Cousera IBM Data Science Capstone

**1. Introduction/Problem**

This paper will employ machine learning models to predict the severity of automobile accidents based on data collected by the city of Seattle regarding traffic incidents.Per the National Safety Council there were 38,800 fatal auto accidents in 2019 in the United States, by understanding the factors that contribute to more severe accidents we can look for solutions that may mitigate the fatalities. Including signs that warn drivers when conditions (weather, road surface, visibility etc) are poor to induce safer driving and allocating emergency resources more efficiently to increase response time and lower fatalities.

**2. Data Source Description**

**2.1 Data Description**

The data I will be using was provided by Coursera from the city of Seattle in a csv format. There are 194,673 observations in the file beginning on 01/01/2004 until 05/20/2020. There are 37 attributes and the label is accident severity (SEVERITYCODE). The labels are unbalanced as 70% of the observations are listed as 1( property damage) . I will need to account for this, so the model does not have bias.

**2.2 Features**

The features I plan to use in the machine learning model are: Location, Weather Condition, Car Speeding, Light Conditions, Road Condition, Junction type, Person Count, Vehicle Count.

The Speeding field only has data in approximately 5% of cases and is null in the other 95%. I am making the assumption that if there is no data in the speeding column either the vehicle wasn’t speeding, or the officer was unable to determine if speeding was involved. Therefore, I made the decision to replace all null values in the speeding column with 0.

**3. Methodology**

Since a serious injury/fatality is a rare event, the data labels are unbalanced, 70% of the incidents are labelled 1 (property damage) and only 30% are labelled 2 (injury/fatality). This topic was not covered in the course materials, so I had to research the problem.

