WESTERN MICHIGAN UNIVERSITY

CS 5610 Advanced R for Data Science

Crash Severity Prediction

Spring 2023

INSTRUCTOR: Dr. Wassnaa Al-Mawee

AUTHORS: Asif Irfanullah Masum, Ifrat Zaman

I. INTRODUCTION

Road traffic crashes are a significant public health issue and a major cause of death worldwide. According to some statistics, approximately 1.25 million people die and 50 million people are injured every year due to vehicular crashes [1]. The economic and social burden associated with these crashes is enormous, and the costs and consequences of the losses are significant. Therefore, it is essential to develop effective interventions to prevent, or at least minimize crash-related fatalities and injuries.

To address this issue, several government agencies, including police, health departments, and education institutions, have implemented many strategies to improve road safety. A few examples of these interventions are designing safer infrastructure, integrating road safety features in transport planning, improving vehicle safety features and driver behavior. These strategies are formulated using road traffic crash data sourced from various organizations.

One such crash data was used to devise a machine learning model to predict real-time crashes in freeway work zones [2]. The study compared Convolutional Neural Network and Binary Logistic Regression using crash and traffic data from several freeways in the Los Angeles region. The Convolutional Neural Network displayed promising results with a global accuracy of 79.50% in predicting these crashes. This study shows that machine learning techniques are revolutionizing the way crash datasets are analyzed and interpreted, providing potential insights to improve road safety and inform the development of more effective interventions and policies to prevent future accidents.

One potential area of improvement in road safety is predicting freeway crash injury severity based on a number of environmental and human factors. Accurate predictions can assist first responders and medical personnels to prioritize the most seriously injured victims and provide them with the necessary medical care. These predictions can also be used by local transportation agencies to identify hazardous conditions and strategize accordingly to avoid severe road crash incidents. Therefore, the proposed project aims to develop a machine learning model that predicts freeway crash injury severity using a crash dataset. By predicting injury severity, this model will provide valuable insights to improve post-crash care for victims of road crashes and help develop accurate diagnosis and remedial measures for road traffic operational problems.

II. PROJECT OBJECTIVE

The project is divided into two components:

i. Creating a machine learning prediction model that predicts road crash injury severity based on *environmental factors* such as weather condition, lighting condition, and road condition

ii. Creating a machine learning prediction model that predicts road crash injury severity based on *human factors* such as alcohol and drug consumption, and hazardous actions taken by the driver

The decision to divide the project into two separate components was based on the different end users of the models. For example, the first component can be used by local transportation agencies and police to take necessary precautions to make roads safer in suboptimal environmental conditions (like snowy road conditions and strong winds). The second component can be developed and deployed as a consumer application that predicts whether it is safe for an individual to drive based on a few Q&As within the application interface.

III. DATASET DESCRIPTION

The freeway crash dataset used for this project is obtained from Western Michigan University Transportation Research Center for Livable Communities (TCRLC). The data was provided by the Michigan State Police, Office of Highway Safety Planning, to WMU for research and learning purposes. The dataset consists of 399,794 observations with the following features:

Environmental and roadway factors			
No. of Features	Features	Description	Note
1	injy_svty_cd	1 | Fatal injury (K) 2 | Incapacitating injury (A) 3 | Non-incapacitating injury (B) 4 | Possible injury (C) 5 | No injury (O) Null | Not Entered	The degree of injury suffered by the involved party as a result of the crash
2	crsh_id	Unique crash identifier	Unique crash identifier

3	unit_num		The number of units deployed to the scene of the event
4	invl_prty_key	Unique ID of the involved individuals	The unique identifier of the individuals involved in the event
5	prty_type	D | Driver	The type of party involved in the event
6	rdwy_area_cd		The roadway area condition assessed from 1-6
7	objectid		
8	rte_no		
9	pr		
10	mp		
11	milt_time		
12	num_unit		The number of units deployed to the scene of the event
13	crsh_type_cd	1 | Single Motor Vehicle 2 | Head On 3 | Head On-Left Turn 4 | Angle 5 | Rear End 6 | Rear End-Left Turn 7 | Rear End-Right Turn 8 | Sideswipe-Same 9 | Sideswipe-Opposite 10 | Other/Unknown Null | Not entered	The type of crash
14	wthr_cd	1 | Clear 2 | Cloudy 3 | Fog/Smoke 4 | Rain 5 | Snow/Blowing Snow 6 | Severe Wind 7 | Sleet/Hail 8 | Other/Unknown	The weather conditions at the time of the crash

