University of California Davis

STA 208 - Statistical Machine Learning

Final Project Report

Using Machine Learning to Advance Policy Implementation: Vision Zero Los Angeles/San Francisco, California

Authors
Weijing Wang
Richard Ly
Zineb Sordo

Advisor
Dr. Bo Y.-C. Ning

April 26, 2023

Contents

1 Introduction

In 2021, the US witnessed 42915 traffic fatalities on roads (NHTSA, 2022), accounting for a more than 10 percent increase from 2020. This striking number of traffic fatalities makes 2021 the deadliest year in the past 16 years. Over years, those traffic fatalities on California streets account for about 10 percent of the nationwide traffic fatalities (3,606 out of 36,096 in 2019, NHTSA). And many more were injured and/or experienced losses of loved ones and/or properties to varying degrees in traffic crashes. Traffic crashes indeed become a public health imperative and is in great need of policy and funding intervention.

Striving for reducing traffic fatalities and serious injuries across the country, the Vision Zero safety policy has gained its momentum in the U.S. since its entry in 2012. The city of Chicago was the first US city to adopt the Vision Zero policy in 2012, and many localities joined the Vision Zero communities in the past decade. To date, more than 50 localities have committed to Vision Zero with a goal of eliminating traffic fatalities and severe injuries. Among those Vision Zero communities, about one-third (i.e., 14 cities or counties) are located in California.

With increasing numbers of traffic fatalities across the country, the federal and state agencies have released a number of transportation safety policies and funding sources in supportive of the implementation of Vision Zero at localities. At the federal level, in February 2022, the US DOT released Secretary of Transportation Pete Buttigieg's National Roadway Safety Strategy and officially committed to a national traffic safety goal of zero deaths and serious injuries (US DOT, 2022). Shortly afterwards, through H.B.3684-Infrastructure Investment and Jobs Act, the US DOT released the Safe Streets and Roads for All (SS4A) grant program providing 5 billion dollars over a five-year time period, to support localities in working towards zero traffic deaths and serious injuries (US DOT, 2022).

At the state level, California has joined this nationwide traffic safety campaign and committed to zero traffic fatalities and severe injuries through senate bills, policy initiatives, and funding resources. In February 15, 2022, California DOT (Caltrans) officially announced the goal of zero traffic fatalities and serious injuries by 2050 through the Director's policy letter. California SB-932 (Bicycle and Pedestrian Safety), recently passed in May 2022, signals a significant step in improving traffic safety for bicyclists and pedestrians. Equally important, California AB-2336 (Vehicle: Speed Safety Pilot Program) and AB-2664 (Pedestrian Crossing Signals) provide additional policies and safety strategies in supportive of the state's goal of zero traffic fatalities and serious injuries.

At the locality level, increasing numbers of cities and/or counties in California have adopted the Vision Zero policy, with a goal of eliminating traffic fatalities and severe injuries within their jurisdictions (Vision Zero Network, 2022). Since Vision Zero San Francisco established in 2013, to date, nearly 15 cities and/or counties in California, including City of Los Angeles, have adopted the Vision Zero safety policy, and many more localities are in preparation of adopting the Vision Zero policy (Vision Zero Network, 2022).

Thus, in supportive of the national and statewide goal of zero traffic fatalities and serious injuries, one motivation of conducting the current study is to promote the policy implementation for the Vision Zero communities and prepare California localities for adopting the Vision Zero policy. With those policy and funding commitments to working towards eliminating traffic fatalities and serious injuries, it is crucial for transportation professionals, stakeholders, and elected officials to be informed of the impact of policy intervention and the potential concrete approaches and strategies in improving the efficiency of policy implementation, including the Vision Zero policy. And the current project aims to make contributions from such aspects.

In doing so, we ask two research questions: 1) Whether and to what extent the adoption of the Vision Zero policy is effective in reducing traffic fatalities and serious injuries? 2) Comparing and identifying the machine learning models that are most suitable and/or powerful in predicting injury severity in traffic crashes. To answer those two research questions, the remainder of the project is organized in the following sections: Literature Review, Study Context and Data Sources, Exploratory Data Analysis, Data Preparation, Predictive Modelling, Results and Findings, Implications, and Discussions.

2 Literature Review

2.1 What is the Vision Zero Safety Policy?

Vision Zero for road safety, invented by the Swedish parliament, was firstly adopted as a traffic safety concept by the Riksdag in 1997 (Swedish Parliament, 1997). The primary goal of Vision Zero is to work towards eliminating traffic fatalities and serious injuries (Swedish Parliament, 1997). In doing so, one essential philosophy of Vision Zero is to emphasize a shared responsibility between road users and street designers (Swedish Parliament, 1997). To understand the impact and effectiveness of the Vision Zero policy, a number of studies have been conducted in European countries (Johansson, 2009; Värnild, Tillgren, Larm, 2020; Värnild, Belin, Larm, Tillgren, 2019; Värnild, Belin, Tillgren, 2016). In general, evidence drawn from those prior studies suggests a relationship between the adoption of Vision Zero and a declining trend in traffic fatalities and serious injuries (Johansson, 2009; Värnild, Belin, Larm, Tillgren, 2019), suggesting the success of the policy intervention.

In the US, Vision Zero has grown rapidly since its first entry in 2012, and the policy has gained its momentum in the last few years (Vision Zero Network), partially due to increasing numbers of traffic fatalities across the country (NHTSA). Since its first adoption in 2012 by the city of Chicago, to date, more than 50 localities across the country have joined the Vision Zero communities (Vision Zero Network). In 2022, the US DOT released the National Roadway Safety Strategy report in supportive of Vision Zero through a nationwide traffic safety goal of zero traffic deaths and serious injuries (US DOT, 2022), signaling further policy and funding commitments to the implementation of Vision Zero policy.

However, recent news and media reports on increasing numbers of traffic fatalities in those Vision Zero communities call for an immediate action on understanding the impact of the Vision Zero policy and identifying concrete approaches and strategies in effectively and efficiently implementing the policy. For instance, in April 2022, one of the well-recognized transportation/planning news website, Planetizen, released an article, asking "Vision Zero Is Largely a Failure in the United States. Why?" (Lonescu, 2022). Earlier of the year, Los Angeles Times released another news article, titled "Hundreds died in L.A. traffic crashes in 2021. Is Vision Zero a failure?" (Smith, 2022), as a response to seeing 2021 as the deadliest year in traffic fatalities.

Motivated by the national and local media reports, we started with a review of current literature on understanding the impact of Vision Zero policy and found that relevant studies remains understudied and are particularly limited with a focus on the U.S. setting.

Thus, our project aims to take a first step in filling this gap and providing some evidence on the topic. Specifically, the purposes of the current project include 1) understanding the impact of the Vision Zero policy on eliminating traffic fatalities and serious injuries and 2) identifying the targeted and concrete approaches and strategies that governments and agencies at differing jurisdiction levels can take to predict injury severity in traffic crashes and promote the policy implementation effectively and efficiently.

2.2 ARIMA in Measuring Policy Intervention

In measuring the impact of a new law and policy, autoregressive integrated moving average (ARIMA) models have been extensively used in prior policy intervention studies and present to be suitable in various fields. Some examples in the literature include the intervention impact analysis of international trade policy (Unnikrishnan Suresh, 2016), monetary policy (Bernanke et al., 2022), and public health (Griffin et al., 2021; Lopez Bernal et al., 2018, Schaffer et al., 2021).

The application of ARMIA models further extends to the policy analysis literature in the transportation field (Lavrenz et al., 2018; Washington et al., 2020). Specifically, prior studies using ARIMA models in understanding the impact of transportation policies range from driving licensing policies (Neyens et al., 2008), the camera enforcement at intersections (Carnis Blais, 2013; Vanlaar et al., 2014), the mandatory safe belt law (Houston Richardson, 2002; Masten, 2007; Wagenaar et al., 1988), speed limit changes (Rock, 1995; Vernon et al., 2004), to pedestrian stops and/or yield policies at crosswalks and intersections (Kweon et al., 2009). Overall, research results in current literature suggest that ARIMA models particularly lead to informative outcomes in the time series modelling (Lavrenz et al., 2018; Washington et al., 2020). Thus, we use ARIMA modelling to answer our first research question on the impact of the Vision Zero safety policy in the following analyses.

ARIMA models are suitable in our project for the following reasons: 1) Compared to cross-sectional studies, a time-series study using ARIMA models captures the underlying causal impacts/relationship and/or systematic patterns over time by eliminating noises or outliers; 2) Compared to segmented regression models and difference-in-difference (DID) models (Griffin et al., 2021), often known as before and after study designs, ARIMA has gained its popularity in evaluating policy interventions, particularly with the presence of seasonality and autocorrelation in transportation safety studies (Lavrenz et al., 2018); 3) Intervention analysis of ARIMA models is considered as one of the best designs for establishing causal effects analysis (Cook Campbell, 2002); 4) Differing from segmented regression models, ARIMA models return the outcome variables as a function of the outcome measure at previous time points instead of the time variable, making the reproducitivity of policy intervention analysis more feasible for localities and transportation agencies in practices.

2.3 Machine Learning for Injury Severity Prediction

In implementing policy interventions to address traffic safety issues, one primary step is to predict injury severity in traffic crashes given a combination of contributing factors. In doing so, identifying the concrete methods and models that are suitable and powerful in injury severity predication plays an important role in achieving the goal of zero fatalities and serious injuries. For this purpose, increasing numbers of studies compare various machine learning algorithms and identify the models that are suitable and present a high performance in predicting injury severity in differing targeted communities (Al Mamlook et al., 2019; Chakraborty et al., 2021; Wahab Jiang, 2019; Zhang et al., 2018).

Specifically, a study conducted in Michigan suggests that in classifying and predicting traffic injury severity, Random Forests (RF) and Logistic Regression (LR) models present to have a better performance with a 75.5 percent accuracy and a 74.5 percent of accuracy, respectively, compared to Naive Bayes (NB) with a 73.1 percent of accuracy (Al Mamlook et al., 2019). The study defines the predication accuracy of machine learning algorithms as the AUC value, that is Areas Under Curve (AUC) (Al Mamlook et al., 2019). Another study in Texas, found that Support Vector Machine (SVM) presents a more effective and accurate performance in predicating the occurrence of traffic crashes, compared to NB, in which the authors use the values of Mean Absolute Deviation (MAD), and Mean Squared Predictor Error (MSPE) as the model performance measures for the comparison purpose (Li et al., 2008). More recently, one study in Florida, comparing the model performance in predicating injury severity, found that compared to K-Nearest Neighbor (KNN) and Support Vector Machine (SVM), Random Forest (RF) presents to have a more accurate model performance (Zhang et al., 2018).

Building upon prior studies, we identify and select a few machine learning models that present to be effective and accurate in predicting injury severity to carry out our analyses. Specifically, in our project, we choose Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest (RF) as machine learning algorithms of interest and use K-Nearest Neighbors (KNN) as the model for the purpose of comparing predication performance of the chosen machine learning algorithms.

3 Study Context and Data Sources

3.1 Study Context

The current project aims to contribute to existing literature through 1) providing some evidence on the impact of the Vision Zero policy on addressing traffic safety concerns, 2) identifying the machine learning models that are present to be effective and accurate in predicting traffic injury severity. To do so, we identify and choose the city of Los Angeles and San Francisco as the study areas for the following reasons. First, both cities are one of the first Vision Zero communities. San Francisco officially adopted the policy in 2013, and the city of Los Angeles joined the communities in 2015. With the policy implementation in place for a seven-years time period or longer, these two cities present to be suitable to study the impact of the Vision Zero policy. Second, in 2021, nearly 300 people died in car crashes in Los Angeles, making 2021 the deadliest year in the past decade for the city. Every year, San Francisco, on average, witnessed about 30 people who were killed on the city's streets. And each year, thousands of people were seriously injured to varying degree on the streets in those two cities.

