# Understanding The Factors Affecting Safety of E-Scooter and Bicycle Users in Urban Environments: An Injury Severity Analysis Using Machine Learning and Natural Language Processing

A Thesis

by

Pranik Koirala

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

MASTER OF SCIENCE

| | |
|---|---|
| Chair of Committee | Dr. Yunlong Zhang |
| Committee Member | Dr. Ipek Sener |
| Committee Member | Dr. Guni Sharon |

March 2023

Major Subject: Civil Engineering

## TABLE OF CONTENTS

# 1 INTRODUCTION

Streets and roads are part of a community, and a safe and healthy community must value maintaining a safe roads/street for all people of the community and not just those driving vehicles [1]. This includes providing safe road for mixed form of transportation such as micromobility. The Federal Highway Administration broadly defines micromobility as any small, low-speed, human- or electric-powered transportation device, including bicycles, scooters, electric-assist bicycles, electric scooters (e-scooters), and other small, lightweight, wheeled conveyances. The importance of micromobility use is becoming more widely recognized in urban transportation, especially in terms of its ability to decrease emissions, congestion, and noise pollution, as well as its convenience. Moreover, micromobility can save energy and space [2].

E-scooters are becoming an increasingly popular mode of transportation in major cities around the world. E-scooters offer a convenient and environmentally friendly alternative to vehicles for short trips. In fact, it has become the most popular form of shared micromobility in many cities [3]. E-scooters have become popular in large cities around the world, leading to increased sharing of urban space with bicycles, which raises important questions about the safety and about conflict between both modes of transportation. Moreover, with increasing private e-scooter ownership and shared e-scooter programs in major urban centers, legislation has been pushing e-scooters to bike lane or road lane [3]. Understanding the differences and similarities between the crashes, the factors influencing the crashes and the severity of injury to the user can help designers and policy makers make better decisions about how to use urban space in a way that is safe and efficient for all road users.

There are only handful of studies looking into the similarities and differences between e-scooter and bicycle in terms of safety. One such study, by Namiri, et.al. (2021), found that e-scooter and bicycle riders are both prone to serious injuries if intoxicated. The study analyzed the National Electronic Injury Surveillance System (NEISS) data concerning bicycle and e-scooter related emergency cases. It also showed that a higher proportion of e-scooter users are intoxicated, as compared to cyclist, calling for more awareness programs and training [4]. Another study, by Meyer, et.al., (2022), compared injuries and contributing factors for e-scooter and bicycle riders in Germany. The researchers utilized level 1 trauma center data to analyze the injuries and found that the injury severity was higher in e-scooter users than in bicycle users. They also found that helmet use was less common among e-scooter users compared to bicycle users [5].Survey conducted by Useche, et.al, (2022) study found e-scooter users in Spain had a higher rate of perceived risky behaviors compared to bicycle users [6].

A fairly new form of micromobility, the e-scooter's characteristics and interaction with the environment are not well established. Although traditional crash reports are limited in terms of e-scooter data, many studies have utilized alternative sources of data, such as National Electronic Surveillance System (NESS) data, Hospital emergency data, emergency medical service telephone records, patient interview surveys [7]. Most descriptive studies on e-scooter usage and safety using NESS data showed that younger populations were more involved in crashes, and a positive relation between e-scooter crashes and intoxication. These studies also found that about 10-20% of e-scooter crashes involved motor vehicles and most motor vehicle involved crashes causes more severe injuries [4] [8] [9].

In a recent study done on e-scooters in Austin, Texas, English, et.al. (2020), found that about 10% of the e-scooter related crashes involved a moving vehicle, and the majority involved a white male in his 30s falling off the e-scooter in an intoxicated state [10]. A 2022 study in Austin also found that males younger than 34 years were more prone to crashes [11]. Studies in Washington, DC using hospital e-scooter related crash data and patient interviews showed that around 10-15% of e-scooter crashes involved moving vehicle. The characteristics and behaviors of e-scooter riders in Washington, DC, do not follow the characteristics and behaviors from other larger cities [12] [13]. Similarly, in Copenhagen the demographic of e-scooter user is mostly teenagers [14]. This study focuses on understanding the factors affecting safety of e-scooters and bicycle users in just one city.

To the author's knowledge, there is a lack of e-scooter injury severity studies in present literature. It is because of this reason; this study has evaluated bicycle related safety studies and its methodologies. Many studies have used traditional statistical methods such as logistic regression, proportional odds models, partial odds models and other models in micro mobility safety analysis [15], [16], [17] , whereas others have also made use of machine learning models such as Random Forest (RF), XGboost, Support Vector Machine (SVM), etc., [18], [19], [20] in micro mobility analysis. Although these studies demonstrate the supremacy of machine learning models in classification accuracy, some literature still found that traditional models like logistic regression perform better than machine learning models [21].

As severely injured or fatally injured cases are comparatively rare, the crash severity data suffers from class imbalance. The problem with class imbalance data is that standard classification algorithms tend to be biased towards the majority classes [22]. Recent studies have identified different techniques to treat imbalanced crash severity data, including over-sampling, under-sampling, and hybrid methods [23] [24] [25]. This research also aims to provide insight in terms of using different sampling methodology to overcome the problem of data imbalance as well as compare traditional statistical model and tree-based machine learning models for modeling e-scooter and bicycle injury severity in addition to comparing e-scooter and bicycle injury severity contributing factors.

# 2 LITERATURE REVIEW

This literature review section is divided into three parts, the first part investigates the studies looking into the similarities and differences between e-scooters and bicycle. The second section reviews papers using machine learning in injury severity analysis in different mode of transportation specially bicycle. Finally, the third part deals with use of topic modeling and natural language processing in transportation related research.

## 2.1 E-scooter and bicycle

Several studies have been conducted to examine various aspects of e-scooters and bicycles. This section will discuss some of the studies that have investigated the similarities and differences between the two modes of transportation. One of these studies took a sample of 2000 e-scooters in the campus of Virginia Tech and collected trajectory data via GPS devices. The researchers implemented recursive logit route choice model to identify travel characteristics of e-scooter users. They found e-scooter users travel on bikeways (59%), multi-use paths (29%), tertiary roads (15%) and the travel distance does not affect the travel. Moreover, slope is also not a determinant for e-scooter route choice [26]. In addition to e-scooters competing for urban space with bicycle, e-scooters are also replacing some types of bicycles like e-bike. Researchers studied the impact of e-scooter sharing on bike sharing in Chicago using publicly available e-scooter data from 2019 and bike sharing data from Divvy bikes. The study uses difference-in-difference (DID) method to analyze before and after effect of introduction of e-scooter sharing program to bicycle sharing program. They found that bike sharing decreased after the introduction of e-scooters but was not affected during the peak hours [27].

The purpose of use of e-scooter and bicycle is not the same according to some studies. Bielinski, et. al. in their 2020 study compared the users of e-scooter sharing user and e-bike sharing users.

The study was based on survey done in Tricity metropolitan area in northern Poland. They found that e-scooters are mostly used for leisure purposes whereas people use e-bikes for mostly first and last mile transport and getting to point of interest [28]. Similar research also found the same finding. The researchers in the study conducted a survey to understand the differences and similarities between e-scooter sharing and bike sharing users in the city of Trinity in northern Poland. The main finding of the study was that e-bikes are usually used as a first and last mile transport and to commute directly to various places of interest, whereas e-scooters are more often used for leisure rides [29].

Similarly, Stray, A, et. al., in their study characteristics of electric scooter and bicycle injuries after introduction of electric scooter rentals in Oslo, Norway compared e-scooters and bicycles. The study used 3191 patient records (850 e-scooter riders' and 2341 bicyclists) and observed the riders characteristics such as age distribution, helmet use, day of week and time are different from that of bicyclists [30]. Likewise, a study used emergency department data for e-scooter and bicycle related accident patient from the Oslo University Hospital between January 1, 2019, and March 31, 2020. The research compared the characteristics to both the mode of transportations and found that e-scooter injuries were associated with younger age and a higher rate of intoxication compared to bicycle injuries. E-scooter injuries were more likely to occur on weekends and during evening or nighttime hours, while bicycle injuries were more likely to occur during weekdays and daytime. Annual incidents of injury were higher for bicycle users, but e-scooter riders were less likely to use a helmet [31].

Study done in Spain used road user's risk perception in the form of rating for e-scooter and bicycle users. The data was collected for 950 non-cyclist and non-e-scooter users in Spain. In the survey e-scooter users were found to have more dangerous and risky behavior on the road compared to cyclist [32]. Additionally, many studies suggest e-scooter have a risky behavior such as no helmet use and intoxicated riding. A study in Finland compared facial injuries in e-scooter or bicycle related accidents between January 2019 and October 2020. Higher proportion of e-scooter users had a facial injury compared to proportion of bicycle users with facial injury. Alcohol involvement was significantly higher in e-scooter patients (88.9%) than in bicycle patients (31.5%) with a p<0.001. The researchers also conducted a multivariate analysis and concluded that e-scooter incidents were 18times more likely to occur under the influence of alcohol [33]. Similar study in Germany also concluded that e-scooter riders have an increased risk of facial injury compared to cyclists [34]. The study used e-scooter and bicycle patient data from the Department of Oral and Maxillofacial Surgery and Department of Trauma Surgery of the Technical University of Munich, Germany. The data showed that no e-scooter patient was wearing a helmet and the number of riders under influence was also higher in e-scooter compared to cyclists.

Another study investigating the risky behavior of e-scooters, bicycle and e-bike found e-scooter users to be more likely involved in a secondary task like using headphones while riding [35]. The observational study was performed in Germany where researchers collected direct data from the streets of Braunschweig. They also found that e-scooter users were slightly more likely to violate one or more traffic rules compared to e-bike or bicycle riders.

Although early research has found that e-scooter users show a risky behavior while driving [11], [12], [13], [29], some studies are investigating this issue more closely. In such study's recent finding suggest that e-scooter maneuvering is safer than bicyclist in lower speeds [36]. This study used field collected data from bicycle and e-scooters in Sweden to compare the two modes of transportation with respect to safety. The researchers developed a framework to analyze the data from the different sensors and developed a model to generalize the results. The findings from the analysis showed that e-scooters might be more easily maneuverable and comfortable than bicycle. The sensors were used to collect data on braking, steering, speed as performance indicators. A survey was also conducted comparing e-scooter and bicycle use. It found that users found it less comfortable to brake on the e-scooters than bicycle, while they felt more comfortable for the steering. At higher speed bicycle was easier to maneuver but at lower speed of 10kmph user felt e-scooter to be easier.

Another study found that e-scooter related facial injury is not more than bicyclist but comparable to bicyclist related facial injury [37]. The study used emergency department' patients record data from Liverpool for 2020 and 2021 and analyzed the injury sustained by e-scooter users. The study was designed to study the effects of legislative change in e-scooter related policy in 2020 affecting e-scooter ridership. It found that the increase in e-scooter usage also increase in e-scooter related musculoskeletal injuries, but the rate was comparable to bicyclist. Similarly, A comparative study [38] was conducted in Washington, DC between e-scooter and bicycle using emergency department data. The data related to e-scooter injury was from 2019 and that of bicycle user was from 2015-2017. It was observed that e-scooter crash less frequently involved moving vehicles (13.1% vs 37.7%) or occur on roads (24.5% vs 50.7%). Head injury rates were similar but e-scooter riders more often experienced concussion with loss of consciousness (4% vs 6%) and are far less likely to were helmets (2% vs 66%). The researchers conclude that the type of injury sustained by the two modes of transportation are different and the frequency is greater for e-scooters but the rate may decrease as people gain experience using the new mode of transportation.

A recent study conducted in the city of Brisbane, Australia, where e-scooters are only permitted on footpaths alongside pedestrians, delved deeper into this issue by comparing the helmet usage of both shared and private e-scooters and bicycles. The study collected data directly on-site and the findings reveal that, regardless of whether they are riding e-scooters or bicycles, individuals are less likely to use helmets when utilizing shared vehicles as opposed to privately-owned ones [39].

In conclusion, this section of the literature review has investigated studies that have compared e-scooters and bicycles. Studies have found that e-scooters and bicycles have different characteristics such as age distribution, helmet use, and time of use. E-scooters were found to be associated with younger age and higher rate of intoxication compared to bicycle injuries, and e-scooters injuries were more likely to occur on weekends and during evening or nighttime hours. Additionally, research showed that e-scooters are mostly used for leisure purposes whereas e-bikes are mostly used for first and last mile transport and getting to point of interest. The research also showed that E-scooter riders were less likely to use a helmet. Overall, this literature review provides insight into the characteristics and usage patterns of e-scooters and bicycles and the comparison between them. Much research shows that e-scooters can be a convenient and safe mode of transportation with responsible user especially private e-scooter users. One research even pointed e-scooter to be easier to maneuver in lower speeds. Since the findings are not definitive and clear there is a need for further research to help policy makers make informed decisions on e-scooters and bicycles.

## 2.2 Machine learning in injury severity analysis

Many transportation safety researchers are highly interested in modeling the severity of injuries that occur in crashes. By predicting the expected severity of a crash, it can help identify the factors that contribute to injury severity, which can help to reduce the severity of crashes and improve road safety. Contributing factor analysis can also help identify infrastructure related characteristics between different mode of transportation. Additionally, such prediction can also aid hospitals in providing appropriate medical care quickly. This section of the literature review investigates the use of different statistical methods and machine learning models in injury severity analysis.

In a comprehensive literature review of studies dealing with crash injury severity Savolainen et al. presented different statistical methodology used in injury severity as well as the nature of injury severity data [41]. According to the review injury severity data consists of many characteristics such as multicollinearity, within crash correlation and unobserved heterogeneity. These issues in the injury severity data can affect in model prediction and sometimes contradicts statistical models assumptions [42]. Similarly, Santos et al. in their review of machine learning algorithms for crash injury severity prediction reviewed 56 studies and concluded that random forest, support vector machine and decision tree performed the best in most pf the studies reviewed. More specifically, random forest preformed best in 70% of the studies when it was applied [43].

Another recent paper compared different machine learning and traditional statistical models in injury severity of motor vehicle crashes in rural highway. The study used XGBoost, logistic regression, random forest and decision tree to model the data and found XGBoost to outperform all other models [44].

[40] A study from Sweden, the researchers present a review of current literature and proposes a preliminary framework for developing a Level of Service (LOS) for e-scooters. The study highlights that need for more studies related to infrastructure and e-scooters as e-scooter are

becoming more and more popular. The researchers also suggest more comparative studies between different modes of transportation and e-scooter for better understanding of e-scooters impact. Considering all these research’s findings and recommendation this thesis will try to understand the differences and similarities of e-scooter crashes and bicycle crashes using different techniques.  This thesis intends to provide additional information regarding e-scooter and bicycle to policy makers and planners for more informed decision making and design.

## 2.3 Topic modeling and natural language processing

Few research has made use of natural language processing and topic modeling in analyzing crashes in transportation research. Most studies that have done it use it in analyzing the description of the crash scene and extract more information and unobserved trends.

Das, et. al. in their 2021 study explores the use of text mining and topic modeling to understand and extract insights from detailed crash narratives available in the Motorcycle Crash Causation Study (MCCS) sponsored by NHTSA. The dataset contained 351 injury crashes. The study used Latent Dirichlet allocation (LDA) method for topic modeling of the crash narrative. The language analysis was done on two separate clusters of data, namely fatality related dataset and non-fatality related dataset. Different topic clusters were identified with these datasets whose composition was studied and compared. Highly representative keywords or risk factors in fatal crash reports were compared for both the clusters (fatality related and non-fatality related) (Das, Dutta, & Tsapakis, 2021).

Another study which used LDA in transportation related, more specifically e-scooter related study is Aman, et. al., 2021. The study utilized LDA model to analyze over 12,000 rider0-generated reviews to understand the satisfaction factor and concerns of users [46]. Similarly, Kwayu, 2020, analyzed crash reports which had 'failed to yield' or 'disregarded traffic control' in hazardous action in signal-controlled intersection. The purpose of the research was to discern and difference between these two types of hazardous action as it is sometimes used interchangeably and can be confusing. They successfully inferred in what cases police reported a 'disregarded traffic control' or 'fail to yield' [47]. Likewise, Zhang et. al., 2016, claims that the information on hazardous actions in a crash report is not always accurate and the use of the descriptive narrative to identify hazardous actions is warranted. The experiment focused discerning if hazardous actions was 'none' or 'hazardous' using NLP and ML [48] .

# 3 DATA

## 3.1 CRIS

The data for the study is collected from Texas Department of Transportation (TxDOT) Crash Record Information System (CRIS) database. The database contains information collected by Texas police officers throughout the state and is maintained by TxDOT. The study also had access to police reports which was used for text mining to extract e-scooter relate data. 2195 crash data from 2018 to 2021 was available for analysis, of which 153 data points were identified to be involving e-scooter crashes. 819 bicycle related crash data points were filtered from the dataset. A distribution of variables selected by PhiK and domain knowledge is tabulated below.

Table 1 CRIS variable summary

| SN | Name | Column name | Definition/Detail | Frequency | | |
|---|---|---|---|---|---|---|
| | | | | E-scooter | Bicycle | Combined |
| 1 | Alcohol | Prsn_Alc_Rslt_ID | Alcohol content in e-scooter/bicycle user | | | |
| | | | 1=alcohol | 2 | 0 | 2 |
| | | | 0=no-alcohol | 151 | 819 | 970 |
| 2 | Roadway system | Rpt_Rdwy_Sys_ID | Road system on which crash occurred | | | |
| | | | 1=Interstate | 10 | 25 | 35 |
| | | | 2=US Highway | 1 | 19 | 20 |
| | | | 3=State Highway | 1 | 22 | 23 |
| | | | 4=Farm to Market | 0 | 3 | 3 |
| | | | 5=Local Road/street | 141 | 771 | 912 |
| 3 | Roadway | Road_Algn_ID | 1= straight, level | 150 | 694 | 844 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | alignment | | 2= straight, grade | 5 | 136 | 141 |
| | | | 3 = straight, hillcrest | 4 | 43 | 47 |
| | | | 4 = curve, level | 5 | 21 | 26 |
| | | | 5 = curve, grade | 0 | 21 | 21 |
| | | | 6 = curve, hillcrest | 1 | 5 | 6 |
| 3 | (Manner of collision) | FHE_Collsn_ID | 1=one motor vehicle – going straight | 61 | 446 | 507 |
| | | | 2= one motor vehicle – turning right | 30 | 177 | 207 |
| | | | 3= one motor vehicle – turning left | 15 | 204 | 219 |
| | | | 4= on motor vehicle - backing | 0 | 4 | 4 |
| | | | 5=one motor vehicle - other | 0 | 6 | 6 |
| | | | 10=Both Angle | 26 | 3 | 29 |
| | | | 20=Both Same direction | 8 | 0 | 8 |
| | | | 30=Both Opposite direction | 13 | 0 | 13 |
| 4 | Traffic Control Device | Traffic_Cntl_ID | 1=None | 24 | 146 | 170 |
| | | | 5=Signal light | 39 | 177 | 216 |
| | | | 20=Marked line | 34 | 211 | 245 |
| | | | 8=stop sign | 16 | 103 | 119 |
| | | | 15=crosswalk | 15 | 45 | 60 |
| | | | 21=signal light with red light running camera | 8 | 14 | 22 |
| | | | 11=center divider | 8 | 51 | 59 |
| | | | 16=bike lane | 6 | 54 | 60 |
| | | | 17=other | 2 | 20 | 22 |
| | | | 9=yield sign | 1 | 9 | 10 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  |  | 10=warning sign | 0 | 6 | 6 |
| 5 | Roadway Type | Road_Type_ID | 0=other type | 0 | 2 | 2 |
|  |  |  | 1=2lane,2way | 0 | 3 | 3 |
|  |  |  | 2=4 or more lane, divided | 20 | 102 | 122 |
|  |  |  | 3=4 or more lane, undivided | 133 | 712 | 845 |
| 6 | Ethnicity | Prsn_Ethnicity_ID | 0=Unknown | 5 | 17 | 22 |
|  |  |  | 1=White | 88 | 553 | 641 |
|  |  |  | 2=Hispanic | 25 | 141 | 166 |
|  |  |  | 3=Black | 22 | 58 | 80 |
|  |  |  | 4=Asian | 6 | 29 | 35 |
|  |  |  | 5=Others | 3 | 7 | 10 |
| 7 | Light Condition | Light_Cond_ID | 0=Unknown | 1 | 6 | 7 |
|  |  |  | 1=Daylight | 89 | 616 | 705 |
|  |  |  | 2=Dawn | 0 | 5 | 5 |
|  |  |  | 3=Dark, not lighted | 8 | 45 | 53 |
|  |  |  | 4=Dark, lighted | 54 | 152 | 206 |
|  |  |  | 5=Dusk | 1 | 14 | 15 |
|  |  |  | 6=Dark, unknown lighting | 0 | 2 | 2 |
| 8 | Weather Condition | Wthr_Cond_ID | 0=Unknown | 2 | 6 | 8 |
|  |  |  | 2=Rain | 8 | 24 | 32 |
|  |  |  | 5=Fog | 0 | 1 | 1 |
|  |  |  | 8=Other | 0 | 1 | 1 |
|  |  |  | 11=Clear | 126 | 667 | 793 |
|  |  |  | 12=Cloudy | 17 | 141 | 158 |
|  |  |  | 4=Winter | 0 |  |  |
| 9 | Weekend binary | weekend_binary | 0=weekday | 72 | 522 | 594 |
|  |  |  | 1=weekend | 81 | 318 | 399 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 10 | Helmet use | Prsn_Helmet_ID | 0=unknown | 5 | 102 | 107 |
| | | | 1=worn | 3 | 289 | 292 |
| | | | 2=not worn | 144 | 438 | 582 |
| 11 | Intersection | Intrsct_Relat_ID | 0=No | 49 | 324 | 373 |
| | | | 1=Yes | 104 | 516 | 620 |
| 12 | Vehicle Body Style | Veh_Body_Styl_ID_1 | 0=Unknown | 6 | 50 | 56 |
| | | | 1=Passenger car | 95 | 472 | 567 |
| | | | 2=SUV | 27 | 157 | 184 |
| | | | 3=Pickup | 18 | 106 | 124 |
| | | | 4=Van | 3 | 25 | 28 |
| | | | 5=Heavy vehicle (truck, bus) | 4 | 30 | 34 |
| 13 | Age of rider | Prsn_Age | Age of micromobility / bicycle rider | | | |
| 14 | Age of driver | Prsn_Age_unit1 | Age of motor vehicle driver | | | |

The injury severity data available in the CRIS database is of ordinal nature. The Injury classes are more severe as the level increases. Usually around 2-3% fatal injuries are observed related to e-scooter and bicycle involving crashes. The CRIS data has five level of injury severity which is aggregated into severe and non-serious injuries. There are 21 severe and 132 non-severe e-scooter related injury severity data points and 103 severe injury cases and 737 non-severe cases in case of bicycle crashes.