		Null | Not Entered	
15	lit_cd	1 | Daylight 2 | Dawn 3 | Dusk 4 | Dark-Lighted 5 | Dark-Unlighted 6 | Other Unknown Null | Not Entered	The lighting conditions at the time of the crash
16	rd_cond_cd	1 | Dry 2 | Wet 3 | Icy 4 | Snowy 5 | Muddy 6 | Slushy 7 | Debris 8 | Other/Unknown Null | Not Entered	The road conditions at the time of the crash.
17	num_lns		The total number of lanes of the street at the site of the crash, including continuous turn lanes (excluding flare, temporary or parking lanes)
18	spd_limt		The speed limit on the primary street at the site of the crash
19	mdot_regn_cd	0 | Statewide Multi-Region 1 | Superior 2 | North 3 | Grand 4 | Bay 5 | Southwest 6 | University 7 | Metro	The MDOT region in which the crash occurred
20	lane_dprt_cd		Condition of the lane assessed between 0-3
21	vehc_dfct_cd	1 | Brakes 2 | Lights/reflectors 3 | Steering 4 | Tires/Wheels 5 | Windows 6 | Other Null | Not Entered	The defective vehicle part which was a contributing cause of the crash

22	prty_age		The involved party's age at the time of the crash
23	rstr_not_used_fail	1 | Yes 0 | No	A variable which indicates the restraint usage of the involved party at the time of the crash
24	gndr_cd	M | Male F | Female Null | Not Entered	The involved party's gender
25	hzrd_actn_cd	0 | None 1 | Speed too fast 2 | Speed too slow 3 | Failed to yield 4 | Disregard traffic control 5 | Drove wrong way 6 | Drove left of center 7 | Improper passing 8 | Improper lane use 9 | Improper turn 10 | Improper/no signal 11 | Improper backing 12 | Unable to stop 13 | Other 14 | Unknown 15 | Reckless driving 16 | Careless/negligent Null | Not Entered	The involved party's hazardous action which contributed to the crash
26	alch_susp_ind	0 | No 1 | Yes Null | Not Entered	An indicator of whether (in the opinion of the officer) the involved party had been drinking before the crash
27	drug_susp_ind	0 | No 1 | Yes Null | Not Entered	An indicator of whether (in the opinion of the officer) the involved party had been using drugs before the crash
28	year		Year the event occurred
29	traffic_volume		Volume of traffic at the time of the event

IV. ARCHITECTURE & DIAGRAM

Machine learning techniques will be implemented to complete this project. The following is a flowchart architecture for this project:

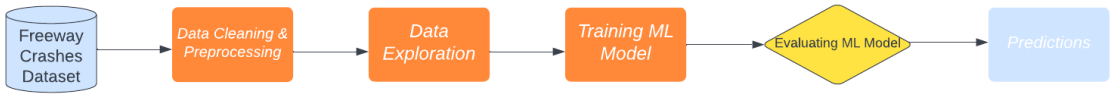


Figure 1: Project Architecture

1. **Data Cleaning and Preprocessing:** The first step of the project is to prepare the dataset for the rest of the sections. The dataset will be processed and cleaned to remove any irrelevant information and eliminate data discrepancies to ensure accurate data analysis. The dataset is checked for missing values, outliers, and errors. The clean dataset will be divided into two subsets, each with its respective feature variables; environmental factors subset and human factors subset.
2. **Data Exploration:** This section will involve the analysis of the clean data to identify patterns and relationships between variables. This will be the majority part of the project since it is crucial to understand the effect of the feature variables in crash predictions. Most of this section will involve descriptive statistics and visualization techniques, like scatter plots, histograms, and correlation matrices. Furthermore, the dataset will be split into train and test sets to enable the evaluation of the machine learning model.
3. **Model Training:** Using the Multinomial Logistic Regression algorithm, the model will be trained using the cleaned, processed training dataset for each subset.
4. **Model Evaluation:** Using the test dataset, the trained model will be evaluated in terms of accuracy scores.

V. DATA CLEANING & PREPROCESSING

The data cleaning and preprocessing is a critical step in any data analysis and machine learning project. The goal of data cleaning is to identify and correct any errors and inconsistencies in the dataset, whereas preprocessing is done to transform the data into a format that is suitable for analysis. However, since this project has two aspects to it (environmental & human factors), the first step of data cleaning and preprocessing is to create two subsets of the

dataset. This is done at this stage of the project to make sure that no observations are lost prior to the subsetting of the dataset. The following are the features in each of the subsets:

i. Environmental factors subset: *injy_svty_cd, wthr_cd, lit_cd, rd_cond_cd, mdot_regn_cd, invl_prty_key*

```
natural_factor_crash <- subset(crash_dataset, select = c(injy_svty_cd, wthr_cd, lit_cd,
rd_cond_cd, mdot_regn_cd, invl_prty_key))
```

ii. Human factors subset: *injy_svty_cd, vehc_dfct_cd, rstr_not_used_fail, hzrd_actn_cd, alch_susp_ind, drug_susp_ind,invl_prty_key*