Third, to date, there are in total 14 Vision Zero communities in California, and more than 50 across the country. Thus, carrying out our project aims to provide suggestions and recommendations to the Vision Zero policy community peers in effectively implementing the policy in practices. In addition, in 2022, the Director of California Department of Transportation (Caltrans) officially committed to the Vision Zero policy, with a goal of eliminating traffic fatalities and serious injuries on California streets by 2050 (Caltrans, 2022). In addition, a recently passed bill of SB-932 (Bicycle and Pedestrian Safety) further signals the urgency of addressing traffic safety issues in California. Thus, the study results in our project are expected to reveal some evidence on the impact of policy interventions on improving traffic safety and provide suggestions to the state and localities in working towards the goal of zero fatalities and serious injuries by 2050.

3.2 Data Sources

Carrying out the current project relies on a few publicly-accessible data sources. Two main data sources used to retrieve traffic crash records are 1) the Los Angeles Open Data from City of Los Angeles and 2) the TransBase Dashboard from San Francisco. We focus on the crash victim files for the purpose of this project. We also use the websites of Vision Zero

Network, Vision Zero Los Angeles, and Vision Zero San Francisco to further understand the study context and the policy implementation of Vision Zero at the national and local levels. Also the US Census American Community Survey (ACS) 5-Year, 2019 is used to indicate/measure sociodemographic characteristics of the study areas as needed.

Specifically, the raw data of crash victim records provides a number of features that are meaningful for our analyses, for instance, collision date and time, injury severity, types of transport modes, intersections. The raw data also provides the indicators of sociodemographic characteristics of crash victims, including age, gender, and race and ethnicity. Some travel behavior features include whether crash victims drove under influence or not. Building upon prior studies, these features present to be meaningful in capturing the impact of contributing factors on the occurrence of traffic fatalities and serious injuries and sociodemographic characteristics of traffic crash victims.

4 Exploratory Data Analysis

Overall, the preliminary data in our project includes 400,633 traffic crash victim records covering the study period from January 2013 to December 2019. Those crash records are grouped into the following categories, 1) 1787 records (i.e., 0.4 percent) as Fatal Injury, 2) 9717 records (i.e., 2.4 percent) as Serious Injury, 3) 56415 records (i.e., 14.1 percent) as Visual Injury, 4) 108453 records (i.e., 27.1 percent) as Complaint of Injury, 5) 19517 records (i.e., 4.9 percent) as Complaint of Pain, and 6) 204744 records (i.e., 51.1 percent) as Others.

4.1 Summary Statistics

In this section, we start with analyzing a few features that present to be contributing factors in the literature on traffic crashes. For instance, many studies reveal that vulnerable road users including pedestrians and bicyclists, elderly populations, people with color are often over-represented in traffic fatalities and serious injuries. Building upon prior studies, we group the predictor variables indicating injury severity into two classes: 1) Fatality or Severe injury, and 2) Minor Injury or Others. Below we will present the summary statistics of those features of interest, including types of road users, and the victim's gender, age, and race, as a function of injury severity.

Overall, Figure 4.1 - 4.4 provide evidence on traffic fatalities and serious injuries are not equally distributed among crash victims with different sociodemographic characteristics. Beginning with types of road users, as shown in Figure 4.1, about one-third crash victims who died or were seriously injured were pedestrians or bicyclists. These types of victims can sometimes be difficult to see and can be more unpredictable for cars especially if they drive in the same lanes.

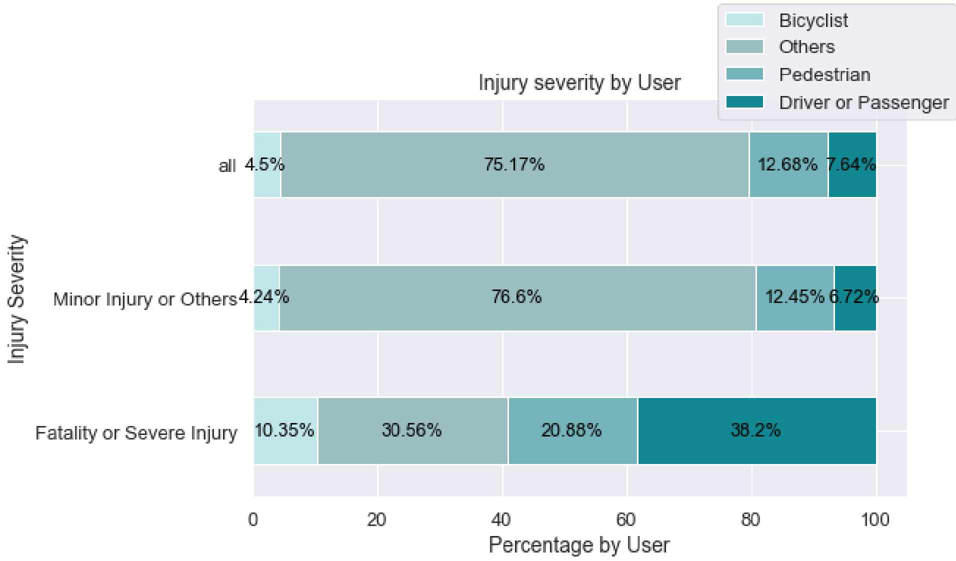


Figure 4.1: Plots of the distribution of the victim's role by severity injury.

Figure 4.2 shows that male victims were more likely to be killed or seriously injured compared to overall crash victims. Specifically, male victims accounted for 70.59 percent of fatal or severe injuries, while they made up about 60 percent of crash victims and 50 percent of the total populations in the study areas. The figure highlights the relationship (even if it may be a small one) between the victim's gender and they're injury severity in a potential collision crash.

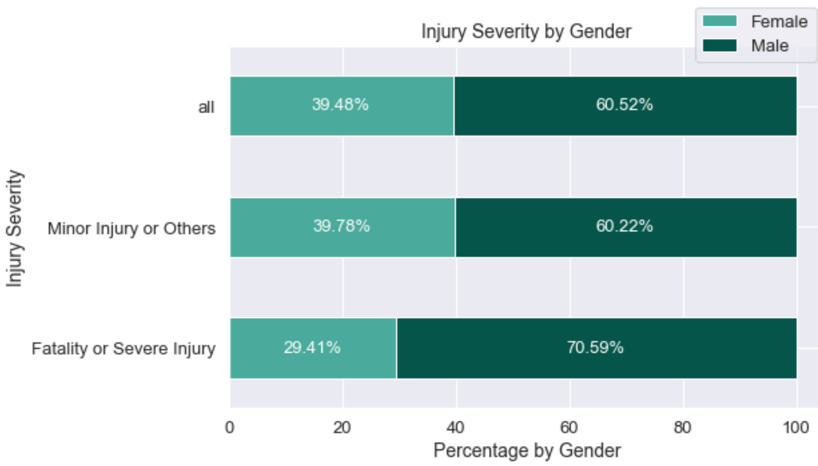


Figure 4.2: Plots of the distribution of the victim's sex in the dataset

Figure 4.3 shows that compared to their young counterparts, crash victims who are aged 65 years old or above, were much more likely to be killed or seriously injured. Specifically, elderly populations accounted for 13.27 percent of fatal or severe injuries, while only 8.5 percent of crash victims were aged 65 years old or above. In practices, when an elderly victim is in a collision, they are much more likely to end up with a severe (or fatal) injury.

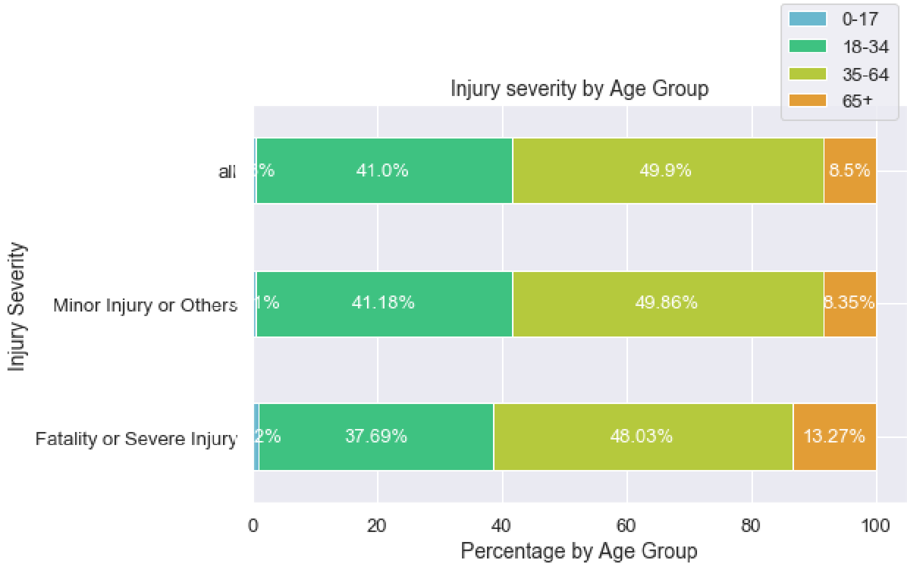


Figure 4.3: Plots of the distribution of the victim's age in the dataset

Similarly, in Figure 4.4, Black populations in traffic crashes accounted for 17.49 percent of fatalities and serious injuries while they made up about 16.54 percent of all crash victims. Yet Black populations only account for less than 10 percent of the total populations in Los Angeles and about 6 percent of the total populations in San Francisco. In addition, we notice that in Los Angeles and San Francisco, nearly 50 percent of crash victims are Hispanic populations, signaling traffic safety concerns among Hispanic residents.

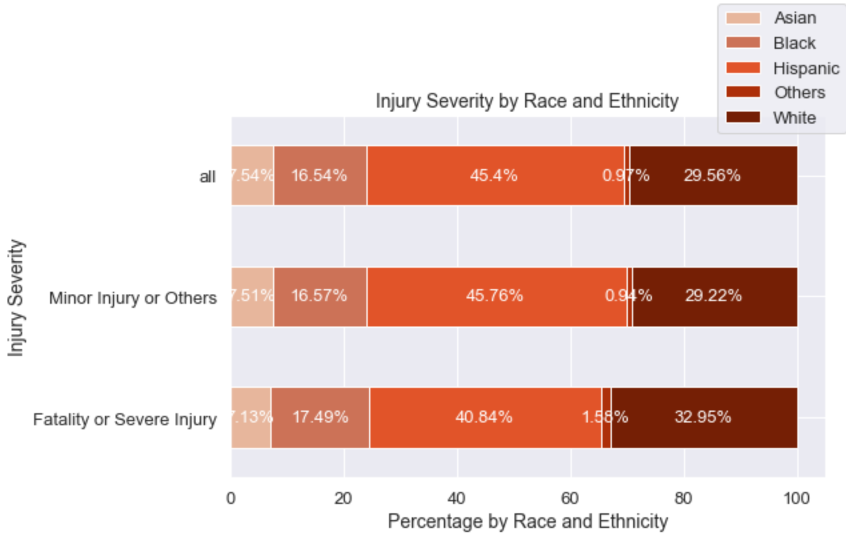
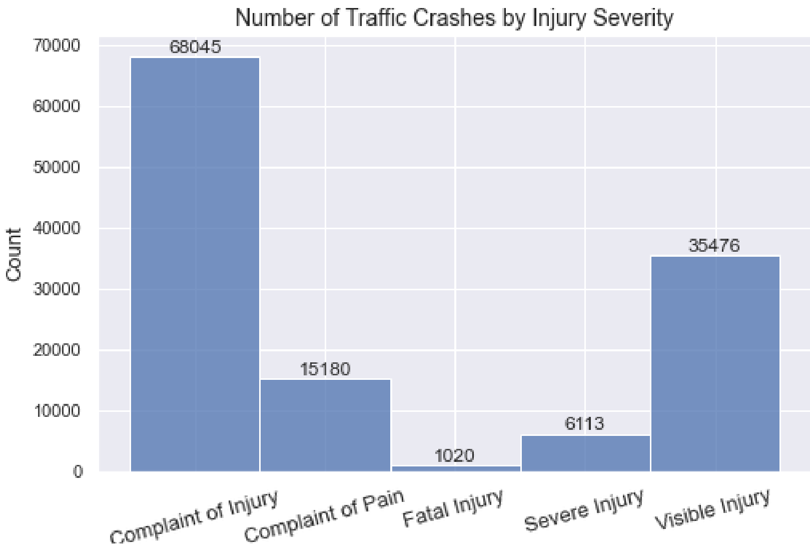


Figure 4.4: Plots of the distribution of the victim's race in the dataset

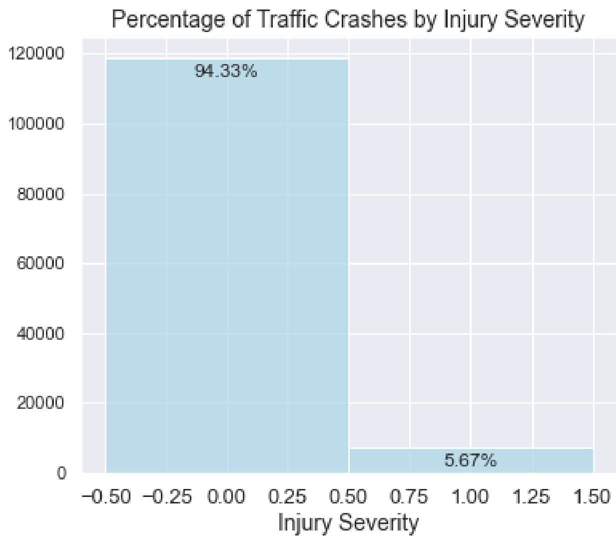
4.2 Balance Analysis

To answer our first research question on policy intervention analysis, we particularly focus on traffic fatalities and serious injuries. Because the primary goal of the Vision Zero policy is to eliminate fatalities and serious injuries (US DOT; Vision Zero Network). Thus, it is reasonable to expect the impact of adopting Vision Zero on the occurrence of fatalities and serious injuries instead of minor injuries or property damage. As a result, the crash victim records for building ARIMA models is 11505 in total.

