1.Introduction

With rising trend of motorization, road safety has become a major concern. According to World Health Organisation, approximately 1.35 million people die each year as a result of road traffic crashes. Not only it costs human lives, but it costs taxpayers in the form of damaged property, post accident healthcare, disability allowance etc. Road traffic crashes cost most countries 3% of their gross domestic product.

Extensive studies have been conducted to find the cause and circumstances under which the accident occurred. For example, Rautela and Sharma (2004) did an analysis on accidents in the Himalayan terrain of Uttaranchal in India using the Fatality Index (FI) and suggested mitigation strategies for accidents on hill roads. Similarly, Johnson et al (2004) studied the rate of usage of cell phones during day and night conditions along-with other demographic characteristics. They found that the most frequent distraction to a driver was the usage of cell phone. In this project we will try to find factors or combination of factor which could cause an accident.

1.1 Business Understanding

The objective of this project is to develop a machine learning model to predict the severity of an accident depending on various factors. There are many potential causes of an accident. Some of the accidents are caused by the people who defy the laws by speeding, drunk driving, distracted driving etc. which can be prevented by incorporating harsher penalties. On the other hand, there are some uncontrollable circumstances such as road conditions, visibility, weather conditions etc. We will try to find certain patterns or correlations between attributes to predict the severity of an accident.

The target audience will be the government, the emergency response services, the police and the driver. Utilizing the model created in this task, the government can plan and prepare themselves better to manage any crisis. They can likewise alert the public, the emergency services and the police to be cautious in case any circumstance that can cause an accident, emerges.

2.Data

2.1 Data source

The data was collected from *https://data-seattlecitygis.opendata.arcgis.com*. The data came with a PDF file which contains the description about the attributes. There are 37 attributes and 194673 rows of data.

2.2 Data Processing

The downloaded data had lots of missing values, duplicate columns, redundant attributes and in some columns, data was not standardized and normalized. The process of data cleaning is discussed in the following steps.

- I started with dropping the duplicate columns. Our target variable had two columns namely "SEVERITYCODE.1" and "SEVERITYCODE". Hence, I decided to drop one of them.
- Upon examining further, I found some attributes that were redundant. For example, “INCKEY” which describes a unique key for the incident, was of no use. Similarly, the “LOCATION” attributes were of no use since we already have the coordinates of the accident. A total of 12 such attributes were dropped.
- The columns “EXCEPTRSNCODE” and “EXCEPTRSNDESC” were entirely empty and the metadata had no description about them, so I went ahead and dropped them from the data frame.
- Next, I used Isnull() function to detect missing data. Missing data from the columns 'ADDRTYPE’, ‘ST_COLDESC', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'LIGHTCOND', 'WEATHER', 'ROADCOND', 'X', and 'Y' were dropped since it accounted for mere ~5% of the total data. After this step, there were 180067 rows of data and 24 attributes left for analysis.
- There were few columns were data was not in same format. For example, "UNDERINFL" had four different values such as 1, Y, 0 and N. Y and N were replaced with 1 and 0 respectively.
- Attribute "INATTENTIONIND" and “SPEEDING” had only one type of response i.e. Y. Others were left empty. It was assumed N for others.
- “INCDTTM” is in datetime64[ns] format which can’t be used for analysis. Therefore, data was extracted from it and few more attributes were created such as “Hour”, “Day”, “Day of the Week”,” Year”, “Weekend” and “Month”.

2.3 Feature selection

2.3.1 Target Variable

Our target variable is ‘SEVERITYCODE’ and it has two type of values.

1	Property Damage
2	Injury

2.3.2 Attribute selection

For the selection of attributes, we ran **Pearson's chi-square test of association.** This test is used when we have categorical data for two independent variables, and we want to see if there is any relationship between the variables. Upon running the test, following attributes were found to have some relationship with severity of an accident. In the next section we will carry out some exploratory data analysis and try to understand the relationship.

Chi square test

```
ADDRTYPE is IMPORTANT for Prediction
JUNCTIONTYPE is IMPORTANT for Prediction
LIGHTCOND is IMPORTANT for Prediction
WEATHER is IMPORTANT for Prediction
ROADCOND is IMPORTANT for Prediction
COLLISIONTYPE is IMPORTANT for Prediction
SDOT_COLDESC is IMPORTANT for Prediction
INATTENTIONIND is IMPORTANT for Prediction
UNDERINFL is IMPORTANT for Prediction
PEDROWNOTGRNT is IMPORTANT for Prediction
SPEEDING is IMPORTANT for Prediction
ST_COLDESC is IMPORTANT for Prediction
HITPARKEDCAR is IMPORTANT for Prediction
Day is NOT an important predictor. (Discard Day from model)
Month is IMPORTANT for Prediction
Year is IMPORTANT for Prediction
day_of_week is IMPORTANT for Prediction
weekend is IMPORTANT for Prediction
hours is IMPORTANT for Prediction
```

3.Exploratory Data Analysis