5-2023

Investigating the Impact of Covid-19 on Mobility Condition.

Sandip Acharya

INVESTIGATING THE IMPACT OF COVID-19 ON MOBILITY CONDITION

THESIS

Presented in Partial Fulfillment of the Requirements for

the Master of Science Degree in the Graduate School

of Texas Southern University

By

Sandip Acharya, B.S.

Texas Southern University

2023

Approved By

Mehdi Azimi

Chairperson, Thesis Committee

Gregory H. Maddox

Dean, The Graduate School

Approved By:

Mehdi Azimi	03/28/2023
Ph.D., Chairperson of Thesis Committee	Date
Yi Qi	03/28/2023
Ph.D., Committee Member	Date
Fengxiang Qiao	03/28/2023
Ph.D., Committee Member	Date
Subasish Das	03/28/2023
Ph.D., Texas state University	Date
Ismet Sahin	03/28/2023
Ph.D., Graduate School Representative	Date

INVESTIGATING THE IMPACT OF COVID-19 ON MOBILITY CONDITION

By

Sandip Acharya, M.S.

Texas Southern University, 2023

Assistant Professor Mehdi Azimi, Advisor

Having large number of vehicles operating in the freeways of Houston daily, the mobility concern is high as some of the freeways in Houston are among the most congested freeways in United States. During the COVID-19 pandemic, the less congested freeways led to over speeding resulting in various crashes and even fatality. This resulted in changing of drivers; and ultimately the mobility patterns were changed during the study years of 2019, 2020 and 2021. To better understand how this mobility pattern changed over the three years, this research used Machine Learning algorithms to examine the mobility of freeways in Houston during that time. For this purpose, a model was developed using python coding which considered operating speed and other independent variables to understand the change of the traffic mobility. Several methods were used in the study to check the effectiveness of Artificial Intelligence modeling. To check how the mobility was impacted over the years, Violin Plots were also plotted to illustrate the change of operating speed from year 2019 to 2021.

The results of this research demonstrated that there are eight factors that have significant effects on the vehicular mobility. Among them, annual average daily traffic is the most influencing in traffic mobility study whereas K-factor is the least effective among the selected variables. Relative countermeasures were recommended according to the influencing factors that were identified.

Keywords: Traffic Mobility, Machine Learning, COVID-19, Speed

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABBREVIATIONS

ADT:	Average Daily Traffic
VMT:	Vehicles miles Travelled.
CRIS:	Crash Records Information System
GIS:	Geographic Information System
HSIS:	Highway Safety Information System
NPMRDS:	National Performance Management Research Data Set
KNN:	K-Nearest Neighbors
AADT:	Annual Average Daily Traffic
K-fac:	K-factor
DVMT:	Daily Vehicle Miles Travelled
TxDOT:	Texas Department of Transportation

VITA

2014-2018	B. E, Tribhuvan University, Nepal
2021-2023	Graduate Research Assistant and Student, Texas Southern University, Houston, Texas
Major Field	Transportation Planning and Management

ACKNOWLEDGEMENT

Firstly, I would like to thank my academic advisor, Dr. Mehdi Azimi, for the continuous support of my study and related research at TSU, and for patience, motivation, and immense knowledge. Without his, assistance, I would not have achieved the level in my academic and research studies that I am at today. I could not have imagined having a better advisor and mentor for my master's study. His guidance helped me not only in all the time of research and writing this thesis, but also in writing recommendation letters for my scholarship applications. His trust and continuous support made my days at TSU easier and happier. Dr. Azimi is a highly respected advisor with excellent professional experience and great accomplishments which I would like to attain in the future. Working side-by-side with him was a great experience, and I really learned a lot from him. I am grateful for his guidance, which helped me to achieve my goal of completing the master's study at TSU. Thank you for sharing your research and life experience with me, which would benefit my whole life.

I would like to express my sincere appreciation to my thesis committee members. I truly appreciate Dr. Yi Qi for her guidance and helpful suggestions in my entire thesis. I would like to thank Dr. Fengxiang Qiao for his support and help during my studies at TSU. And special thanks go to Dr. Subasish Das and Ismet Sahin for serving as my committee members and all their assistance in this thesis.

I would like to thank all my friends for their help in various ways. Thank you for your understanding and encouragement in so many moments of crisis. Your friendship makes my life a wonderful experience.

Last but not the least; I have to thank my family specially my dad, mom and relatives for their unconditional love and support throughout my life. Thank you for giving me strength to reach for the stars and chase my dreams. This journey would not have been possible without the support of my family.

CHAPTER 1

INTRODUCTION

1.1 Background of Research

There are many challenges in developing future commuting options and making traffic mobility effective. In the United States, most of the people prefer to use cars even when there are public transportation systems available. Texas is one of the most populous and fast-growing states in the United States, and as such, it has a high level of traffic congestion, particularly in its major metropolitan areas such as Houston, Dallas, and Austin. Due to high economic growth, employment opportunities, and available facilities, the concentration of the population is high in those metropolitan areas resulting in high traffic volume in urban areas.

Harris County is in the Gulf Coast region of Texas, the southeastern part of the state. It is the third-most populous county in Texas and the most populous county in the Houston metropolitan area. Harris county covers an area of 1,778 square miles and has a population of over 4.7 million people. The estimated daily vehicle miles traveled (VMT) in the Houston-Galveston region, which includes Houston, was approximately 113.4 million miles per day in 2019. (Urban Mobility Report, 2019).

Traffic mobility studies have been conducted by researchers for many years, but the formal study of traffic engineering began in the early 1900s, with John F. O'Connell being one of the pioneers in this field (Institute of Transportation Engineers, 2023). The advent of automobiles and growth in cities led to increased traffic congestion; and traffic engineering and transportation planning became crucial, with highway agencies investing in research and education in these fields. Traffic mobility studies have evolved over time and incorporated new technologies and techniques to analyze and improve traffic flow. As a result, they are now a crucial aspects of transportation planning for both urban and rural areas. The difficulty lies in improving mobility while simultaneously decreasing traffic congestion and accidents.

Houston is a major part of Harris County; and this study is focused on major freeways in Houston. The city is served by several major highways providing important links to other cities in Texas and neighboring states. The large volume of traffic movement occurs to and from the airports in Houston. The city has two major airports, George Bush Intercontinental Airport, and William P. Hobby Airport, offering domestic and international flights. The Metropolitan Transit Authority of Harris County (METRO) provides bus and light rail service in the Houston area, offering residents and visitors alternatives to driving. Houston also has a growing bike share program and is working to improve pedestrian walkways and bike lanes to enhance mobility options for residents and visitors.

Freeways in Houston

Harris County is home to a large and complex network of roads, including major freeways, highways, and local streets. Some of the major highways and freeways that pass through or are in the county include IH-610 Loop, IH-10, IH-45, IH-69, US-290 and SH-288. IH-610 forms a loop around the inner-city sector of the city of Houston. IH-10, IH-45, IH-69, and US-290 are major north-south and east-west highways that run through the county, connecting Houston to other parts of Texas and the region. Similarly, SH-288 runs from downtown Houston to the southern part of the county, providing a major route for traffic between Houston and the Gulf Coast. Beltway 8, the Sam Houston Parkway along with the Sam Houston Tollway is an 88-mile beltway around the city of Houston, lying entirely within the county and providing an alternate route for those willing to pay toll for the use of the road. In addition to these major highways and freeways, Harris County also has many local streets and roads that provide access to neighborhoods and other areas within the county.

Interstate Highway 10

Traffic congestion on I-10 in Houston is a common issue, particularly during rush hours and major events. This I-10 is among some of the most congested freeways in the United States in spite of having some of the widest sections of freeways up to 26 lanes (both sides). The traffic mobility condition on I-10, specifically during peak hours, is not good due to its congested road segments as vehicles experience long delays during those time of the day.

Interstate Highway 69

In Houston, IH-69 follows US 59 (Southwest Freeway) from Fort Bend County to the west loop of I-610. I-69 then follows US 59 (Eastex Freeway) from the north loop of I-610 to the Montgomery–Liberty County line. It provides important transportation links for the region, connecting various cities as well as linking Houston to Mexico. This is also important from an economic point of view as it is a connector to various ports resulting in high traffic volume per day.

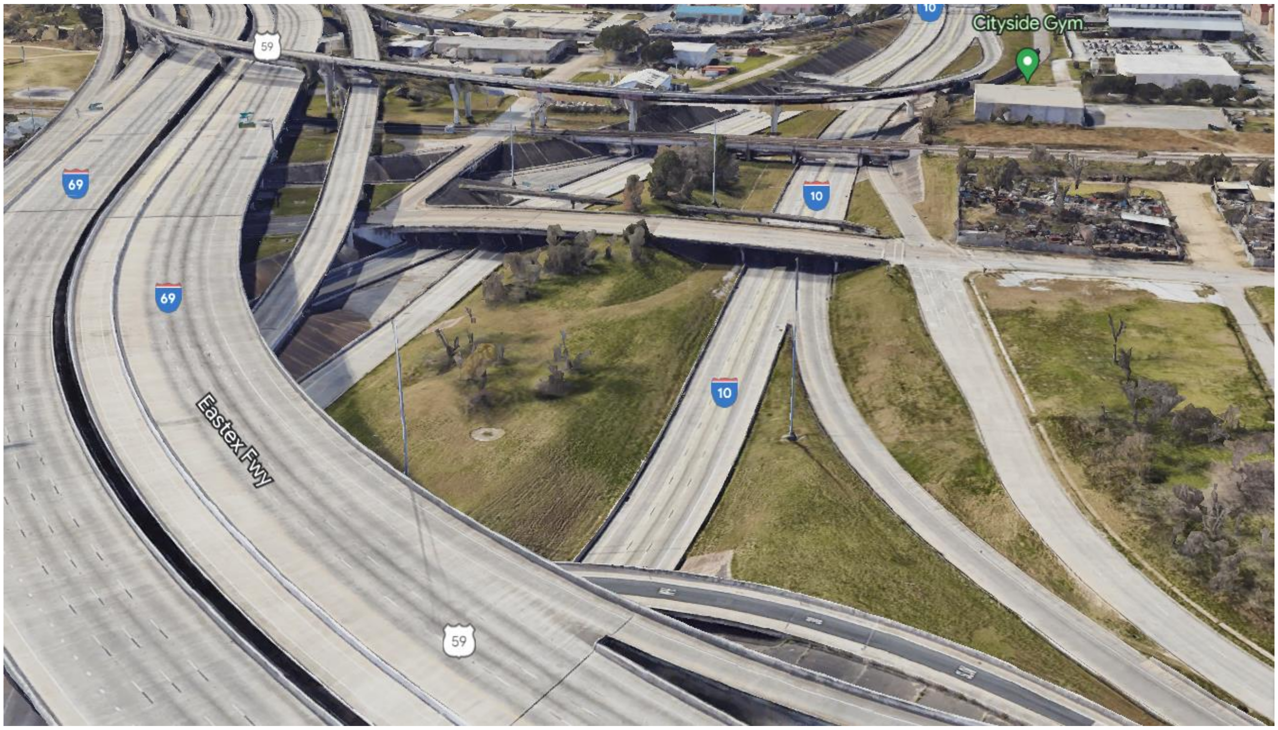


Figure 1. IH-10, IH-59/69 Near Houston Downtown (Source: Google maps)

Interstate Highway 610 Loop

It is also known as the Inner Loop, and it is the inner of two concentric freeways that encircle the city of Houston; and has a total length of approximately 38.6 miles. 610 loop serves as important route for internal mobility inside Houston and is heavily congested during rush hours.

Interstate Highway 45

IH-45 is a major highway in the state of Texas, running from Dallas to Houston. In Houston, IH-45 connects several major suburbs and the city's central business district, providing a key route for travel and commerce in the region. IH-45 connects the north and south parts which is important for trucking and connecting to other big cities whereas IH-10 is an important interstate connecting the east part and west parts of Houston with a large amount of traffic per day.

Figure 2. IH-45 and IH-610 Loop Intersections in Houston (Source: Wikipedia)

The number of registered vehicles in the county is increasing along with the county's population growth. In recent years, Harris County has been one of the fastest-growing counties in the United States, with a population that has been rapidly increasing due to migration, both domestic and international. As more people move to the area, there will be more demand for vehicles to travel around the county, which would lead to an increase in the number of registered vehicles. Additionally, economic growth in the area, such as an increase in jobs and businesses are also contributing factors to an increase in the number of registered vehicles.

COVID-19 and Harris County

Harris County was significantly affected by the COVID-19 pandemic. The County recorded high numbers of cases and deaths, and local officials have implemented various measures such as mask mandates and capacity restrictions to slow the spread of the virus. During the COVID-19 pandemic, Harris County went through changes in traffic patterns and mobility. With many businesses closed or operating at reduced capacity, and many people working from home, there was less traffic on the roads, in general. As a result, some drivers took advantage of the reduced traffic to travel more, which led to over speeding and more traffic incidents and crashes. Furthermore, a part of the residents who worked remotely, returned to work as the pandemic progressed, which increased the traffic.

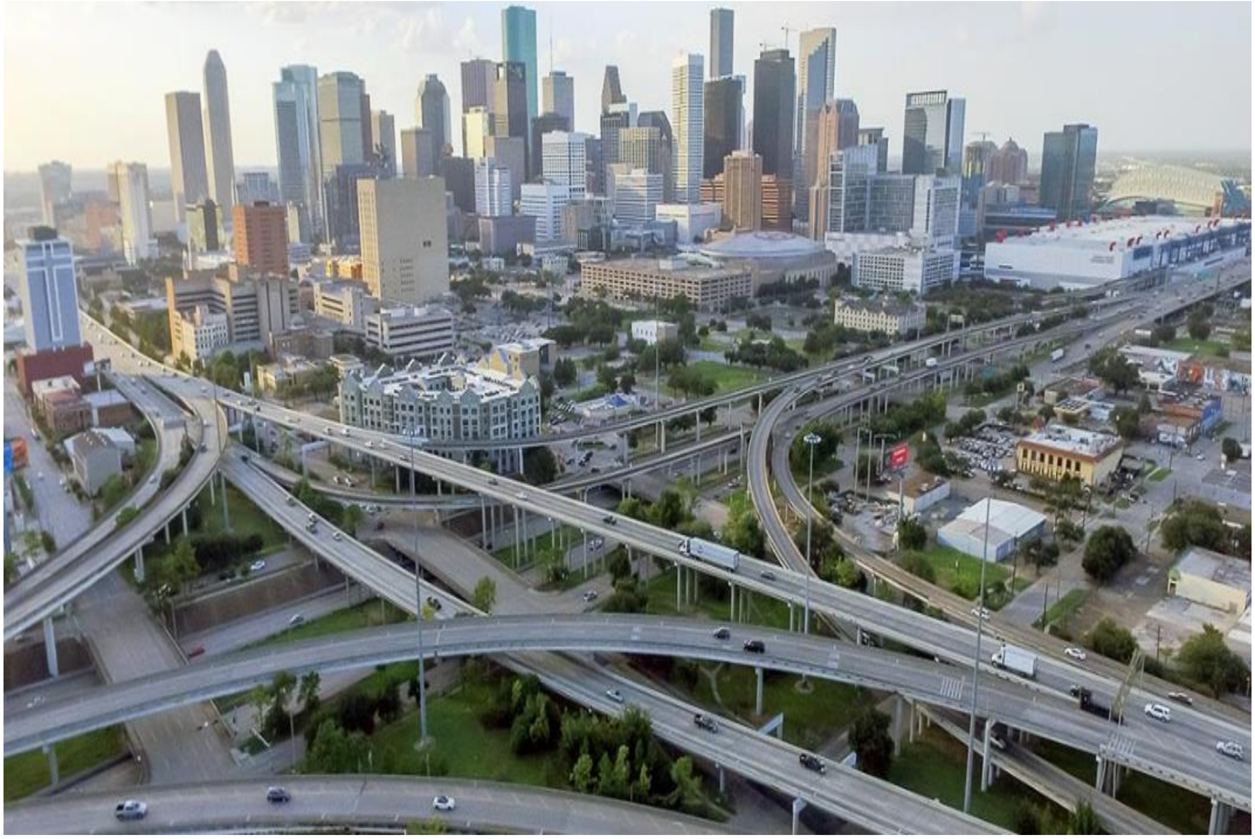


Figure 3. Traffic in Houston During COVID-19 Off-Peak Hour (Source: Rice University)

Traffic Mobility in Houston

The condition of traffic mobility in Houston varies and can be influenced by several factors such as time of day, weather, road closures, and accidents. During rush hours, traffic can become congested in certain areas, particularly in the downtown and IH 10, loop 610 and more. The city has made efforts to improve traffic flow, including the expansion of freeways and the implementation of high-occupancy vehicle lanes. Various strategies such as building new highways and roadways, expanding public transportation options, and using technologies such as traffic management systems and Intelligent Transportation Systems (ITS) are widely implemented to enhance the smooth flow of traffic in Houston. Furthermore, the strategies such as carpooling, toll lanes, and Texas Clear Lanes initiative have also been adopted which aim to reduce congestion and improve traffic mobility on key freeways in Houston. The driving speed in Houston for normal traffic is around the speed limit itself; and the operating speed changed drastically during the COVID-19 pandemic. In 2020, while the traffic volume decreased, the average speed increased compared to previous years. However, this did not result in a proportionate reduction in fatal and severe crashes. The frequency of crashes is more influenced by measurements of speed, and the impact of speed on crashes varies across different severity levels (Das et al., 2023).

Artificial Intelligence (AI) in Traffic Mobility Studies

Artificial Intelligence (AI) can play a significant role in improving traffic mobility by providing intelligent and automated solutions for traffic management, transportation planning, and traffic prediction. AI integrates innovative mobility solutions and uses cutting-edge technologies and methodologies to understand and improve the efficiency, safety, and sustainability of transportation. AI has the potential to make transportation systems more efficient, safer, and sustainable by providing real-time insights and automated decision making, helping traffic mobility to be more fluid and reducing congestion and travel time.