Turning to our second research question on injury severity predication, we choose to group the crash records into two-classes: 1) Fatality and Severe Injury, and 2) Minor Injury or Pain, building upon prior studies. Specifically, the first category includes the **Fatality** and the **Sever injury**. And the second category includes the **Visible injury**, the **Complaint of injury**, and the **Complaint of pain**. Therefore, the `victim_severity_injury` variable defined as the predictor in our analyses becomes a binary target variable. In both cases, this variable had the following distributions:



(a)



(b)

Figure 4.5: Distribution of the categories of the target variable

As shown in Figure 4.5, in both cases using the count and percentage as indicators, the target variable is highly imbalanced with a striking majority of minor injuries (more than 95% in the case of binary classes). Therefore, this issue will be important to address in the predictive machine learning models, notably through methods like oversampling, undersampling and appropriate models. We will conduct the data treatment in the section below.

4.3 Feature Selection

Building upon the literature and with the results of summary statistics, we choose the following features as the variables of interest in building predicative models:

Selected features	Description
Injury Severity	Fatality and/or Severe Injury Minor Injury and/or or Pain
Gender	Male Female
Age group	Children (17-) Young Adult (18-34) Middle-aged Adult (35-64) Elderly populations (65+)
Race or ethnicity	Black Hispanic Asian White Other
User role	Drive or Passenger Pedestrian Bicyclist Others
At Intersection	Yes No
DUI (Drive under Influence)	Yes No
Population density	Continuous variable

5 Data Treatment

5.1 Data Preprocessing

In order to prepare the data for model predicting, the following steps were taken:

- Split the train and test set
- Create 2 sets of encoded predictor variables, the dummy variables for 3 of the classifiers and the Label-Encoded variables for the Random Forest Classifier.
- Scale the dummy or One-Hot encoded variables

The first step of the data preprocessing was to split the data into a train and test set in a stratified fashion by taking into account the distribution of each label in the both sets, and thus help improve imbalance when training and test our models. In addition, creating a validation set was not necessary since we used a GridSearch with a 5-fold cross-validation before training the machine learning models on the trainset and then evaluating the optimized final model on the test set. In doing so, overfitting was less likely to happen.

Then, before fitting the data to the model, we use the process of One-Hot encoding to encode the categorical predictors used in classifying the injury severity of new observations. In doing so, a dummy variable X_{ij} is created for each class j in each of the categorical predictors X_i. Here, $i = 1, \dots, p$ indexes the p predictor variables, and $j = 1, \dots, r$ indexes the number of classes for each categorical predictor variable. These dummy variables X_{ij} takes values in $\{0, 1\}$, where 1 corresponds to the observation being of the class of the dummy variable, and 0 corresponds to the observation being of a different class than that dummy variable. We choose to employ One-Hot encoding as there was significant improvement in classification accuracy over simply using one categorical variable to encompass the several classes for any one predictor variable. These One-Hot encoded features will then be used for training the classifiers Logistic Regression, SVM and KNN. Then, on the other hand, the features will be Label encoded for the Random Forest classifiers since this classifier is more adapted to multi-class categorical features. In doing so, all the categorical features that are in the *string* format will be converted to categorical multi-class numerical variables and directly fed to the model as the training set.

5.2 Oversampling and Undersampling

Next, we recall the imbalance of the original dataset that we consider. In order to circumvent the issue of an imbalanced dataset, we consider performing both random over-sampling of the minority class and random under-sampling of the majority class. Thus, we randomly duplicate observations of class 1 injury severity and we randomly delete observations of class 0 injury severity to obtain a resampled dataset. In doing this, a more balanced injury severity class distribution is obtained such that the model achieves higher accuracy in predicting the minority class: class 1 injury severity, consisting of severe and fatal injuries. This resampling method is applied to each model during a GridSearchCV while making sure that during the cross validation, the examples being trained are not repeated. To do so, we used a the Imbalance Learn package's Pipeline class where the parameters given were a resampling method, which in this case was SMOTEENN (a combination of over and undersampling) and the ML model to train. The SMOTE algorithm has been used in prior studies to address the same data imbalance issue in traffic crashes (Al Mamlook et al., 2019), which presents to be suitable in our project. Thus, the resampling is only applied during the classifier fitting and therefore avoiding overfitting the model.

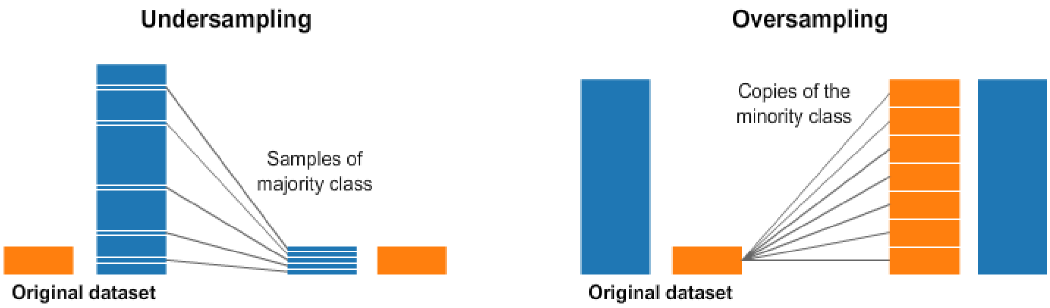


Figure 5.1: Oversampling and Undersampling data

6 Predictive Modelling

6.1 ARIMA Modelling

In this project, we first use the autoregressive integrated moving average (ARIMA) modelling to understand the impact of the Vision Zero safety policy on traffic fatalities and serious injuries. The application of ARIMA requires a time-series dataset. To prepare the modelling-ready dataset, we sub-select the crash records with a consequence of fatality or severe injury. And we then calculate the numbers of traffic fatalities and severe injuries by month from January 2013 to December 2019. The ready-data is a monthly time-series dataset. To apply the ARIMA models, the following provides a detailed description of the assumptions and equations of ARIMA modelling, and the workflow of modelling building and diagnostic measures.

6.1.1 Assumptions

Stationarity

One standard/fundamental requirement of implementing ARIMA models is stationary. Assuming stationarity builds on three properties: 1) constant mean, 2) constant variance, and 3) constant covariance solely depending on the time interval between measures. Whether the data is stationary or not is often detected by the Autocorrelation Functions (ACFs) and the Partial Autocorrelation Functions (PACF), and data-differencing presents to be suitable in addressing non-stationarity. These approaches are also used in our project.

Autocorrelation

Time series (including ARIMA) observations are often autocorrelated, in which the observations are correlated with the values at previous time points. In other words, time series observations are not interdependently distributed, which violates the standard assumptions of regression analysis. In addressing this autocorrelation issue, differencing the data is used and presents to be suitable. In addition, the ACFs and PACFs are frequently used to detect autocorrelation issues. Specifically, the ACF plot presents the correlation between one observation with its peers at previous time spots, and the previous time spots are determined by lags which is defined as the number of time points. And the PACF presents the correlation between one observation with its peers that cannot be explained by the correlations at any lower order lags.

Seasonality

The seasonality captures those variations of a pattern or trend, occurring at regular time intervals. Some examples of the seasonality include season of the year, month of the year, day of the week, time of day. In this project, we use the monthly time-series data and the ACFs and PACFs to detect the trend and/or seasonality.

6.1.2 Specifications

The purpose of ARIMA models is to return the outcome measure as a function of its measures at previous time points and the error terms. To elaborate, ARIMA models is a combined format of autoregressive (AR) models, moving average (MA) models, and integrated (also known as differencing). Below are the specifications for each component.

The specification of AR is defined as:

$$Y_t = \mathrm{c} + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \ \Phi_p Y_{t-p} + \epsilon_t$$

(6.1)

Where Y_t is the outcome measure, which is predicted by one or multiple values at previous time points of Y_t. c is a constant. θ denotes the autocorrelation between the observation and its past values. p is the number of time lags, and ϵ is the error term.

The specification of MA is defined as:

$$Y_t = \mathrm{c} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \ \theta_q \epsilon_{q-1}$$

(6.2)

Where Y_t is the outcome measure, which is predicted by error terms of one or multiple lagged values. θ denotes the autocorrelation of the error terms, and q is the number of time lags.

The specification of integrated/differencing functions is defined as:

$$Y'_t = Y_t + Y_{t-1}$$

(6.3)

In which the differencing function is used to achieve the stationarity and address the autocorrelation in the application of ARIMA model building.

6.1.3 Model Building

Below Figure 6.1 shows the workflow we use for model building and diagnostic measures. The Step 1 is essential to detect whether the data meets the assumptions of ARIMA models, and the Step 2 applies appropriate data transformation and treatment to address the issues present in the Step 1. The Step 3-5 aim to identify the appropriate model and use diagnostic measures to ensure model validity. Lastly, in the Step 6, we will present a comparison between counterfactual predication and observed data after the policy intervention.

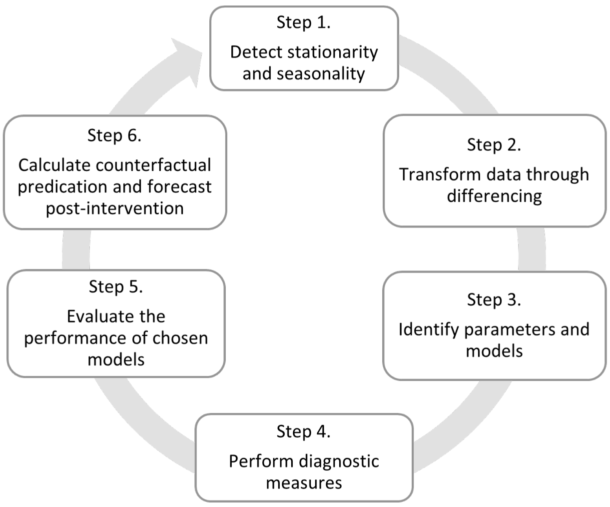


Figure 6.1: Workflow for Model Building and Diagnostic Measures

Notes: ACF = Autocorrelation Function, PACF = Partial Autocorrelation Function

Next, we introduce the results of each step that we take to meet the assumptions of implementing ARIMA models and the performance measures of the final model. Figure 6.2 is the constructed monthly time-series data, which shows the trend of traffic fatality and serious injury in Los Angeles and San Francisco from January 2013 to December 2019.

