IBM Data Science Professional Certification Capstone: *US Accident Severity*

Author: Avinash Wilson John Peter

Date: 14-Sep-2020

Table of Contents:

Introduction :

- The vehicles cause many problems globally. Car accidents are responsible for **1.35M deaths** on roadways every year.
- In this project we will try to construct an optimal model for predicting **The Severity of Road Accidents**. Specifically, the road accident on a given day (with given factors) in and around New York.
- The severity falls under 5 categories, where **0 indicates less severity** and **4 indicated more severity**.

Data :

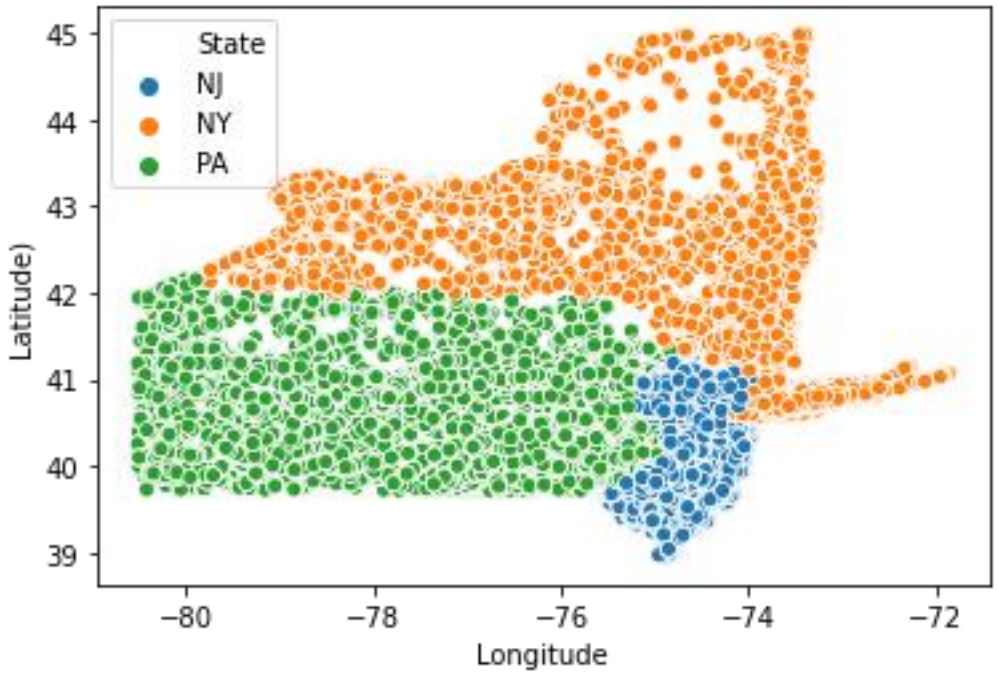
- All the accident data that are collected from February 2016 to June 2020 for the Contiguous **United States**.
- **3.5 million** accident records in this dataset.
- ***Source:***
  - KAGGLE: https://www.kaggle.com/sobhanmoosavi/us-accidents

Methodology :

