Data-Driven Analysis of Traffic Crashes in Chicago

Kentaro Osawa, Bharat Premnath

The George Washington University

DATS 6501: Data Science Capstone

Dr. Abdi Awl

May 1, 2024

.

Table of contents

Table of contents 2
Table of figures 3
Table of tables 5
Glossary of Terms 6
Introduction 8
Introduction/Background 8
Problem Statement 8
Problem Elaboration 9
Motivation 9
Project Scope 9
Literature Review 10
Investigating the significant factors in traffic crash severity 10
Estimating the probability distribution of the count of traffic crashes 11
Methodology 13
Data Collection 13
Dataset Description 13
Data Processing and Feature Engineering 16
Data Processing and Feature Engineering for Classification Models 16
Data Processing and Feature Engineering for GLM 19
Data Modeling and Visualizations 21
Data Modeling and Visualizations for Classification Models 21
Data Modeling and Visualizations for GLM 24
Results and Analysis 28
Results and Analysis of classification models 28
Results and Analysis of GLM 33
Development of dashboard 35
Data Pipeline Architecture and Components 36
Data Flow Process 38
Conclusion 42
Conclusion 42
Project Limitation and Future Research 43
References 44
Appendix 46
Statistical test 46
Complement figures 48
AIC and BIC from Negative Binomial models 56

Table of figures

Figure 1 Class imbalance in the target variable "Severity"..... 19
Figure 2 The number of traffic crashes by month (left column), day of the week (middle column) and hour (right columns). The first, second and third row contains only Severity 1, 2 and 3 accidents, respectively. The value of 1 on the day of the week stands for Sanday..... 22
Figure 3 Geographical distribution of traffic crashes from 2015 to January 21, 2024. Green, yellow and red dots stand for Severity 1, 2 and 3 accidents, respectively…..... 23
Figure 4 Proportion of each level of weather condition in each severity category..... 23
Figure 5 Histogram of the count of traffic crashes in a 2-hour interval in Austin and Near West Side..... 25
Figure 6 Histogram of the count of traffic crashes in a 2-hour interval in Austin for each severity..... 25
Figure 7 Histogram of the count of traffic crashes in a 2-hour interval in Austin for each day of the week. The value of 1 on the day of the week stands for Sunday..... 26
Figure 8 Histogram of the count of traffic crashes in a 2-hour interval in Austin for each month..... 27
Figure 9 Histogram of the count of traffic crashes in a 2-hour interval in Austin for each timeframe..... 28
Figure 10 Feature importances of the best model (CatBoost trained from undersampled data)..... 31
Figure 11 Proportion of each level of first crash type in each severity category..... 31
Figure 12 Proportion of each level of trafficway type in each severity category..... 32
Figure 13 Proportion of each level of the number of units in each severity category..... 32
Figure 14 Architecture of the pipeline..... 36
Figure 15 Screen of the classification models part in dashboard..... 40
Figure 16 Screen of the GLM part in dashboard..... 41
Figure A1 Proportion of each level of traffic control device in each severity category..... 48
Figure A2 Proportion of each level of device condition in each severity category..... 48
Figure A3 Proportion of each level of lighting condition in each severity category..... 49
Figure A4 Proportion of each level of alignment in each severity category..... 49
Figure A5 Proportion of each level of roadway surface condition in each severity category..... 50
Figure A6 Proportion of each level of road defect in each severity category..... 50
Figure A7 Proportion of each level of street direction in each severity category..... 51
Figure A8 Proportion of each level of day of the week in each severity category..... 51
Figure A9 Proportion of each level of month in each severity category..... 52
Figure A10 Proportion of each level of day in each severity category..... 52
Figure A11 Proportion of each level of hour in each severity category..... 53
Figure A12 Proportion of each level of minute in each severity category..... 53
Figure A13 Proportion of each level of posted speed limit in each severity category..... 54

Figure A14 Proportion of each level of the primary contributory cause in each severity category..... 54
Figure A15 Proportion of each level of the secondary contributory cause in each severity category..... 55

Table of tables

Table 1 Columns in the original dataset..... 13
Table 2 Summary of evaluation metrics from classification models trained from the undersampled data..... 29
Table 3 Summary of evaluation metrics from classification models trained from the oversampled data..... 29
Table 4 Summary of all the combinations of independent variables..... 33
Table 5 Summary of AIC and BIC..... 34
Table A1 Summary of the test of independence between the target and categorical features..... 46
Table A2 Summary of the Kruskal-Wallis test between the target and categorical features 47
Table A3 AIC from Negative Binomial models. The α changes from 0.1 to 2 by 0.1, and then we change α by 0.01 between 0.1 and 0.3 because AIC takes the smallest value at α = 0.2..... 56
Table A4 BIC from Negative Binomial models. The α changes from 0.1 to 2 by 0.1, and then we change α by 0.01 between 0.1 and 0.3 because AIC takes the smallest value at α = 0.2..... 58

Glossary of Terms

Term	Definition
Akaike Information Criteria (AIC)	A metric used for model selection, which contains a penalization term for the number of parameters. This metric is used to select the model predicting well.
Bayesian Information Criteria (BIC)	Another metric used for model selection, which also contains a penalization term for the number of parameters. This metric is used to select the correct model.
CatBoost	A machine learning algorithm, which develops tree models one after another so as for a new model to decrease errors made from the previous model. In this algorithm, you do not need to encode categorical features.
Decision Tree	A machine learning algorithm, which finds a rule for dividing data so as to decrease the impurity in each divided data.
Feedforward Neural Network (FNN)	A deep learning algorithm, which does not have a loop structure. Information flows only one way: from the input nodes to the output nodes.
Generalized Linear Model (GLM)	A linear regression model, which can adopt not only gaussian but also other distributions as the distribution of its residuals.
Negative Binomial model	A type of GLM model, which assumes the Negative Binomial distribution as the distribution of its residuals. In Negative Binomial distributions, the value of the mean becomes smaller than that of the variance.
One-hot encoding (OHE)	An encoding method, in which each categorical feature is encoded to have the same number of rows as the number of its levels. Each new row corresponds to a level of the original feature and has the value of 0 (when an observation does not have the corresponding value) or 1 (when an observation has the corresponding value).
Poisson model	A type of GLM model, which assumes the Poisson distribution as the distribution of its residuals. In

	Poisson distributions, the value of the mean is equal to that of the variance.
Random Forest	A machine learning algorithm, which develops many tree models and makes predictions using a majority vote of the models.
Synthetic Minority Oversampling Technique (SMOTE)	A sampling method, which synthesizes new samples for minority classes to make the class balanced.
Zero-inflated model	A type of GLM model, which is often used when the count of zeros is larger than expected from the Poisson or Negative Binomial models.

Introduction

Introduction/Background

According to the World Health Organization (2023), approximately 1.19 million people die, and 20 - 50 million people suffer non-fatal injuries from road traffic crashes worldwide each year. Governments take countermeasures to decrease the number of traffic crashes. This project aims to address an essential problem in the field of traffic administration, which is reducing traffic crashes, by leveraging data science techniques to analyze and model traffic crash data. This project aligns with the ultimate goal of reducing traffic crashes.

Previous studies investigated significant responsible features for traffic crash severity, and several factors are suggested as significant responsible features. For example, some research suggested weather conditions, and some suggested the day of the week (Bhuiyan et al., 2022). All these factors should be liable for their severity, but the significant factors may change depending on the country or city. This project will investigate traffic accidents in a city in the United States. The findings in this project will help make effective policies and decisions to eliminate fatal injuries from traffic crashes.

Problem Statement

In the United States, approximately 6 million traffic crashes happen each year (National Highway Traffic Safety Administration, 2023c), which lead to human suffering as well as economic losses. For example, An estimated 48 thousand people died in traffic crashes in 2022, and traffic fatalities have been increasing over the past decade (National Highway Traffic Safety Administration, 2023b). According to the National Highway Traffic Safety Administration (2023a), traffic crashes cost American society 340 million dollars in 2019.

Problem Elaboration

This project addresses the specific problem of determining the major factors affecting the severity of car crashes. It involves developing an interpretable model for predicting crash severity and conducting statistical analysis. The scope includes descriptive and inferential statistical analysis, handling missing data, feature engineering, model selection, and validation.

Motivation

This project is motivated by a desire to decrease severe traffic accidents. Implementing effective countermeasures seems to be efficient for achieving this goal. We aim to gain insights into the main factors affecting the severity of traffic crashes and impact effective policy-making by utilizing data science knowledge and techniques for traffic crash data.

Project Scope

Our main objective is to discover the main factors affecting the severity of traffic crashes. For this purpose, we will develop several machine-learning models to predict the severity of each traffic accident. After finding the best model based on model evaluation metrics, we will show the significant features in the best model prediction and recommend what countermeasures might be useful to decrease severe car accidents.

In addition, we will estimate the probability distribution of the count of traffic crashes by utilizing generalized linear models (GLM). By knowing the probability distribution, we can predict the risk of car accidents.

Literature Review

Investigating the significant factors in traffic crash severity

Bhuiyan et al. (2022) investigated the significant factors in traffic accident severity using data about car accidents in Bangladesh. They divided the accidents into three parts: Severe (fatality is greater than or equal to 3), Medium (fatality is 1 or 2), and Mild (fatality is zero). They utilized feature selection techniques and developed several machine learning models to predict the severity of accidents using each set of features. The models were Decision tree, Random forest, Multinomial Naive Bayes, and Gaussian Naive Bayes. They evaluated the models by accuracy, precision, recall, specificity, and F1 score. After evaluating each model, they listed the significant factors in the severity of traffic crashes, which were the features from the well-performing model: day of the crash (month), residential location, vehicle type, license type, seat belts, gender, time of the crash, road surface type, and road classification.

Ghandour et al. (2022) analyzed the traffic accidents in Lebanon. They set Fatality occurrence, a two-level feature, Fatal or Not Fatal, for their target variable and developed binary classification machine learning models: Sequential minimal optimization (SMO), Random forest, Artificial neural network, Logistic regression, and Naive Bayes. They adopted the F1-score, AUC-PR, and Kappa as the performance metrics and evaluated the models. As a result, Vote SMO with Bagging J48 was identified as their best model. From the best model, they enumerate the influencing factors: crash type, injury severity level, spatial cluster ID, hour of road cash, day of the week, and road type.

Fiorentini and Losa (2020) developed models to predict the severity of road crashes based on road accident data in Great Britain. They used Crash Severity as their target variable, which had two levels: Fatal + Injury (F+I) and Property Damage Only (PDO). Their target

variable was highly imbalanced. Therefore, they handled this imbalance utilizing the random undersampling the majority class (RUMC) technique and developed models including Random Tree, K-Nearest Neighbor, Logistic Regression, and Random Forest from both the initial and undersampled training datasets. From the evaluation metrics, including the confusion matrix, they asserted that the models trained from the balanced data were better than those trained from the imbalanced data. In addition, they showed the significant input features: day of the week, the number of casualties, the first road class, and the number of vehicles.

As explained above, Random Forest and Decision Tree models are often used in this field. It is interesting that the suggested significant factors differ among the studies, although there are a few common factors across multiple studies. This indicates that the significant factors are different by country or city.

Estimating the probability distribution of the count of traffic crashes

In the attempt to model traffic crash data, Poisson and Negative Binomial processes are often assumed. The zero-inflated probability model is another choice. In Poisson models, the mean and the variance are equal, while the variance becomes greater than the mean in Negative Binomial models. In zero-inflated models, the count of traffic crashes shows more zeros than expected from the previous two models. Lord et al. (2005) summarized the theoretical principles for these models and reviewed previous studies. In addition, they simulated crash data from Bernoulli trials and discussed the conditions where each model can be used as statistical approximations to the crash process. They concluded that Poisson models reproduce the distribution of the count of traffic crashes well when they happen under nearly homogenous conditions, whereas Negative Binomial models do under heterogeneous conditions. On the other

hand, they suggested that zero-inflated models are the results of one or more of the following four conditions:

1. "sites with a combination of low exposure, high heterogeneity, and sites categorized as high risk"
2. "analyses conducted with small time or spatial scales"
3. "data with a relatively high percentage of missing or mis-reported crashes"
4. "crash models with omitted important variables"

Lord et al. (2008) indicated that the Conway-Maxwell-Poisson (COM-Poisson) generalized linear model (GLM) was an alternative choice of the Negative Binomial model by showing COM-Poisson GLM worked as well as Negative Binomial models. The COM-Poisson GLM can model traffic crash data with a smaller variance compared to the mean and the opposite. Since sometimes there are traffic crash datasets with underdispersion, COM-Poisson GLMs are better than Negative Binomial models (Lord et al., 2008).

In this project, we modeled traffic crash count data using Poisson, Negative Binomial, and zero-inflated models. The reasons why we did not utilize COM-Poisson GLMs were as follows:

1. We used Python to develop models, but no package provided COM-Poisson GLMs.
2. The cases where COM-Poisson GLMs show their advantages seem not to be many because "crash data have often been shown to exhibit over-dispersion" (Load et al., 2008).

Methodology

Data Collection

The data was retrieved from the Chicago Data Portal (https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if/about_data). It was about the traffic crashes that occurred in the city of Chicago from 2015. The data source updates the data on a daily basis.

Dataset Description

The dataset contained traffic crashes data from 2015 to January 21, 2024 because we acquired the dataset from the above data source on January 22, 2024. It had 799,526 rows and 48 columns. The following table shows the column name and description of each column.