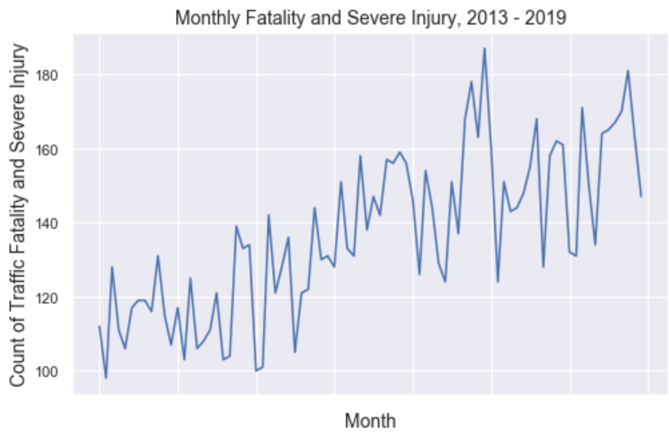


Figure 6.2: Monthly fatality and serious injury Trend, 2013-2019

Detect stationarity and seasonality

Before applying ARIMA models into the policy intervention analysis, we begin with detecting the two fundamental assumptions in modelling: stationarity and seasonality, demonstrated above. To do so, we use the autocorrelation functions (ACFs) and partial autocorrelation (PACFs). The ACF plot in Figure 6.3 - 6.4 show that many ACF values are above the autocorrelation boundaries, and the majority of ACF values are clustered on one side of the zero-benchmark line (above the line in this case). The PACF plot in Figure 4 presents a similar pattern. The results of ACFs and PACFs overall suggest that the current data is non-stationary, and data transformation and treatment are needed.

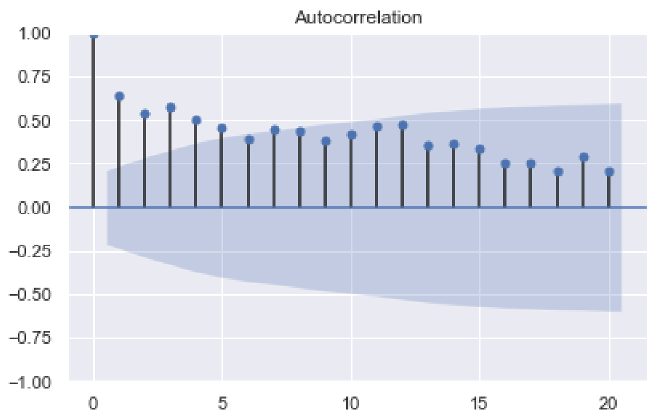


Figure 6.3: ACF Plot of Original Data

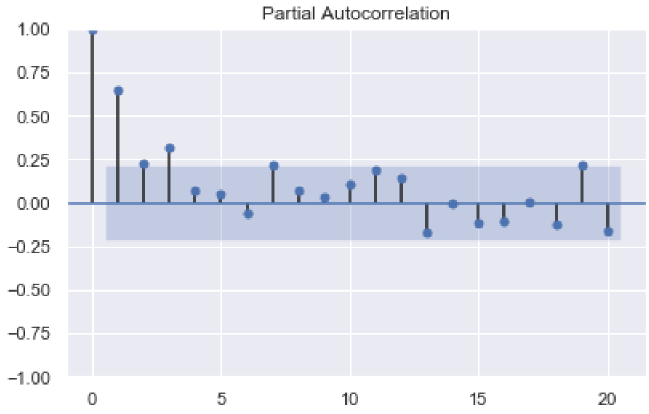


Figure 6.4: PACF Plot of Original Data

Transform data through differencing

To address the non-stationarity in the original data, we take the first-order differencing for data transformation. The results of ACF and PACF after the first differencing show significant improvements (see Figure 6.5 - 6.6). Specifically, Figure 6.5 - 6.6 show that the majority of ACF and PACF values are within the autocorrelation boundaries and are fairly equally distributed on both sides of the zero-benchmark line. These suggest the data after differencing is stationary.

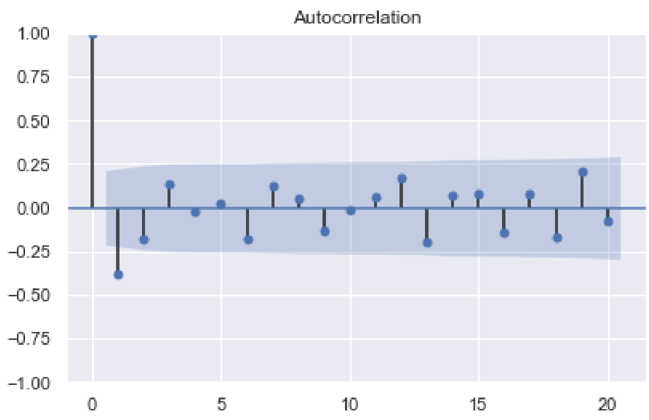


Figure 6.5: ACF Plot after First-order Differing

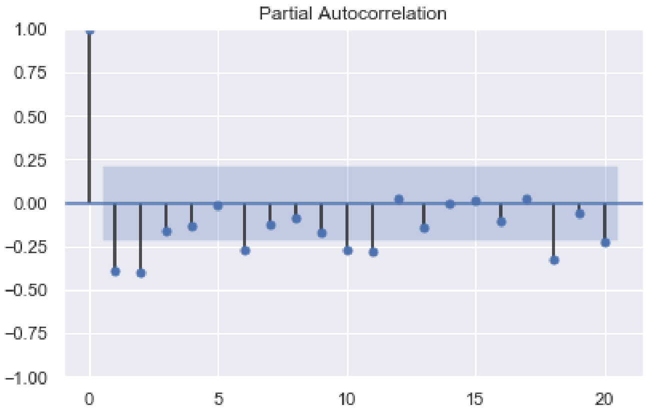


Figure 6.6: PACF Plot after First-order Differing

Identify parameters and model(s)

One critical step of using ARIMA models is to choose the appropriate parameters and final model to ensure the validity of counterfactual predication. In doing so, we first identify the parameter candidates. As mentioned above, the monthly time-series data used in the current study presents seasonality (see Figure 6.2), which suggest that the Seasonal ARIMA (SARIMA) modelling is suitable for the following analyses. To capture the seasonality in the modelling, the SARIMA includes two portions of parameters: the non-seasonal portion and the seasonal portion, which can be displayed as:

SARIMA (p, d, q) (P, D, Q)m

Where the lowercase portion (p, d, q) stands for the non-seasonable portion and the uppercase portion (P, D, Q)m stands for the seasonal portion. And m denotes the number of observations per season. According to the trend and pattern shown in Figure 6.2, we select the monthly data as the unit of the season, indicating a yearly seasonality. As a result, we choose the number 12 for m in model building below indicating a 12-months.

To preselect the potential models, we use the ACFs and PACFs for the non-seasonal portion (Figure 6.5 - 6.6) and the seasonal portion (Figure 6.7 - 6.8) to retrieve the potential parameter candidates. We then construct the alternative models with a combination of parameter candidates.

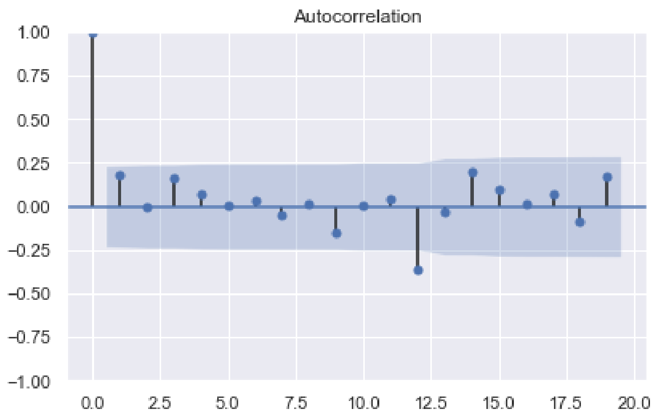


Figure 6.7: ACF Plot after First-order Differing, Lag = 12

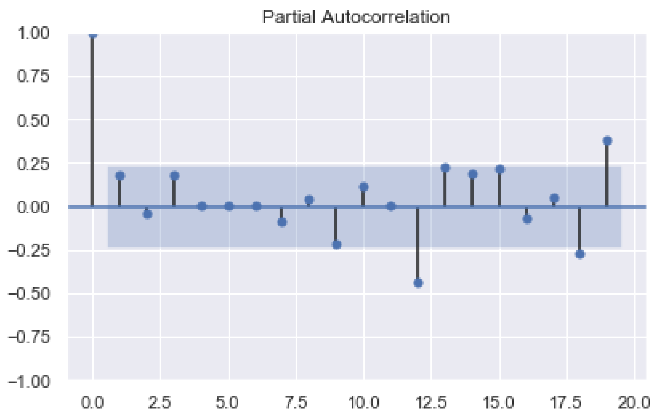


Figure 6.8: PACF Plot after First-order Differing, Lag = 12

The preselection process returns a number of model candidates, and we present two of them here for the simplicity: Model 1. (p = 0, d = 1, q = 1), (P = 1, D = 1, Q = 1, m = 12), and Model 2. (p = 0, d = 1, q = 1), (P = 0, D = 1, Q = 1, m = 12).

The SARIMA results for each model are shown in Figure 6.9 - 6.10, respectively.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                      injury_severity   No. Observations:                  84
Model:             SARIMAX(0, 1, 1)x(1, 1, 1, 12)   Log Likelihood                -292.995
Date:                              Sun, 05 Jun 2022   AIC                            593.990
Time:                                      17:03:24   BIC                            603.041
Sample:                                           0   HQIC                           597.589
                                               - 84
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8152      0.084     -9.658      0.000      -0.981      -0.650
ar.S.L12      -0.0389      0.197     -0.198      0.843      -0.424       0.347
ma.S.L12      -0.9994    133.697     -0.007      0.994    -263.040     261.041
sigma2       157.6680   2.11e+04      0.007      0.994   -4.11e+04    4.14e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.52   Jarque-Bera (JB):                 0.10
Prob(Q):                              0.47   Prob(JB):                         0.95
Heteroskedasticity (H):               2.06   Skew:                            -0.09
Prob(H) (two-sided):                  0.08   Kurtosis:                         2.97
===================================================================================
```

Figure 6.9: SARIMA Results - Model 1

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                      injury_severity   No. Observations:                  84
Model:            SARIMAX(0, 1, 1)x(1, 1, [], 12)   Log Likelihood                -301.211
Date:                              Sun, 05 Jun 2022   AIC                            608.423
Time:                                      00:01:56   BIC                            615.211
Sample:                                           0   HQIC                           611.122
                                               - 84
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8022      0.078    -10.224      0.000      -0.956      -0.648
ar.S.L12      -0.4567      0.096     -4.758      0.000      -0.645      -0.269
sigma2       268.2592     41.465      6.469      0.000     186.989     349.530
===================================================================================
Ljung-Box (L1) (Q):                   0.26   Jarque-Bera (JB):                 3.45
Prob(Q):                              0.61   Prob(JB):                         0.18
Heteroskedasticity (H):               2.38   Skew:                            -0.50
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.39
===================================================================================
```

Figure 6.10: SARIMA Results - Model 2

To choose the final model, we use a few criteria. 1) The smaller the AIC value, the better the model performs. 2) The value of Ljung-Box is larger than 0.05, suggesting the model is valid. 3) The P-values are significant (< 0.05). Taking those criteria into consideration, we choose Model 2 as the desired model. For the clarification, although the AIC value in Model 1 is smaller than that in Model 2, the P-values in Model 1 are not statistically significant. Thus we proceed with Model 2 for the next step.

Perform diagnostic measures

With the model identified above, we then conduct diagnostic measures to further determine whether the model is appropriate for the post-policy intervention predication. Figure

6.11 shows the results of diagnostic measures for Model 2 selected above. The results suggest that: 1) The residual errors fluctuate fairly well around a mean value of zero with a uniform variance; 2) The density distribution suggests the normality with its mean slightly towards the left-hand side; 3) No significant deviations present in the quantiles plot; 4) The ACF plot shows a white noise, in which all autocorrelations fall within the threshold boundaries. Overall, the results of diagnostic measures suggest that the selected model is appropriate and valid for the predication purpose.

