Part 2: Description of the Data

Car Accident Severity Analysis

(Data Science Coursera Capstone Project)

Prepared By:

Rajkumar Bathani

1. Understanding Data

1.1 Data Sources

The dataset used for this project is about car accidents which have happened within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding the severity of car accidents, the severity of each car accidents along with the time and conditions under which each accident occurred.

1.2 Data Cleaning

The dataset has many null values for several attributes. The target variable column 'Severity Code' has values categorized into 1 and 2, that needs to be transformed into 0 and 1 where 0 indicates 'Property damage only' and 1 represents 'Injury Collision', in order to avoid ambiguities while model fitting and predictions. Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables Inattention, Speeding and Under the influence. For lighting condition, Light was given 0 along with Medium as 1 and Dark as 2. For Road Condition, Dry was represented by 0, and 1 represent Mushy and Wet was assigned to 2. If talking about Weather Condition, 0 represents Clear, 1 indicates Overcast, Windy is 2 and Snow was given 3. 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to an extreme level of accident severity. Whereas, there were unique data which is not preferred to remove them such as above-mentioned attributes with value like 'Unknown' and 'Other'.

1.3 Feature Selection

The Severity Code is set as target variable and with respect to target variable other 6 features are selected such as; INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, and SPEEDING. The below table describes how each feature contributes to the overall severity of accidents.

Table.1 Description of how each feature affects the severity of accident

Feature Variables	Description of Features
INATTENTIONIND	Whether or not the driver was inattentive (Yes/No)
UNDERINFL	Whether or not the driver was under the influence (Yes/No)
WEATHER	Weather condition during the time of accident (Rainy/Overcast/Snowy/Clear)
ROADCOND	Road condition during the time of accident (Wet/Dry)
LIGHTCOND	Light condition during accident (Lights On/Dark with light on)
SPEEDING	Was the car above the speed limit at the time of accident (Yes/No)

2. Exploratory Data Analysis

Exploratory Data Analysis is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task, as it helps in analyzing data sets to summarize their main characteristics, often the visual methods.

The below image shows that that target variable 'SEVERITYCODE' is not equally distributed, as it can be said that it is highly imbalanced distribution of attribute. The values such as '0' represents 'property damage' and '1' signifies 'physical injury'. It is very important to have a balanced dataset, otherwise, the machine learning algorithm will produce biased results. Hence, SMOTE was used to balance the target variable in equally distributed sets.

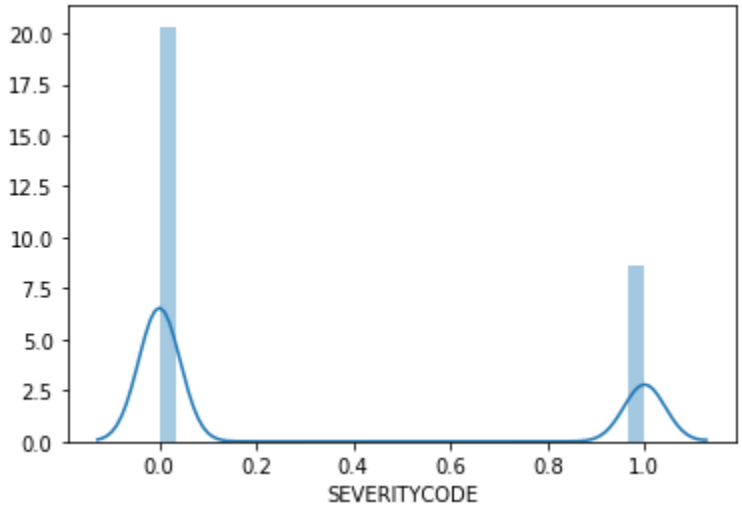

*Figure.**Error! No text of specified style in document.**..1 Distribution of Target Variable*

The point where the accident happened also contribute to the severity of an accident. There are three different values registered for the address type that shows the point where the accident happened such as; Block, Intersection and Alley. The counts for all the three different types of categories are as follows 126926, 65070, and 751 respectively.

There is a need to encode all the selected features into a format which is interpretable by ML algorithm. For example, encoding weather conditions into numeric range of 0 to 3, where '0' represents the best weather condition like 'Clear', and '1' is for Overcast and Cloudy weather. '2' is for Windy and '3' is for Raining and Snowy conditions. All these encoding procedures should be applied to the selected features.

According to the below image, the independents variable which has contributed the most is adverse lighting condition and the variables such as under the influence and overspendings were the least contributors to the severity of an accident. This is how all features can be used to determine that which independent variable plays a vital road in causing the accidents.

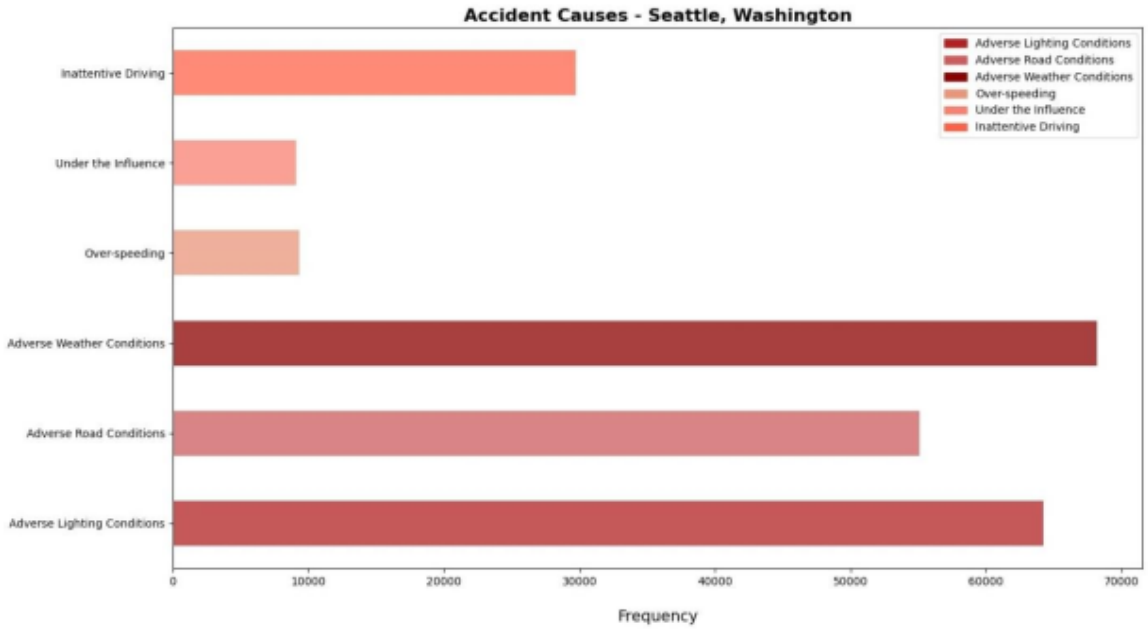

Figure.2 Conditions that contribute to the severity of an accident

Even considering the real life scenarios, it is proven that factors like bad light conditions, worst road conditions and adverse weather conditions influence the most.