Predicting Car Accident Severity in Seattle City

Daniel Orozco Venegas

October 10, 2020

1. Introduction

1.1 Background

The Seattle City population in 2019 was 747, 300 people and had grown at a 22.78% rate since 2010, at the same time, 81% of Seattle households owned at least one vehicle in 2019. Furthermore, in 2019 Seattle City reported 9412 car accidents. Therefore, it is of great interest to reduce car accidents, and it can be achieved by making citizens aware of how likely they are to be implicated on a severe car accident. However, predicting car accidents and their severity is not a task humans can accomplish when thousands of cars, pedestrians and pedalcyclists are involved in the prediction.

1.2 Problem

Data that might help to determine car accident severity might include weather condition, speeding, light condition, road condition, and how crowded a place is. This project aims to predict whether and how severe a car accident will be based on these data.

1.3 Interest

The Seattle City government is interested in reducing car accidents because it negatively impacts quality of life, and has high costs, both human and economical. Other major cities governments may also be interested in the project as they could implement it in their cities. The National Health Service may as well be interested because reducing car accidents will improve public health, and reduce health assistance costs.

2. Data

2.1 Data Source

The Seattle City car accident statistics, including weather condition, speeding, light condition, road condition, and number of involved people, can be found in one Kaggle dataset here, and it has 194, 673 events. This dataset, however, only have data since 2004 until 2020, but there is no data available

before 2004. Also, only ~5% of the observations have values for the speeding feature, which is considered as one of the most important features to be included on a car accident severity prediction.

2.2 Data cleaning

There is no other dataset about Seattle City car accidents specifying speeding for enough observations to be combined or replace the dataset from here. Therefore, as the studied dataset only has speeding values for ~5% of the events, the speeding feature was discarded for further consideration.

Some of the columns that had different values to express the same were normalized. Also, any row containing null values was removed from the dataset, this is because the number of rows with null values was small and the resulting dataset still preserved a large number of samples. No normalization was required as the dataset had its numerical values already normalized. One outlier with 81 people involved in the accident was removed too.

2.3 Feature selection

After data cleaning, there were 182, 895 samples and 37 features in the data. By examining the meaning of each feature, it was determined that there was redundancy in the features and that some of them were irrelevant for the car accident severity prediction, some columns were simply used like keys.

Now, other features might have seem like redundant, for example the number of people involved in the accident, number of pedestrians involved, and the number of pedalcyclists, but there is a big difference in the probabilities of suffering serious injuries if people involved in the accident were inside of a car or not. Therefore, the different counts of different kind of people involved in car accidents were preserved.

After discarding redundant and irrelevant features or columns, 12 features were preserved, this is the list of selected features: number of people involved in the accident, number of pedestrians, number of pedalcyclists, number of vehicles, if a parked car was hit, address type, collision type, junction type, under influence of drugs, weather, road condition, and light condition.