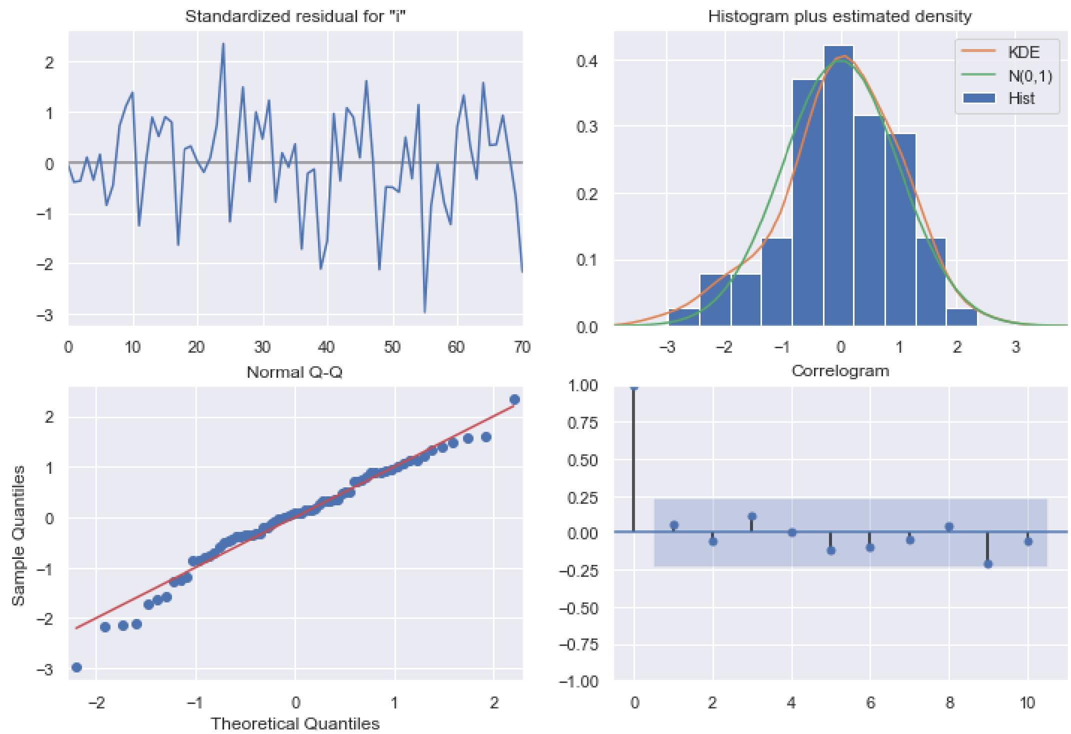


Figure 6.11: Diagnostic Measures for Model 2

The results of the chosen ARIMA model and corresponding counterfactual predication will be present in the results section.

6.2 Logistic Regression

6.2.1 Theory

Logistic Regression is a widely used binary and multi-class classification machine learning model and is associated to the following model: p(y|$\mathbf{x}$,θ) $= Bern(y|\sigma(\mathbf{w}^T x + b))$

where w are the weights, b is the bias and $\theta = (\mathbf{w}, b)$ and σ is the Sigmoid function which is in the case of binary classification is the following:

p(y=1|$\mathbf{x}$,θ) $= \frac{1}{1+e^{-(\mathbf{w}^T x + b)}}$

The Sigmoid function output the probability that an observation belongs to class 1 and classifies with a default threshold of 0.5 (default decision rule). The prediction function is therefore the following:

f($\mathbf{x}$,θ) $= \mathbf{w}^T x + b = b + \sum_{n=1}^{N} w_n * x_n$

which means that this function defines a hyperplane (which represent the decision boundary) with $w \in R^N$ being the normal vector orthogonal the this last and with an offset from the origin as the bias $b \in R$ There are several optimization algorithms that can be used for optimizing the log-likelihood function:

$$L(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} [y_n log(p_n) + (1 - y_n) log(1 - p_n)]$$

(6.4)

where p_n is the probability of belonging to class 1.

Among these algorithms, we took interest in the following:

- Newton's method which minimizes the cost function using its first and second derivative
- Limited-memory Broyden–Fletcher–Goldfarb–Shanno Algorithm or L-BFGS
- A Library for Large Linear Classification or Liblinear which applies automatic parameter selection using a Coordinate Descent algorithm by successively approximating minimisation of the cost function along the coordinate hyperplanes.

6.2.2 Model Fitting

For the model fitting, we used a balanced Logistic Regression (to overcome the imbalanced dataset) with the One-Hot encoded data, which was used to train the 5-fold cross-validation gridsearch with 42 combination of parameters based on the previously mentioned optimization algorithms. In addition to these parameters, we chose to test the L1/L2 regularization terms (to avoid overfitting and select the best predictors), and C the regularization coefficient that goes with L1/L2.

Then, for the SMOTEENN Pipeline method (SPM) was used to train a 5-fold GridSearch on similar parameters (Table 6.1) to take into account the imbalance of the dataset.

	Original	Resampled
Penalty	L1	L2
Loss	Squared Hinge	Squared Hinge
C	0.01	0.01
Balanced	Yes	No

Table 6.1: Logistic Regression parameters chosen for original and resampled datasets

Finally these 2 best models were trained on the entire train set, and in order to find the best hyperplane or decision boundary, the optimal threshold was obtained based on the precision-recall curve to maximize the number of true positives predicted by the model. The classification results were obtained using a confusion matrix (Table 6.2) with rows 1-2 representing the predictions with the trained model on the original data and rows 3-4 the resampled data:

		Predicted	
		0	1
Actual	0	19838	3902
	1	647	780
	0	17747	5993
	1	476	951

Table 6.2: Confusion Matrix: Logistic Regression trained on original (Top) and resampled (Bottom)

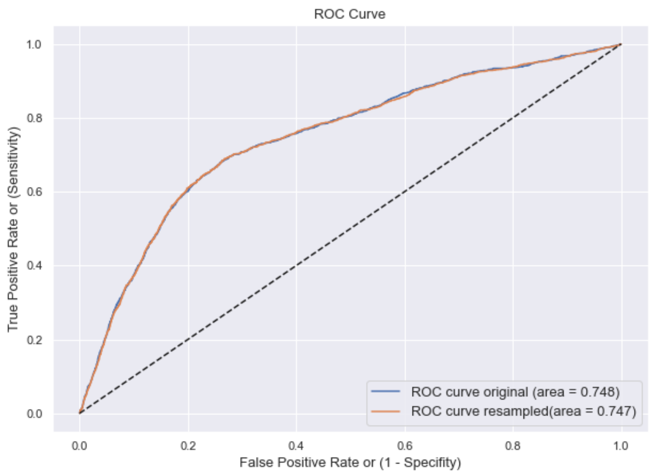


Figure 6.12: ROC Curve Logistic Regression for original and resampled data

To begin with, we can see that both models (using original and resampled data) are similar which can be explained by the additional parameter *balanced* of the Logistic Regression

	F1-score	Precision	Recall
0	0.90	0.97	0.84
1	0.26	0.17	0.55

Table 6.3: Test results for Logistic Regression for the original data

	F1-score	Precision	Recall
0	0.90	0.97	0.85
1	0.26	0.17	0.53

Table 6.4: Test results for Logistic Regression on the resampled data

model used for the original dataset. In addition, all the evaluation metrics for both data types are almost identical with a relatively low Precision for class 1 and therefore a low F1 score, i.e. the models have trouble predicting true positive classes. This is a consequence of the lack of data on the minority class.

6.3 Random Forest Classifier

6.3.1 Theory

Random Forest is an ensemble learning methods in which a number of weak models (decision trees) that perform poorly independently are combined to form a powerful model.

The training of the model works in the following manner:

- N samples are selected at random with replacement
- At each node:
  - m predictor variables are selected at random
  - the variable that give the best split based on an question or function is used to do a binary split on that node (the question's goal is to partition the data into a as unmixed or pure classes as possible and this purity is measured by the Entropy and Information Gain of the model)
  - Iteratively do the same for the next node selecting another m variables at random

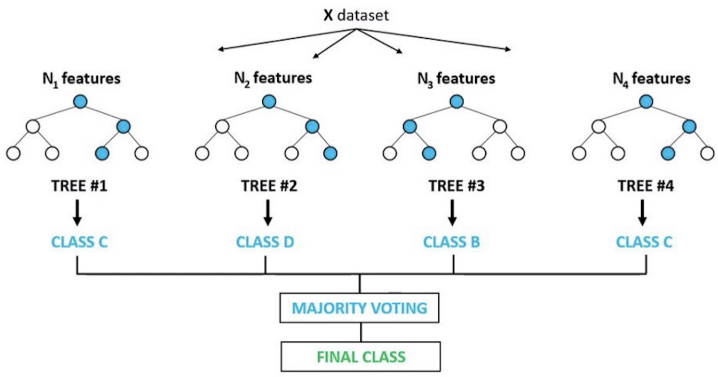


Figure 6.13: Random Forest Classifier

In the case of classification, an input is given a the top of the forest and as it traverses down the tree the data gets bucketed into smaller and smaller sets and the forest chooses the the classification with the most votes from each tree.

6.3.2 Model Fitting

In order to fit the Random Forest classifier, the first of Label-encoding was used to create the training and testing data as explained previously. Then the same process was used: first the original label-encoded dataset was given as input to a 5-fold GridSearch to find the optimal parameters which are the following:

	Original	Resampled
max depth	10	10
min samples leaf	4	4
min samples split	5	5
n estimators	300	300

Table 6.5: Random Forest parameters chosen for original and resampled datasets
We can see that the parameters chosen are identical in both cases.

		Predicted	
		0	1
Actual	0	23740	0
	1	1427	0
	0	18628	5112
	1	559	868

Table 6.6: Confusion Matrix: Random Forest trained on original (Top) and resampled (Bottom)

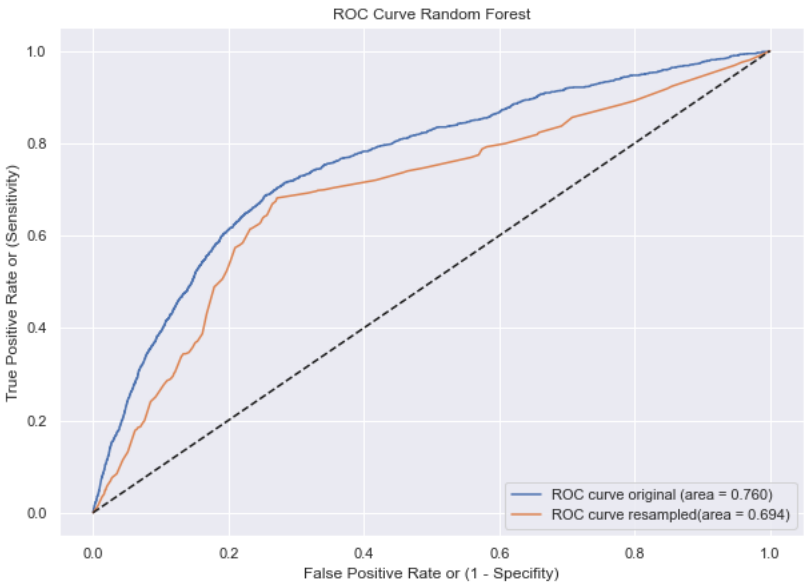


Figure 6.14: AUC plot for Tuned Random Forest Classifier trained on original and resampled data

	F1-score	Precision	Recall
0	0.97	0.92	1.00
1	0	0	0

Table 6.7: Test results for Random Forest for the original data

	F1-score	Precsion	Recall
0	0.87	0.97	0.78
1	0.23	0.15	0.61

Table 6.8: Test results for Random Forest for resampled data

We can see that there is a clear difference between the training on the original data and the resampling method used since in the first case the F1 score for class 1 is 0 whereas for the latter, the results are similar the Logistic Regression results with a much higher Recall value. Thus the RFC classifiers better the 2 classes thanks to the label-encoded data and the numerous categorical variables (Ensemble methods and decision trees are particularly adapted to this type of data).

