Analysis of Injury Severity in Traffic Collision based on Surrounding Factors

Rushi Raval[1]

I. Introduction

In recent years, open datasets have become vital for researchers. One dataset that intrigued me is the Traffic Collision dataset on Ottawa's website. It offers detailed information on traffic accidents, including factors like location and environmental conditions. The main aim is to analyze injury severity, using "Max Injury" to classify injuries from minor to fatal. The focus is on understanding how factors such as location type, road conditions, and environmental factors affect injury severity.

Data visualization and statistical modeling were used to investigate traffic collisions.

To assess the influence of environmental conditions on accident rates, a pie chart displayed the distribution of accidents across various environments. Additionally, a bar chart depicted the temporal distribution of accidents throughout the day.

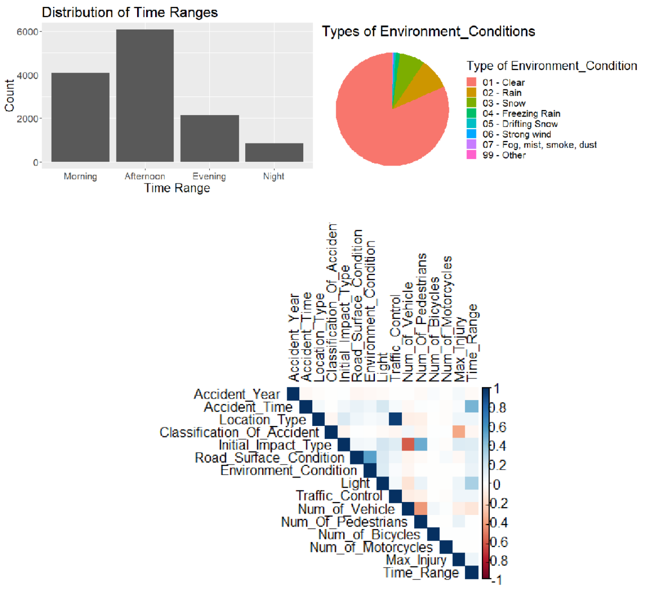

Fig. 1. Plots for different features

Through analysis by correlation plots, potential interactions were uncovered among various variables influencing accident dynamics. For instance, the correlation plot revealed associations between the type of initial impact and factors such as environmental conditions, road surface conditions, and lighting. This suggests that certain types of accidents may be more prevalent under specific environmental and lighting conditions, potentially influencing accident severity. Moreover, the correlation between the time of day and lighting conditions highlighted how variations in lighting throughout the day may interact with traffic patterns to affect accident outcomes. These insights underscore the importance of considering the interplay between different factors in understanding accident occurrence and severity.

Several factors could indicate the presence of moderating effects within the context of accidents. Firstly, lighting conditions may moderate the impact of driver visibility on accident severity, with reduced visibility during nighttime hours potentially amplifying the consequences of driver errors. Adverse weather conditions like rain or snow worsens the effects of speeding or reckless driving, increasing the likelihood of accidents. Moreover, the time of day may moderate the relationship between traffic volume and accident risk, with peak traffic hours.

II. Methods

A. Linear Regression : for feature selection

Linear regression was utilized initially to understand the relationship between various factors and the severity of injuries in traffic collisions. To identify the most important predictors, a stepwise regression method was applied. This process aided in identifying the key features that significantly impact injury severity.

Through the stepwise regression method, the predictors were narrowed down to a final set that demonstrated significant associations with injury severity. The selected features include Traffic Control, Light, Road Surface Condition, Initial Impact Type, and Classification of Accident.

The formula for linear regression using the selected features as predictors is given by:

$$\begin{aligned}\text{Max Injury}_i = {} & \beta_0 + \beta_1 \times \text{Traffic Control}_i \\ & + \beta_2 \times \text{Light}_i + \beta_3 \times \text{Road Surface}_i \\ & + \beta_4 \times \text{Initial Impact Type}_i \\ & + \beta_5 \times \text{Classification Accident}_i\end{aligned}$$

In this formula:

- Max Injury_i represents the severity of injury for observation i.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are the coefficients corresponding to the selected features: Traffic Control, Light, Road Surface Condition, Initial Impact Type, and Classification of Accident, respectively.

These features will serve as predictors in subsequent modeling endeavors, where the outcome variable will be Max

Injury. By concentrating on these key features, the objective is to formulate a model that precisely predicts the severity of injuries in traffic collisions.

