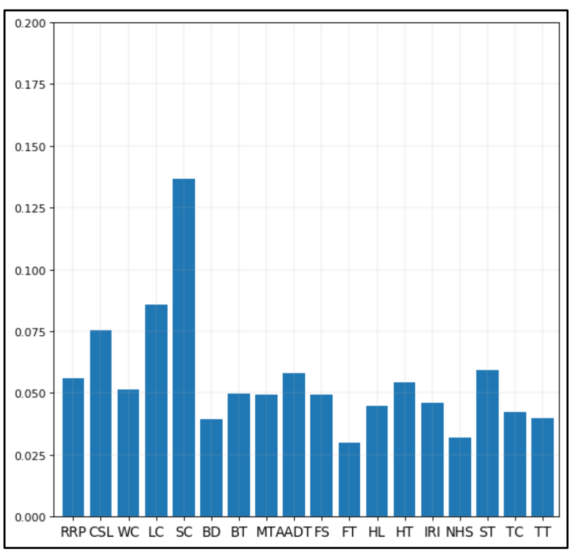
Feature Selection. After choosing the model with highest performance accuracy, the selected Machine Learning model is used to train crash data, and to individually analyze the importance of attributes regarding crash results. The crash data is divided into six groups with three safety device types and two areas. XGBoost with the best model performance was used to train the six data sets. The feature selection is performed during modeling process to rank the impact level of selected attributes on crash severity or property damage. The feature selection results through modeling on crash severity is shown in Figure 9.



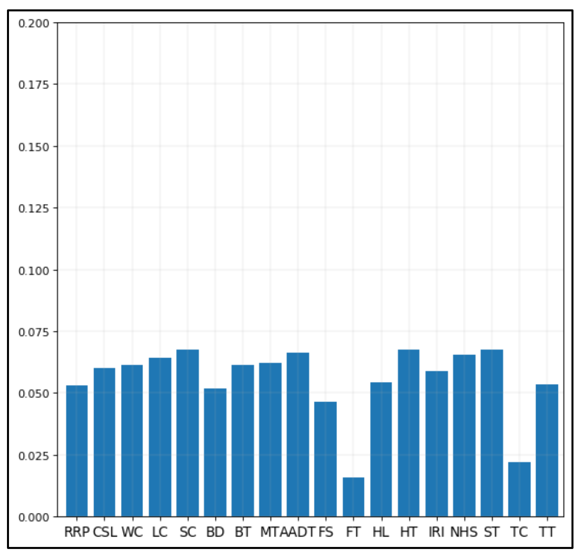
(a) Guardrail–Metro



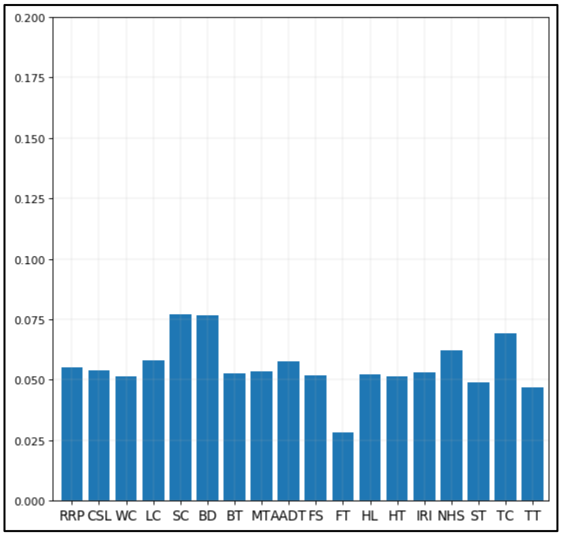
(b) Guardrail–Rural

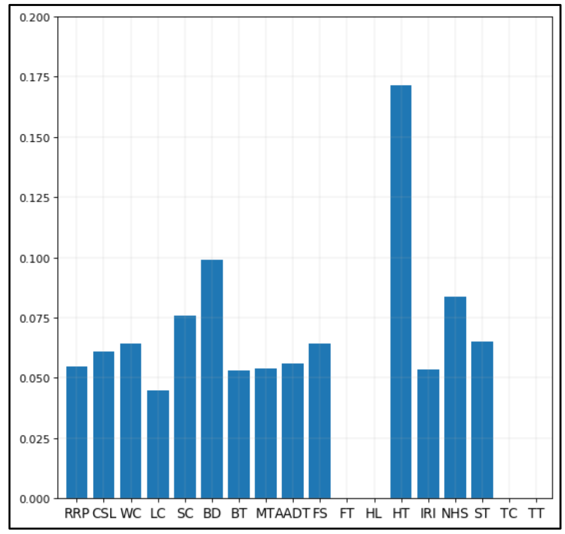


(c) Median barriers–Metro



(d) Median barriers–Rural





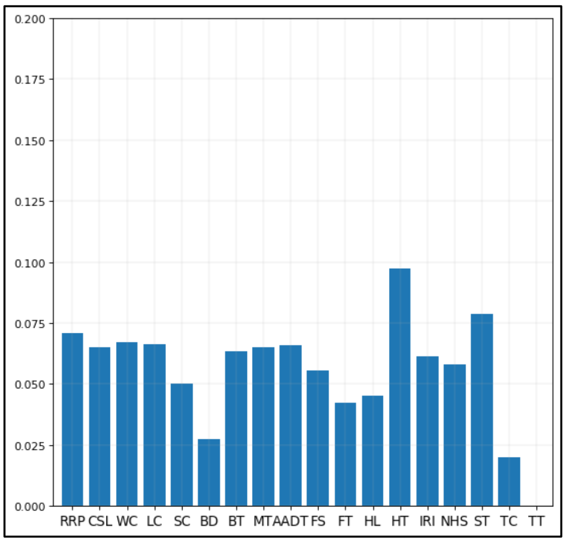
(e) Bridge Rail-Metro (f) Bridge Rail-Rural