6.4 SVM Classifier

6.4.1 Theory

Support Vector Machine (SVM) is a supervised machine learning algorithm commonly used in classification. SVM constructs a hyper-plane with an intention to achieve the largest distance, or margin, to the nearest training data points of any class. Given training vectors $x_i \in R^p, i = 1, \ldots, n$ with two classes, and a vector of classes $y \in \{1, -1\}^n$, it is desired to find $w \in R^p$ and $b \in R$ such that the prediction given by $sign(w^T x + b)$ is correct for most observations. The primal problem to solve is:

$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$ subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$, with $\zeta_i \geq 0, i = 1, ..., n$

where C is penalty parameter, ζ_i is the distance for an observation from from their correct margin boundary, and $\phi(\cdot)$ is the kernel function. In our application, we use the linear kernel such that $\phi(x)$ is the identity function. By trying to minimize $||w||^2 = w^T w$, we try to maximize the margin. Solving the optimization problem, we obtain a decision function for a given observation x:

$$\sum_i y_i \alpha_i K(x_i, x) + b, \text{ where } K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

and the predicted class correspond to its sign. The α_i are the dual coefficients from the dual problem to the primal problem:

$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$ subject to $y^T \alpha = 0$, with $0 \leq \alpha_i \leq C, i = 1, ..., n$

where e is the vector of all ones, and Q is an n by n positive semidefinite matrix with $Q_{ij} \equiv y_i y_j K(x_i, x_j)$

6.4.2 Model Fitting

To optimize the parameters of the SVM model, use 5-fold cross validation and grid search the following sets of parameters in order to maximize the AUC of the ROC curve: C = {1e-5,1e-4,1e-3,1e-2,0.1,1,10}, penalty = {L1,L2}, loss = {hinge, squared hinge}. We test several values of penalty parameters C, alongside L1 and L2 penalties and hinge and squared hinge loss functions. We perform the 5-fold CV and grid search on both the original dataset and the resampled dataset. We list the selected model parameters that maximize the AUC for the ROC curve below, for each dataset. These selected parameters are used to train a respective model for each of the original and resampled datasets, and these models are used to predict new data.

Using the models trained with the parameters listed in the table above, the confusion matrices in the table below were obtained by predicting the test dataset. The model trained on the original dataset obtained an overall accuracy of aprproximately 0.94.

However, the recall value for class 1 injury severity was found to be 0. For the model trained on the resampled dataset, we find there to be a significant improvement in the class 1 recall value (0.23) for a notable decrease in the overall accuracy (0.74).

		Predicted	
		0	1
Actual	0	23740	0
	1	1427	0
	0	17677	6063
	1	470	957

Table 6.9: Confusion Matrix: Tuned SVM trained on original (Top) and resampled (Bottom) datasets, predicting the test datasets

Below we give the plots the ROC and Precision-Recall (PR) curves. We note the similarity in ROC and PR curves in comparing the predictive power of the models trained on the original and resampled datasets, respectively. However, we note that significant increase in recall value for class 1 injury severity prediction by using a model trained on a resampled dataset instead of the original, imbalanced dataset. While the overall predictive accuracy has decreased, and the ROC curves may suggest similar predictive power, we find that the model now becomes more effective in predicting class 1 injury severity by training the model on the resampled dataset.

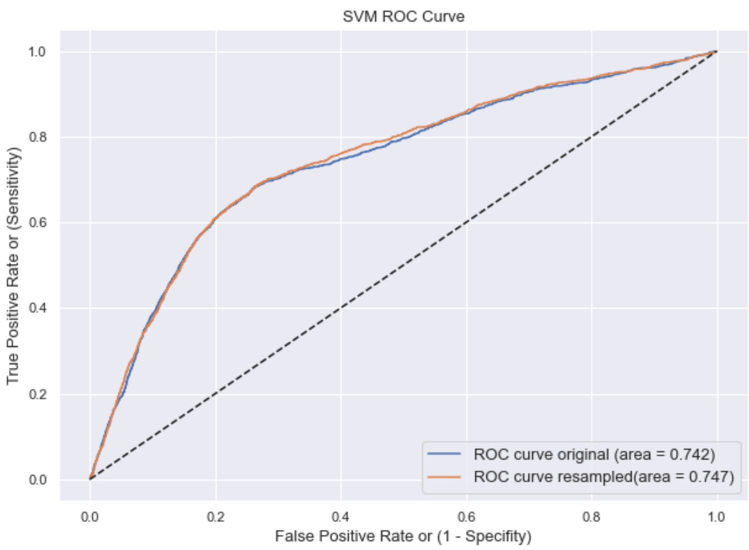


Figure 6.15: SVM ROC curves of the models trained on the original (blue) and resampled (brown) dataset on predicting the test dataset

	F1-score	Precision	Recall
0	0.97	0.94	1.00
1	0.00	0.00	0.00

Table 6.10: Test results for SVM Classifier on original dataset

	F1-score	Precision	Recall
0	0.84	0.97	0.74
1	0.23	0.67	0.23

Table 6.11: Test results for SVM Classifier on resampled dataset

6.5 KNN Classifier

6.5.1 Theory

The K Nearest Neighbors algorithm is a supervised machine learning algorithm that is given labeled input data in order to appropriately classify new, unlabeled data. In a prediction task, the KNN algorithm classifies an observation by weighing the labels of the closest K observations. In weighing the labels of the K nearest neighboring observations, the model may consider the weight of each of the K neighbors uniformly. Alternatively, the model may give more weight to closer neighbors and less weight to farther neighbors, among the K neighbors. In this case, a metric of distance is used to calculate the weights. Below we formally state the KNN model equation.

Suppose we are given a dataset of features X and a corresponding vector of labels Y. We choose a $k \in N$. For a test observation x_0, there exists a set of k nearest neighbors N_0. Let j index the number of labels, and let i index the set N_0. In the weighted KNN model, consider weights $w_i := \frac{1}{d(x_0,y_i)}$, where $d(\cdot)$ is a measure of distance, such that greater magnitudes of distance between x_0 and a $x_i \in N_0$ equates to less weight. In the KNN model using uniform weights, we may take each $w_i := 1$. The conditional probability that the test observation x_0 is of class label j is then given by the equation below.

$$P(Y = j|X = x_0) = \frac{1}{k} \sum_{i \in N_0} w_i I(y_i = j)$$

We assign the test observation x_0 to the class j that yields the highest conditional probability.

In addition to deciding how to assign weights to the K nearest neighbors, the value of K itself must be assigned. In the following section, we detail the process of selecting the value of K and the process of selecting the weight assignment of the K neighbors.

6.5.2 Model fitting

To select the optimal value of K and weight assignment to achieve the best Area Under the Curve (AUC) value for the Receiver Operating Characteristic (ROC) curve, we use cross validation and grid search methods on our training dataset. We use 5-fold cross validation, and grid search among the following parameter sets: weights = {uniform, distance}, n_neighbors = {1,3,5,. . . ,19}. Here, weights is the weighting used for the K neighbors. Uniform weighting considers the weight of the labels of each of the K neighbors equally, and here distance weighting uses the Euclidean distance as the distance metric. By grid searching among each permutation of (weights,distance), we fit 20 models with unique permutation of weights and distance. In addition to 5-fold cross validation, this results in fitting 100 models for each of the original dataset and the resampled dataset. The parameters of the model with highest AUC obtained from grid searching and cross validation is chosen to train a KNN model and predict the classes of new observations. We list the parameters chosen for the original dataset and resampled data set here, and note that the AUC values for each parameter permutation may be referenced in the provided code.

	Original	Resampled
Weights	Uniform	Distance
K	19	19

Table 6.12: KNN training parameters chosen for original and resampled datasets

After training models with the parameters listed in the table above, we use the models to predict the test dataset. We obtain the confusion matrices in the table below. The model trained on the original dataset obtained an overall accuracy of approximately 0.94. However, the recall value for class 1 injury severity was again found to be 0, similar to the results from the SVM model. For the model trained on the resampled dataset, in predicting the test dataset we find there to be a large improvement in the class 1 recall value (0.59) for a sizable decrease in the overall accuracy (0.75).

		Predicted	
		0	1
Actual	0	23739	1
	1	1427	0
	0	18073	5667
	1	583	844

Table 6.13: Confusion Matrix: KNN trained on original (Top) and resampled (Bottom) datasets, predicting the test dataset

Below we give the plots the ROC and Precision-Recall (PR) curves. Interestingly, the ROC curve for the model trained on the resampled dataset has a lower AUC value than the ROC curve for the model trained on the original dataset. While the resampled ROC for curve may indicate a worse overall predictive performance, the significant increase in recall value obtained by using the resampled training dataset indicates that the resampling technique is effective in increasing predictive power of class 1 injury severity.

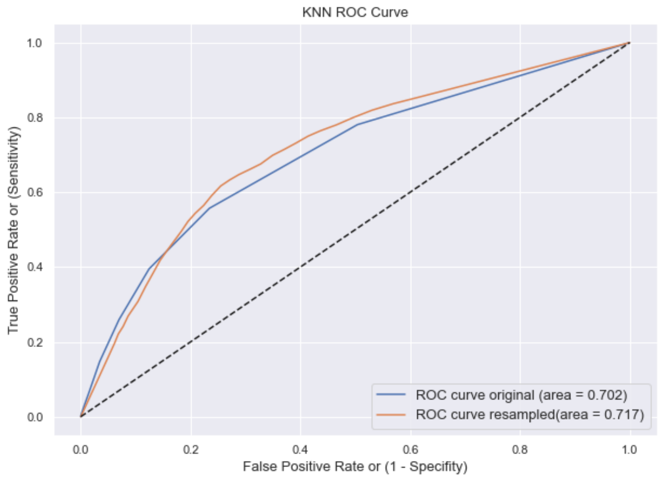


Figure 6.16: KNN ROC curves of the models trained on the original (blue) and resampled (brown) dataset on predicting the test dataset

	F1-score	Precision	Recall
0	0.97	0.94	1.00
1	0.00	0.00	0.00

Table 6.14: Test results for SVM Classifier on original dataset

	F1-score	Precision	Recall
0	0.85	0.97	0.76
1	0.21	0.13	0.59

Table 6.15: Test results for SVM Classifier on resampled dataset

7 Results and Findings

7.1 Role of Vision Zero in Addressing Fatality/Severe Injury

To answer our first research question, whether the Vision Zero policy is effective in reducing/mitigating traffic fatalities and severe injuries, we now turn to a comparison between the counterfactual scenario using the SARIMA model identified above and the observed crash records from the police department. To do so, in our analyses, we define January 2016 as the time when the policy intervention occurred for the simplicity reason.

Figure 7.1 below shows the trend of the counterfactual predication and the observed crash data during the study period. Specifically, the y-axis indicates the number of monthly traffic fatalities and serious injuries; the x-axis is the time indicated by each month from January 2013 to December 2019; the blue line shows the observed traffic fatalities and serious injuries reported by the police department, and the red line shows the predicated traffic fatalities and serious injuries in a scenario without the Vision Zero policy intervention.

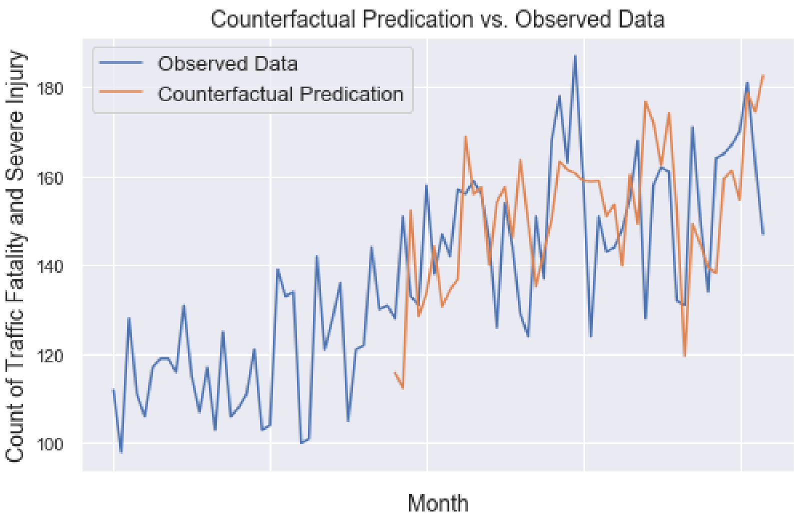


Figure 7.1: Counterfactual Predication vs. Observed Data

Figure 7.2 - 7.3 show the SARIMAX results and diagnostic measures of the chosen model for the counterfacutal predication. Specifically, the P-values in Figure 7.2 show

the statistical significance (< .05), and the ACF plot in Figure 7.3 presents white noise. Overall, we conclude that the chosen model is valid for the predication purpose.

```
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                       injury_severity   No. Observations:                84
Model:             SARIMAX(0, 1, 1)x(1, 1, [], 12)   Log Likelihood             -301.211
Date:                               Sun, 05 Jun 2022   AIC                         608.423
Time:                                       00:01:56   BIC                         615.211
Sample:                                            0   HQIC                        611.122
                                                - 84
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1         -0.8022      0.078    -10.224      0.000      -0.956      -0.648
ar.S.L12      -0.4567      0.096     -4.758      0.000      -0.645      -0.269
sigma2       268.2592     41.465      6.469      0.000     186.989     349.530
===================================================================================
Ljung-Box (L1) (Q):                   0.26   Jarque-Bera (JB):                 3.45
Prob(Q):                              0.61   Prob(JB):                         0.18
Heteroskedasticity (H):               2.38   Skew:                            -0.50
Prob(H) (two-sided):                  0.04   Kurtosis:                         3.39
===================================================================================
```