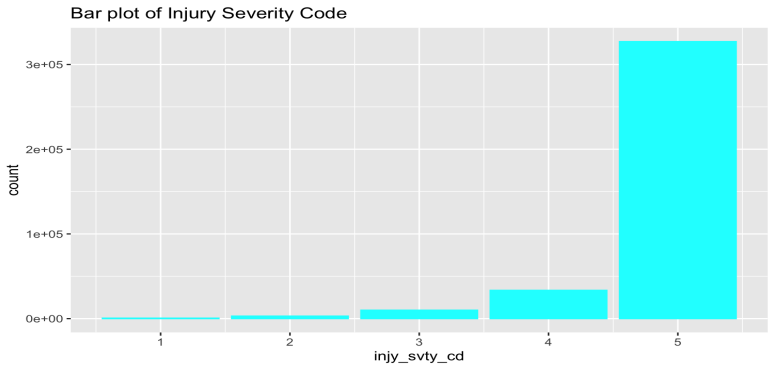
```
human_factor_crash <- subset(crash_dataset, select = c(injy_svty_cd, vehc_dfct_cd,
rstr_not_used_fail, hzrd_actn_cd, alch_susp_ind, drug_susp_ind,invl_prty_key
))
```

The *injy_svty_cd* variable is the target vector for the machine learning algorithm that will be used afterwards, and the rest of the variables are used in the feature matrix for their respective subsets.

Upon dividing the dataset into two subsets, data cleaning techniques are applied to remove duplicates and null values from each subset. The environmental factors subset consisted of **21,655** null and **4,147** duplicate values, whereas the human factors subset consisted of **31,125** null and **4,147** duplicate values. The dimensions for environmental factors subset and human factors subsets upon data cleaning are **374,099 x 6** and **371,684 x 7**, respectively.

VI. DATA EXPLORATION

This section will outline all the important features used in this project along with their effects on crash severity. However, it is first important to investigate the number of observations available for the different categories of injury severity:



1 | Fatal injury (K)
2 | Incapacitating injury (A)
3 | Non-incapacitating injury (B)
4 | Possible injury (C)
5 | No injury (O)
Null | Not Entered

Figure 2: Bar plot of injury severity categories

There is an imbalance in the number of observations in each category of injury severity. Therefore, it is clear from Figure 2 that a majority of freeway crashes resulted in no injuries.

ENVIRONMENTAL FACTORS DATASET

Weather Condition: Recall the code descriptions of the weather condition variable:

1 | Clear
2 | Cloudy
3 | Fog/Smoke
4 | Rain
5 | Snow/Blowing Snow

The graph on the left shows that there is an imbalance in the number of observations in each weather condition category with wthr_cd = 1 dominating the effect of weather condition on injury severity. However, the graph on the right provides an interesting insight. Fog/Smoke condition has fewer observations of incapacitating (2), non-incapacitating (3), and possible (4) injuries. This is unusual since foggy conditions are more likely to produce dangerous crashes. However, it can also be likely that fewer drivers were on the road due to the foggy conditions.

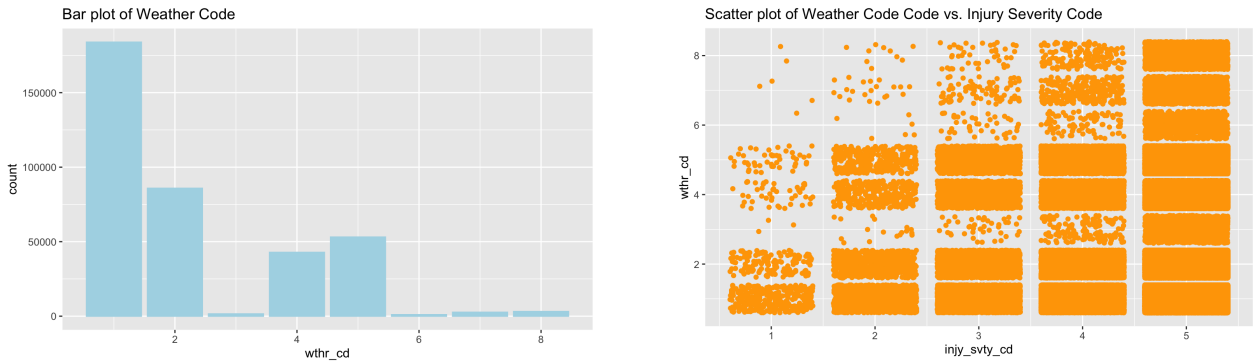


Figure 3: Effects of Weather Conditions on Injury Severity

Lighting Condition: Recall the code descriptions of the lighting condition variable:

1 | Daylight

2 | Dawn

3 | Dusk

4 | Dark-Lighted

5 | Dark-Unlighted

6 | Other Unknown

Null | Not Entered

The graph on the left shows that there is an imbalance in the number of observations in each lighting condition category with lit_cd = 1 dominating the effect of lighting condition on injury severity. However, the graph on the right provides an interesting insight. Dawn (2) and Dusk (3) conditions have fewer observations of incapacitating (2), non-incapacitating (3), and possible (4) injuries compared to other lighting conditions. This is unusual since dawn and dusk conditions are more likely to produce dangerous crashes. However, it can also be likely that fewer drivers were on the road on those times of day.

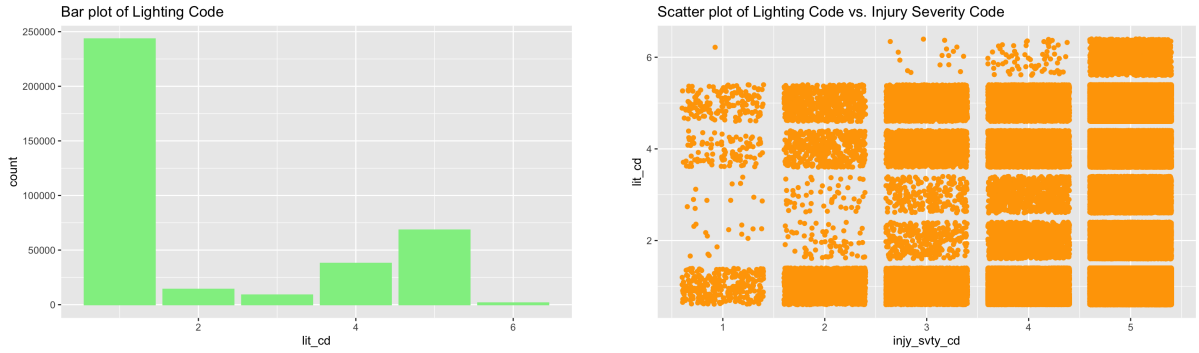


Figure 4: Effects of Lighting Conditions on Injury Severity

Road Condition: Recall the code descriptions of the road condition variable:

1 | Dry

2 | Wet

3 | Icy

4 | Snowy

5 | Muddy

6 | Slushy

7 | Debris

8 | Other/Unknown

Null | Not Entered

The graph on the left shows that there is an imbalance in the number of observations in each road condition category with rd_cond_cd = 1 dominating the effect of road condition on injury severity. However, the graph on the right provides an interesting insight. Muddy (5) and Debris (7) conditions have fewer observations of incapacitating (2), non-incapacitating (3), and possible (4) injuries compared to other road conditions. This is unusual since muddy and debris conditions are more likely to produce dangerous crashes. However, it can also be likely that fewer drivers go off-roading and venture out on muddy and debris-filled roads.

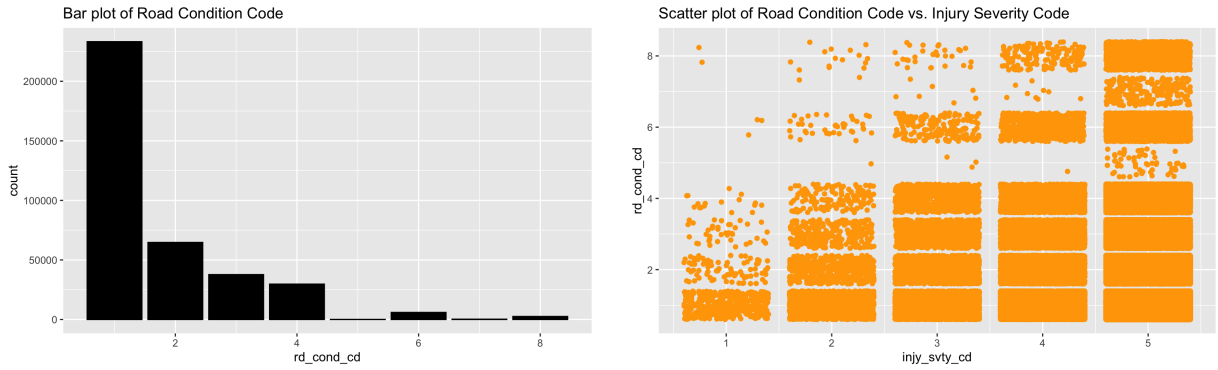


Figure 5: Effects of Road Conditions on Injury Severity

MDOT Region: Recall the code descriptions of the Mdot region variable:

0 | Statewide Multi-Region

1 | Superior

2 | North

3 | Grand

4 | Bay

5 | Southwest

6 | University

7 | Metro

The graph on the left shows that there is an imbalance in the number of observations in each MDOT Region category with mdot_regn_cd = 7 dominating the effect of MDOT Region on injury severity. The graph on the right provides a snippet of the effect of MDOT Region on injury severity. Fatal injury condition has the least number of data points but that is understandable since the count of fatal injury observations is the least compared to other degrees of severity (Figure 2). Otherwise, the distribution of Figure 6 aligns with the distribution in Figure 2.

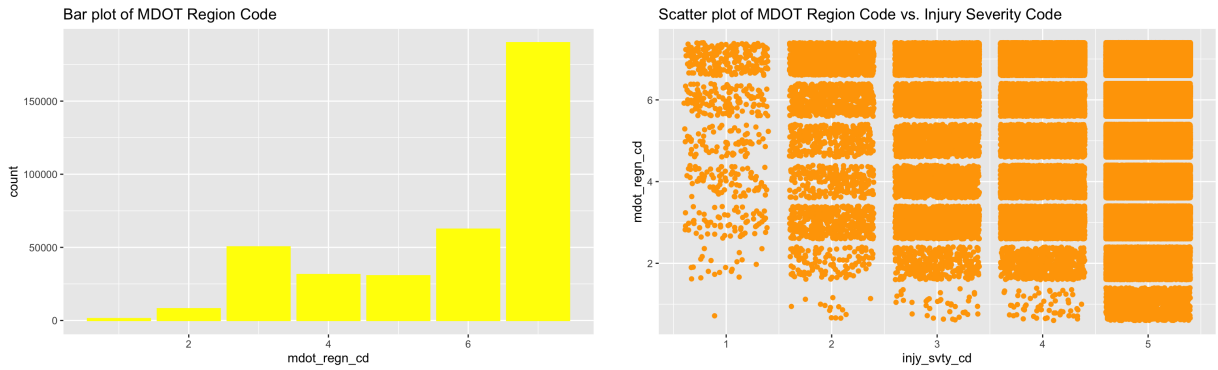


Figure 6: Effects of MDOT Region on Injury Severity

HUMAN FACTORS DATASET

Vehicle Defect Condition: Recall the code descriptions of the Vehicle Defect Code variable:

1 | Brakes

2 | Lights/reflectors

3 | Steering

4 | Tires/Wheels

5 | Windows

6 | Other

The graph on the left shows that there is an imbalance in the number of observations in each vehicle defect category with vehc_dfct_cd = 6 dominating the effect on injury severity. However, the graph on the right provides an interesting insight. Brakes (1), Steering (3) and Tires/Wheels (4) conditions have more observations of incapacitating (2), non-incapacitating (3), and possible (4) injuries compared to other vehicle defect conditions. This is understandable since these defects can often result in dangerous crashes.

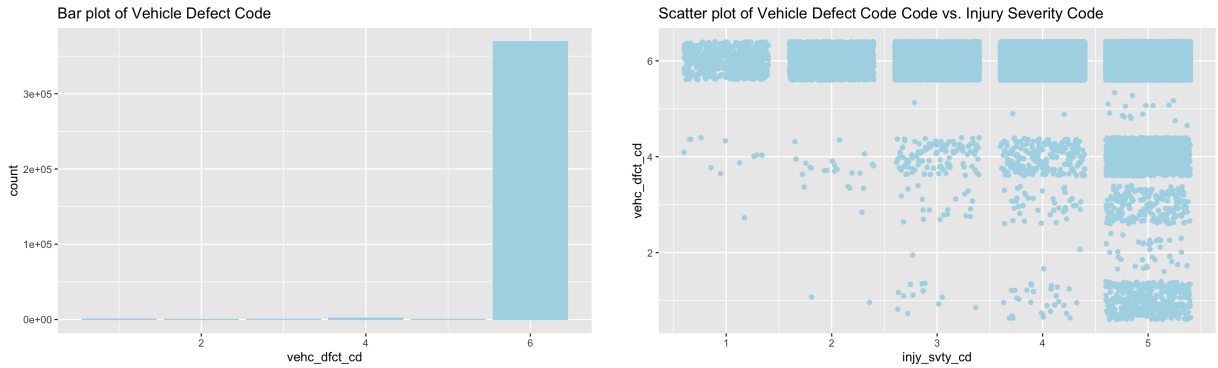


Figure 7: Effects of Vehicle Defect Condition on Injury Severity

Restraint Usage Condition: Recall the code descriptions of the restraint usage variable:

1 | Yes

0 | No

The graph on the left shows that there is an imbalance in the number of observations in each restraint usage category with rstr_not_used_fail = 0 dominating the effect on injury severity. However, the graph on the right provides an interesting insight. There exists more data points of fatal accidents in rstr_not_used_fail = 0, which is counter-intuitive since seatbelts are designed to prevent fatality in accidents. However, it would be safe to assume that a significant majority of the drivers on the road use seat belts out of sheer habit which would explain this phenomena.