Figure 9 Linked data attributes importance ranking on crash severity by XGBoost model for each safety device

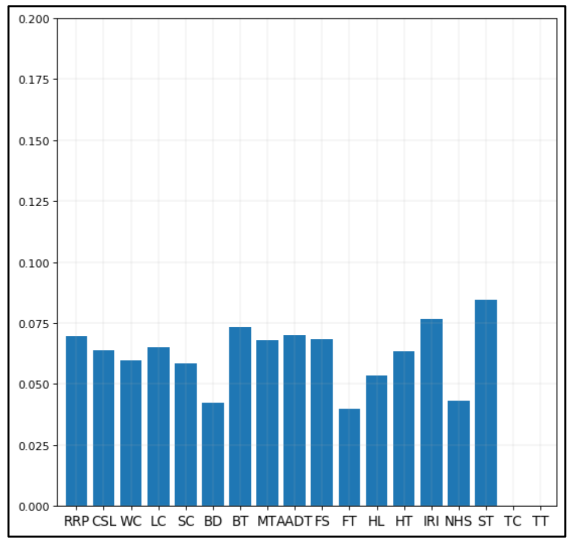
According to the results presented by the attributes importance ranking through the Gradient Boosting modeling, the impacting attributes on the severity of injuries during crashes are diversified on the specific type of safety devices. Generally speaking, by comparing different areas, it can be demonstrated that in metropolitan areas, crash attributes tend to have more influence on the crash severity compared to roadway attributes. Surface condition and light condition are especially having higher importance scores in metropolitan counties. The attributes from roadway database have minor impact crash severity such as AADT and IRI in metropolitan areas.

When focusing on rural areas, the impact from crash attributes is decreasing significantly. Importance of attributes become more similar especially on guardrail and median barrier related crashes which also suggests the rising importance of roadway attributes. Relatively, HOV types have the highest importance rank on rural bridge rail crashes, the impact of bridge detail information has also increased compared to metropolitan areas.

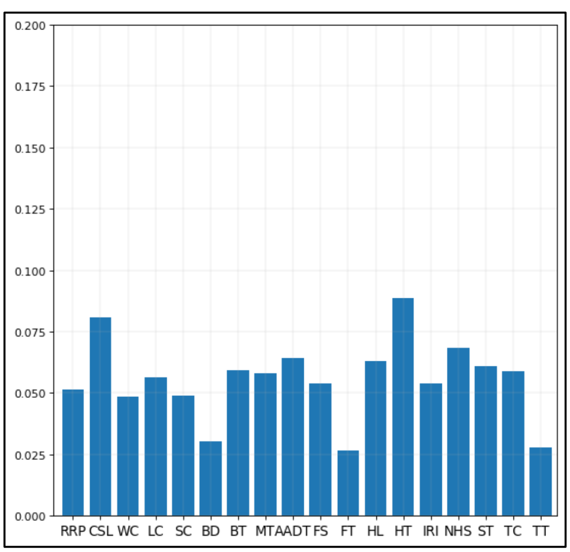
To emphasize the attribute impacts on the property damage, the same datasets are trained again with another output. The analysis results are shown in Figure 10.