AI is being increasingly used in traffic mobility studies to enhance our understanding of traffic patterns. AI can be used to analyze large amounts of data, including traffic flow, travel time, road conditions, and travel behavior, to gain insights into transportation patterns. AI algorithms can be used to predict traffic congestion and reroute vehicles to reduce delays, or to optimize traffic signal timings to improve traffic flow. AI can also be used to simulate traffic scenarios and evaluate the impact of various interventions, such as road construction or the implementation of new transportation technologies. AI algorithms can be trained to analyze historical traffic data and predict traffic patterns in real-time, providing accurate information on traffic conditions and helping to reduce congestion. AI-powered ITS can provide real-time traffic management and control by analyzing traffic data and adjusting traffic signals, rerouting vehicles, and providing real-time information to drivers and public transportation operators. AI-based algorithms can be used to optimize traffic flow by analyzing sensor data and making real-time adjustments to traffic signals, tolls, and speed limits in order to reduce congestion and improve travel times.

This thesis is the outcome of an in-depth research of pre-COVID, COVID, and post-COVID phase impacts on traffic mobility inside Harris County. Pre-COVID is the phase where there was no positive COVID cases in the county. COVID phase is the time where the COVID started and reached to the peak. Post-COVID refers to the time when the COVID cases started dropping after the vaccination phase. Post-COVID-19 mobility refers to the changes and adaptations in the traffic in response to the COVID-19 pandemic. There is a research gap in the study of traffic mobility considering the three stages of COVID-19. This study is leaned to investigate the impact of different phases of COVID-19 on mobility. The change of the travel pattern on mobility before COVID-19, the time during COVID-19, and after the COVID-19 phase turned into a new normal phase are to be analyzed for Harris County. AI and Machine Learning techniques are used during the development the model, training of the model, and validation of the model.

The study covers a total of 179.1 miles of freeway on each side of the IH-45, IH-69, IH-10, and IH-610 loop inside Harris County. The traffic mobility from the year 2019 to 2021 is to be studied during the analysis; and the impacts caused by COVID-19 is to be compared with the traffic mobility in the previous year.

1.2 Research Objectives

The goal of this traffic mobility study is to investigate traffic mobility on freeways during pre-pandemic, pandemic, and post-pandemic. To identify issues, and provide recommendations, this study involves gathering and analyzing data related to traffic volume, speed, travel time, and transportation features which can focus on a larger area, such as the entire transportation network of a city like Houston. Some specific objectives of a traffic mobility study include:

- To investigate how mobility patterns have changed due to the pandemic, by analyzing changes in driving speed of 2019, 2020 and 2021,
- Understand the impact that the pandemic has had on traffic mobility on freeways of Houston,
- Develop a model to investigate the relationship between pre-pandemic (2019), pandemic, (2020), and post-pandemic (2021), traffic mobility,
- Demonstrate how the proposed method can provide new insights about changes in traffic mobility due to the pandemic,
- Illustrate how driving speed is influenced during a pandemic and how it has changed people's travel behavior,

1.3 Outline of the Study

In documenting the achievement of these objectives, this thesis is comprised of five chapters. The first chapter provides a background to identify the problem, defines research goal and objectives, and describes the layout of the study. The second chapter presents a literature review that provides an overview of various reports, articles, journals, and manuscripts related to the topic. This step helps to identify gaps in the current knowledge and to determine what is already known about the topic. Various databases and search tools such as TRB-TRID, ScienceDirect, and ASCE online library were used during the literature review process. The third chapter describes the methodology and how the research and overall analysis was conducted. This section also provides the information on the data processing, the variables that were used for model development, and the model validation. The data collection methods will depend on the research design and methodology chosen. The data is collected through various private companies, the Texas Department of Transportation, Various federal agencies, and other open sources. Once the data is collected, it will be analyzed to identify mobility patterns and develop a model to analyze the mobility trends. Various statistical techniques will be used to validate the developed model. The fourth chapter covers the results and discussion, which presents the findings of the study in details including the interpretation that can be made. The last chapter includes the conclusion and recommendation, where a summary of the results and the main findings of the and also the recommendations for future research are presented.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a brief summary of reviewed literatures on traffic mobility as well as their obtained results. Then, the studies that applied Machine Learning algorithms for traffic mobility analysis will be introduced along with the Machine Learning models which are used in traffic mobility studies. Some literatures are also included to better understand the impact of COVID-19 on mobility along with traffic safety. At the end, a summary of the reviewed literature will be presented.

2.1 Studies on Traffic Mobility

The Houston-Galveston Area Council (H-GAC), the region's Metropolitan Planning Organization (MPO), conducted a study to define mobility needs for the historic East End community near Downtown Houston (East End Mobility Study, 2019). It aimed to improve mobility and access in the study area by leveraging existing transportation networks and developing new infrastructure to support current and future population, development, and land uses. The study presented mobility improvement opportunities to support and stimulate future development in the area, which included neighborhoods and management districts. To address existing gaps and opportunities, potential improvements were identified through various means such as in-depth observations of operations, land use and travel demand modeling, stakeholder recommendations, and review of existing project plans. The potential improvements focused on improving operations for vehicles, addressing potential network bottlenecks, supporting improved mobility for walking, bicycling, and transit use, and proactively addressing issues that might arise from future development.

The 2021 Urban Mobility Report published by the Texas A&M Transportation Institute (TTI) analyzes urban transportation trends in the United States and provides insights into travel behavior, congestion, and other urban mobility challenges (Urban Mobility Report, 2021). The trends from 1982 to 2020 show that congestion was persistently growing until 2020, when it varied greatly across cities, roads, and hours. The 2020 congestion costs and travel delays were about half of the 2019. Considering various parameters such as delay per auto commuter, annual person hour of delay, annual extra travel time for commuter, excess fuel consumption, annual congestion cost, excess truck travel time, travel time index and commuter stress index, the report ranks Houston in top five of the United States most congested cities. It shows the urge for a detailed study of mobility in Houston, which this thesis is to fill the research gap.

Urban mobility has been impacted due to COVID-19 pandemic; and it has been a topic of research in recent years. You et al. (2022) conducted a review of existing literature on impact of the pandemic on urban mobility. Their review included both quantitative and qualitative evidence from scientific papers, reports from public and private organizations, and news articles. The review was supplemented by interviews with both experts in the field of urban mobility and representatives from public and private transportation agencies. The interviews were to explore the impact of the pandemic on urban mobility, as well as to identify the role of transportation policies in mitigating the disruption caused by the pandemic. The study highlighted the needs for cities to be more resilient and to plan for potential disruptions due to future pandemics.

Mobility changes due to COVID-19 were observed in various parts of the world. A study conducted in Los Angeles found that there was a significant decrease in traffic mobility across the city, particularly in retail, recreation, and transit, while residential activity remained relatively consistent (Lu and Giuliano, 2023). The research used descriptive analysis to determine the change in the mobility across different neighborhoods, age groups, and types of mobility. Furthermore, it was found that the mobility decreased more for younger people compared to older people. Difference-In-Difference (DID) regression was conducted to estimate the impact of dependent and independent variables. The study findings revealed that individuals living in low-income and ethnic minority neighborhoods reduced their work and shopping travel to a lesser extent compared to residents of high-income and white neighborhoods during the shelter-in-place order. This pattern is in line with the possibility that high-income individuals have more work-from-home opportunities, while low-income/minority groups may face more challenges related to household and work-related responsibilities.

Sana et al. (2022) used data from multiple sources and an interactive web-based map to determine the effects of pandemic on traffic patterns in San Francisco. It was found that the pandemic had a significant impact on traffic patterns, with some streets becoming even more congested than before. However, most areas showed resilience; and mobility decline was attenuated despite an upsurge in COVID-19 cases. The research identified areas of improvement in terms of traffic management and suggested that the data and map can be used for long-term monitoring and analysis of traffic congestion. The study provided a practical reference for establishing future epidemic countermeasures by corroborating that the pandemic effect was ineffectual and transitory.

In similar context, study conducted by Owen et al. (2020), provides an analysis of the impact of the COVID-19 lockdown on traffic flow in the London, noting a significant decrease in flow following the lockdown, as well as a decrease in flow in the week preceding it due to non-essential travel advice. The results are presented in the form of a change in average daily flows per 10,000 vehicles, with the percentage change varying depending on the characteristics of the individual sites in each speed limit. There is an increase in speeds leading up to the introduction of the lockdown. The comparison of changes in average and 85th percentile speeds in different speed limits, indicates a trend towards higher speeds by a larger percentage of drivers.

A study was conducted to find the impact of the COVID-19 pandemic on public and private transportation in Madrid (Spain), as well as the emergence of cycling as an alternative form of transportation (Sanz et al., 2021). To understand how various regions were impacted during the COVID-19 crisis, the mobility study was supplemented by an examination of socioeconomic factors, land usage, and transportation networks. Both qualitative and quantitative methods were used for this study. The qualitative method included interviews with transportation and air quality experts, as well as surveys of residents. The quantitative method included analysis of regional mobility data and air quality data. The study showed that both overall public transportation usage and private vehicle usage decreased significantly during the pandemic. The region also saw an increase in the use of cycling as an alternative form of transportation. Potential implications of the new mobility patterns for the region in the future were suggested, as well.

Another study showed that pandemic had both long-term and short-term impacts on traffic mobility (Kellermann et al., 2022). The research, conducted in Berlin of Germany, found that the pandemic had a significant impact on urban mobility behavior during a 20-month pandemic period and comparing them to the pre-pandemic situation. A combination of qualitative and descriptive analysis is used, and mode shift was observed from using public transportation to walking more and using private transportation mode. The study also found that the pandemic has had a significant impact on the distribution of mobility. The effects of the pandemic on mobility are likely to be long-term, and there are important implications for urban planning. The results are found measuring how people's mobility patterns have changed due to the pandemic, and identifying any lasting impacts that may create a completely different urban mobility environment referred to as a new normal.

2.2 Traffic Mobility and Safety

Higher operating speed of vehicles may lead to crashes. In this context, Das et al. examined the relationship between operating speeds on freeways and traffic crashes during the COVID-19 pandemic. This study used average speed, standard deviated speed and various levels of crashes to indicate the severity of fatality. R programing was used to check the relationship between operating speed and crashes. The results showed that operating speeds on freeways in 2020 were significantly higher than the previous two years, but the differences varied by the posted speed limits (PSL). The study finds that higher operating speeds are associated with more crashes, and daily-level operating speed measures were positively associated with crash occurrences. The study also showed that drivers during the pandemic engaged in more risky behaviors including speeding and driving while impaired.

Das et al. conducted similar study on crash modeling to understand the impact of freeway operating speeds on crashes during COVID. The study showed that the increase in speed impacted the crash types. During the low traffic duration of 2020, higher operating speed had more critical impacts on crash frequencies and severity of crashes compared to the years 2018 and 2019.

2.3 Machine Learning for Mobility Analysis

Deng et al. developed predictive models for traffic mobility using Machine Learning approaches, focusing on traffic speed and volume. The collected data was processed and divided into training and validation datasets for constructing, validating, and comparing model fitting and prediction performance. Different models were applied in the study, including autoregressive integrated moving average model (ARIMA), MLP, CNN, and LSTM. ARIMA was used as representative of statistical models whereas CNN and LSTM were selected as the representatives of Deep Learning models to train the dataset. The models were evaluated based on indicators such as R2, RMSE, ACF, PACF, confidence interval, and Shapiro-Wilk statistic. Results showed a similar and satisfactory performance for predicting traffic volume, while predicting traffic speed was more challenging due to the data characteristics. The researchers summarized the results of the model comparison and evaluation, highlighting the advantages and limitations of the models for future applications.

Prokhorenkova et al. (2017) presented key algorithmic techniques used in CatBoost, the gradient boosting toolkit that outperformed other publicly available implementations on a variety of datasets. CatBoost introduces two algorithmic advances, ordered boosting and an innovative algorithm for processing categorical features, to combat a prediction shift caused by a target leakage that are present in all existing implementations of gradient boosting algorithms. The study provides a detailed analysis of the problem and shows that the proposed algorithms solve it effectively, leading to excellent empirical results.

2.4 Summary

Mostly researchers focused on studying trend of mobility over the years. The gap in research was found as most of the studies was focused on analyzing the mobility trend only on either of before, during and after COVID-19. The change in mobility patterns is impacted by COVID-19 which is not considered in depth by any of the previous studies. The studies have not considered average and standard speed and also probable related variables such as right of way, annual average daily traffic, vehicles miles travelled, crash data, K-factor and so on during mobility analysis. In addition to that the previous studies also don't consider the implementation of Machine Learning on traffic mobility studies on urban freeways. There haven't been any studies on this specific topic in this area before, so it's important to look into it. To fill this research, gap this thesis is the outcome of in dept study of mobility pattern and how COVID-19 have impacted the mobility on major freeways inside Houston. The use of Machine Learning approach is highly relevant in the present context to better understand the change in pattern of mobility.

CHAPTER 3

METHODOLOGY

In this chapter, the data required for the study will be introduced in two sections. The first section describes the databases from where the data were taken. The used data are listed, and types of data are illustrated which will be used for detailed analysis phase. This section lists the dependent variables used in the research as well as the types of in-dependent) used in the research. The second section presents detailed statistics of the variables that is used in this search. The third section describes the implemented algorithms whereas the fourth section includes the analysis performed for the study.

3.1 Data Description

This research aims to investigate the mobility condition of Harris County during the pre-COVID phase (i.e. the year 2019), COVID-19 phase (i.e. the year 2020), and post-COVID phase (i.e. the year 2021). The dataset utilized for the analysis comprises a total of 2427 rows of freeway segments of each analysis year. Out of these, 2057 rows, or 84.76% of total data, have valid mobility-related information recorded for the respective road segments. Conversely, there is a lack of mobility-related information for the remaining 15.24% of the road segments within the dataset.

The current study focuses on three performance measures in the analysis of freeway segment speeds within Houston: average operating speed, the standard deviation of operating speed, and 85th percentile speed. The speed information utilized for this analysis was collected over a period of five years (2019 to 2021).

This study utilized multi-source data (data from NPMRDS, INRIX, and Wejo) to collect mobility-related information. Among 185 columns from big data, 19 columns were used during the analysis. Those columns contained information about speed, specific identifications of stations, traffic data, freeway features, and accident-related data from three years (2019, 2020, 2021). For each analysis year, route id, object id, land use data (for which this study considered only urban area, median width, number of lanes, right of way, annual average daily traffic (AADT), K-factor, length of the sections, unique id, speed data (average, standard deviation and 85th percentile), posted speed limit, and accident-related data.

Data for the years 2019, 2020, and 2021 were selected and extracted from the large dataset. The data were saved in three spreadsheets, each containing 19 columns including different variables. Then, the data was processed and cleaned by removing the records that had missing data (variables); and the processed files of each year were prepared for analysis.

3.2 Variable Selection

From a large number of available variables from the big dataset, a subset of relevant variables is chosen to build a Machine Learning model. The goal of variable selection is to identify the most important and informative predictors that have a significant impact on the outcome variable while discarding the less important ones. Selecting the appropriate variable selection method depends on the characteristics of the data being analyzed and the objectives of the traffic mobility study. It is crucial to use a method that is suitable for the data and can generate a model that is both accurate and concise.

A dependent variable is a variable that is being measured or observed in response to changes in the independent variable. The dependent variable is what we are measuring or observing to determine the effect of independent variables on it. In this study, average speed, speed standard deviation, and 85th percentile speed was used as dependent variables.

- **Average speed**: Average speed is a measure of the distance traveled by an object over a specified time. It is calculated by dividing the total distance traveled by the object by the time it took to travel that distance. In the analysis, the average speed for the years 2019, 2020, and 2021 is represented by *SpdAve19*, *SpdAve20*, and *SpdAve21*, respectively.
- **Speed standard deviation**: Speed standard deviation is a statistical measure of the variability or spread of the speeds in each set of data which measures how much the individual speeds in the dataset differ from the average speed. In the analysis, standard deviated speed for the years 2019, 2020, and 2021 is represented by *SpdStd19 SpdStd20*, and *SpdStd21*, respectively.
- **85th percentile speed:** 85th percentile speed is a statistical measure commonly used to describe the speed at which 85 percent of vehicles are traveling at or below in a given segment. In the analysis, the 85th percentile for the years 2019, 2020, and 2021 is represented by *Spd85_19*, *Spd85_20*, and *Spd85_21*, respectively.

This paragraph below provides descriptive statistics for a set of independent variables of freeway segments that are used for Traffic mobility measures. After conducting variable importance analysis, the following key variables are considered for the analysis:

- a number of lanes (num_lanes): A number of lanes influences traffic speed and finally impacting on mobility on freeways. The value ranges from 1 to 6 lanes with a mean of 3.74 lanes and a standard deviation of 1.184 lanes across all segments.
- annual average daily traffic (adt_adj): It is a measure used in transportation engineering and planning to estimate the average number of vehicles that travel on a roadway over a

year, divided by the number of days in that year. The value ranges from 499 to 272,758 with a mean of 147034.67 and a standard deviation of 63201.98 across all segments. Annual average daily traffic (AADT) can have a significant impact on traffic mobility as the volume of vehicles on a roadway exceeds its capacity, and traffic congestion can occur, leading to slower travel times impacting mobility.

