Predicting Crash Severity of Chicago Streets

Hawraa Salami

SPRINGBOARD Capstone Project 1

2018 Chicago Crash Facts

Total Crashes	**98,859**
Total Injuries	**24,400**
Total Incapacitating Injuries	**2,609**
Total Deaths	**132**

- What conditions lead to severe crashes?

- Data-driven models aim to understand the severity of crashes.

How can the models help?

City of Chicago

Develop better traffic control policies

Car Manufacturer

Incorporate more safety features

Insurance Companies

Perform better risk assessment

Chicago Crashes Dataset

- The data is available from Chicago online portal*.
- It contains information of Chicago crashes from 2015 to present.

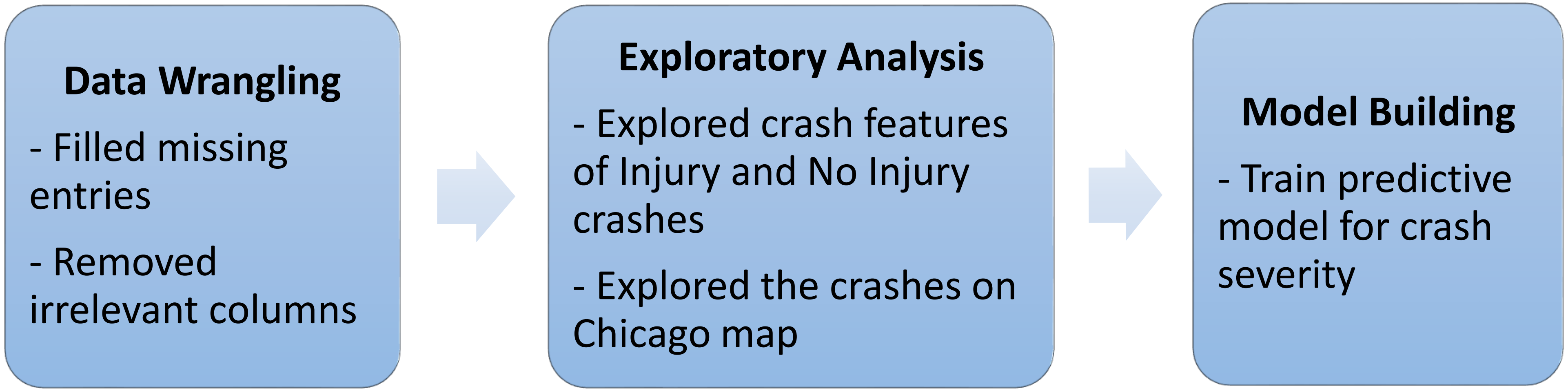
Crash Location and Time	External Conditions	Crash Cause and Description
Crash hour, day, month	Weather & Lighting Conditions	Primary Cause (driving behavior)
Crash address	Road alignment, type and surface	Type of Collision
	Speed Limit and Control Device	Type of crash: Injury or No injury

* https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

Data Analysis Steps

- **Goal:** Build a model that predicts the type of crash (Injury or No Injury)

