Capstone Project: Car Accident Severity Prediction

1. Introduction / Business Problem

Road safety is a serious public concern, thus it is important to carry out studies relating to accident analysis and prediction. Through data analytics, we are able to extract meaningful information and provide insights on key factors that lead to road accidents.

A lot of road accident data collected by government agencies are public available. This project focuses on data obtained from Seattle City to predict the severity of an accident based on time, collision type, road conditions, and other environmental factors. This would be useful in the development of traffic rules and accident prevention policies, as well as provide practical information to the general public.

2. Data

2.1 Data Source

The dataset contains collision data collected by the Seattle Department of Transportation from 2004 to 2020. It contains attributes such as location, date, time, weather, light condition, road condition, type of collision, and severity of collision, to name a few. Since the goal is to provide a measure of the severity of accidents at different points in time and space, the severity of collision was the label used to train the model.

2.2 Data Preparation

2.2.1 Missing Values

The dataset has missing data that needed to be dealt with. The junction type, weather, road condition, and light condition columns have observations classified as ‘Unknown’. The missing values are practically unknown values, so it made sense to replace the values with the same class. Observations for address type and collision type were dropped as the most frequent values for each column were not high enough in count percentage to replace the missing values. All other attributes not relevant to my analysis were dropped.

2.2.2 Data Formatting

Attributes with incorrect data types were converted to the proper format, namely date and time of incident, and further processed and replaced by dummy variables containing the hour and day of incident.

The dataset mostly contains categorical data. Most machine learning algorithms cannot operate on categorical data directly, therefore these were converted to numerical data using one hot encoding technique.

2.2.3 Class Imbalance

Imbalanced classes have a negative impact on the accuracy of machine learning models. This occurs when there is a disproportionate ratio of observations in each class. The dataset contains more observations for accidents that resulted to property damage (severity code #1) than accidents resulting to injury (severity code #2). To fix the imbalance, I decided to under-sample the data by randomly removing observations from the majority class (severity code #1) to match the number of observations for the minority class (severity code #2).

2.3 Feature Selection

After pre-processing the data, there were 113,954 samples and 52 features (which increased due to one hot encoding). Some attributes were duplicates or redundant in value, while others were not relevant to my analysis. I'm particularly interested in attributes that relate to time, collision type and other environmental factors, thus attributes influenced by human behaviour were removed. The following features were selected - address type, collision type, junction type, weather, road condition, light condition, vehicle count, hour of day, and day of week.