## 3.2 Demographic and Socioeconomic data

The American Community Survey's five-year average demographic and Socioeconomic data is used in this study. The American Community Survey is a demographics survey program conducted by the U.S. Census Bureau. The data is downloaded from social explorer [45] as it provides a cleaner format of the dataset. The different demographic and socioeconomic variables used in the study is mentioned in Table 2.

## 3.3 Built Environment data

### 3.3.1 City of Austin

City of Austin’s open data portal provides access to several built environment datasets. This study has utilized the dataset available in the open data portal for analysis. Among other datasets, the study makes use of the Land Use dataset. The Land Use dataset is maintained by the Housing and Planning Department of City of Austin.

The land inventory is based on several sources. The polygon geography is taken from appraisal district parcel layers merged and the land use is calculated by classifying land according to a coding system that reflects the primary improvements (building and structure) on each parcel [46].

The Land Use data is used to calculate the Land Use Mix Index or land use entropy for each count point within 1 mile around the point or location. The idea behind mixing urban land uses is that a variety of applications or activities that are near together may serve complementary purposes, each of which increases the usefulness of its neighbors [47]. To measure the land use mix, Entropy Index is used in this study. The Entropy Index is an adaptation of Shanon’s Entropy. The equation of land use mix index (ENT) is shown as follows.

$$ENT = -\frac{\left[\sum P^{i}\ln\left(\mathrm{P}^{\mathrm{i}}\right)\right]}{\ln(\mathrm{k})}$$

Where $P^i$ is the percentage of each type of land use within the buffered area around the count point $i$; $k$ is the number of land-use types within the buffer zone $i$, and $k \geq 2$.

The bike facility dataset made available in the City of Austin's open data portal by The Austin Transportation Department. The dataset has rich information regarding the presence of bike lane and the type of bike lane such as buffered bike lane, protected bike lane, trail lane, or no bike lane. Similarly, Urban trail dataset prepared by Urban Trails Program, Department Neighborhood Connectivity Division, City of Austin Public Works is also utilized in the study. The dataset describes a specific type of urban infrastructure for bicycle and pedestrians mostly used for recreational purposes mostly within park or similar area.

### 3.3.2 Smart Location data

The U.S. General Services Administration (GSA) and U.S. Environmental Protection Agency (EPA) Smart Location Database (SLD) satisfies the rising demand for data products and technologies that systematically assess the location efficiency of distinct locations. For each Census block group (CBG) in the US, the SLD provides a summary of a number of demographics, employment, and built environment factors [48]. The smart location database's variables are highly correlated with each other as seen in Figure 1. Since network related variables are calculated using roadway network related dataset from TxDOT, only transit related (D4A) and walkability features are used in model training.

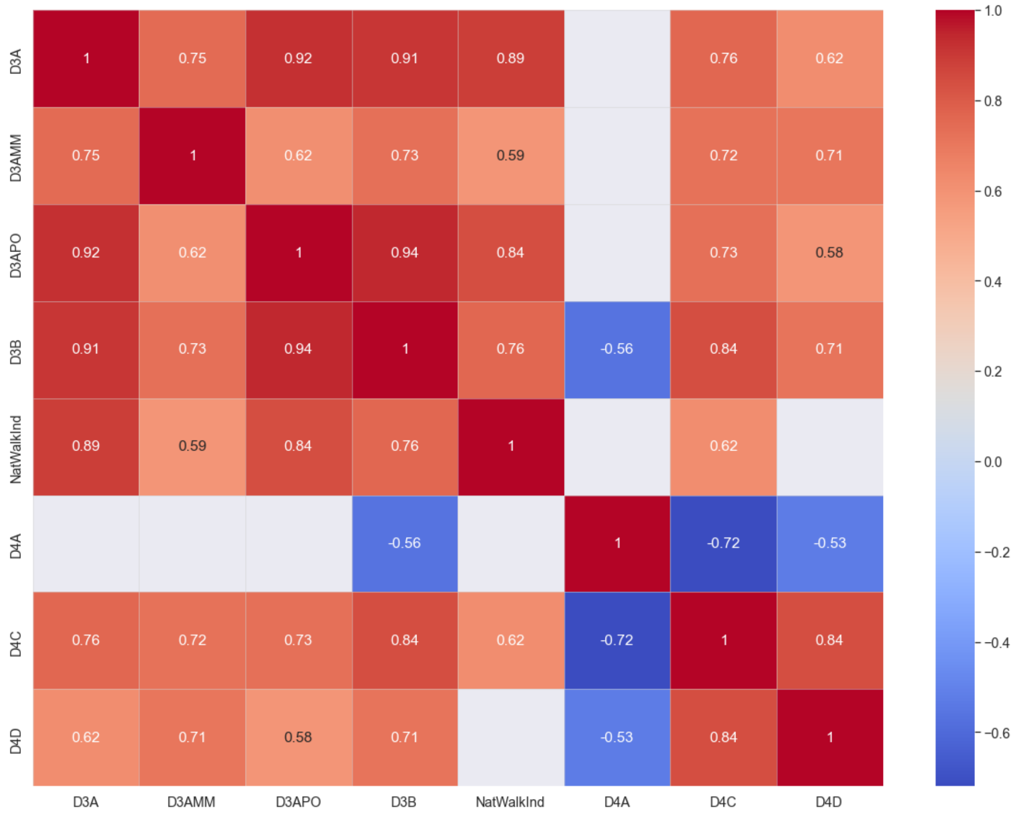


Figure 1 Pearson correlation coefficient more than |0.5| between smart location variables

### 3.3.3 Roadway inventory

The roadway related data is collected from TxDoT's roadway inventory datasets. TxDoT annually publishes the dataset which contains GIS line work, and all roadway inventory attributes [3]. The variables available and used for the study from this dataset is mentioned in the variable table below from SN 22 to SN 30. The variables in Table 2 is calculated using the data available in the roadway inventory dataset with respect to 1 mile radius around count station location. Density of roadway network within 1 mile of buffer around a location is calculated by dividing total length of network (ft.) by the area of buffer (sq.ft.).

Different variables from the different data sources used in the study and mentioned above are mentioned in Table 2.

Table 2 Description of variables from different data sources

| | Name | Column name | Description |
|---|---|---|---|
| Demographic | | | |
| 1 | Total Population | TotPop | The attribute indicates the total population within 1 mile radius of the station |
| 2 | Total Male Population | TotPopMale | The attribute indicates the total male population within 1 mile radius of the station |
| 3 | Total Female Polulation | TotPopFemale | The attribute indicates the total female population within 1 mile radius of the station |
| 4 | Male population 5 to 14 years of age | Male_5_14 | The attribute indicates the total male population between the age 5 and 14 within 1 mile radius of the station |
| 5 | Male population 15 to 24 years of age | Male_15_24 | The attribute indicates the total male population between the age 15 and 24 within 1 mile radius of the station |
| 6 | Male population 25 to 34 years of age | Male_25_34 | The attribute indicates the total male population between the age 25 and 34 within 1 mile radius of the station |
| 7 | Male population 35 to 54 years of age | Male_35_54 | The attribute indicates the total male population between the age 35 and 54 within 1 mile radius of the station |
| 8 | Male population 55 to | Male_55_64 | The attribute indicates the total male |

| | | | |
|---|---|---|---|
| | 64 years of age | | population between the age 55 and 64 within 1 mile radius of the station |
| 9 | Male population 65 and over | Male_65_over | The attribute indicates the total male population between the age 65 and more within 1 mile radius of the station |
| 10 | Total White population | Pop_Race_White | The attribute indicates the total population identifying as White race within 1 mile radius of the station |
| 11 | Total Black population | Pop_Race_Black | The attribute indicates the total population identifying as Black race within 1 mile radius of the station |
| 12 | Total Asian population | Pop_Race_Asian | The attribute indicates the total population identifying as Asian race within 1 mile radius of the station |
| 13 | Total native American population | Pop_Race_Native | The attribute indicates the total population identifying as Native American race within 1 mile radius of the station |
| 14 | Total population – two or more race | Pop_Race_Two | The attribute indicates the total population identifying as two or more race within 1 mile radius of the station |
| 15 | Number of Household | House | Number of households occupied by owners |
| 16 | Bachelors or higher education | Edu_Bachelor_ more | The attribute indicates the population 25 year and over with a bachelor’s degree or higher degree |

| | | | |
|---|---|---|---|
| 17 | Households with less than $60k | HHI_less_60 | The attribute indicates the number of households with household income less than $60k within 1 mile of the station |
| 18 | Households with income between $60k - $100k | HHI_60_100 | The attribute indicates the number of households with household income between $60k and $100k within 1 mile of the station |
| 19 | Households with income between $100k - $150k | HHI_100_150 | The attribute indicates the number of households with household income between $100k and $150k within 1 mile of the station |
| 20 | Households with income more than $150k | HHI_150_more | The attribute indicates the number of households with household income more than $150k within 1 mile of the station |
| 21 | Per capita income | Income | The attribute indicates per capita income (adjusted for the year mentioned in the year column) |
| Built Environment | | | |
| 22 | Urban Interstate/Freeway density within 1 mile radius | UI_1mile | The attribute indicates urban Interstate/Freeway highway density (length/area) within 1 mile radius |
| 23 | Urban arterial road density | UA_1mile | The attribute indicates Urban Arterial highway density (length/area) within 1 mile |

| | | | |
|---|---|---|---|
| | | | radius |
| 24 | Urban connector road density | UC_1mile | The attribute indicates Urban Collector highway density (length/area) within 1 mile radius |
| 25 | Urban local road density | UL_1mile | The attribute indicates Urban Local highway density (length/area) within 1 mile radius |
| 26 | Rural Interstate/Freeway density within 1 mile radius | RI_1mile | The attribute indicates Rural Interstate/Freeway highway density (length/area) within 1 mile radius |
| 27 | Rural arterial road density | RA_1mile | The attribute indicates Rural Arterial highway density (length/area) within 1 mile radius |
| 28 | Rural connector road density | RC_1mile | The attribute indicates Rural Collector highway density (length/area) within 1 mile radius |
| 29 | Rural local road density | RL_1mile | The attribute indicates Rural Local highway density (length/area) within 1 mile radius |
| 30 | Maximum speed within 1 mile radius | MaxSpd_1mile | The attribute indicates the maximum speed in any roadway within 1 mile radius buffer of the location |
| 31 | Most common land use type | Land_use | The most common land use type within 1 mile of the crash location |
| 32 | Land use entropy | Lu_entropy | Land use entropy within 1 mile of the crash |

| | | | |
|---|---|---|---|
| | | | location |
| 33 | Bicycle facility type | Bike_fac12 | Presence of Bicycle facility and its type<br>0 = no bike lane<br>1 = bike lane<br>2 = buffered bike lane<br>3 = protected bike lane<br>4 = trail lane |
| 34 | Unprotected land | Ac_Unpr | Total unprotected land (i.e., not park, not national park, etc.) |
| 35 | Total road network | D3A | Total road network density |
| 36 | Multimodal link network density | D3AMM | Network density in terms of facility miles of multi-modal links per square mile |
| 37 | Pedestrian oriented link network density | D3APO | Network density in terms of facility miles of pedestrian oriented links per square mile |
| 38 | Street intersection density | D3B | Street intersection density (weighted, auto-oriented intersections eliminated) |
| 39 | Walkability index | NatWalkInd | Walkability index comprised of weighted sum of the ranked values of other smart location variables such as D2A. |
| 40 | Distance to transit stop | D4A | Distance from the population-weighted centroid to nearest transit stop (meters) |
| 41 | Transit frequency per 0.25 miles | D4C | Aggregate frequency of transit service within 0.25 miles of CBG boundary per hour during evening peak period, 2020 GTFS |

| | | | |
|---|---|---|---|
| 42 | Transit frequency per mile | D4D | Aggregate frequency of transit service [D4C] per square mile |

# 4 METHOD

The bicycle and e-scooter data used in this study are preprocessed in Jupyter Notebook using python programming language. In addition, it is also used for model building as well as handling spatial data. PhiK correlation coefficient is used as variable selection method to select important variables for model training. The cleaned datasets are resampled using different techniques as further discussed in the paper. The resampled data are then evaluated based on an RF model using repeated stratified K-fold cross-validation technique. The best-performing sample is then split into train data and validation data before training the models for classification. The methodology steps implemented in this study are shown in Figure 1. The individual steps of the study are discussed in detail after the figure.

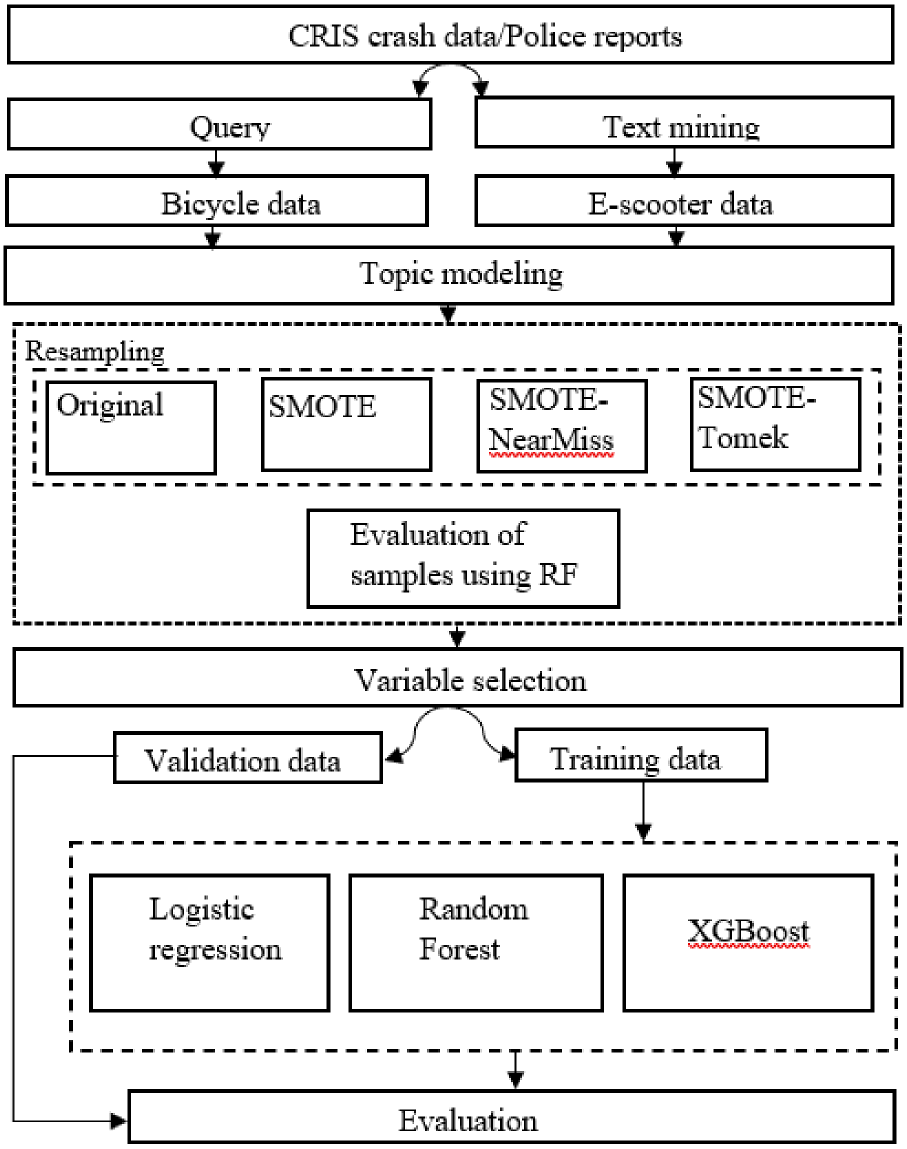


Figure 2 Flow chart of study method for identifying contributing factors.

## 4.1 Text Mining

There have been many advances in text mining and text source interpretation in transportation safety analyses. Traditionally, crash reports contain highly detailed information including linguistic narratives with details about crash events and contexts. Many studies have compared different methods of text mining in transportation context [49]. There is no specific code for e-scooter in the crash dataset, so text mining is recognized as a suitable tool for identifying e-scooter involved crashes in police reports. This study utilizes an optical character recognition (OCR) tool, Tesseract OCR Engine, for reading the descriptive section of police reports. All the recognized words are then separated and checked for predetermined keywords related to e-scooters. The keywords are determined using manual analysis of a fraction of crash reports involving e-scooters.

## 4.2 Topic modeling

Natural language processing technique is used to generate new features and understand the descriptive section of the police reports. As the police report's description section is digitized for identifying e-scooter related cases, the digitized descriptive section is used to cluster the incident cases based on the use of language in the descriptive section. All police report's descriptive section is clustered into some categories which can be analyzed to investigate if different language is used for e-scooter and bicycle related accidents.

Firstly, the digitized text is split into pieces based on space in the process known as Tokenization. The Tokenized text is then cleaned using the technique called Stopword removal. In which common English words with no meaning like 'the', 'in', etc. are removed. After cleaning, the remaining words are reduced to the base form of the word, for example 'stopping' is reduced to 'stop' in the process called lemmatization. Finally, the words are trained using two models to identify clusters. Again, Using the N-gram technique adjacent words are also combined to give the model more information.  Two models, Latent Dirichlet Allocation (LDA) and Bidirectional Encoder Representation from Transformers (BERT) is used to calculate the clusters among the reports.