- **Steps:**



Data Wrangling

- Filled missing entries

- Removed irrelevant columns

Exploratory Analysis

- Explored crash features of Injury and No Injury crashes

- Explored the crashes on Chicago map

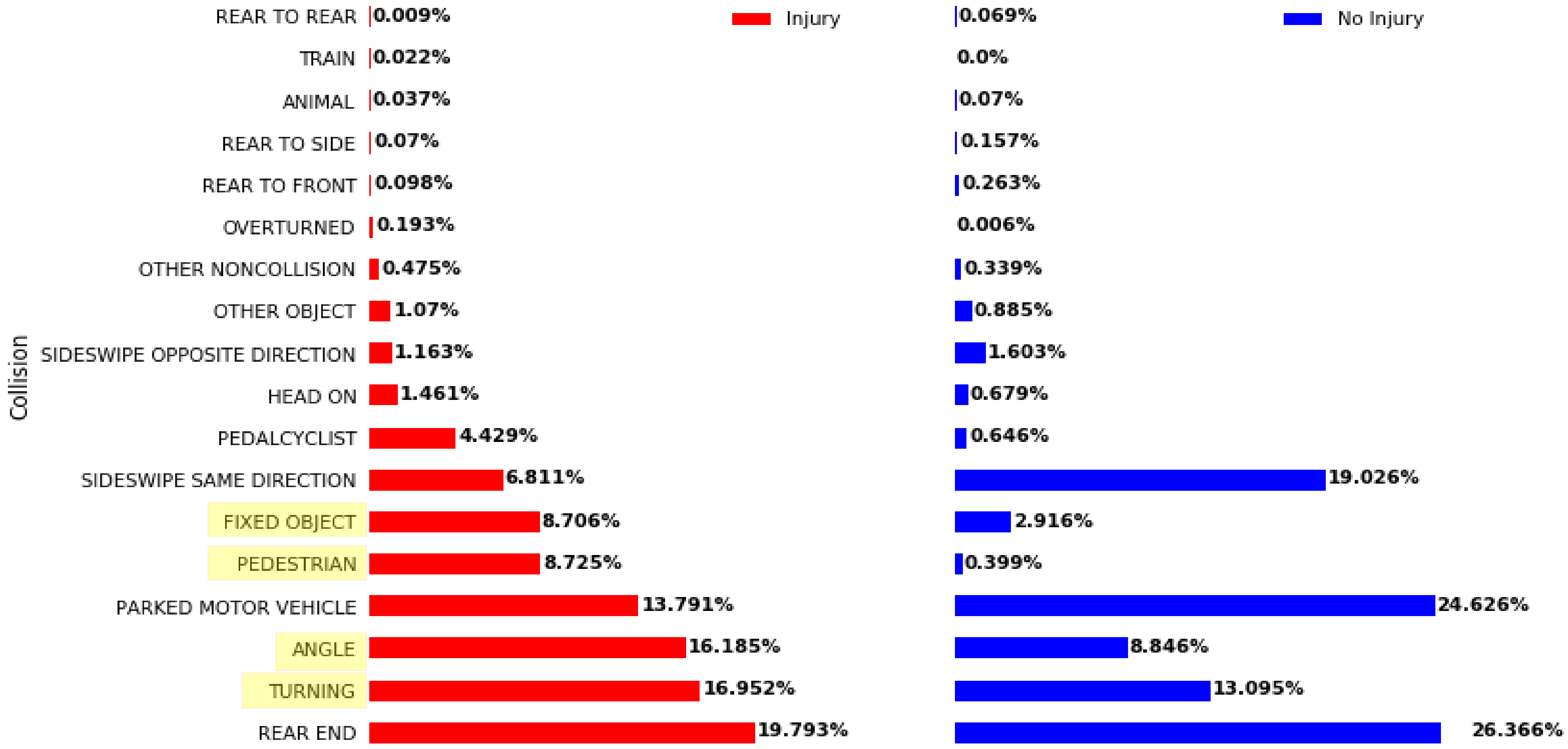
Model Building

- Train predictive model for crash severity

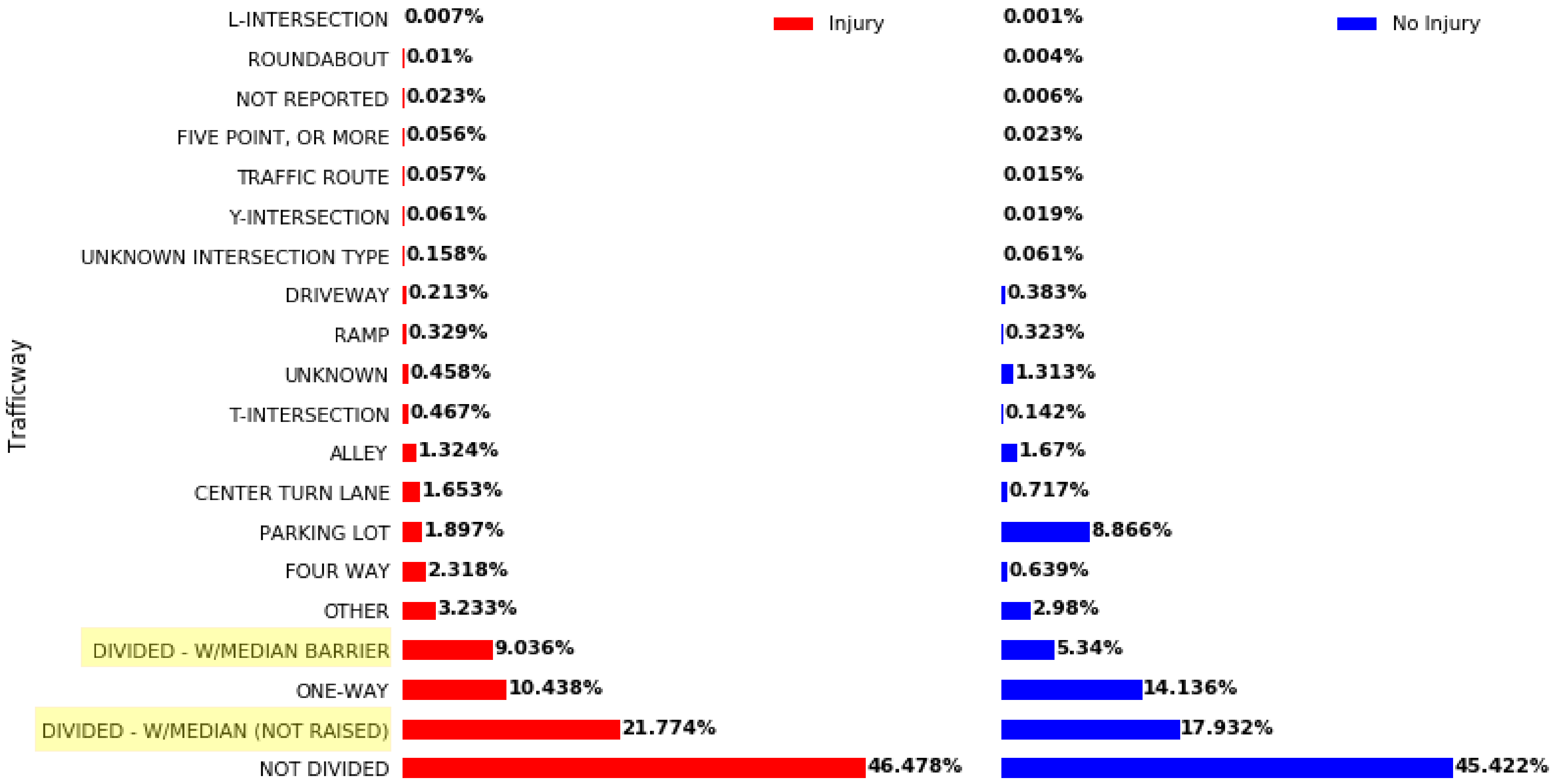
Exploratory Analysis

