# Predict the Severity of a Traffic Accident

Monica

9 October 2020

## 1 Introduction

### 1.1 Background

Every year car accidents cause hundreds of thousands of deaths worldwide. According to a research conducted by the World Health Organization (WHO) there were 1.35 million road traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15–29 years. Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally and are predicted to become the seventh leading cause of death by 2030[1]. Leveraging the tools and all the information nowadays available, an extensive analysis to predict traffic accidents and its severity would make a difference to the death toll. Analysing a significant range of factors, including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accidents can be performed. Thus, trends that commonly lead to severe traffic incidents can help identifying the highly severe accidents. This kind of information could be used by emergency services, to send the exact required staff and equipment to the place of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this severe accident situation can be warned to nearby hospitals which can have all the equipment ready for a severe intervention in advance. Consequently, road safety should be a prior interest for governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety.

### 1.2 Problem

Data that might contribute to determining the likeliness of a potential accident occurring might include information on previous accidents such as road conditions, weather conditions, exact time and place of the accident, type of vehicles involved in the accident, information on the users involved in the accident and off course the severity of the accident. This projects aims to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services.

### 1.3 Interest

Governments should be highly interested in accurate predictions of the severity of an accident, to reduce the time of arrival and to make a more efficient use of the resources, and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.

# 2 Data

## 2.1 Data Source

The dataset is a part of an example dataset in IBM Data Science Professional Certificate. It can be downloaded from my GitHub repository https://github.com/monica110394/Coursera_Capstone. The name of the file is Data-Collisions.csv.

## 2.2 Variable Selection

After reading the data in pandas, analysis is performed to find the number of samples present in the dataset, also the number of attributes. The dataset has 194,673 records and 38 attributes including the target variable. The columns present in the dataset are:

```
'SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO,
'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
'EXCEPTRSNDESC', 'SEVERITYCODE.1', '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'
```

Going through the description of every attribute, which can also be found at my GitHub repository with the file name Metadata.pdf, intuitionally it can be seen that many of the attributes will not contribute to the prediction. The attributes that have been ignored for this prediction are:

```
'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO, 'STATUS', 'INTKEY',
'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE.1',
'SEVERITYDESC', 'INCDATE', 'INCDTTM', 'SDOT_COLDESC', 'PEDROWNOTGRNT',
'SDOTCOLNUM', 'SPEEDING', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY'
```

The attributes that are kept for the prediction are (17 variable including target variable):

```
'SEVERITYCODE', 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT',
'PEDCYLCOUNT', 'VEHCOUNT', 'JUNCTIONTYPE', 'SDOT_COLCODE',
'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
'SPEEDING', 'ST_COLCODE', 'HITPARKEDCAR'
```

Target variable: **'SEVERITYCODE'**

## 2.3 Feature Engineering

After knowing about the dataset, feature engineering is performed such as:

- Null Values are detected
  - **SEVERITYCODE** **0**
  - ADDRTYPE 1926
  - COLLISIONTYPE 4904
  - PERSONCOUNT 0
  - PEDCOUNT 0
  - PEDCYLCOUNT 0
  - VEHCOUNT 0
  - JUNCTIONTYPE 6329
  - SDOT_COLCODE 0
  - INATTENTIONIND 164868
  - UNDERINFL 4884

    - WEATHER 5081
    - ROADCOND 5012
    - LIGHTCOND 5170
    - SPEEDING 185340
    - ST_COLCODE 18
    - HITPARKEDCAR 0

- Attributes that have most of the values, more than 70% values, as null, are dropped
    - INATTENTIONIND 164868
    - SPEEDING 185340

- Current number of rows, columns and attributes
    - Number of rows: 194673
    - Number of Columns 15
    - **'SEVERITYCODE'**, 'ADDRTYPE', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'ST_COLCODE', 'HITPARKEDCAR'

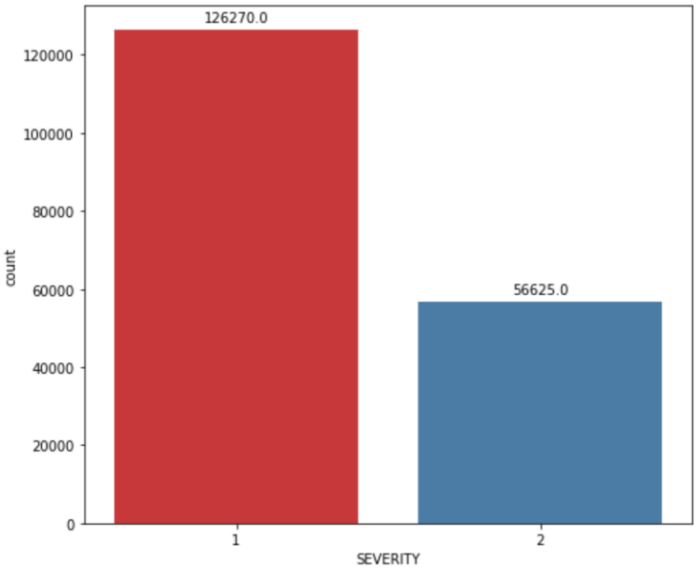
- Records or rows that have missing values are dropped as well.
    - Number of rows: 182895
    - Number of Columns 15

- Label Encoded all the categorical columns converting the 'string' labels into numbers and then changed their data types from 'integer' to 'category'.
    - 'ADDRTYPE'
    - 'COLLISIONTYPE'
    - 'JUNCTIONTYPE'
    - 'UNDERINFL'
    - 'WEATHER'
    - 'ROADCOND'
    - 'LIGHTCOND'
    - 'HITPARKEDCAR'

- Numeric columns
    - 'PERSONCOUNT'
    - 'PEDCOUNT'
    - 'PEDCYLCOUNT'
    - 'VEHCOUNT'

- Renaming columns to reasonable names

- On plotting a count plot of the target variable, **'SEVERITYCODE'**, it is found that the data set is highly unbalanced, which can cause the prediction to be skewed. Hence, the dataset is balanced by down sampling the category which as greater number of samples, in this case, when **'SEVERITYCODE'** is 1.




# 3 Exploratory Data Analysis