1. INTRODUCTION | Business Understanding

Seattle Department of Transporation (SDOT) develops, maintains, and operates a transportation system that enhances the quality of life, environment and economy of Seattle, as well as making sure people get around the city safely. Moving goods in, out of, and around Seattle is key to the regional economy and everyday life.

1.1 Problem Statement

Weather condition, visibility, roads condition, and other reasons are major factors in road accidents in the city, resulting in the following direct (or indirect) consequences:

a) fatalities
b) property damage
c) traffic delays
d) other indirect consequences (not covered under the current scope of this project.)

The scope of this project is limited to the dataset that contains car accidents that occurred in the city of Seattle, Washington, US between the years Jan 2004 and May, 2020

1.2 Audience/Stakeholders

The following are the major stakeholders that are directly/indirectly impacted:

- Drivers of vehicles that are involved in the accident
- Other passengers who are traveling in the vehicle that was involved in the accident
- Insurance companies (life and non-life)
- Seattle Police Department (SPD)
- Accident Traffic Records Department
- Other parties (owners of public/private property)
- People living in Seattle

These stakeholders will have to pay close attention to the problem described above. Therefore, in the larger interest of all the stakeholders, it is highly recommended to use a predictive collision model.

1.3 Objective

The objective of this project is to **build a robust predictive collision model** that will help to **predict severity of accidents** and to **reduce the frequency of car collisions** in the city of Seattle. The model should predict the severity of an accident given the conditions: weather, road, and visibility, or a combination of these conditions. Based on the prediction, the stakeholders and/or designated authorities can take appropriate measures/actions to mitigate the impending dangers or risks.

2. DATA

2.1 Data understanding

- **Data Source:** Historical data has been collected from the Seattle Police Department (SPD) and Accident Traffic Records Department, from the year 2004 till May 2020. The dataset is available on IBM cloud as a CSV file.

Atrributes:

- In total, there are 37 attributes (columns or features or independent variables), and not all attributes are useful. We decide to keep some and drop some.
- Some attributes have missing data.
- There are numerical and categorical types of data.
- The following is a list of attributes or features taken from the raw dataset. We need to do feature-engineering in order to improve the predictability of our model.

Attribute	Data type, length	Description
LOCATION	Text	Description of the general location of the collision
SEVERITY CODE	Text	A code that corresponds to the severity of the collision 0 – unknown 1-prop damage 2-injury 2b-serious injury 3-fatality
SEVERITY DESC	Text	A detailed description of the severity of the collision
COLLISON TYPE	Text	Type of collision *(entering at an angle; opposite direction etc*)
PERSON COUNT	Double	Total number of people involved in the collision
WEATHER	Text	Description of weather conditions at the time of collision *(overcast; raining; clear)*
ROAD COND	Text	Condition of the road during collision *(wet; dry; unknown)*
LIGHT COND	Text	Description of the general location of the collision *(daylight; dark: street lights on; dark: no street lights)*
VEH COUNT	Double	Number of vehicles involved in the collision *(0, 1, 2, 3, 4)*

2.2 Data Preparation

- To be able to build a good model you need a rich dataset, and our **dataset is rich**, as it contains many observations *(about 20,000 rows)* and various attributes *(37 columns).*
- This is a supervised machine learning model, therefore we need labeled data to train and validate the model. The **first column** in the dataset is the **labeled data (*dependent variable* y)**.

Label/Target column:

- Label/target variable for the dataset is **severity** *(dependent variable y)*, which describes the **fatality of an accident**. This variable is in the first column of the dataset.
- The remaining columns *(independent variables **X**)* have different types of attributes. Some or all can be used to train the model. I see that most of the observations are good to train and test the machine learning model.
- The data has **unbalanced labels**, therefore we should **balance the data. O**therwise, we will create a biased machine learning model.

To get the data set ready for analysis, the following tasks will be carried out as a part of data preparation.

Data cleaning – extract and convert: this activity does one or more of the following to make it suitable for data analysis and model building.

1. Replacing undesirable strings dropping unwanted features (columns).
2. Changing column data types to appropriate types using encoding.
3. Handle the missing data.