- median width (med_wid): Median width refers to the width of the central reservation or divider between opposite directions of travel on a freeway or highway. The value ranges from 0 to 58 feet with a mean of 17.93 feet and a standard deviation of 11.83 feet across all segments. The median width of a roadway can have an impact on traffic mobility, particularly in terms of safety and capacity to accommodate the flow of vehicles.
- K-factor (k_fac): It is the proportion of AADT on a roadway segment or link during the design hour, i.e. the hour in which the 30th highest hourly traffic flow of the year takes place and the value ranges from 6.1 to 22.1 with a mean of 10.17and a standard deviation of 0.81 across all segments.
- length of the section (len_sec): It is the length of each section under the study area. The study area contains various study sections, and the value differs based on the length of the section considered. The value of the length of the section ranges from 0.005 to 2.392 miles with a mean of 0.23 miles and a standard deviation of 0.27 miles across all segments.
- vehicle miles traveled (dvmt): It is a measure of the total distance traveled by all vehicles within a specified geographic area and time period. The value which ranges from 5.489 to 382146.6 with a mean of 30875.41874 and a standard deviation of 42758.10166 across all segments. Vehicle miles traveled can have a significant impact on traffic mobility, as

it is a measure of the total distance traveled by all vehicles on a given road or network over time. Congestion, capacity utilization of freeway infrastructure, and travel time are influenced due to vehicle miles traveled which are important parameters in the mobility study.

- crash data (All): The total crashes from the analysis segments are taken into account and summarized in this section. The yearly number of accidents in the segment ranges from 0 to 1020 with a mean and standard deviation of 32.46520681 and 60.05183188 respectively.

Table 1: Summary of Independent Variables

Variables	**abbreviations**	**Min**	**Max**	**Mean**	**STD**
median width	med_wid	0	58	17.92895	11.837
number of lanes	num_lanes	1	6	3.745012	1.183912
Right of way	row_w_usl	0	514	294.0175	106.6164
annual average daily traffic	adt_adj	499	272758	147034.7	63201.98
K-factor	k_fac	6.1	22.1	10.17358	0.805684
length of the section	len_sec	0.005	2.392	0.228943	0.274014
vehicle miles traveled	dvmt	5.489	382146.6	30875.42	42758.1
crash data	All	0	1030	32.46521	60.05183

3.3 Model Selection

Machine Learning is a type of artificial intelligence that enables computer systems to learn and improve from experience without explicit programming. It uses statistical techniques to identify patterns and relationships in data to make predictions or decisions on new data. There are several types of Machine Learning, including supervised, unsupervised, semi-supervised, reinforcement, and deep learning, each with its own unique approaches and applications and is capable of using neural networks to acquire hierarchical understandings of data, and has achieved state-of-the-art performance in various tasks. Shallow learning algorithms can be trained on historical traffic data to predict future traffic patterns based on linear relationships between the input features and the output target. Machine Learning is widely used in the field of traffic mobility to improve the efficiency, performance, and safety of transportation systems. The following algorithms were used for data processing and analysis in this study.

3.3.1 Random Forest Regressor

Random Forest is a Machine Learning algorithm that belongs to the ensemble learning family of algorithms. It is widely used for both regression and classification tasks. The algorithm creates a collection of decision trees, where each tree is trained on a random sample of the data and a random subset of the features. The final prediction is obtained by combining the predictions of all trees, either by voting (in classification) or averaging (in regression). The idea behind using multiple trees is that by aggregating the predictions of many models, the overall performance is improved, and the risk of overfitting is reduced. Additionally, Random Forest also provides feature importance scores, which can be used to determine the most important features in the data. Random Forest is a type of supervised Machine Learning algorithm used for both classification and regression. It is an ensemble method, which means that it combines multiple decision trees to make predictions.

A decision tree is a flowchart-like tree structure, where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome. In Random Forest, multiple decision trees are created using different subsets of the training data and features, and the predictions of all the trees are combined to make the final prediction.

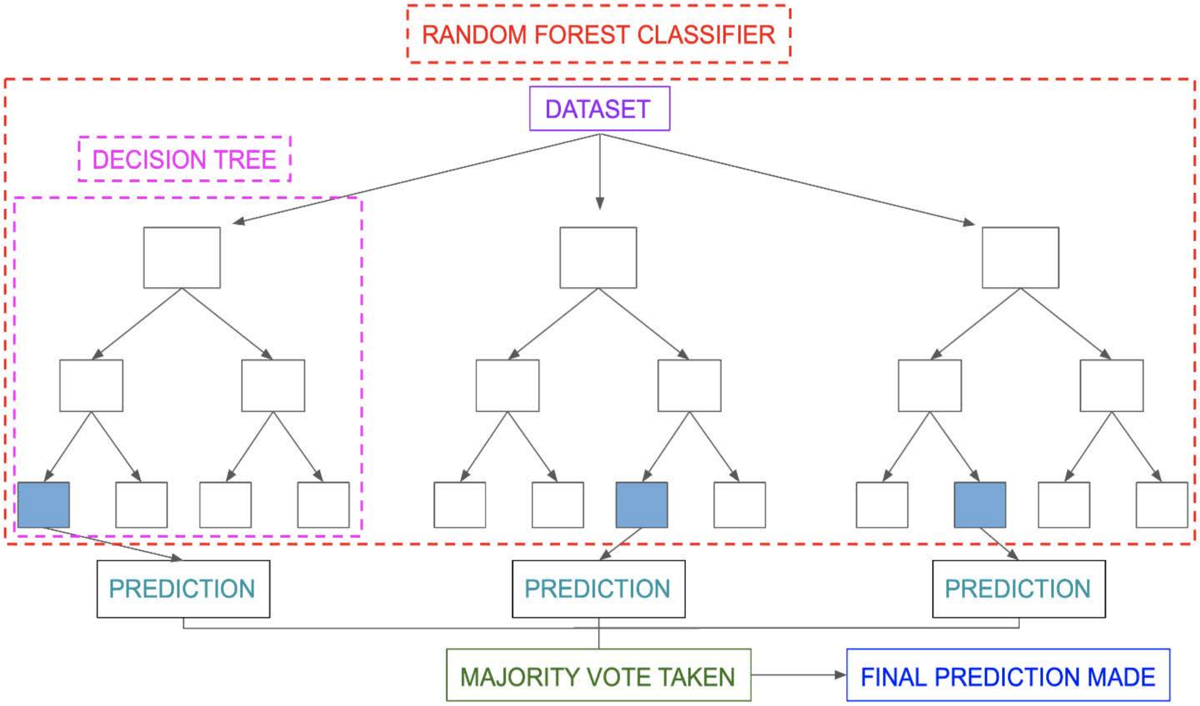


Figure 4. Random Forest Classifier (Source: Section Engineering, 2020)

The margin function in random forest is given by.

$mg\,(X, Y) = av_k\; I(h_k\,(X) = Y\,) - \max\, av_k\; I(h_k\,(X) = j)_{\;j \neq Y}$

where h1 (x), h2 (x), . . ., h_k (x) are ensembles of classifiers, X, and Y are random vectors, and I is the indicator function.

The margin function evaluates the difference in the average number of votes between the right class that exceeds and the average vote for all other classes. A larger margin indicates a higher level of confidence in the classification. The error generalization is given by:

$PE * = P_{X,\,Y}\,(mg\,(X, Y) < 0)$

where X, Y indicates that the probability is over the X, Y space.

The basic idea behind Random Forest is to randomly select a subset of the data and a subset of the features to train each decision tree. By training multiple decision trees with different subsets of data and features, Random Forest can reduce the variance and increase the accuracy of the predictions. Random Forest has several advantages like handling high dimensional data as well as handling missing data and it can handle categorical variables as well as numerical variables. It also can estimate feature importance, which can be used in feature selection. This study used the Python ML library scikit-learn function 'sklearn.ensemble.RandomForestRegressor' to utilize this algorithm.

3.3.2 Gradient Boost Regressor

Gradient Boosting (GB) is a Machine Learning technique used for both classification and regression problems. Friedman (2001) introduced the GB technique, also known as Multiple Additive Trees (MAT), which is an advancement in data mining that builds on Decision Trees (DT). GB employs stochastic GB to enhance the performance of the learning algorithm and increase its accuracy. Boosting is a strategy that uses multiple models with low error rates to enhance accuracy and combining them into an ensemble result in better performance. GB Regressor is a specific type of GB that is used for regression problems. It is an ensemble learning method that combines several models to create a more powerful model. In GB Regressor, weak regression models such as decision trees are trained sequentially, and each new model learns from the errors of the previous model.

GB Machine Learning approach is illustrated in a Figure below. The ensemble classifiers are made up of a group of weak classifiers, and the weight of incorrectly predicted points is raised for the next classifier. The ultimate conclusion is reached by calculating the weighted average of the individual forecasts. GB model is a series expansion that approximates the true functional relationship. The equation for the GB model algorithm is outlined below:

$$f(x) = \sum_{n}(f_n(x)) = \sum_{n}(\beta_n g(x, \gamma_n)) \qquad (1)$$

The estimate of the response variable, denoted as f(x), is based on a set of predictors, x. The single decision trees, g(x, γn), are characterized by the parameter γn, which identifies the split variables. Coefficients βn (n = 1, 2,,n) determine how the single trees are combined. The values of βn are determined by minimizing a specific loss function, L (yi, f(xi)). The performance of the prediction model is evaluated using a loss function, such as deviance. A functional gradient descent optimization method was proposed to improve numerical efficiency. The algorithm to initialize f0(x) is presented below.

i) For n=1,2, 3,m (number of trees)

- For i = 1 to m (number of observations), calculate the residuals

$$\tilde{y}_{in} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f(x)=fm-1(x)} \qquad (2)$$

- Fit a decision tree to $\tilde{y}in$to estimate γ_n
- Estimate βn by minimizing $L\ (y_i, fn - 1(x_i) + \beta_n g(x,\gamma_n))$
- Update $f_n(x) = f_n - 1(x) + \beta_n g\ (x, \gamma_n)$

ii) Calculate $f(x) = \sum_{n} f_n(x)$

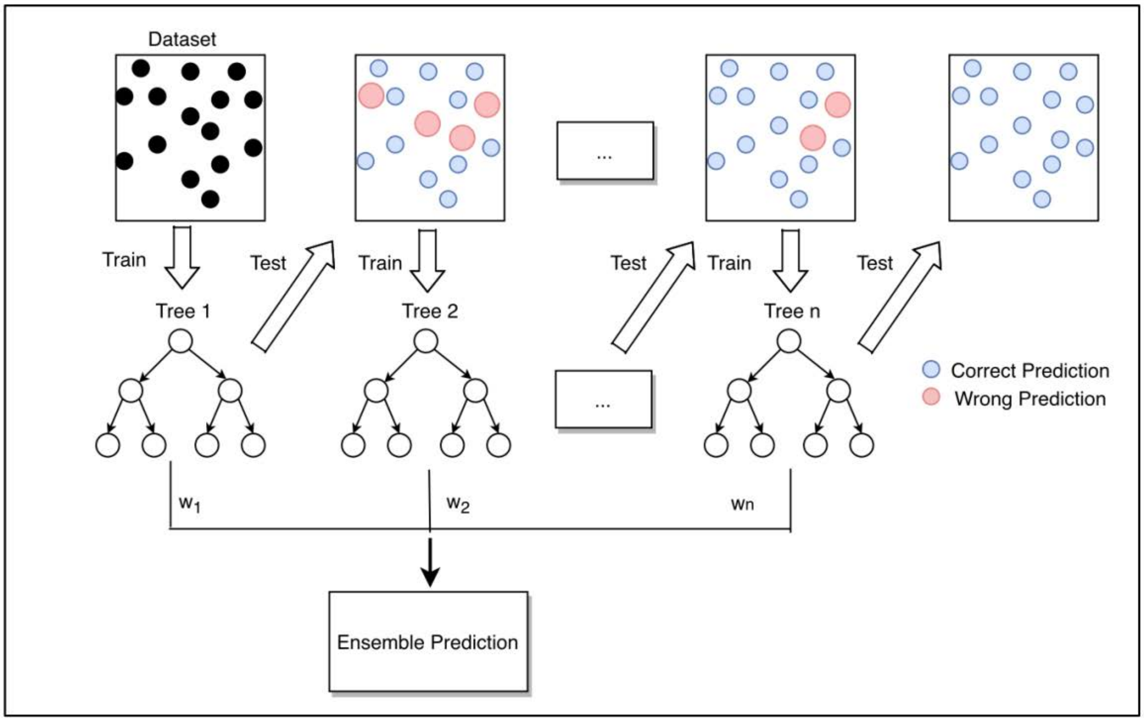


Figure 5. Gradient Boosting (GB) (Source: Zhang et al., 2021)

The process of building decision trees involves adding trees iteratively until the best fit is achieved. Overfitting occurs when the trained models perform well only on the training data but have low prediction accuracy when tested with other datasets. To avoid overfitting, the model is also tested by fitting a test dataset. The iterative training process stops when the model performs well for both the training and test datasets. Regularization parameters can help overcome overfitting and improve model performance. These parameters have two components as learning rate and tree complexity. The learning rate determines the rate at which the model is updated or improved after each stage, with values ranging from 0.0001 to 1.0. Lower learning rates minimize the loss function but require more data and time to run the model, while values closer to 1.0 need less training data but result in overfitting and poor performance. Tree complexity refers to the number of nodes in a single decision tree, with the simplest tree having only two nodes and one split. Balancing both the learning rate and tree complexity rate is crucial to avoid overfitting. The scikit-learn Python ML library provides the function '*sklearn.ensemble.GradientBoostingRegressor*' to implement the GB regression model.

3.3.3 K-Nearest Neighborhood (KNN)

K-Nearest Neighbors (KNN) is a type of supervised Machine Learning algorithm used for classification and regression. The KNN algorithm was originally a classification algorithm proposed by Cover et al. (1967). In recent years, it has been widely used as a non-parametric regression method. The non-parametric method means that it does not make any assumptions about the underlying data distribution. The basic idea behind KNN is that an object or data point is classified based on the majority class of the K nearest points to it. The number of nearest points (K) is a user-specified parameter, which can be chosen based on the specific problem or dataset.

The target value is predicted using state vectors which are in turn constructed using historical and current data. Using Euclidean distance between the present state vector and each previous state 13 vector, K historical moments with the smallest distances are selected as the K-nearest neighbors. The prediction result of the target time can be obtained by calculating the average value of K neighbors at the next time point. Figure 3 illustrates the detailed process of KNN.

The algorithm described here is a valuable data mining technique that utilizes prior data samples with known output values to approximate the output value of a new data sample with unknown output. Rather than making broad assumptions, this algorithm compares new problem sets with previous ones stored in memory. A significant advantage of the neighbor algorithm is its adaptability to handle large amounts of data. Known as memory-based learning, KNN predicts a value or class for a new sample while calculating the distances or similarities to previous training examples. The process involves computing the distances from each point in the KNN master data set to a point in the test data set whose core value is unknown, with neighbors determined by selecting K observations with the shortest distance. Euclidean distance, represented by an equation for points i and j, is typically used in this method for calculating distances.

$$D(i, j)=\sqrt{\sum_{k=1}^{p}(X_{ik} - Xjk)^2} \qquad (3)$$

where i and j are data points on the graph.

Figure below shows an example of the KNN method using K=3 in a scenario with a total of 12 observations comprising six orange and six blue observations. On the left side of Figure 3, a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly occurring class, in this case, blue. On the left side of below, the KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.

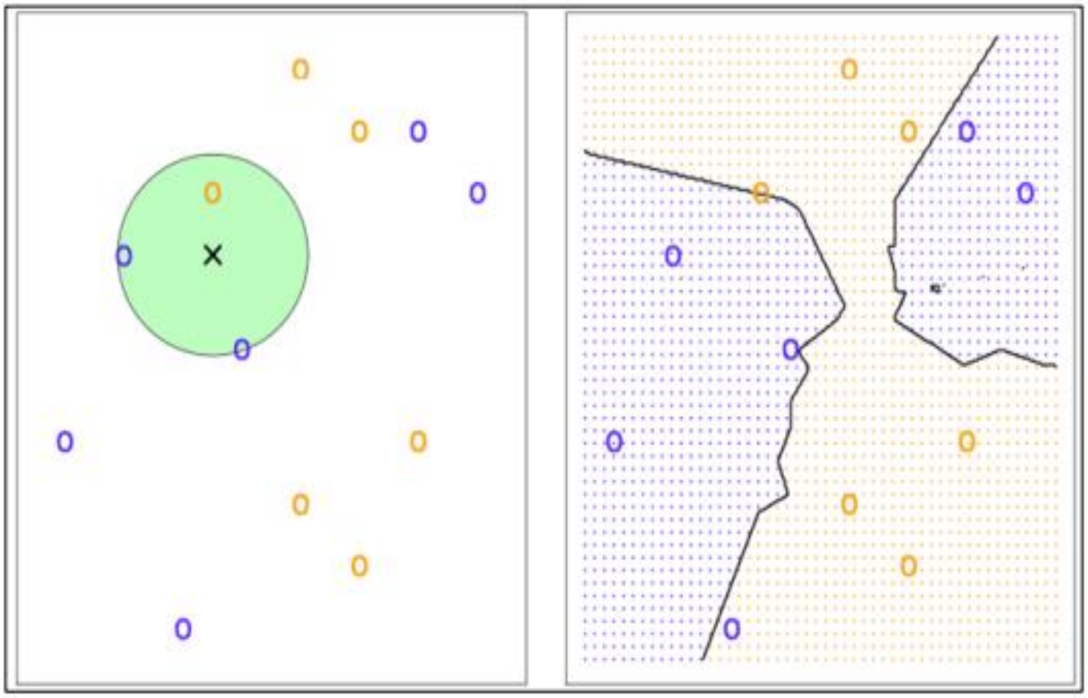


Figure 6. KNN Approach (Source: KNN, Open Source Enthusiast, 2020)

The algorithm first stores all the data and their corresponding classes. When a new data point is encountered, the distance between it and all the stored data points is calculated. The K data points with the smallest distance to the new point are chosen and the majority class among them is assigned to the new point as its class.