Table 1

Columns in the original dataset

Column Name	Description
CRASH_RECORD_ID	Unique ID in the dataset
CRASH_DATE_EST_I	Crash data estimated by desk officer or reporting party (Only used in cases where crash is reported at police station days after the crash)
CRASH_DATE	Date and time of crash as entered by the reporting officer
POSTED_SPEED_LIMIT	Posted speed limit, as determined by reporting officer
TRAFFIC_CONTROL_DEVICE	Traffic control device present at crash location, as determined by reporting officer
DEVICE_CONDITION	Condition of traffic control device, as determined by reporting officer
WEATHER_CONDITION	Weather condition at time of crash, as determined by reporting officer
LIGHTING_CONDITION	Light condition at time of crash, as determined by reporting officer
FIRST_CRASH_TYPE	Type of first collision in crash

TRAFFICWAY_TYPE	Trafficway type, as determined by reporting officer
LANE_CNT	Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection)
ALIGNMENT	Street alignment at crash location, as determined by reporting officer
ROAD_SURFACE_COND	Road surface condition, as determined by reporting officer
ROAD_DEFECT	Road defects, as determined by reporting officer
REPORT_TYPE	Administrative report type (at scene, at desk, amended)
CRASH_TYPE	A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away
INTERSECTION_RELATED_I	A field observation by the police officer whether an intersection played a role in the crash. Does not represent whether or not the crash occurred within the intersection.
NOT_RIGHT_OF_WAY_I	Whether the crash began or first contact was made outside of the public right-of-way.
HIT_AND_RUN_I	Crash did/did not involve a driver who caused the crash and fled the scene without exchanging information and/or rendering aid
DAMAGE	A field observation of estimated damage.
DATE_POLICE_NOTIFIED	Calendar date on which police were notified of the crash
PRIME_CONTRIBUTORY_CAUSE	The factor which was most significant in causing the crash, as determined by officer judgment
SEC_CONTRIBUTORY_CAUSE	The factor which was second most significant in causing the crash, as determined by officer judgment
STREET_NO	Street address number of crash location, as determined by reporting officer
STREET_DIRECTION	Street address direction (N,E,S,W) of crash location, as determined by reporting officer
STREET_NAME	Street address name of crash location, as determined by reporting officer
BEAT_OF_OCCURRENCE	Chicago Police Department Beat ID. Boundaries available at https://data.cityofchicago.org/d/aerh-rz74
PHOTOS_TAKEN_I	Whether the Chicago Police Department took photos at the location of the crash
STATEMENTS_TAKEN_I	Whether statements were taken from unit(s) involved in crash

DOORING	Whether crash involved a motor vehicle occupant opening a door into the travel path of a bicyclist, causing a crash
WORK_ZONE_I	Whether the crash occurred in an active work zone
WORK_ZONE_TYPE	The type of work zone, if any
WORKERS_PRESENT_I	Whether construction workers were present in an active work zone at crash location
NUM_UNITS	Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.
MOST_SEVERE_INJURY	Most severe injury sustained by any person involved in the crash
INJURIES_TOTAL	Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer
INJURIES_FATAL	Total persons sustaining fatal injuries in the crash
INJURIES_INCAPACITATING	Total persons sustaining incapacitating/serious injuries in the crash as determined by the reporting officer. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they were capable of performing before the injury occurred. Includes severe lacerations, broken limbs, skull or chest injuries, and abdominal injuries.
INJURIES_NON_INCAPACITATING	Total persons sustaining non-incapacitating injuries in the crash as determined by the reporting officer. Any injury, other than fatal or incapacitating injury, which is evident to observers at the scene of the crash. Includes lump on head, abrasions, bruises, and minor lacerations.
INJURIES_REPORTED_NOT_EVIDENT	Total persons sustaining possible injuries in the crash as determined by the reporting officer. Includes momentary unconsciousness, claims of injuries not evident, limping, complaint of pain, nausea, and hysteria.
INJURIES_NO_INDICATION	Total persons sustaining no injuries in the crash as determined by the reporting officer
INJURIES_UNKNOWN	Total persons for whom injuries sustained, if any, are unknown
CRASH_HOUR	The hour of the day component of CRASH_DATE.
CRASH_DAY_OF_WEEK	The day of the week component of CRASH_DATE. Sunday=1
CRASH_MONTH	The month component of CRASH_DATE.
LATITUDE	The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash

LONGITUDE	The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash
LOCATION	The crash location, as determined by reporting officer, as derived from the reported address of crash, in a column type that allows for mapping and other geographic analysis in the data portal software

Data Processing and Feature Engineering

Data Processing and Feature Engineering for Classification Models

Removing an identifier and some features. We removed the identifier "CRASH_RECORD_ID" as well as "REPORT_TYPE", "DATE_POLICE_NOTIFIED", "PHOTOS_TAKEN_I", and "STATEMENTS_TAKEN_I", which did not affect the severity of the crash. "CRASH_HOUR" and "CRASH_MONTH" are also removed because new features that were the same as them were created, as explained in the next paragraph.

Handling date time feature. It is convenient that date information was separated into year, month, day, hour, minute, and second. Therefore, we converted "CRASH_DATE" to six new features: "CRASH_DATE_year", "CRASH_DATE_month", "CRASH_DATE_day", "CRASH_DATE_hour", "CRASH_DATE_minute", "CRASH_DATE_second".

Handling missing values. First, we checked the proportion of missing values in each column and removed features with greater than 50% of the missing value proportion. As a result, "CRASH_DATE_EST_I", "LANE_CNT", "INTERSECTION_RELATED_I", "NOT_RIGHT_OF_WAY_I", "HIT_AND_RUN_I", "DOORING_I", "WORK_ZONE_I", "WORK_ZONE_TYPE", and "WORKERS_PRESENT" were removed from the dataset. Next, rows with any missing values were dropped.

Removing constant features. We checked whether there were any features whose value was constant. "INJURIES_UNKNOWN" and "CRASH_DATE_second" turned out to have constant values, which led to their removal..

Removing unreliable observations. We found several observations had the value of 2013 or 2014 in "CRASH_DATE_year". However, since the collection of traffic crash data started in 2015, they must have been wrong. In addition, the latitude and longitude in some observations were 0, which must have been erroneous because all the traffic crashes occurred in Chicago. We removed all these observations.

Feature engineering. We created a new feature named "Severity" using "INJURIES_TOTAL" and "INJURIES_FATAL". "Severity" has three levels as follows:

- Severity 1: The value of "INJURIES_TOTAL" is 0.
- Severity 2: The value of "INJURIES_TOTAL" is 1 or 2, and "INJURIES_FATAL" is 0.
- Severity 3: The value of "INJURIES_TOTAL" is greater or equal to 3, or "INJURIES_FATAL" is greater than 1.

We also grouped the levels of "PRIME_CONTRIBUTORY_CAUSE" and "SEC_CONTRIBUTORY_CAUSE" because they had as many as 40 levels, which led to the expansion of the dimension. After the grouping, the number of their levels became 8.

Reducing dimension. Considering the model development, the small dimension of the dataset is preferable. Therefore, we removed several features.

Our purpose in developing machine-learning models was to find the significant factors in the severity of traffic crashes, which could help make effective policies to reduce severe car accidents. Therefore, we removed several features that were not helpful for the above purpose: "CRASH_TYPE", "DAMAGE", "MOST_SEVERE_INJURY", "INJURIES_TOTAL", "INJURIES_FATAL", "INJURIES_INCAPACITATING", "INJURIES_NON_INCAPACITATING", "INJURIES_REPORTED_NOT_EVIDENT", "INJURIES_NO_INDICATION", and "CRASH_DATE_year".

This dataset had various location-related features, such as "STREET_NAME" and "BEAT_OF_OCCURRANCE". These were categorical features, leading to the expansion of the dimension of the dataset. Therefore, we decided to focus on "LONGITUDE" and "LATITUDE" and removed the other location-related features: "STREET_NO", "STREET_NAME", "BEAT_OF_OCCURRENCE", and "LOCATION".

In addition, we dropped "CRASH_DATE_day" and "CRASH_DATE_minute". The former turned out to be independent from "Severity" as a result of the test of independence (see the Appendix). The latter was unreliable because the proportion of the rounded numbers, like 5 minutes and 10 minutes, was greater than the others (see Figure A12), indicating biases in this feature.

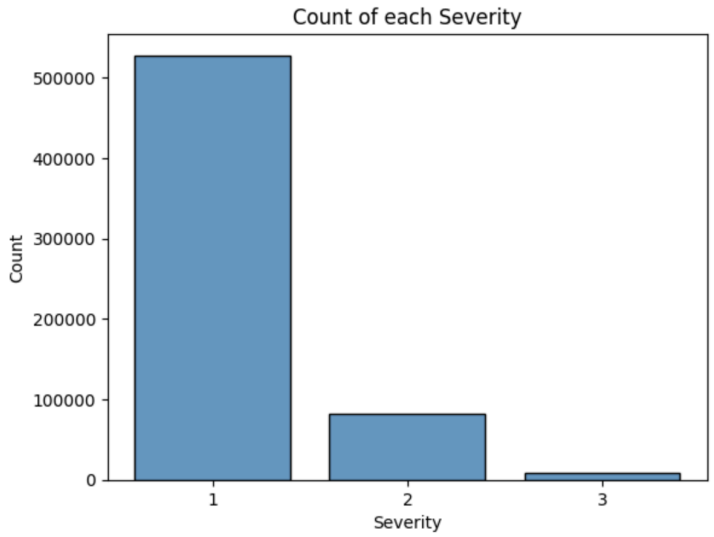
Splitting the dataset. We split the dataset into training, validation, and test data in a 6:2:2 ratio, ensuring that the proportions of each target level are equal across all the data.

Standardizing the numerical features. We standardized the numerical features in all the data using the mean and standard deviation of the numerical features in the training data.

Handling the class imbalance. The dataset had a class imbalance in the target classes as shown in Figure 1. Therefore, we addressed this problem by utilizing an oversampling and undersampling method. For the oversampling method, we used the Synthetic Minority Oversampling Technique (SMOTE). For the undersampling method, we exploited the Random Under Sampler. In general, oversampling methods have a risk of causing overfitting, whereas undersampling methods tend to lead to data losses. In this project, we developed models using the oversampled and undersampled data and compared them.

Figure 1

Class imbalance in the target variable "Severity"



Encoding the categorical features. We intended to develop four machine-learning models: CatBoost, Random Forest, Decision Tree, and Feedforward Network (FNN) models. Except for when developing the CatBoost model, which does not require the encode of categorical features, we encoded the categorical features. For "CRASH_DATE_month", "CRASH_DATE_hour", and "CRASH_DAY_OF_WEEK", we utilized the cyclic encoding, which preserves their cyclic characteristic (Axel, 2023). For the other categorical features, we used the one-hot encoding (OHE).

Data Processing and Feature Engineering for GLM

Removing features. We aimed to develop GLMs to estimate the probability distribution of the count of traffic crashes in several community areas in a 2-hour interval using month, day of the week, timeframe, severity, and community area as independent variables. Therefore, we removed all the features except for those related to the above variables. We kept only the following features: "CRASH_DATE", "INJURIES_TOTAL", "INJURIES_FATAL", "CRASH_DAY_OF_WEEK", "LATITUDE", and "LONGITUDE".

Handling date time feature. We converted "CRASH_DATE" to six new features: "CRASH_DATE_year", "CRASH_DATE_month", "CRASH_DATE_day", "CRASH_DATE_hour", "CRASH_DATE_minute", "CRASH_DATE_second".

Removing observations prior to September 2017. The data in our dataset started being collected in 2015, but the data from 2015 to August 2017 were collected from a part of the city. Since keeping the data from this period might cause biases, we removed them.

Handling missing values. There was no feature with more than 50% of the proportion of missing values. We just removed rows with any missing values.

Removing constant features. We checked whether there were any features whose value was constant. "CRASH_DATE_second" turned out to have constant values, which led to their removal.

Removing unreliable observations. Some observations had a latitude and longitude of 0, which must have been erroneous because all the traffic crashes occurred in Chicago. We removed all these observations.

Feature engineering. We created "Severity" in the same way as explained in the previous section. We also created a new feature named "Community area", which corresponded to community areas in Chicago. In Chicago, there are 77 community areas, and we designated the corresponding community area to each observation according to its longitude and latitude. A few observations did not fall into any community areas, and they were removed.

Aggregating the dataset. We aggregated the data so that each row has a year, month, day of the week, timeframe, severity, community area, and the count of traffic crashes as its features. We utilized this data for exploratory data analysis. When developing GLMs, we removed the

year and encoded the categorical features, which were all the features except for the count of traffic crashes, using the OHE.

Data Modeling and Visualizations

Data Modeling and Visualizations for Classification Models

Data Modeling. We planned to construct four types of machine learning classification models to predict the severity level of traffic crashes: CatBoost, Random Forest, Decision Tree, and FNN models. We trained each type of model using the oversampled and undersampled training data.

In the validation process, we utilized GridSearchCV for the Random Forest and Decision Tree models. We changed “min_samples_split” and “min_samples_leaf” in the Decision Tree models. In the Random Forest models, we tuned “n_estimators” in addition to the above two hyperparameters. For the CatBoost models, we could not use GridSearchCV because of the lack of memory. Therefore, we built several models with different hyperparameters, which were “iteration”, “learning rate”, “l2_leaf_reg”, and “depth”, and checked the macro F1 scores on the validation data. For the FNN models, we also built several models with different numbers of layers and nodes and checked the losses on the validation data.

In the evaluation process, we looked into accuracy, macro precision, macro recall, and macro F1 score for each model. Because of the class imbalance in the target, we focused on the macro F1 score to select the best model.

Visualizations. We created several figures to see the dataset characteristics. In the following, we showed a few figures. Figure 2 shows the number of traffic accidents by month, day of the week, and hour for each severity. Notably, Severity 1 and 2 accidents often happened

on Fridays, but Severity 3 crashes frequently occurred on Saturdays and Sundays. Figure 3 shows the geographical distribution of car accidents from 2015 to January 2024. The more severe crashes are, the more localized their distributions are. Figure 4 shows how the proportion of each level in "WEATHER_CONDITION" changes by severity. There are notable differences, although the proportions of "RAIN" and "CLOUDY/OVERCAST" in Severity levels 2 and 3 are slightly larger than those in Severity level 1. Figures like Figure 4 for other features are shown in the Appendix.

Figure 2

The number of traffic crashes by month (left column), day of the week (middle column) and hour (right columns). The first, second and third row contains only Severity 1, 2 and 3 accidents, respectively. The value of 1 on the day of the week stands for Sanday.

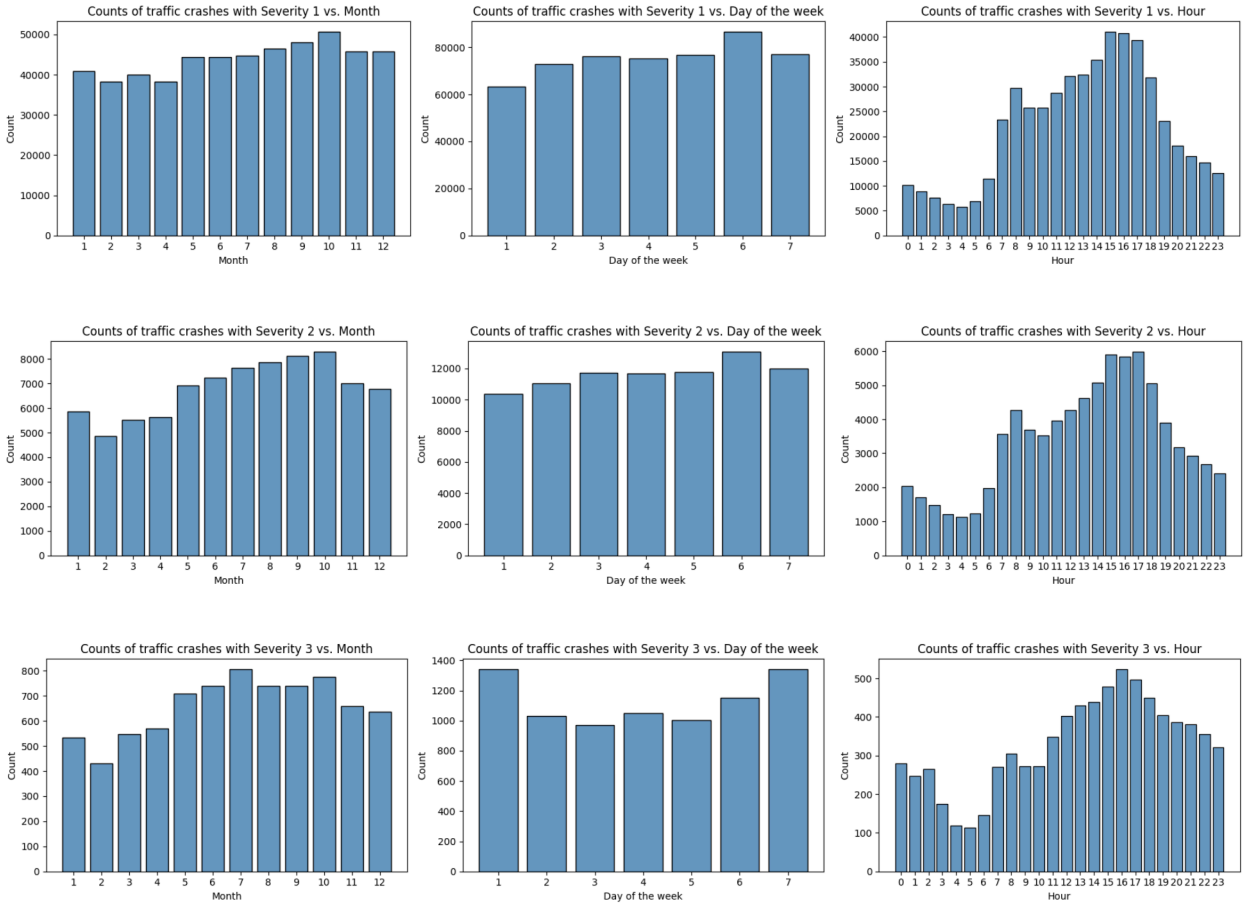


Figure 3

Geographical distribution of traffic crashes from 2015 to January 21, 2024. Green, yellow and red dots stand for Severity 1, 2 and 3 accidents, respectively.

