1. Introduction

1.1 Background

In most cases the major uncontrollable factors for an accident to occur are weather, light and road conditions. Accidents can be mitigated by studying the underlying patterns in the data and giving this invaluable information as a warning and or consideration to the local government, traffic police and the respective drivers of the vehicles. Also this information extracted from the data can also be shared with city planning and urban development construction department who will be taking into consideration the future designs and constructions of the motorways in order to reduce the uncertainty of accidents moreover from the financial point of view the cost and labor involve in these sorts of projects can be managed in an efficient way by the policymakers.

1.2 Problem Statement

Data that commits to determining the severity of accident that includes addresstype, junction type, weather condition, Road and Light conditions. This project aims to predict Severity of accident based on these metrics**.**

1.3 Interest

The target audience of the project is local government, police, rescue groups, car insurance companies and last but not the least the drivers themselves. The model and its results are going to provide some key insights for the target audience that will enables these authorities to leverage on the model in making data-driven decision and reduce the number of accidents and injuries in their localities

2. Data Acquisition and Cleaning

2.1 Data Source

Data provided by the Seattle Department of Transportation (SDOT) on vehicle accidents along with its severity is used to derive insights and patterns on how and when these accidents have taken place with the environmental factors like weather, road conditions etc. The dataset

consists of 40 columns having different features of data like, collision severity, road conditions, number of people involved, location of collision, weather etc.

Source:
https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset

2.2 Data Cleaning

I have dropped certain columns based on problem statement and understanding of the data. Handling Null/Missing values: There are some null values for ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND attributes which are replaced with value 'others'

2.3 Feature selection

Feature selection/engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms.

Based on the problem statement, our predictor or target variable will be 'SEVERITYCODE' because it is used to measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'ADDRTYPE', JUNCTIONTYPE', 'WEATHER', 'ROADCOND' and 'LIGHTCOND'. Regarding exploratory data analysis I thought to choose some of numerical variables such as (PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT)

Table below shows the selected and dropped features from the dataset

KEPT FEATURES	DROPPED FEATURES
‘SEVERITYVODE’,‘ADDRTYPE’, JUNCTIONTYPE’,‘WEATHER’, ‘ROADCOND’,‘LIGHTCOND’ ‘PERSONCOUNT’,‘PEDCOUNT’ ‘PEDCYLCOUNT’,’VEHCOUNT’	‘X’, ‘Y’, ‘COLDETKEY’, ‘REPORTNO’, ‘INTKEY’, ‘LOCATION’,‘EXCEPTRSNCODE’, ‘EXCEPTRSNDESC’,‘SEVERITYCODE.1’, ‘SEVERITYDESC’,‘INCDATE’,’INCDTTM’, ‘SDOT_COLCODE’,‘SDOT_COLDESC’, ‘INATTENTIONIND’,‘UNDERINFL’, ‘PEDROWNOTGRNT’,SDOTCOLNUM’, ‘SPEEDING’,‘ST_COLCODE’, ‘ST_COLDESC’,’SEGLANEKEY’, ‘CROSSWALKKEY’,‘HITPARKEDCAR’, ‘PERSONCOUNT’,‘PEDCOUNT’, ‘PEDCYLCOUNT’, ‘VEHCOUNT’, ‘OBJECTID’, ‘COLLISIONTYPE’,’STATUS’