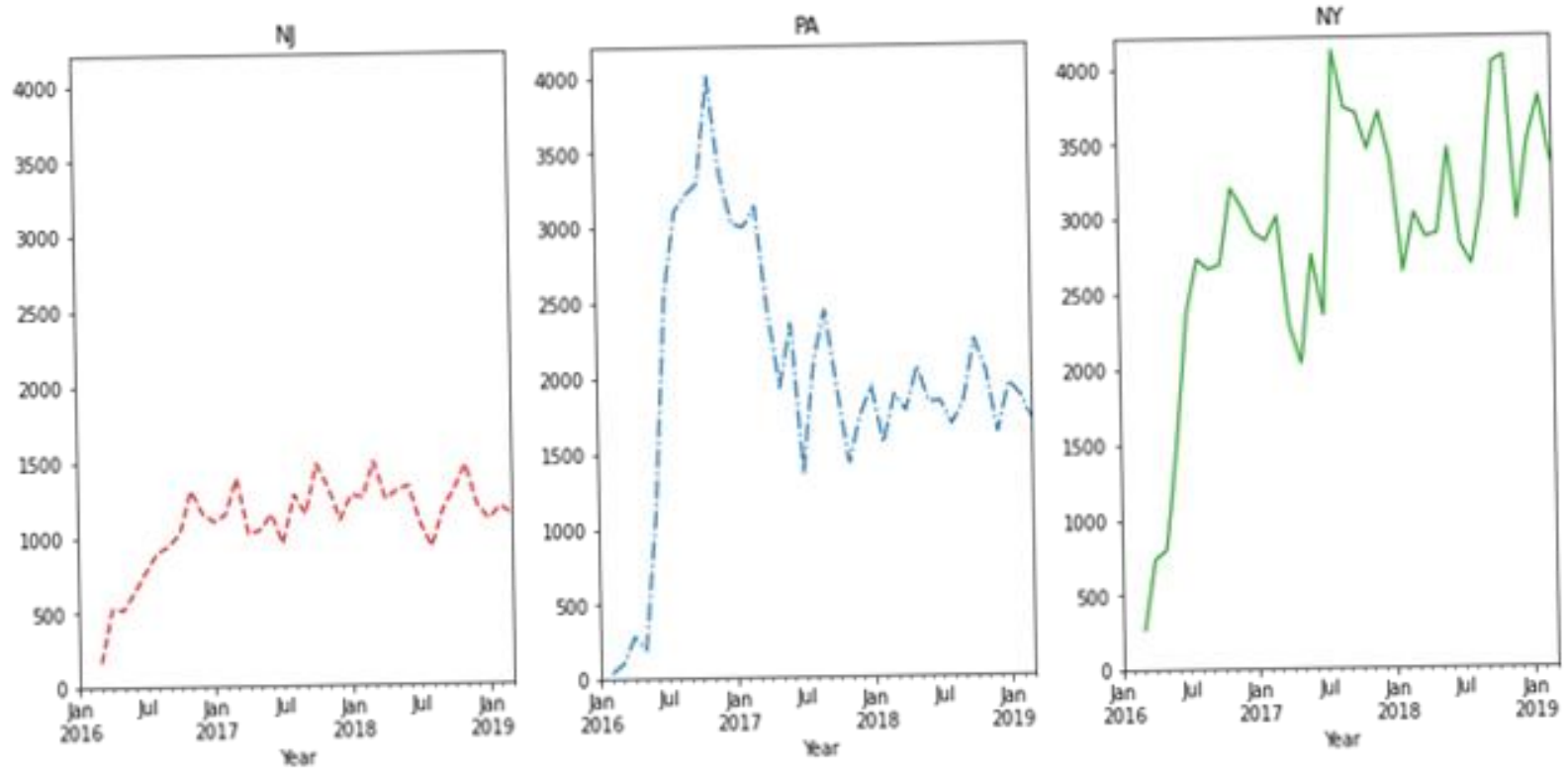
- **Github** as a repository
- **Jupyter Notebook - Python**
- Packages such as *Pandas, NumPy and Sklearn*.
- Once I have load data into Pandas Dataframe, used ***'info'*** attribute to check the **feature** names and their data types.
- Then I have *cleaned* the data and re-built the Attributes in a **Standardized Formats** to ease the usage.
- Then I have presented few **statistics** inferences coupled with **visuals** in the *Explanatory Data Analysis* Section.
- Because of my PC's **less computation capacity**, I have used only the records corresponding to the *New York* to proceed further.
- I have chosen the **Random Forest** machine learning model; I have also built and *evaluated* the model and shown the results with accuracy.