3.3.4 CatBoost Regressor

CatBoost is a popular GB Machine Learning algorithm derived from Gradient Boosting Decision Tree (GBDT). It is used for supervised learning tasks, such as classification and regression. CatBoost employs a technique called "ordered boosting," which uses the natural ordering of categorical features to improve training and prediction speed. Boosting algorithms are a type of ensemble method that combines several weak models into a strong model, where each model is trained on the residual errors of the previous model. CatBoost is a deep boosting algorithm because it employs many decision trees, each of which is typically deep. It also incorporates several other features, such as built-in handling of missing values, and has a variety of hyperparameters that can be tuned to optimize performance.

CatBoost is a Machine Learning method that originates from Gradient Boosting Regressor, and it was introduced by Yandex (2017). Gradient Boosting is a robust Machine Learning technique that can address challenges associated with noisy data, complex dependencies, and heterogeneous features. Compared to other Gradient Boosting Regressor algorithms, CatBoost offers several advantages. First and foremost, it is proficient in handling categorical data. Traditional Gradient Boosting Regressor algorithms may replace categorical features with average label values, and node splitting is based on these average label values. This approach is known as Greedy Target-based Statistics (Greedy TBS) which can be represented in following equation:

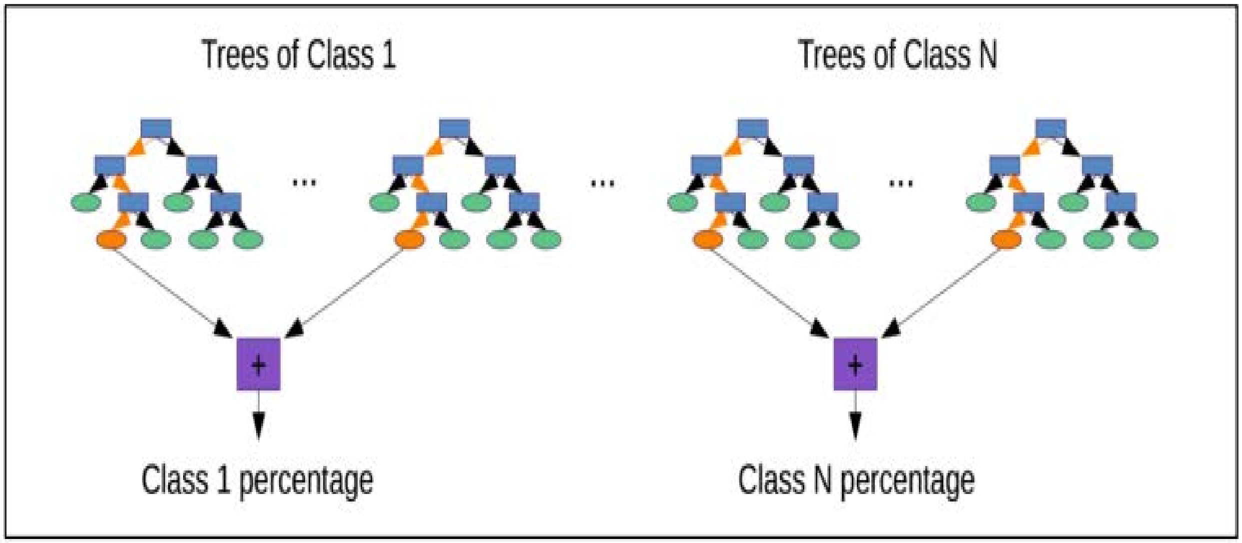


Figure 7. Gradient Boosting Decision Trees used for CatBoost

CatBoost excels at handling categorical data, which is traditionally challenging for GBDT algorithms. Instead of replacing categorical features with average label values as in

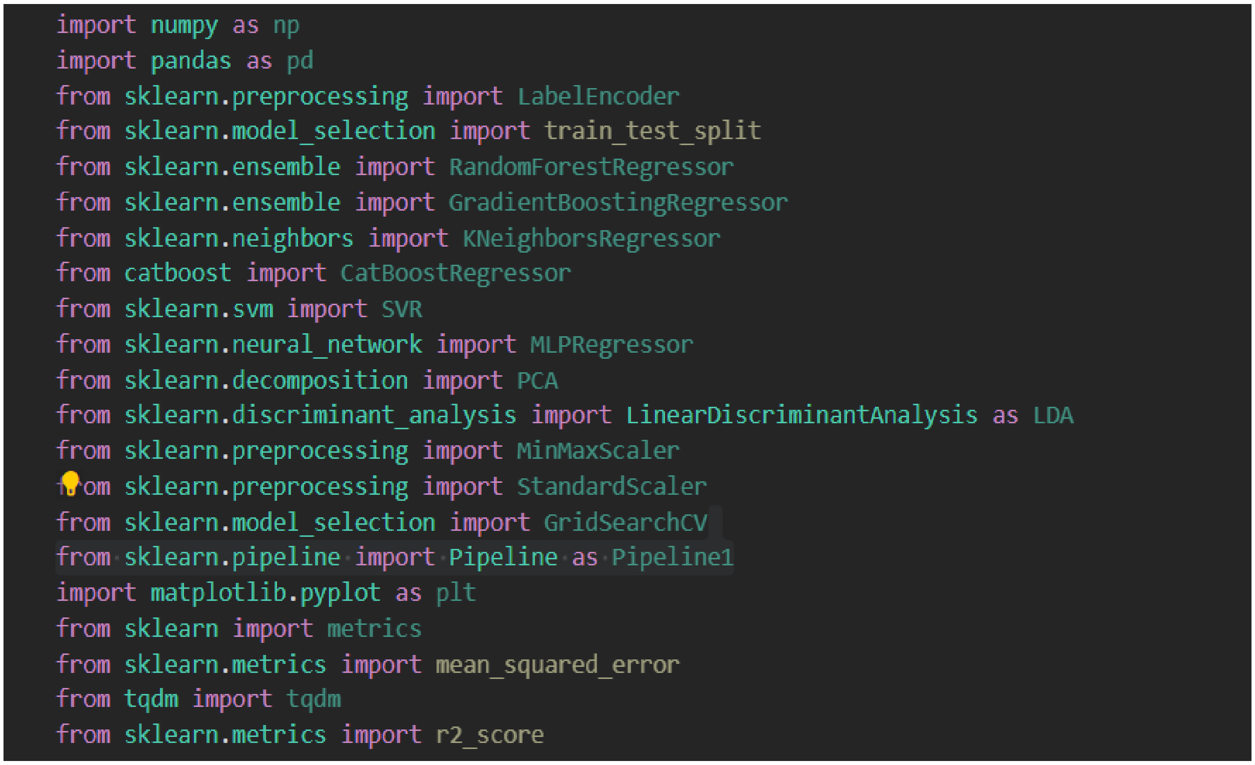
GBDT, CB uses a technique called Greedy Target-based Statistics (Greedy TBS), where node splitting in decision trees is based on the distribution of the target variable within each category. This approach leverages the natural ordering of the categories to better capture the relationship between the features and the target variable.

3.4 Analysis

In this study, Machine Learning techniques are used to analyze traffic mobility inside Harris County. In comparison to traditional statistical models, Machine Learning models utilize more intricate and sophisticated model structures and algorithms. Traditional statistical models assume a linear relationship between the independent and dependent variables, which may not always be held in real-world data. Machine Learning methods can capture non-linear relationships, making them more accurate in modeling complex data. These state-of-the-art methods were designed not just to enhance performance but also capable to grasp the complex nature of traffic flow and mobility. Machine Learning methods can provide more accurate and robust models compared to traditional regression models, making them more useful in many real-world applications where complex and large datasets need to be analyzed. Random Forest regressor, Gradient Boost regressor, KNN regressor, CatBoost regressor, and Support Vector method regressor are used in various stages of Machine Learning during the analysis process.

Machine Learning is increasingly being used in traffic studies to improve the understanding of traffic patterns, predict traffic congestion, traffic mobility study and identify areas where traffic improvements can be made. Machine Learning models provide valuable insights into traffic flow patterns, trends of mobility, traffic congestion, and safety. By using these insights, traffic management can be optimized to improve traffic flow, reduce travel time, and enhance overall system performance. In this thesis the use of Intelligent Transportation Systems and Machine Learning is effectively implemented during data analysis, predictive modeling, and optimization of parameters to better understand the mobility of traffic on freeways inside Harris County.

In this research, several models were utilized to estimate freeway mobility measures. While working on data analytics where this research also focuses to visualize data, all mobility performance measures were fully taken into account. The regression was carried out using different packages from Scikit-learn library, such as GradientBoostingRegressor, KNeighborsRegressor, , MLPRegressor, and the CatBoostRegressor package from the CatBoost library. Visual studio was used for writing the codes for the study. The following steps were taken during the analysis for the study.

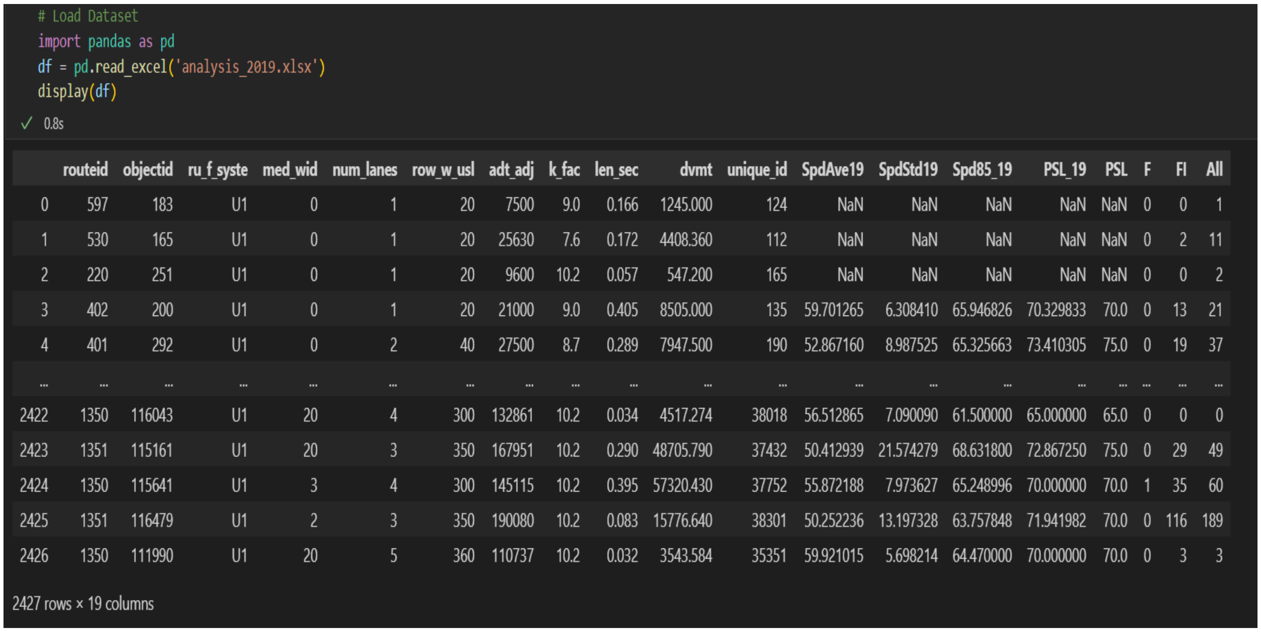


```
import numpy as np
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.neighbors import KNeighborsRegressor
from catboost import CatBoostRegressor
from sklearn.svm import SVR
from sklearn.neural_network import MLPRegressor
from sklearn.decomposition import PCA
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import Pipeline as Pipeline1
import matplotlib.pyplot as plt
from sklearn import metrics
from sklearn.metrics import mean_squared_error
from tqdm import tqdm
from sklearn.metrics import r2_score
```

Figure 8. Packages used for the Analysis

3.3.2 Defining the Packages

Data loading: total of 2427 rows and 19 columns are loaded followed by data cleaning stage.



```
# Load Dataset
import pandas as pd
df = pd.read_excel('analysis_2019.xlsx')
display(df)
```

✓ 0.8s

	routeid	objectid	ru_f_syste	med_wid	num_lanes	row_w_usl	adt_adj	k_fac	len_sec	dvmt	unique_id	SpdAve19	SpdStd19	Spd85_19	PSL_19	PSL	F	FI	All
0	597	183	U1	0	1	20	7500	9.0	0.166	1245.000	124	NaN	NaN	NaN	NaN	NaN	0	0	1
1	530	165	U1	0	1	20	25630	7.6	0.172	4408.360	112	NaN	NaN	NaN	NaN	NaN	0	2	11
2	220	251	U1	0	1	20	9600	10.2	0.057	547.200	165	NaN	NaN	NaN	NaN	NaN	0	0	2
3	402	200	U1	0	1	20	21000	9.0	0.405	8505.000	135	59.701265	6.308410	65.946826	70.329833	70.0	0	13	21
4	401	292	U1	0	2	40	27500	8.7	0.289	7947.500	190	52.867160	8.987525	65.325663	73.410305	75.0	0	19	37
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2422	1350	116043	U1	20	4	300	132861	10.2	0.034	4517.274	38018	56.512865	7.090090	61.500000	65.000000	65.0	0	0	0
2423	1351	115161	U1	20	3	350	167951	10.2	0.290	48705.790	37432	50.412939	21.574279	68.631800	72.867250	75.0	0	29	49
2424	1350	115641	U1	3	4	300	145115	10.2	0.395	57320.430	37752	55.872188	7.973627	65.248996	70.000000	70.0	1	35	60
2425	1351	116479	U1	2	3	350	190080	10.2	0.083	15776.640	38301	50.252236	13.197328	63.757848	71.941982	70.0	0	116	189
2426	1350	111990	U1	20	5	360	110737	10.2	0.032	3543.584	35351	59.921015	5.698214	64.470000	70.000000	70.0	0	3	3

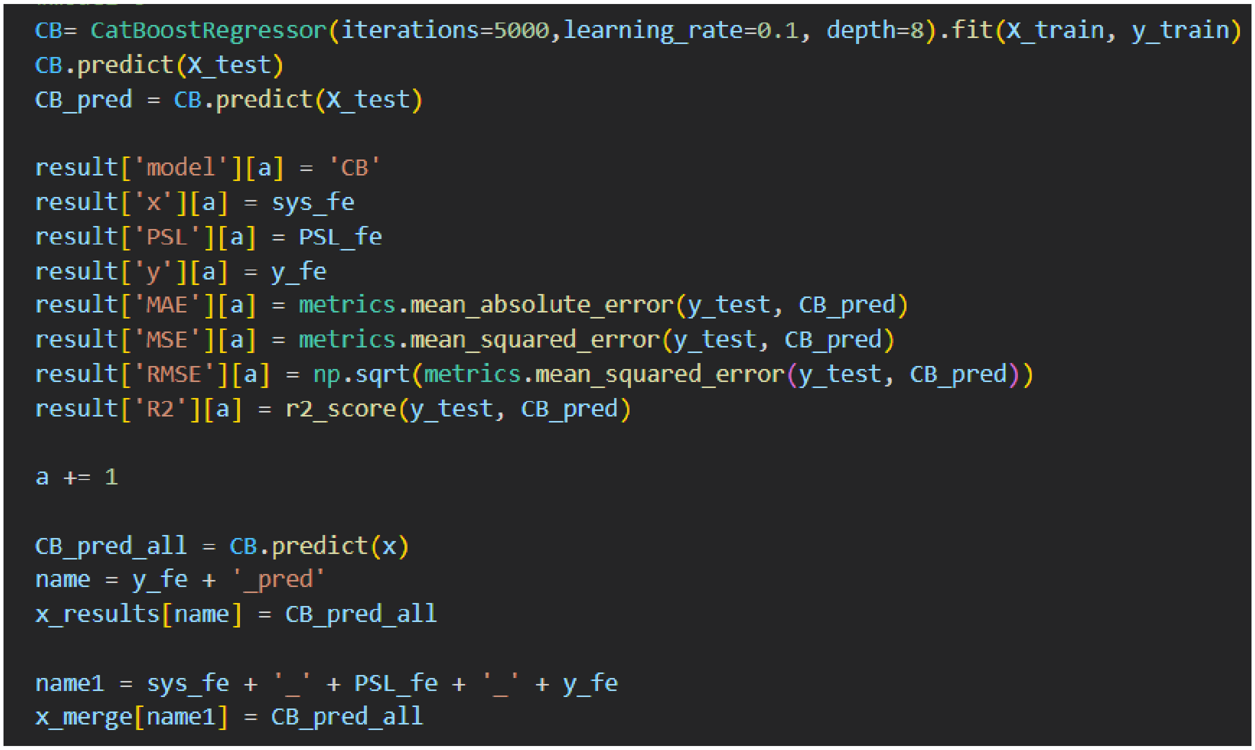
2427 rows × 19 columns

Figure 9. Data Loading in Visual Studio

The Random Forest (RF) model has an optimized parameter of n_estimators = 500, max_depth=8, 19 and random_state=1. The GB model has an optimized parameter of n_estimators = 1000, max_depth=8, learning_rate = 0.01, and random_state=0. The KNN model has an optimized parameter of n_neighbors=50 and weights='distance'. The SVR model has an optimized parameter of C=1 and epsilon=0.1. The Multi-layer Perceptron (ANN) model has an optimized parameter of activation= 'tanh', solver = 'lbfgs', max_iter=1000. The CatBoostRegressor (CB) model has an optimized parameter of iterations=5000, learning_rate=0.1, and depth=8. In the model selection process, only the different types of land use are considered.

