**IBM Data Science Capstone Project v1.0**

**USA Accident – Traffic prediction Analysis**

**Prepared By: Vijayaragavan Narendran**

**1. Introduction**

The aim of this project is predicting the traffic delay duration with given weather condition in United states.   This model will be useful for who are new comers on the state for safe journey. A model to predict accident severity is built using these predictors as input vectors. Once the model is validated using a Machine Learning algorithm approach, we can be confident in predicting accident severity. In this proactive approach, the results of the analysis would be useful to various Entities like the Police and Insurance Companies. The goal is to reduce the fatalities and economic losses from accidents.

Approximately 1.35 million people die each year as a result of road traffic crashes. The 2030 Agenda for Sustainable Development has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2020. Road traffic crashes cost most countries 3% of their gross domestic product. More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists. 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles. Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.

**1.1 Data Description**

USA accident data has taken from <https://www.kaggle.com/>. This dataset contains several features and indicates that severity based on more features. Aim of this project is predict the severity with applicable features. These will be helpful and reduce the traffic delays. Severity data has 4 different values. 1-Minor, 2-Medium, 3-High,4-Very High.

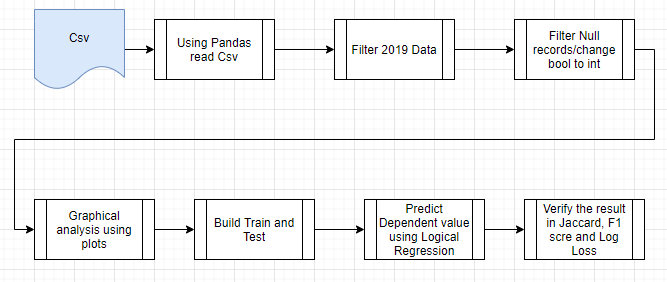
**2. Methodology**

Dataset has 1.2gb size, and not taken all the data for prediction. Have filtered only 2019 data for prediction. Most of the data are having null values. Those null records were eliminated for better prediction.

Some features ('Bump', 'Crossing', 'Junction', 'Roundabout', 'Stop') are having bool datatypes (True/False). True – data are replaced by 1 and False data are replaced by 0. Wind\_Direction has more similar duplicate values hence replaced the values and keep unique and clear values, example (East to E, West to W, North to N, South to S). Start\_date and End\_date columns are string and converted as DateTime. “Wind Direction” and “weather condition” are having string values. Using pd dummies, those columns are converted as new separate rows with features which values are either 1 or 0. Used Logical Regression for getting prediction value.

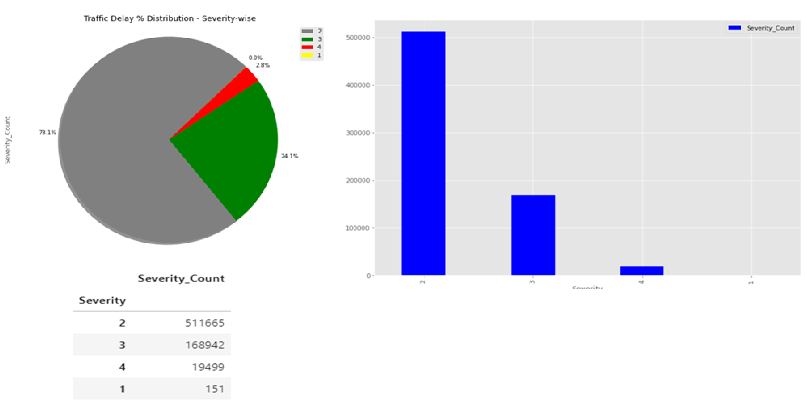
|  |  |  |
| --- | --- | --- |
| **Feature** | **Datatype** | **Required Feature for Prediction** |
| ID | object |  |
| Source | object |  |
| TMC | float64 |  |
| Severity | int64 | Y |
| Start\_Time | object | Y |
| End\_Time | object | Y |
| Start\_Lat | float64 | Y |
| Start\_Lng | float64 | Y |
| End\_Lat | float64 | Y |
| End\_Lng | float64 | Y |
| Distance(mi) | float64 |  |
| Description | object |  |
| Number | float64 |  |
| Street | object |  |
| Side | object |  |
| City | object | Y |
| County | object | y |
| State | object | Y |
| Zipcode | object |  |
| Country | object |  |
| Timezone | object |  |
| Airport\_Code | object |  |
| Weather\_Timestamp | object |  |
| Temperature(F) | float64 | Y |
| Wind\_Chill(F) | float64 | Y |
| Humidity(%) | float64 | Y |
| Pressure(in) | float64 | Y |
| Visibility(mi) | float64 | Y |
| Wind\_Direction | object | Y |
| Wind\_Speed(mph) | float64 | Y |
| Precipitation(in) | float64 |  |
| Weather\_Condition | object | Y |
| Amenity | bool |  |
| Bump | bool | Y |
| Crossing | bool | Y |
| Give\_Way | bool |  |
| Junction | bool | Y |
| No\_Exit | bool |  |
| Railway | bool | Y |
| Roundabout | bool | Y |
| Station | bool | Y |
| Stop | bool | Y |
| Traffic\_Calming | bool |  |
| Traffic\_Signal | bool |  |
| Turning\_Loop | bool |  |
| Sunrise\_Sunset | object |  |
| Civil\_Twilight | object |  |
| Nautical\_Twilight | object |  |
| Astronomical\_Twilight | object |  |
|  |  |  |