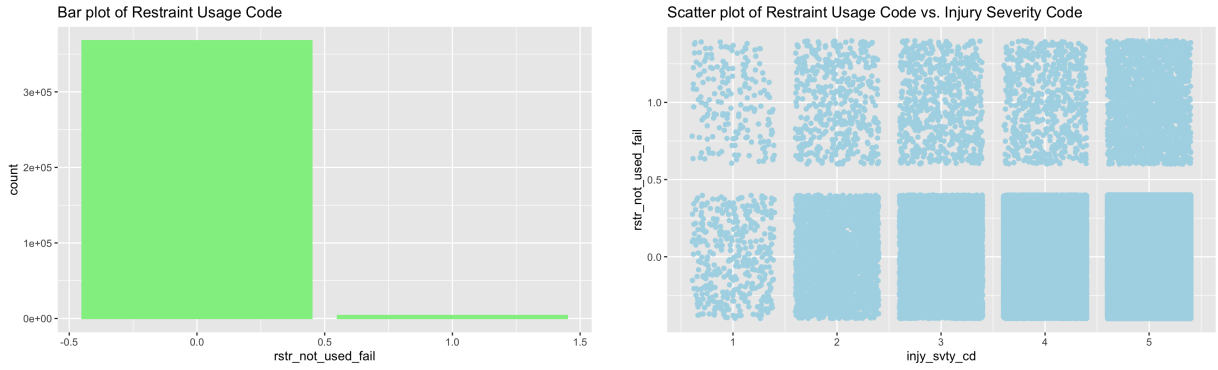


Figure 8: Effects of Restraint Usage Condition on Injury Severity

Hazardous Action Code: Recall the code descriptions of the hazardous action code variable:

0 | None

1 | Speed too fast

2 | Speed too slow

3 | Failed to yield

4 | Disregard traffic control

5 | Drove wrong way

6 | Drove left of center

7 | Improper passing

8 | Improper lane use

9 | Improper turn

10 | Improper/no signal

11 | Improper backing

12 | Unable to stop

13 | Other

14 | Unknown

15 | Reckless driving

16 | Careless/negligent

Null | Not Entered

The graph on the left shows that there is an imbalance in the number of observations in each hazardous action category with hzrd_actn_cd = 0 dominating the effect on injury severity. However, the graph on the right provides an interesting insight. There exists more data points of fatal accidents in 11 < hzrd_actn_cd <16 conditions, which leaves cases with mostly unknown/other hazardous actions that could be worth investigating.

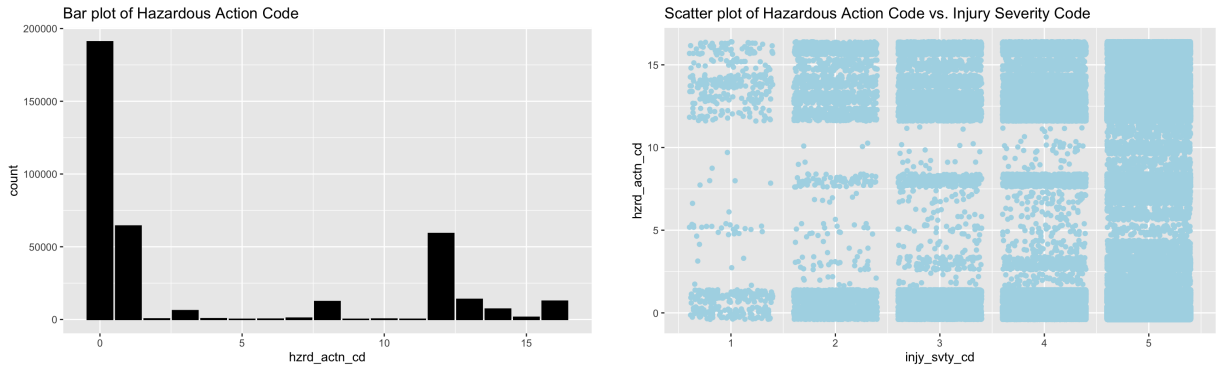


Figure 9: Effects of Hazardous Action on Injury Severity

Alcohol Usage: Recall the code descriptions of the alcohol usage variable:

0 | No

1 | Yes

Null | Not Entered

The graph on the left shows that there is an imbalance in the number of observations in each alcohol usage category with alch_susp_cd = 0 dominating the effect on injury severity. This makes sense as only a fraction of all crashes are due to alcohol consumption.

