Assignment on

Car Accident Severity – Prediction using Machine Learning Algorithms

Submitted in partial fulfillment of the course

Applied Data Science Capstone

by

Sanket Amdalli

Date of submission – 22th September 2020

1. Introduction

1.1 Background

Road accidents cause upwards of 1.35 million deaths, and between 20-50 million injuries every year globally, as reported by the World Health Organization. The major factors that cause road accidents and increase the risk of injury from accidents, stated by the WHO, are - Speeding, Driving under the influence of alcohol and other psychoactive substances, Nonuse of motorcycle helmets, seat-belts, and child restraints, Distracted driving (using mobile phones), Unsafe vehicles, Unsafe road infrastructure, and Inadequate post-crash care. While most of these factors are the responsibility of the vehicle driver, the last two (road infrastructure and post-crash service) are under control of the government, and road authorities can certainly take many steps to ensure that accidents are averted, and when they do happen, the casualties are attended to as quickly as possible.

1.2 Problem

We see warning signboards of "Accident Prone Area, Drive Carefully" in places where a lot of accidents have happened in the past. This is a rudimentary example of how a road authority has used a simple data point (number of accidents) to derive an insight (accident-prone area) and set-up a warning sign for drivers. The idea of this project is to use the same principle of deriving insights from different data points, which can be used by the authorities to improve the infrastructure and post-crash service. With various attributes about the accident – the time of day, location, road conditions, light conditions, weather conditions, etc. – available to us, a deeper analysis can be performed to determine the extent to which these factors impact road accidents, and consequently a model can be developed using the key factors to predict the severity of potential accidents.

1.3 Interest

The primary consumers of our model (target audience) will be the governments – the road/transport authorities, who can use insights from the model and build a mechanism to alert vehicle drivers about difficult driving conditions and potential hazards, thereby reducing the frequency of accidents. By identifying the severity conditions, they can also inform the police and medical authorities to be prepared for quick action in case an accident happens. In the long run, insights from the model can be used while planning the construction of new roads (to answer questions like 'Do we need more streetlights on this road?' 'Is it better to construct a junction with traffic lights or simply a roundabout?', etc.). For the scope of this project, the government authorities of the city of Seattle, Washington are the target audience.

2. Data Understanding

2.1 Data Source

The data source used for the project is collected by the Seattle Police Department, and maintained and owned by the Seattle Department of Transport (SDoT). It contains the records of vehicle collisions in the city of Seattle since year 2004, with attributes such as the accident severity, location, timestamp, number of people involved, number of vehicles involved, number of injuries/fatalities, whether a driver was speeding, whether a driver was driving under influence, road conditions, light/visibility conditions and weather conditions among others.

Total number of records (as on 20 Sep 2020): 221,525
Number of fields: 40
Update frequency: Weekly

The homepage for the dataset is here:
https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

Detailed attribute information can be found at this link:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

Metadata for the dataset is here:
https://www.arcgis.com/sharing/rest/content/items/5b5c745e0f1f48e7a53acec63a0022ab/info/metadata/metadata.xml?format=default&output=html

2.2 Data Cleaning and Preparation

In order to prepare the dataset for analysis, the necessary cleaning processes need to be performed. This includes – handling Missing/Null values, converting the variables into the appropriate data type, removing outliers (if any), and balancing the dataset.

2.3 Feature Selection

After cleaning the data, the features were examined, and I decided to retain only those which would help in the modelling, dropping the ones that were redundant or irrelevant to my analysis. Only the attributes that could be possible 'causes' for affecting the severity of accidents need to be included in the model. For instance, the No. of fatalities/injuries/vehicles attributes have a direct correlation to the severity of accident, however, they are a consequence of the accident, not a cause for the accident. So they have been excluded from the model.

Also, attributes such as whether the accident happened due to driving under influence, speeding, inattention are factors that may affect the severity of accidents, but these are factors which are in control of the drivers, and it is impossible for the road/transport authorities to ascertain beforehand whether a person is driving under influence and send out alerts of potential accidents to other drivers. There can certainly be a correlation between say, an area which has a lot of pubs/bars, and the possibility of severe accidents happening in that area at late evening/weekends because of more people driving under influence. And such a correlation must be explored. But it is currently beyond the scope of my analysis, and hence I have excluded these attributes from my model. Some other attributes which are simply long-form descriptions of other coded attributes in the dataset, and hence are redundant, have also been excluded.