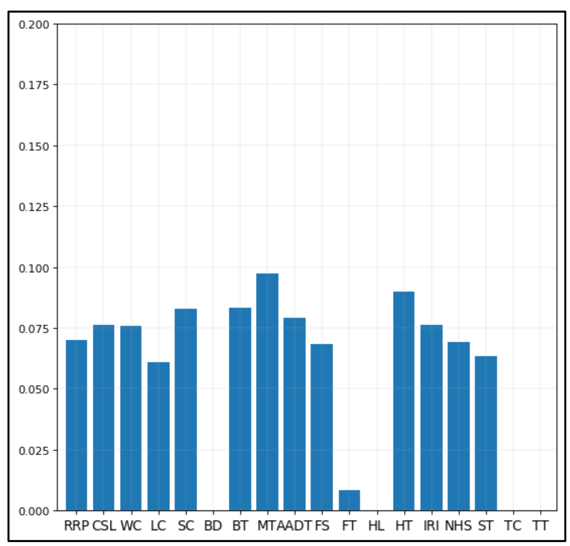


(a) Guardrail–Metro

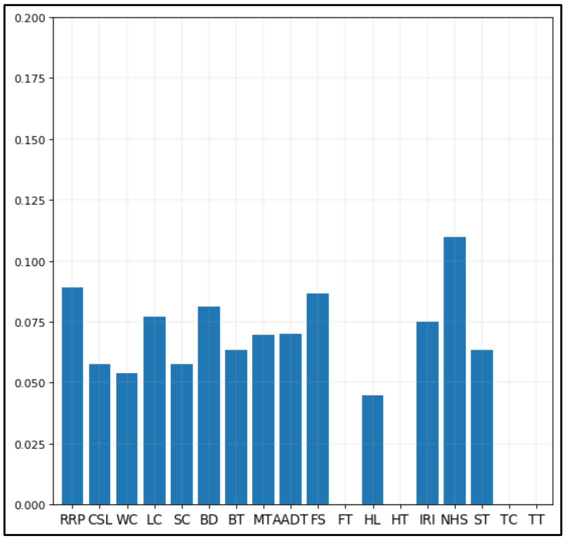


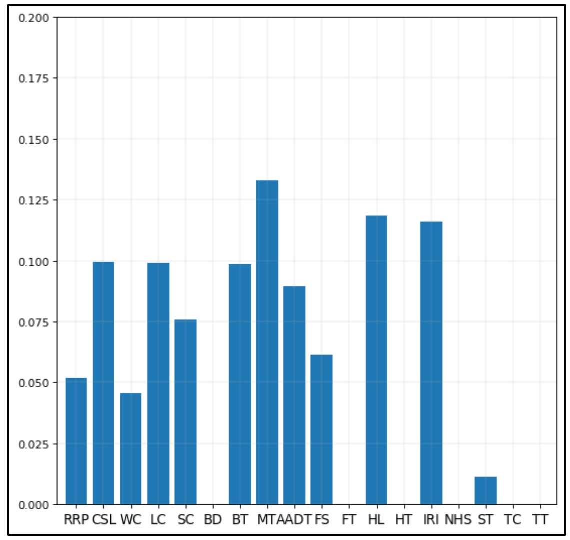
(b) Guardrail–Rural





(c) Median barriers–Metro (d) Median barriers–Rural





(e) Bridge Rail-Metro (f) Bridge Rail-Rural

Figure 10 Linked data attributes importance ranking on property damage by XGBoost model for each safety device

Since property value amount damaged also reflects the performance of roadside safety devices, and the modeling result that set property damage as output is also analyzed to assess the performance difference. In metropolitan areas, the attributes with the highest important score on the crash result for each selected safety device are from roadway attributes. More specifically, types of HOV lanes impact the performance of both guardrail and median barriers on property damage in metropolitan counties. Highway systems which specify located transportation infrastructure affect the performance of bridge rail.

The results in rural areas have some differences as median type is the major affecting attribute for median barriers and bridge rail. As for guardrail, IRI and highway network type rank the highest as important parameters in rural crashes. Although some attributes from crash conditions increase the importance scores in rural areas, the major impacting attributes are focused on transportation infrastructure and roadway design instead of crash conditions.

4.2.2 Modeling Analysis on Crash Data under Pandemic

Since the COVID-19 pandemic has significant influence over transportation field during 2020. One year of CRIS crash data in year 2020 is sampled to study the difference of attributes influencing performance of selected safety attributes. Due to the decrease of traffic flow during pandemic period, the crash number in 2020 has been remarkably changed. Detailed crash number and population with normalized trend in recent ten years in Texas are shown in Table 11 through CRIS crash data. The normalized general crash trend is increasing over the first nine years. However, the total crash in 2020 has decreased by 15.9% compared to year 2019 and is even lower than the crash number in 2014. In considering of the population growth, crash number per 1,000 population in 2020 is significantly dropped and close to the crash number per 1,000 population in 2011.

Table 11 Normalized Crash Trend in Texas from 2011 to 2020

Year	**2011**	2012	2013	**2014**	2015
Crash Number	456,150	495,893	521,475	**555,298**	601,175
Population	25,645,504	26,084,120	26,479,646	26,963,092	27,468,531

Crash Number over Every 1,000 Population	**17.8**	19.0	19.7	20.6	21.9
Year	2016	2017	2018	2019	**2020**
Crash Number	632,288	620,860	629,116	649,024	**545,736**
Population	27,914,064	28,291,024	28,624,564	28,986,794	**29,360,759**
Crash Number over Every 1,000 Population	22.7	21.9	22.0	22.4	**18.6**

Model Performance. Since the size of one year crash data is different compared to seven years of linked data used in the last modeling analysis part. Performances of candidate models are required to be evaluated again on crash data set to select the suitable Machine Learning model. The dataset is categorized under each safety device similarly in the previous section. The results of model performance evaluation based on different output attributes are shown in Table 12 and Table 13.

Table 12 Accuracy Scores for Candidate Models on Crash Severity in Crash Data Analysis

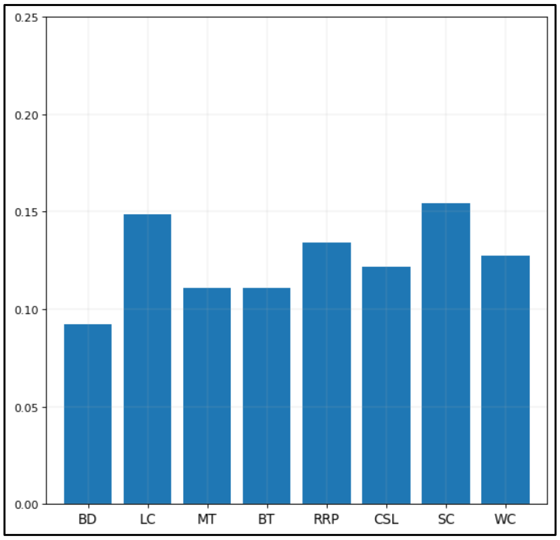
Area	Decision Tree	XGBoost	Random Forest	MLP Classifier	Gaussian NB	K-Nearest Neighbors
Rural	0.575	0.631	0.600	0.630	0.563	0.614
Metropolitan	0.595	0.625	0.606	0.625	0.589	0.578

Table 13 Accuracy Scores for Candidate Models on Property Damage in Crash Data Analysis

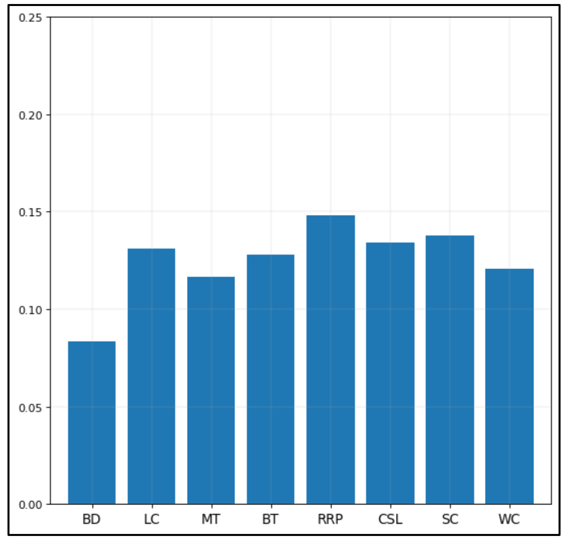
Area	Decision Tree	XGBoost	Random Forest	MLP Classifier	Gaussian NB	K-Nearest Neighbors
Rural	0.967	0.981	0.979	0.979	0.953	0.979
Metropolitan	0.972	0.978	0.977	0.978	0.956	0.978