LDA is a statistical natural language model which helps separate text document into different topic or classes by analyzing frequency of word counts. Likewise, BERT is a transformer-based machine learning technique for natural language processing.

### 4.3 Resampling

The categorical nature of the crash injury severity data and disproportional distribution of observation in each category makes the dataset suffer from class imbalance problem. Classes with smaller sample size are called minority class. In this study, crash severity categories - fatal injury and severe injury were combined as severe injury severity class and the remaining severity categories are combined to be non-serious injury severity class. The severe injury severity class is the minority class. There are different approaches to overcome this problem in dataset. Over-sampling and under-sampling are two popular techniques to help balance the dataset.

Over-sampling is a method in which samples of a minority class is synthetically produced using different techniques. The method is used to increase the sensitivity of a classifier to the minority class. One of the most widely used over-sampling methods is the Synthetic Minority Over-Sampling Technique (SMOTE) algorithm [51]. The working principle of this method is it finds two nearest neighbor minority class data points, draws a line between them and generates a synthetic data point belonging to the minority class in that line. In this study, an open sourced, MIT-licensed library for python programming language, Imbalanced-learn is used to implement the SMOTE algorithm. The SMOTE algorithm uses unsupervised nearest neighbors learning algorithms like Multidimensional binary search trees (KD Tree) [52] and Ball Tree [53].

Several modified versions of SMOTE have been proposed such as Borderline-SMOTE [54], ADASYN [55]. There are also techniques that employ a hybrid approach for resampling, using both over-sampling and under-sampling. One such method is SMOTE-Tomek Links which combines over-sampling method from SMOTE and under-sampling from Tomek Links. Tomek links is an under-sampling method that identifies pair of data points called 'Tomek Link'. 'Tomek Link' are pair of data points belonging to different class such that their Euclidean distance is minimum and one of the points belongs to the minority class.

The Tomek link used in the Tomek algorithm is mentioned below. Here, $d(x_i, x_j)$ denote the Euclidean distance between $x_i$ and $x_j$, where $x_i$ denotes the sample that belongs to the minority class and $x_j$ denotes sample that belongs to the majority class. If no sample satisfies the following condition:

$$d(x_i, x_k) < d(x_i, x_j), or$$

$$d(x_j, x_k) < d(x_i, x_j)$$

then the pair of $(x_i, x_j)$ is a Tomek Link.

The method identifies 'Tomek Link' pairs and eliminates the data point belonging to the majority class but near/identical to the minority class.

Similarly, a hybrid approach in which over-sampling is paired with under-sampling is SMOTE-NearMiss sampling. In this hybrid method of resampling, over-sampling of the minority class is paired with under-sampling of the majority class based on its average distance with the minority class. The smallest averaged distance sample is selected for under-sampling.

The resampled datasets are evaluated before analyzing any further. The evaluation is done based on evaluation score on RF model using a repeated stratified K-fold cross-validation technique.

The resampled dataset is divided into 10 folds and the cross validator is repeated three times to get the average scores for each sample.

## 4.4 Logistic Regression (classification)

### 4.4.1 Variable Selection

As part of feature selection, initially many variables are eliminated in the preprocessing stage of the analysis. Empty variables, variables with too many null values and irrelevant variables were eliminated. In addition to eliminating empty and unusable variables, collinearity test is also conducted using Pearson correlation coefficient. Variables which are correlated with each other with more than 0.6 correlation coefficient are removed.

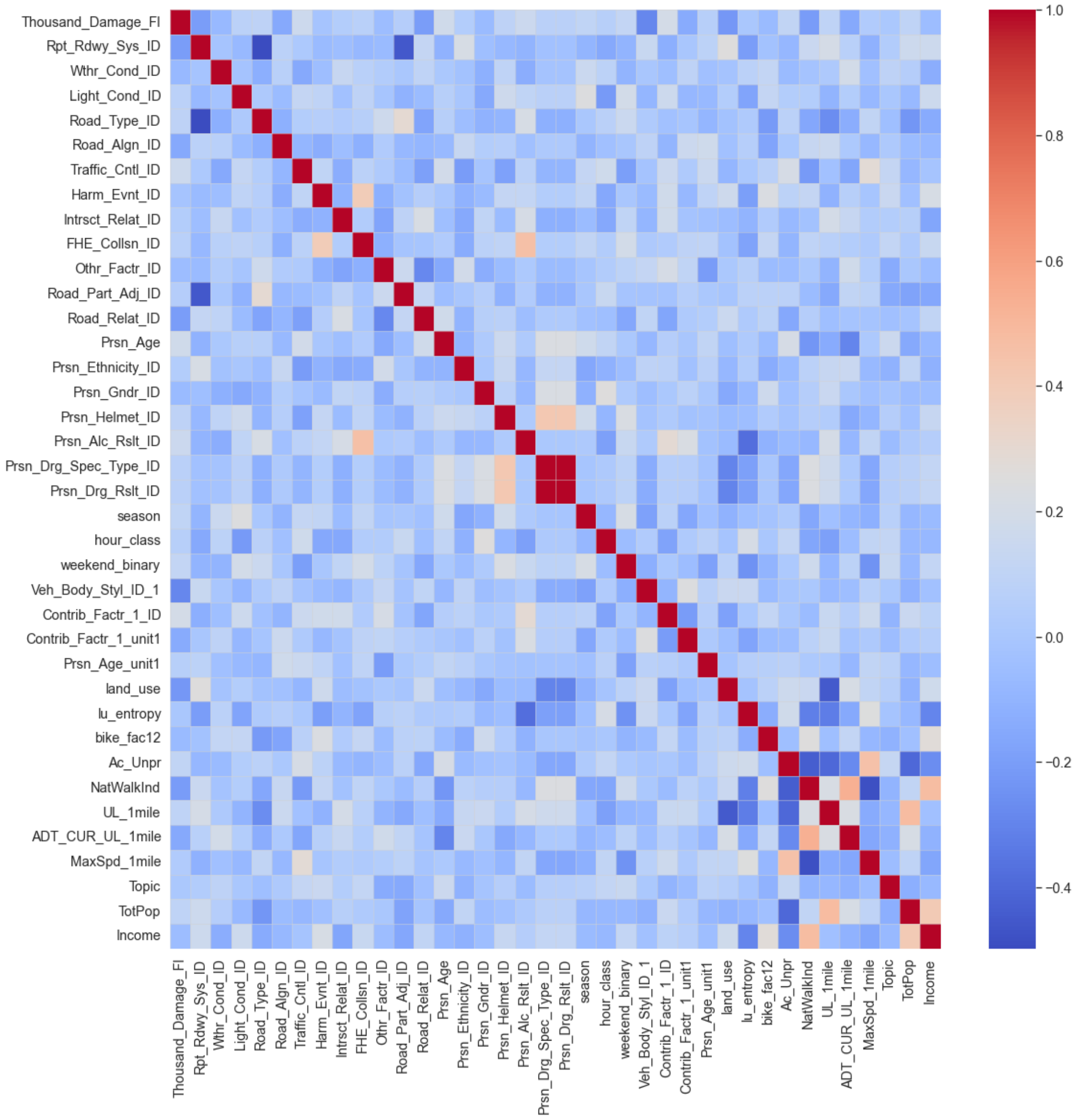


Figure 3 Heatmap of correlation coefficient of independent variables and injury severity of e-scooter crashes after removing variables with >0.6 coefficient with one another.

Figure 3 depicts the heatmap of Pearson correlation coefficient between multiple variables after elimination of variables with more than 0.6 or -0.6 coefficient for e-scooter related crash data. Similarly, Figure 4 shows Pearson correlation coefficient for bicycle related crash data variables.

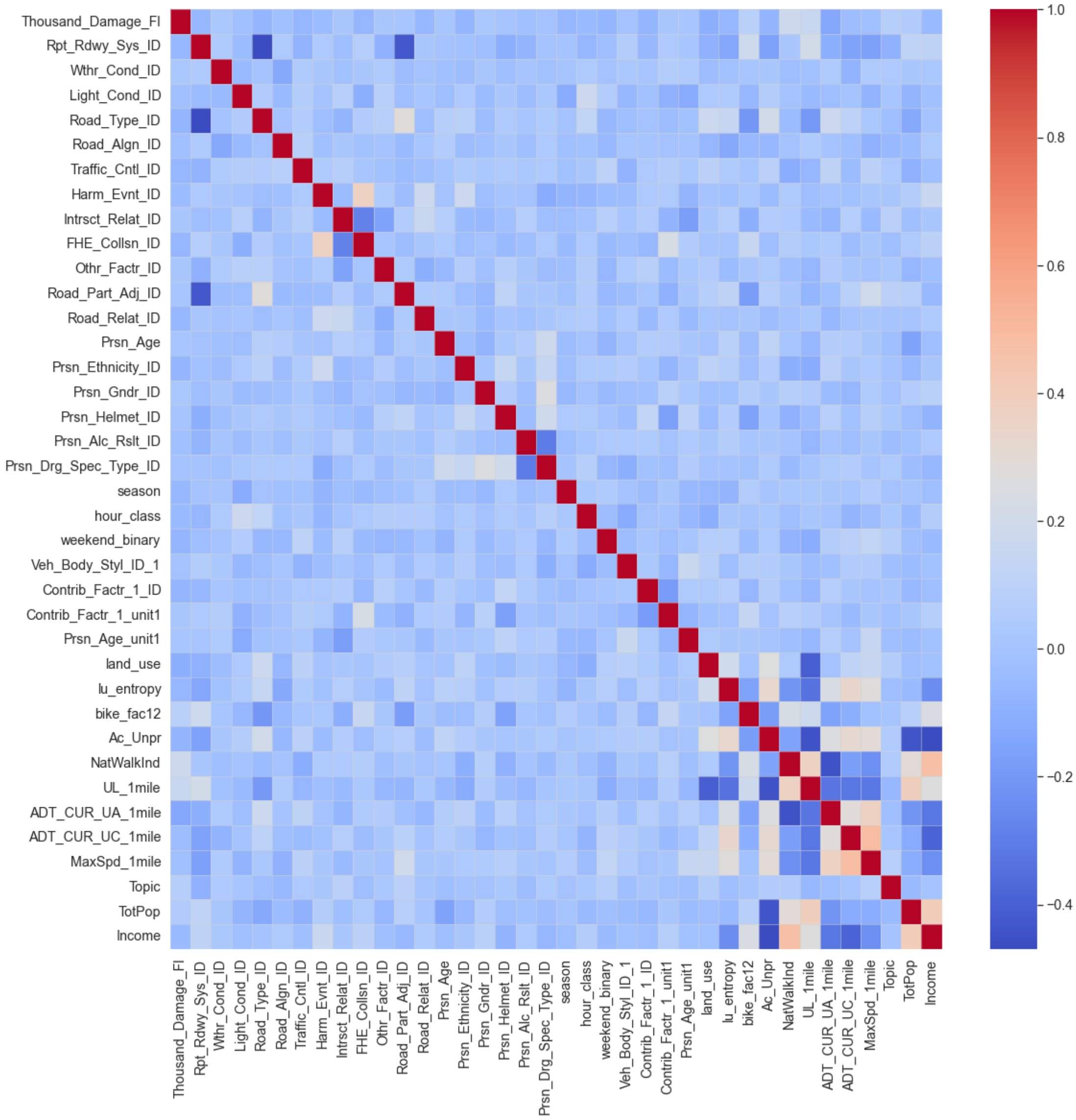


Figure 4 Heatmap of correlation coefficient of independent variables and injury severity of bicycle crashes after removing variables with >0.6 coefficient with one another.

In addition to checking for colinearity, variable selection is done using chi- square contingency rest and PhiK correlation coefficient. PhiK is a new and useful correlation coefficient that captures non-linear dependency, consistently works with categorical, ordinal, and interval variables, and reverts to the Pearson correlation coefficient in the case of a bivariate normal input distribution [50]. The correlation PhiK is derived from Pearson's $\chi$ 2 contingency test i.e., the hypothesis test of independence between two (or more) variables in a contingency table. The correlation PhiK follows a uniform treatment for interval, ordinal and categorical variables [50].

The PhiK coefficient is calculated by firstly calculating Pearson's chi-squared contingency test statistic $\chi^2$.

$$X^2 = \sum_{0}^{ij} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Where, i and j are rows and columns of the contingency table. O is the observed frequency and E is the expected frequency.

There are some drawbacks to using PhiK coefficients such as there are no closed-form formula for a PhiK coefficient. The coefficient ranges from 0 to 1, therefore does not indicate directional characteristics and the coefficient is not as precise as other coefficients such as Pearson's correlation coefficient if working with numeric-only variable. Despite these shortcomings, this method is sufficient and appropriate for preliminary variable selection in this study.

### 4.4.2 Logistic model

Logistic regression classifier is used to analyze the contributing factors for injury severity classification. Statsmodels api was used to analyze the datasets using logistic regression. Logistic regression classifier is a transformation of a linear regression using the sigmoid function. The vertical axis stands for the probability for a given classification and the horizontal axis is the value of *x*. It assumes that the distribution of y|x is Bernoulli distribution. The formula of LR is as follows:

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \sum_1^i \beta_i x_i)}}$$

Here $\beta_0 + \sum \beta_i x_i$ is similar to the linear model $y = ax + b$. The logistic function applies a sigmoid function to restrict the y value from a large scale to within the range 0–1 which in this study represent severe and non-severe injury.

### 4.4.3 Performance Evaluation

The logistic regression model is evaluated using a confusion matrix. The metrics of accuracy, precision, recall and F score were derived from the confusion matrix presented in Table 3.

Table 3 Confusion matrix for 2-class classification

| | Predicted Positive | Predicted Negative |
|---|---|---|
| Actual Positive | True Positive (TP) | False Negative (FN) |
| Actual Negative | False Positive (FP) | True Negative (TN) |

The metrics generated from the confusion matrix are calculated as:

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{TP + TN + FN + FP}}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP + FP}}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP + FN}}$$

$$F\ score = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

The recall score of the test dataset was used as a basis for evaluating the models, as recall is sensitive to false negatives. A false negative occurs when a true positive is misclassified as a negative. In other words, this study places a greater emphasis on correctly identifying serious or fatal injuries and is more sensitive to falsely classifying serious injury as non-serious injury.

## 4.5 Machine learning model

In this paper, two tree based models are used to train the data. The models are discussed in detail below. The first model is a random forest, which is an ensemble learning method that uses multiple decision trees to make predictions. The second model is a XGBoost, which is a powerful and widely used classification algorithm that uses gradient boosting to improve model performance. Each of these models has its own strengths and weaknesses, and the results of the training process are discussed in the subsequent sections of this paper.

### 4.5.1 Variable selection

A separate collinearity analysis is done while preparing the dataset for machine learning model as tree-based models are robust against collinearity compared to logistic regression classification model. A threshold of |0.8| Pearson correlation coefficient is considered while selecting features for model training.

After excluding features with high correlation with each other, PhiK correlation coefficient was conducted to select the best features for model training. The top variables according to the Phi_K coefficient, as shown in Figure 5, were selected for the model. In addition, some additional variables were included based on our literature review and domain knowledge.

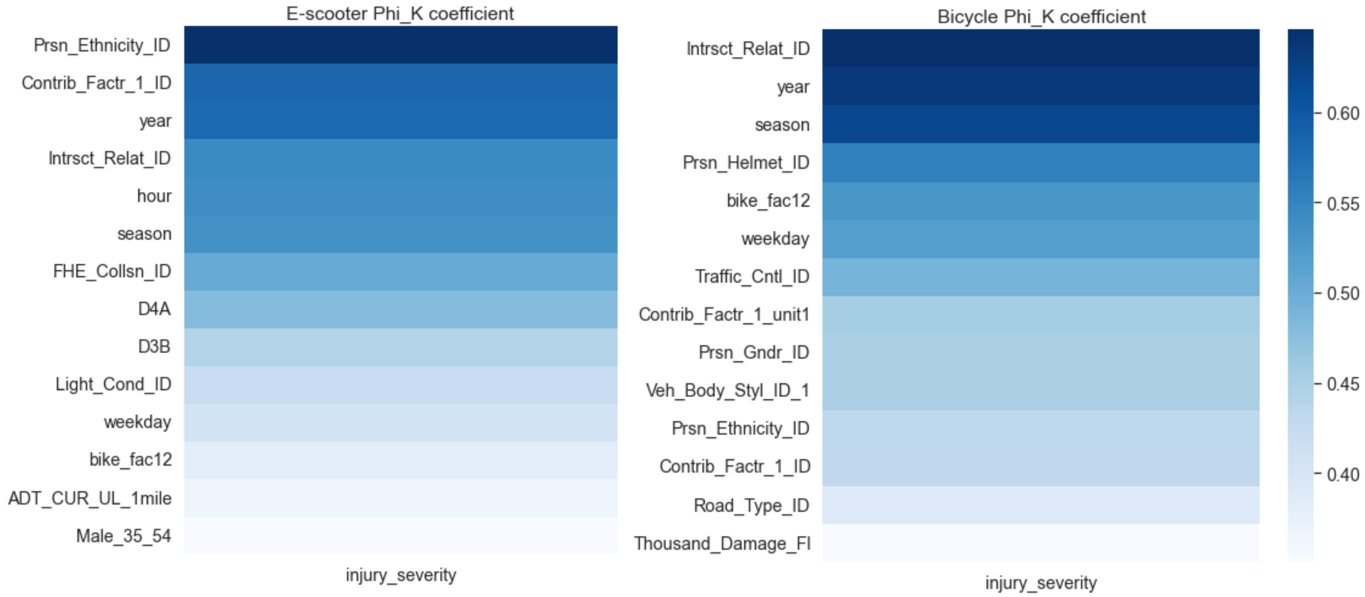


Figure 5 Heatmap of PhiK correlation coefficient for e-scooter (left) and bicycle (right) with respect to injury severity

### 4.5.2 Parameter tuning

The whole dataset for each mode of transportation is divided into training data and validation data. The validation data is used for calculating optimal hyperparameter. The hyperparameter of the models are tuned using cross validation method. Cross validation technique divides the data into multiple subsets (folds) and train the model using different combination of subsets. The remaining subsets are used to test the performance of the trained model. This technique does not require separate training and testing dataset but using the whole data to train and test by making use of the subsets. Grid search technique involves specifying a grid of hyperparameters and a range of values for each parameter on which model is trained and evaluated. The optimal hyperparameters for each model and each mode of transportation are mentioned in Table 4.

Table 4 Optimal hyperparameter for Random Forest and XGBoost

| Model | Random Forest Parameter | E-scooter | Bike |
|---|---|---|---|
| Random Forest | class_weight | balanced | balanced |

|  |  |  |  |
|---|---|---|---|
|  | criterion | gini | gini |
|  | max_depth | 15 | 25 |
|  | min_sample_leaf | 2 | 2 |
|  | min_sample_split | 5 | 2 |
|  | n_estimators | 100 | 100 |
| XGBoost | alpha | 0 | 0 |
|  | eta | 0.01 | 0.01 |
|  | gamma | 0 | 0 |
|  | lamda | 0 | 0 |
|  | max_depth | 2 | 2 |
|  | min_child_weight | 1 | 0.1 |
|  | subsample | 0.2 | 0.3 |

### 4.5.3 Random Forest Model

Random forest (RF) is a commonly used machine learning model which utilizes multiple decision trees to reach a single result [56]. Decision trees are supervised learning algorithms typically trained through the Classification and Regression Tree (CART) algorithm. Decision tree can be understood as a web of decisions that spilt the data into two groups at each node. Metric such as Gini impurity, information gain, or mean square error (MSE) are used to evaluate the quality of the split.

RF is a set of decision trees combined with bagging method and feature randomness or feature bagging. The individual decision trees in the set of decision trees are built using a subset of data points selected by bagging method, the individual decision- tree’s features are also a subset of the actual features selected at random. These techniques make a random forest superior to a simple decision tree algorithm in terms of classification and regression as it is less susceptible to over-fitting, noise, and biases in the data.