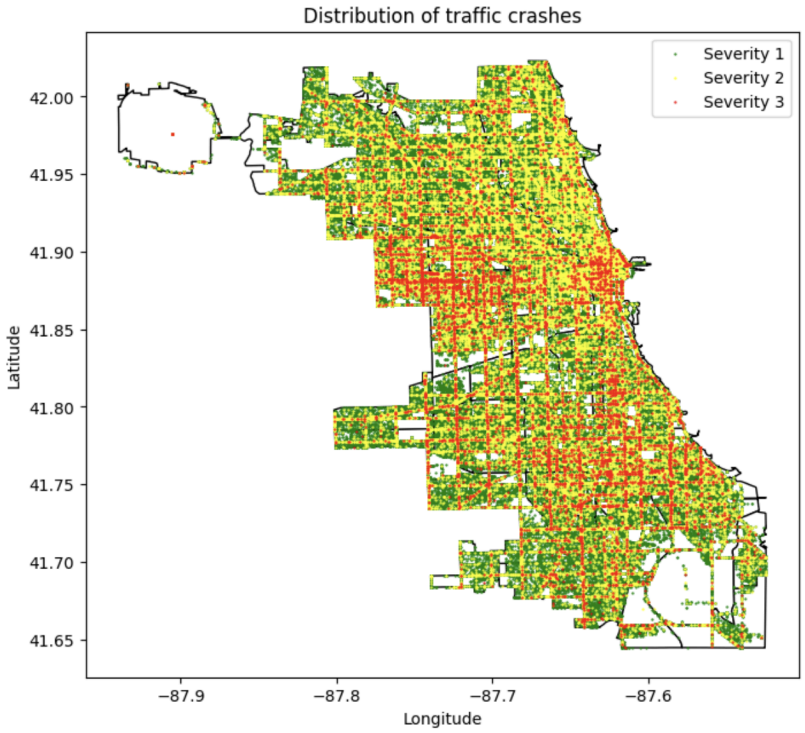
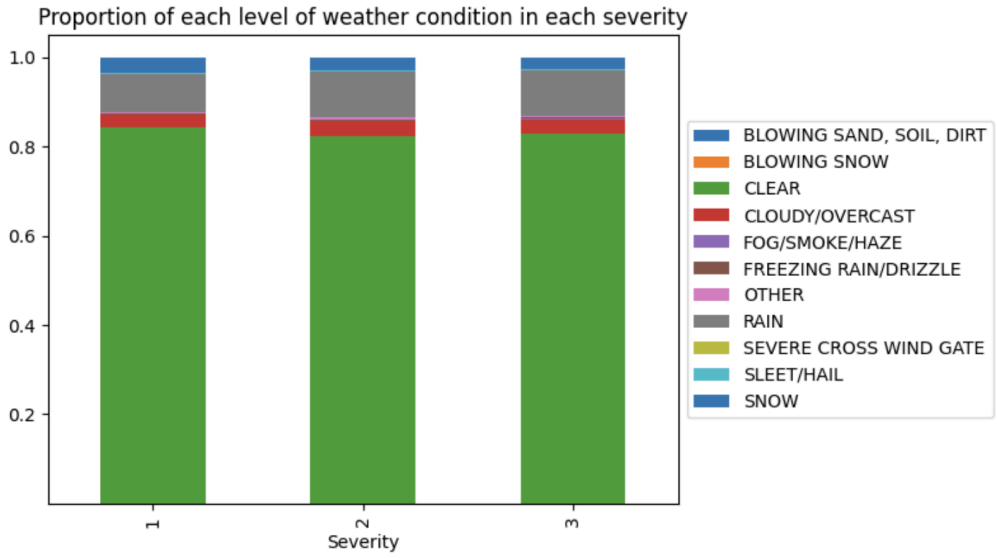


Figure 4

Proportion of each level of weather condition in each severity category.



Data Modeling and Visualizations for GLM

Data Modeling. We modeled the distribution of the count of traffic crashes in a 2-hour interval using Poisson, Negative Binomial, and Zero-Inflated models. In these models, independent variables were the month, day of the week, timeframe, severity, and community area. The number of community areas is 77, leading to the expansion of the dimension. Therefore, we only focused on the five community areas that accounted for 20% of Chicago's total number of accidents: Austin, Near West Side, Near North Side, Loop, and West Town. Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) were calculated for each model, and the model with the smallest AIC and BIC was selected as the best model.

Visualizations. The following figures show how the distribution of the count of traffic crashes varies by each independent variable. Figure 5 shows Austin has a more frequent number of crashes = 1 and 2 compared to Near West Town, indicating the distribution of the count of car accidents varies among community areas. From Figure 6, it is evident that each severity level crash has a different distribution. Figure 7 suggests the variance of the distributions by the day of the week. Figure 8 shows how the month has an impact on the distribution of the count of car crashes, although the differences in the distribution are not as evident as the other independent variables. Figure 9 indicates that the shape of the histogram also changes by the time frame.

Figure 5

Histogram of the count of traffic crashes in a 2-hour interval in Austin and Near West Side.

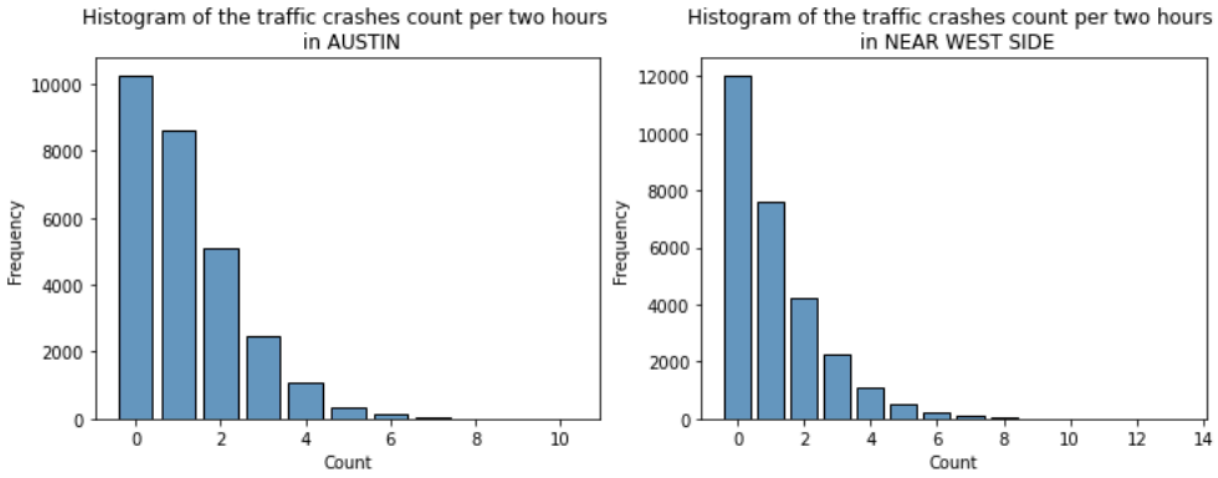


Figure 6

Histogram of the count of traffic crashes in a 2-hour interval in Austin for each severity.

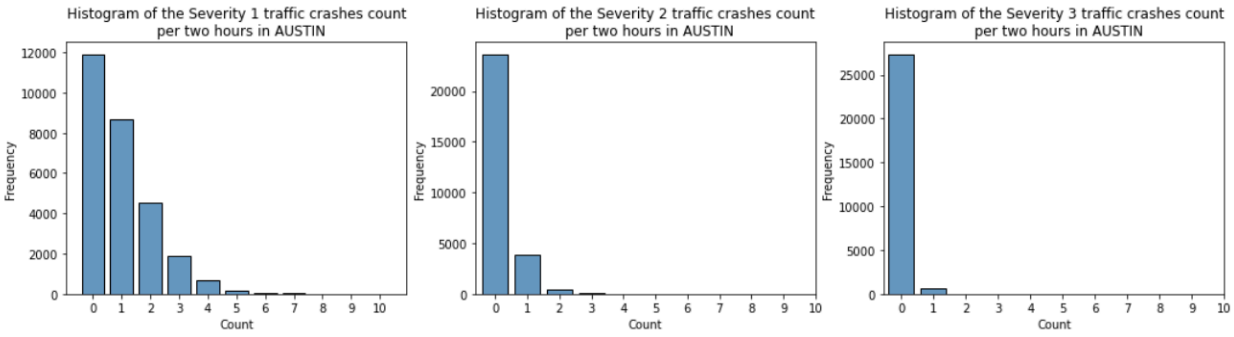


Figure 7

Histogram of the count of traffic crashes in a 2-hour interval in Austin for each day of the week. The value of 1 on the day of the week stands for Sunday.

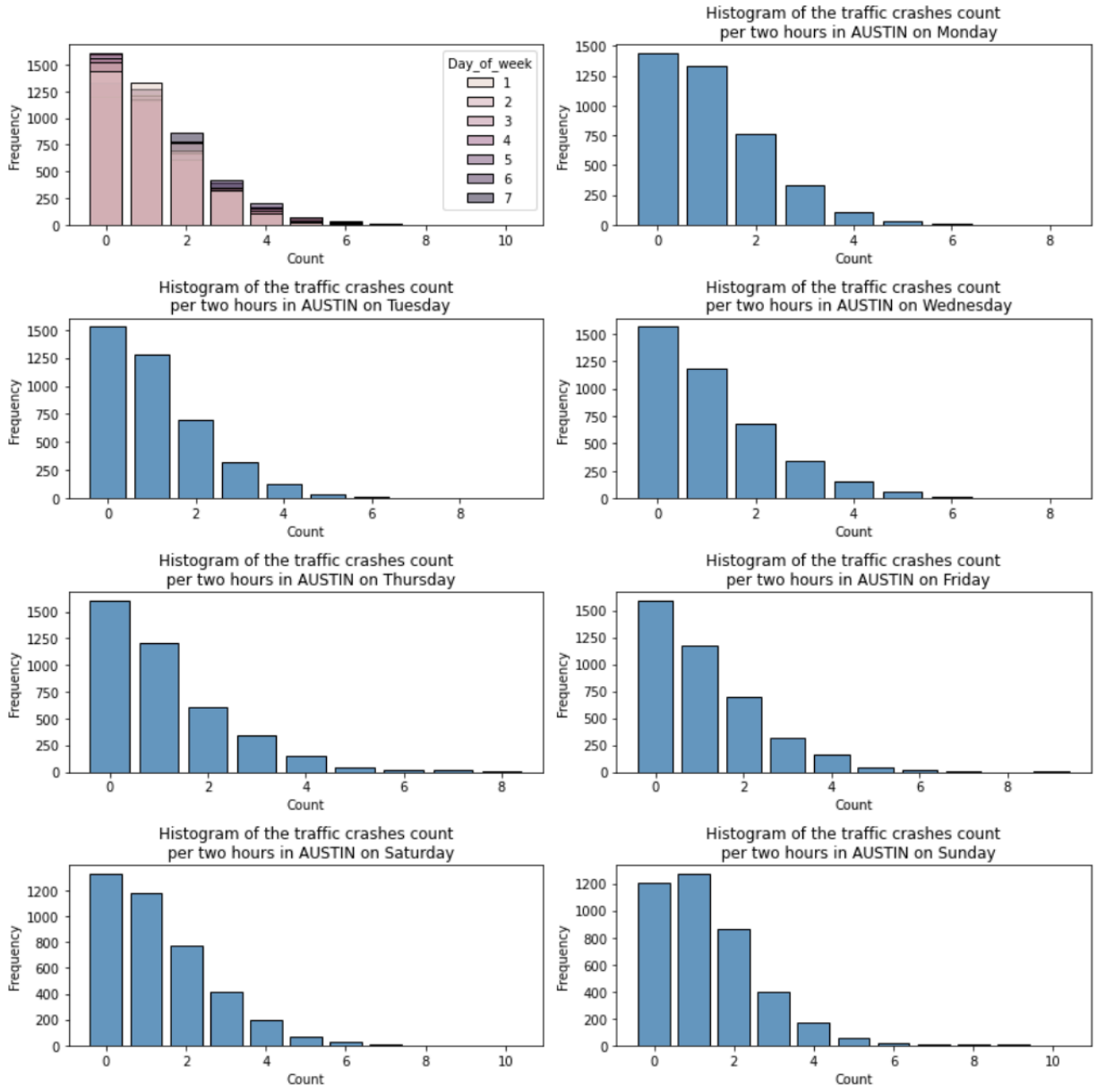


Figure 8

Histogram of the count of traffic crashes in a 2-hour interval in Austin for each month.