Figure 7.2: SARIMA Results of the Chosen Model

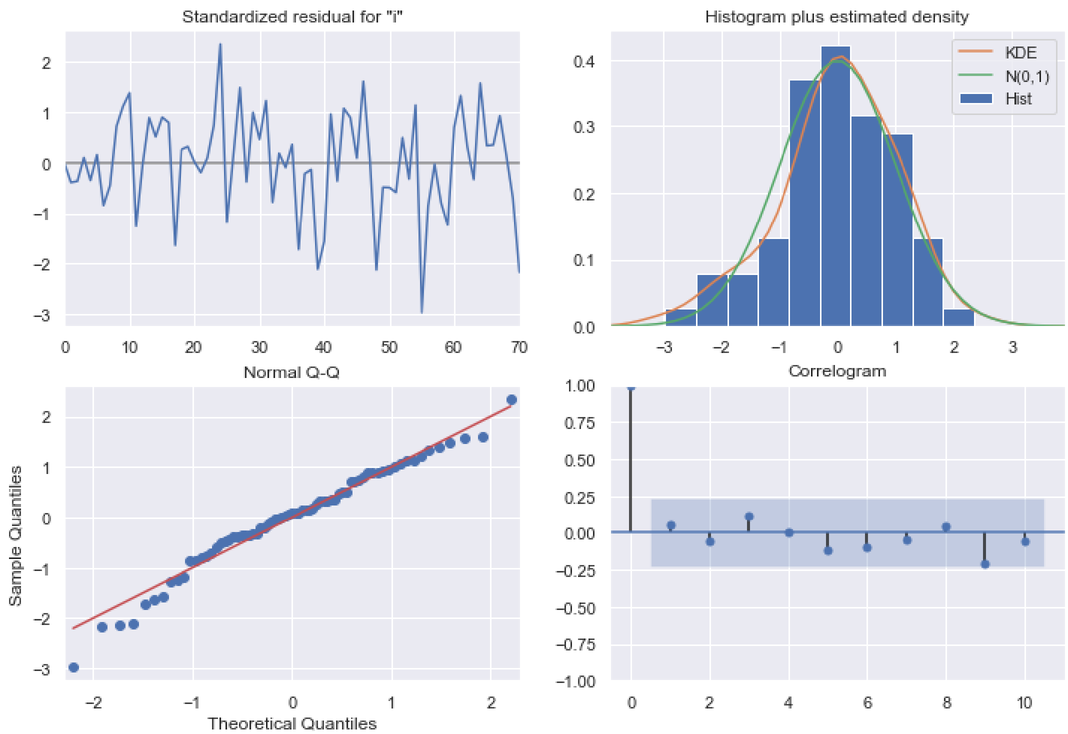


Figure 7.3: Diagnostic Measures for the Chosen Model

Overall, our study results suggest during the post-policy intervention time period (January 2016 - December 2019), the occurrence of traffic fatalities and severe injuries does not present significant changes. However, a close look indicates the mean value of observed data is slightly smaller than that of the counterfactual predication, suggesting potential improvements on addressing traffic safety.

Therefore, we conclude that based on our analyses in Los Angeles and San Francisco, the Vision Zero policy does not present a significant impact on the occurrence of traffic fatalities and severe injuries. More importantly, our analyses take a first step in measuring the impact of policy intervention, and provide a concrete tool to these Vision Zero communities with a purpose of implementing the policy in a more effective and efficient way. For this purpose, in the next section, we will further present how classification/machine learning models can be a help in advancing transportation safety and improving the implementation of Vision Zero policy.

7.2 Comparison of Modeling Performances in Injury Severity Prediction

To summarize the results of the four machine learning models, we analyzed the following plots:

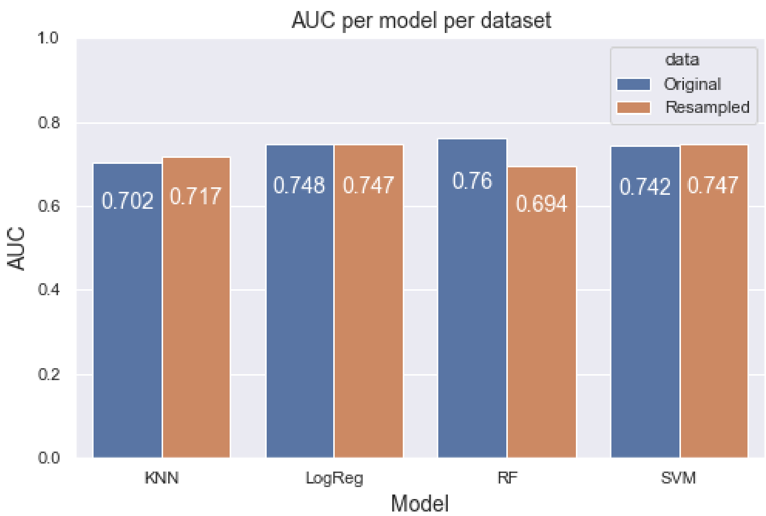


Figure 7.4: Comparison of AUC for each model, trained on original and resampled datasets

To better focus our comparison of the performances of the various classification models used, we first consider the difference between training each model on the original dataset and the resampled dataset. In the plots above, we compare the AUC of each model, trained on the original and resampled datasets. We find that there is little difference in AUC for KNN, Logistic Regression, and SVM. There is a small decrease in AUC for the resampled dataset for Random Forest.

Further, we can again examine the confusion matrices of each model trained on the two datasets. We find a consistent result for each of the 4 classification models that we consider: In using undersampling and oversampling techniques, we obtain a significant increase in class 1 injury severity recall value for a small loss in overall prediction accuracy. Critically, in Random Forest, KNN, and SVM, using the resampled dataset to train the models enabled the models to correctly classify class 1 injury severity with a nonzero recall value. This is a significant result, as class 1 injury severity consists of severe and fatal injuries, and it is essential to be able to predict class 1 injuries with a nonzero accuracy for a practical predictive model. With this in mind, we focus our comparison of metrics solely on the 4 models each trained on resampled dataset, visualized in the below figure.

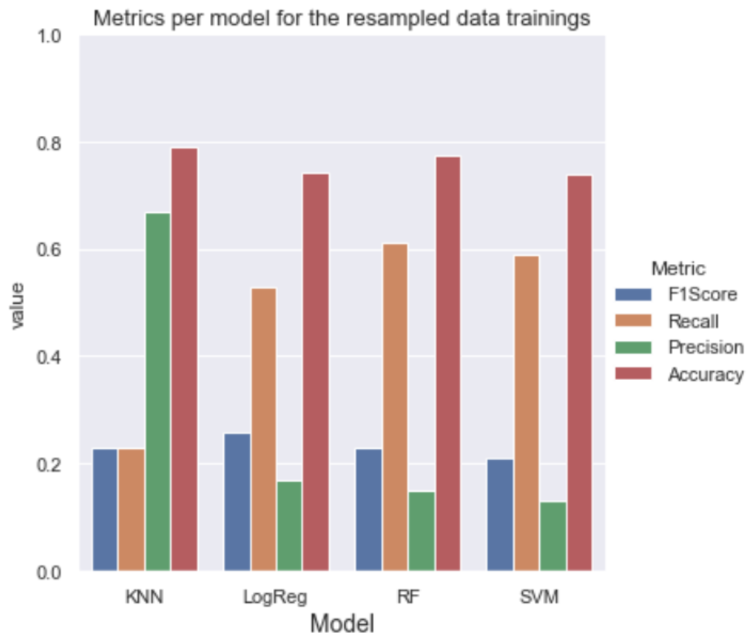


Figure 7.5: Comparison of different metrics for each model trained on resampled data

We first consider the F1 score of each model that is considered. We find that each model performs similarly with respect to this metric. Similarly, we find that the overall accuracy of each model is not largely different from one another. There are, however, notable differences between the models with respect to the precision and recall metrics.

We recall the equations of both precision and recall here.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

With this in mind, we see that KNN, with the highest precision, is the model which gives the greatest proportion of: correctly predicted class 1 injuries over all predicted class 1 injuries. The remaining three models perform similarly in this regard. In other words, KNN tends to produce the least false positives of the models we consider, while the remaining three models produce a similar proportion of false positives.

With respect to recall, however, KNN perform the worst of the 4 models we consider. Recall can be understood as the proportion of: correctly predicted class 1 injuries over all true class 1 injuries. Logistic Regression, Random Forest, and SVM have similar values of recall, with Random Forest having the highest recall value. In other words, Random Forest tends to produce the least false negatives for the models we consider. We note that Logistic Regression and SVM only fall slightly behind in this regard. KNN tends to produce the most false negatives.

Since there is no large difference in F1 score and accuracy among the four models we consider, these metrics lose importance in deciding which model is best in predicting injury severity. In addition, it is essential to predict class 1 injury severity for practical prediction purposes. With these considerations in mind, we instead focus on the precision and recall metrics. With respect to precision, KNN performs much better than the remaining three models. With respect to recall, Random Forest had the best predictive performance, with SVM only slightly worse in this regard.

8 Implications

The current project and analyses present contributions to both the academic field and the transportation safety practices in many aspects. Since its first entry in the US in 2012, Vision Zero has gained its momentum in the last few years. To date, more than 50 localities across the country has adopted the Vision Zero safety policy, and about one-third of those Vision Zero communities are in California. However, little is known about the impact and effectiveness of the Vision Zero safety policy on addressing traffic safety at localities.

Thus, through answering the first research question, our project contributes to existing literature on Vision Zero with a purpose of raising the awareness of measuring the policy implementation and evaluating the policy impacts through concrete methods, approaches, and strategies. We acknowledge that applying ARIMA modelling in policy intervention analysis builds on certain assumptions as mentioned in the method section above. And we do not intend to generalize the results in our study to other peer Vision Zero communities due to the limitations on the study areas and the impacts of additional contributing factors in traffic crashes. Thus, our study aims to take a first step in measuring the impact of the Vision Zero policy, and further efforts and contributions to the literature are much needed and expected.

From the transportation practice perspective, our study first provides some evidence to those Vision Zero communities on the impact of the policy implementation and what concrete actions and strategies those localities can consider to improve the effectiveness of policy implementation. Second, the study results also help the federal and state agencies, such as California, to provide the targeted policy and funding support to localities to further promote the policy implementation, such as SB-932. Overall, through answering our first research question, we argue that for those Vision Zero localities and the state and federal agencies that are supportive of the Vision Zero policy, it is crucial to take concrete methods and tools to measure the policy implementation process and work towards eliminating traffic fatalities and severe injuries in a more effective and efficient approach.

To further promote the effective implementation of the Vision Zero policy, we compare the performances of a number of machine learning models in predicting injury severity in traffic crashes. Our study results suggest that given the chosen study areas, the results of comparing modeling performance are largely consistent with prior studies on the topic (Al Mamlook et al., 2019; Wahab Jiang, 2019l Zhang et al., 2018). Specifically, Random Forest, Logistic Regression, KNN, and SVM present to be relevantly accurate in predicting injury severity in different perspectives and/or with different accuracy metrics/measures.

In attempting to address our second research question, we have found that there is no universally best model for predicting the injury severity of traffic collisions. While each of the four considered classification models perform fairly well in overall prediction accuracy, it remains essential to be able to predict severe and fatal injuries in an effort to reduce, and potentially eliminate, these injuries in Vision Zero communities. With this in mind, we cannot naively rank the model with best overall accuracy as the best model to use in predicting injury severity.