3.3.3 Optimized Parameters used in CatBoosting:

This research used CatBoost modeling and used iterations of 5000 with learning rate of 0.1 and depth of 8. Iterations learning rate and depth are important hyperparameters. Each boosting iteration involves adding a new decision tree to the model that corrects the errors of the previous iteration. In this case, the model is trained with 5000 boosting iterations, which means that 5000 decision trees will be added to the model. Learning rate is a hyperparameter that determines the step size at which the algorithm updates the model with each new tree. The depth is a hyperparameter that controls the complexity of the individual decision trees in the model.



```
CB= CatBoostRegressor(iterations=5000,learning_rate=0.1, depth=8).fit(X_train, y_train)
CB.predict(X_test)
CB_pred = CB.predict(X_test)

result['model'][a] = 'CB'
result['x'][a] = sys_fe
result['PSL'][a] = PSL_fe
result['y'][a] = y_fe
result['MAE'][a] = metrics.mean_absolute_error(y_test, CB_pred)
result['MSE'][a] = metrics.mean_squared_error(y_test, CB_pred)
result['RMSE'][a] = np.sqrt(metrics.mean_squared_error(y_test, CB_pred))
result['R2'][a] = r2_score(y_test, CB_pred)

a += 1

CB_pred_all = CB.predict(x)
name = y_fe + '_pred'
x_results[name] = CB_pred_all

name1 = sys_fe + '_' + PSL_fe + '_' + y_fe
x_merge[name1] = CB_pred_all
```

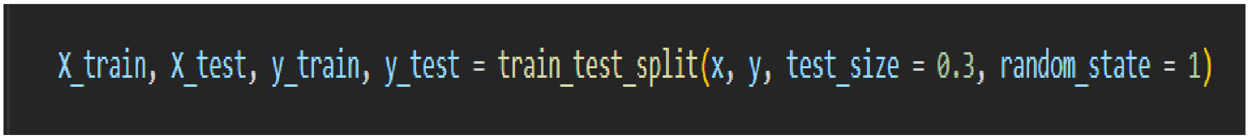
Figure 10. Parameters Used for Optimization

3.3.4 CatBoost Modeling

The results of a CatBoost model applied to predict speed-related measures (average speed, standard deviation of speed, and 85th percentile speed) for urban land use at all Posted Speed Limit(PSL) = and three = Posted Speed Limit = (65, 70, and 75) are illustrated in the result section. The R-squared values determine how effectively the model clarifies the variation in the target variable, ranging from 0 to 1, where higher values imply a superior fit. On the other hand, the MAE, MSE, and RMSE serve as metrics to gauge the accuracy of the model's predictions, with smaller values denoting better performance. The MAE, MSE, and RMSE results for the models presented in the table suggest excellent prediction accuracy, with values that are relatively low when compared to the target variable's range. From the results, it is noteworthy that the prediction accuracy appears to enhance when developing models based on various PSL levels, implying that PSL levels might serve as a useful categorization or filtering alternative for congestion measures related to speed.

3.3.5 Data Training and Testing

Data training and testing are required in Machine Learning to ensure that the Machine Learning model is accurate and can generalize to new data which are useful in model building and evaluate a model's performance. The following figure shows the percentage of data used during training and testing. This research used 70% of the total data to training the model and remaining 30 % of the data to test the model.

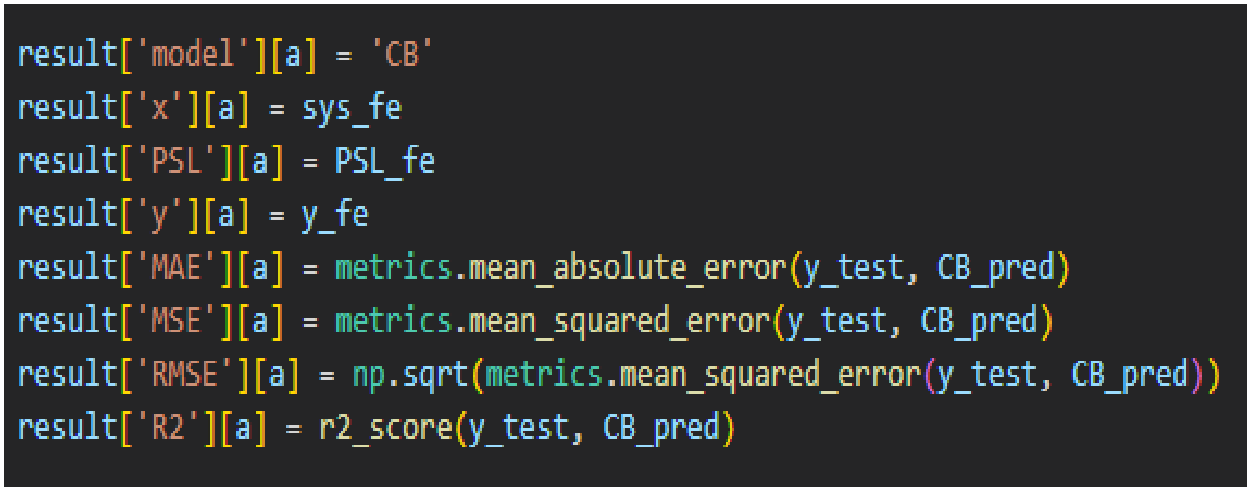


```
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 1)
```

Figure 11. Training and Testing Parameters

3.3.6 Model Evaluation

Model evaluation is an essential aspect of Machine Learning. It involves measuring the performance of a Machine Learning model by assessing how well it can generalize to new data. This involves checking of overfitting, and underfitting, bias variance tradeoff evaluating the matrices and cross validation. This helps in understanding effectiveness about model evaluation for developing accurate and reliable AI algorithms.



```
result['model'][a] = 'CB'
result['x'][a] = sys_fe
result['PSL'][a] = PSL_fe
result['y'][a] = y_fe
result['MAE'][a] = metrics.mean_absolute_error(y_test, CB_pred)
result['MSE'][a] = metrics.mean_squared_error(y_test, CB_pred)
result['RMSE'][a] = np.sqrt(metrics.mean_squared_error(y_test, CB_pred))
result['R2'][a] = r2_score(y_test, CB_pred)
```

Figure 12. Errors calculation

3.5 Shapley Additive Explanations (SHAP)

SHAP method is used to clarify the results obtained from the XGBoost model. In the past, Machine Learning techniques were often referred to as "black boxes" because it was challenging to comprehend how each factor affected the model's outcomes. However, the SHAP approach utilizes a game-theoretic strategy to estimate the specific contribution of each feature to the model's prediction. This technique is suitable for explaining tree-based Machine Learning models. To determine the Shapely value of a particular feature for a given prediction, we can use the following equation.:

$$\emptyset_i \quad = \quad \sum_{S \subseteq F \setminus \{i\}}^{T} \frac{|\mathrm{S}|!(|\mathrm{F}|-|\mathrm{S}|-1)!}{|\mathrm{F}|!} [\, f_{SU\{i\}}\big(X_{SU\{i]}\big) - (f_s(X_s)] \qquad (4)$$

where:

ϕi: shapely value of feature i

S: a possible feature subset

$|S|$: feature counts in subset S

F: the set of all features

$|F|$: feature counts in set F

$fS\cup\{i\}(xS\cup\{i\})$: model prediction based on the features in subset S and feature i

$fS(xS)$: model prediction based on subset S features.

Here, $fS\cup\{i\}(xS\cup\{i\}) - fS\ (xS)$ is the distinction between a model prediction generated only using features from subset S and a model prediction made using features from subset S and feature i. This difference shows how the model's prediction outcomes may change with the inclusion of feature I. The Shapely value for a specific characteristic in any given prediction is calculated using the weighted average of all differences across all possible subsets.

The SHAP method was utilized to understand the importance of each feature in the CatBoost model. The impact of each explanatory variable on the model's predictions has been evaluated and ranked accordingly. Features with higher levels of importance indicate that they have a more substantial effect on the SpdAve measures. For rural freeways, the top factors in order of importance are traffic volume (adt_adj), K-factor, and median width.

CHAPTER 4

RESULT

After training and testing the model, the outcomes were received in terms of predicted speed for different years under various speed limits in urban land use. Based on the analysis of traffic data using artificial intelligence, the results indicate that AI can effectively predict traffic mobility. The AI algorithm was able to accurately predict traffic flow speeds. The scatter plots below are plotted between observed and predicted mobility performance measures for different subsets based on posted speed limits and different study years. The linear trends of the majority of the models indicate that AI models are well-suited for determining the predicted values. Linear trends are typically considered to be a positive indication of the effectiveness of a model, as they suggest that there is a strong correlation between the input data and the predicted values.

4.1 Model Performances

The scatter diagrams for the predicted and observed mobility performance measures in different subsets are categorized by the posted speed limits. Most of the scatter plots diagram predicted speed and operating speed re closer to each other, demonstrating that used AI models is appropriate for computing predicted speed values. One of the key achievements of this research is to compute mobility measures for study area (selected freeways) lacking historical data on speed or mobility.

At PSL= 65 mph

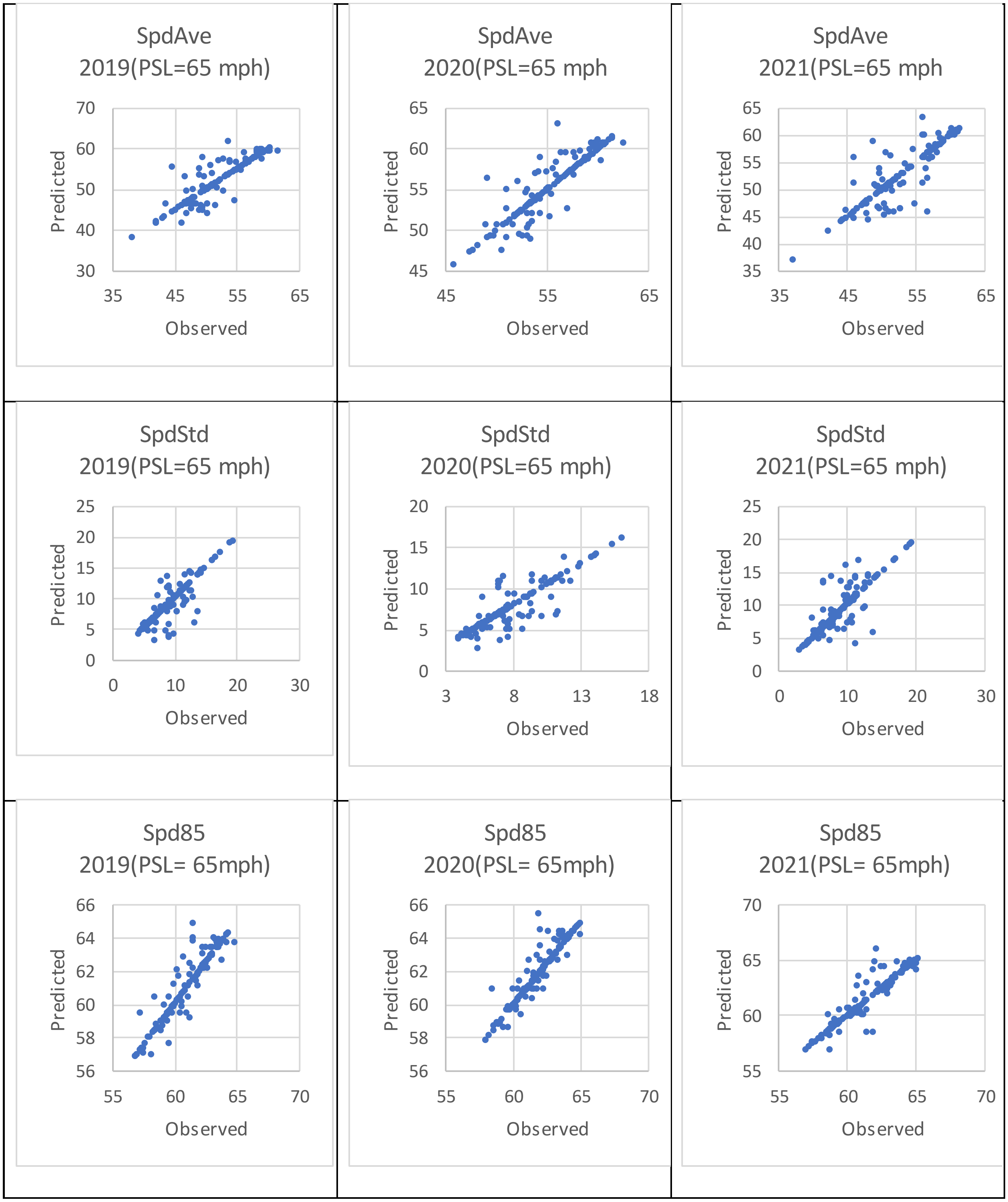
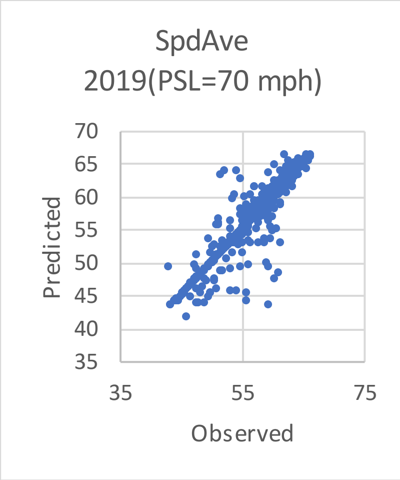
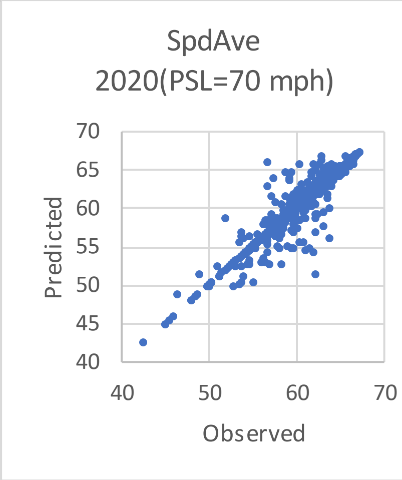
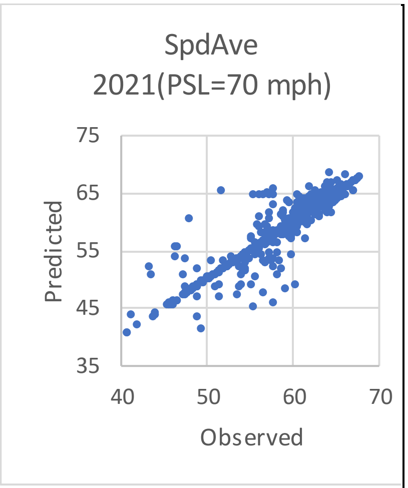


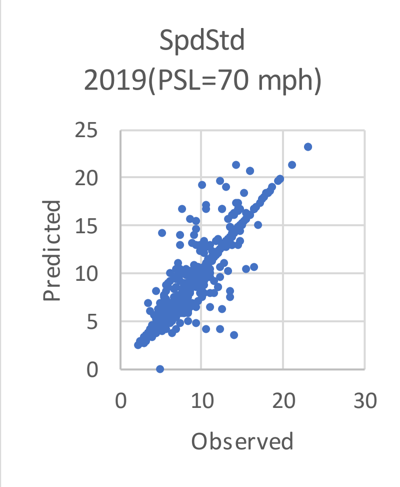
Figure 13. Observed and predicted speed (at PSL= 65 mph)