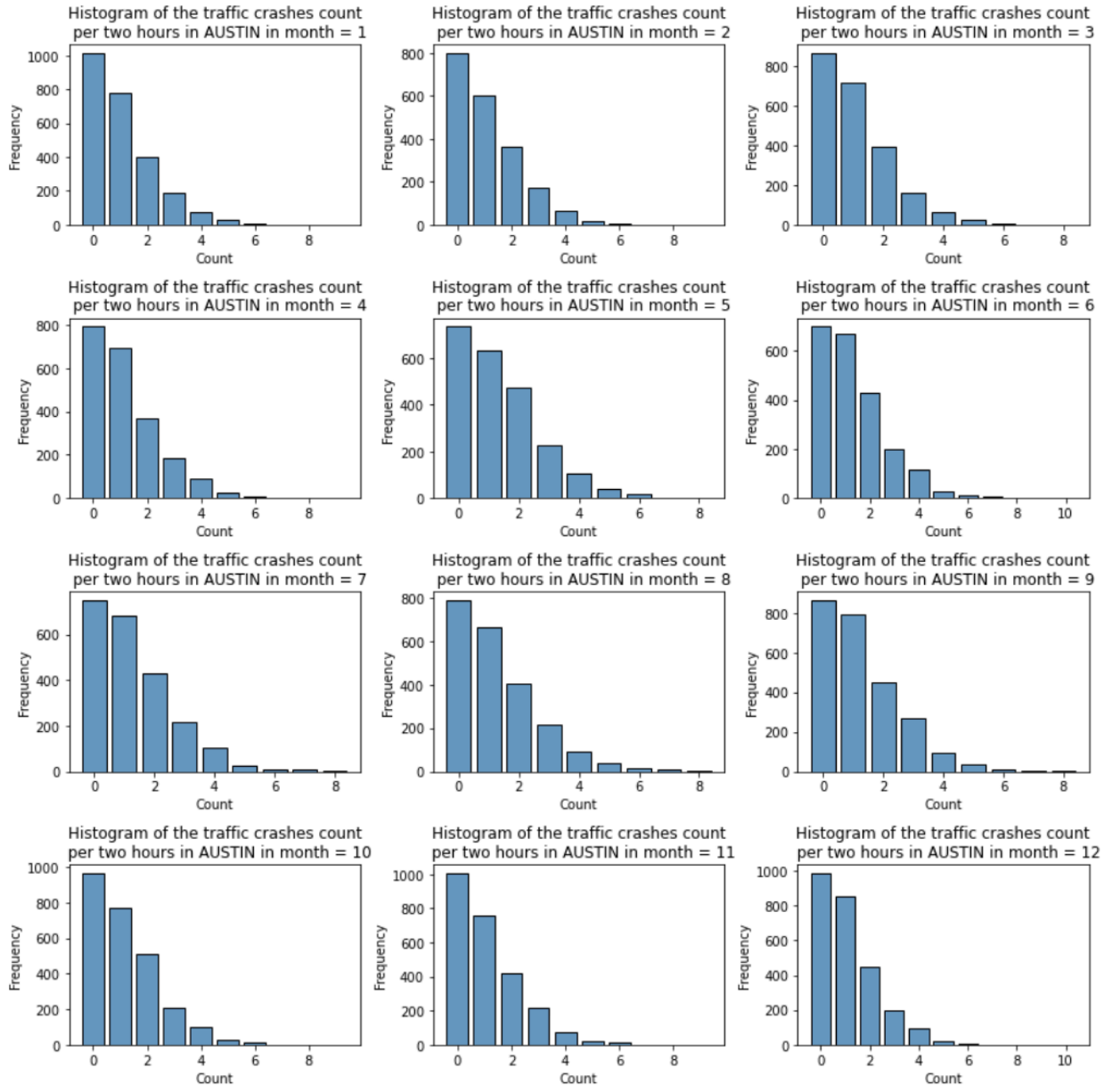
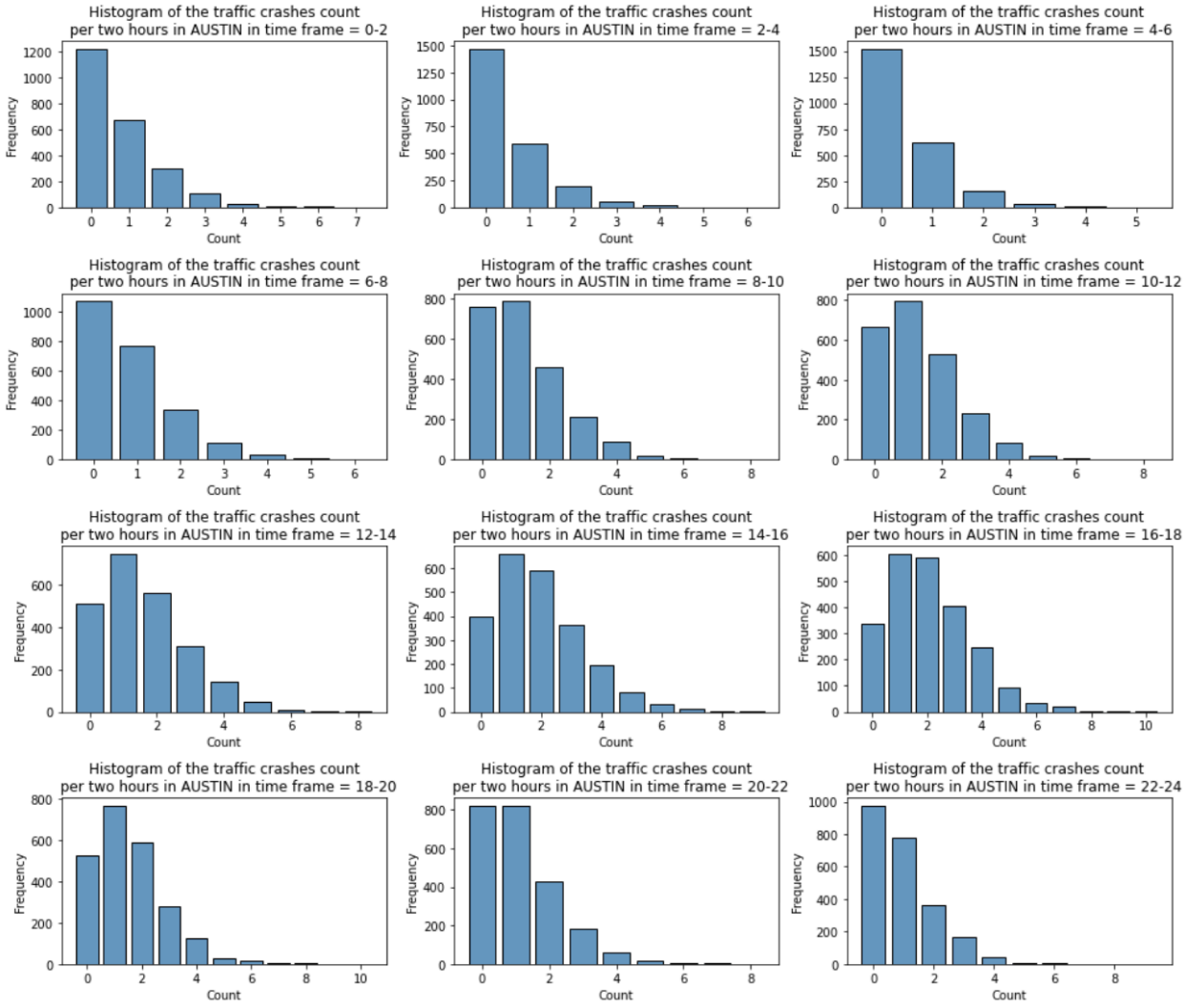


Figure 9

Histogram of the count of traffic crashes in a 2-hour interval in Austin for each timeframe.



Results and Analysis

Results and Analysis of classification models

Tables 2 and 3 summarize the evaluation metrics from models trained from the undersampled and oversampled data, respectively. Both tables indicate that the ensemble models, the CatBoost and Random Forest models, are better than the other models based on the macro F1

score. Given the class imbalance in the target variable, we regarded the macro F1 score as more important than the other metrics. This led to picking the CatBoost model trained from the undersampled data as the best model. The best model's F1 score, 0.480, was a bit better than that from Ghandour et al. (2022), 0.44, although their model was a binary classification model.

Table 2

Summary of evaluation metrics from classification models trained from the undersampled data.

	CatBoost	Random Forest	Decision Tree	FNN
Accuracy	0.876	0.698	0.585	0.691
Precision (macro)	0.445	0.567	0.557	0.452
Recall (macro)	0.613	0.440	0.421	0.562
F1 (macro)	**0.480**	0.427	0.381	0.420

Table 3

Summary of evaluation metrics from classification models trained from the oversampled data.

	CatBoost	Random Forest	Decision Tree	FNN
Accuracy	0.811	0.834	0.708	0.693
Precision (macro)	0.461	0.449	0.464	0.409
Recall (macro)	0.447	0.462	0.404	0.477
F1 (macro)	0.452	0.455	0.408	0.409

We looked into the feature importances of the best model, shown in Figure 10. The most significant feature was "FIRST_CRASH_TYPE", which indicates the type of first collision, and "LATITUDE", "LONGITUDE", "TRAFFICWAY_TYPE", and "NUM_UNITS", which is the

number of units involved in the crash, follow. Figure 11 shows how the proportion of each level in "FIRST_CRASH_TYPE" changes by severity. Severity 1 crashes had a more significant proportion of "PARKED MOTOR VEHICLE" than the other levels. Severity 2 accidents had a notable percentage of pedestrian-related crashes compared to the others. Severity 3 traffic crashes happened more often when the first crash type was "ANGLE" and "TURNING". As shown in Figure 3, more severe crashes tended to be localized than less severe ones, which makes the fact reasonable that the best model had high feature importances for "LATITUDE" and "LONGITUDE". Figure 12 shows the variation of the proportion of levels in "TRAFFICWAY_TYPE" in each severity level. The proportion of "ONE-WAY" and "PARKING LOT" is larger in Severity 1 crashes than in other severities. In contrast, the "FOUR WAY" proportion increases as the severity level increases. Figure 13 displays the number of units involved in crashes for each severity level. As expected, the number of units tends to increase as the severity level increases.

All the developed codes for the classification models are stored in our Github repository (https://github.com/kosawa26/Capstone/tree/main).

Figure 10

Feature importances of the best model (CatBoost trained from undersampled data).

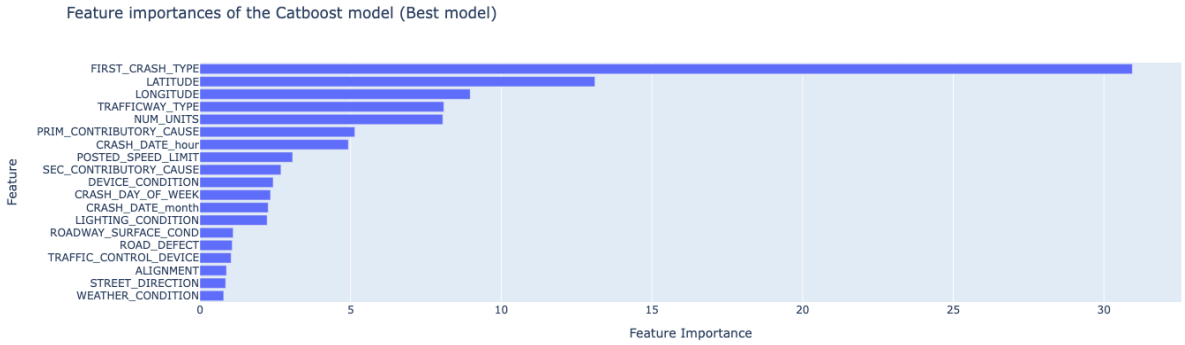


Figure 11

Proportion of each level of first crash type in each severity category.

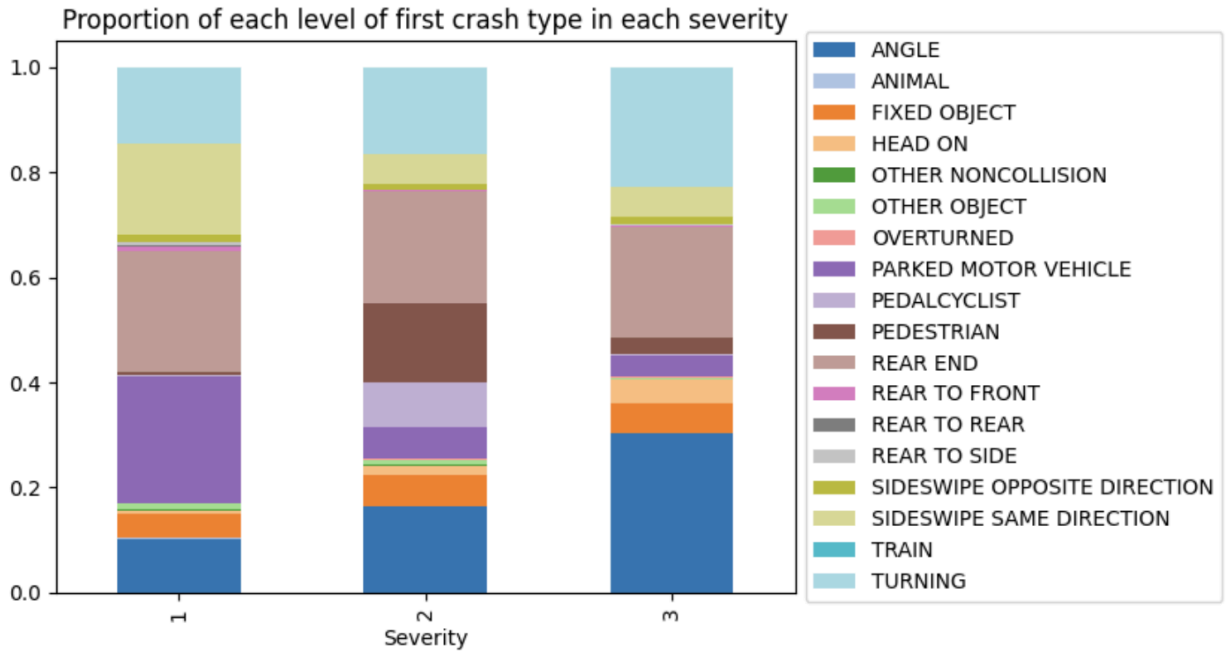


Figure 12

Proportion of each level of trafficway type in each severity category.

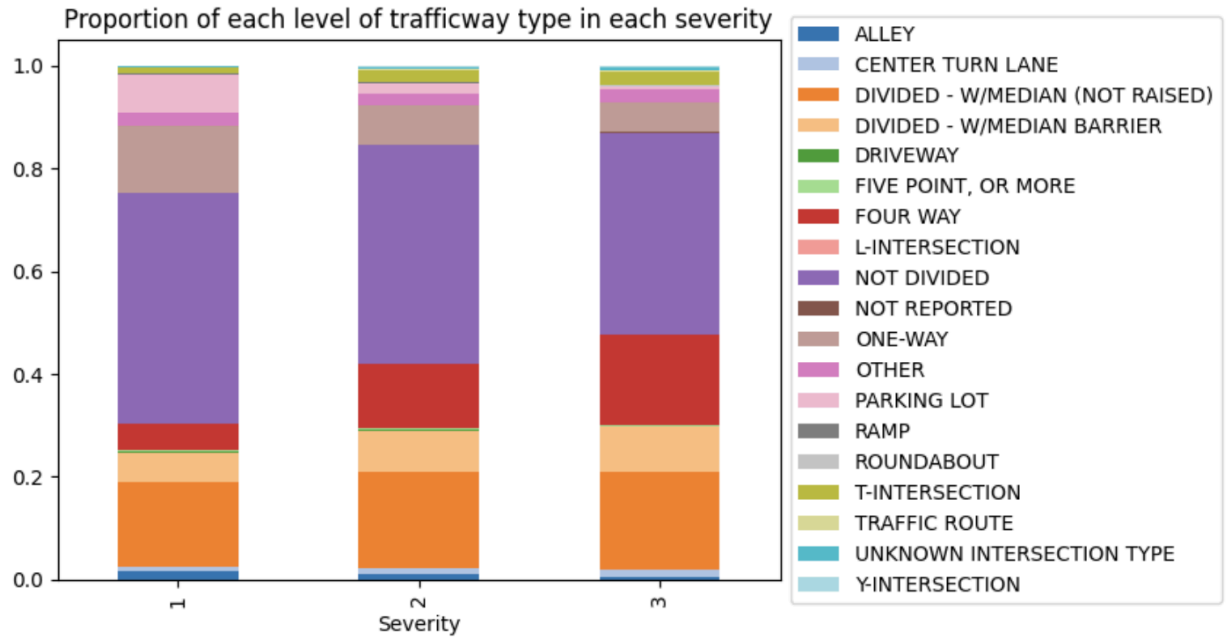
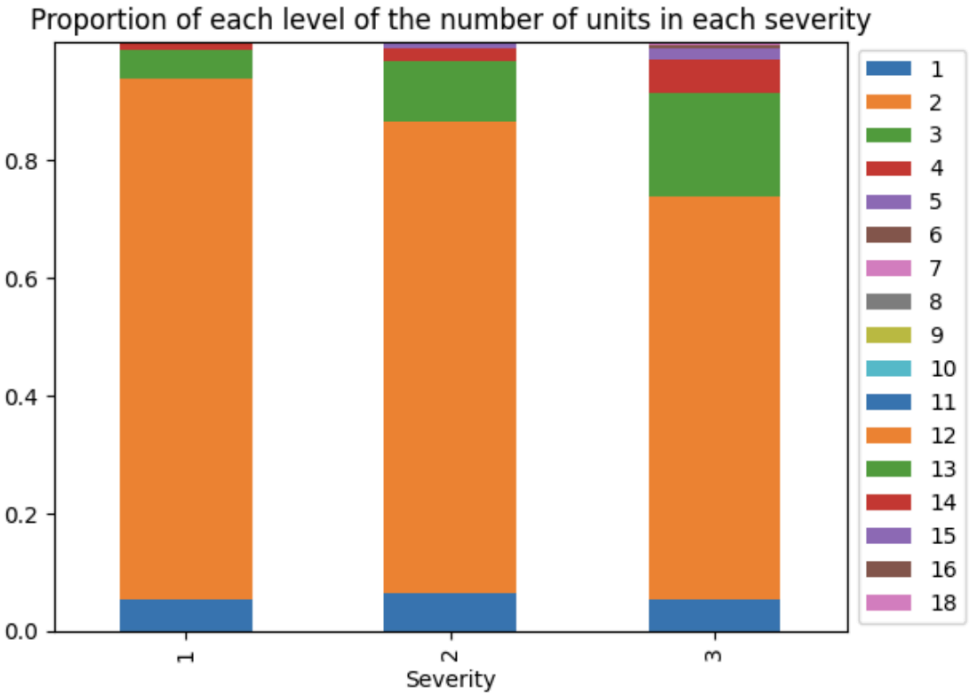


Figure 13

Proportion of each level of the number of units in each severity category.



