Capstone Project - Car Accident Severity

1. Introduction

1.1 Background

Car accident causes damages to both properties and injuries, even death. Approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities[1].

Everyone wants to have a safe road journey, however, various of factors could lead to an accident in a second and human beings cannot take 100% perfect measures to prevent the accident from happening.

1.2 Business Problem

The problem that we are going to solve is to predict the severity of having a car accident by taking into consideration multiple factors that could lead to a car accident.

This prediction can help to warn the driver in advance and can therefore, hopefully decrease the real accident rate.

1.3 Interest

The government could be potentially interested in the project as it could help to decrease local the injury or fatality rate.

Besides, insurance companies could be a potential user as they can prevent the car accident and therefore, decrease the insurance policy claim cost.

2. Data acquisition and cleaning

2.1 Data source

I got the data from coursera suggested list: you can also find the data in

Applied Data Science Capstone – week1 – Downloading Example Data Set.

The Data has 38 types of car accident relevant factors and include 1 946 734 rows of records.

The data set looks very comprehensive and should have enough sample to train the ML model and get a decent prediction result.

2.2 Feature selection

After analyzing all the attributes, I decide to use the following 7 attributes as input:

- INCDTTM: The date and time of the incident.
- INATTENTIONIND: Whether or not collision was due to inattention. (Y/N)
- UNDERINFL: Whether or not a driver involved was under the influence of drugs or alcohol
- WEATHER: A description of the weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision
- SPEEDING: Whether or not speeding was a factor in the collision

[1] Data from Association for Safe International Road Traveling, weblink

The output is about PERSONCOUNT - The total number of people involved in the collision. I strongly believe that the number could be a good measurement for the severity of the car accident and I categorize the severity as below:

- Level 1 means only property damage (0 person involved)
- Level 2 means light car accident(1-2 persons involved)
- Level 3 means medium (3-5 persons involved)
- Level 4 means high (5-10 persons involved)
- Level 5 means heavy (more than 10 persons involved)

One common point is that almost all the data is categorial and not continuous.

2.3 Data cleaning

There are several problems with the dataset.

Firstly, there are many missing values for different attributes. I mainly used the following three ways to handle with it:

- For attribute with Yes and No value, as there is no NO answer, therefore, I assume all the missing value means No
- For attribute with few missing values (less than 5%) and hard to predict the value, I dropped the line directly
- For attribute with few missing values (less than 5%) but easy to predict the value, like the time per day, I use the mean of the time series to fill the empty value.

Secondly, most attributes are in text format. Therefore, I choose to define the text with numbers and convert them into integer for further modeling.