IBM Data Science Professional Certification Capstone: Coursera

Project: *US Accident Severity*

Author: Avinash Wilson John Peter
Date: 14-Sep-2020

Table of Contents:

1.Introduction:

Business Problem :

All around the world, roads are shared by many motorized vehicles that have made transportation faster and more comfortable while supporting the country's economic status and social development. However, these vehicles cause many problems globally. Car accidents are responsible for 1.35M deaths on roadways every year. Almost 3.7k people are killed globally in road traffic crashes, where more than half of those killed are pedestrians, motorcyclists, and cyclists etc.

In this project we will try to construct an optimal model for predicting **The Severity of Road Accidents**. Specifically, this report will be targeted to stakeholders interested in knowing the chances of them, encountering into a 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**.

Since the roadways is preferred by many as a mode of transportation, the vehicles are prone to accident especially when the weather is not favorable.

So here, we will use our data science skills to predict whether the person has chances to encounter a Severe Collision based on few criteria. Upon analyzing the data set and building a model to predict the severity, the people in New York who usually go by car might think of other alternatives.

2.Data

Description:

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset.

Content:

This dataset has been collected in real-time, using multiple Traffic APIs. Currently, it contains accident data that are collected from February 2016 to June 2020 for the Contiguous United States. Check here to learn more about this dataset.

Inspiration:

US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of

precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the impact of COVID-19 on traffic behavior and accidents.

Source:

KAGGLE: https://www.kaggle.com/sobhanmoosavi/us-accidents

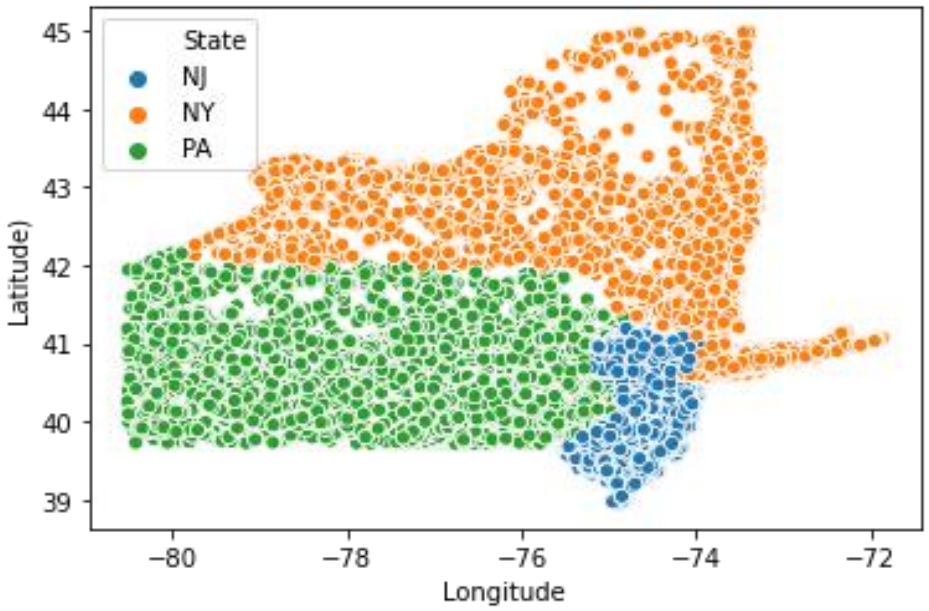
The given dataset has 3513617 number of records with 48 attributes. Here, Attributes like *Zipcode, ID, Source* etc. are not necessary while training the model. Also, we are not going to consider all the records to fit our model because of **memory concerns**. So, we will remove the unwanted attributes and NA records and work with a **filtered** dataset *(based on States)*.