As the results shown in Table 12 and Table 13, the highest accuracy scores are shown in red colors for both output attributes. The XGBoost also has the highest accuracy scores on crash severity and property damage output in this dataset. Unlike the linked data, the performance of candidate Machine Learning Models are not necessarily better in metropolitan scenario. The performance of MLP Classifier for sample crash data is close to XGBoost. Thus, XGBoost is selected as the Machine Learning model in this section.

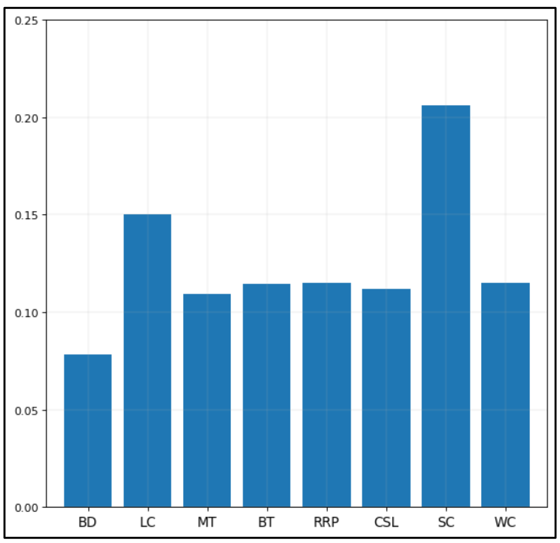
Feature Selection. The target data has fewer attributes and smaller data size compared to linked data with additional roadway attributes. This part of data analysis focuses on parameters in crash database, the outputs used for assessment of devices performance are still crash severity and whether over $1,000 property damage. The XGBoost modeling results on the crash severity output are shown in Figure 11.