Data Visualization & Statistical Analysis

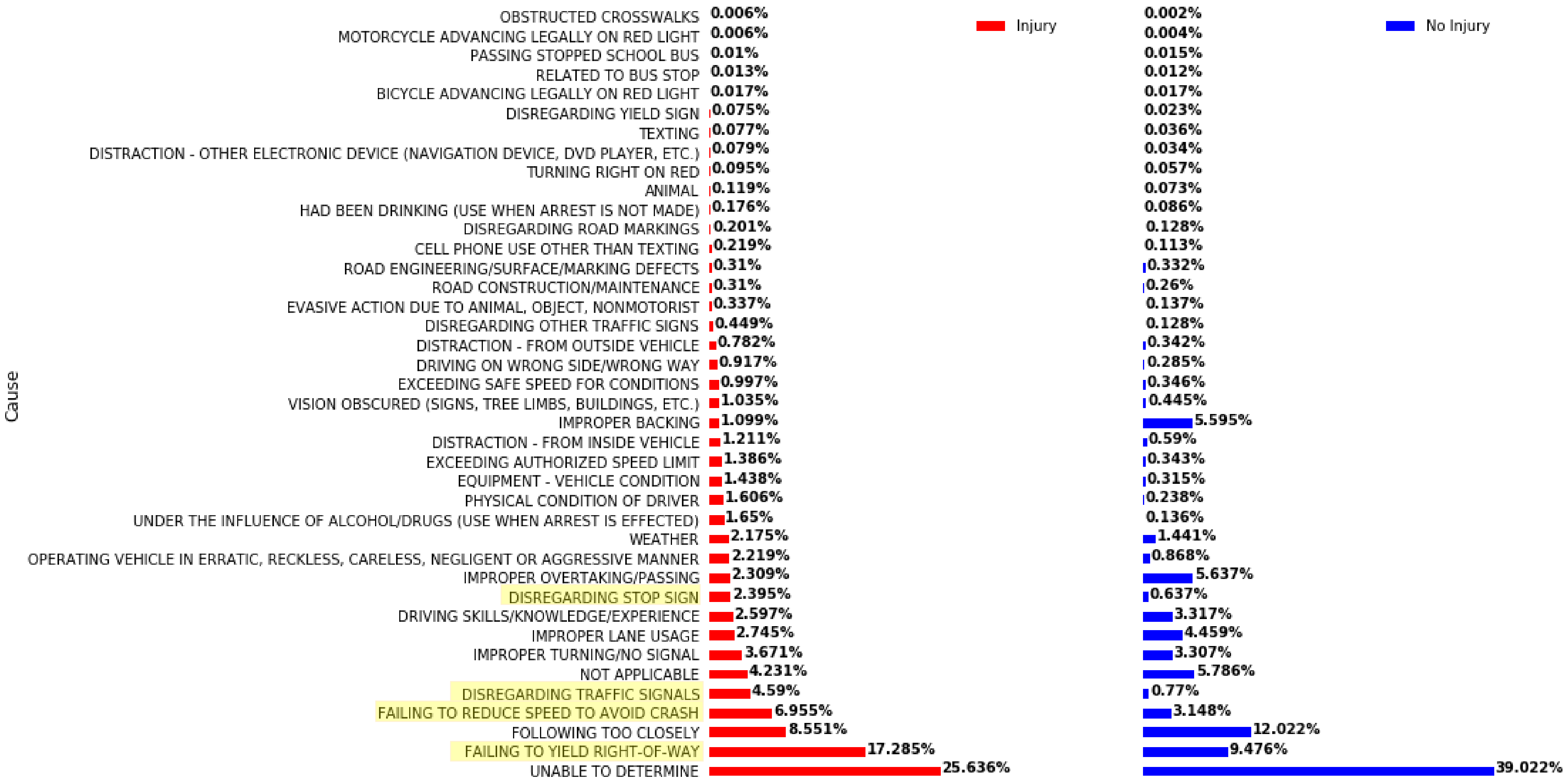
Collision Distribution for Each Crash Type



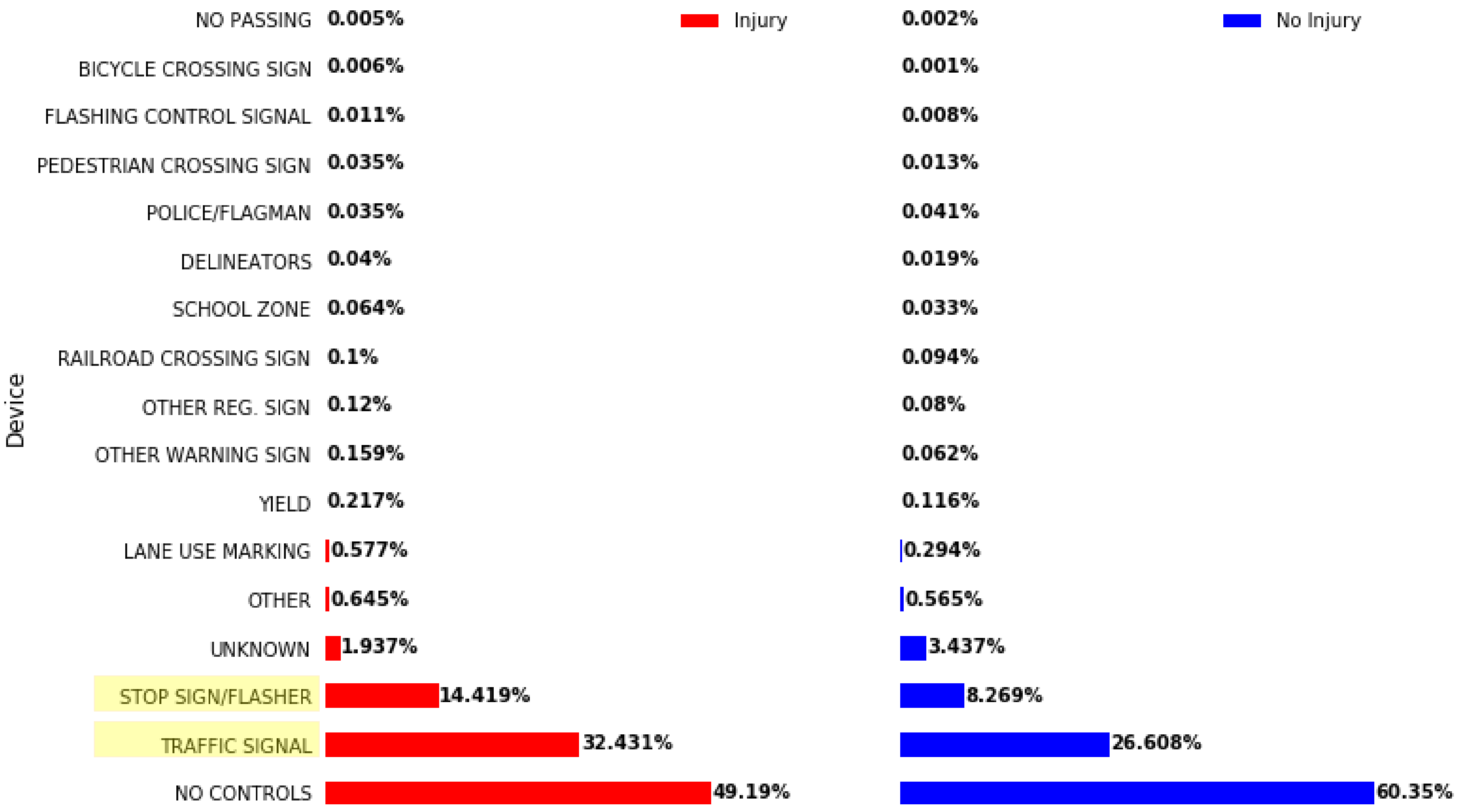
Trafficway Distribution for Each Crash Type



