Vinit Deswal
Coursera_Capstone
17 August 2020

DATA DESCRIPTION AND PROCEDURE

ROAD ACCIDENT SEVERITY

Background:- We need a dataset which has a large combination of the data related to a particular place which can be used to create a best suitable model and predict the severity of accident by using the required data. The dimensions of the dataset should be large and should have a high number of entries for better accuracy of model.

Data Description:- The data we will be using for the project is from Seattle, Wasington, US named as "Data-Collisions.csv" provided by -" SDOT GIS Analyst" It has stored data from the year 2004-Present. It is a large dataset with dimension 193673 x 38 to work on. It has a special column showing the **severity of the collision** which can be used for training and predicting the model. The severity codes are shown in this <------table.

```
df.shape
(194673, 38)
```

code	Severity
0	unknown
1	serious damage
2	injury
2b	serious injury
3	fatality

Some of key attributes that we will be using and many more are here------->

Location	Road condition
Weather condition	Junction junction
Car Speeding	number of people involved
Light conditions	number of vehicles involved in

It consists of all the attributes related to whether like rain, fog, snow , road related like sand or gravel ,road condition, light condition, for the best prediction of the model, further it has the date and time of each collison. The count of cars, pedestrians and pedestrian cyclists. Driver condition

data is also available like drunk or not, speeding or not , was attentive or not. And finally it has where the accident occurred and how it occurred.

HOW WE WILL USE THIS DATA FOR OUR PROBLEM?

- This dataset consists of 193673 entries making it suitable for better accuracy of the model. So firstly we will import the data file to our notebook using pandas library function ‘pd.read_scv(“file name”)’

```
df= pd.read_csv('Data-Collisions.csv')
```

- Later we will analyse and process the data to select best attributes to be selected for the making the machine learning model

```
df_pred = df[['OBJECTID','ADDRTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','INCDTTM','SDOT_COLCODE','INATTENTIONIND',
             'UNDERINFL','WEATHER','ROADCOND','LIGHTCOND','PEDROWNOTGRNT','SPEEDING','ST_COLCODE']].copy(deep=True)
```

- Then we will process the data and remove the empty/null/NaN values and relace them with suitable data like mode or 0

```
df_pred.isnull().sum
```

OBJECTID	0
ADDRTYPE	0
PERSONCOUNT	0
PEDCOUNT	0
PEDCYLCOUNT	0
VEHCOUNT	0
INCDTTM	0
SDOT_COLCODE	0
INATTENTIONIND	0
UNDERINFL	0
WEATHER	0
ROADCOND	0
LIGHTCOND	0
PEDROWNOTGRNT	0
SPEEDING	0
ST_COLCODE	0

- Later with more processing data, modifying, normalising and One hot coding we will get best data for our model

- Create the model and check it’s score.
- If we will be satisfied with the score we will accept the model for further use otherwise send it back for more training

This is how we will use the data to predict the severity of a accident