3.Methodology

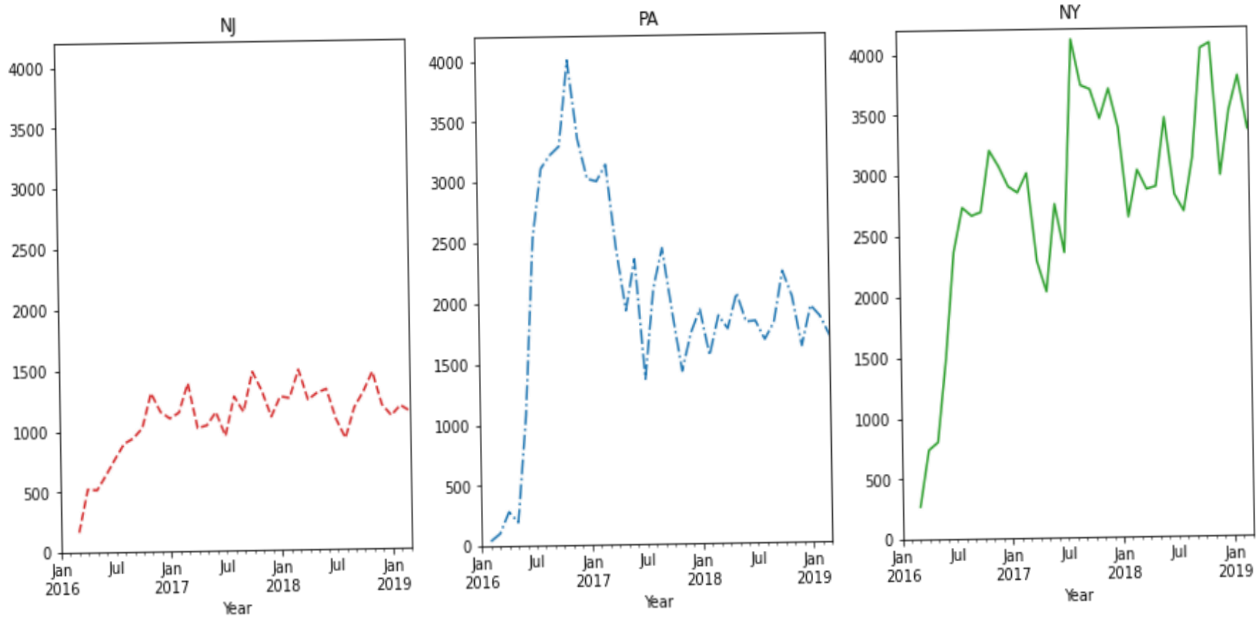
- For implementing the solution, I have used **Github** as a repository and running **Jupyter Notebook** to preprocess data and build Machine Learning models. Regarding coding, I have used **Python** and its popular packages such as *Pandas, NumPy and Sklearn*.
- Once I have load data into Pandas Dataframe, 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.

4.Explanatory Data Analysis

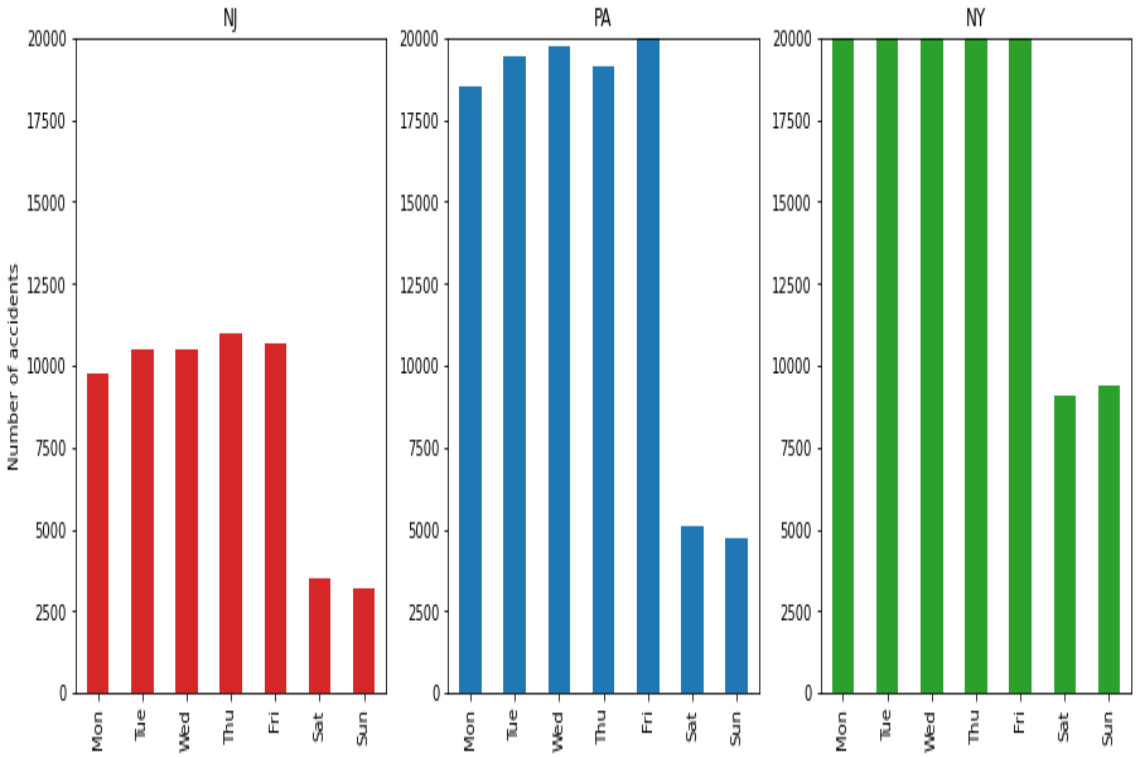
Plotting the States with The Dots referring to the Accident Occurring Spots among NY, NJ and PA :