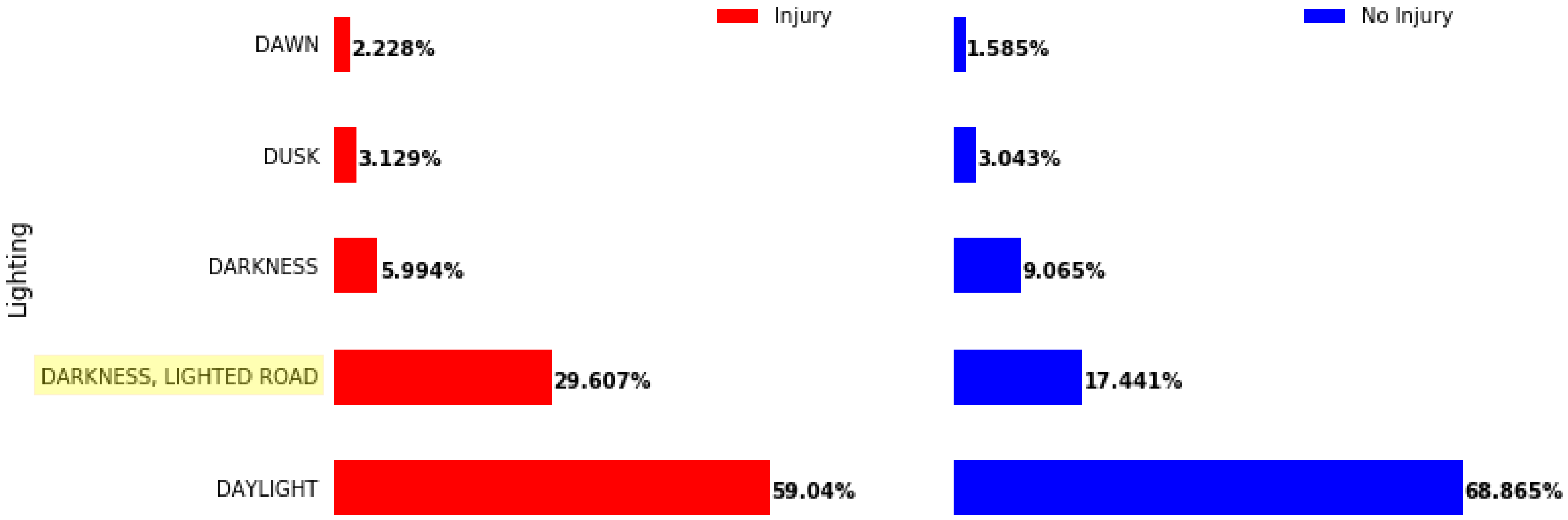
Driving Behavior for Each Crash Type



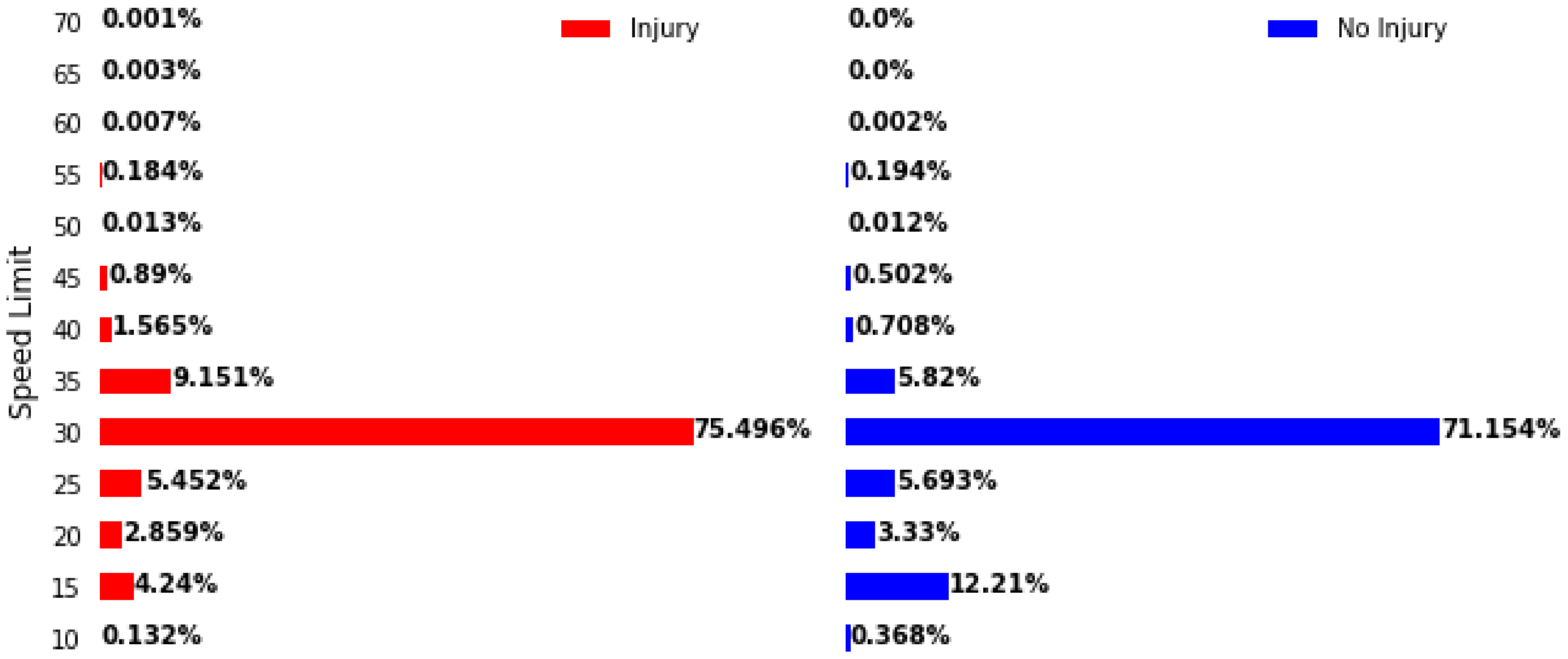
Control Device for Each Crash Type