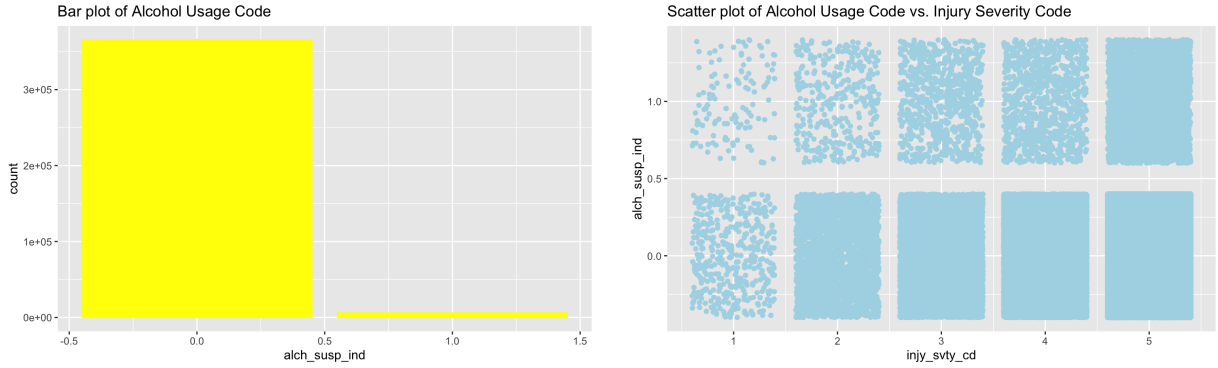


Figure 10: Effects of Alcohol Usage on Injury Severity

Drug Usage: Recall the code descriptions of the drug usage variable:

0 | No

1 | Yes

Null | Not Entered

The graph on the left shows that there is an imbalance in the number of observations in each drug usage category with drug_susp_ind = 0 dominating the effect on injury severity. This makes sense as only a fraction of all crashes are due to drug consumption.

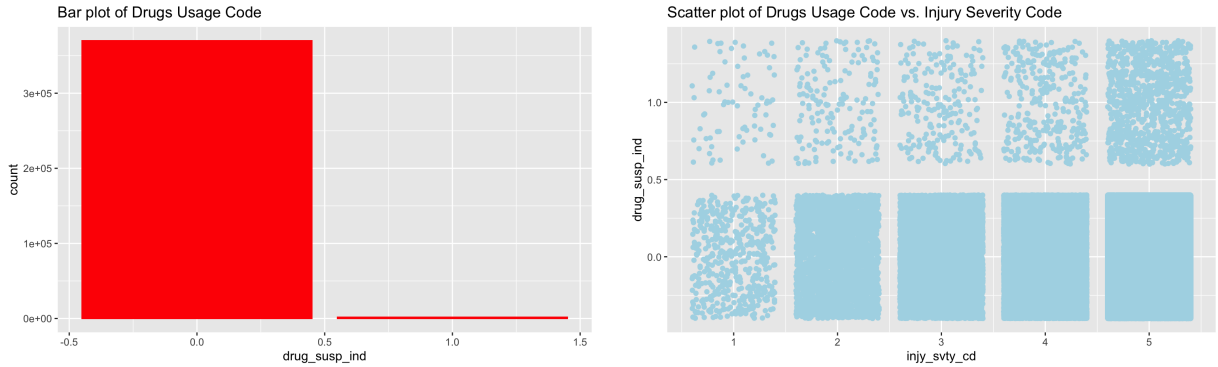


Figure 11: Effects of Drug Usage on Injury Severity

VII. TRAINING ML MODELS AND EVALUATION

It is now time to train the ML model of choice, which is the Multinomial Logistic Regression algorithm, to predict freeway crash injury severity. The crash dataset containing the

environmental and human factors, such as weather conditions, road type, and driver behavior, will be used for the ML model. Each dataset is split into train-test datasets using the 80-20 split ratio, respectively. Here is a glimpse of the code:

```
## Spliting Data for Training set and Testing set for Natural Factor Subset

```{r}
set.seed(123)
split <- sample.split(natural_factor_crash$injy_svty_cd, SplitRatio = 0.8)
nfcTrainData <- subset(natural_factor_crash, split == TRUE)
nfcTestData <- subset(natural_factor_crash, split == FALSE)
```

Spliting Data for Training set and Testing set for Human Factor Subset

