Predicting Car Crash- Seattle

BACKGROUND: The number of traffic collisions and their victims has been a rising trend globally due to increases in population and motorisation. Traffic collisions disturb the traffic operations, break down the traffic flow, and cause severe urban problems worldwide. Major traffic accidents can sometimes lead to irreparable damages, injuries, and even fatalities. In order to take necessary actions to control this ever-growing problem, extensive research has been carried out into the prediction of traffic collisions in both developed and developing countries using various statistical techniques. Different factors involved in traffic collisions have a substantial effect on each other, thus making it difficult to individually consider any of the parameters when explaining the severity of traffic collisions. Realising traffic accidents as a preventable problem developed countries have implemented different policies and measures to reduce this problem. These include enforcement, education, training and engineering improvements. Any part of this report can be utilised by the government authorities for making necessary policy changes to avoid collisions or to minimise their severity.

OBJECTIVES OF THIS PROJECT: The main objective of the research is to investigate the role of factors in collision severity using Seattle Department of Transportation data and predictive models. Specific objectives include: 1) Exploring the underlying variables such as human characteristics, vehicle characteristics, roadway characteristics, and environmental characteristics that impact collision severity. 2) Predicting collision severity using Decision Tree and Logistic Regression

DESCRIPTION OF THE DATASET: Governments, states, provinces and municipalities collect and manage data for their internal operations. In the last decade, an open data movement has emerged that encourages governments to make the data they collect available to the public as “open data”. Open data is defined as “structured data that is machine-readable, freely shared, used and built on without restrictions. The data set used here is taken from the open data website of the Seattle City. It is published by the Seattle Department of Transportation. The dataset contains information about 194673 collisions, recorded between 2004-01-01 00:00:00 and 2020-05-20 00:00:00.

The data can be accessed through the following link: https://s3.us.cloud-objectstorage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/DataCollisions.csv

The metadata for the same can be accessed through: https://s3.us.cloud-objectstorage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/ Metadata.pdf

2 APPROACH: First of all, the data set will be analysed using data visualisation tools and libraries in python to identify trends in collisions and parameters affecting the collisions. Then the data set will be modelled to predict collision severity. The data set mentions 2 levels of collision severity: 1- Property Damage Only Collision 2- Injury Collision The approach for modelling collision severity involves statistical modelling considering severity as a dependent variable while road conditions, speeding, driver attention, influence of drugs/ alcohol on driver, junction type where the collision occured and a few environmental factors as the independent variables. ASSUMPTIONS: A few of the columns in the data set contained categorical values, 'Y': Yes, and NaN. It is assumed that the NaN values correspond to 'N':No. It is also assumed that the data values -'Other' and 'Unknown' correspond to Null as they tell us nothing about the features in the dataset. Exploratory Data