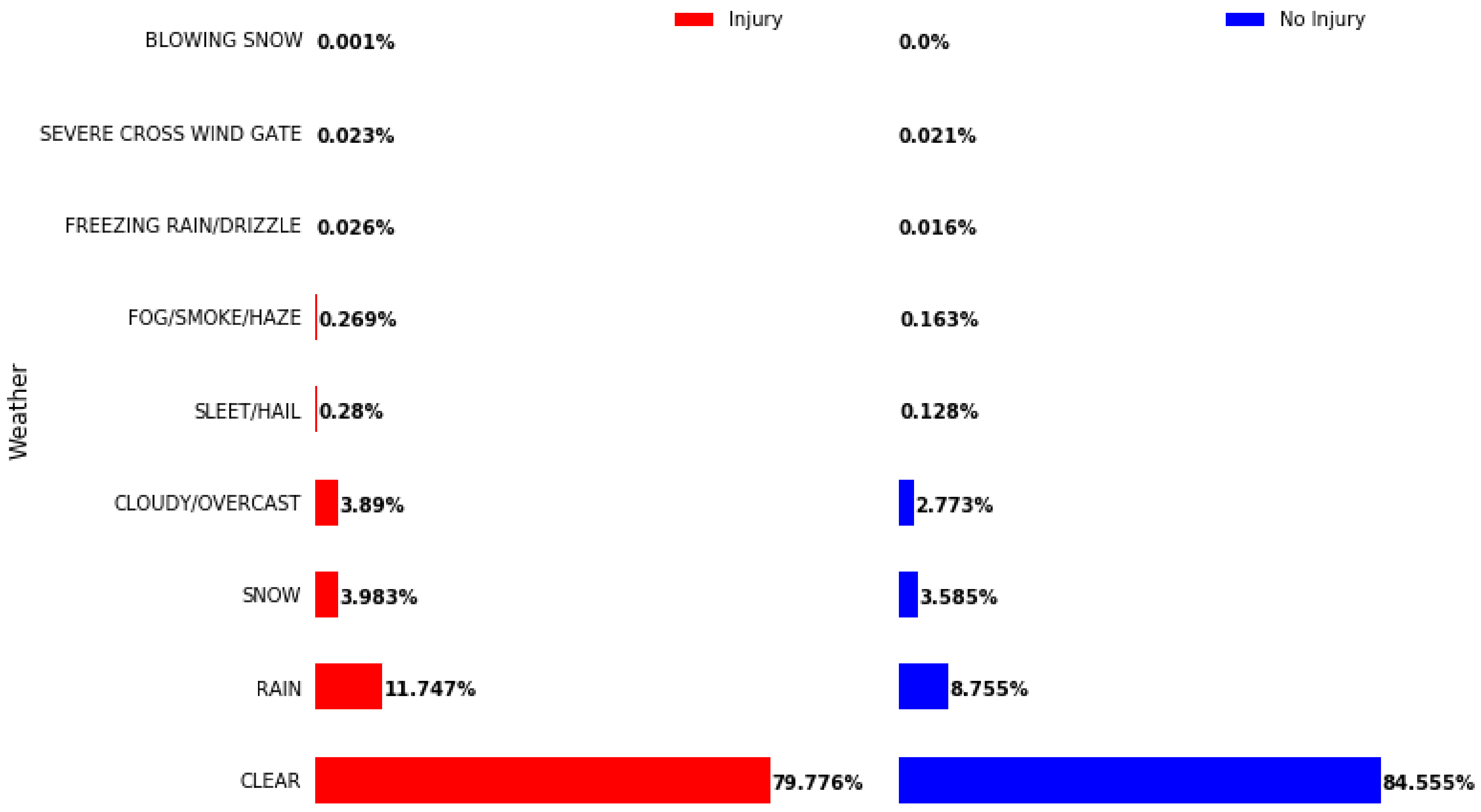
Lighting Conditions for Each Crash Type



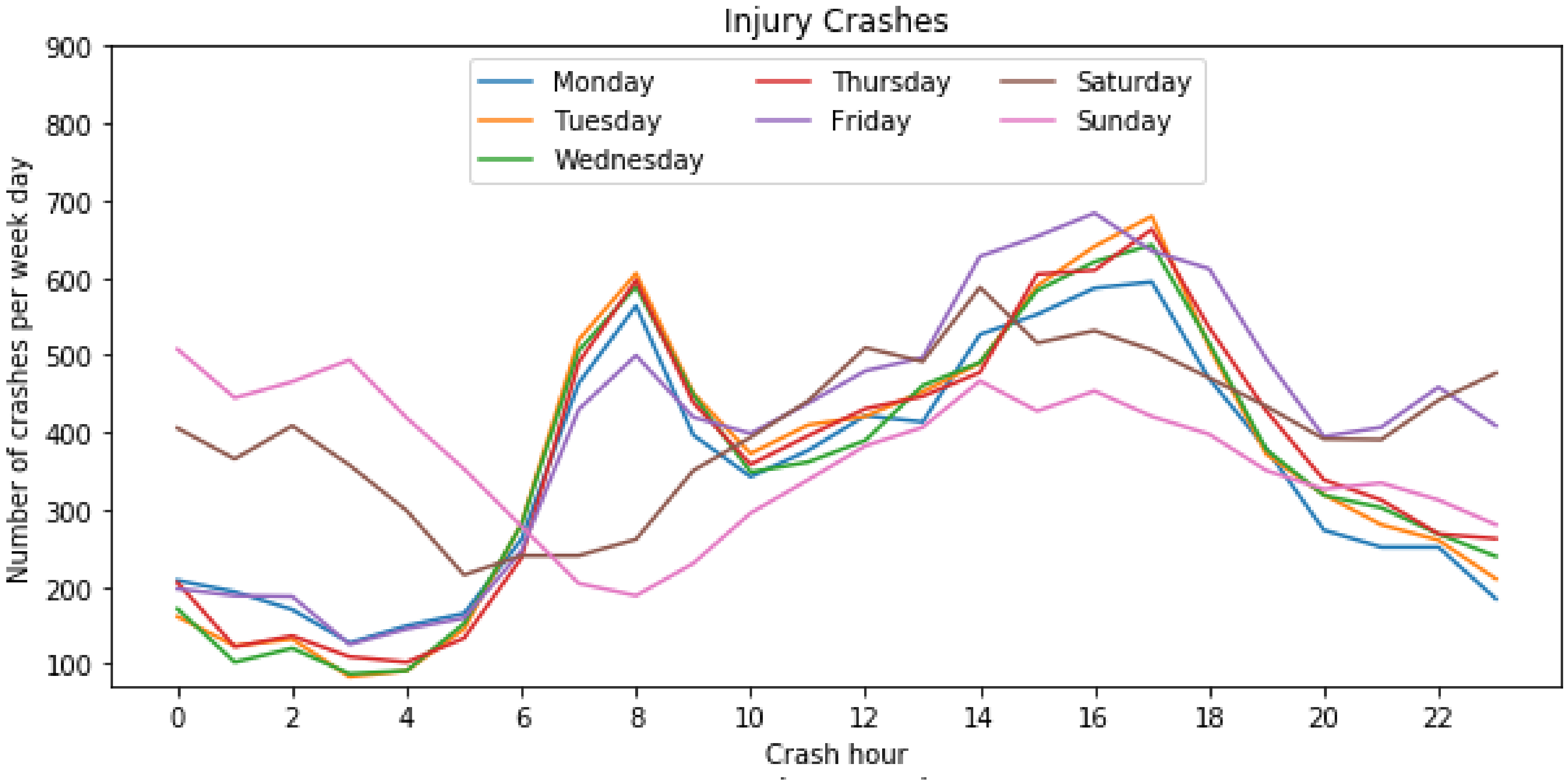
Speed Limit Distribution for Each Crash Type