### 4.5.4 Extreme Gradient Boosting (XGBoost)

XGBoost is a tree boosting machine learning algorithm under the framework of gradient boosting [57]. Just as RF combines the advantages of bagging and decision trees, gradient boosting method (GBM) forms a powerful tree-based learner through continuous iteration of the gradient optimization algorithm. The combination results in XGBoost, first proposed by Chen and Guestrin 2016 [58]. The algorithm is more efficient and powerful than other tree-based methods since it combines software and hardware optimization techniques and can be classified and regression.

SHAP (SHAPley Additive exPlanations) is a technique used to evaluate the importance of features in a machine learning model. It is a game theory-based model implemented as a python library that provides explanations for the output of any machine learning model. SHAP uses SHAPley values, a concept from cooperative game theory, to determine the contribution of each input feature to the model's output. This technique helps to allocate "credit" for the model's predictions among its input features, allowing for a more detailed understanding of how the model is making its predictions [59].

# 5 RESULTS

## 5.1 Resampling

Three resampling methods were implemented as discussed in the method section. The resampled datasets were evaluated using RF model, which showed that resampling done after variable selection performed better. The evaluation was done using cross validation simultaneously during the training of the model. The results of different resampled data using random forest cross validation is shown in the Table 4. Figure 6 shows the scattered plot of different sampled data with respect to age and network density. 0 here indicates non-serious injury and 1 indicates serious injury.

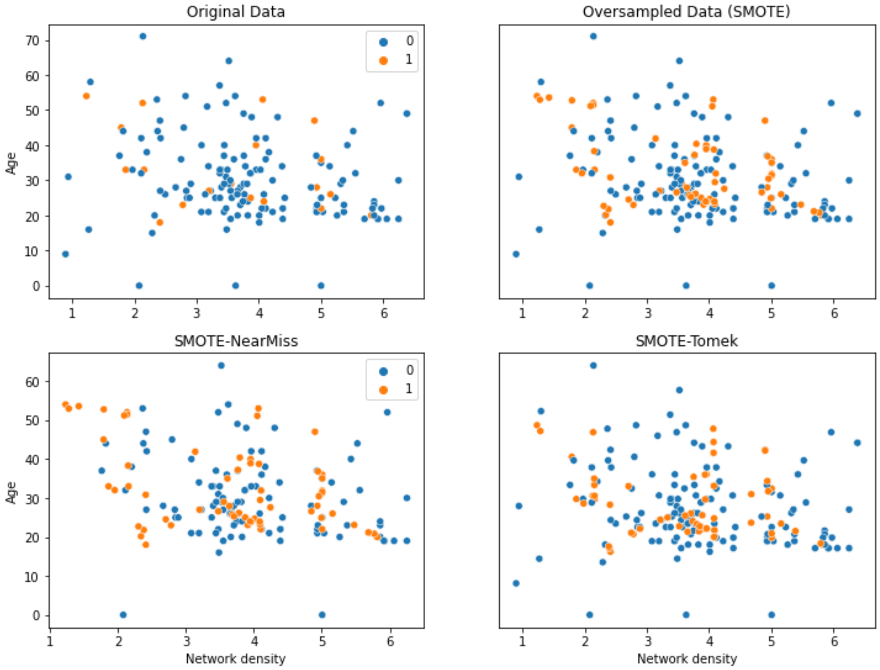


Figure 6 Scatter plot of different resampled dataset

Table 5 Summary of sample evaluation

| Mode | Sampling Technique | Sample size (non-severe, severe) | Mean Accuracy | Mean Precision | Mean Recall |
|---|---|---|---|---|---|
| E-scooter | Original | (132, 21) | 0.8672 | 0.1000 | 0.0667 |
| | SMOTE | (132, 66) | 0.8972 | 0.9540 | 0.7278 |
| | SMOTE-NearMiss | (99, 66) | 0.8759 | 0.9232 | 0.7643 |
| | SMOTE-Tomek | (128, 66) | 0.9118 | 0.9738 | 0.7571 |
| Bicycle | Original | (737, 103) | 0.8750 | 0.0833 | 0.0094 |
| | SMOTE | (737, 368) | 0.8745 | 0.8915 | 0.7110 |
| | SMOTE-NearMiss | (552, 368) | 0.8438 | 0.8545 | 0.7393 |
| | SMOTE-Tomek | (726, 368) | 0.8728 | 0.8891 | 0.7183 |

The results show that resampling of the imbalanced data produced a much accurate classification model. Among the resampling methods we can see the performance of model trained with SMOTE and SMOTE-Tomek sample gave a higher precision and accuracy. Recall score is more sensitive to false negative which is more concerning to this study as it is studying the rare case of severe injury incidents. Hybrid resampling method SMOTE-NearMiss produces higher recall scores in both e-scooter and bicycle dataset. The study uses resampled dataset using SMOTE-NearMiss method for further analysis based on the above results.

## 5.2 Topic Modeling

LDA was used to cluster topics for analysis for both the mode of transportation. Number of topics was fixed to 3 topics after testing for 3 t 8 topics as more topics ended up having more common words. Comparing the most common words in each topic it was evident that the difference in each cluster was not very prominent. Using LDA topic modeling, a term-dictionary consisting of a list of terms with their probability of occurrence in the documents corresponding to each topic is generated. The following table contains the top eight terms for each topic with its probability of occurrence.

Table 6 Top eight words in a topic and its probability for e-scooter using a LDA model

| Topic | Top eight terms | Mode |
|---|---|---|
| 0 | 0.020*"right" + 0.019*"travel" + 0.014*"driver" + 0.014*"turn" + 0.013*"lane" + 0.011*"leave" + 0.011*"state" + 0.011*"stop" | e-scooter |
| 1 | 0.023*"travel" + 0.022*"state" + 0.018*"pedestrian" + 0.016*"right"+ 0.015*"stop" + 0.015*"turn" + 0.015*"driver" + 0.012*"lane" | e-scooter |
| 2 | 0.019*"right" + 0.019*"travel" + 0.015*"turn" + 0.013*"driver" + 0.012*"stop" + 0.012*"leave" + 0.012*"pedestrian" + 0.011*"bound" | e-scooter |
| | | |
| 0 | 0.023*"turn" + 0.021*"travel" + 0.021*"state" + 0.020*"right" + 0.018*"stop" + 0.016*"leave" + 0.014*"lane" + 0.013*"driver" | Bike |
| 1 | 0.029*"travel" + 0.026*"turn" + 0.024*"right" + 0.020*"lane" + 0.019*"stop" + 0.016*"state" + 0.015*"driver" + 0.015*"leave" | Bike |
| 2 | 0.030*"travel" + 0.025*"turn" + 0.022*"state" + 0.020*"leave" + 0.017*"right" + 0.016*"front" + 0.016*"lane" + 0.013*"stop" | Bike |

Comparing just the top eight words and its probability in each of the topics in both the mode of transportation we can observe there are a lot of repetitive words. The word “right”, “travel”, “driver”, “turn” are common on all three topics of e-scooter related crash narrative. Similarly, “turn”, “travel”, “state”, “right”, “stop”, and “leave” are common in all the topics in bike related crash narrative. Many words are also common within two of the topics in both mode of transportation. The words are also not indicative of any unusual behavior or action that might give clue related to severity of injury.

Since the topics generated by LDA was of no use, another method of topic modeling was implemented. The BERTopic library for python was implemented to cluster the documents into a number of topics for both the mode of transportation. Different number of topics were made using the documents. Different topics for e-scooter related crash as well as bike related crash text were analyzed for PhiK correlation with injury severity. The PhiK correlation coefficient was compared with the Phik correlation coefficient between age of rider and injury severity. The "nr_topics" parameter was changed to get different number of topics for the documents. The parameter was varied from 3 to 18 and the PhiK was compared. The topics which had the highest PhiK coefficient was chosed as the most optimal topic with respect to injury severity. The e-scooter documents were clustered into six topics and the topics had a PhiK correlation coefficient of 0.237 with the injury severity which is similar to PhiK correlation coefficient between injury severity and age of rider which is 0.281. Similarly, bike documents were clustered into 11 topics and the Phik correlation coefficient between injury severity and topics was 0.15 which is more than the coefficient of 0.14 between rider's age and injury severity.

The top eight words and its probabilities from the selected topics are listed in Table 7 and Table 8. The words and its probability are distinct in each topic with some words like travel, vehicle, repeated in many topics.

Table 7 Top eight words in a topic and its probability for e-scooter dataset.

| Topic | Top eight terms (e-scooter) |
|---|---|
| 10 | 0.059 * right + 0.056* travel + 0.052* stop + 0.051* turn + 0.047* driver + 0.046* leave + 0.044* st + 0.043* state |
| 0 | 0.187* pedestrian + 0.069* way + 0.067* travel + 0.065* vehicle + 0.063* front + 0.057* right + 0.056* intersection + 0.053* sidewalk |

| | |
|---|---|
| 1 | 0.086* officer + 0.061* scene + 0.06* state + 0.057* respond + 0.056* vehicle + 0.05* front + 0.049* block + 0.047* crash |
| 2 | 0.099* state + 0.086* driver + 0.062* intersection + 0.059* travel + 0.057* cross + 0.049* green + 0.049* st + 0.048* lane |
| 3 | 0.123* bound + 0.094* lane + 0.088* travel + 0.086* nit + 0.085* turn + 0.084* gin + 0.07* right + 0.066* leave |
| 4 | 0.171* red + 0.154* crosswalk + 0.153* light + 0.077* strike + 0.077* turn + 0.074* attempt + 0.071* cross + 0.068* proceed |

Table 8 Top eight words in a topic and its probability for bicycle dataset.

| Topic | Top eight terms (bike) |
|---|---|
| 10 | 0.044* turn + 0.043* travel + 0.039* right + 0.037* lane + 0.036* leave + 0.036* state + 0.033* stop + 0.031* front |
| 0 | 0.051* right + 0.049* turn + 0.047* state + 0.046* lane + 0.042* travel + 0.039* driver + 0.036* front + 0.034* private |
| 1 | 0.101* light + 0.088* green + 0.079* intersection + 0.067* red + 0.057* travel + 0.053* turn + 0.044* st + 0.039* leave |
| 2 | 0.129* stop + 0.104* sign + 0.073* intersection + 0.05* driver + 0.046* sidewalk + 0.041* right + 0.04* rider + 0.039* state |
| 3 | 0.073* respond + 0.068* officer + 0.051* state + 0.049* report + 0.045* crash + 0.045* call + 0.034* collision + 0.034* scene |
| 4 | 0.121* crosswalk + 0.104* cross + 0.083* st + 0.074* state + 0.071* strike + 0.066* travel + 0.053* signal + 0.043* person |
| 5 | 0.067* injury + 0.06* transport + 0.056* dell + 0.05* shoulder + 0.05* rider + 0.046* right + 0.045* seton + 0.044* strike + 0.035* damage |

| | |
|---|---|
| 6 | 0.161* manor + 0.076* lane + 0.065* block + 0.048* travel + 0.048* leave + 0.048* nit + 0.047* yield + 0.043* advise |
| 7 | 0.124* strike + 0.109* travel + 0.103* fail + 0.1* nit + 0.088* yield + 0.083* front + 0.083* stop + 0.065* ave |
| 8 | 0.218* riverside + 0.052* turn + 0.048* stop + 0.045* freeman + 0.044* travel + 0.044* cross + 0.044* right + 0.043* sidewalk |
| 9 | 0.236* lot + 0.2* parking + 0.074* turn + 0.073* exit + 0.072* one + 0.066* park + 0.064* head + 0.064* leave |

Figure 7 represents the percentage distribution of e-scooter topics across injury severity before resampling. Each cell in the non-severe row represents the percentage of data points with the respective topic among non-severe injury type. Topic-0 and Topic-1 are seen to have disproportionate numbers of severe injury type data. Similarly, Figure 8 represent the distribution of topics in different injury severity type. The distribution is not very distinct in bicycle topics, but Topic-1 has a slightly more severe injury type data.

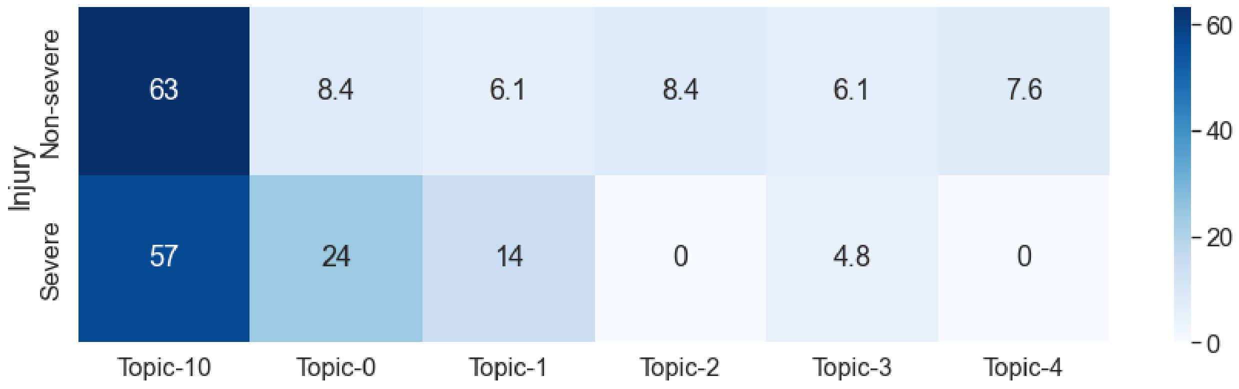


Figure 7 Percentage distribution of e-scooter topics across injury severity before resampling

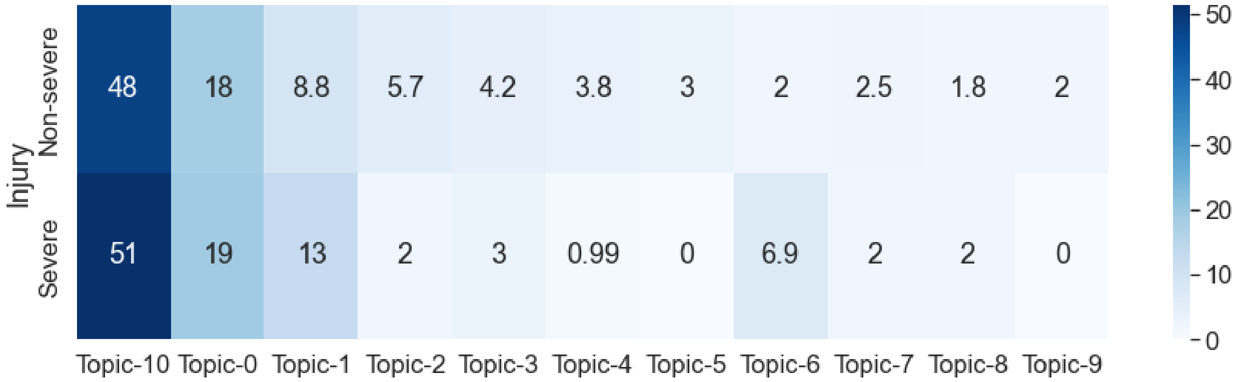


Figure 8 Percentage distribution of bicycle topics across injury severity before resampling

Figure 9 and Figure 10 are percentage distribution of topics with respect to injury severity type in e-scooter and bicycle after resampling using Smote-NearMiss. Like original data Topic-0 has higher presence in severe type injury severity whereas Topic-1 is neutralized as seen in Figure 9. The distribution for bicycle has also changed slightly after resampling, Topic-1 is still highly present in severe injury type whereas Topic-6's disproportional presence in two types of injury severity is decreased.

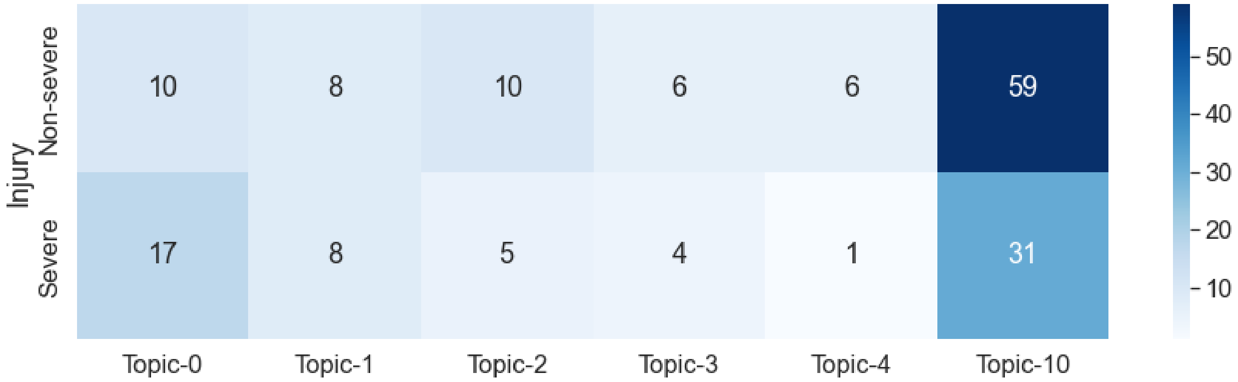


Figure 9 Percentage distribution of e-scooter topics across injury severity after resampling

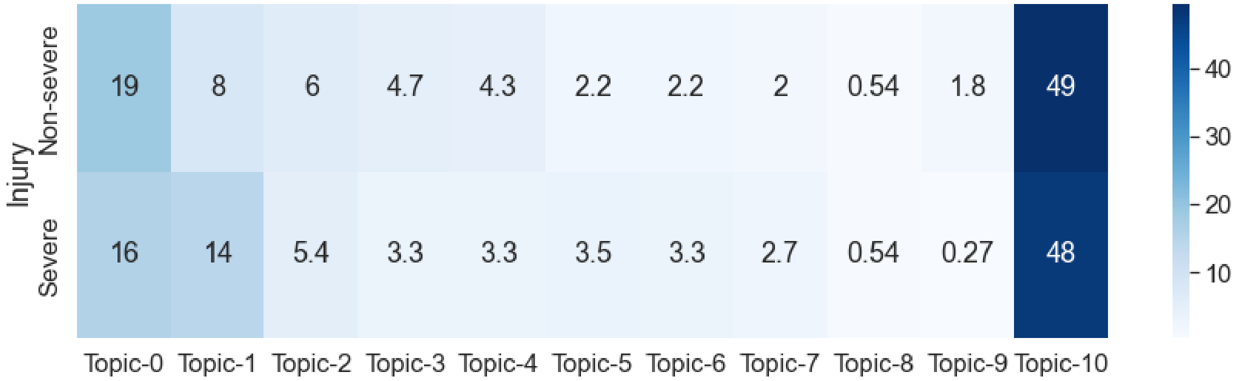


Figure 10 Percentage distribution of bicycle topics across injury severity after resampling

## 5.3 Logistic Regression (classification)

Logistic regression model was conducted for both the e-scooter and bicycle dataset. The accuracy of the logistic regression classification model for e-scooter dataset is found to be 0.618. Similarly, the recall and precision are 0.30 and 0.54. The F1-score calculated from the recall and precision is 0.388.

Likewise, the same analysis with the bicycle dataset gave an accuracy score of 0.65. The recall and precision of the dataset was 0.4 and 0.6 respectively. Using the recall and precision, the F-1 score was calculated which turned out to be 0.48.

The evaluation measure of the bicycle dataset is slightly better than the e-scooter dataset. The number of datapoints in both the datasets are not equal which can also be a reason for the slightly lower scores of e-scooter datasets. From the model, coefficient, and p-values of each of the variables are extracted and mentioned below.

The coefficient of all the variables with a p-value less than 0.1 for the e-scooter dataset is listed below. Observation shows the cost of the damage is the highest contributing variable for e-scooter model followed by weekday-weekend indicating variable. Other variables contributing are hour class, ethnicity of rider, manner of maneuver of rider, traffic control type, age of rider, contributing factor of driver, contributing factor of rider, income, total population.

Injury severity (e-scooter) = cost of damage * 1.74 - weekend/weekday * 1.52 – hour class * 0.49 + ethnicity of rider * 0.28 – manner of maneuver of rider * 0.1 - traffic control type * 0.079 + age of rider * 0.031 – contributing factor of driver * 0.04 + contributing factor of rider * 0.026 + Income * 0.00002 - Total population * 0.00001

Similarly, the coefficient of all the variables with a p-value less than 0.1 for the e-scooter dataset is listed below. It is observed that cost of damage is the highest contributing factor for bicycle model followed by helmet use and the type of vehicle in the crash and the type of bicycle facility. Other variables are season of the year, road alignment, maximum speed and contributing factor. The maximum speed is also correlated with the density of interstate/freeway highways around the crash location.