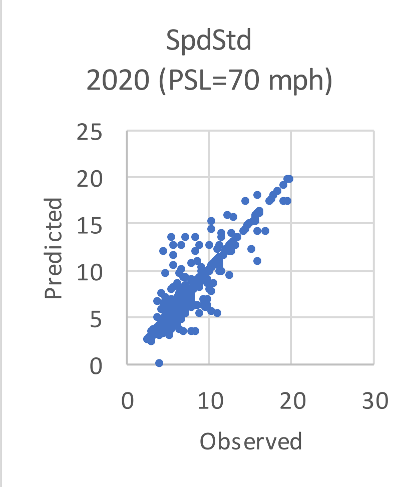
At PSL= 70 mph

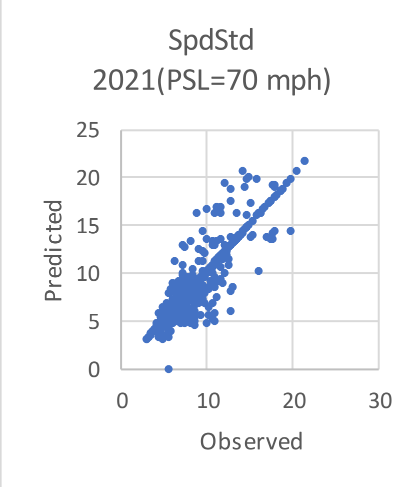

SpdAve
2019(PSL=70 mph)
Predicted
70
65
60
55
50
45
40
35
35
55
75
Observed


SpdAve
2020(PSL=70 mph)
Predicted
70
65
60
55
50
45
40
40
50
60
70
Observed


SpdAve
2021(PSL=70 mph)
Predicted
75
65
55
45
35
40
50
60
70
Observed


SpdStd
2019(PSL=70 mph)
Predicted
25
20
15
10
5
0
0
10
20
30
Observed


SpdStd
2020 (PSL=70 mph)
Predicted
25
20
15
10
5
0
0
10
20
30
Observed


SpdStd
2021(PSL=70 mph)
Predicted
25
20
15
10
5
0
0
10
20
30
Observed

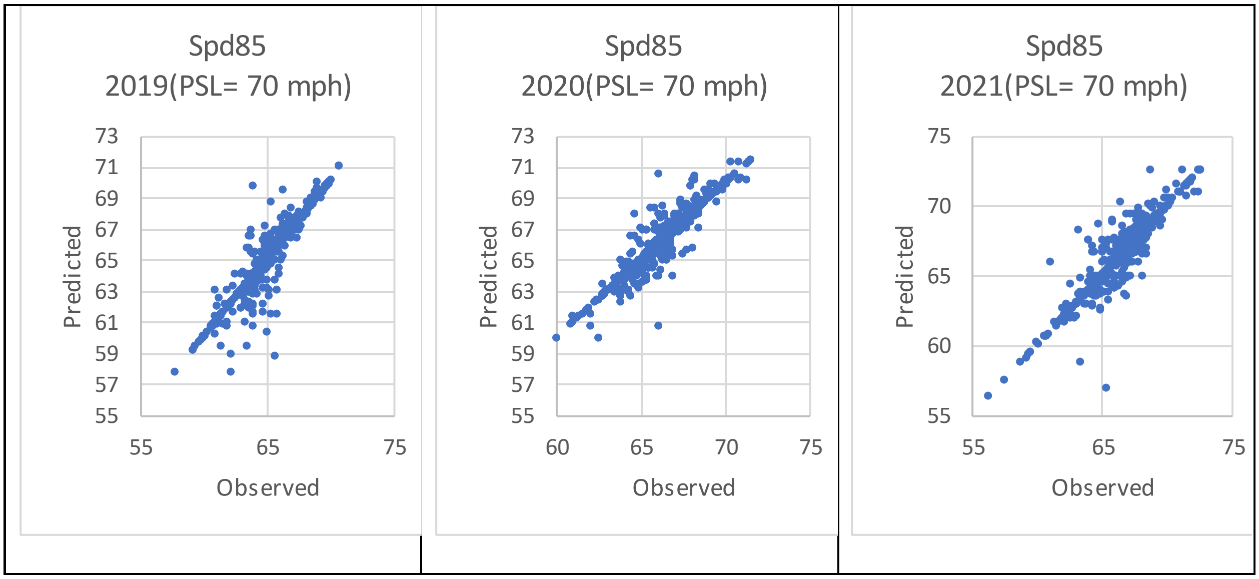
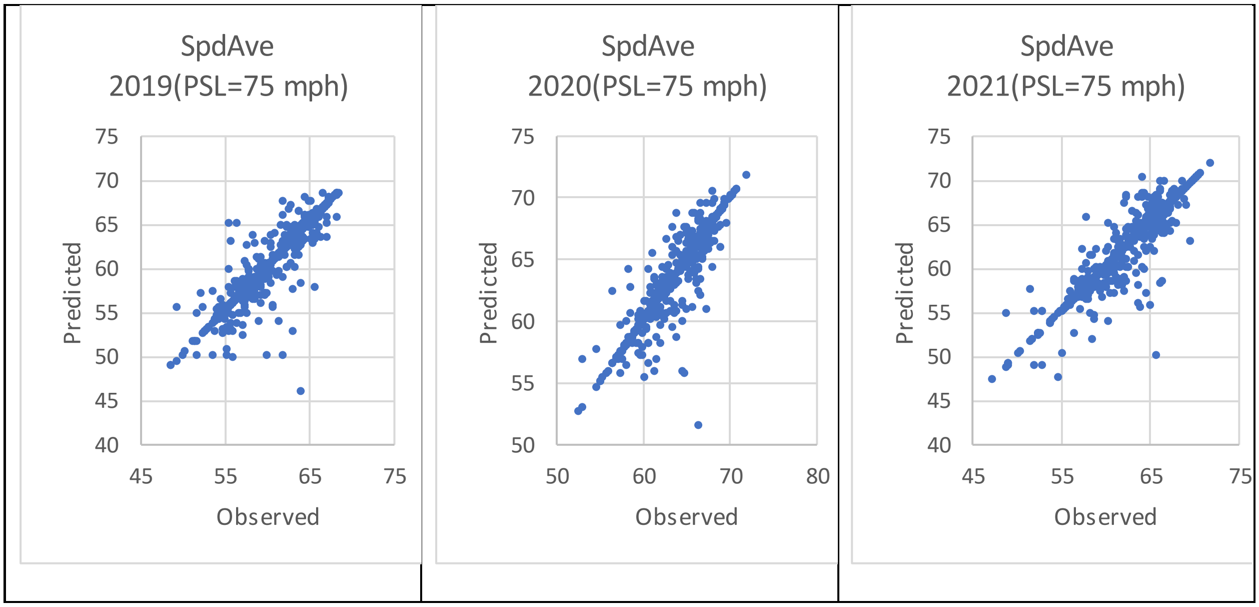


Figure 14. Observed and predicted speed (at PSL= 70 mph)

At PSL= 75 mph



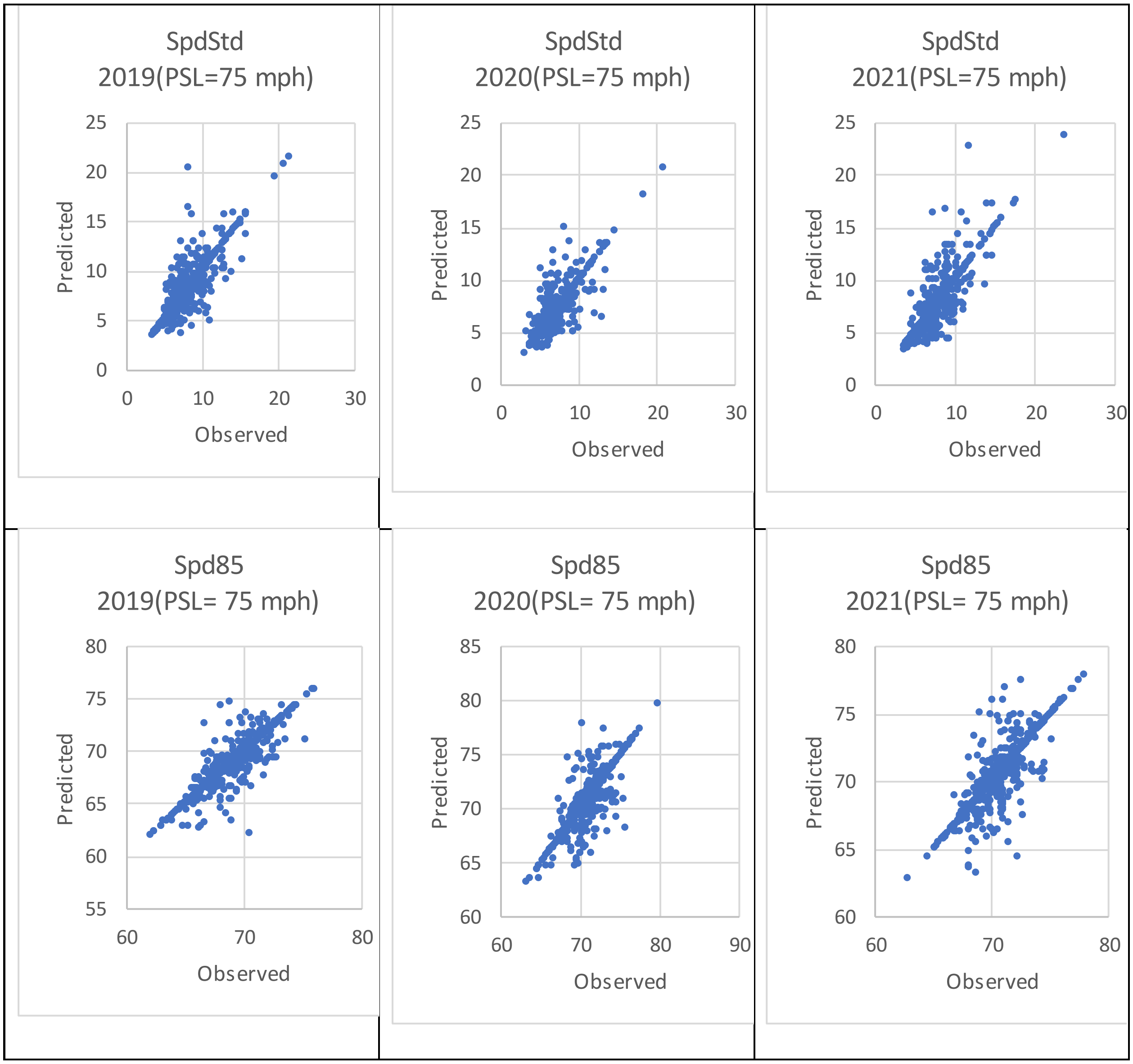
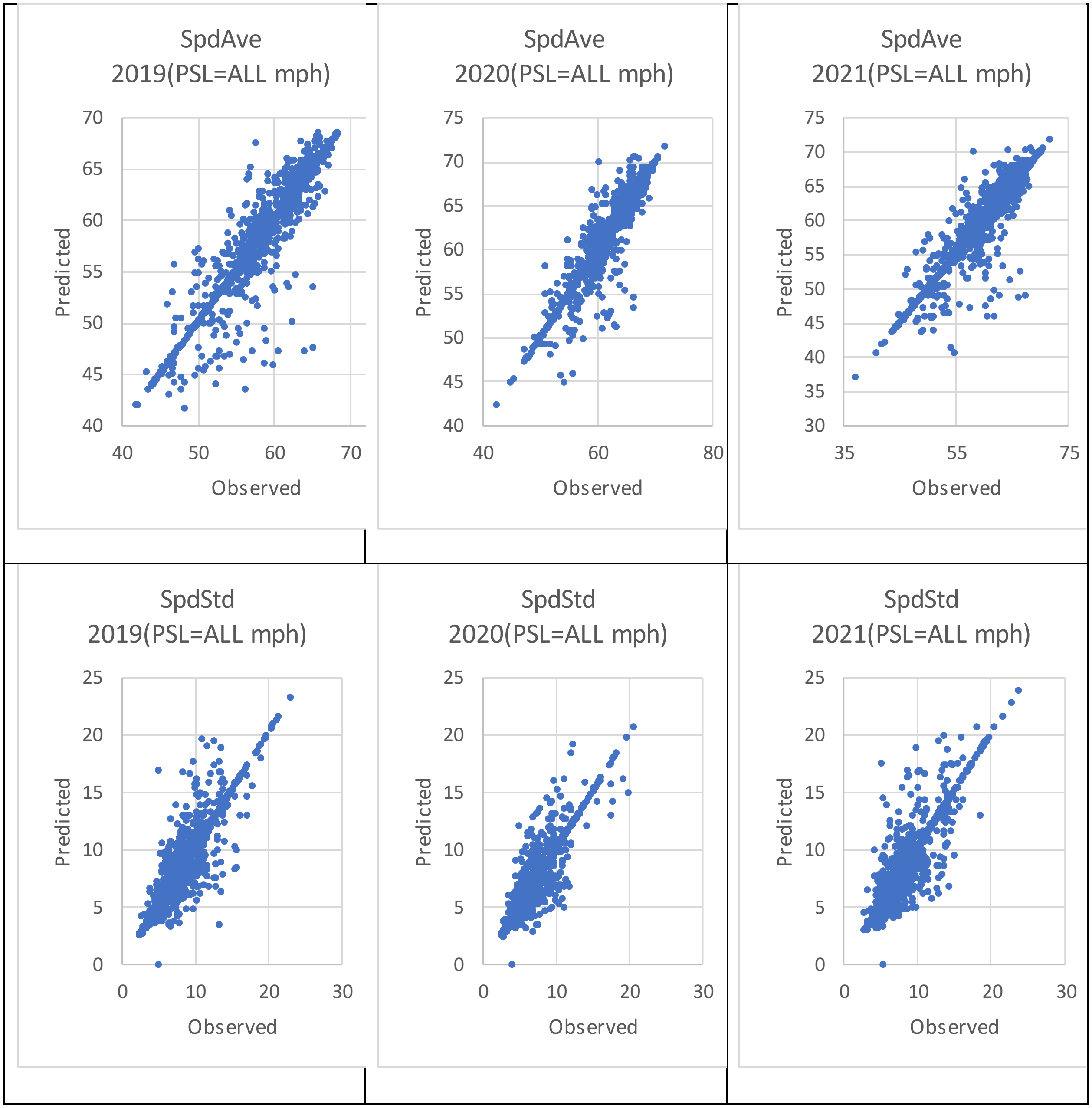


Figure 15. Observed and predicted speed (at PSL= 75 mph)

At PSL= All


SpdAve
2019(PSL=ALL mph)
Predicted
Observed
70
65
60
55
50
45
40
40
50
60
70
SpdAve
2020(PSL=ALL mph)
Predicted
Observed
75
70
65
60
55
50
45
40
40
60
80
SpdAve
2021(PSL=ALL mph)
Predicted
Observed
75
70
65
60
55
50
45
40
35
30
35
55
75
SpdStd
2019(PSL=ALL mph)
Predicted
Observed
25
20
15
10
5
0
0
10
20
30
SpdStd
2020(PSL=ALL mph)
Predicted
Observed
25
20
15
10
5
0
0
10
20
30
SpdStd
2021(PSL=ALL mph)
Predicted
Observed
25
20
15
10
5
0
0
10
20
30

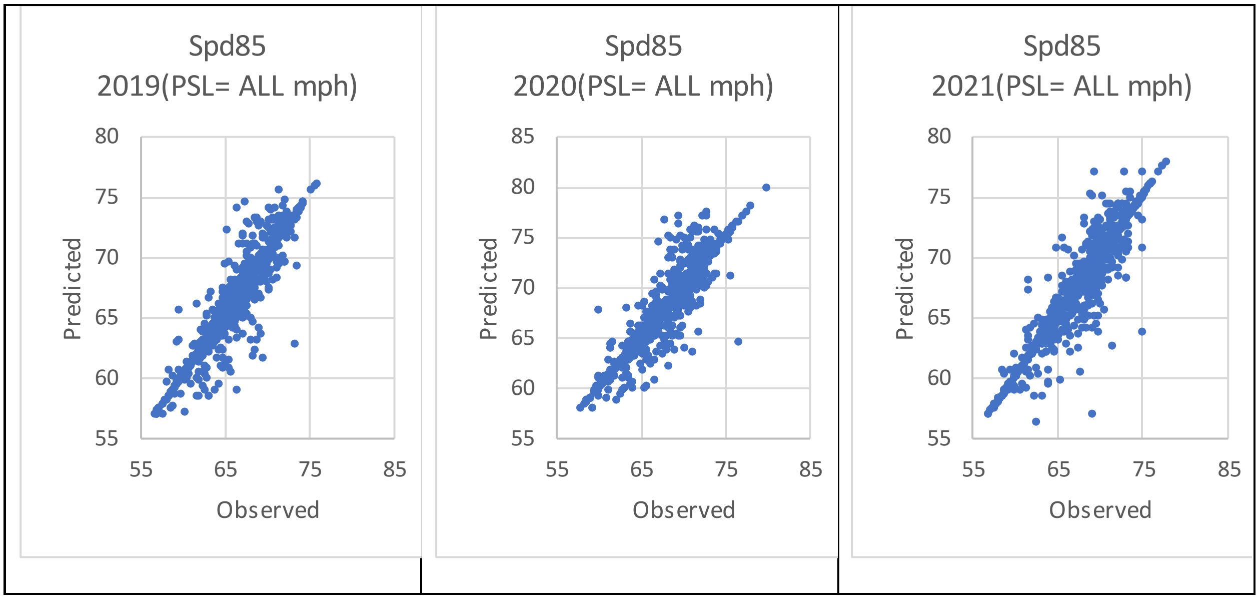


Figure 16. Observed and predicted speed (at PSL= ALL)

4.2 Errors and Model Evaluation Criteria

The model performance evaluation relied on critical indicators, including the root mean square error (RMSE) Mean Absolute Error (MAE), which stands for Mean Squared Error (MSE), and the coefficient of determination, R2 (also known as R-squared). Ultimately, the model with the best-predicted performance was recommended. The modeling results from CatBoost algorithms is tabulated below for different years in urban land use.

Table 2: CatBoosting Modeling Results

	Year	PSL (mph)	**MAE**	**MSE**	**RMSE**	**R2**
SpdAve	2019	65	2.715795	13.18961	3.63175	0.611071
		70	1.81909	9.322579	3.05329	0.684335
		75	1.840055	8.080825	2.842679	0.586292
		ALL	1.977069	9.408856	3.067386	0.686266
	2020	65	1.971632	6.349697	2.51986	0.658872
		70	1.348943	4.72096	2.172777	0.704527
		75	1.515728	5.199044	2.280141	0.560099

		ALL	1.642825	6.648176	2.578406	0.673655
	2021	65	2.86325	15.27001	3.907687	0.513545
		70	1.927446	9.607671	3.099624	0.661669
		75	2.021466	8.270874	2.875913	0.568991
		ALL	2.068708	10.85316	3.294414	0.64533
SpdStd	2019	65	1.793072	6.590817	2.567259	0.283187
		70	1.607836	5.838905	2.416383	0.581464
		75	1.491993	4.43915	2.106929	0.285246
		ALL	1.581241	5.058954	2.249212	0.48174
	2020	65	1.544446	4.102793	2.025535	0.423535
		70	1.162202	3.271934	1.808849	0.641159
		75	1.307428	3.329886	1.824798	0.184001
		ALL	1.310564	3.49912	1.870593	0.487366
	2021	65	2.091253	8.372337	2.893499	0.260685
		70	1.614802	5.170156	2.273798	0.618243
		75	1.607576	4.797151	2.19024	0.339224
		ALL	1.57532	5.050776	2.247393	0.426432
Spd85	2019	65	0.834991	1.259448	1.122251	0.761673
		70	0.721623	1.431637	1.19651	0.693647
		75	1.290028	3.230023	1.797227	0.433256
		ALL	1.160223	3.22862	1.796836	0.717375
	2020	65	0.716074	0.969628	0.984697	0.730634
		70	0.639258	0.928965	0.963828	0.758272
		75	1.504375	4.414752	2.101131	0.203469
		ALL	1.279377	3.872402	1.967842	0.685415
	2021	65	0.827943	1.49452	1.222506	0.744662
		70	0.850194	1.681947	1.296899	0.653133
		75	1.568677	4.486767	2.118199	0.261303
		ALL	1.28585	3.764432	1.940214	0.684804

4.2.1 Mean Absolute Error (MAE)

Absolute error represents the amount of error in a prediction and is defined by the difference between the predicted value and the true value. The mean absolute error (MAE) is the average of all absolute errors and is defined by Equation 10. The scale or unit of the MAE is the same as the original data and therefore can be named a scale-dependent accuracy measure. The

MAE has only nonnegative values since absolute values are considered, and therefore can avoid mutual cancellation of the positive and negative errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |p_i - p_o|$$

where n is the number of errors or sample size, and O_i and P_i are the true value and predicted value of the i^{th} observation, respectively. The bar graph below shows how the model is performing in various speed limits taken into consideration by calculating mean absolute error.