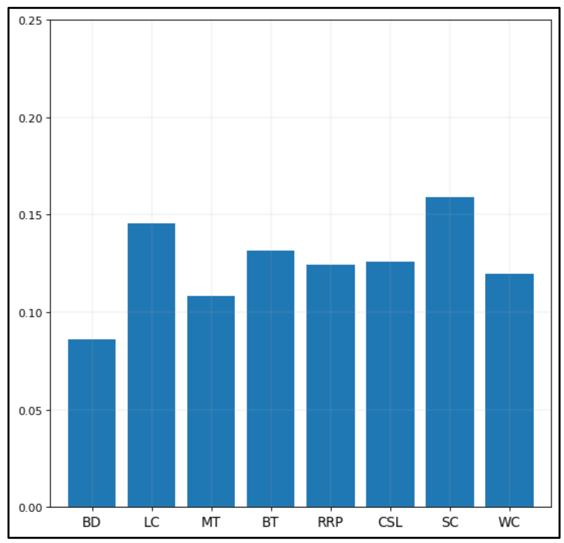
(a) Guardrail–Metro



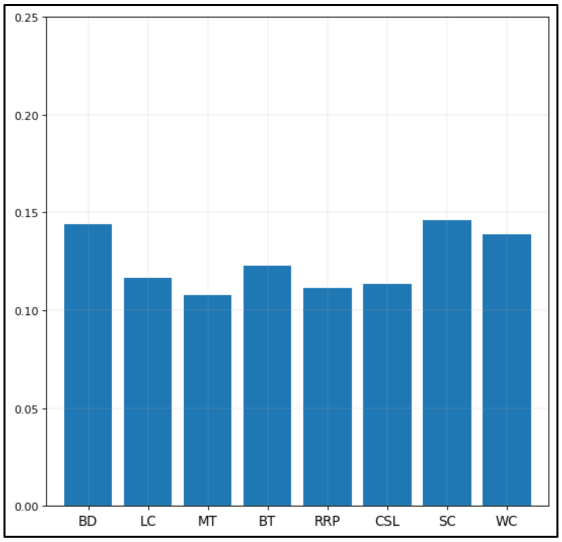
(b) Guardrail–Rural



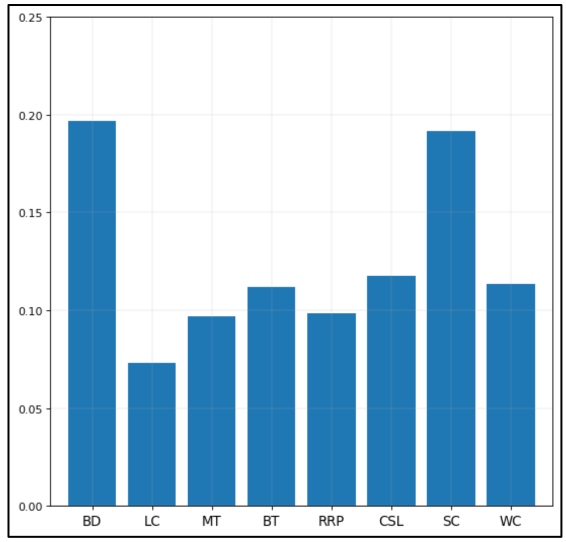
(c) Median barriers–Metro



(d) Median barriers–Rural



(e) Bridge Rail–Metro



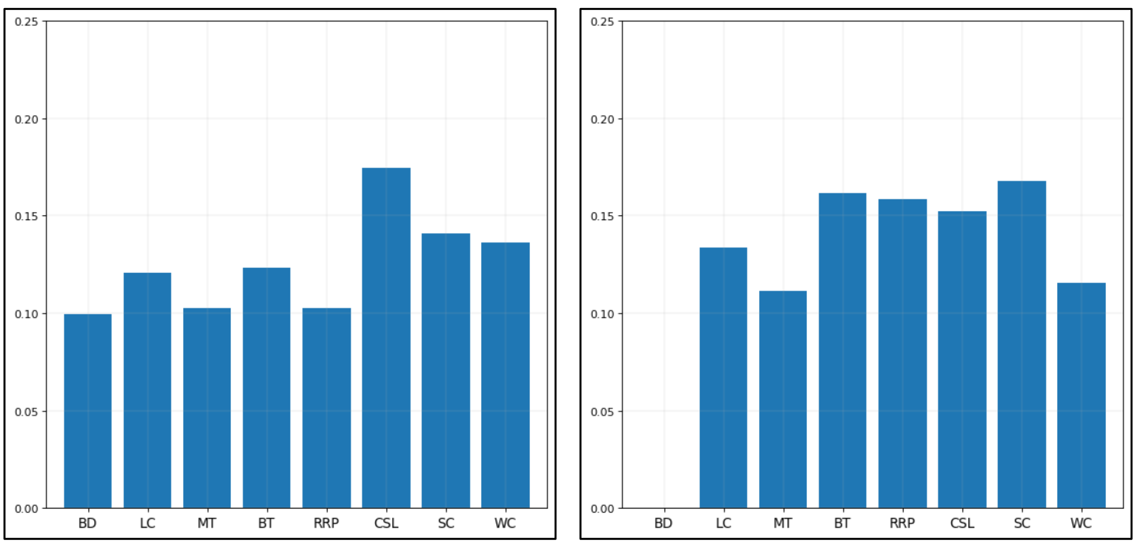
(f) Bridge Rail–Rural

Figure 11 Crash data attributes importance ranking on crash severity by XGBoost model for each safety device

From modeling results on crash severity when data source is one year of Texas crash data under the pandemic period, the dissimilarity of the importance scores for related attributes between different safety devices and areas is reduced. Generally, the highest impact attribute on crash severity regardless of data groups is the road surface condition. The light condition is also having relatively higher impact over crash severity in this dataset. Due to the decrease in total crash number, the impact of crash attributes in metropolitan crashes tend to be more average and the important crash parameters in metropolitan crashes tend to be included in the crash conditions.

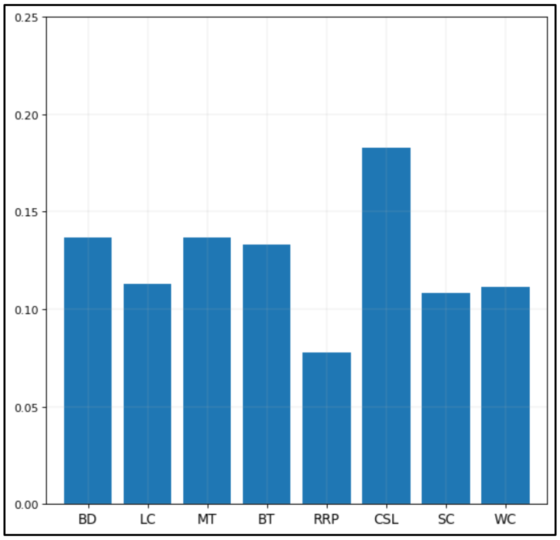
Compared to metropolitan areas, rural crash attributes have similar impact on crash severity. Apart from surface condition, bridge detail attribute that indicates the lane type of crash location is also vital especially for rural bridge rail crashes under pandemic. The attributes that are related to roadway maintenance and basic infrastructure have higher influence in rural areas under the pandemic.

The XGBoost model is used to train same groups of data again for presenting performance through property damage. The results are shown in Figure 12.

