Background and Research Questions

ffic accidents are a widespread problem, resulting in many injuries and worldwide.

S, in 2022, car crashes resulted in 42,795 deaths, incurring costs of 00 billion.

t a crash data analysis and identify the trend of leading factors the fol- earch questions are identified:

actors contributed to severe injury or fatality in recent years?
actors contributed to minor or no injury in recent years?
anges in leading factors influence the injury severity over the years?
nodeling scheme can more accurately represent the statistical property ata and precisely predict the injury severity?

Methodology

the research questions, the following steps are considered:

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

Data Engineering

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

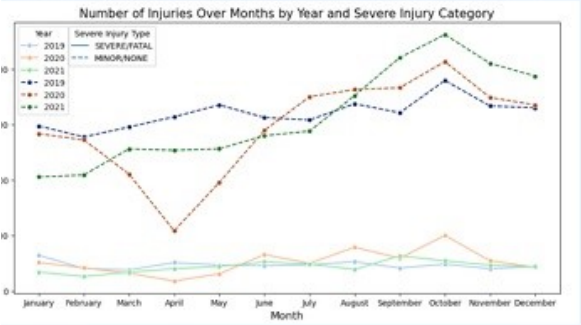


Figure 1. No. of crash injuries per month over year by injury severity

dings toward modeling

imbalanced:
13.3% of the data represent fatal/serious (label 1).
hetical model classifying all cases as Minor/no injury could achieve 86.7% accuracy.
iques for dealing with imbalanced data such as sampling are required.
mpling with SMOTE resulted in computational challenges.
sampling was identified as the remedy for imbalanced data.
020 recorded the lowest number of crashes in each category.
r experienced the highest frequency of crashes.
er 2019 had the most major/fatal injuries.
er 2021 had the most minor/no injuries.

Factor Distribution-Alcohol

Alcohol impaired crashes by age group:

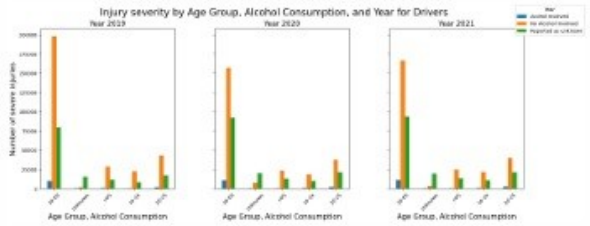


Figure 2. Variable distribution

Alcohol impaired crashes by region:

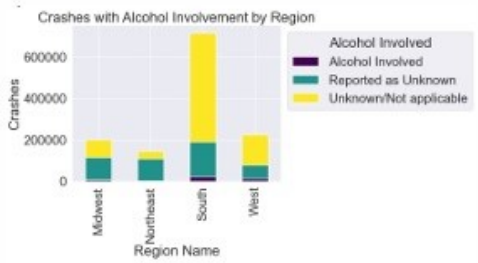


Figure 3. Variable distribution

Correlation Analysis

Correlation analysis to decipher the relationships between independent variables:

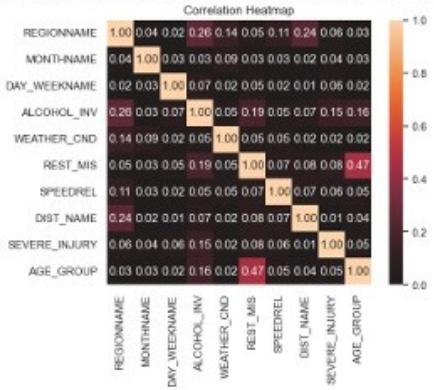
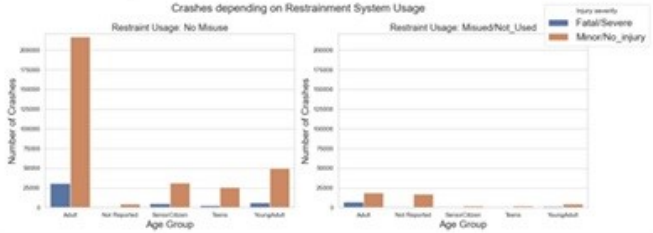


Figure 4. Correlation Matrix of Variables

- 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)

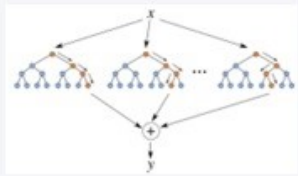
LR models the relationship between predictors and a categorical respo able as follows:

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

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

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

RF uses labeled data to learn and classify unlabeled data, with architec



Modeling Challenge

Downsampling can enhance the precision and recall of the model, hov

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

Results and Conclusions

Trade-off between downsampling and evaluation metrics of modeling:

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

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
  - LR with 100% sampling-rate performs the best among LR cases (with an accura
- This analysis showed that the South region experienced the highest alcohol-involved crashes.
  - West region reported the second highest.
- Distraction by Outside/Others is the main category of known distrac 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 rec