Predicting the Severity of Car Accidents in Seattle (2015-2020)

Fathan Mubina

October 8, 2020

1. Introduction

1.1 Background

Car accidents often occur anytime and anywhere, causing minor to severe impacts to the passengers, the car(s) involved, public facility, and people around the area. The impacts are not only physically, but also economically. The severity of car accidents is driven by many contributing factors related to human, vehicle, and environment of the incidents. It is highly important to identify the contributing factors to the severity of the incidents so that the response to the incidents can be predicted and the prevention and mitigation can be planned well. Hence, it is beneficial to comprehend the relationship between car accident severity with its contributing factors and build machine learning models to predict car accident severity. For example, the model can be used by the government or road managers to improve the infrastructure of the road or the area.

1.2 Problem

The data might contribute to predicting the severity of car accidents, driven by its various contributing factors, such as location, road condition, lighting condition, weather condition, and human condition. The objective of this project is to predict how the factors of the car accidents contribute to the severity of the car accidents.

1.3 Interest

This project might interest the government or road managers to identify the key items to improve in managing the roads or areas to prevent other incidents to happen. Emergency units might also be interested in this project that enables them to response to the incidents quickly and effectively. App developers might be interested in creating an app for users to alert them about the driving conditions and to improve the users' driving experience.

2. Data Acquisition and Data Wrangling

2.1 Data Sources

The dataset used in the project is acquired from Coursera's Applied Data Science Capstone course that can be found in this link. The dataset contains car accident severity data in Seattle, Washington, USA from 2004 to 2020. The dataset consists of 38 columns and 194,673 rows. The severity code of the dataset are categorized into 1 (Property Damage Only Collision) and 2 (Injury Collision)

2.2 Data Wrangling

The dataset contains data points from 2004 to 2020. Given the project wants to predict the severity of car accidents based on their contributing factors, it is assumed that the data older than 5 years might not be relevant as improvement or change in environment and regulation

might have been implemented. Hence, the dataset is filtered for data points of car accidents in Seattle from 2015 to 2020.

In the dataset, there are a lot of data points that have missing values of location coordinates (X and Y). Because the project wants to map the location of incidents and might provide insights on which area that has more or less car accidents, the data points with missing coordinate values are then eliminated. In other features, missing values are also found. Given the big dimension of the data, the missing values from contributing features are also eliminated, assuming that the remaining data points are representative to predict the severity of car accidents based on the contributing factors. However, in some features, such as weather, light condition, and road condition, the data values are identified as “Unknown” or “Other”. These values might not be representative to determine the contributing factors of car accident severity. Therefore, the values are considered as missing values and eliminated from the dataset.

After data filtering and data cleaning, there is 42698 sample in the data set. As the project wants to build a model that predicts the severity of car accidents by its distributing factors, some features are selected to be included in the modeling. The twelve features include severity code, location (X and Y coordinates), date of incident, collision type, junction type, number of person involved, number of vehicle involved, weather, light condition, road condition, and address type.