Results and Analysis of GLM

We developed several models, in which the independent variables were different, for Poisson, Negative Binomial, and zero-inflated models. Table 4 summarizes all the combinations of independent variables explored in this project. For example, Model 1 is a full model containing constant term, month, timeframe, day of the week, severity, and community area as the independent variables. Model 2 has all the above variables except for the month. Table 5 shows AIC and BIC from each model. In terms of both AIC and BIC, the full model of the Negative Binomial model showed the best performance. Note that Negative Binomial models have a parameter α, which determines the difference between the mean and the variance. We only showed the results from the Negative Binomial model with $\alpha = 0.24$ in Table 5 because it showed the best AIC and BIC among the other Negative Binomial models. The results from the other Negative Binomial models are shown in the Appendix.

When you look at Table 5 closely, you will notice that Models 3, 5, 7, and 8 have worse AIC and BIC than the other models. All these models excluded timeframe from their independent variables, indicating that timeframe played a more important role than month and day of the week.

All the developed codes for the GLMs are stored in our Github repository (https://github.com/kosawa26/Capstone/tree/main).

Table 4

Summary of all the combinations of independent variables.

	Constant	Month	Timeframe	Day of week	Severity	Community area

Model 1	○	○	○	○	○	○
Model 2	○	×	○	○	○	○
Model 3	○	○	×	○	○	○
Model 4	○	○	○	×	○	○
Model 5	○	×	×	○	○	○
Model 6	○	×	○	×	○	○
Model 7	○	○	×	×	○	○
Model 8	○	×	×	×	○	○

Table 5

Summary of AIC and BIC

	Poisson model		NB model ($\alpha = 0.24$)		Zero-Inflated model	
	AIC	BIC	AIC	BIC	AIC	BIC
Model 1	459,117	459,500	**456,279**	**456,662**	457,517	457,900
Model 2	459,800	460,062	456,874	457,137	458,128	458,391
Model 3	493,646	493,909	485,544	485,806	485,145	485,408
Model 4	460,470	460,787	457,284	457,602	458,612	458,930
Model 5	494,329	494,471	486,152	486,294	485,699	485,842
Model 6	461,153	461,350	457,879	458,076	459,216	459,413
Model 7	494,999	495,196	486,679	486,876	486,268	486,465
Model 8	495,681	495,759	487,286	487,363	486,815	486,891

Development of dashboard

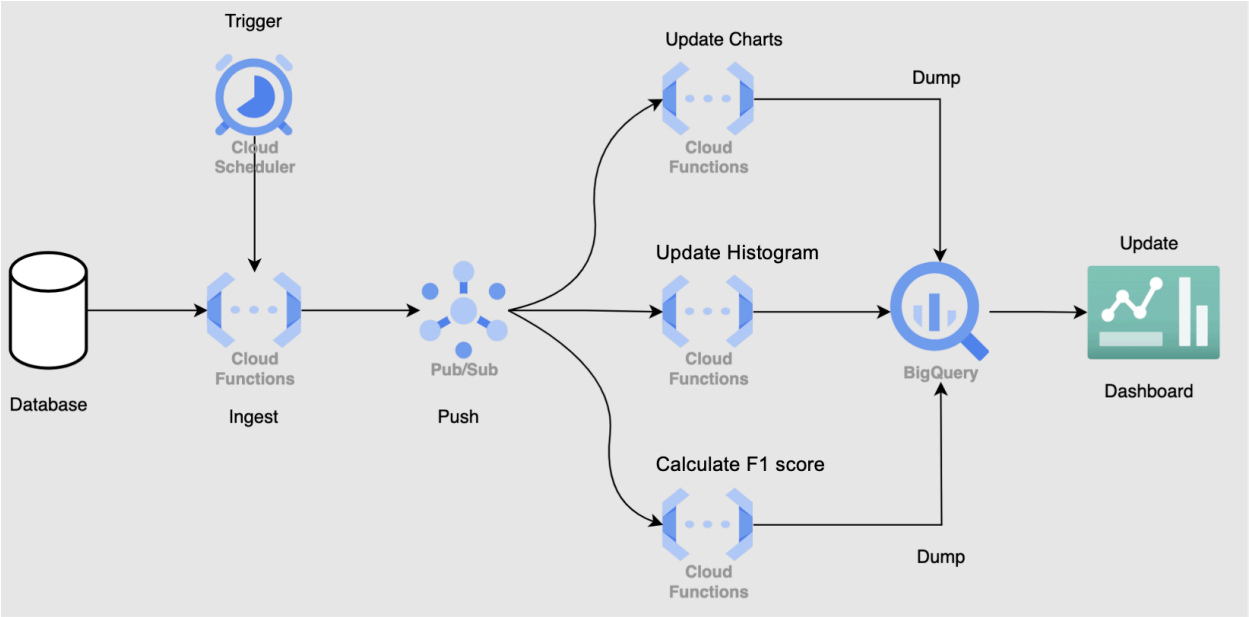
We developed a dashboard that shows the results of our EDA and models utilizing Google Cloud Platform. In addition, we created a pipeline that retrieves and processes new data automatically.

Figure 14 shows the architecture of the pipeline. In the pipeline, Cloud Scheduler invokes Cloud Functions, which retrieves the latest 1,000 observations from the Chicago Data Portal daily at 12 p.m. (CDT). After acquiring the latest data, the Cloud Functions filter the data and keep only the previous day's data. Then, the Pub/Sub component works and passes the filtered data to the other Cloud Functions, where the data is preprocessed and stored in the BigQuery.

When clients access the dashboard URL, the Cloud Run works and accesses the data stored in the BigQuery and the Cloud Storage. The Cloud Storage contains previous data that is not stored in the BigQuery. The Cloud Run creates figures from the data and shows the updated dashboard.

Figure 14

Architecture of the pipeline



Data Pipeline Architecture and Components

Cloud Scheduler. This component serves as the cron job scheduler. It activates the data ingestion process at specified time intervals, set to run daily. This regular activation ensures consistent updates to the data according to the schedule.

Cloud Functions. Cloud Functions are serverless execution environments that run code in response to events triggered by the Cloud Scheduler. The primary functions within the pipeline include extracting data from the source, transforming it, and loading it into the data warehouse. This setup provides a scalable solution for managing bursts of data processing without the overhead of server infrastructure management.

Pub/Sub. Pub/Sub acts as the real-time messaging system within the pipeline. It supports decoupled communication between different parts of the system, thereby enhancing both scalability and reliability. The configuration includes:

- **Publisher**: This component publishes messages to a topic, which, in this case, are handled by Cloud Functions responsible for data extraction.
- **Subscriber**: Subscribers receive messages from a topic. In this architecture, there are three subscribers, each corresponding to different transformation tasks like updating the charts, GLM inference and F1 score calculation in the pipeline.
- **Topics**: These serve as the channels through which messages are sent from publishers to subscribers.
- **Subscriptions**:
    - **Push**: In push subscriptions, messages are automatically sent to subscribers as soon as they are published.
    - **Pull**: In pull subscriptions, subscribers request messages from the topic when ready to process them. Push subscriptions are utilized in this architecture to ensure efficient and timely data processing.

BigQuery. BigQuery serves as the centralized data warehouse for storing all transformed data. It supports rapid SQL queries and is capable of handling large volumes of data, making it well-suited for analytical and reporting needs.

Dashboard.

- **Constructed in Dash**: Dash, a Python web application framework, is utilized for building the analytical web application that visualizes traffic crash data. It enables the creation of highly interactive user interfaces using pure Python.

- **Deployed on Cloud Run**: The Dash application is deployed on Cloud Run, a serverless platform that facilitates the running of stateless containers invokable via web requests or Pub/Sub events. Cloud Run is fully managed, which encompasses all aspects of infrastructure management, including automatic scaling in response to traffic. This deployment ensures that the dashboard is scalable and available with minimal operational overhead.

Data Flow Process

- **Trigger**: The data pipeline is initiated by the Cloud Scheduler, beginning with data extraction.
- **Extract**: Cloud Functions retrieve raw crash data from the data source, which is a public data portal or an internal database.
- **Transform**: The raw data undergoes processing and transformation to meet analytical needs. This stage involves data cleaning and aggregation.
- **Load**: The transformed data is loaded into BigQuery. This step is facilitated by a dedicated set of Cloud Functions designed to handle data dumping efficiently.
- **Dashboard Update**: Once the data is available in BigQuery, the dashboard retrieves this data to refresh its visualizations, ensuring that the displayed information is based on the most recent dataset.

The dashboard can be accessed from the following URL (https://dashapp-wnihzixt4q-uc.a.run.app/). Figure 15 displays the screen of the classification models part in the dashboard.

1. Users can examine how the proportion of each level in a feature changes with the severity levels. They can select a feature from the dropdown menu and change the period of years from the slider.
2. Users can see the geographical distribution of traffic crashes each year.
3. Users can see the monitoring result of the best model. The best model is compared to the base model, the dummy classification model, using the macro F1 score. If the metric from the best model is worse than the base model's, an alert requesting to retrain the model appears.

Figure 16 shows the screen of the GLM part in the dashboard:

1. Users can see how the distribution of the count of traffic accidents in a 2-hour interval changes by month, day of the week, or timeframe.
2. Users can check the more specific distribution of the count of traffic crashes in a 2-hour interval by selecting a community area, severity, month, day of the week, and timeframe.
3. Users can see the probability of the count of car accidents in a 2-hour interval in a specific community area, severity, month, day of the week, and timeframe.

Figure 15

Screen of the classification models part in dashboard

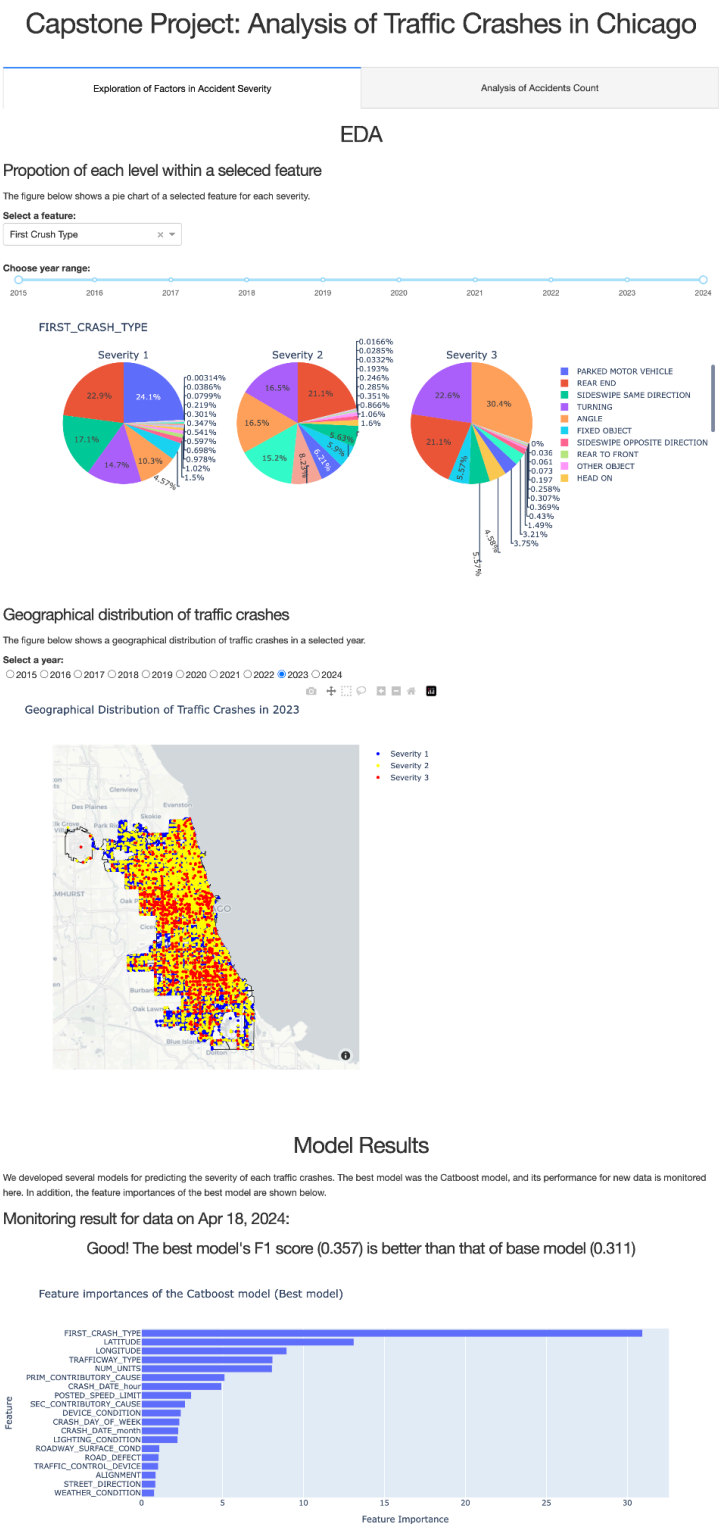
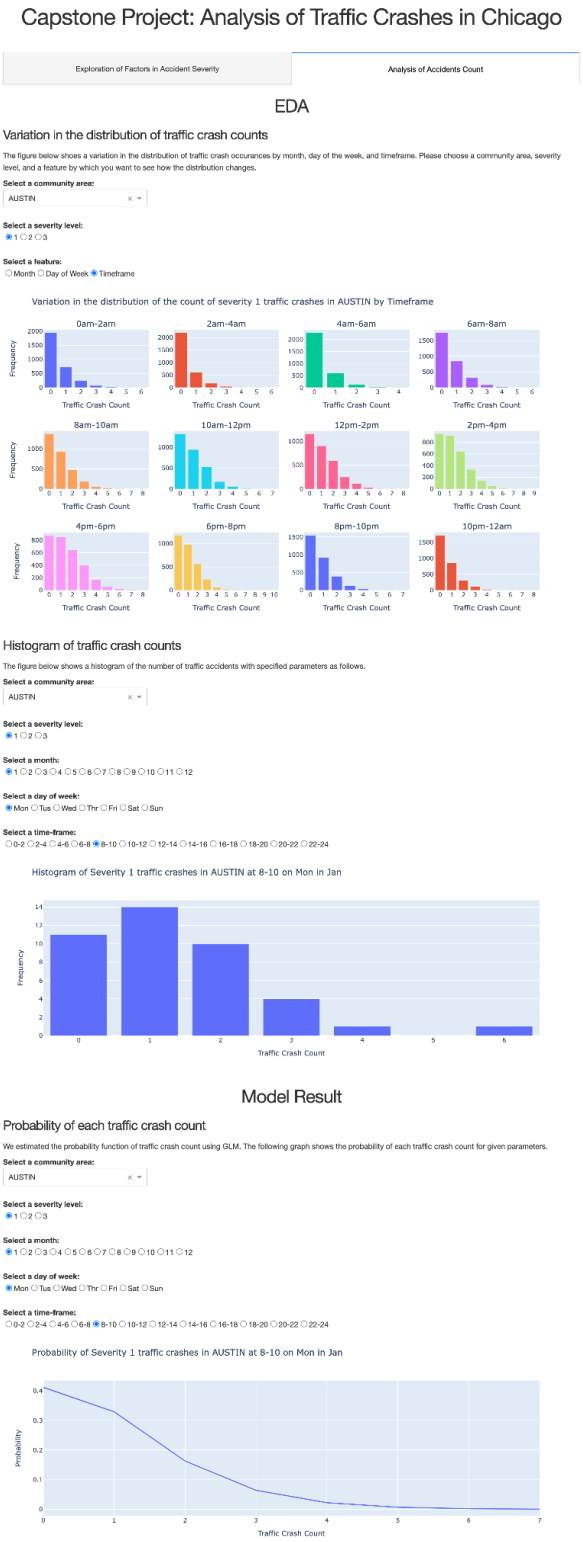


Figure 16

Screen of the GLM part in dashboard



Conclusion

Conclusion

In this project, we analyzed the traffic crash data in Chicago to find the significant factors determining the traffic crash severity and to estimate the probability of the number of traffic crashes. For the first purpose, we developed several machine learning models to predict the severity of traffic crashes and looked into the feature importances of the best model. Since severe accidents happen less often than less severe accidents, it is often the case that there is a class imbalance in the severity level. We dealt with the class imbalance in our dataset by utilizing an undersampling and oversampling method. We evaluated the models by the macro F1 score, and the CatBoost model trained from the undersampled data was the best. Among the feature importances of the best model, a feature representing the type of first crash and features related to the location of the accident were significant. Since the most severe accidents often happened when the first collision type was the side collision, and they were more localized than the less severe crashes, urging drivers to exercise extreme caution when engaging in behaviors that may cause side-impact collisions, such as entering intersections and roads, in high-accident areas could be the most effective way to reduce the most severe traffic accidents.