```{r}
set.seed(123)
split <- sample.split(human_factor_crash$injy_svty_cd, SplitRatio = 0.8)
hfcTrainData <- subset(human_factor_crash, split == TRUE)
hfcTestData <- subset(human_factor_crash, split == FALSE)
```
```

Figure 12: Splitting datasets into train-test sets using 80-20 split ration

Once the splitting is completed, the train set is used to train the Multinomial Logistic Regression model to make injury severity predictions. Here is how the model was setup for learning:

```
## Model Training on Natural Factor Subset

```{r}
# Train a multi-class logistic regression model
logRegModel1 <- multinom(injy_svty_cd ~ ., data = nfcTrainData)

# Make predictions on the test data
predictions1 <- predict(logRegModel1, nfcTestData)

# Check the metrics of the model
cat("Model Accuracy for Human Factor Subset", mean(predictions1 == nfcTestData$injy_svty_cd), "\n")
```
```

```
# weights:  35 (24 variable)
initial  value 481670.968993
iter  10 value 195259.691310
iter  20 value 187470.499893
iter  30 value 177805.371975
iter  40 value 157134.563691
iter  50 value 153802.686075
iter  50 value 153802.686075
iter  60 value 143782.867490
iter  70 value 143483.359575
iter  70 value 143483.359575
iter  80 value 143075.877534
iter  90 value 142975.526679
iter 100 value 142828.019455
final  value 142828.019455
stopped after 100 iterations
Model Accuracy for Natural Factor Subset 0.874325
```

```
## Model Training on Human Factor Subset

```{r}
# Train a multi-class logistic regression model
logRegModel2 <- multinom(injy_svty_cd ~ ., data = hfcTrainData)

# Make predictions on the test data
predictions2 <- predict(logRegModel2, hfcTestData)

# Check the accuracy of the model
cat("Model Accuracy for Human Factor Subset", mean(predictions2 == hfcTestData$injy_svty_cd), "\n")
```
```

```
# weights:  40 (28 variable)
initial  value 478561.534946
iter  10 value 181692.448513
iter  20 value 157729.037185
iter  30 value 142862.211717
iter  40 value 140793.214581
iter  50 value 139896.700515
iter  60 value 139352.319374
iter  70 value 138937.125891
iter  80 value 138806.308586
iter  90 value 138727.782865
iter  90 value 138727.782865
iter 100 value 138692.286656
final  value 138692.286656
stopped after 100 iterations
Model Accuracy for Human Factor Subset 0.8744905
```

Figure 13: Training Multinomial Logistic Regression Model using the train dataset

The accuracy score is used as the evaluation metric to measure the performance of the model.

The model achieved an accuracy score of **87.43%** for the environmental factors subset whereas the accuracy score achieved for the human factors subset is **87.45%**. These results demonstrate the potential of using machine learning techniques to improve road safety by identifying hazardous conditions and informing the development of effective interventions and policies to prevent future accidents.

VIII. CONCLUSION

In this project, a machine learning model was developed using the Multinomial Logistic Regression algorithm to predict freeway crash injury severity based on various environmental and human factors. The model achieved an accuracy score of **87.43%** for the environmental factors subset whereas the accuracy score achieved for the human factors subset is **87.45%**; therefore, demonstrating its potential to accurately predict the severity of freeway crashes and provide valuable insights to improve post-crash care for victims of road crashes. The results indicate that machine learning techniques can be used to analyze and interpret crash datasets effectively, providing insights to improve road safety and inform the development of more effective interventions and policies to prevent future accidents. By predicting injury severity, the model can assist first responders and medical personnel in prioritizing the most seriously injured victims and providing them with the necessary medical care. In conclusion, this project highlights the potential of machine learning techniques in predicting freeway crash injury severity and improving road safety. Future work could involve exploring other machine learning algorithms and ensembling methods to further improve the accuracy and robustness of the model. Overall, this project has shown that machine learning techniques can play a vital role in enhancing road safety and saving lives.

REFERENCES

[1] Abdulhafedh, A. (2017) Road Traffic Crash Data: An Overview on Sources, Problems, and Collection Methods. *Journal of Transportation Technologies*, **7**, 206-219. doi: 10.4236/jtts.2017.72015.

[2] Wang, J., Song, H., Fu, T., Behan, M., Jie, L., He, Y., & Shangguan, Q. (2022). Crash prediction for freeway work zones in real time: A comparison between convolutional neural

network and binary logistic regression model. *International Journal of Transportation Science and Technology*, *11*(3), 484–495. https://doi.org/10.1016/j.ijtst.2021.06.002