# Data-A-Team

## Project 5: Predicting Injury and Fatalities in Car Accidents

#### About the Data:

Fatality Analysis Reporting System (FARS) Analytical User’s Manual, 1975-2020: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813254 This multi-year analytical user’s manual provides documentation on the historical coding practices of FARS from 1975 to 2020. In other words, this manual presents the evolution of FARS coding from inception through present. The manual includes the data elements that are contained in FARS and other useful information that will enable the users to become familiar with the data system.

Fatalities and Coding and Validation manual (https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813251) (1225 pages) Provides more detailed definitions for each data element and attribute for a given year.

#### Problem Statement:

First Responders want to be better prepared by the time they reach an accident site

>\* Based on number of vehicles, people, time of day, weather conditions, how many fatalities and what level of injury severity to expect

>\* What are the worst days/times when they should be appropriately staffed

>\* What can be done to reduce the number of fatalities and severity of injuries

---

#### Methodology:

>\* Feature and Preprocessing

Feature Selection: I used a correlation matrix and compared all features to Injury severity. From there I chose the top ten positive and top 10 negative correlated features. If the features were categorical, I dummified the features.

Dummifying: To avoid collinearity between features through One hot encoding for features that showed to be correlated to injury severity and dropping the first value for each. Before dummifying, I had \_ features.. After dummifying I had \_ features

---

#### Model Selection:

>\* Model selection: Logistic Regression, Random Forest Classifier

>\* Logistic Regression

o How much do each feature lead to fatalities

>\* Random Forest Classifier

- Same variables as LR; predict Fatalities

- Hyper tuning using GridSearch

>\* Metrics: Accuracy

Both models can be used for easier interpretability.

---

#### Results

Baseline Score:

|Baseline Score| Log Reg Score (test) | Random Forest(test) |

|---|---|---|

|60.70%| 66% | 70% |

Based on the results, the Random Forest performed better than the lr model although the performance is on slightly above the baseline score. This may be due to the fact that the Vehicle dataset has features that are relevant that \*\*can\*\* help improve our model's accuracy, they may not be the \*\*best\*\* features for predictions.

---

#### Recommendations

With this, we do know that coefficients like the speed of the vehicles before the accident, make, model, etc. can help us infer the severity of a person’s injury

To help avoid accidents in the future and make the jobs of first responders easier, more enforcements of speed limits with higher patrolling should be in effect in highways that are higher prone to accidents

Next Steps:

Continue to create and select better features from other sources

Try to combat the unbalanced class of the target variable by getting accident data where no deaths were involved. This could help us predict non fatal injuries more accurately