Explanatory Data Analysis

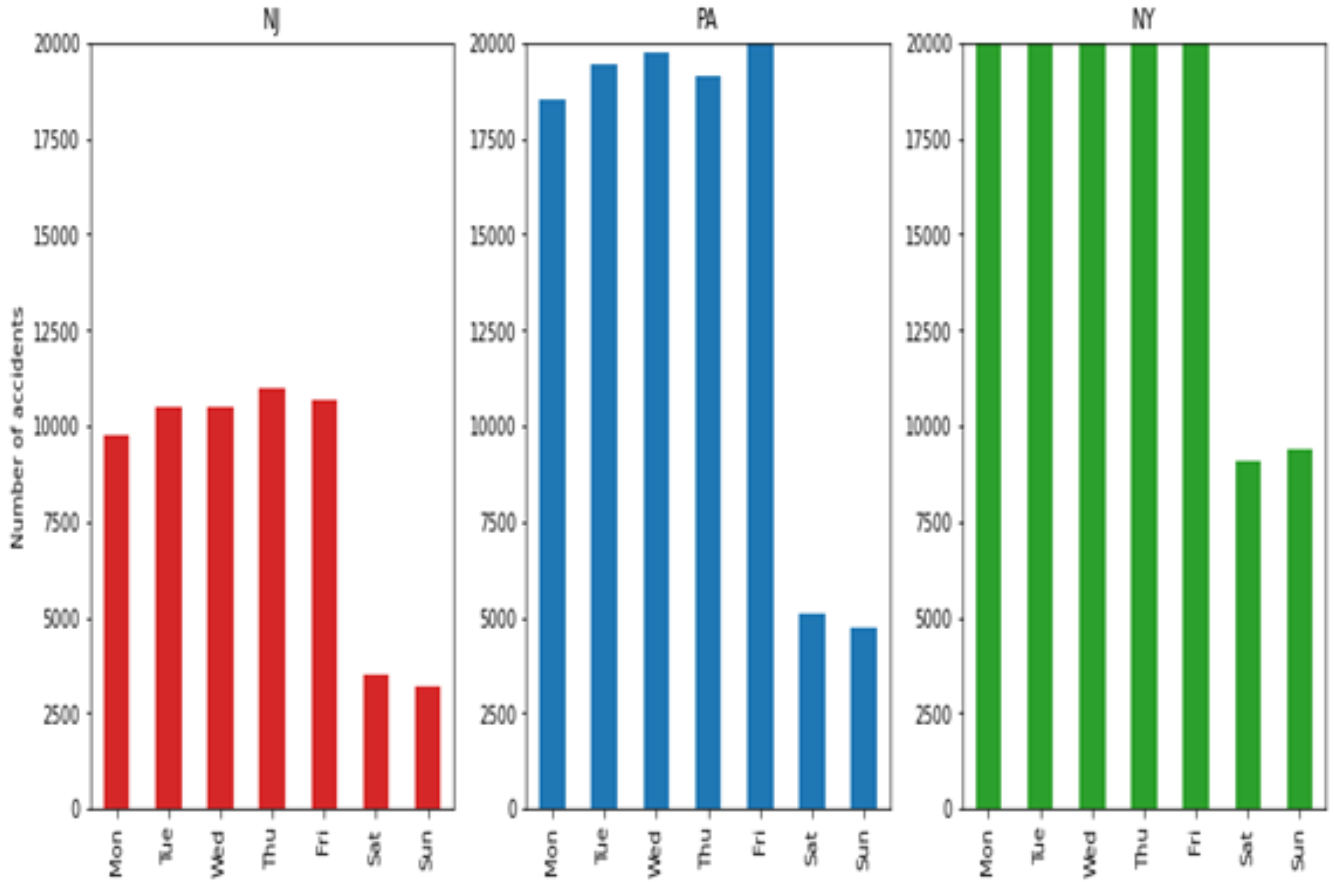
Plotting the States with The Dots referring to the Accident Occurring Spots among NY, NJ and PA :



