Bishal Kumar Sahoo

Seattle Road Collisions Severity Predictor

IBM Data Science Capstone

INTRODUCTION:

Background

Since the invention of the automobile road and highways became one of the largest life takers as mortality rates are higher than ever in the past 10 years. Each year around 20-30 Million people get into a road accident in which around 10% of those lose their lives. Road accidents are a serious shame for our society and still we are not in a state to reduce it. Most of the innocent casualties are of the pedestrians , cyclists and the bikers and between the age of 20-35 years, the future of any country and the solo earners of a family. Naturally, a question to be asked is what is the cause of such a high percentage of car accidents.

Problem Description

We have to gather old accident record and its severity for a place with other informations like *location of accident, number of people involved, number of pedestrians, number of vehicles time and date of accident, way of accident, road condition, lighting and whether* at place of accident and create a machine learning model with this data so that later if we pass the following details , the model can predict the severity to us.

Interest

- Predicting the data will help us to get cautious in advance and then it will be the person's choice to whether they should proceed on the same road with precaution or do a detour to a safer road.
- In this project we will try to predict Injury Collision and Property Damage Only Collision, this report are aim to stakeholders interested preventing and reducing injury collisions and Property Damage Only Collision.

DATA:

Data requirement

According to the problem description 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.

The quality we need in our data are -

- It should have a large amount of data for better model training
- The data should have a column which shows the severity level of accident.
- Other necessary traits are:
  - Condition of road, weather, light at place of accident
  - Driver behaviour and consciousness
  - Detailed description of collision with date and time and location Number of person and vehicles involved.

DATA Description

The data we will be using for the project is from Seattle, Wasington, US named as "DataCollisions.csv" which is publicly available at this link: https://s3.us.cloud-objectstorage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv. 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.

Some Data attribute description,

Attribute	Data type, Length	Description
OBJECTID	ObjectID	ESRI unique identifier
ADDRTYPE	Text, 12	Collision address type: Alley, Block, Intersection
LOCATION	Text, 255	Description of the general location of the collision
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision
PEDCYLCOUNT	Double	The number of bicycles involved in the collision.
VEHCOUNT	Double	The number of vehicles involved in the collision.
INCDTTM	Text, 30	The date and time of the incident
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.

ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision.

- **Import Basic Libraries**

```
# firstly, let's import the basic data
libraries import pandas as pd import numpy as
np
```

- **Load Data From CSV File** `df= pd.read_csv('Data-Collisions.csv')`
- After importing the data from the source, taking a first look at first 5 rows of the data

METHODOLOGY:

In this project we will direct our efforts on creating a Machine Learning model for the prediction of the severity of an accident only by getting some description data of the situation and later plan our future plan.

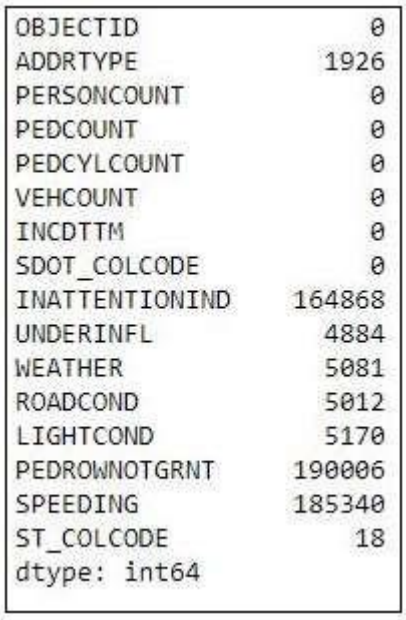
- Firstly, now we will separate between necessary data need for model creating and training and the non essential data for the model.
- Then, We will remove all the NaN(not a number) values and replace them with appropriate solutions.
- Later, take care with the non accurate data types and get them back like with the date case(it's datatype is object)
- After that we will convert the categorical data with numerical data. Simple ones with just replacing commands and the difficult ones with one hot encoding method.
- Then, try to find some pattern by splitting the datetime into daysofweek , month and hours. directly or by plotting a graph
- When satisfied with the data selected we will split it into train and test data
- Train our decision tree model with training and predict the test data and calculate the accuracy of our model.
- If the F1-score is satisfactory, we will train our final model with whole data, as every data is important