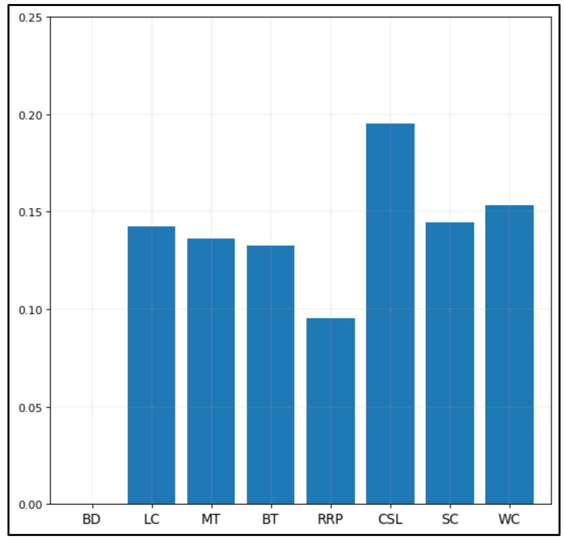


(a) Guardrail–Metro

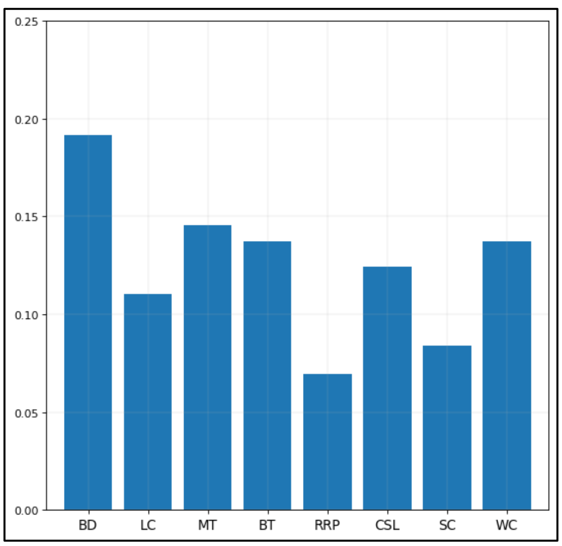
(b) Guardrail–Rural



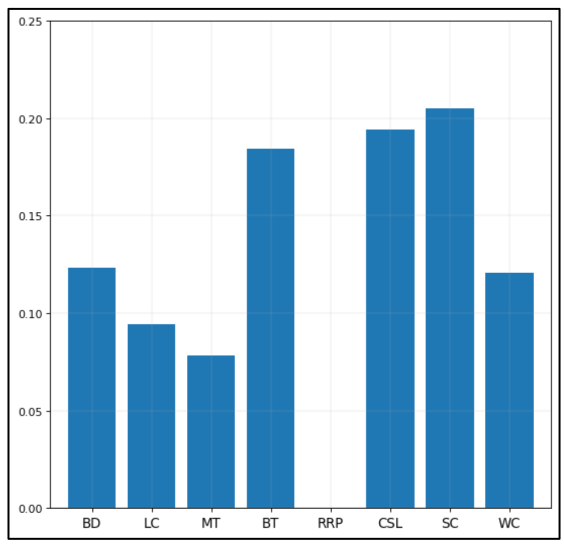
(c) Median barriers–Metro



(d) Median barriers–Rural



(e) Bridge Rail–Metro



(f) Bridge Rail–Rural

Figure 12 Crash data attributes importance ranking on property damage by XGBoost model for each safety device

According to the modeling results for property damage output, the speed limit is presented to be the most important parameter for each safety device in rural and metropolitan areas, except for metropolitan crashes related to bridge rail. The bridge detail still has the influence over property damage for bridge rail in metropolitan crashes, while the importance rank of this attribute is higher even related to median barrier crashes. The importance ranking of attributes in rural crashes are different compared with metropolitan crashes with increased impact from attributes related to roadways.

4.3 Discussion

By conducting a thorough analysis on crash and roadway data sets, part of Level I ISPE process were implemented to address the performance of selected safety devices through crash results outputs. This study went through the statistical analysis which ensure the significance difference between individual input and output attributes in the target databases. Follow-up analysis only makes sense when the results of ANOVA test suggest the dissimilarities within studied data.

Through comparing the statistical counting and correlation results with Machine Learning based modeling results, it can be concluded that, performance of roadside safety devices differs in rural and metropolitan areas. While distribution of serious crashes tends to be higher in rural counties, the impact factors that influence the performance of safety devices are focused on roadway information and transportation infrastructure in rural areas, instead of crash conditions as in metropolitan areas. However, the roadway surface condition has been a fairly important parameter in affecting safety devices. Studies have been conducted to draw the conclusion that, when further feature selection on more detailed scenarios is conducted, it can be demonstrated that the improvement and maintenance of roadway condition in rural areas are insufficient compared to metropolitan areas, which tends to leave bad surface conditions for the roadways. In addition, some transportation infrastructure including medians in rural roadways are much less installed and developed (Wang, *et al.*, 2021). In this way, the transportation equity can be addressed to highlight the importance of balancing between different communities.

The result of this study emphasizes the focusing point of impacting parameters when performing ISPE are varied. It is based on located communities and additional attributes from other data sources apart from crash data itself. The corporation of other related infrastructure also plays a necessary role for roadside safety devices to perform as they designed to. In addition, to back up reviewed needs under pandemic period, the individual crash analysis result suggests the roadway condition and related infrastructure statue in rural counties need to be paid attention to by local authorities. In a word, transportation equity is still vital under pandemic period.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

The major objectives of this thesis include the comparison between two principal communities in transportation safety aspect and addressing the needs for transportation equity by conducting In-Service Performance Evaluation on safety devices. Candidate Machine Learning models with their own algorithms and theories are introduced in this study. By performing assessment on potential suitable models, one Gradient Boosting model was applied to train target datasets for more precise analysis results. Multiple analysis approaches were conducted in this study along with proper comparisons. Additional analysis under specific year is also performed to prove the needs for transportation equity and Environmental Justice practices under pandemic situations.

The findings of this thesis study can be summarized into the following points.

(1) As an important evaluation to transportation safety, crash analysis is essential in transportation studies. The distribution of crashes results in fatal or incapacity injuries and high value of property damage is higher in rural communities, even though the total crash number there is much lower compared to metropolitan counties.

(2) Apart from the maintenance and development of safety devices themselves, the needs for improving surrounding roadway conditions and related transportation infrastructure cannot be ignored. It is a vital consideration to ensure the roadside safety devices are properly in-service before the determination of transportation decisions.

(3) It is rather essential to follow the requirement of transportation equity and EJ practices in related polices. The demand in fulfilling equity in transportation studies should be satisfied all time including the pandemic period.

Illustrated from the conclusions of this study, several recommendations are presented for improving the transportation equity:

(1) Projects and studies focused on maintaining and improving transportation infrastructure and roadway conditions in rural or low population communities are recommended to conduct. Related transportation agencies and local authorities are encouraged to spend more resources and make high-level polices or acts on the development of basic public facilities and infrastructures equally in all regions.

(2) ISPE practices are recommended to be incorporated into proper analysis on transportation infrastructure. Additional data sources and attributes are suggested to be included in crash related analysis. High-end analysis method including Machine Learning modeling or other computer technologies are recommended to be considered in further data analysis.