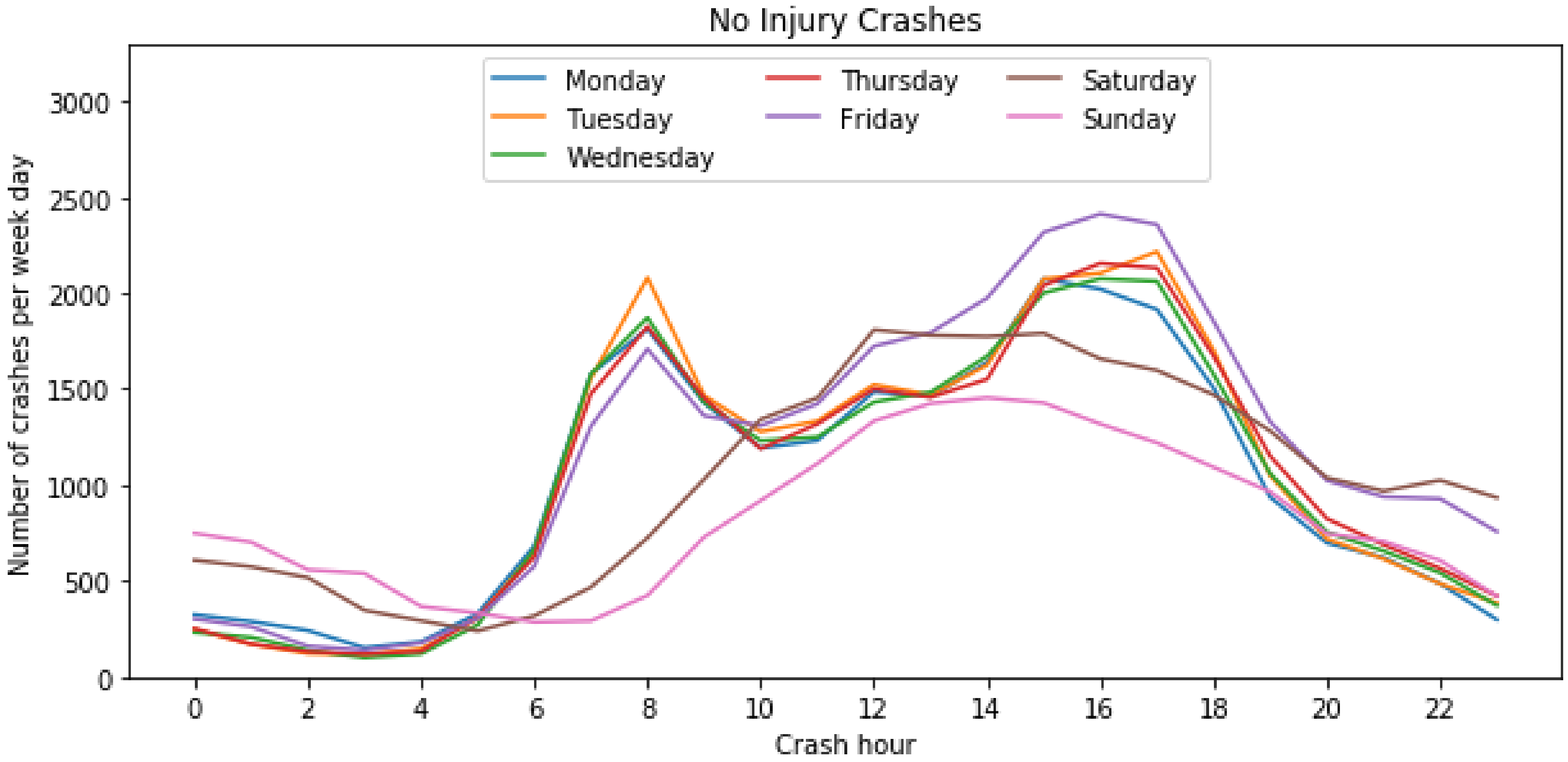
Weather Condition for Each Crash Type



Number of Injury Crashes per Hour



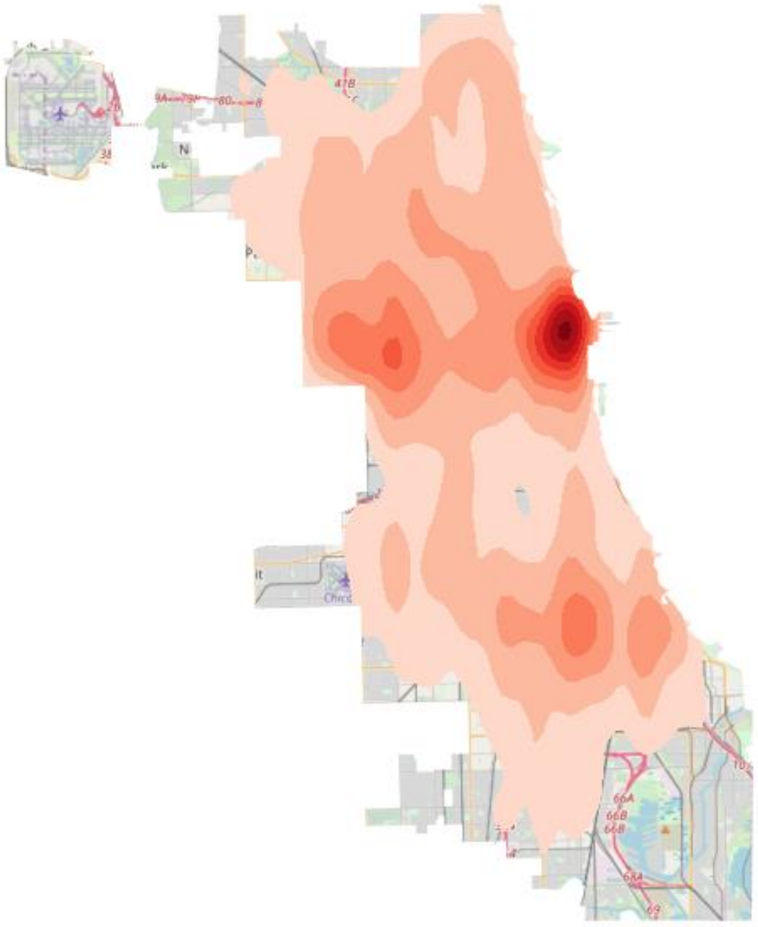
Number of No Injury Crashes per Hour



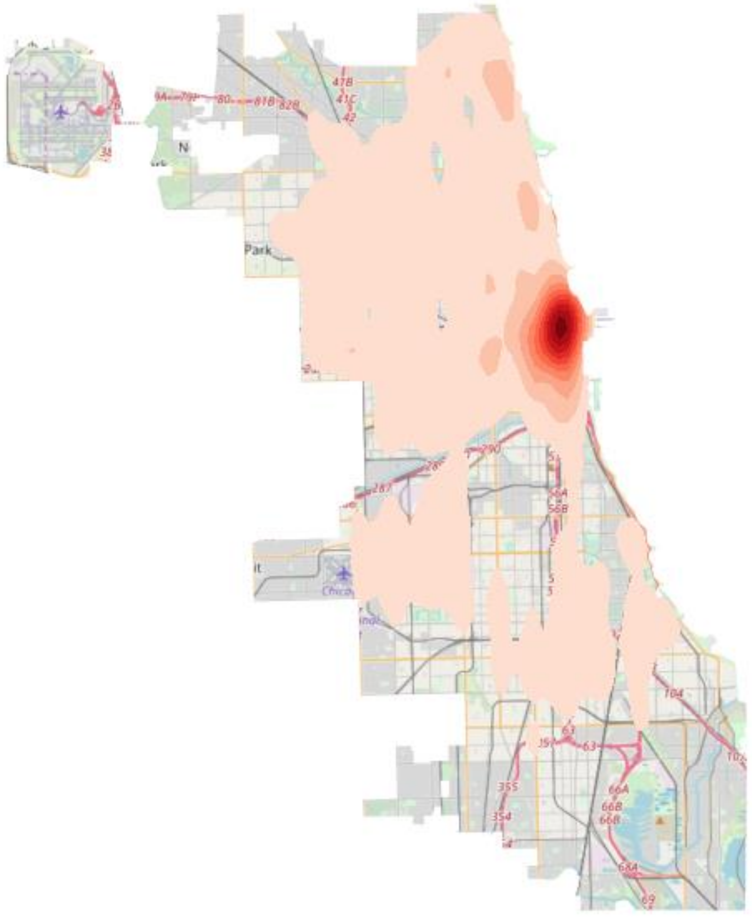
Association Measure between Crash Features and Crash Type

Crash Feature	Cramer’s V Coefficient
Collision	0.354896
Primary Cause (Driving Behavior)	0.305708
Trafficway Type	0.166856
Crash Hour	0.140647
Lighting Conditions	0.132704
Posted Speed Limit	0.127492
Control Device	0.121500
Road Surface Condition	0.064132
Weather Condition	0.057022
Road Alignment	0.056438
Crash Day	0.039466
Crash Month	0.022364

