**COURSERA IBM CAPSTONE PROJECT**

I**ntroduction/Business Problem**

Problem Description and Background:

The city of Seattle collects data on motor vehicle collisions, and the data sets are available for the last several years. It is the objective of this project to analyze the data and use the insights gained from the data to predict the patterns in conditions that lead to accidents, and the severity of the accidents.

These insights can be made available to the city and the public, who are the audience/ stakeholders, and can be used to enhance public safety and minimize accidents and their severity. For example, if accidents are found to occur at a higher rate under some weather conditions and in some locations or intersections, the members of the public may try to avoid driving under those weather conditions in those locations in future, whenever possible. City officials may also take measures to improve public safety. For example, if more accidents occur after dark in the absence of street lighting, the city may provide street lighting to minimize accidents in those locations.

In this report, the data set provided will be analyzed and recommendations will be made based on the conclusions from the data analysis and machine learning techniques applied to the data.

**Data**

Description of Data and How to Solve the Problem:

The city of Seattle has provided the data set on collisions in Seattle from the year 2004 to present.

The data set provided contains the following 38 columns:

['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',

'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',

'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',

'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',

'INCDTTM', 'JUNCTIONTYPE', 'SDOT\_COLCODE', 'SDOT\_COLDESC',

'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',

'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST\_COLCODE', 'ST\_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']

The severity code describes the severity of the collision, with 1 being property damage and 2 being injury. The initial analysis of the data found the following attributes to be important:

• ADDRTYPE: intersection, alley or block

• LOCATION: location, from which the locations having large number of accidents can be identified.

• COLLISIONTYPE, which describes the type of collision, e.g., rear-ended, parked car, etc.

• PERSONCOUNT, which describes the number of people involved in the accident

• PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, which describe the number of pedestrians, bikes or vehicles involved in the accident

• JUNCTIONTYPE, which describes the type of junction

* SDOT\_COLDESC, which describes the accident
* INATTENTIONIND, which describes whether the driver was inattentive
* UNDERINFL, which describes whether the driver was driving under the influence of alcohol
* WEATHER, ROADCOND, LIGHTCOND, which describe the weather, road and light conditions
* SPEEDING, which describes whether the driver was speeding
* In addition, the season (fall, winter, summer or spring) will be extracted from the incident date column to describe the seasons in which the most collisions occur.

A supervised learning approach will be used to fit and predict the severity of collisions based on the above attributes, and to identify the most important conditions occurring in the collisions more frequently. Pearson correlation coefficients may also be applied to test which attributes correlate most strongly with the occurrence and severity of collisions. Classification methods like nearest neighbors, decision tree, SVM and logistic regression will be applied.

The fits and predictions will be used to make recommendations that may help reduce collisions.