(3) The installation and maintenance conditions of safety devices themselves should also be equally evaluated in both metropolitan and rural areas. Inventory of major types of safety devices such as median barriers and guardrails are suggested to be updated with regular time period

REFERENCES

American Association of State Highway and Transportation Officials (AASHTO). (2009). *Manual for assessing safety hardware (MASH)*. Washington, DC.

AASHTO. (2010). *Highway safety manual*. Washington, DC, USA.

AASHTO. (2011). *Roadside design guide*. Transportation Officials. Task Force for Roadside Safety.

AASHTO. (2016). *Manual for assessing safety hardware (MASH)*. Washington, DC.

Abdoli, N., & Hosseinzadeh, A. (2021). Assessing spatial equity of public transit demand amid COVID-1. *International Conference on Transportation and Development 2021*, (pp. 513-520).

Ahmed, N. K., Atiya, A. F., Gayar, N. E., & El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric Review*, *29*(5-6), 594-621.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, *46*(3), 175-185.

Association for Safe International Road Travel (ASIRT). (2020). *Annual global road crash statistics*. Retrieved May 2021, from https://www.asirt.org/safe-travel/road-safetyfacts/#:~:text=More%20than%2038%2C000%20people%20die,for%20people%20aged%201%2D54

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24-31.

Bills, T. S., Sall, E. A., & Walker, J. L. (2012). Activity-based travel models and transportation equity analysis: Research directions and exploration of model performance. *Transportation research record*, *2320*(1), 18-27.

Brownlee, J. (2016, April 11). *Naive Bayes for Machine Learning*. Retrieved June 2021, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503-514.

Cantisani, G., Di Mascio, P., & Polidori, C. (2017). New research opportunities for roadside safety barriers improvement. *IOP Conference Series: Materials Science and Engineering*. *236*, p. 012097. IOP.

Car, Z., Baressi Šegota, S., Anđelić, N., Lorencin, I., & Mrzljak, V. (2020). Modeling the spread of COVID-19 infection using a multilayer perceptron. *Computational and mathematical methods in medicine 2020*.

Castiglione, J., Hiatt, R., Chang, T., & Charlton, B. (2006). Application of travel demand microsimulation model for equity analysis. *Transportation Research Record*, *1977*(1), 35-42.

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

Civil Rights Act. (1964). *Title VI of The Civil Rights Act of 1964*. The 88th United States Congress.

Cole, H. V., Anguelovski, I., Baró, F., García-Lamarca, M., Kotsila, P., Pérez del Pulgar, C., . . . Triguero-Mas, M. (2020). The COVID-19 pandemic: power and privilege, gentrification, and urban environmental justice in the global north. *Cities & Health*, 1-5.

Cooner, S. A., Rathod, Y. K., Alberson, D. C., Bligh, R. P., Ranft, S. E., & Sun, D. (2009). *Performance evaluation of cable median barrier systems in Texas*. Texas Transportation Institute. No. FHWA/TX-09/0-5609-1

Davis, C., & Stacy, C. P. (2021, January 4). *Four Lessons for Cities to Help Advance equitable transportation during the COVID-19 pandemic and beyond*. (Urban Institute) Retrieved June 2021, from https://www.urban.org/urban-wire/four-lessons-cities-help-advance-equitable-transportation-during-covid-19-pandemic-and-beyond

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, 143-148.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B., & Stoddart, G. L. (2005). Economic evaluation using decision analytic modelling. *Methods for the economic evaluation of health care programmes*, 277.

Du, J., Qiao, F., Wang, H., Zhang, Y., & Yu, L. (2021). Frequent pattern analysis of the roadside safety devices related. *International Journal of Engineering Science Invention*, *10*(5), 35-46. DOI: 10.35629/6734

Du, J., Wang, H., & Qiao, F. (2020). Transportation-Related toxic emissions influenced by public reactions to the COVID-19 pandemic. *Journal of Environmental and Toxicological Studies*, *4*(1). DOI: 10.16966/2576-6430.128

Eliasson, J., & Mattsson, L. G. (2006). Equity effects of congestion pricing: quantitative methodology and a case study for Stockholm. *Transportation Research Part A: Policy and Practice*, *40*(7), 602-620.

Environmental Protection Agency (EPA). (1993). *National Environmental Justice Advisory Council*. Retrieved June 2021, from https://www.epa.gov/environmentaljustice/national-environmental-justice-advisory-council

Federal Highway Administration (FHWA). (2012). *FHWA Order 6640.23A*. Retrieved from https://www.fhwa.dot.gov/legsregs/directives/orders/664023a.cfm

FHWA. (2015). *Environmental justice reference guide*.

FHWA. (2016). *Highway performance monitoring system field manual*. Office of Highway Policy Information. No. 2125-0028

FHWA. (2017, June 28). *Planning for transportation in rural areas*. Retrieved May 2021, from https://www.fhwa.dot.gov/planning/publications/rural_areas_planning/page03.cfm

FHWA. (2019). *Functional system lane length - 2019*. Retrieved May 2021, from https://www.fhwa.dot.gov/policyinformation/statistics/2019/pdf/hm60.pdf

FHWA. (2020, August 11). *Roadside safety hardware identification methods*. Retrieved May 2021, from https://safety.fhwa.dot.gov/roadway_dept/countermeasures/reduce_crash_severity/id_methods/ch1.cfm

Gutowski, M., Palta, E., & Fang, H. (2017). Crash analysis and evaluation of vehicular impacts on W-beam guardrails placed behind curbs using finite element simulations. *Advances in Engineering Software*, 114, 85-97.