Injury severity (bike) = cost of damage * 1.01+ helmet use * 0.38 + vehicle body type* 0.334 - bike facility type* 0.23 - season * 0.21 + road alignment * 0.15 - maximum speed * 0.017 - contributing factor of vehicle * 0.01

Although the accuracy is not high the contributing factors are different for e-scooters and bicycles. For instance, weekday or weekend and hour of crash has an impact on e-scooter crash severity whereas season of year has an impact on bicycle crash. Similarly, Ethnicity, total population around the crash location and income of people around crash location only have an impact on e-scooter injury severity classification. Built environment properties such as bicycle facility type and maximum speed or density of interstate/freeway around crash location has a contribution on bicycle crash injury severity classification and has no impact on e-scooter severity classification.

The contributing factor variable which indicates the fault of rider or driver just before the crash contributes to both the classification. But both driver's contributing factor and e-scooter's contributing factors has an impact on e-scooter severity classification whereas in case of bicycle motor vehicle driver's contributing factor has an impact on severity classification.

The data on helmet on e-scooter was mostly empty, might also be because most e-scooter users do not were helmet as shown by multiple studies[31]. In case of bicycle use of helmet was also important feature in classifying if a crash was severely injured or not.

## 5.4 Machine learning models

As mentioned in the method section, three models were trained to classify serious injury and non-serious injury causing crashes in e-scooters and bicycles. The accuracy, precision and recall score of each model calculated using its confusion matrix is mentioned below.

Table 9 Summary of model evaluation

| Mode | Model | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| E-scooter | Logistic | 0.618 | 0.54 | 0.3 | 0.54 |
| | Random Forest | 0.902 | 0.900 | 0.890 | 0.884 |
| | XGBoost | 0.815 | 0.733 | 0.898 | 0.799 |
| Bicycle | Logistic | 0.65 | 0.6 | 0.4 | 0.48 |
| | Random Forest | 0.876 | 0.917 | 0.762 | 0.830 |
| | XGBoost | 0.705 | 0.595 | 0.838 | 0.695 |

The results demonstrate that machine learning models, specifically RF and XGBoost, outperform the logistic regression classification model in accurately classifying injury severity type. For instance, the random forest model achieves an accuracy of 0.902 and an overall F1 score of 0.875 in predicting injury severity for E-scooter riders. Similarly, the XGBoost model attains an accuracy of 0.854 and F1 score of 0.840. Similarly for bicycle riders RF scores an accuracy of 0.876 and F-1 score of 0.83. Likewise, XGBoost scores an accuracy of 0.705 and F-1 score of 0.695. The recall of XGboost is greater than any model in both the datasets. Since this study is more concerned about identifying severe injury severity type than non-severe injury severity type, recall is considered to select the model for further analysis. A higher recall score indicates that the model can identify most of the positive cases which in this analysis is severe injury.

severity type. SHAP analysis is conducted for the XGBoost model to understand the contributing features.

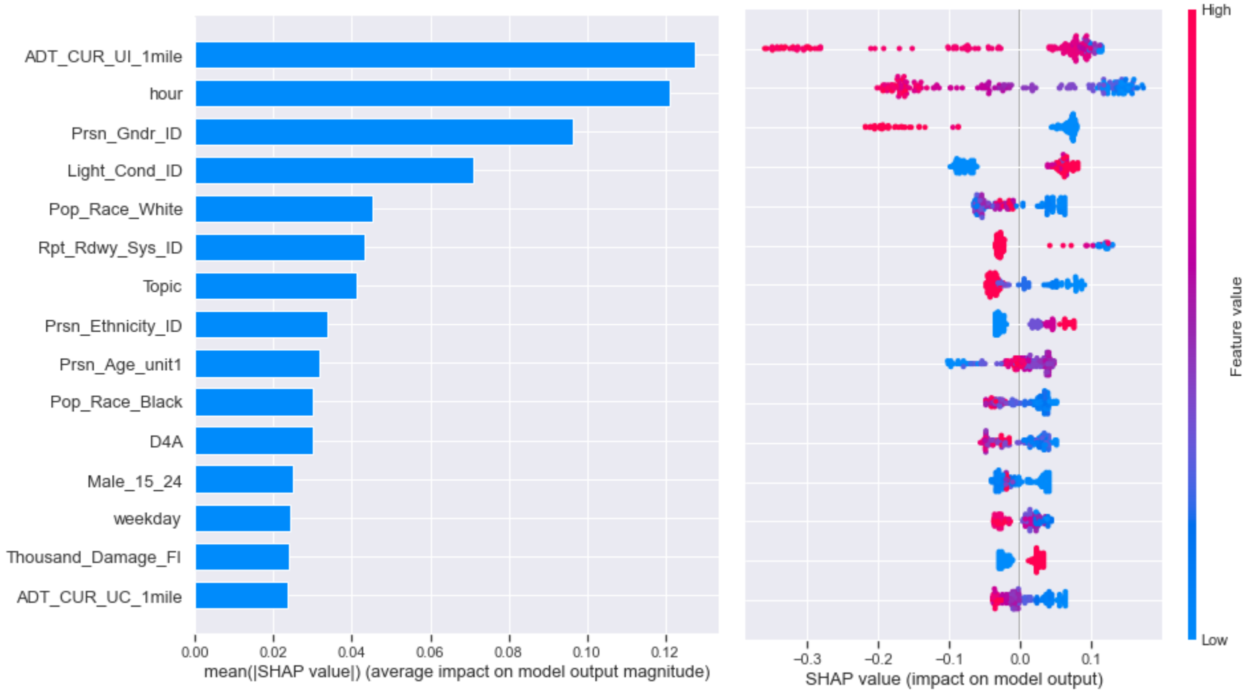


Figure 11 Top 15 features with its average SHAP value (left) and positive, negative relationship of the top 15 features with the target (right) for XGBoost model trained with e-scooter dataset.

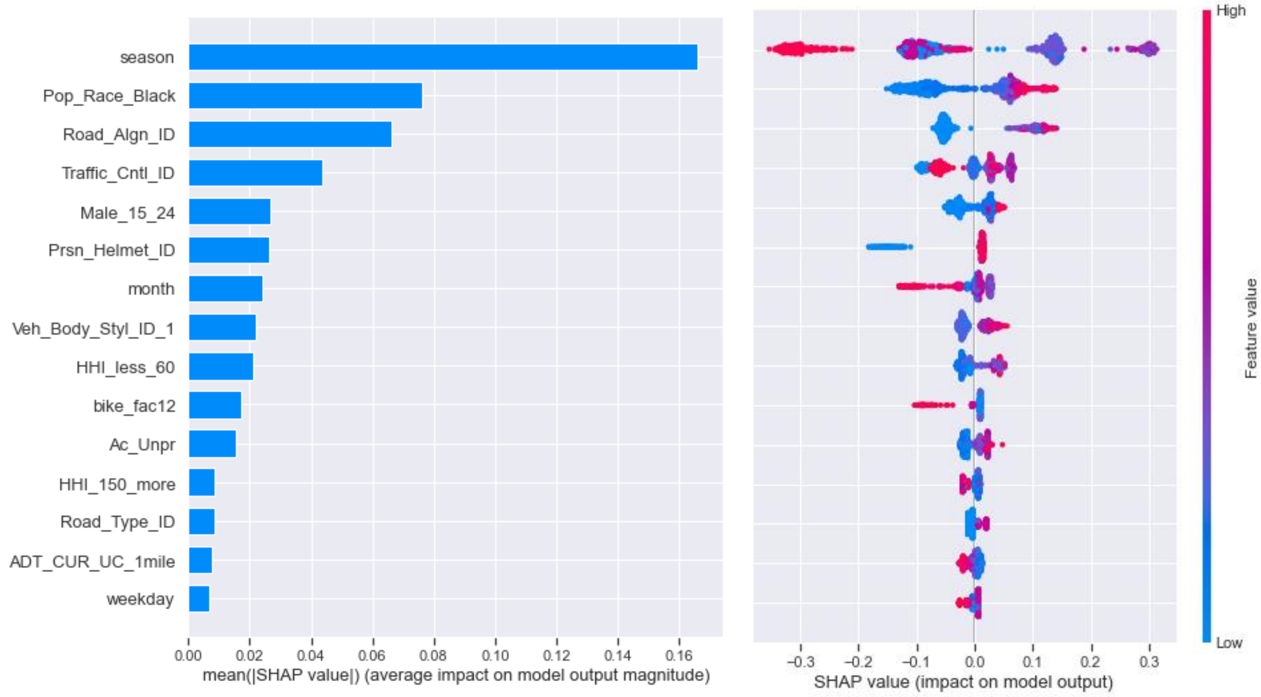


Figure 12 Figure 5 Top 15 features with its average SHAP value (left) and positive, negative relationship of the top 15 features with the target (right) for XGBoost model trained with bicycle dataset.

The vertical bar plot on the left in Figure 11 and Figure 12 shows the top features in descending order with its average of absolute SHAP values. The plot on the right side of the figures shows the positive and negative relationship of the top features with the target. The current average daily traffic, hour of day, gender of the rider and light condition of the location is the major contributing factor in e-scooter injury severity classification. Likewise, season of the year, population of black population, alignment of road, and traffic control type are major contributors in bicycle crash injury classification.

Some major contributing factors are now compared between e-scooter riders and bicycle riders' injury severity classification.

Average daily traffic of interstate/freeway roadways has a high impact on e-scooter injury classification whereas the feature is not within the top 15 important features for bicycle injury classification.

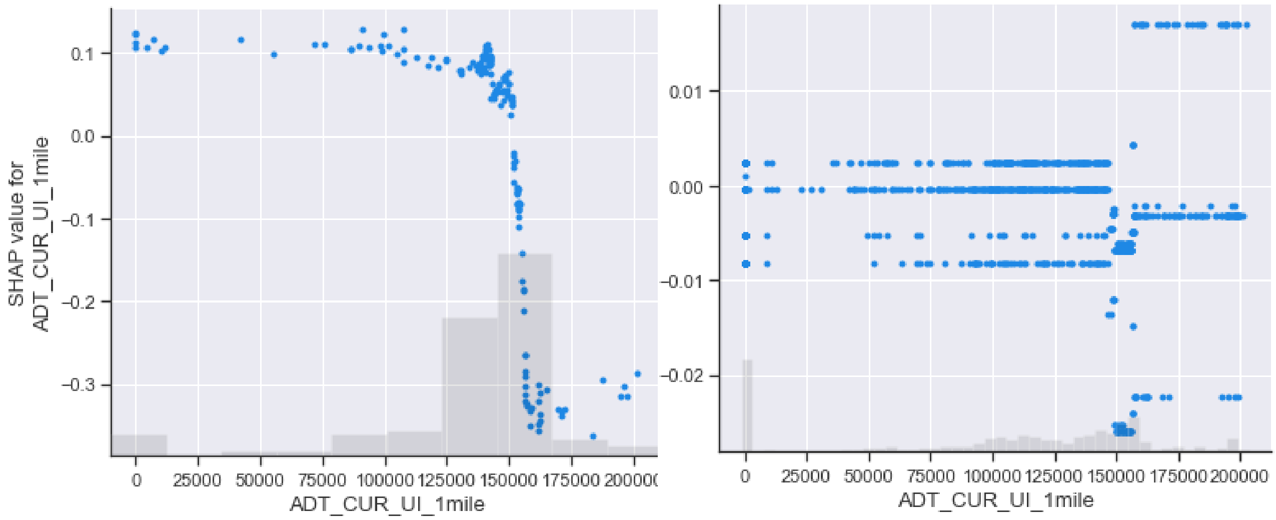


Figure 13 SHAP dependency plot showing the effect of average annual traffic on interstate/freeway road on the prediction of the models. The plot on the left side is from the e-scooter model and the plot on the right side is from the bicycle model.

It is clear from the dependency plot that the average daily traffic in the interstate or freeway within 1 mile radius of the e-scooter crash has a clearer and higher impact on injury severity classification. Average daily traffic less than 150,000 indicates a severe injury. It is also evident from the graphs that crash location with zero or no interstate/freeway within 1 mile radius has a more severe injury. ADT for interstate/freeway is highly correlated with density of interstate/freeway, it might also indicate e-scooter crashes occurring in inner cities are more dangerous than crashes occurring in areas outside of inner city.

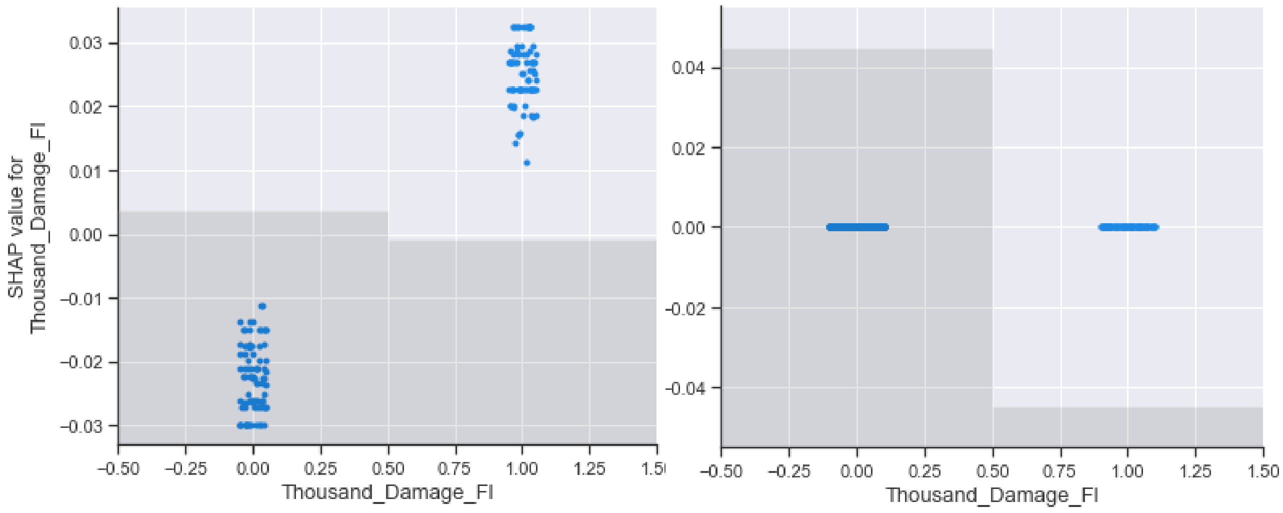


Figure 14 SHAP dependency plot showing the effect of cost of crash on the prediction of the models. The plot on the left side is from the e-scooter model and the plot on the right side is from the bicycle model.

The cost of damage indicated by the variable ‘Thousand_Damage_FI” has no impact on bicycle crash injury model whereas it has a meaningful impact on e-scooter crash injury. In case of e-scooter crash, if a crash’s damage is more than a thousand dollors or equivalent it is more likely to have a severe injury.

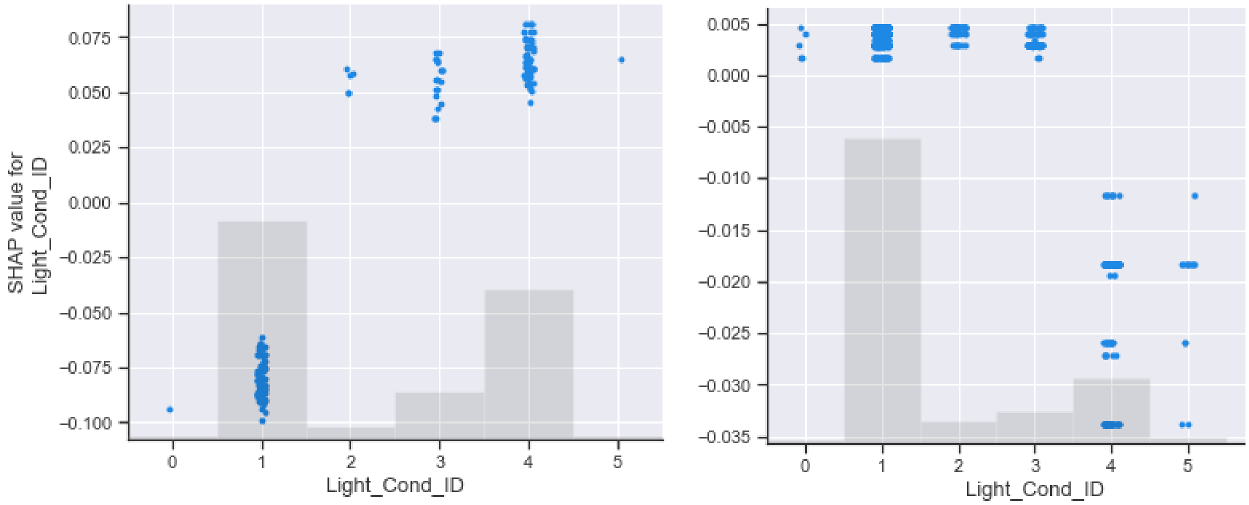


Figure 15 SHAP dependency plot showing the effect of condition of lighting on the prediction of the models. The plot on the left side is from the e-scooter model and the plot on the right side is from the bicycle model.

The condition of lighting has very less impact on bicycle injury classification compared to e-scooter classification. Loght_Cond_ID code 1, which represent daylight has a negative impact on severe classification of injury severity in e-scooters whereas code 4 and 5 which are Dark, lighted and Dusk has a negative impact in case of bicycle. The dependency plot shows indicates that crashes at dark hours contributes to severe e-scooter injury whereas it doesnot have a huge impact on bicycle crash injury.

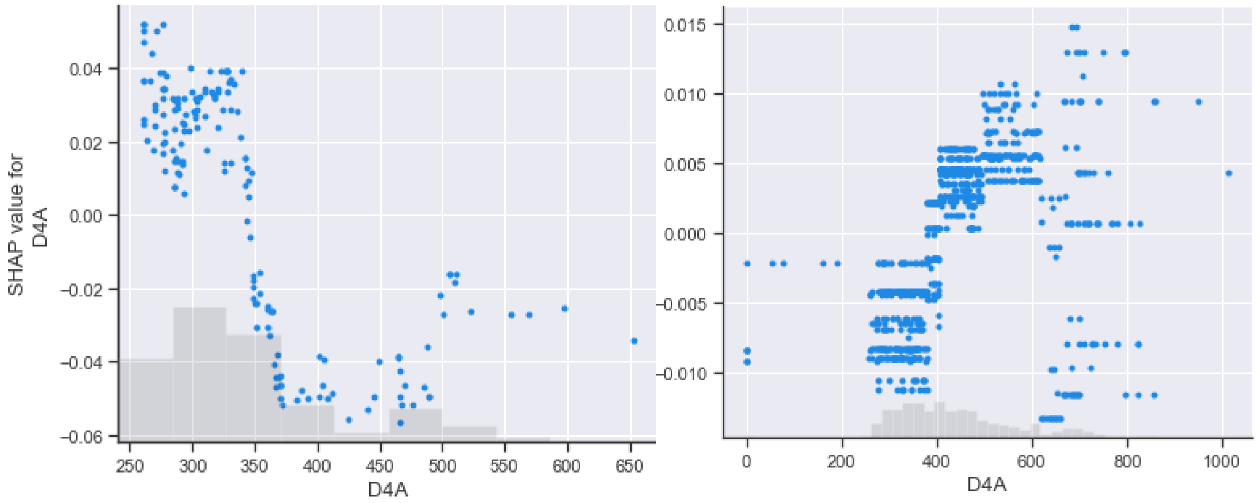


Figure 16 SHAP dependency plot showing the effect of proximity of transit stops on the prediction of the models. The plot on the left side is from the e-scooter model and the plot on the right side is from the bicycle model.

As seen in Figure 16, D4A is the meaure representing the distance from the population-weighted centroid to nearest transit stop (meters). The SHAP dependency plot shows that the crash location with more transit or transit stops nearby contributes to e-scooter's severe injury whereas the severity increases with increasing distance to transit in bicycle injury classification. This might be because most e-scooter crashes occur within core cities where shared e-scooters operate.