Data Analysis

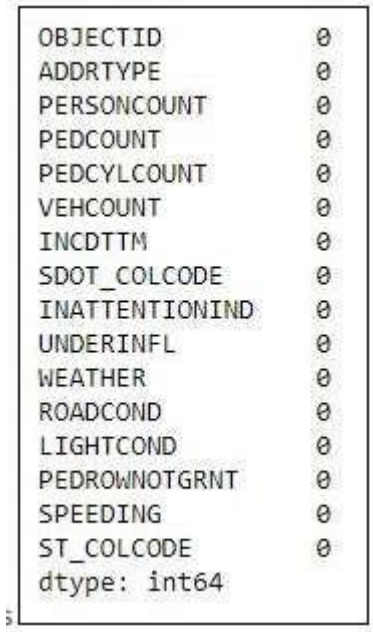
Firstly, we created a temporary copy of the original data with the necessary attributes that we thought would be necessary just by looking at their data type and the data. Later they will be processed and more will be added or removed by taking into consideration of our needs and which seems to be best suitable for our model.

Removing null / NaN values –

NaN values could be a real mess to a classification model, so it is mandatory to remove them ASAP. There are many ways to do this like deleting rows, replacing them with the mean of the whole column, filling them with the nearest neighbour, or in classification problems filling them with most abundant data. So , we used filling the most of the columns with the mode of that column and the columns which were having binary values like yes('Y') and no('N'), they were only filled the yes values, so the NaN were replaced with 0 (later the Ys were converted to 1 too in converting categorical data to numerical). So to achieve a completely filled dataframe.



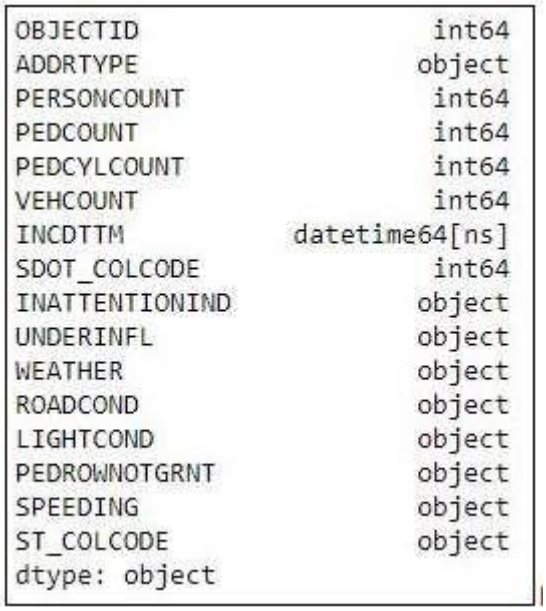
Before removing NaN Values



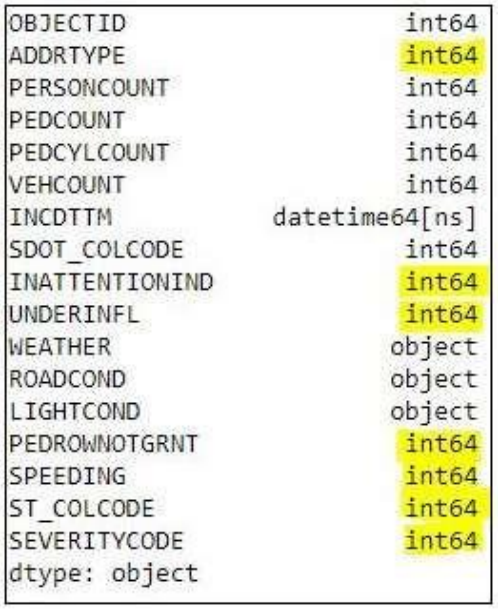
After removing NaN Values

The next step was to convert categorical data to numerical data for ease of computer -

A computer best understands the numeric data, it's not like a human that understands words, so to make the work fast and easy for the computer we convert the categorical data into numeric data. In pandas, the string or text data type is written as object and numeric as int or float.



