Car Accident Severity Analysis Capstone Project

1. Introduction - Business Understanding

1.1 Business problem –

Today world has changed and has higher usage of automobiles for moving from point to point. The movement of vehicles/ automobiles is influenced by many factors like local traffic, weather, inclement conditions, locality, terrain, roads etc. During movement of vehicles different accidents happen which can be avoided if certain precautions are taken by the person who is driving the automobile.

Through this project we would help the agency and the community to avoid car accidents by understanding the different causes for car accidents and provide solutions to overcome accidents and in turn reduce casualties and expenses incurred on people and vehicles. This will result in agencies taking necessary measures, early warnings, road signs, check posts, deployment of emergency vehicles, road barriers etc. to avoid car accidents.

We will be using machine learning algorithms of data science to identify the parameters which cause accidents and also provide solutions on the measurements to be taken to avoid these accidents

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.
In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations. Besides the aforementioned reasons, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.
The target audience of the project is local Seattle government, police, rescue groups, and last but not least, car insurance institutes. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

2. Data -

2.1 Data description –

The data is obtained from the Seattle department of transportation (SDOT) which identifies different parameters when there have been accidents. This can be easily obtained from the SDOT website in CSV files. The data presented here is from February 2006 till May

2020. The data represents various parameters like the street, date, severity, severity, place of accident and several others which are cause of car accidents

2.2 Source of the data – Seattle Department of Transport database

2.3 Data interpretation –

The current data consists of **38 columns** and **194673 rows** of data. The different columns which have been captured are provided below along with data types and descriptions. Descriptions are provided in the attachment – Metadata.pdf

SEVERITYCODE
X
Y
OBJECTID
INCKEY
COLDETKEY
REPORTNO
STATUS
ADDRTYPE
INTKEY
LOCATION
EXCEPTRSNCODE
EXCEPTRSNDESC
SEVERITYCODE
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 dependent variable for our study is **"SEVERITYCODE"** and has measurements for severity of an accident on a **scale of 0 to 3**. It contains several numbers which are as follows –

Severity Codes are as follows –

0: unknown

1: Property Damage

2: Injury

2b: serious Injury

3: fatality

Further we will have to see if the data consists of any NaN and null values which may not be contributing to the solution analysis. We will have to do some data cleansing. There may be certain records which may not be necessary for our problem which can be filtered so that we can obtain more accurate results but while filtering we should be careful not to lose out any data as studied at early stages of our course. Also, the data records of **194673** have at present hold only values of '1' or '2' under the column **"SEVERITYCODE".** The data is obtained in **CSV format** and can be easily uploaded through the "read_csv" function of python

2.3 Data Pre-processing –

The dataset in its original form is not completely fit for data analysis. There are many columns which may not be relevant and may need to be dropped or discarded. To prepare the data, first, we need to drop the non-relevant columns. Also we notice that most of the data types are of type object which need to be converted into numerical data types

The following will be columns for our data analysis and Machine Learning models!

SEVERITYCODE X Y OBJECTID INCKEY COLDETKEY
REPORTNO STATUS ADDRTYPE INTKEY LOCATION
EXCEPTRSNCODE EXCEPTRSNDESC SEVERITYCODE 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

we will check the data types of these new columns through the Python DataFrame and move ahead with the data analysis further.

2.4 Balancing the Dataset –

When we keenly observe we see that the target variable **SEVERITYCODE** only **43% balanced** (out of a total of 194673 datapoints we have code1 = 58188 and code 2 = 136485, hence code1/code2; 58188/136485 42.6%). This can skew our data points and provide wrong results. Hence, we will have to first obtain a balanced data set which can done through simple statistical techniques like downsampling. This will be done for **code1** and **code2**(please refer Metadata.pdf for classification)

After we have used statistical technique, we have now obtained a balanced dataset.