Classify and Predict Injury Severity Using Crash Reporting Sampling System (CRSS) Data of the Most Recent Years

Meghanath Payasam [1] Golnoosh Garakani [1] Ayegba David Haruna [1]

[1]University of Michigan, Dearborn

IMSE 586

Background and Research Questions

- Road traffic accidents are a widespread problem, resulting in many injuries and deaths worldwide.
- In the US, in 2022, car crashes resulted in 42,795 deaths, incurring costs of over $300 billion.

To conduct a crash data analysis and identify the trend of leading factors the following research questions are identified:

- Which factors contributed to severe injury or fatality in recent years?
- Which factors contributed to minor or no injury in recent years?
- How changes in leading factors influence the injury severity over the years?
- Which modeling scheme can more accurately represent the statistical property of the data and precisely predict the injury severity?

Methodology

To answer the research questions, the following steps are considered:

- Employed recent accident data from the CRSS from 2019-2021.
- Data set creation and data engineering to prepare the data set for modeling.
- Categorized the severity of accidents into two groups of No injury/Minor injury as label 0 and Major injury/Fatal as label 1.
- Considered various factors including the year, geographic region, month, weekday, speed, alcohol involvement, weather conditions, age group, use of restraints, and driver distraction.
- Feature correlation analysis to gain insights into the factors and how they relate to each other and the severity.
- Modeling using binary classification methods of Logistic Regression and Random Forest, model analysis, and visualization of the data.

Data Engineering

Frequency of each **injury severity Class** in different years and months:

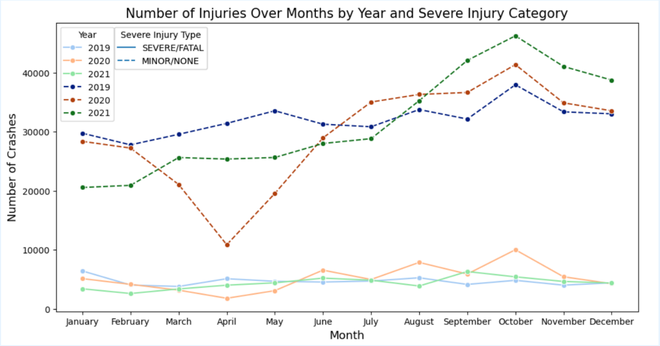


Figure 1. No. of crash injuries per month over year by injury severity

Major findings toward modeling

- Data is imbalanced:
  - Only 13.3% of the data represent fatal/serious (label 1).
  - hypothetical model classifying all cases as Minor/no injury could achieve 86.7% accuracy.
  - Techniques for dealing with imbalanced data such as sampling are required.
  - Upsampling with SMOTE resulted in computational challenges.
  - Downsampling was identified as the remedy for imbalanced data.
- *April 2020* recorded the lowest number of crashes in each category.
- *October* experienced the highest frequency of crashes.
  - October 2019 had the most major/fatal injuries.
  - October 2021 had the most minor/no injuries.

Factor Distribution-Alcohol

Alcohol impaired crashes by age group:

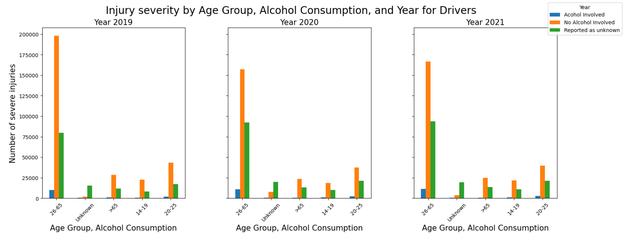


Figure 2. Variable distribution

Alcohol impaired crashes by region:

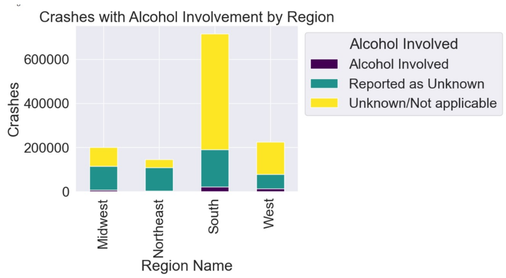


Figure 3. Variable distribution

Correlation Analysis

Correlation analysis to decipher the relationships between independent variables:

