IBM Capstone Project – Modelling Accident Severity Level from a Car Collision

Business Understanding

The impact of road traffic accidents goes beyond property damage. In most cases, human lives are at stake. In more serious cases, fatalities can occur. While road accidents can be reduced through better urban designs and enforcement of the law, more can be done to understand how various factors can influence the severity of the accidents.

Besides the common factors such as weather conditions, road conditions and visibility, other factors such as the type of intersection, the location, the time of day, etc. could also have an impact on both the probability and severity of a car collision if it does happen.

With the increase usage in embedded technology in modern cars, it is possible for cars to have greater awareness of the environment such as current location, weather and visibility condition, route to the destination, etc.

Using these environment and location inputs, it would be beneficial to develop a model to predict the accident severity given the current situation. This would be very helpful to drivers as such a model can give the drivers warnings and encourage them to drive safety in such situations. Hence reducing the chance of an accident.

Data Understanding

The data provides a list of accident traffic records from the Seattle Police Department. Based on the period from 2004 to 2020, 194,673 accidents have been recorded.

Each accident is recorded as a row with 37 independent attributes that provide various details on the accident. The dependent variable SeverityCode is used to indicate the different levels of severity caused by the accident.

There are generally 4 levels of severity based on the meta description of the accidents:

- 3 – Fatality
- 2b – Serious injury
- 2—injury
- 1—property damage

The attributes provide more insights into the details of the accidents and can be broadly categorized into the following area:

1. Location data: The location where the accident took place and characteristic of location such as the junction type.
2. External environment data: Data such as the weather condition, road condition and visibility are available.
3. Impact data: The number of injured persons and other vehicles are also recorded.
4. Driver condition data: In some cases where the accident is caused by driver inattention or influence under alcohol is also recorded.

However, we also note in the dataset that not all information is always available. In some cases, certain key feature values are blank and not available. Further data processing is required to clean up the data before further analysis.

In addition, the dataset is not balance with most of the data (136,485) representing accident of severity 1 and only a small set of data (58,188) representing severity 2. No other severity level is found in the dataset. This could result in a model that is bias due to the lack of proper balance of the various scenarios in the dataset. Sampling of the data might be required to transform the data into a more balanced dataset.