Car Accident Severity Prediction in Seattle

Ludovico Panu

September, 2020

1. Introduction

1.1. Background

Car accidents are one of the biggest cause of injuries and deaths all over the world. They also impact economy with increased commuting times, increased delivery times and costs, and pollution, due to the massive number of cars waiting for the road to be cleared, or heavily slowed down.

1.2. Problem

Building a model to predict car accident severity can help commuters, professional drivers or logistic planners to reduce the personal and/or business impact of car accidents.

1.3. Interest

Described problem and its analysis will be the most interest of two categories of stakeholders:

- Individuals, being them work commuters or professional taxi, truck or bus drivers
- Businesses, like logistic companies, public/private passengers bus companies, taxi companies, government agencies (urban/suburban mobility managers)

2. Data acquisition and cleaning

2.1. Data sources

Collision data had been fetched from Seattle Department of Transportation Open Data Program in CSV format.

Total number of events is around 200k including incomplete rows

2.2. Data cleaning and Feature selection

To predict the severity of potential accident, I had to select the relevant Features within the Data Set. In fact, the severity of the accident could depend on several parameters such as the time of day, the weather, the location of the second car, the type of the car, the number of persons inside the car, and so on. How to select the most important factors that affect the accident severity? How to use them to build a model that can predict from specific conditions (features), the probability of occurrence and the severity of an accident?

To reach the desired objective of this project, I had to choose the features that can affect most the accident and are therefore highly correlated with the labeled target (ie. SEVERITYCODEin the data). The chosen parameters are the following:

DDRTYPE: Collision address type
PERSONCOUNT: The total number of people involved in the collision
VEHCOUNT: The number of vehicles involved in the collision.
INCDATE: The date of the incident.
INCDTTM: The date and time of the incident.
JUNCTIONTYPE: Category of junction at which collision took place.
INATTENTIONIND: Whether or not collision was due to inattention
WEATHER: The weather condition
ROADCOND: The condition of the road
LIGHTCOND: The light condition
SPEEDING: Whether or not speeding was a factor in the collision. (Y/N)

Data Set had been checked for correlation to validate the features choices made. As the Data Set was very imbalanced (70% labels "1" versus 30% labels "2") I reduced the total rows randomly choosing an equal number of labels. Then the data set had been split into a training set and a testing set (80% / 20% of data set).

3. Modeling

Several algorithms had been run, tuned and tested: k-nearest neighbors algorithm, Decision Tree, Support Vector Machine and Logistic Regression.

4. Evaluation

Confronting the metrics of the different models, we look for the highest F1 score, the largest Jaccard score and the smallest log loss.

So I selected the Decsion Tree, and retrained the model over the integral Data Set (without splitting it).

5. Conclusions

The model can non be used to predict the severity of an accident, on the base of features (parameters) observable on site or remotely, to let the stakeholders make an informed decision