Harmon, T., Bahar, G. B., & Gross, F. B. (2018). *Crash costs for highway safety analysis*. Federal Highway Administration. Office of Safety. No. FHWA-SA-17-071

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

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Health Resources and Services Administration (HRSA). (2021, January). *Defining rural population*. Retrieved May 2021, from https://www.hrsa.gov/rural-health/about-us/definition/index.html

Karner, A., & Niemeier, D. (2013). Civil rights guidance and equity analysis methods for regional transportation plans: a critical review of literature and practice. *Journal of Transport Geography*, *33*, 126-134.

Karner, A., London, J., Rowangould, D., & Manaugh, K. (2020). From transportation equity to transportation justice: within, through, and beyond the state. *Journal of Planning Literature*, *35*(4), 440-459.

Karner, A., Rowangould, D., & London, J. (2016). We can get there from here: new perspectives on transportation equity. *UC Davis: National Center for Sustainable Transportation*.

Kim, T. K. (2017). Understanding one-way ANOVA using conceptual figures. *Korean journal of anesthesiology*, *70*(1), 22.

Madeh Piryonesi, S., & El-Diraby, T. E. (2021). Using Machine Learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. *Journal of Infrastructure Systems*, *27*(2), 04021005.

Mak, K. K., & Sicking, D. (2002). *Continuous evaluation of in-service highway safety feature performance*.

McLachlan, G. J., Do, K. A., & Ambroise, C. (2005). *Analyzing microarray gene expression data*. Wiley.

National Highway Traffic Safety Administration (NHTSA). (2018). *Fatality analysis reporting system encyclopedia*. Retrieved May 2021, from https://www-fars.nhtsa.dot.gov/Main/index.aspx

Nicholson, C. (2020). *A beginner's guide to Multilayer Perceptrons (MLP)*. Retrieved June 2021, from https://wiki.pathmind.com/multilayer-perceptron

Pal, S. K., & Mitra, S. (1992). Ultilayer perceptron, fuzzy sets, classifiaction.

Qiao, F., Wang, H., Du, J., Cheng, V., & Odell, W. (2020). *Development and enhancement of in service performance evaluation (ISPE) process for roadside safety devices*. TxDOT Project 0-7018: Tech Memo 3.

Randall, J. (2005). *Texas Reference Marker (TRM) system user's manual. TxDOT*.

Rezapour, M., Molan, A. M., & Ksaibati, K. (2020). Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models. *International journal of transportation science and technology*, *9*(2), 89-99.

Rodier, C., Abraham, J. E., Dix, B. N., & Hunt, J. D. (2010). *Equity analysis of land use and transport plans using an integrated spatial model*. Mineta Transportation Institute. No. CA-MTI-10--2601/2705

Ross Jr, H. E., Sicking, D. L., Zimmer, R. A., & Michie, J. D. (1993). *Recommended procedures for the safety performance evaluation of highway features*. Washington, DC. No.350

Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, *21*(3), 660-674.

Sayad, S. (2010). *K nearest neighbors*. University of Toronto.

Schalkwyk, I. V., Bligh, R. P., Alberson, D. C., Bullard Jr, D. L., Lord, D., & Miaou, S. P. (2004). *Developing an In-Service Performance Evaluation (ISPE) for Roadside Safety Features in Texas*. No. FHWA/TX-05/0-4366-1

State Office of Rural Health (SORH). (2012, April). *Texas county designations*. Retrieved June 2021, from https://www.google.com/url?q=https://www.texasagriculture.gov/portals/0/forms/er/rural-metro%2520counties.pdf&sa=U&ved=2ahUKEwiGzeWer5rxAhWCWM0KHWDCAZEQFjAAegQICRAB&usg=AOvVaw0pnwPcnGxeGaWtozCXwzjt

Texas Department of State Health Services (Texas DSHS). (2020, April 29). *Definitions of county designations*. Retrieved June 2021, from DS

Twaddell, H., & Zgoda, B. (2020). *Equity analysis in regional transportation planning processes, Volume 1*: Guide. No. Project H-54

Texas Department of Transportation (TxDOT). (2014). *Roadside safety field guide 2014*.

United States Department of Transportation (U.S. DOT). (2017, January 18). *Environmental justice strategy*. Retrieved June 2021, from https://www.transportation.gov/transportation-policy/environmental-justice/environmental-justice-strategy

U.S. DOT. (2021, February 11). *Rural opportunities to use transportation for economic success (ROUTES)*. Retrieved May 2021, from https://www.transportation.gov/rural

Wang, H., Qiao, F., Du, J., & Zhang, Y. (2021). In-Service performance comparison of roadside safety devices in rural and metro counties in texas. *International Journal of Engineering Science Invention*, *10*(6), 30-38. DOI: 10.35629/6734

Wang, S., Mo, B., & Zhao, J. (2020). Predicting travel mode choice with 86 machine learning classifiers: an empirical benchmark study. *Transportation Research Board 99th Annual Meeting*. Washington, D.C.

Wemple, E., Colling, T. K., & Systematics, C. (2014). *Improving safety on rural local and tribal roads safety toolkit* (No. FHWA-SA-14-072). United States. Federal Highway Administration. Office of Safety.

Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *Annals of Statistics*, *35*(6), 2450-2473.

Zantalis, F., Koulouras, G., Karabetsos, S., & Kandris, D. (2019). A review of machine learning and IoT in smart transportation. *Future Internet*, *11*(4), 94.

Zhang, L., Zheng, Z., Ren, Y., Zhou, X., Keramati, A., Tolliver, D., & Huang, Y. (2020). A gradient boosting crash prediction approach for highway-rail grade crossing crash analysis. *Journal of advanced transportation 2020*.

Zhang, Y., Cheng, L. C., Qiao, F., & Patel, A. (2019). Enhancement of In-Service Performance Evaluation (ISPE) process for roadside safety devices: A Survey. *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (pp. 328-332). IEEE.