Our findings suggest, with respect to precision and recall, the use of KNN and Random Forest, respectively, in advising Vision Zero road safety policy. In using these models to assist in developing further future traffic and road safety policy, we may hope to better identify and prevent collision related injuries. However, this is not to disregard the predictive capabilities of any of the models we have considered. The predictive of performance of each of the models we have considered is promising, and their potential use in advising traffic and road safety largely depends on the intended effect of the implemented policy, such as trying to reduce solely class 0, solely class 1, or both classes of injuries.

9 Discussions

Although our project provides some evidence on the policy impact and the performances of differing machine learning models, it is important to recognize the model assumptions and the study limitations when one intends to generalize the results to a different context or area. First, the results in our study are drawn from only two cities, Los Angeles and San Francisco, which do not represent the entire Vision Zero communities across the country. However, the ARIMA modelling is a fairly easily replicable method, which makes it feasible to reproduce the policy intervention analysis used in the current study in other communities and obtain a preliminary understanding on the impact of Vision Zero within their jurisdictions.

Second, the ARIMA modelling used in the current study, as part of the causality analysis community, only considers the predictors themselves (i.e., the number of traffic fatalities and serious injuries) and does not take other contributing factors into account. However, at the current phase of Vision Zero, alternative methods to ARIMA modelling, such as the difference-in-difference (DID) model, are neither feasible nor robust in measuring the effects of the policy implementation. Because the DID method requires a comparison between treatment groups and control groups as a function of the policy intervention. However, compared to that of the control groups (i.e., those localities not adopting Vision Zero), the data size of the treatment groups (i.e., those Vision Zero communities) is much smaller. Consequently, applying the DID method will lead to a highly imbalanced dataset, making the policy intervention analysis infeasible. However, with a rapid growth of the Vision Zero communities, applying a variety of models and methods will become feasible. And future efforts on the topic will lead to a more robust and convincing conclusion on the impact of Vision Zero policy on addressing traffic safety issues.

The second goal of this project was to predict the severity injury of collision victims based on specific and representative features. In doing so, the goal on a larger scale is to help create preventive tools and reduce collision injury. Four machine learning models were applied to a highly imbalanced dataset and were each tuned and trained on the raw data and resampled data. Despite this latter method to overcome the skewed classes, the results obtained, while being encouraging are not sufficient yet to make accurate predictions. Indeed, all models have difficulties in predicting true severe injuries. Therefore, the possible improvements for these models are:

- Work on the models and more particularly the most promising (Random Forest and Logistic Regression). Make a fine-tuning gridsearch with a wider set of parameters on a representative subset of the dataset and try a XGBoost classifier which is more efficient and robust than a Random Forest classifier.

- Add data to the enrich the dataset with victims with severe injuries as much as possible or do data augmentation to generate synthetic examples based on the existing ones.
- Use more features after analyzing their correlation to the target variable, that would give more discrimative information the models and help improve the true positives and thus give accurate predictions.

Taking those assumptions and limitations into consideration, practitioners can adapt the methods in measuring the impact of the Vision Zero policy intervention on traffic safety within their jurisdictions, and researchers can take further efforts by advancing methods or generalizing the results in other areas to advance the policy implementation. In addition, our results on injury severity predication will further improve the effectiveness of policy implementation with carefully choosing the appropriate predictive modelling. Different from prior studies with a study unit of road segments (Zhang et al., 2018), our study is carried out at the crash victim level. However, the largely consistent results, further suggests the promising performance and accuracy of Random Forest and Logistic Regression in predicting the occurrence of traffic fatalities and serious injuries across a variety of units of study. Furthermore, our study takes sociodemographic features of crash victims, including age, race, types of road users, into account, signaling the potential impacts on using machine learning algorithms in addressing both transportation safety and equity issues in practices.

10 References

[1] Al Mamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., Frefer, A. A. (2019). Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity. 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 272–276. https://doi.org/10.1109/JEEIT.2019.8717393

[2] Bernanke, B. S., Boivtn, J., Eliasz, P. (2022). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. 37.

[3] Carnis, L., Blais, E. (2013). An assessment of the safety effects of the French speed camera program. Accident Analysis Prevention, 51, 301–309. https://doi.org/10.1016/j.aap.2012.11.022

[4] Chakraborty, M., Gates, T., Sinha, S. (2021). Causal Analysis and Classification of Traffic Crash Injury Severity Using Machine Learning Algorithms.

[5] California Department of Transportation (Caltrans). https://dot.ca.gov/

[6] California Legislature Information. https://leginfo.legislature.ca.gov/

[7] Chen Zeng, Rosa Abnous, Souma Chowdhury, Aerodynamic Modeling and Optimization of a Blended-Wing-Body Transitioning UAV, The AIAA Aviation and Aeronautics Forum and Exposition, 5 - 9 June 2017, Denver, Colorado

[8] David Walker, David Liu, Alan Jennings, Topology Optimization of an Aircraft Wing, AIAA SciTech, 5-9 January 2015, Kissimmee, Florida

[9] Government Offices of Sweden (2020). Renewed Commitment to Vision Zero: Intensified Efforts for Transportation Safety in Sweden.

[10] Griffin, B. A., Schuler, M. S., Stuart, E. A., Patrick, S., McNeer, E., Smart, R., Powell, D., Stein, B. D., Schell, T. L., Pacula, R. L. (2021). Moving beyond the classic difference-in-differences model: A simulation study comparing statistical methods for estimating effectiveness of state-level policies. BMC Medical Research Methodology, 21(1), 279. https://doi.org/10.1186/s12874-021-01471-y

[11] Houston, D. J., Richardson, L. E. (2002). Traffic safety and the switch to a primary seat belt law: The California experience. Accident Analysis Prevention, 34(6), 743–751. https://doi.org/10.1016/S0001-4575(01)00074-4

[12] Johansson, R. (2009). Vision Zero – Implementing a policy for traffic safety. Safety Science, 47(6), 826–831. https://doi.org/10.1016/j.ssci.2008.10.023

[13] Kweon, Y.-J., Hartman, S. E., Lynn, C. W. (2009). Stop versus yield on pedestrian-involved fatal crashes in the United States. Accident Analysis Prevention, 41(5), 1034–1039. https://doi.org/10.1016/j.aap.2009.06.013

[14] Los Angeles Open Data, City of Los Angeles. https://data.lacity.org/

[15] Lonescu (2022). Vision Zero Is Largely a Failure in the United States. Why? Planetizen. https://www.planetizen.com/news/2022/04/116830-vision-zero-largely-failure-united-states-why

[16] Li, X., Lord, D., Zhang, Y., Xie, Y. (2008). Predicting motor vehicle crashes using Support Vector Machine models. Accident Analysis Prevention, 40(4), 1611–1618. https://doi.org/10.1016/j.aap.2008.04.010

[17] Lavrenz, S. M., Vlahogianni, E. I., Gkritza, K., Ke, Y. (2018). Time series modeling in traffic safety research. Accident Analysis Prevention, 117, 368–380. https://doi.org/10.1016/j.aap.2017.11.030

[18] Lopez Bernal, J., Cummins, S., Gasparrini, A. (2018). The use of controls in interrupted time series studies of public health interventions. International Journal of Epidemiology, 47(6), 2082–2093. https://doi.org/10.1093/ije/dyy135

[19] Masten, S. V. (2007). Do states upgrading to primary enforcement of safety belt laws experience increased daytime and nighttime belt use? Accident Analysis Prevention, 39(6), 1131–1139. https://doi.org/10.1016/j.aap.2007.02.010

[20] Neyens, D. M., Donmez, B., Boyle, L. N. (2008). The Iowa graduated driver licensing program: Effectiveness in reducing crashes of teenage drivers. Journal of Safety Research, 39(4), 383–390. https://doi.org/10.1016/j.jsr.2008.05.006

[21] National Highway Traffic Safety Administration (NHTSA). https://www.nhtsa.gov/

[22] New topology optimization method for wing leading edge ribs, DOI 10.2514/1.DC031362

[23] Po Wu, Qihua Ma, Yiping Luo, Chao Tao, Topology Optimization Design of Automotive Engine Bracket, Energy and Power Engineering, 2016,8,230-235

[24] Qi Wang, Zhenzhou Lu, Changcong Zhou, New Topology Optimization Method for Wing Leading-Edge Ribs, Journal of Aircraft, Vol.48, No.5, September-October 2011

[25] Rock, S. M. (1995). Impact of the 65 mph speed limit on accidents, deaths, and injuries in illinois. Accident Analysis Prevention, 27(2), 207–214. https://doi.org/10.1016/0001-4575(94)00058-T

[26] Swedish Parliament (1997). Bill 1996/97:137: Vision Zero and Traffic Safety Society.

[27] Smith (2022). Hundreds died in L.A. traffic crashes in 2021. Is Vision Zero a failure? Los Angeles Time. https://www.latimes.com/california/story/2022-01-09/traffic-deaths-vision-zero-garcetti

[28] Schaffer, A. L., Dobbins, T. A., Pearson, S.-A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating large-scale health interventions. BMC Medical Research Methodology, 21(1), 58. https://doi.org/10.1186/s12874-021-01235-8

[29] Shadish, W. R., Cook, T. D., Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

[30] TransBase Dashboard, San Francisco. https://transbase.sfgov.org/dashboard/dashboard.php

[31] U.S. Department of Transportation (US DOT). https://www.transportation.gov/

[32] Unnikrishnan, J., Suresh, K. K. (2016). Modelling the Impact of Government Policies on Import on Domestic Price of Indian Gold Using ARIMA Intervention Method. International Journal of Mathematics and Mathematical Sciences, 2016, 1–6. https://doi.org/10.1155/2016/6382926

[33] Vision Zero Network. https://visionzeronetwork.org/

[34] Vinayak Kulkarni, Anil Jadhav, P.Basker, Finite Element Analysis and Topology Optimization of Lower Arm of Double Wishbone Suspensionusing RADIOSS and Optistruct,International Journal of Science and Research(IJSR), ISSN:2319-7064

[35] Värnild, A., Tillgren, P., Larm, P. (2020). What types of injuries did seriously injured pedestrians and cyclists receive in a Swedish urban region in the time period 2003–2017 when Vision Zero was implemented? Public Health, 181, 59–64. https://doi.org/10.1016/j.puhe.2019.11.019

[36] Värnild, A., Larm, P., Tillgren, P. (2019). Incidence of seriously injured road users in a Swedish region, 2003–2014, from the perspective of a national road safety policy. BMC Public Health, 19(1), 1576. https://doi.org/10.1186/s12889-019-7937-0

[37] Värnild A, Belin M, Tillgren P. (2016). 763. Vision zero – road traffic effects for severely injured in a Swedish county Injury Prevention, 22.

[38] Värnild A, Belin MA, Tillgren P. (2016). Vision zero – road traffic effects for severely injured in a Swedish county. Safety 2016 World Conference; 18–21 september 2016. Tammerfors: Injury Prevention.

[39] Vernon, D. D., Cook, L. J., Peterson, K. J., Michael Dean, J. (2004). Effect of repeal of the national maximum speed limit law on occurrence of crashes, injury crashes, and fatal crashes on Utah highways. Accident Analysis Prevention, 36(2), 223–229. https://doi.org/10.1016/S0001-4575(02)00151-3

[40] Vanlaar, W., Robertson, R., Marcoux, K. (2014). An evaluation of Winnipeg's photo enforcement safety program: Results of time series analyses and an intersection camera experiment. Accident Analysis Prevention, 62, 238–247. https://doi.org/10.1016/j.aap.2013.09.023

[41] Wahab, L., Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLOS ONE, 14(4), e0214966. https://doi.org/10.1371/journal.pone.0214966

[42] Wagenaar, A. C., Maybee, R. G., Sullivan, K. P. (1988). Mandatory seat belt laws in eight states: A time-series evaluation. Journal of Safety Research, 19(2), 51–70. https://doi.org/10.1016/0022-4375(88)90044-8

[43] Washington, S., Karlaftis, M., Mannering, F., Anastasopoulos, P. (2020). Statistical and Econometric Methods for Transportation Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429244018

[44] Zhang, J., Li, Z., Pu, Z., Xu, C. (2018). Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods. IEEE Access, 6, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979