1) Introduction & Business Problem

Traffic accidents are a significant cause of deaths and severe injuries in the World, and are a major concern for public health and traffic safety:

Traffic accidents have also a severe impact for a country's finances, for instance, according to the National Highway Traffic Safety Administration (NHTSA), U.S. motor vehicle crashes in 2010 cost almost $1 trillion in loss of productivity and loss of life.

Accident severity prediction can provide crucial information for accident management protocols creation, using modeling techniques such as accident Bayesian network and Regression.

The present work focuses on conducting an accident severity modeling in order to find a suitable algorithim to predict the severity of an accident given various factors such as current weather, road and visibility conditions to name but a few.

The goal is to use this model to alert drivers of the increased potential for a car accident specifically using data from Seattle's local government and transportation department.

Identifying the major causes of accidents and and modelling predictions for car accident severity should help the authorities to prevent or at least mitigate accidents, reducing the number of deaths, accidents and injuries for Seattle metropolitan area.

2) Data

The data set for this work contains traffic accident reported to SDOT Traffic Management Division; the traffic accident for Seattle contains All Years Collisions from the year 2004 to date, and is a dataset updated Weekly.

There are 38 attirbutes, some more relevant than others to identify predictor or target variables: There are severity of accident information as well data regarding accident characteristics (accident occurrence time and accident location), environmental factors (such as Weather and LightConD) road conditions RoadCond. Many attributes of the dataset have great descriptive value depicting information about an accident but are not so relevant to predict the severity of an accident more specifically.

Based on a preliminary analysis SEVERITYCODE is the dependent variable and it is clear that most of the features are of type object, so they will need to be converted to numerical type. Furthermore many features of the dataset have null values, it is therefore necessary to understand during the data pre-processing phase if key variables such as pedestrian right of way (PEDROWNOTGRNT), car speed, driver's inattention can be used, after some transformation, or have to be dropped and left out of the analysis.

Upon proper manipulation, the Dataset balancing and preprocessing before been utilised for ML purposes.

Methodology:

The Collisions dataset consists of 194673 recorded accidents. Each row represents an accident, and for each accident, the corresponding circumstances are documented using 38 attributes like severity, type of weather and road condition, location, address type to mention but a few.

- **Exploratory Data Analysis**

In this preliminary part I have analyzed the dataset and summarised some preliminary aspects:

It is clear at this stage that the dependant variable for our data analysis is the attribute SEVERITYCODE, as it measures the severity of an accident from 0 to 5 and this is ultimately what we would like to predict.

On the provided dataset many attributes have null values, and cannot be utilized, many other are of the wrong type and need conversion, and most of all many variables are not useful to predict car accident severity, since they provide descriptive value like *SDOTCOLNUM, X, Y, LOCATION, INCDTTM, INCDATE, REPORTNO, COLDETKEY, INCKEY* and *OBJECTID*.

The correlation between features has been evaluated.

Furthermore also some of the attributes which are suitable as independent variables cannot be utilized as they miss too many values for example

INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND PEDROWNOTGRNT, SDOTCOLNUM and SPEEDING; These variables have been excluded.

- Preprocessing:

Some values of type object, in order to be trqansformed in numerical and be suitable for machine learing algorithm, have been converted from categorical value to numerical values using Label Encoding,replacing the categorical value with a numeric value between 0 and the number of classes -1.

The dataset now suitable for machine learning is:

```
In [103]: maindf.head()
```

Out[103]:

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
0	2	Overcast	Wet	Daylight	4	8	5
1	1	Raining	Wet	Dark - Street Lights On	6	8	2
2	1	Overcast	Dry	Daylight	4	0	5
3	1	Clear	Dry	Daylight	1	0	5
4	2	Raining	Wet	Daylight	6	8	5

The last problem to solve is that the feature SEVERITYCODE has 136485 records with value 1 and only 58188 with value 2 which is a problem since most machine learning algorithms work best when the number of samples in each class are about equal, since they are designed to maximize accuracy and reduce error.

The chosen approach to achieve balancing is to use Down-sampling which involves randomly removing observations from the majority class to prevent its signal from biasing the learning algorithm.

- **Choosing different models:**

With the dataset now ready for processing, different machine learning model have been applied:

· K-Nearest Neighbors

· Decision Trees

· Support Vector Machine

· Logistic Regression

However it is worth mentioning that the predicted values had much smaller range than the actual values, and as a result, the prediction errors were larger as the actual values moved from zero.

The result had large errors for the prediction.

Results and Discussion:

Unfortunately I was only able to apply the inferential statistical testing performed using f1 score and log loss as evaluation metrics.

The model with the highest performance is the Logistic regression

Avg F1-score: 0.54
Avg F1-score: 0.54
Avg F1-score: 0.54
LogLoss: : 0.68
Avg F1-score: 0.5131

Conclusion:

Based on historical data from weather conditions the more severe type 2 accidents happen in the same light, road, weather condition of type 1 severity accidents.

This showed that most vehicle accidents occur during normal conditions with normal driving circumstances.

This means it will be harder for the Seattle transportation department to mitigate accidents. However, as most accidents only involve property damage or minor injuries, there is not a serious problem that needs to be dealt with right away.

This study suggests that the severity of accidents are marginally influenced by the weather, light condition and road condition but happen because of other factors.