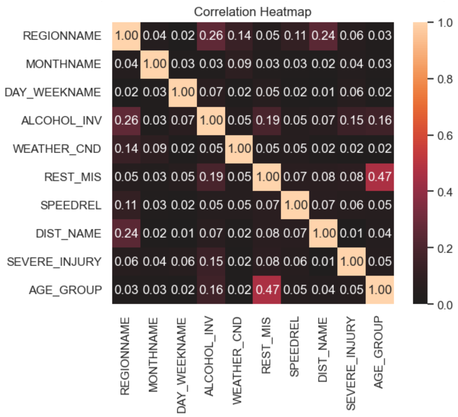
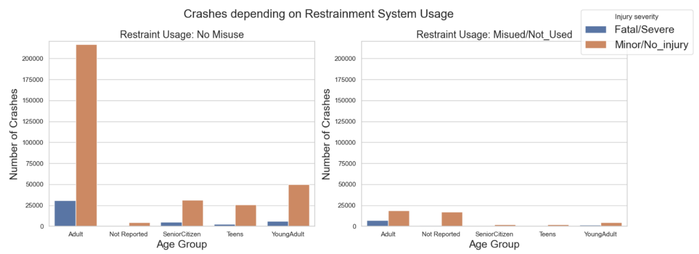


Figure 4. Correlation Matrix of Variables

- restraint showed the strongest correlation with the age group.
  - adults and young adults tend to have a higher incidence of restraint misuse or non-use compared to instances where restraints were not misused.
  - More crashes reported when restraint was used for senior citizens.



Modeling

Modeling schemes

1)Logistic Regression (LR) 2)Random Forest (RF)

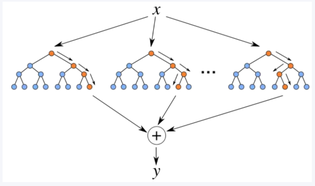
LR models the relationship between predictors and a categorical response variable as follows:

$$Logit(P) = \beta_0 + \beta \cdot X + \epsilon \quad (1)$$

Where $\{x_1, x_2, \cdots, x_n\}$ represent the independent variables used in this modeling. Hence we can derive:

$$P(Y = 1) = \frac{1}{\mathbf{e}^{-(\beta_0 + \beta X)}} \quad (2)$$

RF uses labeled data to learn and classify unlabeled data, with architecture like:



Modeling Challenge

Downsampling can enhance the precision and recall of the model, however:

- Downsampling can lead to a loss of information.
  - Reducing the size of the majority class affects the overall accuracy of the model.
- Comparison of sampling rate vs accuracy is critical to decide the best model.

Results and Conclusions

Trade-off between downsampling and evaluation metrics of modeling:

Modeling	Injury Severity Class	Precision	Recall	F-1 score
LR-100% sampling-rate	0	**0.59**	**0.71**	**0.64**
	1	**0.64**	**0.52**	**0.57**
LR-75% sampling-rate	0	0.63	0.86	0.73
	1	0.64	0.34	0.45
LR-50% sampling-rate	0	0.70	0.94	0.80
	1	0.63	0.21	0.31
LR-25% sampling-rate	0	0.81	0.99	0.89
	1	0.56	0.06	0.11
RF-100% sampling-rate	0	0.79	0.78	0.78
	1	0.78	0.79	0.78
RF-75% sampling-rate	0	**0.81**	**0.84**	**0.82**
	1	**0.77**	**0.73**	**0.75**
RF-50% sampling-rate	0	0.84	0.90	0.87
	1	0.76	0.65	0.70
RF-25% sampling-rate	0	0.89	0.96	0.92
	1	0.76	0.51	0.61

Table 1. Modeling performance in presence of different sampling rates

Key Findings

- RF exceeds the modeling performance of LR.
  - RF with 75% sampling-rate resulted in the best performance (with an accuracy of 0.79).
  - LR with 100% sampling-rate performs the best among LR cases (with an accuracy of 0.61).
- This analysis showed that the South region experienced the highest number of alcohol-involved crashes.
  - West region reported the second highest.
- Distraction by Outside/Others is the main category of known distraction contributing to both minor and major/fatal crashes.
  - This category caused the highest fatalities in 2020.
- Speeding Contributed to the increasing trend in crashes over the recent years.