CAR ACCIDENT SEVERITY REPORT : SEATTLE,WASHINGTON

Submitted by: Abdul Saboor

OCTOBER 1, 2020

IBM CAPSTONE PROJECT

1) Introduction

1.1) Background

Seattle is the largest city in the state of Washington , and is a hub to large two tech giants Microsoft and Amazon. Seattle accounts for nearly 3.4 million population, car accidents has become a major issue a lot these days due to increased car population

Nearly almost 1.25 million people die in road crashes each year. Car accidents are one of the leading causes of death. It took a toll of 518 billion USD on US government. According to Seattle Times, the city's goal is to achieve zero fatalities and serious injuries by 2030.

1.2) Problem

The project aim is to reduce number of accidents by analyzing data that might contribute to the likelihood of potential car accidents. The factors which leads to car accidents can vary a lot , It includes people who are driving very fast due to effect of alcohol, other reasons include weather visibility or road conditions.

1.3) Stakeholders

This will be of huge interest to SDOT(Seattle Department of Transportation) who responsible for the maintenance of the city's transportation systems. Others interested could be car insurance companies , local government of Seattle ,so they can all play important role in decreasing no of accidents in Seattle

2) Data

2.1) Data Source

The data has been provided by SPD(Seattle Police Department) and recorded by Traffic Records Department.The data set has total observations(rows) of 194,673.The main purpose of this report is to predict the accident severity in Seattle, hence the severity code is as follows:

SEVERITY CODE	DESCRIPTION
3	Fatality
2b	Serious Injury
2	Injury
1	Prop Damage
0	Unknown

As the data contains null values and non-relevant columns it is important to clean the data.

2.2) Data Cleaning

As we can see there is a huge imbalance of feature selection 'SEVERITY CODE' which might give us inaccurate results. There is huge difference between first and second row as you can see

```
In [13]: df['SEVERITYCODE'].value_counts()

Out[13]: 1    136485
         2     58188
         Name: SEVERITYCODE, dtype: int64
```

Hence it is important to resample so we can have equal data to work on

```
In [22]: df_firstrow=df[df.SEVERITYCODE==1]
         df_secondrow=df[df.SEVERITYCODE==2]

         df_secondrow_sampling=resample(df_firstrow,replace=False,n_samples=58188,random_state=101)

         df_balanced=pd.concat([df_secondrow_sampling,df_secondrow])
         df_balanced.SEVERITYCODE.value_counts()

Out[22]: 2    58188
         1    58188
         Name: SEVERITYCODE, dtype: int64
```