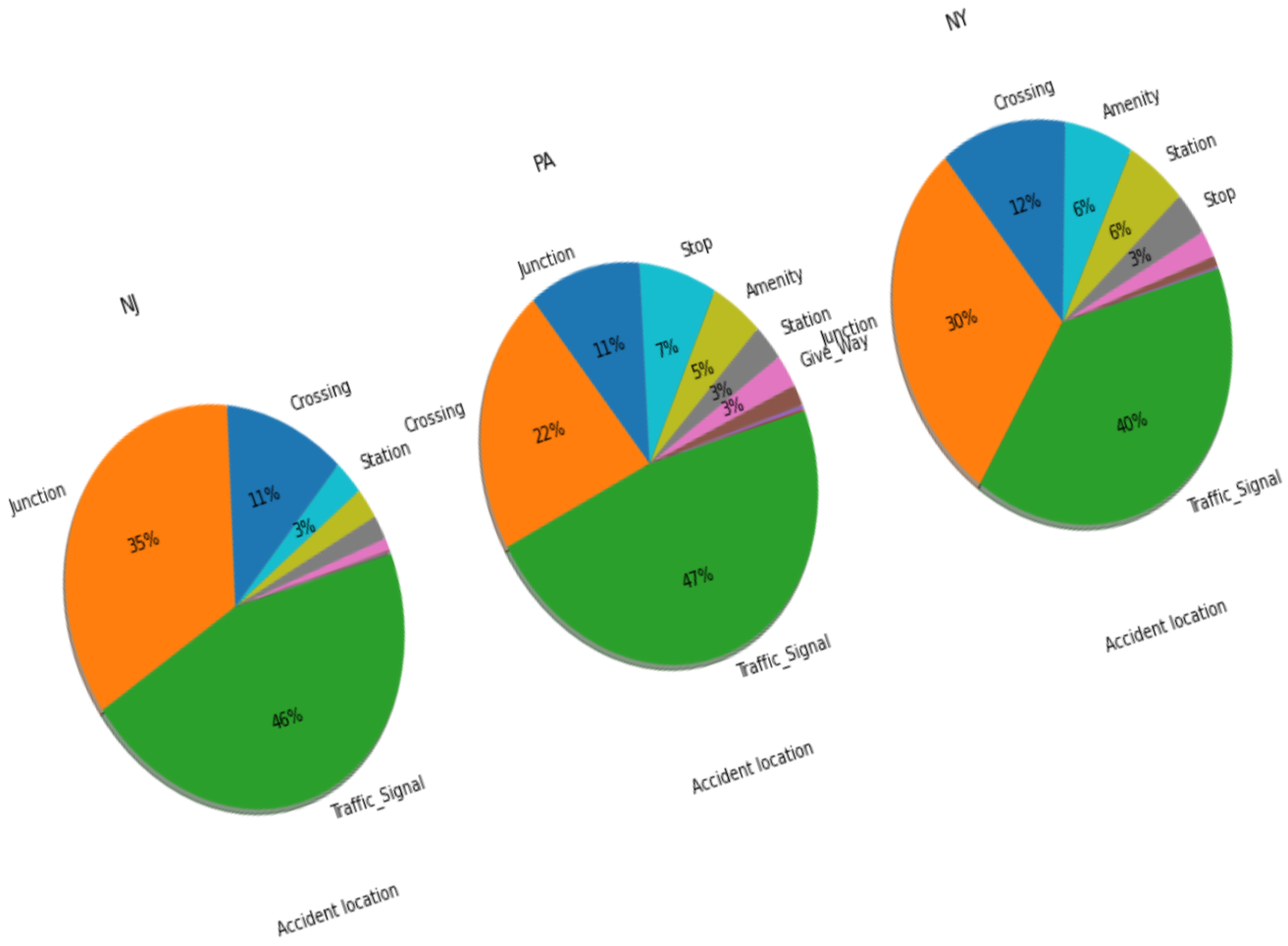
Plotting the States and Accident Occurring Frequency NY, NJ and PA :