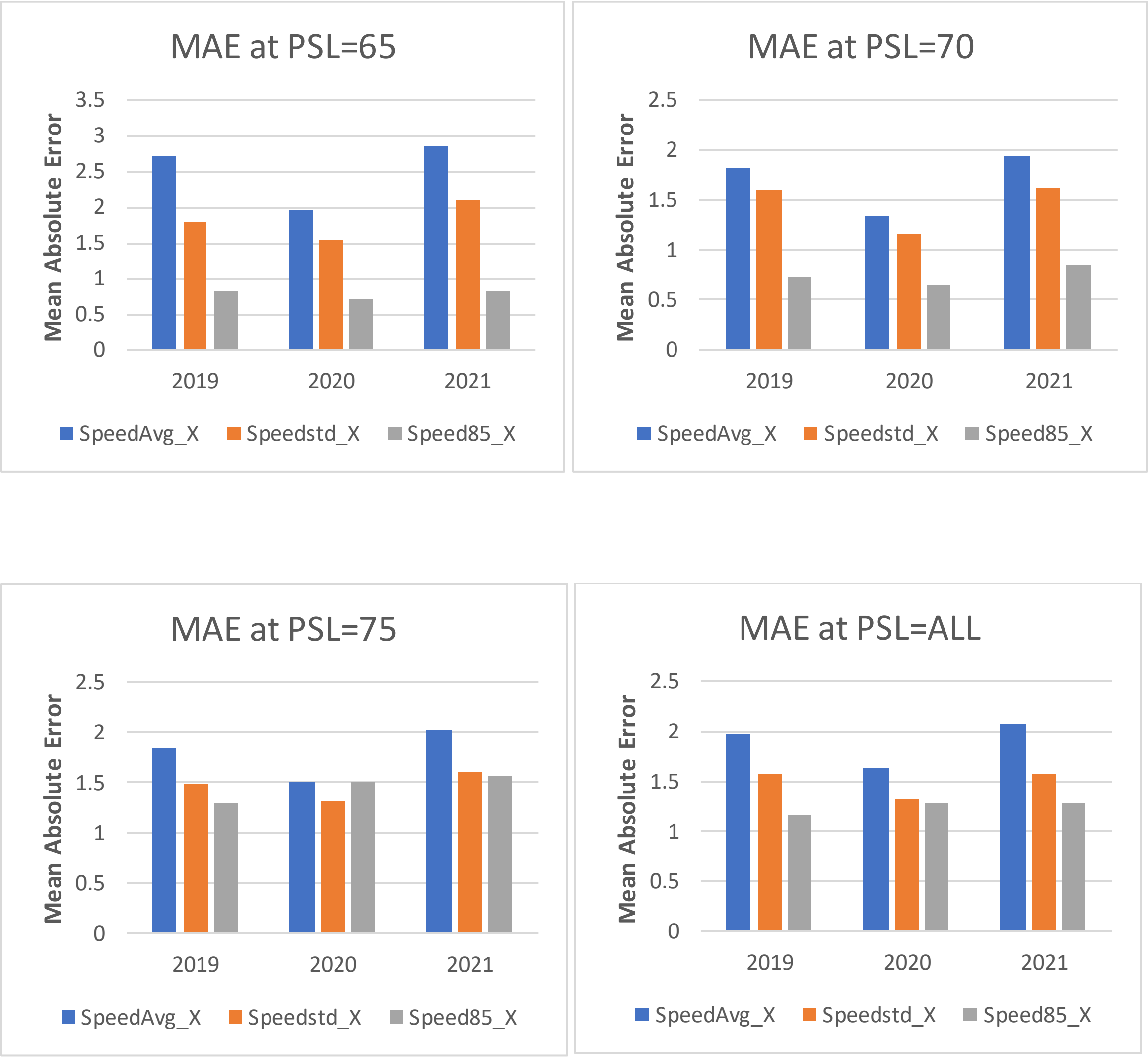


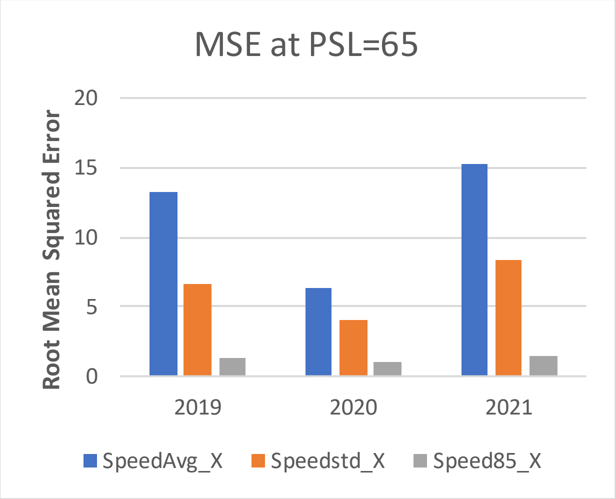
Figure 17. MAE at different PSL

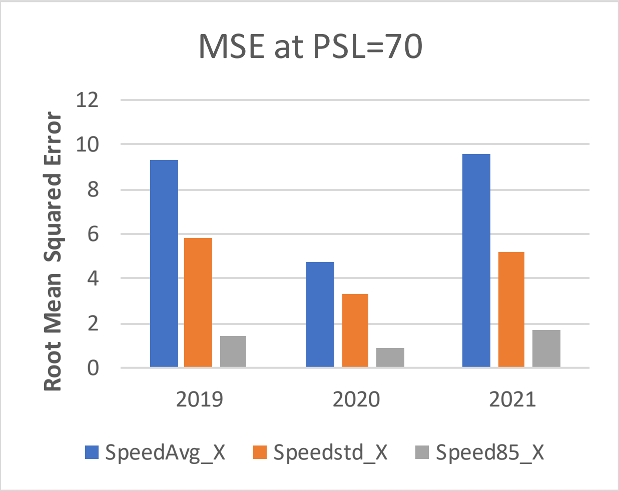
4.2.2 Mean Squared Error (MSE)

Mean squared error (MSE) is calculated by averaging the square of all errors and defined by equation 11. The lower the MSE value, the better the performance of the prediction models. Higher error values are penalized more than the lower ones due to the nature of the square function, and therefore for outliers, MSE will become much larger compared to MAE. Also, the unit of MSE is different than the original data.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - p_o)^2$$

Wwhere n is the number of errors or sample size, and O_i and P_i are the true value and predicted value of the i th observation, respectively.





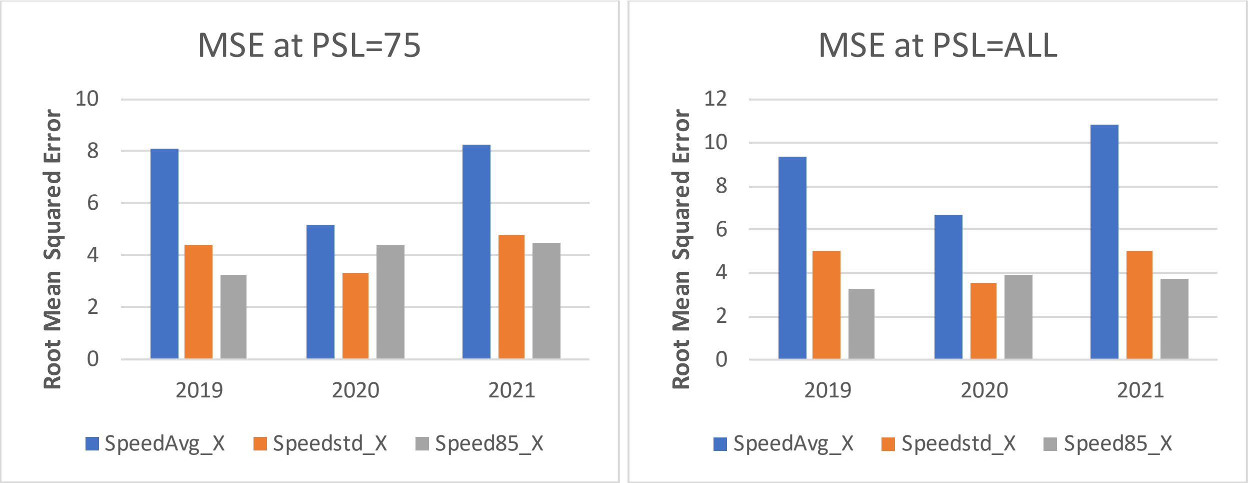


Figure 18. MSE at different PSL

4.2.3 Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) is defined as the square root of the average of the square of all the errors, as described in Equation 12. RMSE is always non-negative, and a lower value indicates a good fit for the data. In general, a lower RMSD is better than a higher one. Although RMSE is a good performance evaluation matrix, it cannot be used to compare between variables as it is a scale-dependent matrix.

$$\mathrm{R}MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_o)^2}$$

where n is the number of errors or sample size, and O_i and P_i are the true value and predicted value of the i $^{\text{th}}$ observation, respectively.

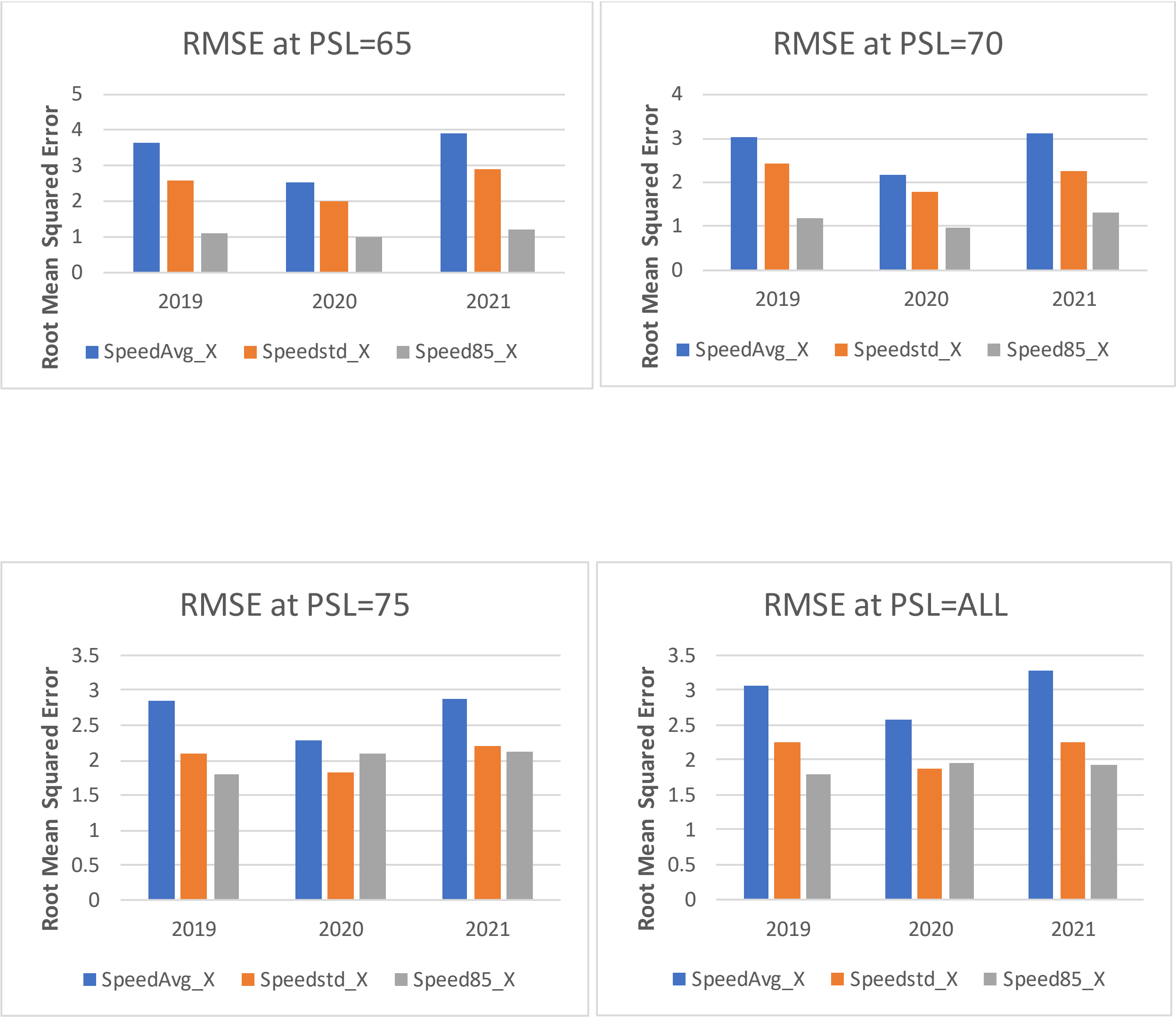


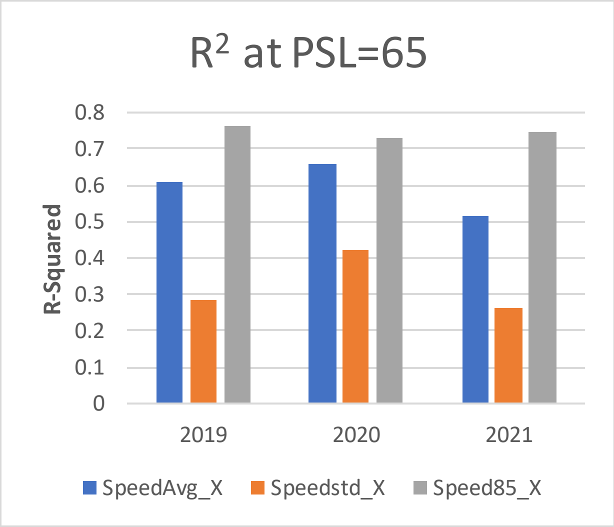
Figure 19. RMSE at different PSL

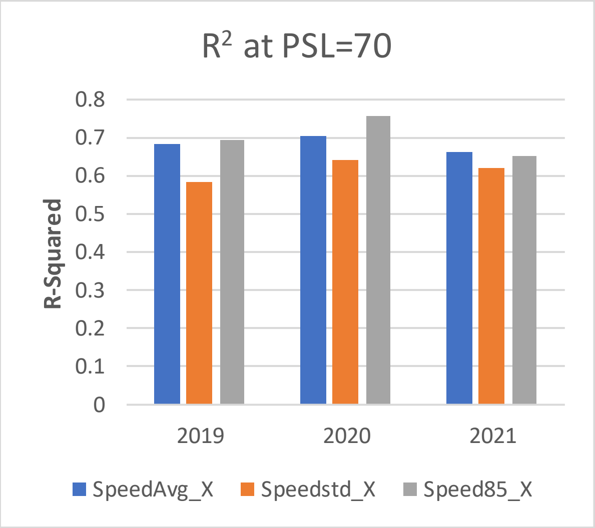
4.2.4 R-Squared

R-squared (R^2) also known as the coefficient of determination is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variable(s) in a regression model. In other words, R^2 measures how well the regression model fits the data, with a value between 0 and 1, where 0 indicates that the model does not explain any of the variability in the data, and 1 indicates a perfect fit.

$$R^2 = 1\text{-}\frac{sum\ of\ squares\ of\ residuals}{total\ sum\ of\ squares}$$

Mathematically, R^2 is calculated as the ratio of the explained variance to the total variance. The explained variance is the sum of the squared differences between the predicted values and the mean of the dependent variable, while the total variance is the sum of the squared differences between the actual values and the mean of the dependent variable.