For the second purpose, we developed several generalized linear models assuming the distribution of the count of the traffic crashes in a 2-hour interval follows the Poisson, Negative Binomial, or zero-inflated distribution. For each assumed distribution, we created several models with different combinations of independent variables. As a result, the Negative Binomial model, which has the constant term, month, day of the week, timeframe, severity, and community area as its independent variables, showed the best performance in terms of AIC and BIC. We

estimated the probability of the count of traffic accidents in a 2-hour interval using the best model, and it can be checked in the dashboard we developed.

Project Limitation and Future Research

The dataset we exploited contained various features compared to the other open data source datasets. However, it did not have information about drivers, such as their gender, age, and driver's license type. Bhuiyan et al. (2022) listed these features as the significant factors determining the fatality level of car accidents. Therefore, using a dataset containing those features would enable us to gain more insight into the significant factors in the Chicago traffic crashes' severity.

We created a dataset used to train GLMs by aggregating the original dataset. Therefore, we were only able to keep features whose values could be identified even after aggregating. For instance, we could identify the value of the month, day of the week, and timeframe when aggregating the count of traffic crashes every two hours. On the other hand, we could not know the weather conditions in a timeframe when there was no accident in that time slot. Therefore, our independent variables were only the month, day of the week, timeframe, severity, community area as well as constant term. We might have developed better models if we had had more variables, such as weather conditions and traffic amounts. In addition, our models were large-scale models that estimated the probability of the count of traffic crashes in a specific community area. The more small-scale models, like models estimating the probability of the count of accidents in a specific street, would be more beneficial.

References

Axel Kud. (2023, December 4). *Why We Need Encoding Cyclical Features.* https://medium.com/@axelazara6/why-we-need-encoding-cyclical-features-79ecc3531232

Bhuiyan, H., Ara, J., Hasib, K. M., Sourav, M. I. H., Karim, F. B., Sik-Lanyi, C., ... & Yasmin, S. (2022). *Crash severity analysis and risk factors identification based on an alternate data source: a case study of developing country. Scientific reports, 12(1), 21243.*

Fiorentini, N., & Losa, M. (2020). *Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), 61*

Ghandour, A. J., Hammoud, H., & Al-Hajj, S. (2020). *Analyzing factors associated with fatal road crashes: a machine learning approach. International journal of environmental research and public health, 17(11), 4111.*

Lord, D., Washington, S. P., & Ivan, J. N. (2005). *Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis & Prevention, 37(1), 35-46.*

Lord, D., Guikema, S. D., & Geedipally, S. R. (2008). *Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention, 40(3), 1123-1134.*

National Highway Traffic Safety Administration. (2023, January 10). NHTSA: *Traffic Crashes Cost America $340 Billion in 2019.* https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019

National Highway Traffic Safety Administration. (2023, April). *Early Estimate of Motor Vehicle Traffic Fatalities in 2022.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813428

National Highway Traffic Safety Administration. (2023, October). *Summary of Motor Vehicle Traffic Crashes.* https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813515

World Health Organization. (2023, December 13). *Road Traffic Injuries.* https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Appendix

Statistical test

We conducted statistical tests to determine if there were differences in a feature across groups divided by Severity levels. For categorical features, we performed the test of independence using chi-square statistics as shown in Table A1. For numerical features, we performed the Kruskal-Wallis test instead of the ANOVA because all the numerical features did not have a normal distribution. The results are shown in Table A2.

Table A1

Summary of the test of independence between the target and categorical features

Feature	Chi-statistics	Critical statistics	p-value
TRAFFIC_CONTROL_DEVICE	10967	49	0.0
DEVICE_CONDITION	9732	21	0.0
WEATHER_CONDITION	540	31	0.0
LIGHTING_CONDITION	1864	16	0.0
FIRST_CRASH_TYPE	121098	49	0.0
TRAFFICWAY_TYPE	17155	51	0.0
ALIGNMENT	274	18	0.0
ROADWAY_SURFACE_COND	717	18	0.0
ROAD_DEFECT	205	18	0.0
PRIM_CONTRIBUTORY_CAUSE	30669	100	0.0
SEC_CONTRIBUTORY_CAUSE	9220	100	0.0
STREET_DIRECTION	442	13	0.0

CRASH_DAY_OF_WEEK	297	21	0.0
CRASH_DATE_month	574	34	0.0
CRASH_DATE_day	69	79	0.2
CRASH_DATE_hour	1956	63	0.0
CRASH_DATE_minute	16084	144	0.0

Table A2

Summary of the Kruskal-Wallis test between the target and categorical features

Feature	Kruskal-statistics	p-value
POSTED_SPEED_LIMIT	4183	0.0
NUM_UNITS	46989	0.0
LATITUDE	1092	0.0
LONGITUDE	164	0.0

Complement figures

We showed figures generated in the EDA process in the following.

Figure A1

Proportion of each level of traffic control device in each severity category.

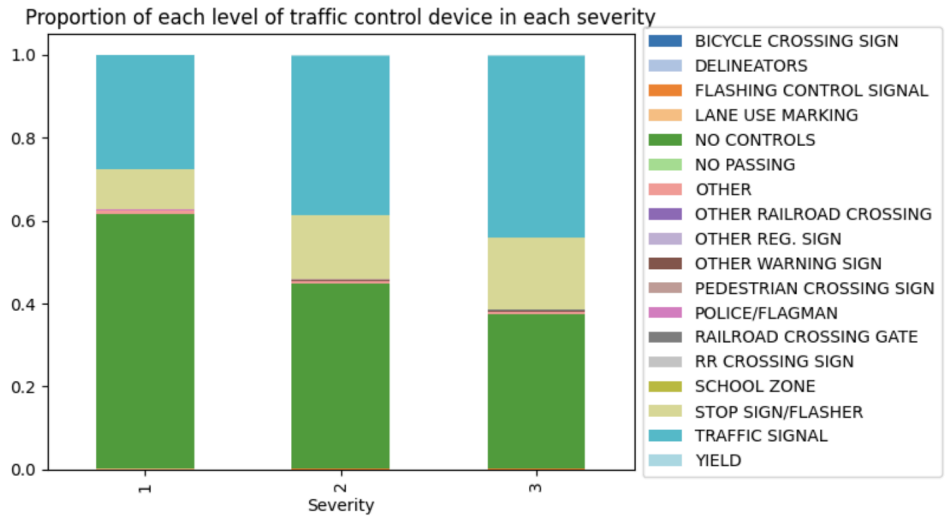


Figure A2

Proportion of each level of device condition in each severity category.

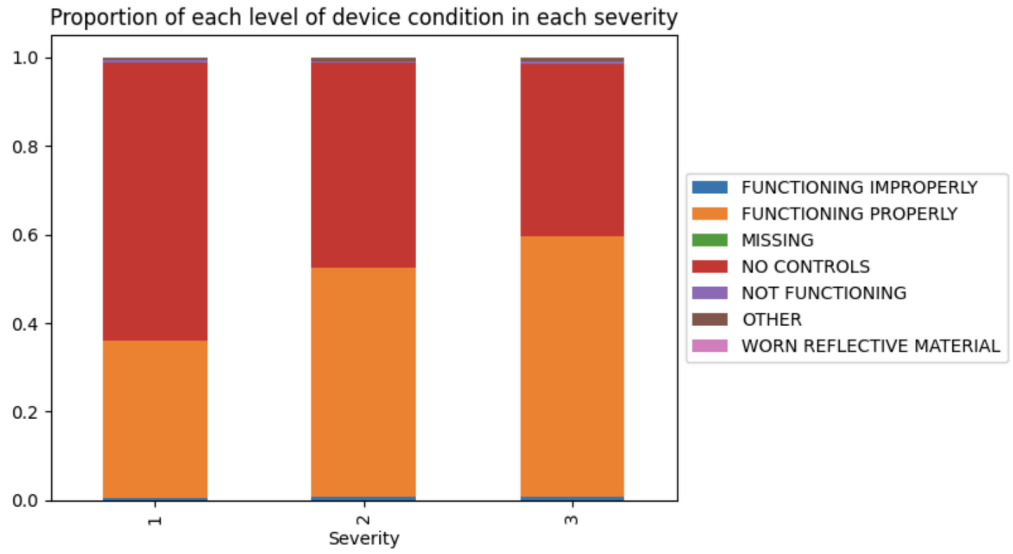


Figure A3

Proportion of each level of lighting condition in each severity category.

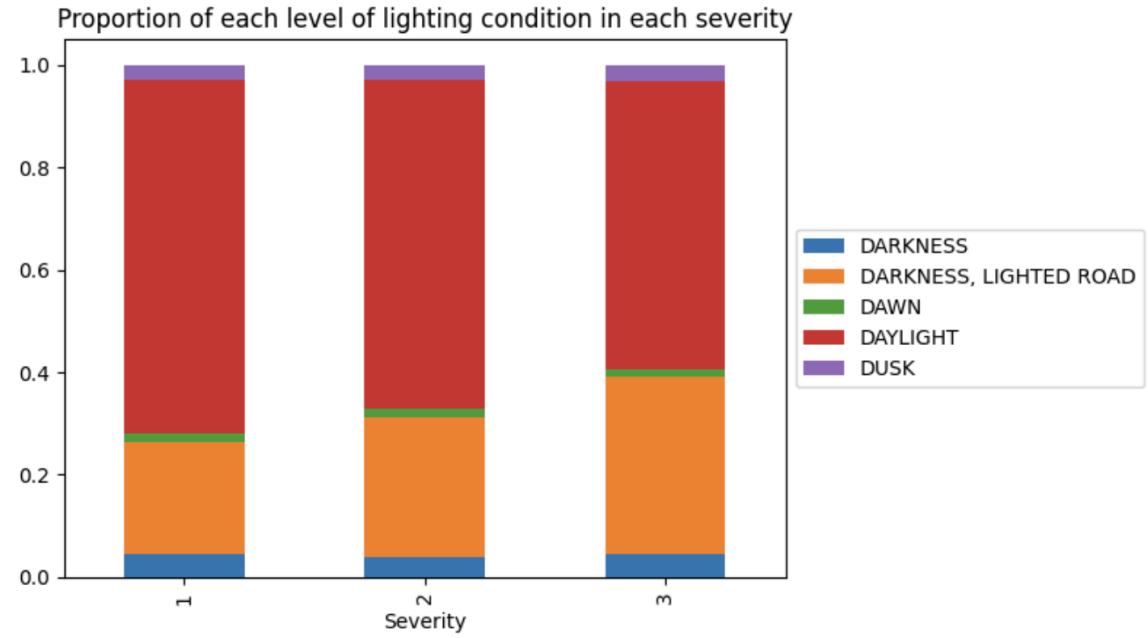


Figure A4

Proportion of each level of alignment in each severity category.

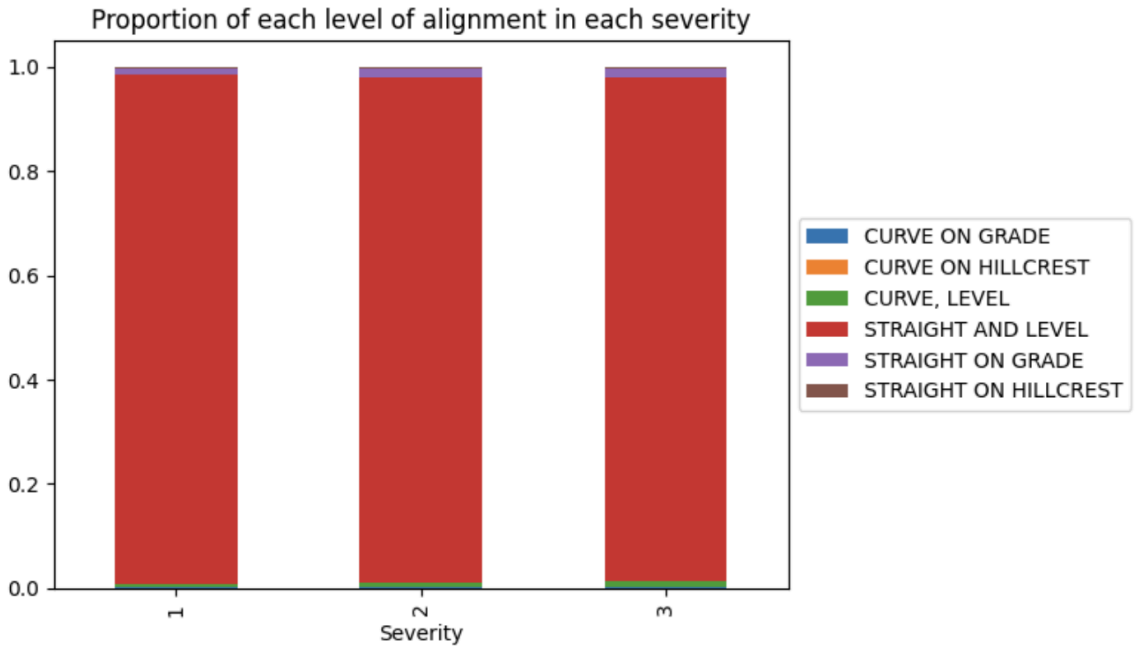


Figure A5

Proportion of each level of roadway surface condition in each severity category.