B. Multiple Linear Regression : comparing different models

Following the selection of predictor variables, multiple linear regression models were constructed to assess their collective influence on injury severity in traffic collisions. Various model configurations were explored by training four distinct models with different combinations of parameters.

In each model iteration, parameters such as feature selection, regularization techniques, and model complexity were systematically adjusted to discern the optimal configuration for accurately predicting injury severity.

The coefficient for Time_Range suggests that accidents occurring later in the day tend to result in more severe injuries, with an estimate of 0.021 (t value = 4.278, $p < 0.001$). Similarly, the coefficient of 0.036 for Initial_Impact_Type highlights the substantial impact of collision type on injury severity (t value = 14.074, $p < 0.001$). Conversely, the negative coefficient estimate of -0.026 for Road_Surface_Condition indicates that accidents on roads with better surface conditions tend to result in less severe injuries (t value = -6.313, $p < 0.001$).

Upon evaluating the performance of each model, variations in their predictive capabilities were observed. Particularly, the fourth model showed a substantial improvement in predictive accuracy, achieving an R-squared value of 18%. This notable enhancement in model performance signifies a significant increase from the modest 1% R-squared value observed in the initial model iteration.

C. Ordinal Logistic Regression : bayesian approach

After exploring multiple linear regression models, a transition was made to a generalized linear regression approach, specifically employing ordinal logistic regression with Hamiltonian Monte Carlo (HMC) simulation using the brms package. This method was chosen to accommodate the ordinal nature of the outcome variable "Max Injury," which categorizes injuries into distinct levels ranging from minor to fatal.

In this framework, "Max Injury" was treated as an ordered categorical variable, representing different levels of injury severity (i.e., minor, minimal, major, and fatal).

Variable	Estimate	Est. Error	Lower CI	Upper CI
Time_Range	0.08	0.02	0.05	0.11
Initial_Impact_Type	0.13	0.01	0.11	0.15
Road_Condition	-0.10	0.02	-0.13	-0.06
Light	0.07	0.01	0.04	0.09
Traffic_Control	0.02	0.01	0.01	0.03
Env_Condition	0.02	0.03	-0.04	0.07

TABLE I

REGRESSION COEFFICIENTS

The coefficient for Time_Range (0.08) suggests that for each unit increase in the time range, there is an estimated average increase of 0.08 units in the response variable, with a standard error of 0.02 and a 95% confidence interval ranging from 0.05 to 0.11. Similarly, the coefficient for Initial_Impact_Type (0.13) indicates that incidents with a specific initial impact type are associated with an average increase of 0.13 units in the response variable compared to incidents with a different initial impact type. Conversely, the coefficient for Road_Surface_Condition (-0.10) implies that a one-unit increase in road surface condition is associated with an average decrease of 0.10 units in the response variable, with a 95% confidence interval from -0.13 to -0.06. Moreover, the coefficient for Light (0.07) suggests that each unit increase in light level corresponds to an average increase of 0.07 units in the response variable, and the coefficient for Traffic_Control (0.02) indicates that incidents under certain traffic control conditions are associated with an average increase of 0.02 units in the response variable compared to incidents under different traffic control conditions. Lastly, the coefficient for Environment_Condition (0.02) suggests that a one-unit increase in environmental condition is associated with an average increase of 0.02 units in the response variable, with a 95% confidence interval ranging from -0.04 to 0.07. All Rhat values being 1.00 indicate good convergence of the Bayesian analysis.

The regression analysis unveiled significant coefficients indicating the relationships between independent variables and the severity of maximum injury (Max_Injury). Time_Range showed that accidents occurring later in the day tend to result in more severe injuries, while Initial_Impact_Type highlighted the substantial impact of collision type on injury severity. Conversely, Road_Surface_Condition suggested that accidents on roads with better surface conditions tend to result in less severe injuries, and Light indicated that accidents in darker conditions are correlated with more severe injuries. Additionally, Traffic_Control and Environment_Condition demonstrated minor impacts, suggesting that areas with more controls or adverse environmental conditions may experience slightly more severe injuries. These findings offer insights into how various environmental factors and collision characteristics contribute to injury severity in accidents.

By leveraging brms and HMC simulation, the ordinal logistic regression model was refined, enhancing its predictive accuracy and robustness.

D. Decision Tree

Following Ordinal Logistic Regression, a transition was made to utilizing the decision tree method to predict Max Injury based on selected predictors. This approach aimed to leverage the algorithm's capacity to capture non-linear relationships and interactions among variables, particularly beneficial for tasks like predicting injury severity in traffic collisions.