OBJECTID	int64
ADDRTYPE	object
PERSONCOUNT	int64
PEDCOUNT	int64
PEDCYLCOUNT	int64
VEHCOUNT	int64
INCDTTM	datetime64[ns]
SDOT_COLCODE	int64
INATTENTIONIND	object
UNDERINFL	object
WEATHER	object
ROADCOND	object
LIGHTCOND	object
PEDROWNOTGRNT	object
SPEEDING	object
ST_COLCODE	object
dtype: object	



OBJECTID	int64
ADDRTYPE	int64
PERSONCOUNT	int64
PEDCOUNT	int64
PEDCYLCOUNT	int64
VEHCOUNT	int64
INCDTTM	datetime64[ns]
SDOT_COLCODE	int64
INATTENTIONIND	int64
UNDERINFL	int64
WEATHER	object
ROADCOND	object
LIGHTCOND	object
PEDROWNOTGRNT	int64
SPEEDING	int64
ST_COLCODE	int64
SEVERITYCODE	int64
dtype: object	

Before Conversion

After Conversion

All data can't be easily converted to numeric , some need further methods like one hot encoding, so the 3 object columns viz, WEATHER, ROADCOND, LIGHTCOND, will be converted through one how encoding where its each categorical data will be converted to new columns and where is yes then 1 and no to 0.

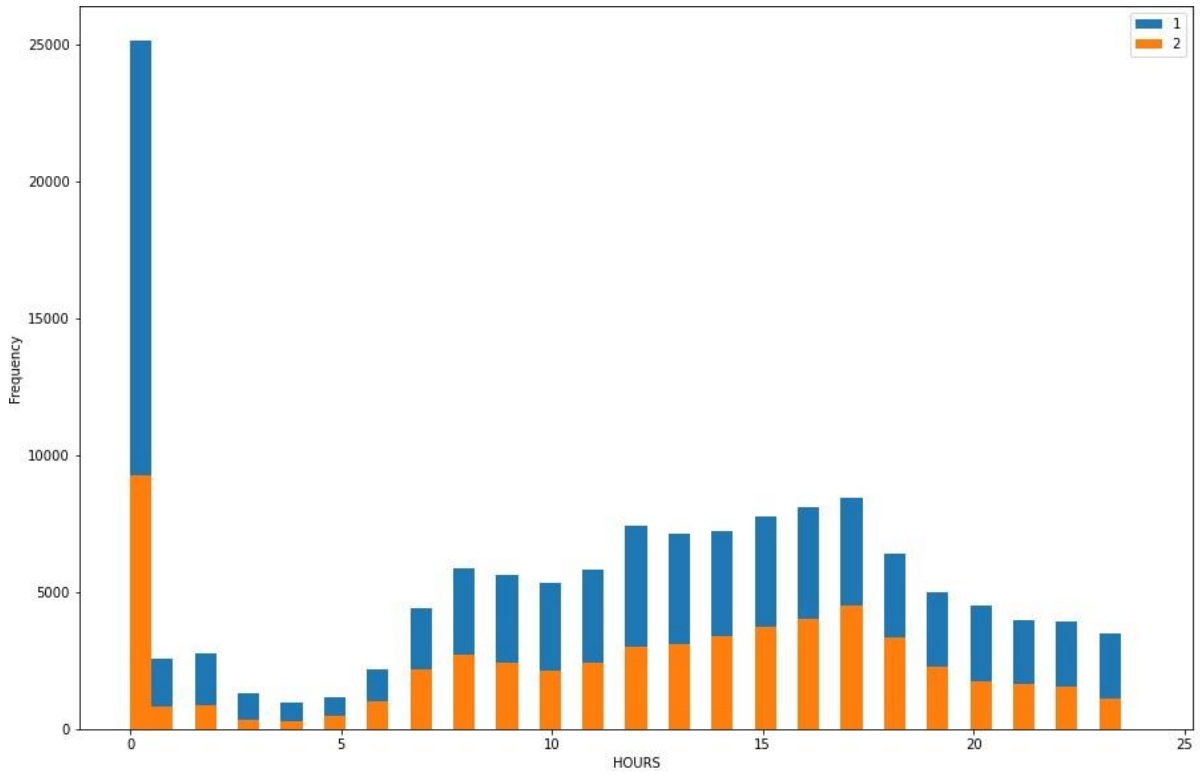
DATETIME data management -

Firstly the date time data was stored in object format, so it was converted to datetime format.

This was done because, after doing this a lot of dt function opens and we can split this data into various parts to get a better view to aur data.

So for further assessment, the datetime data is splitted and new columns were added for months days of week and hours to see a pattern if possible.

With plotting it was found the days and months have an average rate , no peaks and slopes, so they are not good to be considered. While plotting the hours it was found that



There was a peak at 12am. So the data was split in ones and zeros, where ones being the time 12am and from 9am to 6 pm, rest were turned zeros as they don't have much info.

```
HOUR_GP  SEVERITYCODE
0        1                36182
         2                14270
1        1               100303
         2                43918
Name: SEVERITYCODE, dtype: int64
```

Now again a final feature selection took place to select only necessary data -

With one hot encoding the data for WEATHER , ROADCOND, AND LIGHTCOND was splitted into dummies.