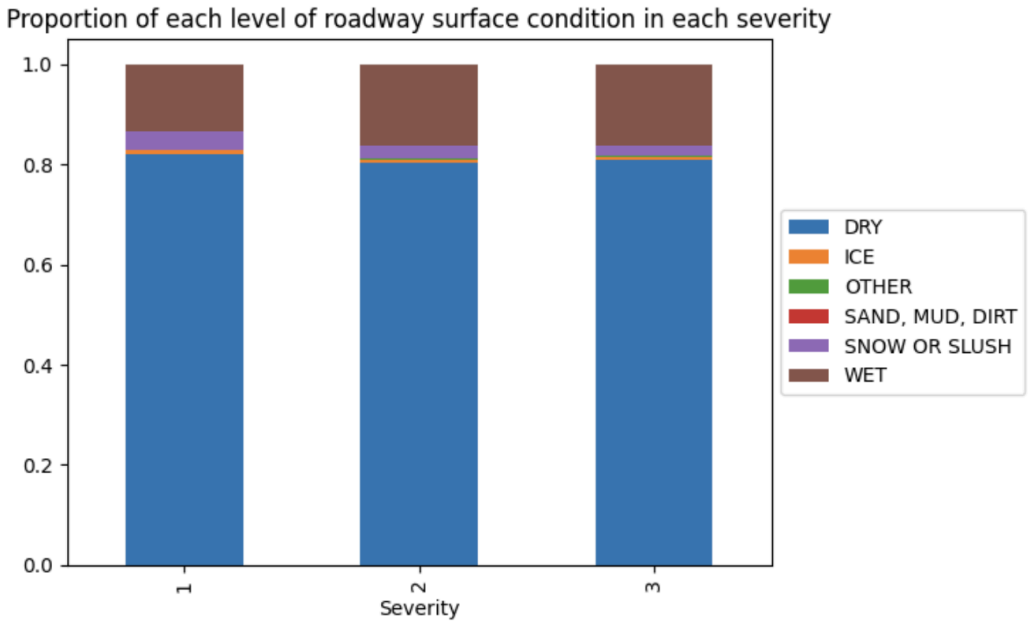


Figure A6

Proportion of each level of road defect in each severity category.

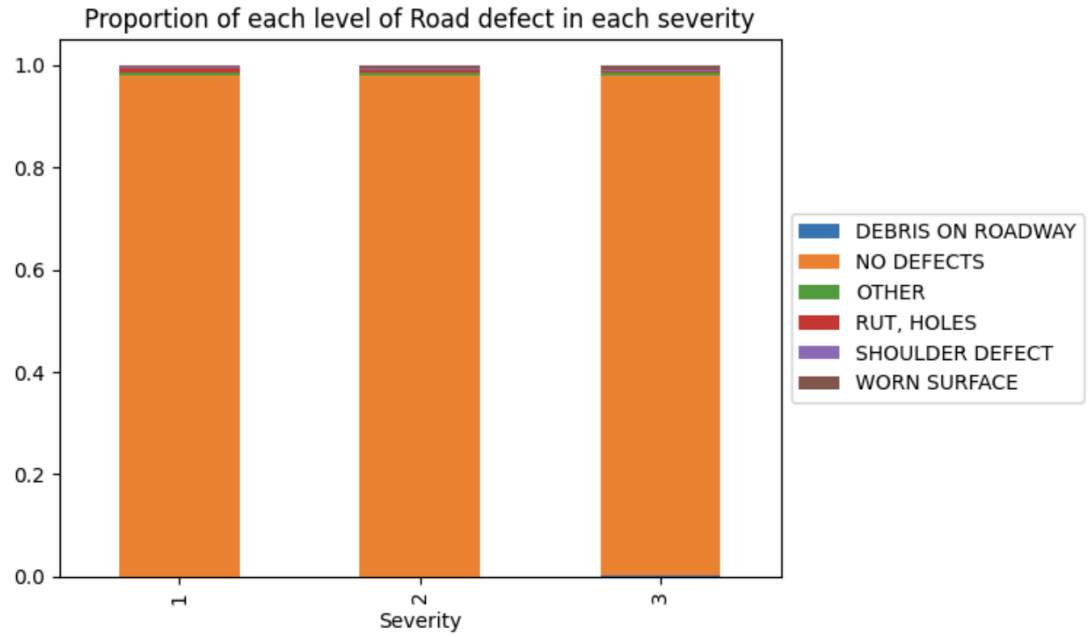


Figure A7

Proportion of each level of street direction in each severity category.

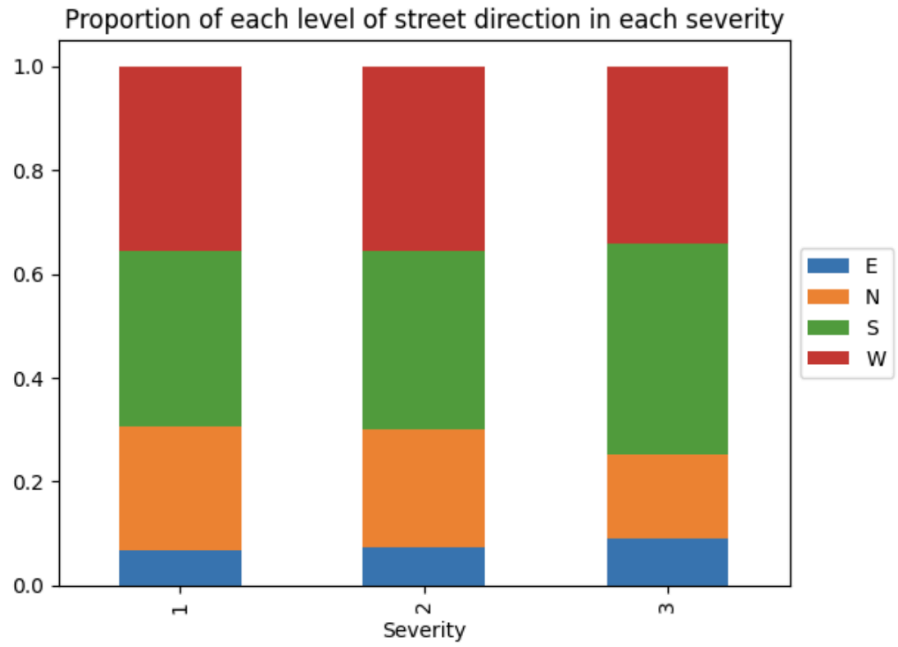


Figure A8

Proportion of each level of day of the week in each severity category.

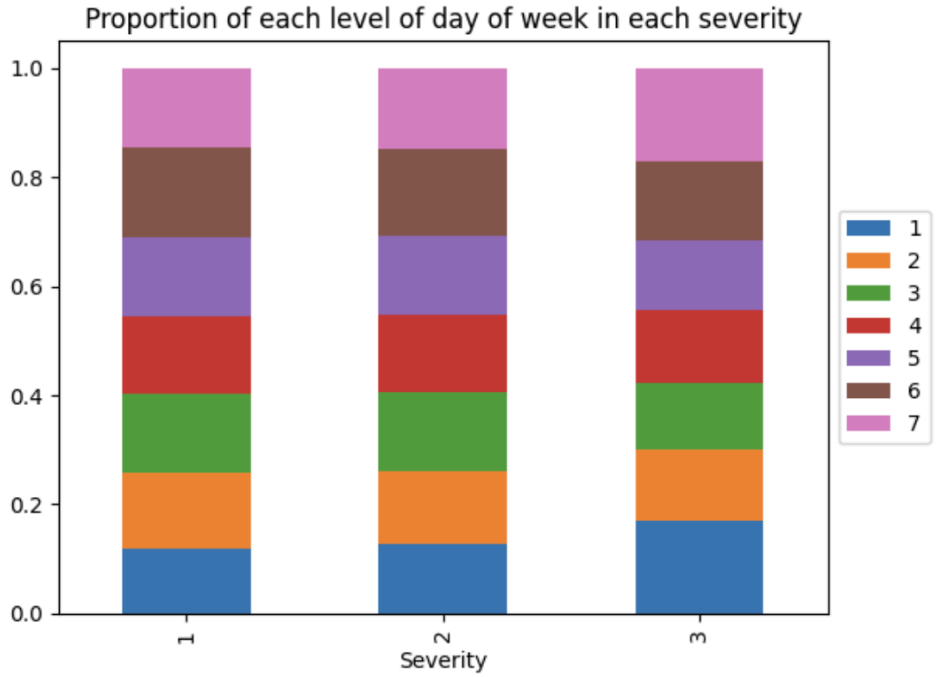


Figure A9

Proportion of each level of month in each severity category.

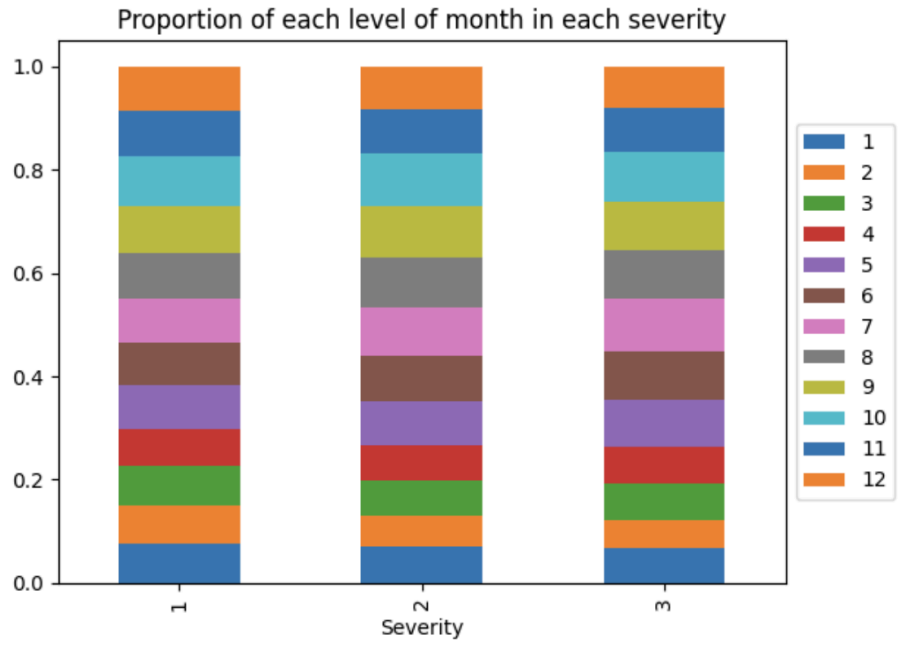


Figure A10

Proportion of each level of day in each severity category.

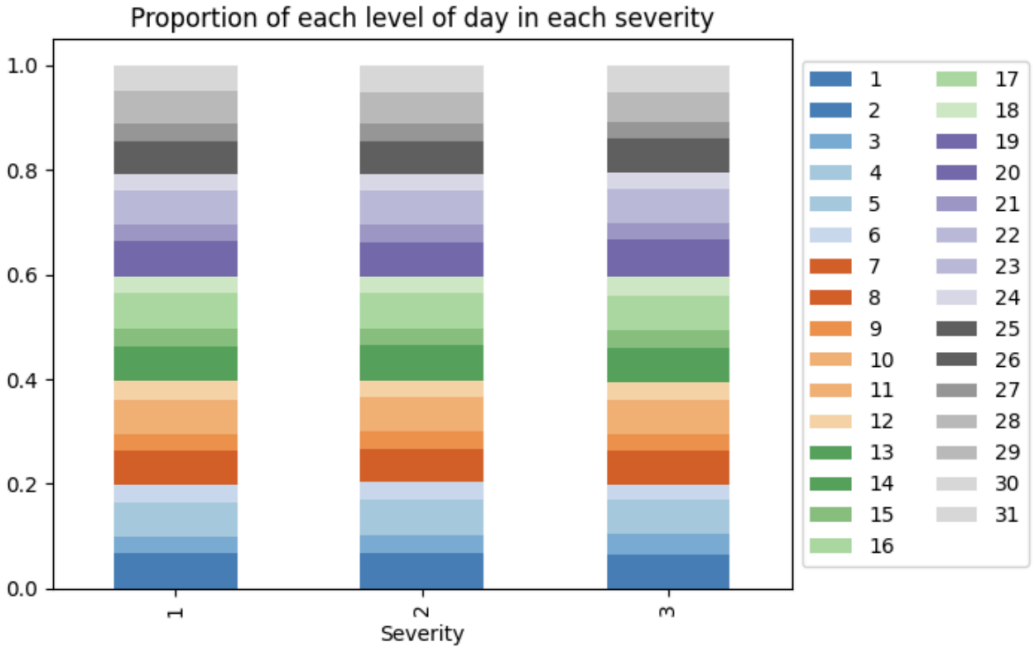


Figure A11

Proportion of each level of hour in each severity category.

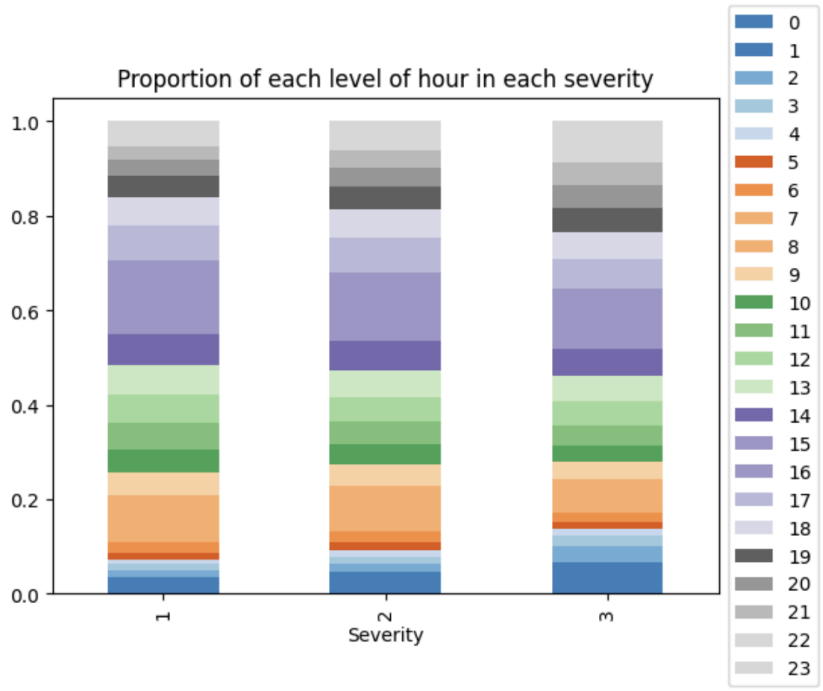


Figure A12

Proportion of each level of minute in each severity category.

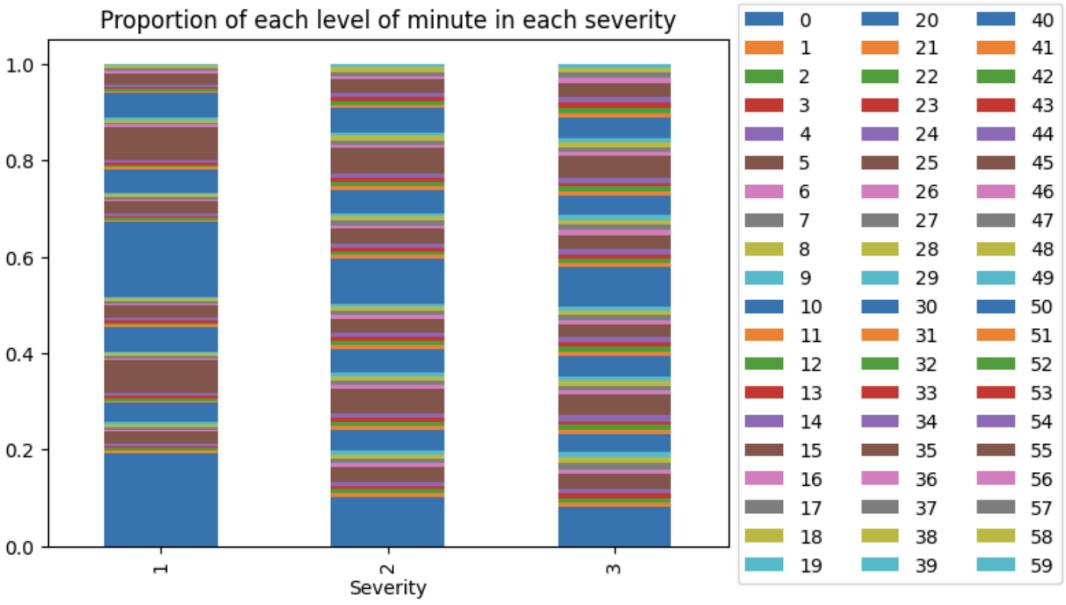


Figure A13

Proportion of each level of posted speed limit in each severity category.