The variables form smart location dataset are mostly correlated so only two variables, namely, D4A and Walkability index was used for modeling. The SHAP dependency plot for walkability index is mentioned in the appendix. The results for the walkability index do not show any distinct trends.

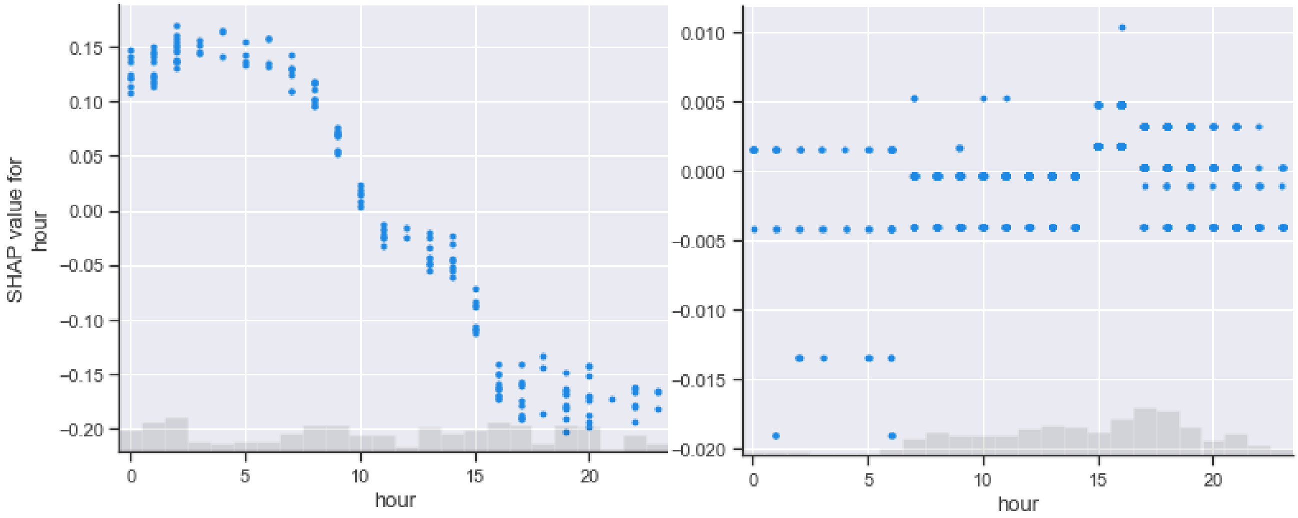


Figure 17 SHAP dependency plot showing the effect of hour of the day on the prediction of the models. The plot on the left side is from the e-scooter model and the plot on the right side is from the bicycle model.

The impact of hour of the day and injury severity is not clear in case of bicycle crash whereas the trend is clear in e-scooter crashes. In can be infered that crashes occuring at midnight and hours after midnight has high contribution in classifying severe injury in e-scooter crashes.

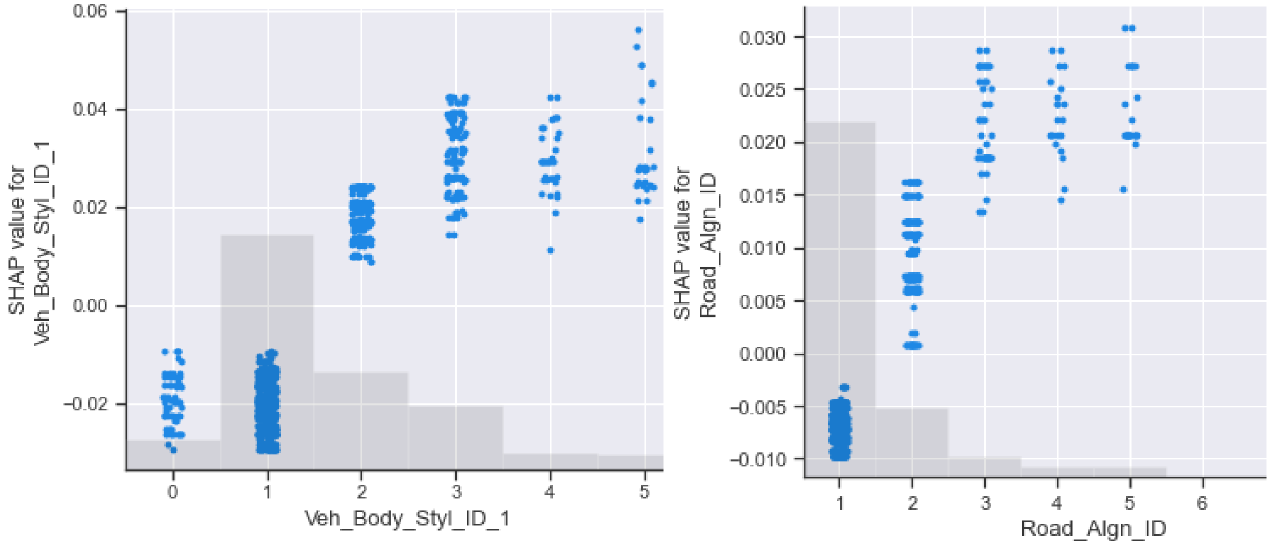


Figure 18 SHAP dependency plot showing the effect of vehicle body type on the prediction of the models. The plot on the left side is from the e-scooter model and the plot on the right side is from the bicycle model.

Vehicle body type is not a significant feature in e-scooter dataset for injury severity classification and was not used in e-scooter model. In case of bicycle, crashes involving SUV, Pickup, Van, and Heavy vehicle (2,3,4,5) are more indicative of severe injury severity than passenger vehicle (1). Similarly, Figure 19 shows the impact of bicycle facilities in injury severity classification. Bicycle facilities doesnot have an impact on e-scooter injury severity classification but has some impact on bicycle classification. Bike facility type 3 which is protected bike lane has negative impact on severe crash injury classification.

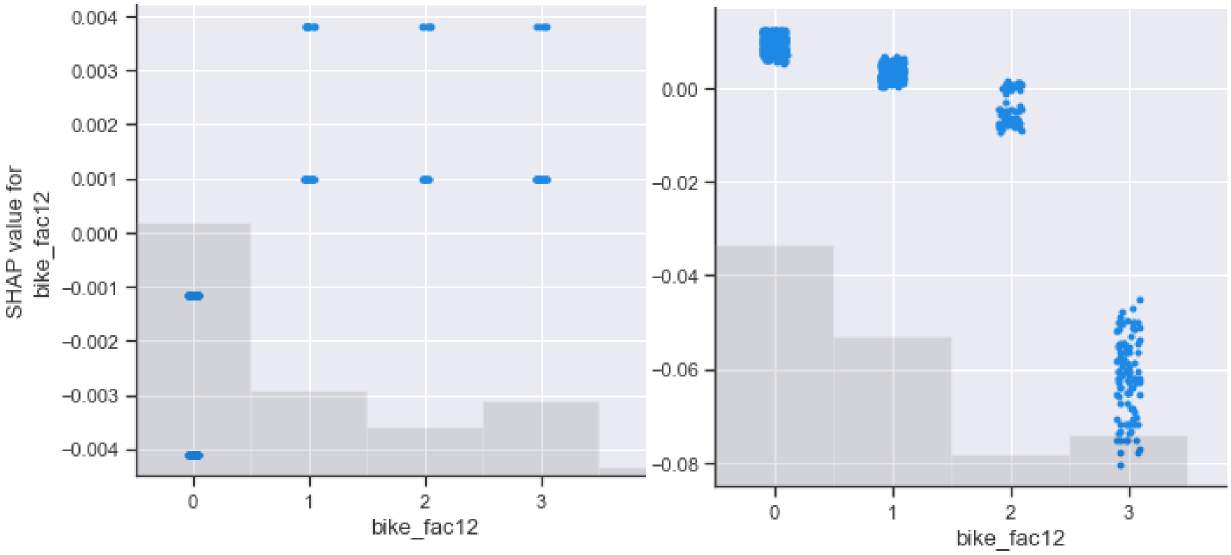


Figure 19 SHAP dependency plot showing the effect of bicycle facility on the prediction of the model. The plot on the left side is for the e-scooter model and the plot on the right side is for the bicycle model.

### 5.4.1 Demographic

Gender of the rider does not have any impact on bicycle crash injury whereas it does have an impact on e-scooter injury severity classification. Figure 20 shows that gender id 1, i.e., male has a higher contribution in classifying a datapoint as severe injury severity. Male (1) contributes in classifying a crash injury as severe whereas Female (2) contributes in classifying a crash injury as non-severe.

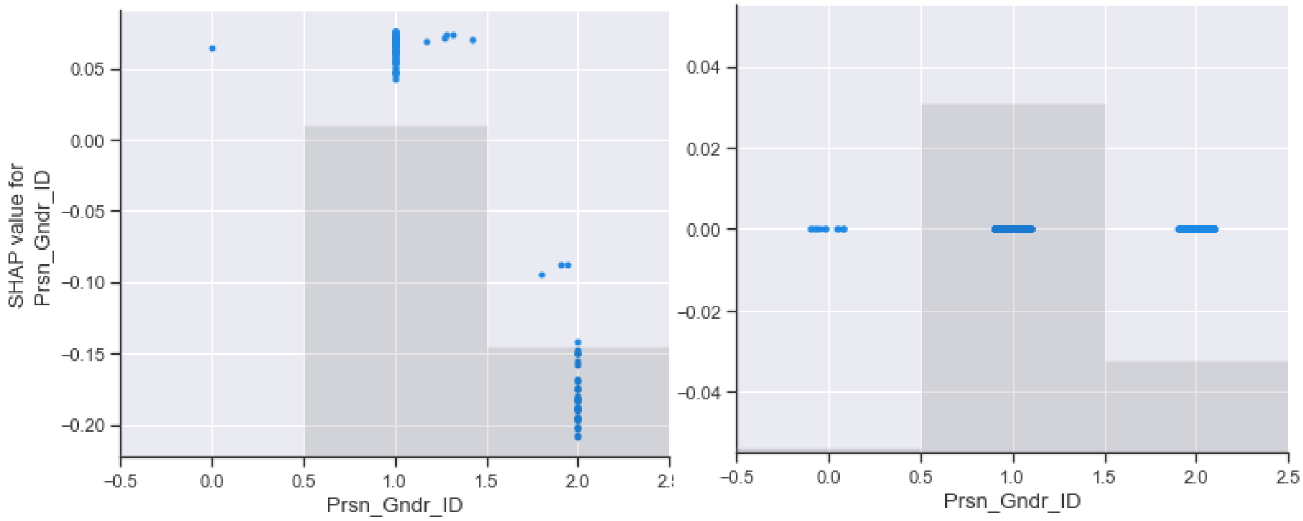


Figure 20 SHAP dependency plot showing the effect of gender on the prediction of the model. The plot on the left side is for the e-scooter model and the plot on the right side is for the bicycle model.

Ethnicity has a higher impact on classification of injury severity in e-scooter related crashes compared to crashes related to bicycles. The code 0 and 1 id for unknown and White population whereas 2 represent Hispanic, 3 Black, 4 Asian and 5 Indian American. Figure 21 shows that non-white ethnicity are more involved in severe crashes involving e-scooters.

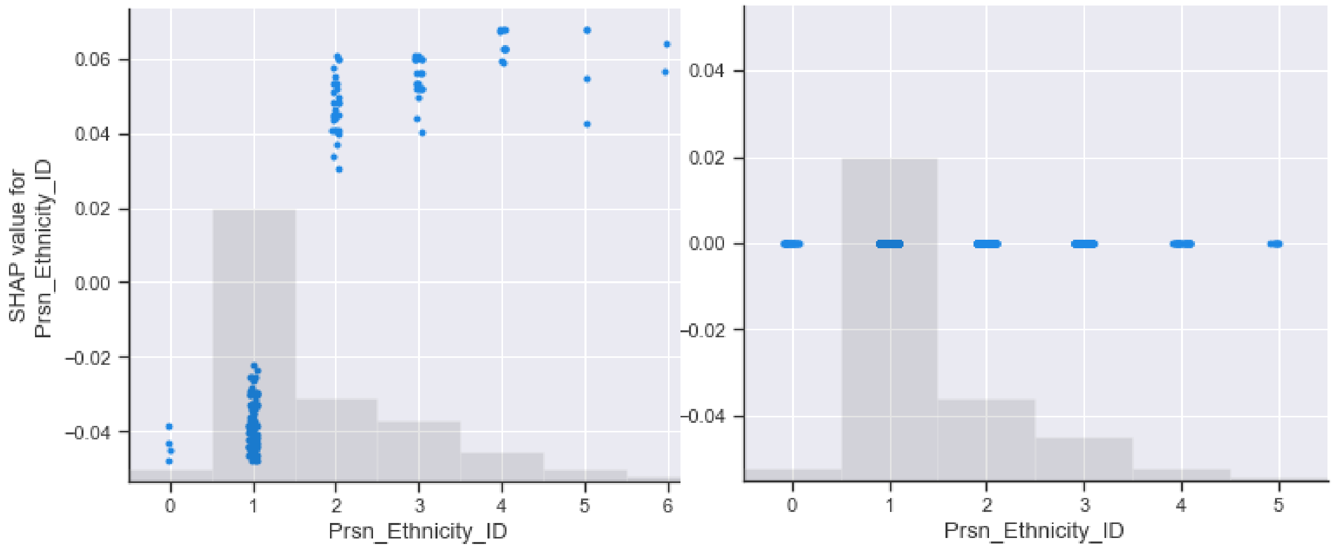


Figure 21 SHAP dependency plot showing the effect of ethnicity on the prediction of the model. The plot on the left side is for the e-scooter model and the plot on the right side is for the bicycle model.

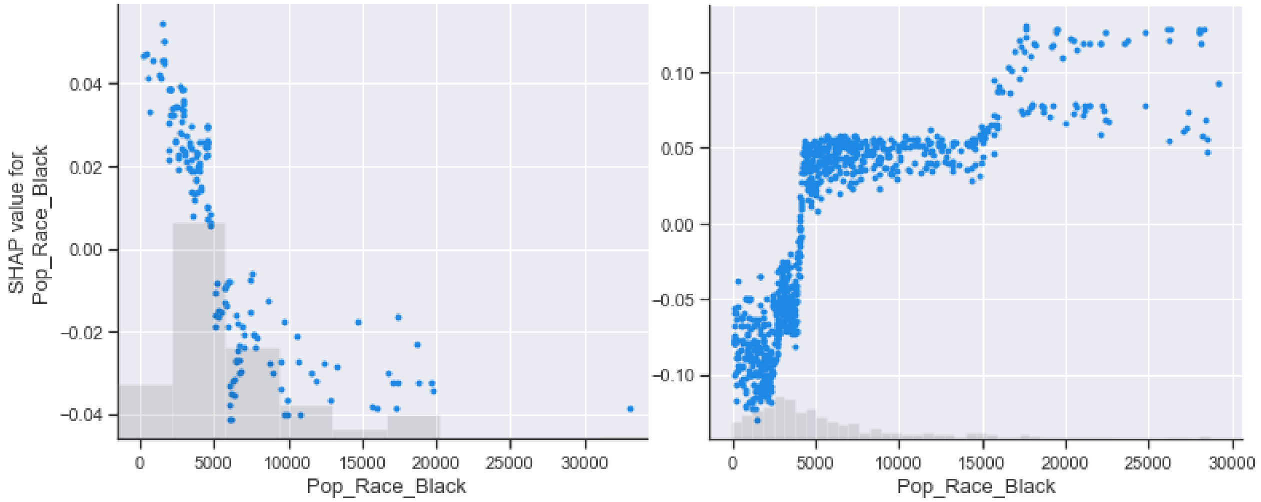


Figure 22 SHAP dependency plot showing the effect of total black population per 1 mile around the crash location on the prediction of the model. The plot on the left side is for the e-scooter model and the plot on the right side is for the bicycle model.

Comparing the population of black population within 1 mile radius the crash location, we see the relation between severe injury and black population is inverse in both cases. Locations with higher black population is less likely to have severe injury in e-scooter related crashes whereas locations with higher black population is more likely to have severe injury in bicycle related crashes.

The shap value are also higher in bicycle crash classification than e-scooter crash, indicating this variable's impact is higher in bicycle crash injury severity classification than e-scooter injury classification.

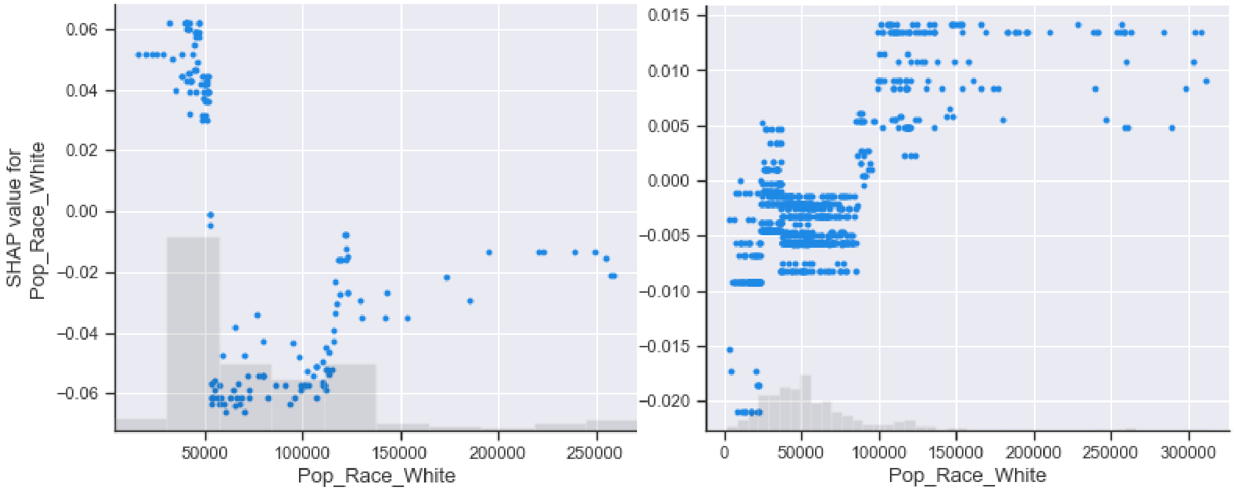


Figure 23 SHAP dependency plot showing the effect of total white population per 1 mile around the crash location on the prediction of the model. The plot on the left side is for the e-scooter model and the plot on the right side is for the bicycle model.

The SHAP value of white population within 1 mile radius of crash location is not as important as black population for bicycle crashes whereas it is more impactful than black population variable in case of e-scooter crash injury classification. Location with white population less than 50000 has higher impact on classifying severe injury in e-scooter crashes whereas it is more ambiguous in bicycle crash injury classification.

### 5.4.2 Topic

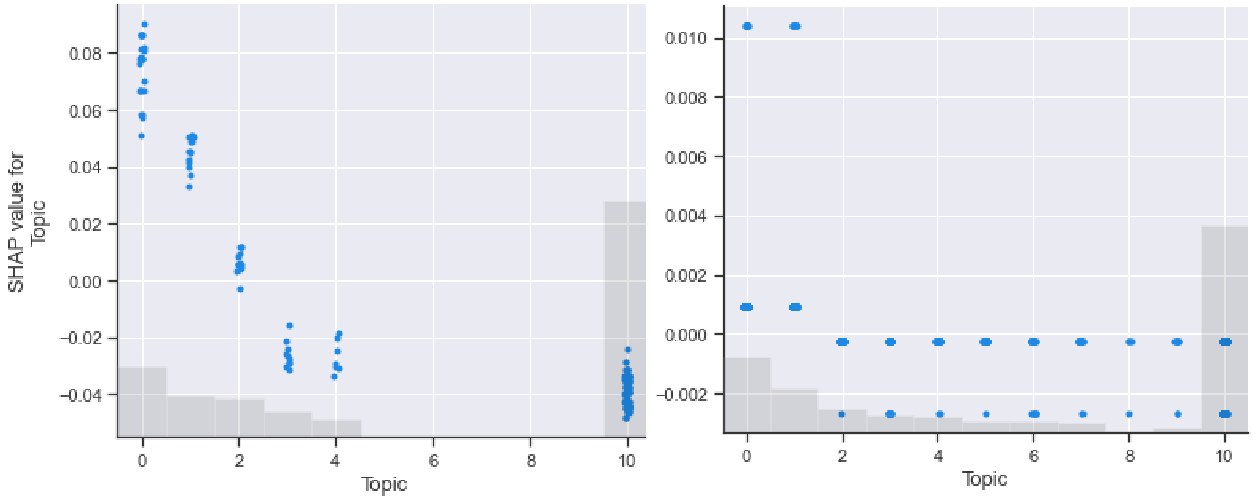


Figure 24 SHAP dependency plot showing the effect of topic on the prediction of the model. The plot on the left side is for the e-scooter model and the plot on the right side is for the bicycle model.