Wet	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	D
1	0	0	0	0	0	1	
1	0	0	1	0	0	0	
0	0	0	0	0	0	1	
0	0	0	0	0	0	1	
1	0	0	0	0	0	1	

Just a section of table to show what one hot encoding did. The data are split into

- X → containing all the data that we will provide to the model for prediction.
- Y → containing the data that our model will predict.

X data have a variety of data, so to make it normal, the data is normalised which pushes the mean of data to zero and variance to 1. Post normalised data →

```
array([[ 1.36398593, -0.33020207, -0.18743029, -0.16958841, 0.12553783,
       -0.88333125, -0.41751024, -0.42518348, -0.2217116 , -0.1567239 ,
       -0.22440165, 0.59145941, -0.01696304, -1.21707436, -0.05414257,
       2.45445634, -0.00506801, -0.45298634, -0.011333 , -0.02409974,
       -0.06841713, -1.4099744 , -0.07905204, -0.01813462, -0.01963186,
       -0.07200071, -0.02431221, 1.76085874, -0.08920831, -0.07872239,  -
       0.576075 , -0.00751719, -0.1141037 , 0.77768637, -0.17682024]])
```

Model creation

Decision tree data machine learning classifier was used to create a prediction model for the given data.

Why Decision Tree?

Because the decision tree is fast and processes a well accurate model. As our data had a dimension of 194673 x 35, training with any other machine learning was not proceeding and the notebook plus kernel kept crashing. So our model should be fast with accuracy to the decision tree provided in that environment.

Train Test split –

The data are primarily split into train and test data.

Why splitting the data?

To get the best accuracy, the model should be tested on unknown data, so as our model had no biasing while predicting. So, the model is trained with train data and the prediction is done with test data which is also used for evaluating the model.

Dimension of Train set: X(155738, 35) , y(155738,)

Dimension of Test set: X (38935, 35) , y(38935,)

The data is fitted and the model was trained and get two types of data one is Training Dataset and other is Test Dataset.

Model evaluation

F1 score was used for the evaluation of model and the F1-score of the model we got is **0.8306** (Test Dataset).

So our model is selected. As our model was trained only with training data so test data is as it is, now we will train our model with whole data.

After training again the data was predicted from the model only and the F1 score we got is **0.8636** (Training Dataset).

Hence our model is ready to predict any data.

RESULT:

As from above methodology and data analysis, we say that our data performed very well with known as well as the unknown data .

- The model shows **0.8306** F1 score on test dataset
- The model shows **0.8638** F1-score on training dataset

So, this model can be used to predict the accident severity just by giving the necessary data and the model will predict the severity for them. A lot of possible casualties and loss of property and life can be minimized by taking precaution measures.

DISCUSSION:

This data was well collected and stored . The data was more than enough to predict the severity. So anyone living in Seattle or someone who is planning to visit can formally use this model to predict the accident severity of that place. They can plan their journey accordingly and mark the safe and unsafe areas. Even more attributes could be used to predict the data but they were not that precise. And even selecting more data will make our model too biased for our data plus start predicting the noise which is just waste. While analyzing the data we saw that the months and days of week had not much difference to the collisions , they were just consistent and that's why they were dropped as we need a pattern to predict the right model. Adding average data will just put load on our machine.

We used **decision tree** in this prediction because the decision tree is fast and processes a well accurate model. As our data had a dimension of **194673 x 35**, training with any other machine learning was not proceeding and the notebook plus kernel kept crashing. So our model should be fast with accuracy to the decision tree provided in that environment.

CONCLUSION:

Purpose of this model was to create a model that can predict the accident severity just by knowing the basic component of that place. Our analysis clearly shows that this data was excellent data which helped our data to achieve a final f1 score of 0.8636. So anyone living in Seattle or someone who is planning to visit can formally use this model to predict the accident severity of that place. They can plan their journey accordingly and mark the safe and unsafe areas further planning depends on them. This model not only will help the general public, but other concerned authorities could also see this and find the areas that are red zones at particular and save precious lives . they can even put the barricades to alert the upcoming traffic. Other concerned authorities who keep the condition of road and light could mark the places which have problems and solve them.

Final decision on journey planning will be made by public based on specific characteristics of that location and past history of Collison.