To assess the model's performance, the dataset was partitioned into training and test sets, with 70% and 30% of the data, respectively. This partitioning ensured robust training while providing an independent dataset for assessing predictive accuracy. Training the decision tree model on the training set and subsequently evaluating its performance on the test set allowed for gauging its generalization capability.

The combination of the decision tree method and the train-test split validation facilitated the development of a robust predictive model for injury severity in traffic collisions. Moreover, this achieved a 61% accuracy rate when evaluated on the independent test set, highlighting its efficacy in accurately predicting injury severity in traffic collisions.

III. Results

A. Confounders, Mediators, Moderators, and Colliders

A confounder is a variable that affects both the predictor and outcome variables, potentially influencing their relationship. To identify confounders, the following method was employed:

1) Compute the correlation matrix to assess relationships between all variables.
2) Correlation coefficients between each predictor and the outcome variable (Max Injury) were calculated.
3) Variables showing high correlation coefficients with the outcome variable were flagged as potential confounders.

In the analysis, it was found that Initial_Impact_Type serves as a confounder. This variable exhibited a notable correlation with both the predictor variables and the outcome variable (Max Injury), indicating its potential to influence the relationship between predictors and injury severity in traffic collisions.

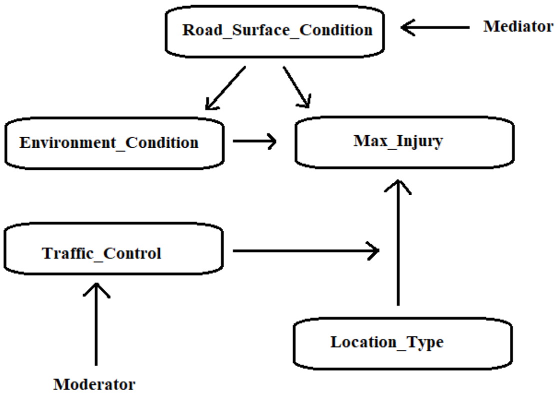

Fig. 2. Confounders, Mediators and Moderators

A mediator is a variable that explains the relationship between an independent variable (predictor) and a dependent variable (outcome) by intervening in the causal pathway between them.

The mediation library was employed, and the mediator function was utilized to identify potential mediators within the model. Through this analysis, Road_Surface_Condition emerged as a mediator, showing a significant p-value of 0.01. This finding suggests that Road_Surface_Condition may play a mediating role in the relationship between the predictor variables and the outcome variable, Max Injury.

By identifying mediators, this analysis provides insights into the underlying mechanisms through which certain variables influence injury severity in traffic collisions. This knowledge can inform targeted interventions aimed at mitigating the impact of mediators and reducing the severity of injuries in road accidents.

A moderator is a variable that influences the strength or direction of the relationship between a predictor variable and an outcome variable. Unlike confounders, which affect both the predictor and outcome variables, moderators specifically impact the interaction between the predictor and outcome.

In addition, the method to identify potential moderators is through visual inspection of interaction plots. Interaction plots display the relationship between the predictor variables and the outcome variable at different levels of the potential moderator. In this analysis, it was discovered that Traffic_Control serves as a moderator between Location Type and Max Injury.

IV. Discussion

A. Ordinal Logistic Model Analysis

The study entailed the development and evaluation of an ordinal logistic regression model to ascertain the severity of injuries based on a predetermined set of predictor variables. Our analysis revealed that the model achieved an overall accuracy rate of 57% in predicting injury severity levels. This performance metric denotes the model's proficiency in effectively stratifying the severity of injuries using the provided predictor variables.

In summary, the ordinal logistic regression analysis unveiled significant associations between predictor variables and higher categories of injury in accidents. Nighttime accidents demonstrated a substantial increase in injury severity, emphasizing the heightened risk during nighttime driving. Certain initial impacts and poor road conditions were also linked to greater injury severity. While traffic control showed a slight increase in severity, factors like lighting and environmental conditions had smaller effects.

References

[1] Harrell, F.E. (2001). Ordinal Logistic Regression. In: Regression Modeling Strategies. Springer Series in Statistics. Springer, New York, NY.

[2] Betancourt, Michael. (2017). A Conceptual Introduction to Hamiltonian Monte Carlo.

[3] Pearl, J. (2009) Causal inference in statistics: An overview. Statistics surveys 3.