The impact of topic in injury severity classification is not mush in bicycle crashes whereas some trend can be observed in e-scooter related crashes. Topic-0, and Topic-1 has an impact in classifying crash injury as severe injury. The common words in Topic-0 are right, turn, state, lane, travel, drive, front and private. Combining the words using domain knowledge we can see some pattern such as right-turn private-drive and front. Manually analyzing police reports with these key words shows a clear patter. Vehicles making a turn from private drive gets in crash with e-scooter hitting the e-scooter by the vehicle’s front portion.E-scooters are driving in pedestrian sidewalks and vehicles are coming out from private parkings in many cases involving such cases.

Similarly, Topic-1 has common words such as light, green, intersection and turn. These words indicates that intersection related crashes where the fault lies either in not following traffic sign properly while making a turn has higher impact on classifying crash as severe.

# 6 CONCLUSION

Employing resampling techniques for an imbalanced class problem is an effective treatment method. Among the four samples evaluated, the SMOTE resampling and SMOTE-Tomek sampling methods produced the highest accuracy but when evaluated in terms of recall score SMOTE-NearMiss performed the best. The recall score increased more than 10-fold after resampling. A combination of domain knowledge and statistical methods for variable selection was used before training the models.

Topic modeling using BERT natural language processing model gave a more distinguishing model than LDA model. There were six topics selected from e-scooter dataset and eleven topics recognized from bicycle crash data using BERT. The topics were used as variables in model training to understand their impact on injury severity classification. The analysis suggests vehicles making a turn from private drive and hitting a e-scooter dring in sidewalk by the vehicle's front portion has a higher chance of getting into severe injury. Similarly, intersection related crashes where the fault lies either in not following traffic sign properly while making a turn has higher impact on classifying crash as severe.

The results from model evaluation shows that the tree-based machine learning models are more accurate in classifying crashes in terms of severity of injury of micro-mobility users, although additional comparison is necessary to compare statistical and machine learning model. Recall score from confusion matrix indicated that the XGBoost model performed the best among the three models. Some inferences derived from comparing the results of SHAP analysis for XGBoost models from both datasets are mentioned below:

- E-scooter crashes occurring in inner cities are more dangerous than crashes occurring in areas outside of inner city.
- The cost of damage has no impact on bicycle crash injury classification but has a meaningful impact on e-scooter crash injury.
- Crashes at dark hours contributes to severe e-scooter injury whereas it doesnot have a huge impact on bicycle crash injury.
- Increasing proximity from transit stops decreases the chance of severe classification in e-scooter data, which might be because most severe e-scooter crashes occur in inner cities with many transits stops as shared e-scooters are mostly available in inner cities in Austin.
- Hour of the day is not a clear indicator of injury severity in bicycle crashes, while it is in e-scooter crashes. E-scooter crashes occurring at late hours, especially after midnight hours are dangerous whereas bicycle injuries in the evening are more dangerous. This coincides with disproportionate e-scooter riders after midnight being intoxicated as seen in many studies. This result calls for serious regulatory evaluation of substance use and e-scooter ridership.
- Vehicle body type is not a significant feature in e-scooter dataset for injury severity classification but is in the case of bicycles. Crashes with large vehicles such as SUV, Van, Trucks increases the chance of severity of crash injury of bicyclist.
- Bicycle crashes occurring at curved roads and straight roads with grades contribute to severe injury severity where it does not have distinctive impact on e-scooter injury.

- Type of bike facility does not seem to impact e-scooter crash severity as much as it impacts bicycle injury severity. This might indicate e-scooters use of bicycle lanes does not impact e-scooter safety in a negative way.
- Gender of the rider has an impact on e-scooter injury severity classification as male contribute to more severe crashes, but not on bicycle crash injury.
- Ethnicity has a significant impact on the model prediction of e-scooter injury severity whereas it has negligent impact on bicycle model predictions, indicating disproportionate use and disproportionate danger of e-scooter use with respect to ethnicity.
- Total population does not show any trend in bicycle crash model but has a distinctive trend in e-scooter related crash data. Locations with lower population has more severe crash injury. Like total population the population of black population within 1 mile radius of the crash location has same characteristic for e-scooter but in case of bicycle more black population area increases the crash severity.

# 7 ACKNOWLEDGEMENT

The thesis is part of the project supported by the Department of Transportation Safety Through Disruption Grant (SAFE-D) research grant under the supervision of Dr. Ipek Sener, Research Scientist, TTI.

# 8 BIBLIOGRAPHY

[1] D. Appleyard, M. S. Gerson and M. Lintell, Livable streets, Berkeley: University of Califronia Press, 1981.

[2] R. L. Abduljabbar, S. Liyanage and H. Dia, "The role of micro-mobility in shaping sustainable cities: A systematic literature review," *Transportation Research Part D: Transport and Environment*, vol. 92, p. 102734, 2021.

[3] N. DuPuis, J. Griess and C. Klein, "Micromobility in Cities: A History and Policy Overview," National League of Cities, 2018.

[4] N. K. Namiri, A. W. Lee, G. M. Amend, J. Vargo and B. N. Breyer, "Impact of alcohol and drug use on bicycle and electric scooter injuries and hospital admissions in the United States," Trauma, 2021.

[5] H.-L. Meyer, M. D. Kauther, C. Polan, D. Abel, C. Vogel, B. Mester, M. Burggraf and M. Dudda, "E-scooter, e-bike and bicycle injuries in the same period—A prospective analysis of a level 1 trauma center," *Nature Public Health Emergency Collection*, pp. 1-10, 2022.

[6] S. A. Useche, S. O. Hern, A. Gonzalez-Marin, J. Gene-Morales, F. Alonso and A. N. Stephens, "Unsafety on two wheels, or social prejudice? Proxying behavioral reports on bicycle and e-scooter riding safety – A mixed-methods study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 89, pp. 168-182, 2022.

[7] N. Iroz-Elardo and K. Currans, "Injury Burden of Introducing E-Scooters: A Review of E-

Scooter Injury Studies Using Retrospective Review ofEmergency Department Records,2015–2019," *TRR,* vol. 2675, no. 12, pp. 1150-1159, 2021.

[8] L. M. Neuroth, K. D. Humphries, J. J. Wing, G. A. Smith and M. Zhu, "Motor vehicle-related electric scooter injuries in the US: A descriptive analysis of NEISS data," *American Journal of Emergency Medicine,* pp. 1-5, 2022.

[9] M. Aizpuru, K. X. Farley, J. C. Rojas, R. S. Crawford, T. J. Moore and E. R. Wagner, "Motorized scooter injuries in the era of scooter-shares: A review of the national electronic surveillance system," *The American Journal of Emergency Medicine,* vol. 37, no. 6, pp. 1133-1138, 2019.

[10] K. C. English, J. R. Allen, K. Rix, D. F. Zane, C. M. Ziebell, C. V. R. Brown and L. H. Brown, "The characteristics of dockless electric rental scooter-related injurys in a large U.S. city," Traffic Injury Prevention, 2020.

[11] . A. Azimian and J. Jiao, "Modeling factors contributing to dockless e-scooter injury accidents in Austin, Texas," *Traffic Injury Prevention,* pp. 107-111, 2022.

[12] J. B. Cicchino, P. E. Kulie and M. L. McCarthy, "Severity of e-scooter rider injuries associated with trip characteristics," Journal of Safety Research, 2021.

[13] J. B. Cicchino, P. E. Kuilie and M. L. McCarthy, "Injuries related to electric scooter and bicycle use in a Washington, DC,emergency department," *Traffic Injury Prevention,* pp. 401-406, 2021.

[14] S. N. F. Blomberg, O. C. M. Rosenkrantz, F. Lippert and H. C. Cristensen, "Injury from

electric scooters in Copenhagen: a retrospective cohort study," BMJ Open, 2019.

[15] Savolainen, P. T. Savolainen, F. L. Mannering, D. Lord and M. A. Quddus, "The statistical analysis of highway crash-injury severitys: A review and assessment of methodological alternatives," *Accident Analysis & Prevention,* vol. 43, no. 5, pp. 1666-1676, 2011.

[16] P. Chen and Q. Shen, "Built environment effects on cyclist injury severity in automobile-involved bicycle crashes," *Accident Analysis and Prevention*, vol. 86, pp. 239-246, 2016.

[17] C. Wang, L. Lu and J. Lu, "Statistical Analysis of Bicyclists' Injury Severity at Unsignalized Intersections," *Traffic Injury Prevention,* vol. 16, pp. 507-512, 2015.

[18] S. Wu, Q. Yuan, Z. Yan and Q. Xu, "Analyzing Accident Injury Severity via an Extreme Gradient Boosting (XGBoost) Model," *Journal of Advanced Transportation,* 2021.

[19] L. Wahab and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," *PLoS ONE,* vol. 14, 2019.

[20] M. Komol, M. Hasan, M. Elhenawy, M. Yasmin, M. Maoud, A. Rakotonirainy and F. Chen, "Crash severity analysis of vulnerable road users using machine learning," *PLoS One,* vol. 16, 2021.

[21] L. Xing, J. He, Y. Li, Y. Wu, J. Yuan and X. Gu, "Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza," *Accident Analysis & Prevention,* vol. 135, 2020.

[22] s. Stephen and J. Nathalie, "The class imbalance problem: A systematic study," *Intelligent*

*Data Analysis*, vol. 6, pp. 429-449, 2002.

[23] H. Jeong, Y. Jang, P. J. Bowman and N. Masoud, "Classification of motor vehicle crash injury severity: A hybrid approach forimbalanced data," *Accident Analysis and Prevention*, vol. 120, pp. 250-261, 2018.

[24] J. Lee and S. Kim, "Ordinal-imbalanced Data Classification through Data Reduction bySingular Value Decomposing TruncationOrdinal-imbalanced Data Classification through Data Reduction bySingular Value Decomposing Truncation," in *IISE*, 2019.

[25] C. Wang, S. Kou and Y. Song, "Identify Risk Pattern of E-Bike Riders in China Based on Machine Learning Framework," *entropy*, vol. 21, p. 1084, 2019.

[26] W. Zhang, R. Buehler, A. Broaddus and T. Sweeney, "What type of infrastructures do e-scooter riders prefer? A route choice model," *Transportation Research Part D*, vol. 94, p. 102761, 2021.

[27] H. Yang, J. Huo, Y. Bao, X. Li, L. Yang and C. R. Cherry, "Impact of e-scooter sharing on bike sharing in Chicago," *Transportation Research Part A* , vol. 154, pp. 23-36, 2021.

[28] T. Bielinski and A. Wazna, "Electric Scooter Sharing and Bike Sharing UserBehaviour and Characteristics," *Sustainability*, vol. 12, 2020.

[29] T. Bielinski and A. Wazna, "Electric Scooter Sharing and Bike Sharing UserBehaviour and Characteristics," *Sustainability*, vol. 12, p. 9640, 2020.

[30] A. V. Stray, H. Siverts, K. Melhuus, M. Enger, P. Galteland, I. Næss, E. Helseth and J. Ramm-Pettersen, "Characteristics of Electric Scooter and Bicycle Injuries After

Introductionof Electric Scooter Rentals in Oslo, Norway," *JAMA Network Open,* vol. 5, no. 8, 2022.

[31] S. A, S. H, M. K, E. M, G. P, N. I, H. E and R.-P. J, "Characteristics of Electric Scooter and Bicycle Injuries After Introduction of Electric Scooter Rentals in Oslo, Norway," *JAMA Network Open,* vol. 5, no. 8, p. e222670, 2022.

[32] U. S, O. S, G.-M. A, J. Gene-Morales, F. Alonso and A. N. Stephens, "Unsafety on two wheels, or social prejudice? Proxying behavioral reports on bicycle and e-scooter riding safety – A mixed-methods study," *Transportation Research Part F: Psychology and Behaviour ,* vol. 89, pp. 168-182, 2022.

[33] O.-J. Murros, T. Puolakkainen, A. Abio, H. Thorén and J. Snäll, "Urban drinking and driving: comparison of electric scooter and bicycle related accidents in facial fracture patient," *Med Oral Patol Oral Cir Bucal,* 2022.

[34] F. D. Grill, C. Roth, M. Zyskowski, A. Fichter, M. Kollmuss, H. Stimmer, H. Deppe, K.-D. Wolff and M. Nieberler, "E-scooter-related craniomaxillofacial injuries compared with bicycle-related injurieseA retrospective study," *Journal of Cranio-Maxillo-Facial Surgery,* vol. 50, pp. 738-744, 2022.

[35] A. K. Huemer, E. Banach, N. Bolten, S. Helweg, A. Koch and T. Martin, "Secondary task engagement, risk-taking, and safety-related equipment use in German bicycle and e-scooter riders – An observation," *Accident Analysis and Prevention,* vol. 172, p. 106685, 2022.

[36] M. Dozza, A. Violin and A. Rasch, "A data-driven framework for the safe integration of

micro-mobility intothe transport system: Comparing bicycles and e-scooters in field trials," *Journal of Safety Research,* vol. 81, pp. 67-77, 2022.

[37] D. Bodansky, M. Gach, M. Grant, M. Solari, N. Nebhani, H. Crouch-Smith, M. Campbell, J. Banks and G. Cheung, "Legalisation of e-scooters in the UK: the injury rate and pattern issimilar to those of bicycles in an inner city metropolitan area," *Public Health,* vol. 206, pp. 15-19, 2022.

[38] J. B. Cicchino, P. E. Kulie and M. L. McCarthy, "Injuries related to electric scooter and bicycle use in a Washington, DC,emergency department," *Traffiic Injury Prevention,* vol. 22, no. 5, pp. 401-406, 2021.

[39] N. Haworth, A. Schramm and D. Twisk, "Comparing the risky behaviours of shared and private e-scooter and bicycle riders in downtown Brisbane, Australia," *Accident Analysis and Prevention,* vol. 152, p. 105981, 2021.

[40] K. Kazemzadeh and F. Sprei, "Towards an electric scooter level of service: A review and framework," *Travel Behaviour and Society,* vol. 29, pp. 149-164, 2022.

[41] P. T. Savolainen, F. L. Mannering, D. Lord and M. A. Quddus, "‘The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives," *Accident Anal. Prevention,* vol. 43, no. 5, pp. 1666-1676, 2011.

[42] J. Zhang, Z. Li, Z. Pu and C. Xu, "Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods," *IEEE Access,* vol. 6, p. 60079, 2018.

[43] K. Santos, J. P. Dias and C. Amado, "A literature review of machine learning algorithms for crash injury severity prediction," *Journal of Safety Research*, vol. 80, pp. 254-269, 2022.

[44] A. Jamal, M. Zahid, M. T. Rahman, H. M. Al-Ahmadi, M. Almoshaogeh, D. Farooq and M. Ahmad, "njury severity prediction of traffic crashes with ensemble machine learning techniques: a comparative study," *International Journal of Injury Control and Safety Promotion*, vol. 28, no. 4, pp. 408-427, 2021.

[45] "Social Explorer," 2022. [Online]. Available: http://www.socialexplorer.com/pub/reportdata/HtmlResults.aspx?reportid=R13167603.

[46] C. o. Austin, "Land Use Inventory Detailed," 2022. [Online]. Available: https://data.austintexas.gov/Locations-and-Maps/Land-Use-Inventory-Detailed/fj9m-h5qy.

[47] J. Jacobs, The Death and Life of Great American Cities, NY: Random House, 1961.

[48] J. Chapman, E. H. Fox, W. Bachman, L. D. Frank, J. Thomas and A. R. Reyes, "SMART LOCATION DATABASE TECHNICAL DOCUMENTATION AND USER GUIDE," U.S. Environmental Protection Agency, Washington, D.C., 2021.

[49] Y. M. Goh and C. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques," *Accident Analysis & Prevention*, vol. 108, pp. 122-130, 2017.

[50] M. Baak, R. Koopman, H. Snoek and S. Klous, "A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics," *arXiv*, 2019.

[51] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research* , vol. 16, pp. 321-357, 2002.

[52] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM,* vol. 18, no. 9, pp. 509-517, 1975.

[53] S. M. Omohundro, "Five Balltree Construction Algorithms," *International Computer Science Institute,* 1989.

[54] H. Han, W.-Y. Wang and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *Advances in Intelligent Computing,* pp. 878-887, 2005.

[55] H. He, Y. Bai, E. A. Gracia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," *IEEE,* 2008.

[56] L. Breiman, "Random Forest," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[57] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics,* pp. 1189-1232, 2001.

[58] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *CoRR,* 2016.

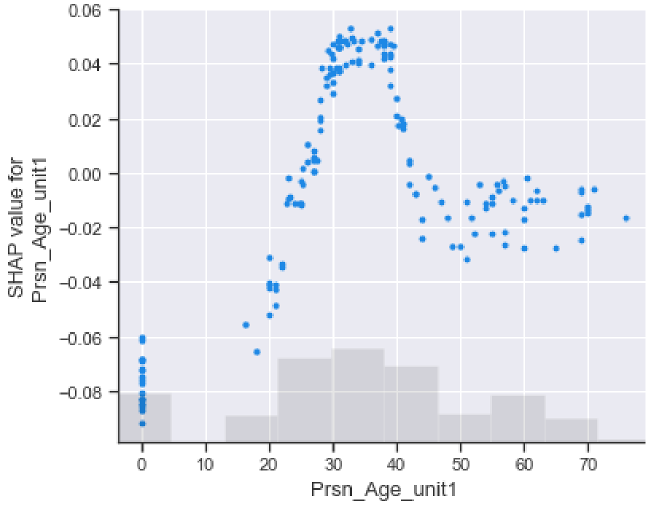
[59] E. Strumbelj and K. Igor, "Explaining prediction models and individual predictions with feature contributions.," *Knowledge and information systems,* vol. 41, no. 3, pp. 647-655, 2014.

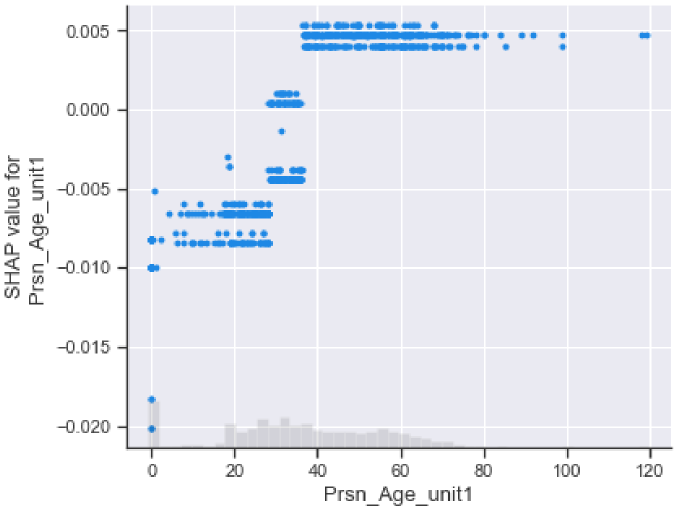
[60] S. Sarkar, A. Pramanik, J. Maiti and G. Reniers, "Predicting and analyzing injury severity:

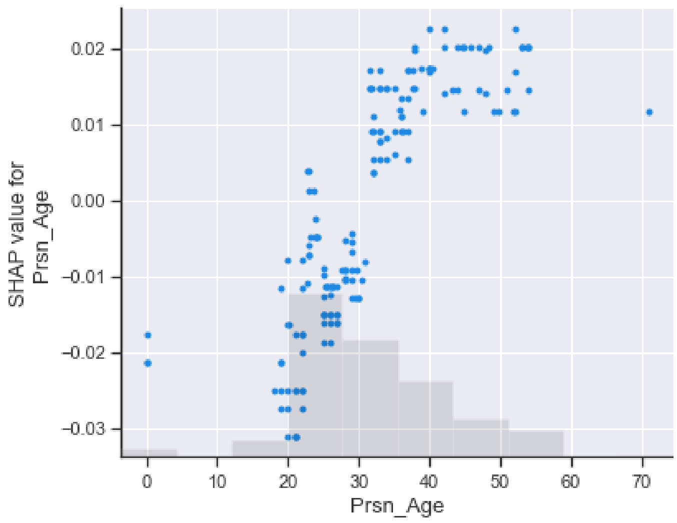
A machine learning-based approach using class-imbalanced proactive and reactive data," *Safety Science*, vol. 125, p. 104616, 2020.