Kernel Density Estimation of Crashes’ Location

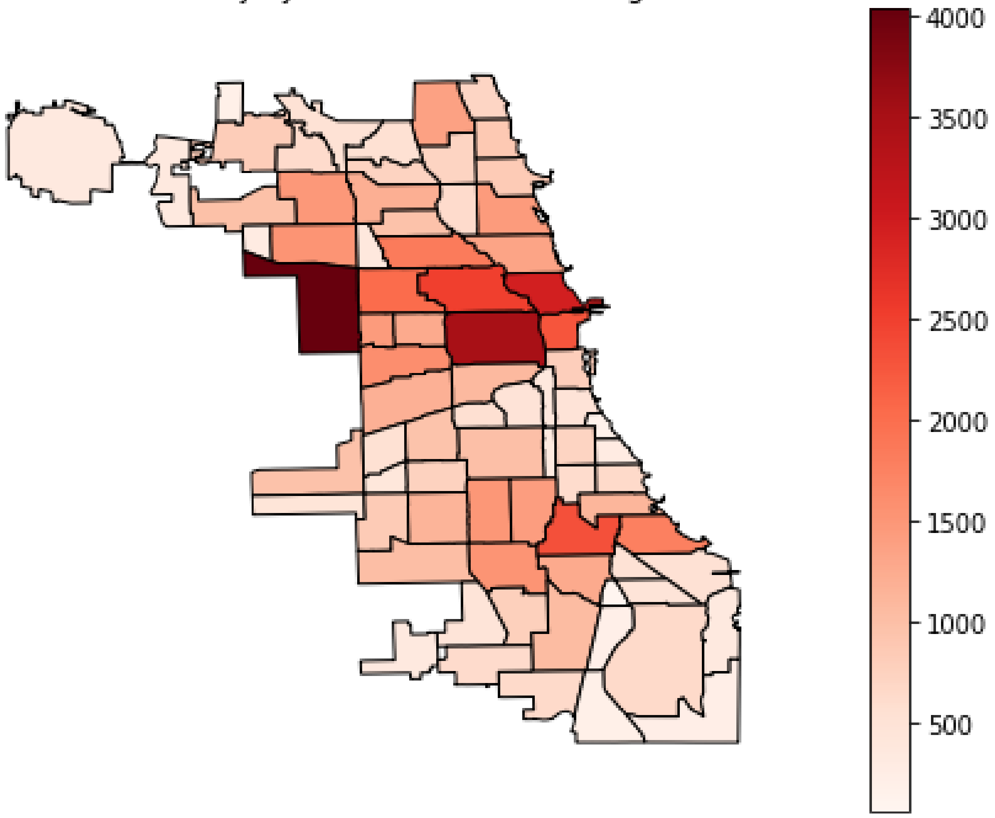


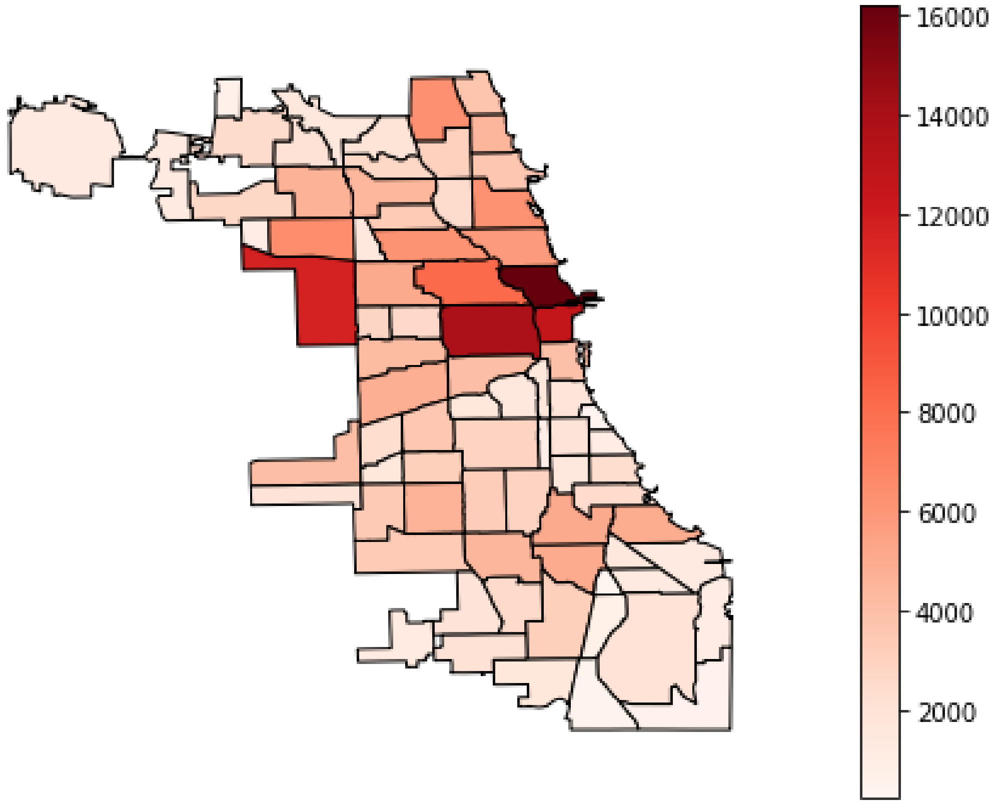
Crashes with injury locations



Crashes with no injury locations

Crashes’ Distribution Across the Community Areas of Chicago


Distribution of Injury Crashes Across Chicago Areas


Distribution of No Injury Crashes Across Chicago Areas

Key Findings

Driving behavior and type of collision showed strongest association with crash severity:

- Example of such behaviors: disregarding traffic signal and stop sign, failing to reduce speed or not giving the right-of-way.

- Example of collisions: angle, turning, collisions with pedestrians.

Key Findings

- Remaining features showed less strong association with crash severity.

- Crashes with Injury do not only happen at rush hours of the weekdays, but also at early hours of the weekend.

- Injury Crashes are not only located in the central part of Chicago but also in the west side of the city.

Building the Predictive Model

In-Depth Analysis

Additional Preprocessing

Additional Preprocessing Steps:

- Added one column “Area” to designate the Chicago area of the crash
- Converted nominal categorical features into numerical entries:
  Tried different encoding schemes (one hot, binary, leave-one-out)

Data Imbalances

- No Injury crashes: 78% ,   Injury crashes: 22%

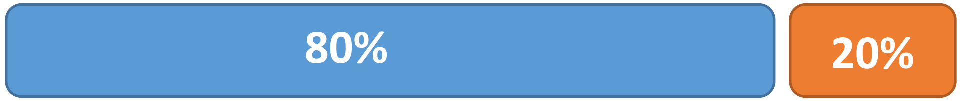
- To address imbalances:

  - Metrics used: precision, recall and F1-score

  - Considered under-sampling of the majority class

Steps of Model Building

- Split the data into training and testing sets:



- Performed 5-fold cross validation on the training set:

  - train different combination of: encoding scheme and training model, with/out under-sampling
  - select the final model using F1-score

- Tested the final chosen model on the testing set: 🟧

Training Models

We tried various models and compared their performance:

- Naïve Bayes
- Logistic Regression
- Linear SVM
- Random Forest
- Ada Boost
- Gradient Boosting
- Balanced Random Forest

Final Model and its Performance

- Selected Final Model (trained after undersampling):

  Gradient Boosting (parameters: *n_estimators=600, max_depth=4*)

  + **Leave-one-out Encoding**

- Performance on the testing set:

Accuracy	**0.747**
Precision	**0.697**
Recall	**0.457**
F1-Score	**0.552**

Features’ Importance

Crash Feature	Importance
Collision	0.418
Primary Cause (Driving Behavior)	0.287
Trafficway Type	0.071
Area	0.054
Crash Hour	0.042
Posted Speed Limit	0.041
Lighting Conditions	0.033
Control Device	0.016
Crash Month	0.0086
Crash Day	0.0066
Road Alignment	0.0062
Weather Condition	0.0059
Road Surface Condition	0.0058

Possible Future Works

Additional Work:

- Incorporate more features related to driver and vehicle’s information
- Consider stacking of the models
- Account for location in terms of zip code instead of the code area
- Perform street level analysis
- Focus on crashes with injury and analyze the conditions of possible types of injuries (fatal, incapacitating and non-incapacitating)

Recommendations

Driving behavior: important feature in predicting severity of crashes especially at *intersections*

Efforts should be focused on:

- Pushing drivers to drive recklessly
- Keeping on educating drivers of defensive driving techniques
- Helping drivers not loosing attention