1. Introduction

1.1 Background

Seattle, a seaport city on the West Coast of the United States, is the largest city in both the state of Washington and the Pacific Northwest region of North America. As of today's date, Seattle is home to a population of 3.4 million people. With the ever-growing population, the total number of personal vehicles had also been increasing exponentially, with the latest number at 457,000, as of 2018. Another research had also shed light on the ownership rate stating that about 81% of Seattle households owned at least 1 car. The increase in car ownership rates will no doubt lead to a higher number of accidents on the road. Statistics done by the World Health Organization showed that approximately 1.35 million people die each year as a result of road accidents, with the average of 3,700 people losing their lives every day on the road.

1.2 Problem

Road traffic accidents cost most countries an estimated 3% of their gross domestic product. The National Highway Traffic Safety Administration (NHTSA) of the USA suggests that the economical and societal harm from accidents can cost up to $871 billion a year to the US. According to the Annual United States Road Crash Statistics journal, every 16 minutes, a car accident that results in death occurs. This project aims to understand the different causes of road accidents, factors to the severity of accidents and eventually predict the severity of accidents for future analysis.

1.3 Interested Stakeholders

The analysis of the severity of accidents will be beneficial to the Public Development Authority of Seattle which works towards improving road factors. This project can also benefit the citizen as they would know how to reduce and mitigate the risk of being involved in a car accident by taking the necessary precautions. Besides that, finding out the causes and possibly predicting the severity of the accident can be useful to car, life and health insurance companies to forecast their strategies on how to produce and market their insurance products.

2. Data Understanding

2.1 Background

The dataset provided by IBM has a total of 194673 observations, with a variation in the total number of observations in each column. Specifically, the dataset consists of 38 features containing both categorical and numeric data and the features provide detailed descriptions of keys provided for each type of incident. However, the dataset had many empty columns which could have been beneficial to the project if the data is present. The columns above are referring to 'pedestrian granted way or not', 'segment lane key', 'cross walk key', and 'hit parked car'.

The machine learning model aims to predict the severity of an accident given the features attached to it. For the variable Inattention, Speeding and Under The Influence, Y is given a value of 1 whereas N and NaN is given a value of 0. For Lighting Conditions, Light is given 0, Medium is given 1 and Dark is given 2. For Road Conditions, Dry is assigned 0, Mushy is assigned 1 and Wet is assigned 2. Lastly, for Weather Conditions, Clear is dictated as 0, Overcast and Partly Cloudy is dictated as 1, Windy is dictated as 2 whereas Rain and Snow is dictated as 3. In some variables, there were unique values such as 'Other' and 'Unknown'. Deleting these rows entirely would lead to a loss of a lot of data which is not preferable.

Hence, to overcome this issue, arrays were made such as to allow each column's unknowns to be encoded based upon the original column and have an equal proportion of the elements. Then, the array was imposed onto the original column to replace the data which had 'Other' and 'Unknown'. The entire process of data cleaning led to a loss of approximately 10000 rows which had NaN data, whereas other rows mentioned before were filled accordingly.

2.2 Feature Selection

After exploring the dataset, a total of 6 variables were choses to be features along with the target variable being Severity Code. These are the features to be used to uncover the insights of the dataset.

Feature variables	Descriptions
INATTENTIONIND	Is driver attentive during driving (Y/N)
UNDERINFL	Is driver driving under influence (Y/N)
WEATHER	Weather conditions during accident
ROADCOND	Road conditions during accident
LIGHTCOND	Light conditions during accident
SPEEDING	Is driver driving above the speed limit when accident occur (Y/N)