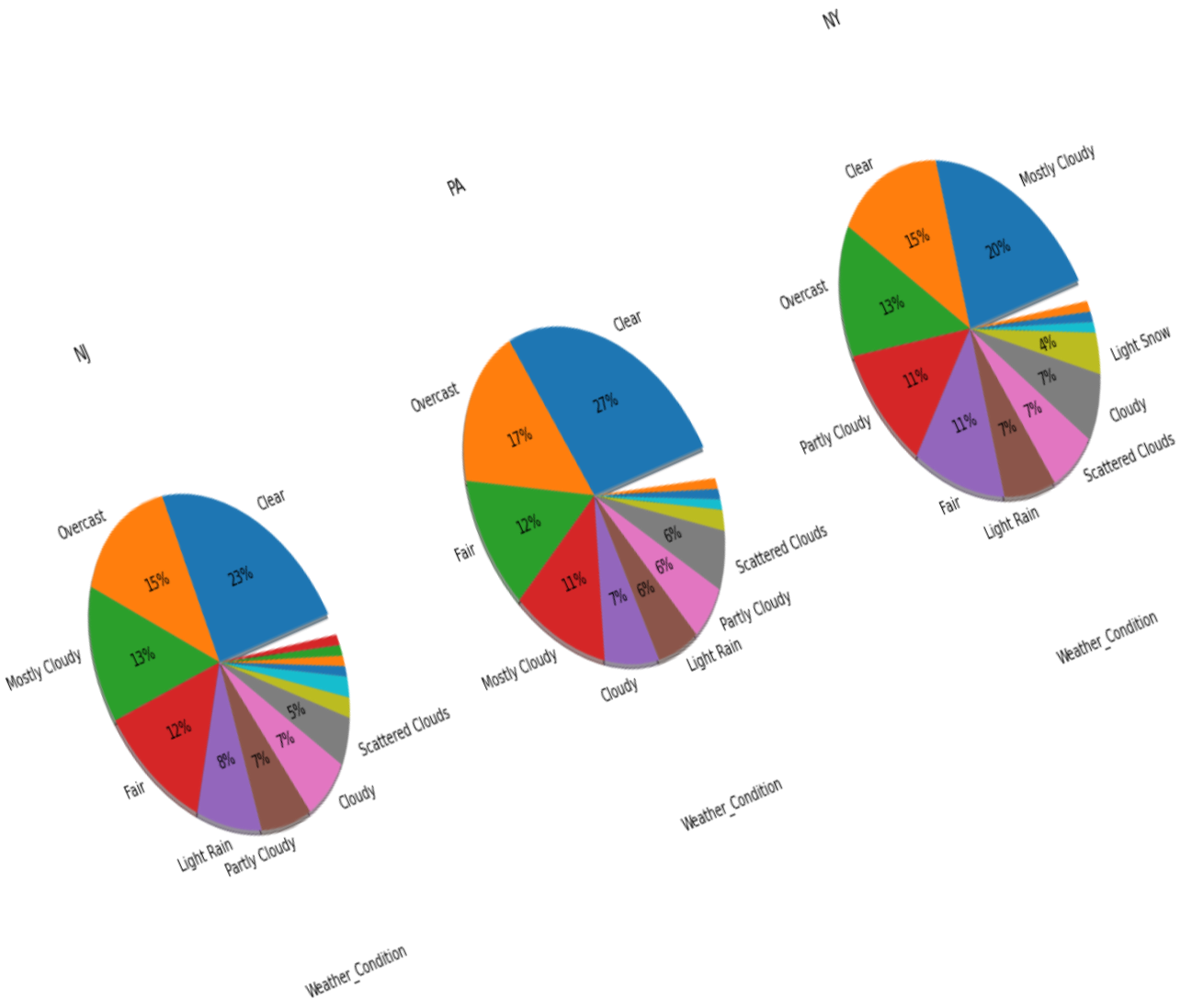
Plotting the depicting Frequency of Accident Occurrence with Days of Week:



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



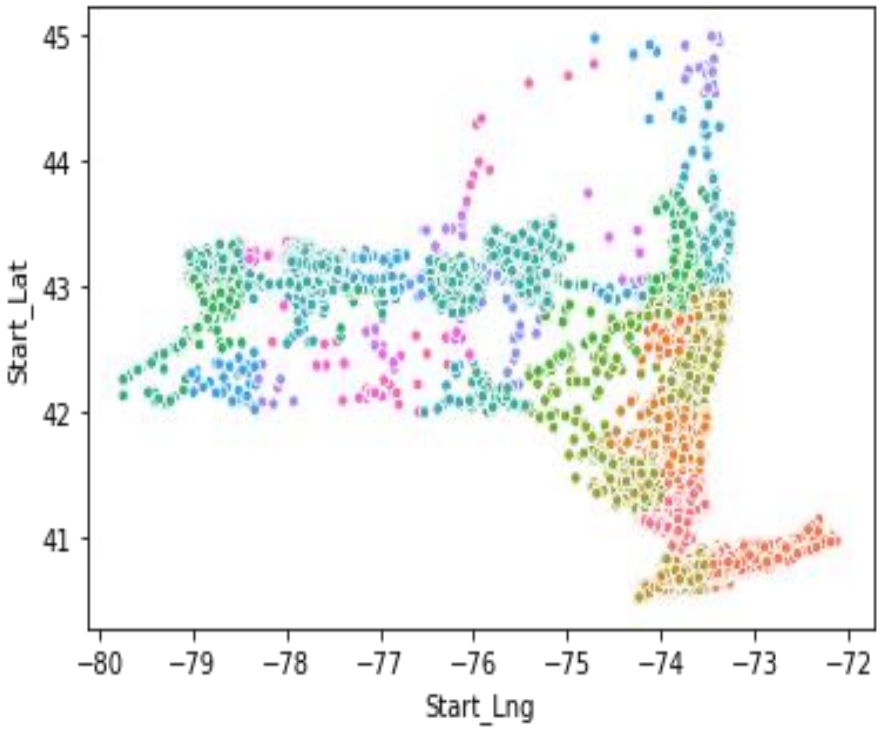
Plotting the depicting ratio of Accident Occurrence with Weather:



!!!**Since the data-set is too large and since I don't have a GPU, I am proceeding to analyze and model for a particular city i.e) New York**!!!

Considering Only New York :

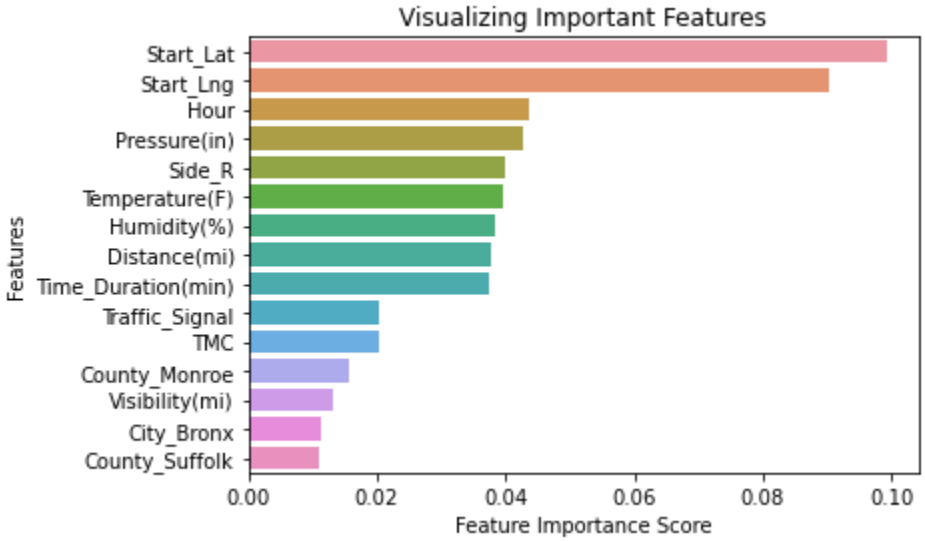
Plotting the depicting the Accident Occurrence within New York :



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:



5.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*.

6.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*.

7.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.