

# COLLISIONS DATA ANALYSIS

*IBM DATA SCIENCE CAPSTONE PROJECT*

E Kim

August 18, 2020



# INTRODUCTION

## Background

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 yrs, the future of any country and the solo earners of a family.

## 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.

# 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 the accident.
- Other necessary traits are -
  - Condition of road, weather, light at place of accident
  - Driver's behaviour and consciousness
  - Detailed description of collision with date and time and location
  - Number of people and vehicles involved.

# Project objective

Creating a prediction system using the old data from a particular place that can predict the severity of a road collision with maximum accuracy.

Several thin, white, parallel lines of varying lengths and slight curves are positioned in the lower right quadrant of the slide, extending from the bottom right towards the center.

# DATA DESCRIPTION

The data we will be using for the project is from Seattle, Washington, 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.

| ROADCOND | LIGHTCOND               | PEDROWNOTGRNT | SDOTCOLNUM | SPEEDING | ST_COLCODE |
|----------|-------------------------|---------------|------------|----------|------------|
| Wet      | Daylight                | NaN           | NaN        | NaN      | 10         |
| Wet      | Dark - Street Lights On | NaN           | 6354039.0  | NaN      | 11         |
| Dry      | Daylight                | NaN           | 4323031.0  | NaN      | 32         |
| Dry      | Daylight                | NaN           | NaN        | NaN      | 23         |
| Wet      | Daylight                | NaN           | 4028032.0  | NaN      | 10         |
| Wet      | Daylight                | NaN           | 4058035.0  | NaN      | 10         |
| Dry      | Daylight                | NaN           | NaN        | NaN      | 32         |

# DATA IMPORTING AND CLEANING

- Firstly useful attributes are selected and transferred to a new dataframe.
- The null values were filled with the most abundant unit of the columns and some were filled with a zero(0) values.
- The categorical data in most of the column is converted to numeric data
- The data type of most of the columns were converted from object → int64 or the date time data was converted to date time format.
- The columns that cannot be converted to easy binary numeric data are passed through one hot encoding which splits them up each unique value to a new column with a binary values (0 or 1).
- Data is again filtered for best attributes and sent to features
- The data is normalised for a better model.

# TARGET

The column named 'SEVERITYCODE' is our target cell which we have to predict. It contains the severity of the collision in form of code. Each increment in the code means an increase in severity of collision.

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

# DATE-TIME DATA ANALYSIS

Date time or the timestamp is one of the best column in a dataset. As with that it can be splitted into various form the way we like it.

So, that's what we did and splitted the data into 3 new columns and tried to form a pattern

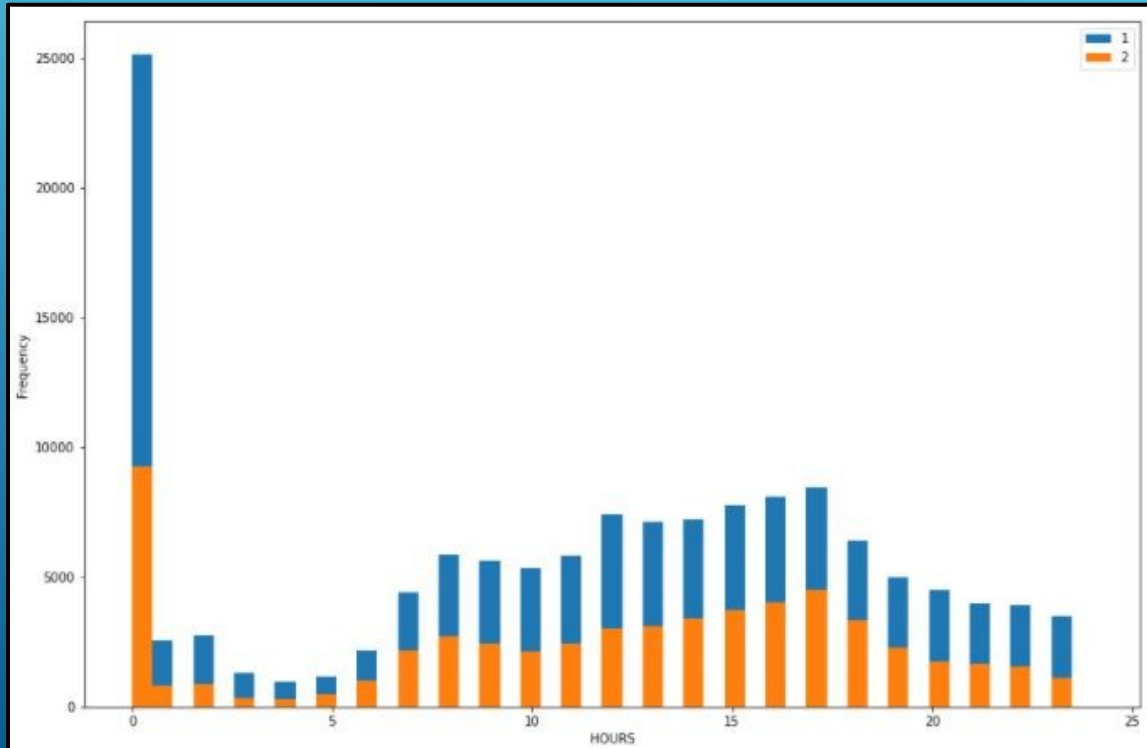
3 parts were,

- From date part we added 2 new columns
  - The day of week column which shows the each day of week with a number range 0-6 , monday-sunday respectively
  - The month column which shows the month of year with each number from 1-12.
- From Time part we added 1 new column
  - The hour column which shows he hour in 24hr format , 0-23.

Graphs were plotted to find a pattern.



# DATE-TIME DATA ANALYSIS



- Day of week and month did not show any pattern.
- The hour plot clearly showed the sudden peak in collision at 0000 hrs time.
- So the hour data converted to new column hour\_gp
- Hour\_gp had a binary output,
  - 1 for collision more than 5000
  - 0 for collision less than 5000

# MACHINE LEARNING AND MODEL EVALUATION

```
Train set: (155738, 35) (155738,)  
Test set: (38935, 35) (38935,)
```

The data was splitted into training and testing data to give out of sample testing to our model, it also prevents biasing.

A decision tree machine learning model was created as it is fast and accurate at the same time. Other models were slow and were crashing the kernel and hanging the machine.

```
f1s = f1_score(y_test,yhat)  
  
print("the f1 score of the dt model is --> ",f1s)  
  
the f1 score of the dt model is --> 0.8310329699996625
```

```
#then, create a model  
dt = DecisionTreeClassifier(criterion="entropy")  
  
#after that, fit the values  
dt.fit(x_train,y_train)  
  
DecisionTreeClassifier(criterion='entropy')
```

Our model has an out of sample f1 score of 0.831 which is acceptable for a good model.

# CONCLUSION AND FUTURE DIRECTIONS

- ▶ Finally the model is re trained with whole data so no data is left wasted.
  - ▶ Built useful model to predict the severity of the collision.
    - The model f1 scored 0.8636
    - But still there is a room for the improvement.
    - Could be increased with more different type of data
  - ▶ Otherwise the model will train on itself as the time goes on and it is used, it will have newdata.
- 