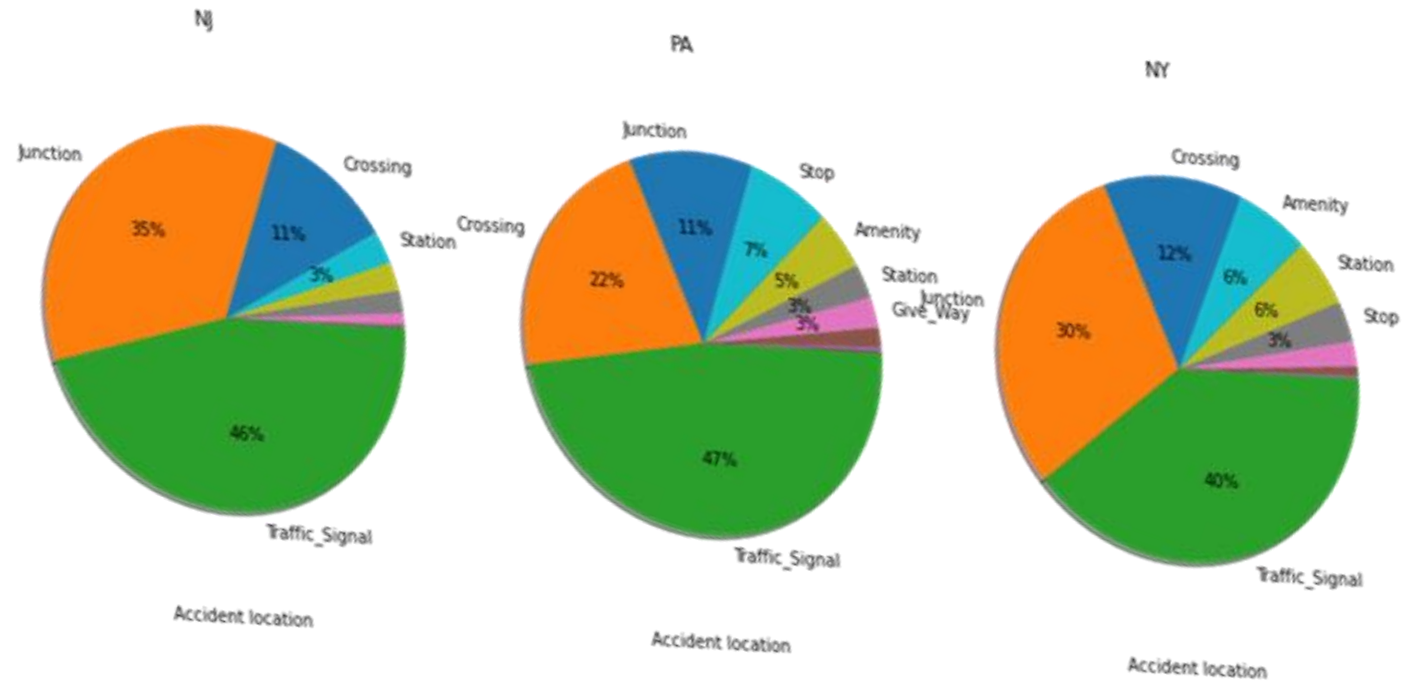
Plotting the States and Accident Occurring Frequency NY, NJ and PA :



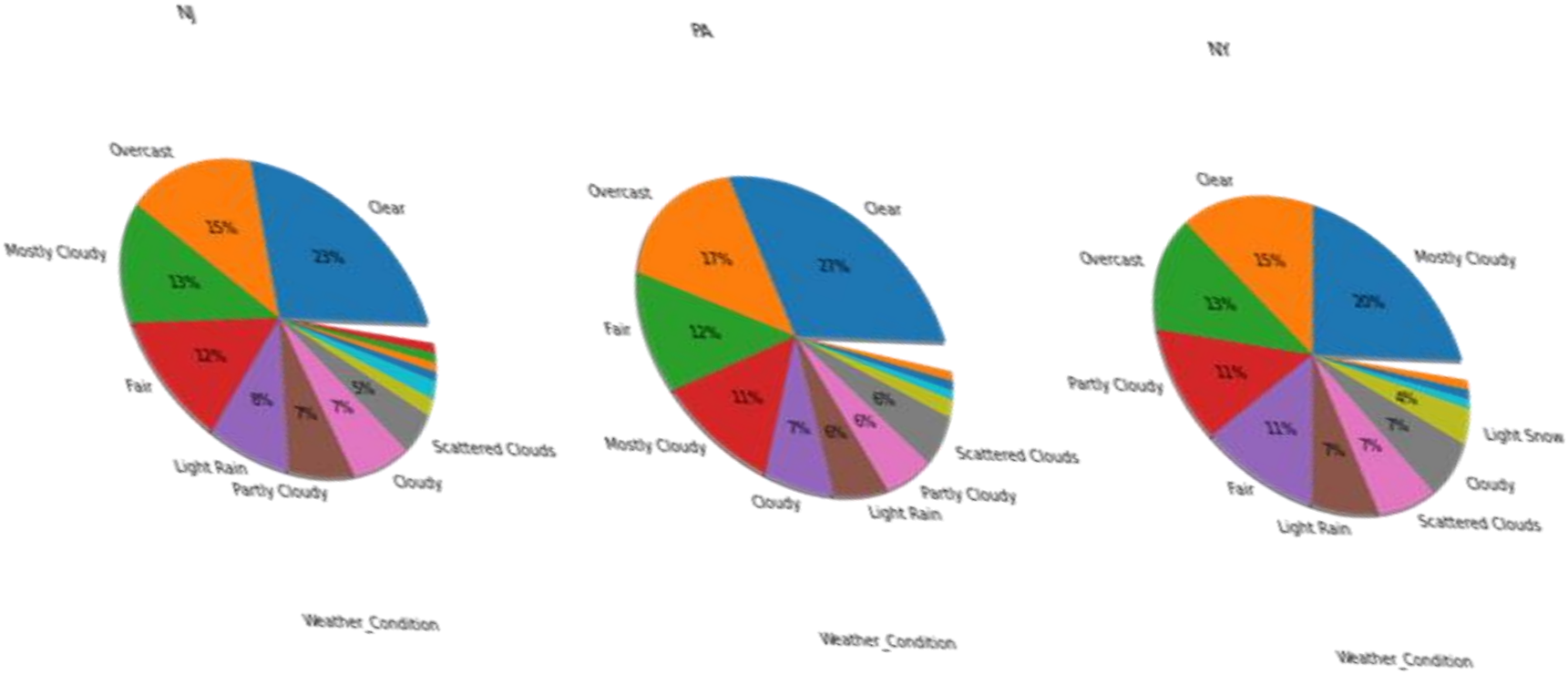
Plotting the depicting Frequency of Accident Occurrence with Days of Week:



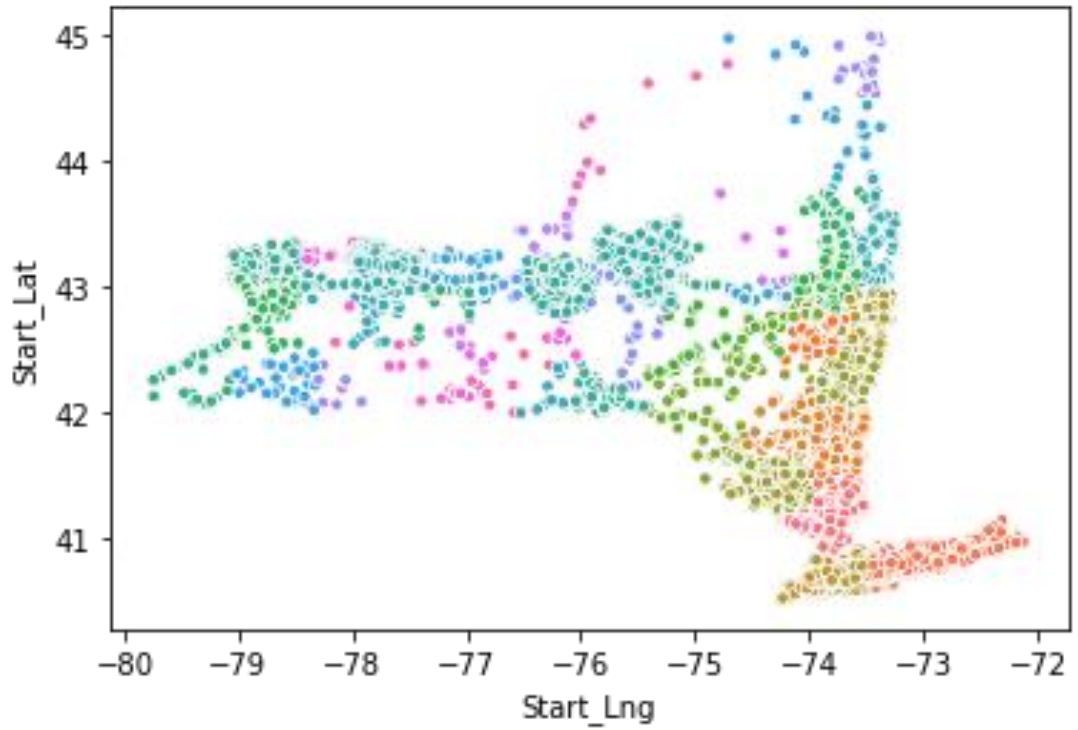
Plotting the depicting Ratio of Accident Occurrence to Type of Location:



Plotting the depicting ratio of Accident Occurrence with Weather:



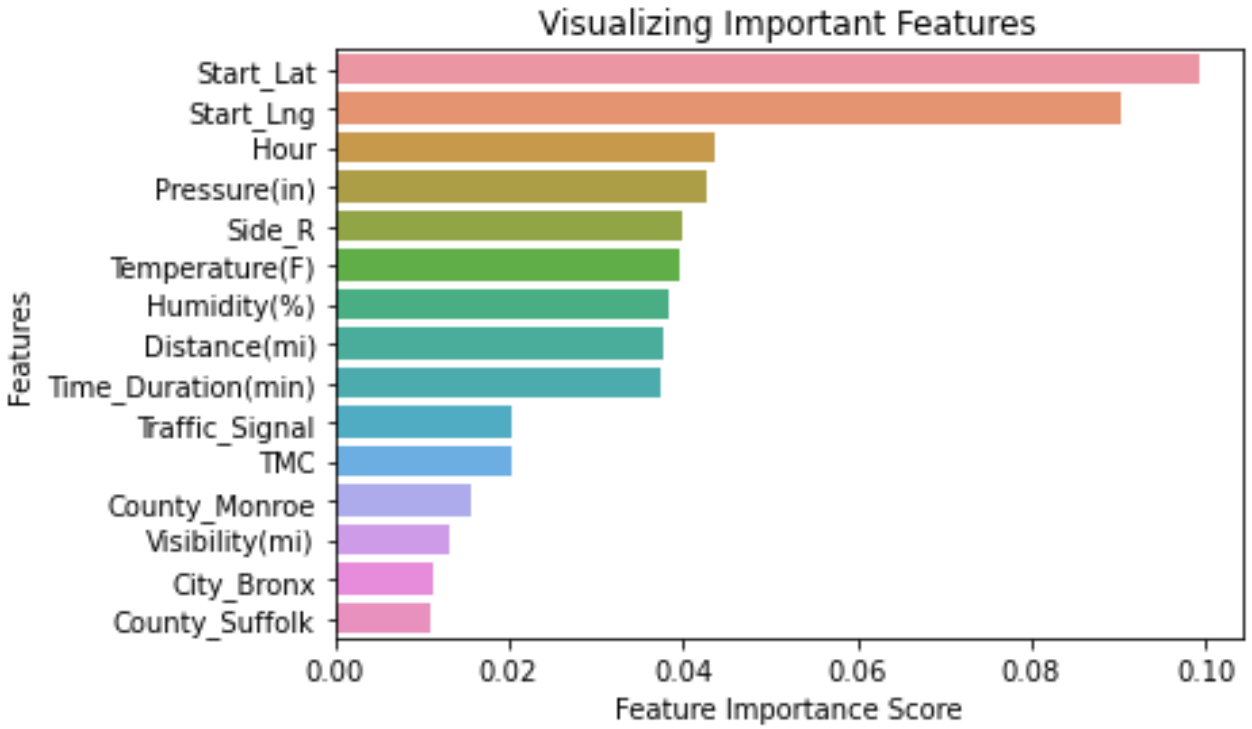
Plotting the depicting the Accident Occurrence within New York :



Random Forest:

- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to the training set.

Plotting the depicting the top Features that can be used to build the model:



Results :

Randon forest algorithm -- Limited feature
: accuracy_score: 0.884.

So, with the **Random Forest Classifier** above, we were able to derive **88.4%** of accuracy with the *test set*.

Discussion :

- The accuracy gained from this model is pretty much good on a scale upto 88.4 in the Test Accuracy.
- As a future work, similar models likes *Decision Trees*, *K-Nearest Neighbour*, *Multiclass Logistic Regression* etc can be built on the same dataset.
- Upon constructing those models, a final model using **Ensemble Learning** can be built to boost up the *accuracy* of the model.
- Also, with a **GPU** its efficient to build the model using *whole of the data-set*.

Conclusion :

- Based on the dataset chosen for this capstone from Time, Weather, and Temperature conditions pointing to certain classes, we can conclude that particular conditions has an impact on whether or not travel could result in less serious (class 0) or more severe accident(class 4), in and around New York.

THANK YOU