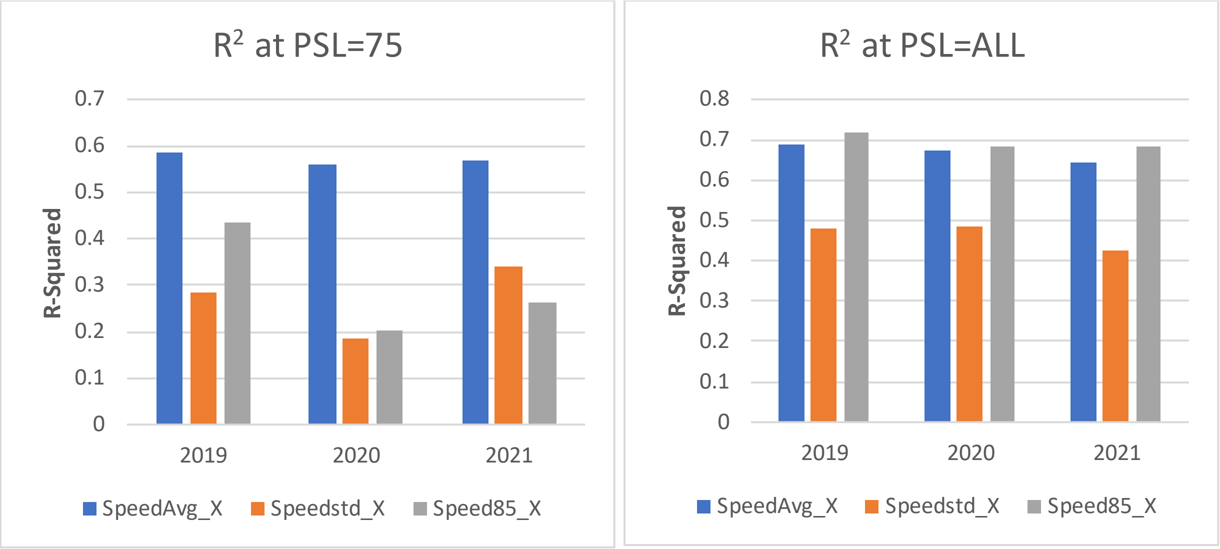


Figure 20. R-squared at different PSL

4.2.5 Violin Plot

A violin plot is a statistical visualization tool that combines the features of a box plot and a kernel density plot. It is used to display the distribution of a continuous variable across different categories or groups. The violin plot consists of a series of kernel density plots, each representing a group or category of the variable being plotted. The violin plot provides a used for a visual representation of the distribution of data across different years allowing for easy comparison between them. It can also reveal any asymmetry or skewness in the distribution that might not be apparent in a traditional box plot. The plot below shows the change in speed during 2020. The violin plot for various PSLs is listed below:

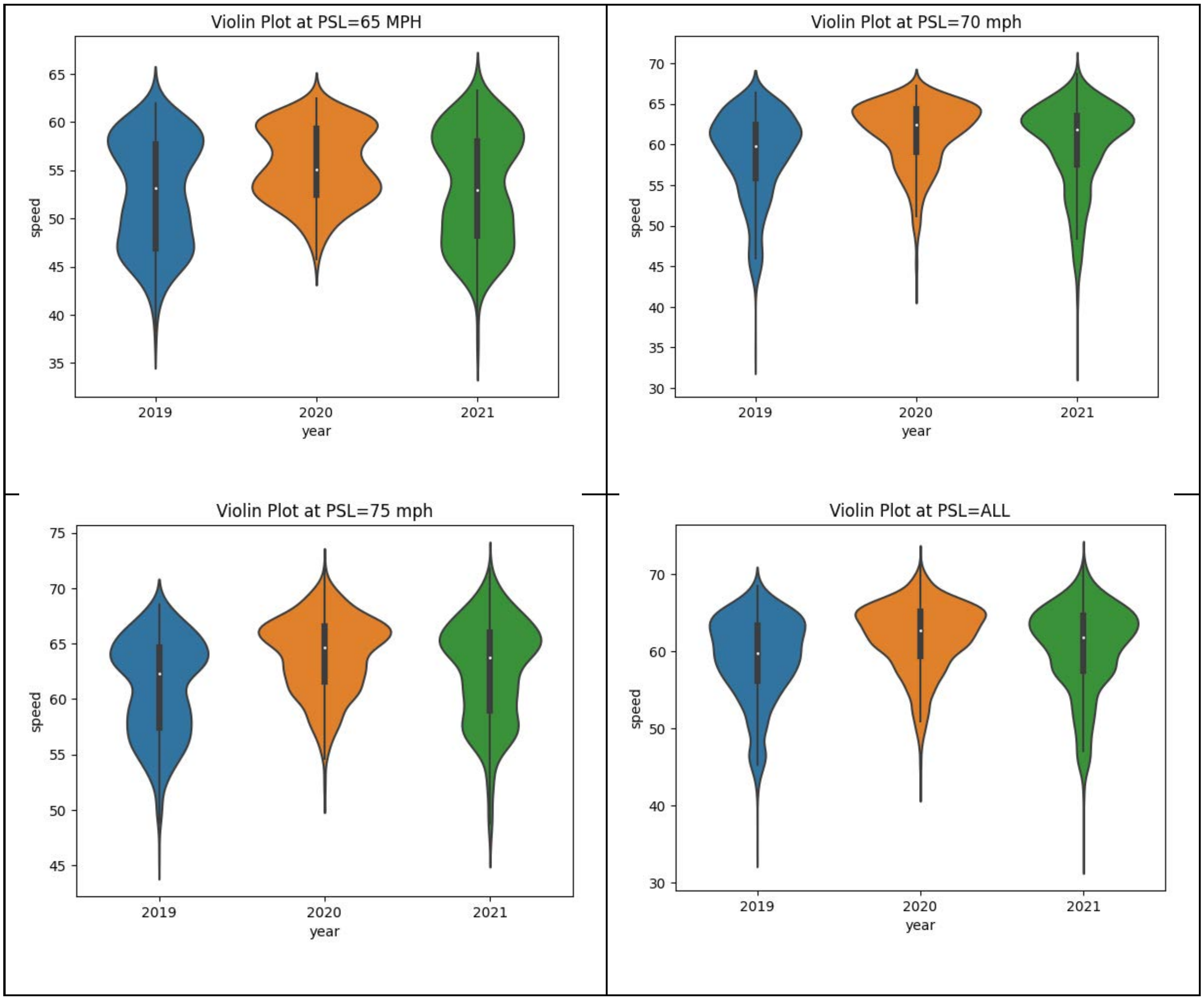


Figure 21. Violin Plot for various PSL for 2019,2020 and 2021

4.2.6 Shapely Plot

Three figures below display partial dependence plots (PDPs) that demonstrate how the independent variables affect mobility measures like SpdAve. The PDPs help to explain the key factors. Among the independent variables, K-factor, len_sec, and all have little impact on mobility measures. However, adt_adj, number_lanes, and median width are associated with mobility measures, with the nature of their association depending on the different clusters of average operating speeds. The SHAP summary pattern of the SpdAve model is shown in Figure for urban freeways. Each explanatory variable has been ranked according to its influence on the model's predictions. A higher level of feature importance implies that the parameter has a significant impact on the measures of average speed (SpdAve).

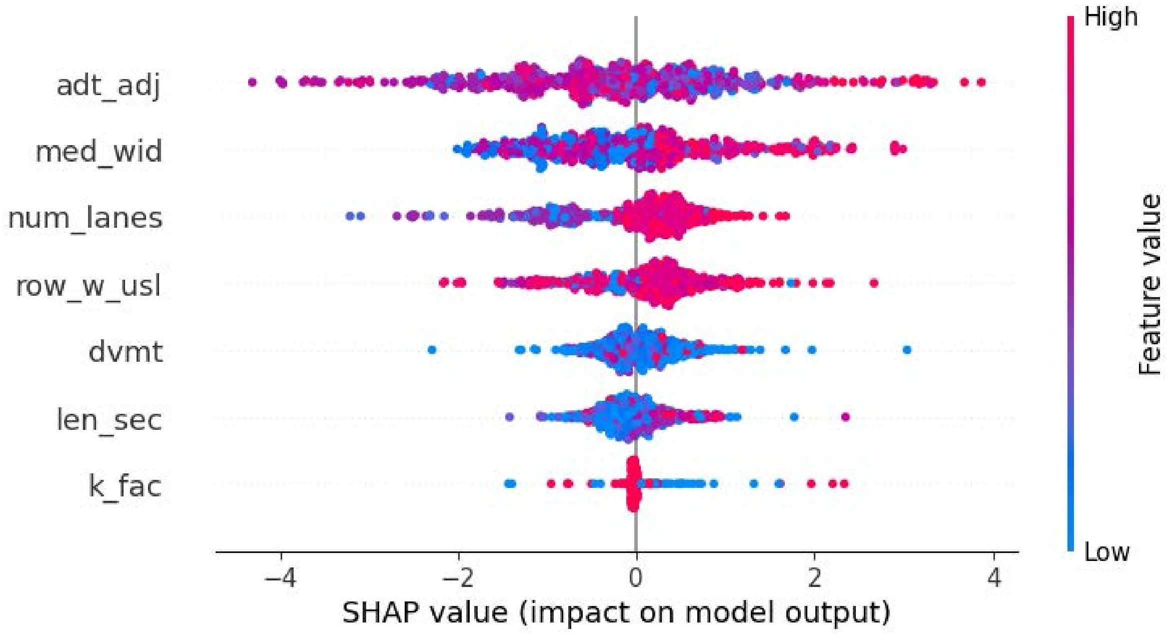


Figure 22. SHAP Values for Year 2019

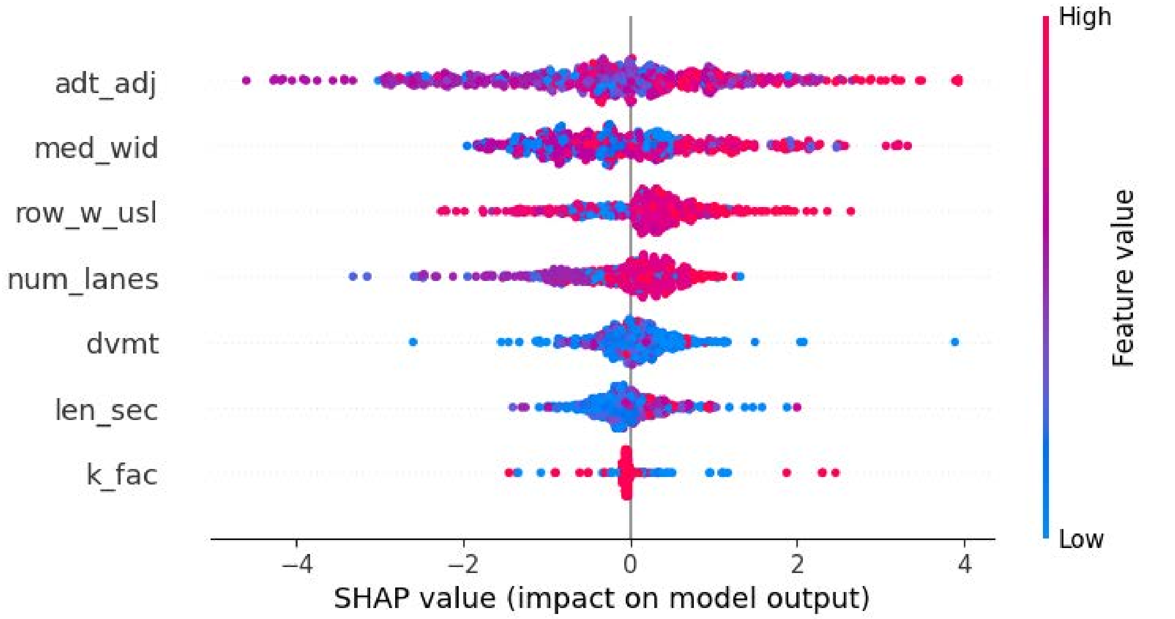


Figure 23. SHAP Values for Year 2020

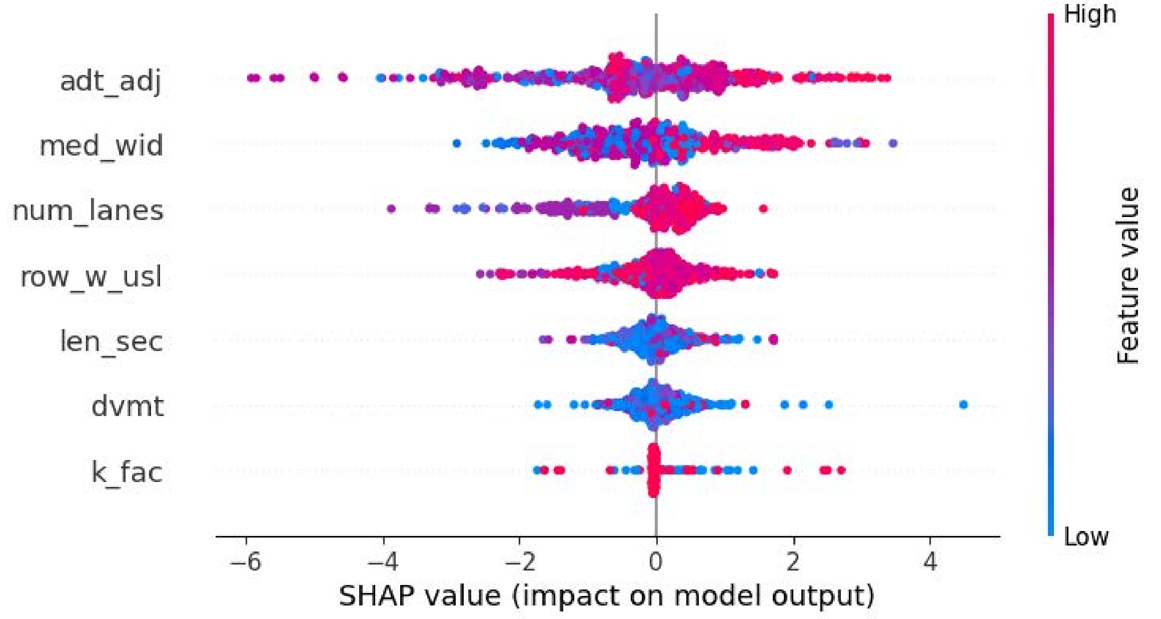


Figure 24. SHAP Values for Year 2021

4.3 Discussions

The outcome from the model and the results are in the acceptable range. The plotted actual speed vs predicted speed at various posted speed limits shows the linear trend during the different years of the study which means the developed model performs well.

CatBoost is capable of detecting complex and non-linear relationships between various dependent and independent variables considered. This means that it identified the traffic patterns based on provided speed which might not be immediately obvious to human analysts, resulting in more accurate traffic flow predictions. And was able to optimize traffic flow in real-time, with high effectiveness. However, there are some limitations associated with using Cat Boost for traffic mobility research. a significant amount of high-quality data is required to train the model effectively. In some stations, the datasets were limited, and the prediction may have been less effective which is also depicted in the errors section. This can be challenging in certain areas where the data sets are limited, resulting in less accurate predictions.

CatBoost modeling is proven to be a powerful tool that has the potential to improve traffic flow in urban areas by creating more efficient and robust modeling and predictions. By utilizing AI-based models to predict traffic mobility and optimize traffic flow, urban areas can reduce travel times and improve safety for all commuters.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

Previous studies have mostly focused on review and trends analyses of COVID-19 on traffic flow. Those studies applied qualitative analyses to study traffic volumes while some used quantitative analyses but checked only traffic volumes to identify the changes in road traffic and to relate them to mobility. The studies didn't consider traffic speed linked to the mobility analyses by considering other variables such as road traffic volume, lane width, and daily miles traveled. In this research, the study was conducted on speed variables through different years with the full consideration of other variables such as number of lanes, annual average daily traffic, median width, length of the sections, vehicle miles traveled, and accident data. A Machine Learning model was used to conduct analysis for model prediction, model accuracy, and effectiveness of all selected variables. CatBoosting was used for data processing and developing of the detailed model. The training and testing of the data were done at 80% and 20%, respectively. Shapely plots were used to illustrate the effectiveness of the parameters in the developed model. Among the independent variables, K-factor, number of lanes, right of way, annual average daily traffic, K-factor, length of the section, vehicle miles traveled, median width, and crash data are associated with higher influence on average operating speeds. The Violin Plots showed the change in speed in 2020 compared to the other two analysis years. The study shows that the COVID-19 pandemic increased average driving speed but did not result in a proportional decrease in crashes. Therefore, changes in driving behavior during the pandemic may have impacted the operating speed of traffic on the freeways of Houston. The results highlighted the importance of understanding the impact of changes in driving behavior during emergency situations and developing effective strategies to promote safe driving habits.

Traffic mobility analyses are conducted by various highway regional and local agencies to better comprehend traffic mobilities and understand the travel trends for planning. It is also crucial for those agencies to understand how mobility patterns could be impacted due to future pandemics similar to COVID-19. The model can be effectively implemented in the future for urban roadway facilities having high traffic volumes and operating speeds if other events and pandemics occur.

Traffic mobility can be influenced by other factors such as weather, construction, flooding, covid cases and work-related activities. Those factors may also have some impacts on traffic mobility, which must be considered in future studies. Furthermore, more optimization during learning steps may also influence the output of the model, which must be investigated through further studies.

REFERENCES

Akioui Sanz, A. A., de Cáceres, A. M., & del Valle, L. Á. (2021). Evolution of mobility during the COVID-19 crisis in the region of Madrid. *Transportation Research Procedia*, *58*, 416–422. https://doi.org/10.1016/j.trpro.2021.11.056

Bian, Z., Zuo, F., Gao, J., Chen, Y., Pavuluri Venkata, S. S. C., Duran Bernardes, S., Ozbay, K., Ban, X. (Jeff), & Wang, J. (2021). Time lag effects of COVID-19 policies on transportation systems: A comparative study of New York City and Seattle. Transportation Research Part A: Policy and Practice, 145, 269–283. https://doi.org/10.1016/j.tra.2021.01.019

Das, S., Le, M., Fitzpatrick, K., & Wu, D. (2022). Did Operating Speeds During COVID-19 Result in More Fatal and Injury Crashes on Urban Freeways? Transportation Research Record: Journal of the Transportation Research Board, 036119812211095. https://doi.org/10.1177/03611981221109597

Deng, Y. (n.d.). PREDICTION OF TRAFFIC MOBILITY BASED ON HISTORICAL DATA AND MACHINE LEARNING APPROACHES.

Kellermann, R., Sivizaca Conde, D., Rößler, D., Kliewer, N., & Dienel, H.-L. (2022). Mobility in pandemic times: Exploring changes and long-term effects of COVID-19 on urban mobility behavior. Transportation Research Interdisciplinary Perspectives, 15, 100668. https://doi.org/10.1016/j.trip.2022.100668

Lu, Y., & Giuliano, G. (2023). Understanding mobility change in response to COVID-19: A Los Angeles case study. Travel Behaviour and Society, 31, 189–201. https://doi.org/10.1016/j.tbs.2022.11.011

Owen, R., Smith, C., Ursachi, G., & Fosdick, T. (n.d.). Analysing the Impact of the COVID19 Lockdown on Vehicle Flow and Speeds in the UK.

Prokhorenkova, L., Gusev, G., Vorobev, A., & Dorogush, A. V. (n.d.). CatBoost: Unbiased boosting with categorical features.

Sana, B., Zhang, X., Castiglione, J., Chen, M., & Erhardt, G. D. (2022). Using Probe-Based Speed Data and Interactive Maps for Long-Term and COVID-Era Congestion Monitoring in San Francisco. Transportation Research Record: Journal of the Transportation Research Board, 2676(6), 48–60. https://doi.org/10.1177/03611981211069961

Das, S. (n.d.). 1 Short-duration Crash Modeling to Understand Impact of Operating Speed on 2 Freeway Crashes during COVID-19.

You, G. (2022). The disturbance of urban mobility in the context of COVID-19 pandemic. Cities, 128, 103821. https://doi.org/10.1016/j.cities.2022.103821