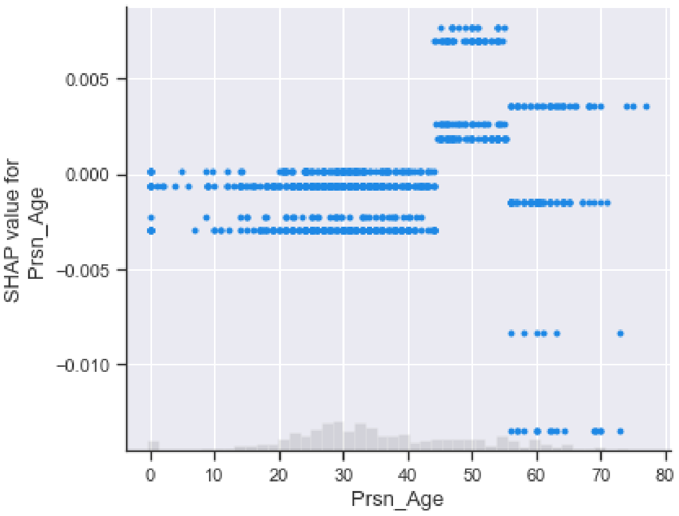
# 9 APPENDIX

## 9.1 Age SHAP dependency plots

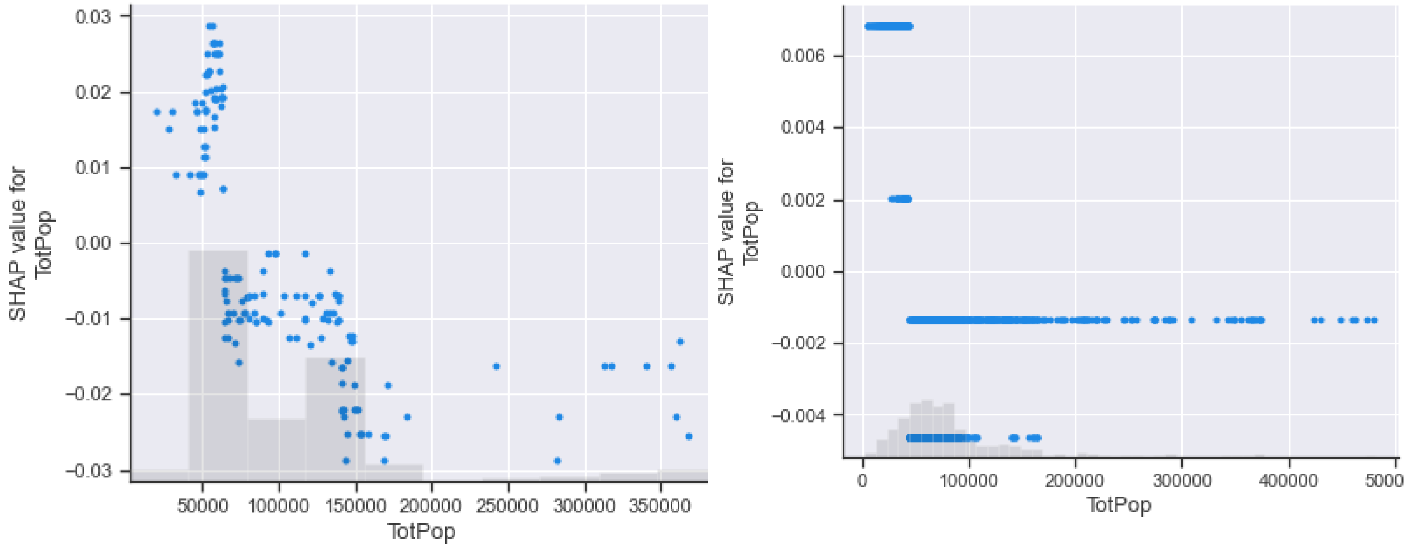




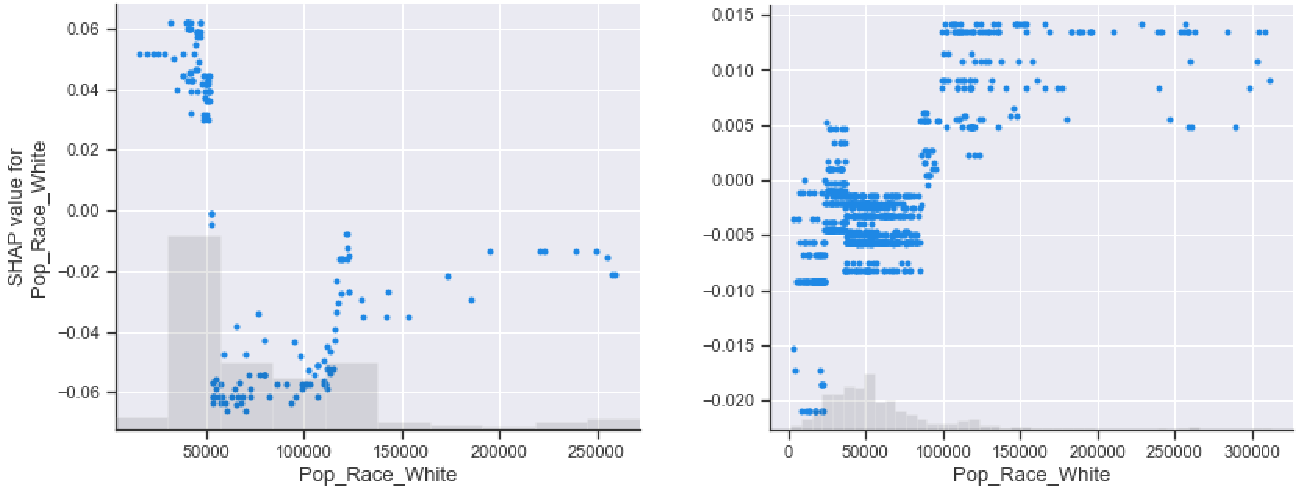




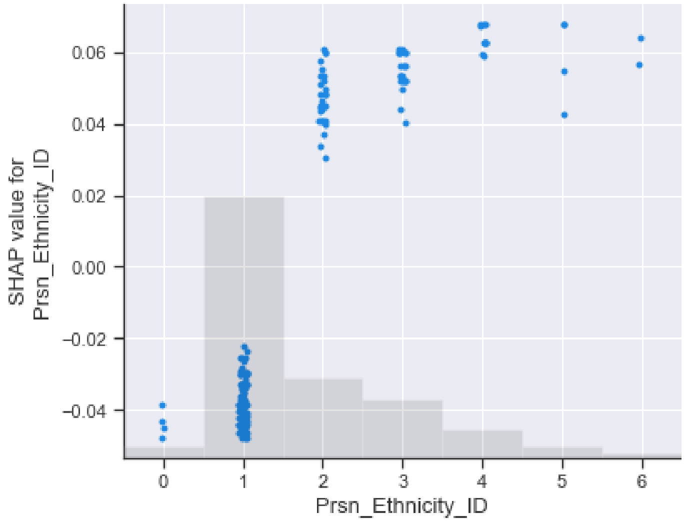
## 9.2 Total population

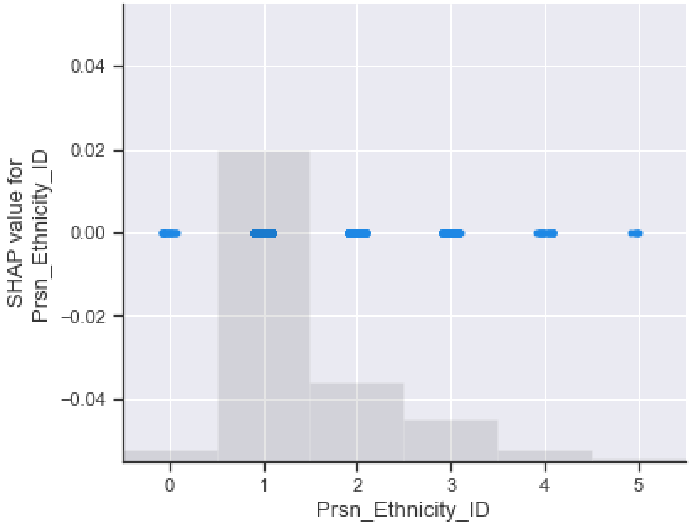


## 9.3 White Population

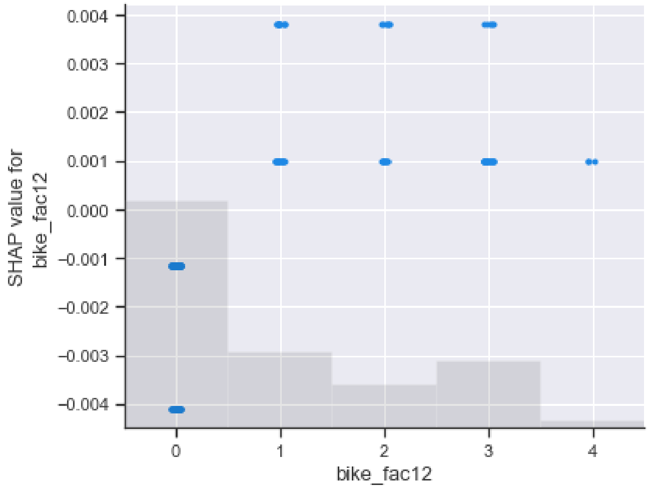


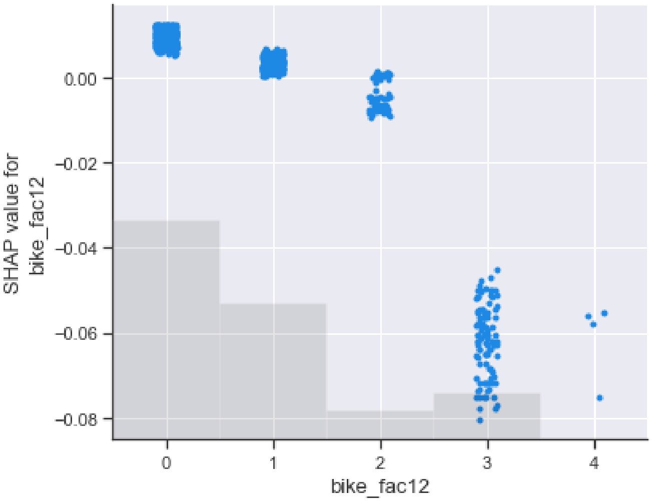
## 9.4 Ethnicity



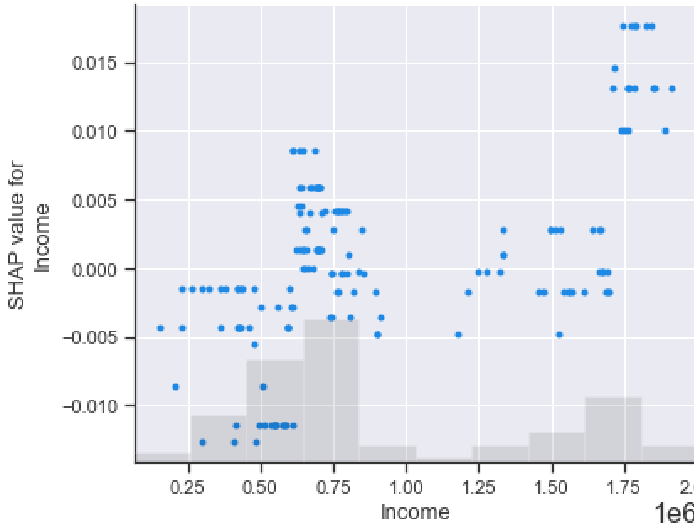


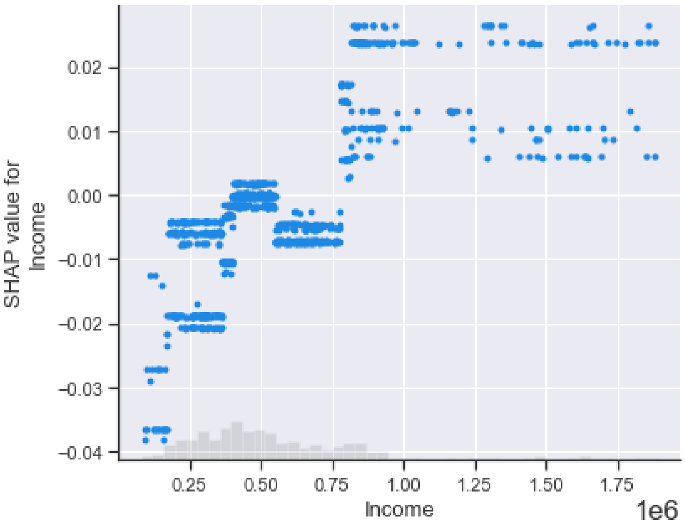
## 9.5 Bike facility



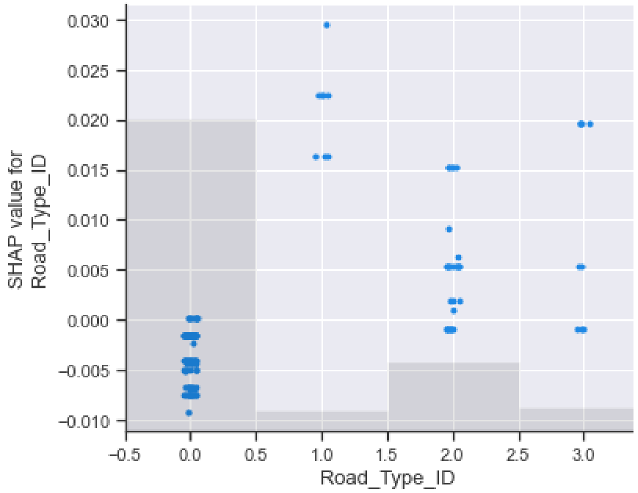


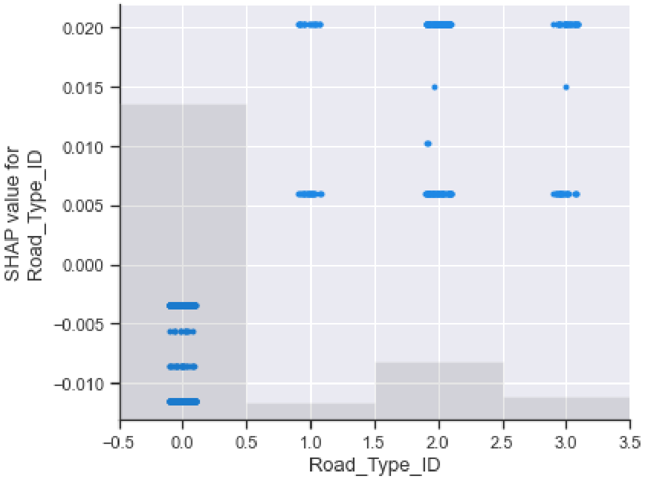
## 9.6 Income



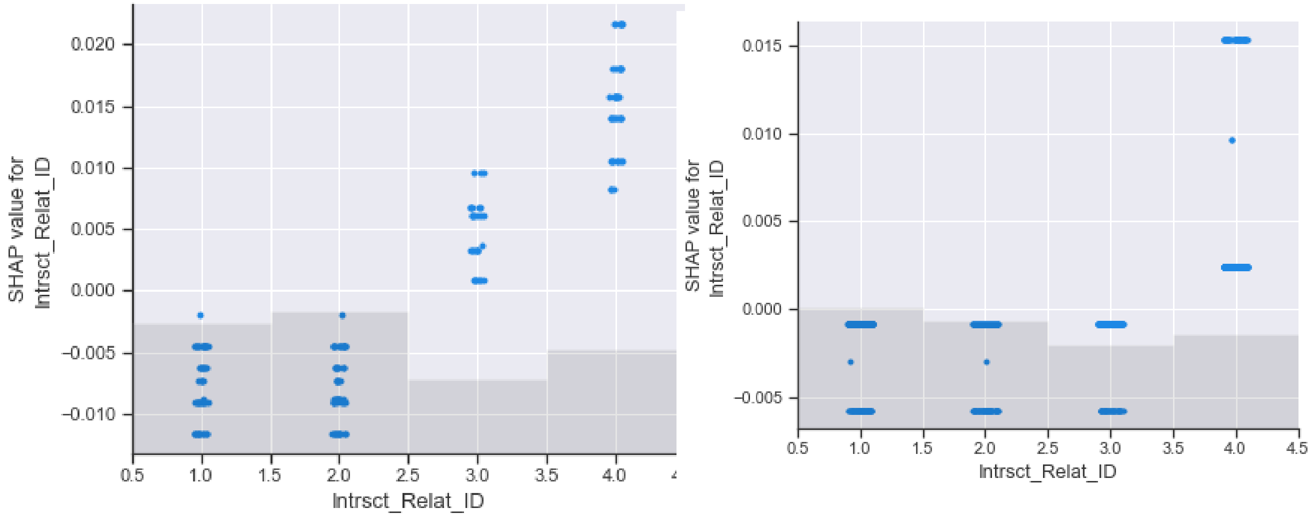


## 9.7 Road type id

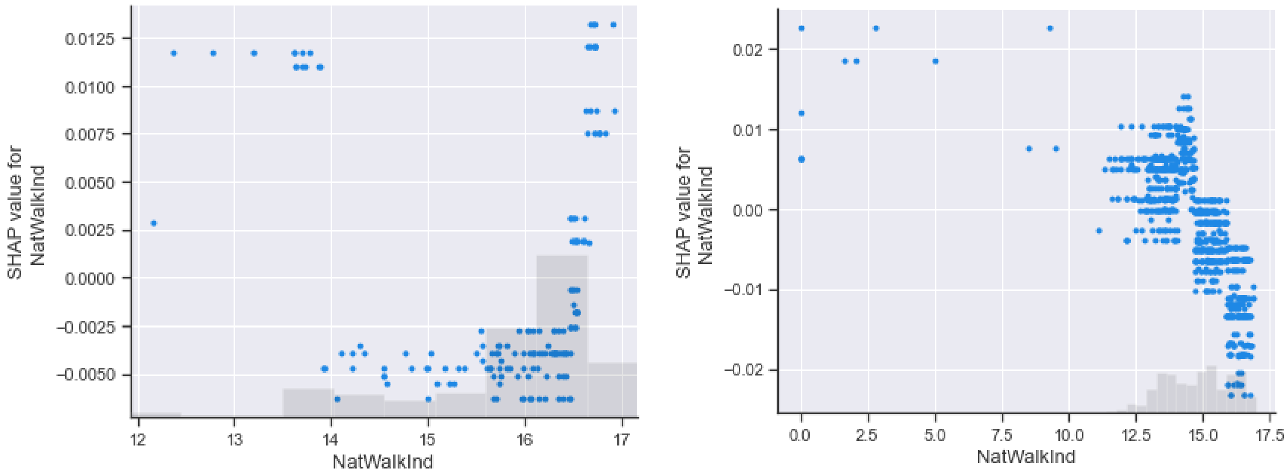




## 9.8 Intersection



## 9.9 Walkability index


SHAP value for
NatWalkInd
NatWalkInd
SHAP value for
NatWalkInd
NatWalkInd