Evaluation of the factors which contribute to Traffic Collisions.

Introduction.

The intention of this report is to evaluate the factors that contribute to Traffic Collisions to see, if we can predict whether a collision will result in an injury. By being able to predict factors in a collision that leads to an injury, governments and the relevant regulatory bodies can then be able develop strategies to decrease the number of injuries caused by road collision. The RSA has reported and increase trend and in both serious injuries and fatality on Irish roads in recent years (Road Safety Authority, 2023).

Methodology

This project will be evaluating a publicly available data from Kagel called “Chicago Car Crash Dataset” (Savorsauce, 2023). This is data is transcribed from reports from the Chicago Police Department following a road collision. Its features include the outcome of a collision, the road and environment conditions, along with time/locations and information of the police activities following the crash. This dataset has 746,498 observations and 49 features.

Some of these features will need to be removed as they will be recording different activities after the collision has occurred. For example, there is a feature called “DATE\_POLICE\_NOTIFIED” which obviously isn’t a factor in the crash itself. Another challenge of this project will be that much of this data is categorical data. Out of the 49 features 32 are categorical. Our Target Variable will be the "MOST\_SEVERE\_INJURY" feature, out of the 746k observations 643k are No indications of injury.

This report will test 3 machine learning models, Decision trees, Random Forest, and Logistic Regression, with a test-train split at 10%, 20% and 30%, to determine the best model at predicting is a collision will result in an injury. We choose Decision Tree and Random Forest, as tree base models are good for categorical features, and they are easier to interrupt which features are important to predicting injuries at collisions (Pratap, 2017). A Logistic Regression model computes the “maximum likelihood estimation” (Pratap, 2017). This is ideal for prediction of whether a injury has or has not occurred.

Data Cleaning and Feature Engineering

As you can see below there are many features which have a high number of Null Values. So the first part of our Data Cleaning was removing this features. Fortunately, most of this data was also data which was recording different activities after the collision has occurred.A black and white bar code

Description automatically generated

Six features were also just other method of recording the target variable. The features were counting the number of fatality, serious and minor injuries, no injury reported, no injury indicated, and total injuries. There was also a number of independent varibles that recording the same or similar data. For example “LATITUDE” and “LONGITUDE” was combined to create “LOCATION”. Due to the redundancy with keeping highly correlated features (Gallatin, 2023) and knowing that we would need to transform the categorical data, we decided to drop these features. All in all we reduce the number of features to 20 before transforming the categorical varibles. And reduce the number of observations to 744,859.

Our target variable is 'MOST\_SEVERE\_INJURY’ As shown in the below graph has five outcomes one for ‘NO INDICATION OF INJURY’ and the other four are some type of an injury. The data is also massively imbalanced. A graph of a number of blue rectangular objects

Description automatically generated with medium confidence

Imbalance data can create poor predictive modelling as most models’ algorithms, including the three we are going to apply, are designed to assume that each classification is equal, and by it not being equal it creates bias in the prediction (Brownlee, 2020). To combat this, we reduce the number of ‘NO INDICATION OF INJURY’ classifications to the sum of all the other classifications and then encoded ‘NO INDICATION OF INJURY’ to ‘0’ and the other four to ‘1’. Next, we needed to Transform all Categorical Data to numerical data, in order for our models to run with them. After we finished the all the Feature Engineering, we have a dataset of 202816 observations and 195 features.   
  
Model Outcomes.

Decision Tree

As stated above we run each model at a train/test split of 10%, 20% and 30%. For decision tree was we used a combination of

Random Forest

Logistical Regression

Entropy, log lost what difference