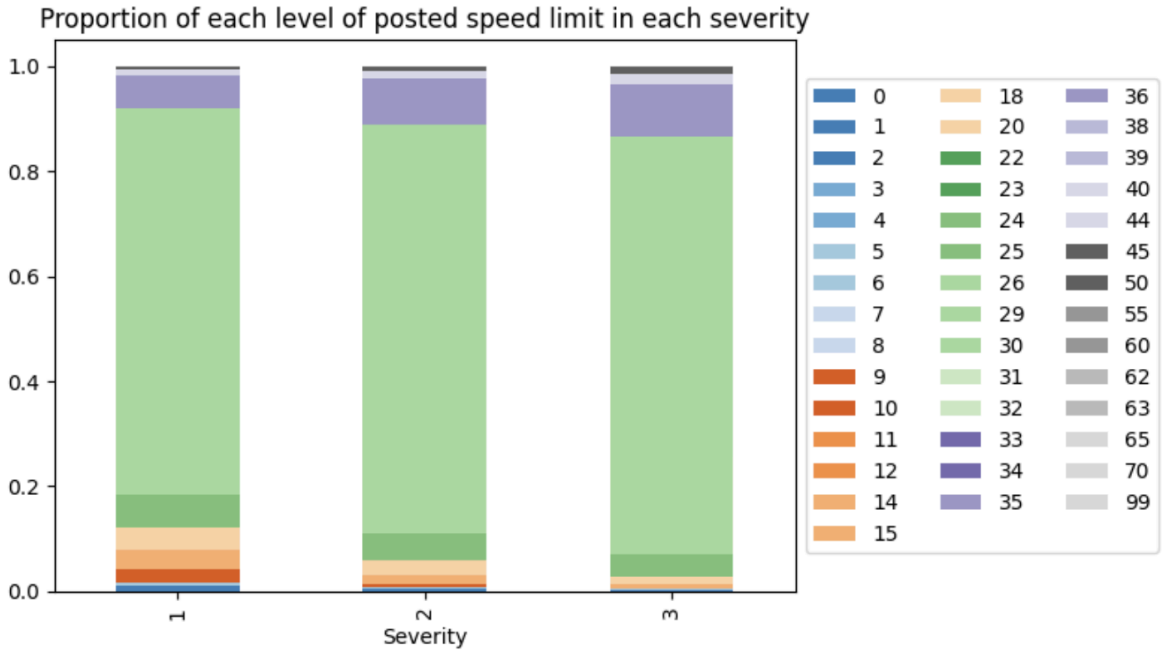


Figure A14

Proportion of each level of the primary contributory cause in each severity category.

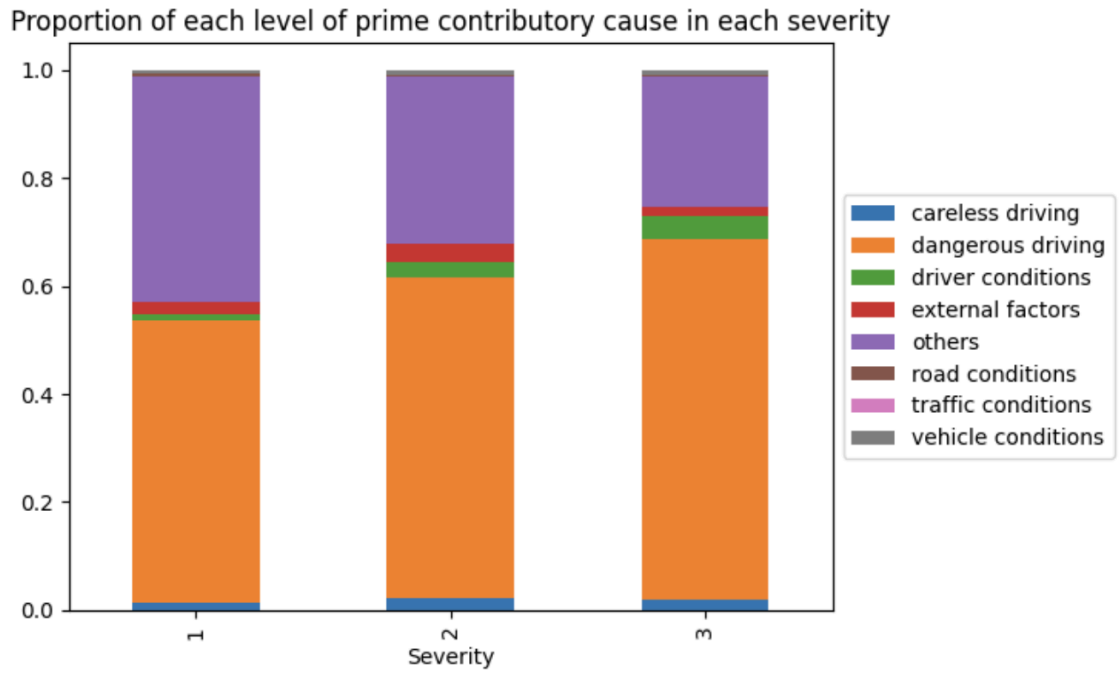
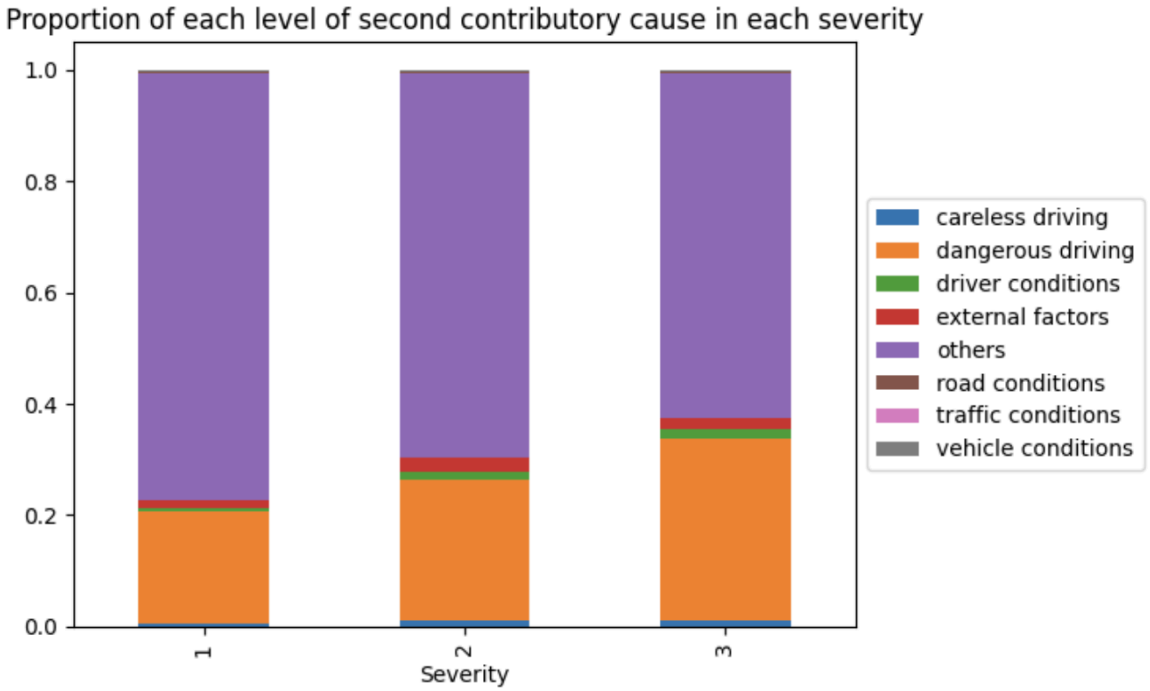


Figure A15

Proportion of each level of the secondary contributory cause in each severity category.



AIC and BIC from Negative Binomial models

As explained in the "Results and Analysis of GLM" section, Negative Binomial models have the parameter α, which determines the difference between the mean and the variance. First, We changed α from 0.1 to 2 by 0.1 and found that the AIC and BIC became the smallest at α = 0.2. Therefore, we changed α from 0.1 to 0.3 by 0.01 to find α that makes the AIC and BIC the smallest. Tables A3 and A4 show the AIC and BIC from all the developed models.

Table A3

AIC from Negative Binomial models. The α changes from 0.1 to 2 by 0.1, and then we change α by 0.01 between 0.1 and 0.3 because AIC takes the smallest value at α = 0.2.

	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
Model 1	457047	456323	456406	457016	457986	459213	460626	462175	463825	465549
Model 2	457688	456931	456985	457570	458520	459729	461125	462659	464294	466006
Model 3	489125	486331	484634	483693	483298	483311	483636	484203	484961	485872
Model 4	458229	457373	457351	457875	458775	459942	461304	462809	464420	466110
Model 5	489774	486950	485227	484264	483848	483843	484151	484703	485448	486345
Model 6	458870	457980	457929	458429	459307	460456	461800	463290	464888	466565
Model 7	490377	487497	485726	484720	484268	484230	484509	485035	485755	486632
Model 8	491025	488115	486318	485288	484815	484759	485021	485531	486238	487101

	α=1.1	α=1.2	α=1.3	α=1.4	α=1.5	α=1.6	α=1.7	α=1.8	α=1.9	α=2
Model 1	467327	469145	470990	472854	474730	476611	478494	480373	482248	484115
Model 2	467773	469580	471415	473270	475137	477010	478885	480757	482625	484485
Model 3	486905	488036	489246	490521	491849	493219	494622	496053	497505	498974
Model 4	467859	469650	471472	473314	475170	477034	478900	480765	482626	484480
Model 5	487366	488486	489686	490952	492270	493632	495027	496451	497896	499357

Model 6	468302	470083	471895	473728	475576	477431	479290	481147	483001	484849
Model 7	487633	488736	489920	491170	492475	493824	495208	496620	498055	499507
Model 8	488091	489182	490356	491597	492893	494233	495609	497013	498441	499887

	α=0.11	**α=0.12**	**α=0.13**	**α=0.14**	**α=0.15**	**α=0.16**	**α=0.17**	**α=0.18**	**α=0.19**	**α=0.20**
Model 1	456926	456817	456721	456635	456559	456494	456438	456391	456353	456323
Model 2	457563	457451	457351	457261	457183	457114	457055	457005	456964	456931
Model 3	488782	488455	488143	487845	487561	487291	487033	486787	486554	486331
Model 4	458093	457970	457860	457760	457671	457593	457525	457465	457415	457373
Model 5	489427	489097	488782	488481	488195	487921	487661	487412	487176	486950
Model 6	458730	458604	458490	458387	458295	458213	458141	458079	458025	457980
Model 7	490025	489688	489367	489061	488769	488490	488224	487970	487728	487497
Model 8	490670	490330	490006	489697	489401	489119	488850	488594	488349	488115

	α=0.21	**α=0.22**	**α=0.23**	**α=0.24**	**α=0.25**	**α=0.26**	**α=0.27**	**α=0.28**	**α=0.29**
Model 1	456301	456287	456279	456279	456285	456298	456316	456341	456371
Model 2	456906	456888	456878	456874	456878	456888	456903	456925	456953
Model 3	486119	485917	485726	485544	485371	485207	485051	484904	484765
Model 4	457340	457314	457296	457284	457280	457282	457291	457305	457325
Model 5	486735	486531	486337	486152	485977	485810	485652	485503	485361
Model 6	457943	457915	457893	457879	457872	457871	457877	457889	457906
Model 7	487277	487068	486868	486679	486498	486327	486165	486010	485864
Model 8	487892	487680	487478	487286	487103	486929	486764	486607	486458

Table A4

BIC from Negative Binomial models. The α changes from 0.1 to 2 by 0.1, and then we change α by 0.01 between 0.1 and 0.3 because AIC takes the smallest value at α = 0.2.

	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
Model 1	457430	456706	456789	457399	458370	459597	461009	462558	464208	465932
Model 2	457950	457193	457248	457833	458783	459991	461387	462921	464557	466268
Model 3	489388	486594	484897	483956	483561	483574	483899	484466	485224	486135
Model 4	458546	457691	457669	458193	459092	460260	461621	463126	464738	466428
Model 5	489916	487092	485370	484406	483991	483985	484294	484846	485590	486488
Model 6	459067	458177	458126	458626	459504	460653	461998	463487	465085	466763
Model 7	490574	487694	485923	484917	484465	484427	484706	485236	485953	486829
Model 8	491102	488192	486394	485365	484892	484835	485097	485608	486315	487178

	α=1.1	α=1.2	α=1.3	α=1.4	α=1.5	α=1.6	α=1.7	α=1.8	α=1.9	α=2
Model 1	467710	469528	471373	473238	475113	476994	478877	480757	482631	484498
Model 2	468035	469842	471678	473533	475400	477273	479148	481020	482888	484748
Model 3	487167	488298	489509	490784	492112	493481	494885	496316	497768	499237
Model 4	468176	469967	471789	473632	475488	477351	479217	481082	482943	484797
Model 5	487509	488628	489829	491094	492413	493774	495170	496593	498038	499704
Model 6	468500	470280	472092	473925	475773	477628	479487	481344	483198	485046
Model 7	487830	488933	490117	491367	492672	494021	495405	496817	498252	499500
Model 8	488168	489259	490433	491674	492969	494309	495685	497090	498518	499963

	α=0.11	α=0.12	α=0.13	α=0.14	α=0.15	α=0.16	α=0.17	α=0.18	α=0.19	α=0.2
Model 1	457309	457201	457104	457018	456943	456877	456821	456775	456736	456706
Model 2	457826	457714	457614	457524	457446	457377	457318	457268	457226	457193
Model 3	489045	488717	488405	488108	487824	487554	487296	487050	486816	486594

Model 4	458411	458288	458177	458078	457989	457911	457842	457783	457733	457691
Model 5	489570	489239	488924	488624	488337	488064	487803	487555	487318	487092
Model 6	458928	458801	458687	458584	458492	458410	458338	458276	458222	458177
Model 7	490222	489885	489564	489258	488966	488687	488421	488167	487925	487694
Model 8	490746	490407	490083	489773	489478	489196	488927	488670	488425	488192

	α=0.21	**α=0.22**	**α=0.23**	**α=0.24**	**α=0.25**	**α=0.26**	**α=0.27**	**α=0.28**	**α=0.29**
Model 1	456684	456670	456663	456662	456668	456681	456700	456724	456754
Model 2	457168	457151	457141	457137	457141	457150	457166	457188	457215
Model 3	486382	486180	485988	485806	485634	485470	485314	485167	485028
Model 4	457657	457631	457613	457602	457597	457600	457608	457623	457643
Model 5	486878	486673	486479	486294	486119	485953	485795	485645	485504
Model 6	458141	458112	458090	458076	458069	458068	458074	458086	458103
Model 7	487474	487265	487065	486876	486696	486524	486362	486207	486061
Model 8	487969	487757	487555	487363	487180	487006	486840	486684	486535