Applied Data Science - Capstone

# **Introduction/Business Problem**

The case study is predicting the severity of the accident. The Outcome of the case study is mainly targeted for Traffic departments, Travellers & Tourist, local residents and others those who use the roadway service for the day to day work.

This will would greatly help the Traffic to monitor the traffic due to accident and take necessary to action to control for hassle free transportation.

For the travellers & tourist, Local residents will be aware of severity and take necessary actions to be safe and avoid accidents.

# Data Section

The Data Set is the Collision details happened over the 2004 to present. It holds all the attributes to the

* Key – Unique Identification
* Location of Collision
* Address Type – Alley, Block, Intersection
* Collision type
* Weather, Road Conditions, Light Conditions
* Fatalities
* No of Persons injured,
* No of pedestrians injured

These attributes will guide in predicting the Target variable the Severity of the Accident.

As the Target variable provided It’s a supervised algorithm with Categorical type (1,2). Also the data set is highly imbalance with below details so we need to balance the dataset before training the models.

1. 136485
2. 58188

There are total of 38 attributes we need to understand the correlation with the target variable and identify the necessary attributes to train the model.

Data Cleaning

* Data Set Count : 194673
* Attributes 37
* Target Variable : SEVERITYCODE

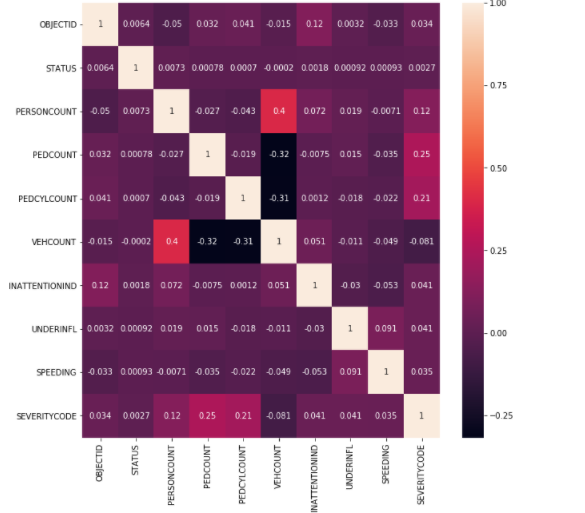
## Identification of the Null/Nan variables.

* The columns which were not important & having NAN have been dropped from dataset.
* Columns with 1 % of Nan compared to complete data set were dropped.
* UNDERINFL,
* JUNCTIONTYPE
* WEATHER
* ROADCOND
* LIGHTCOND
* COLLISIONTYPE
* UNDERINFL
* WEATHER
* ROADCOND
* LIGHTCOND

# Identification Important Features:

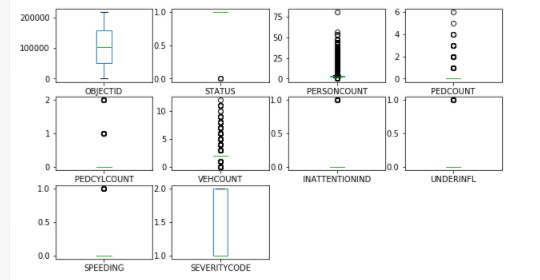
Based on the dataset understanding and correlations have identified the below attributes

['OBJECTID','STATUS','COLLISIONTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','INATTENTIONIND','WEATHER','UNDERINFL','ROADCOND','LIGHTCOND','SPEEDING','SEVERITYCODE']



## Outliers

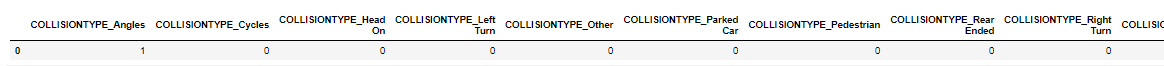
Identification of outliers are performed using the box plot on the selected attributes



# Encoding data

The categorical data need to be converted into numerical data for the model to understand and predict. We used One Hot Encoding on the below fields

* 'LIGHTCOND'
* 'ROADCOND'
* 'WEATHER'
* 'COLLISIONTYPE'



## Train Test Split

Splitting the data for train & test model by splitting 70:30.

# Balancing Data Set

Due to imbalance data set there is high chance the model performance to degrade. So we need to have balanced data in both the class for better model

Library used : IMBLearn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Before Balancing** | |  | **After Balancing** | |
| 1 | 84748 |  | 1 | 75960 |
| 2 | 37980 |  | 2 | 75960 |

# Models Used

* Multinominal NaïveBayes

HyperParameter – Alpha =1.0

* Gaussian NB
* Random Forest Classifier

HyperParameter – max\_depth=15, random\_state=0,n\_estimators=100,max\_features=20

# Results

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Confusion Matrix** | | **Precission** | **Recall** | **F1 Score** | **Accuracy** |
| GaussianNB | 34954 | 6919 | 66 | 68 | 68 | 67 |
| 12228 | 6448 |
| MultinomialNB | 24961 | 16812 | 75 | 66 | 67 | 66 |
| 3685 | 14991 |
| RandomForestClassifier | 21999 | 15261 | 73 | 71 | 71 | 71 |
| 6110 | 31461 |

# Solution

Model designed to automate the accident predictions based on the locality, weather ,road conditions and other attributes with 70 % accurate making it easy for the Traffic Departments & Residents.