Below is the process flow diagram from csv to modelling

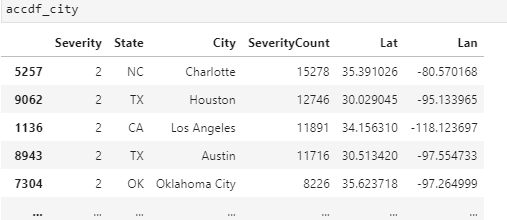


**3. Data analysis:**

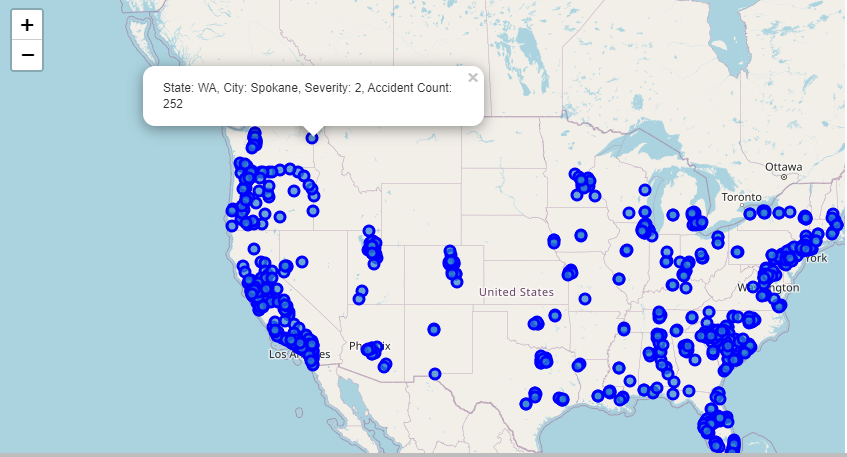
* 1. Have to verify the severity count in DataFrame



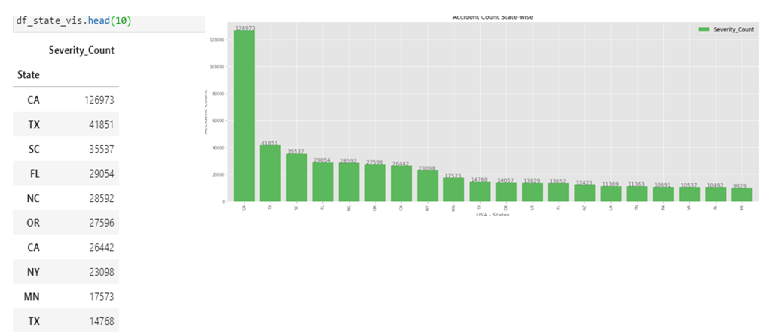
* 1. Have to verify Accident counts with state and city



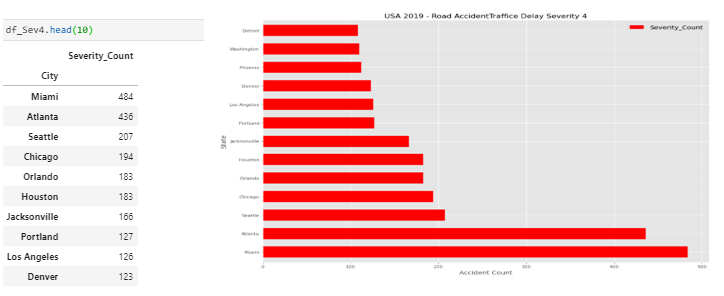
* 1. Use Folium map display the Accident counts in USA map by state, city wise

****

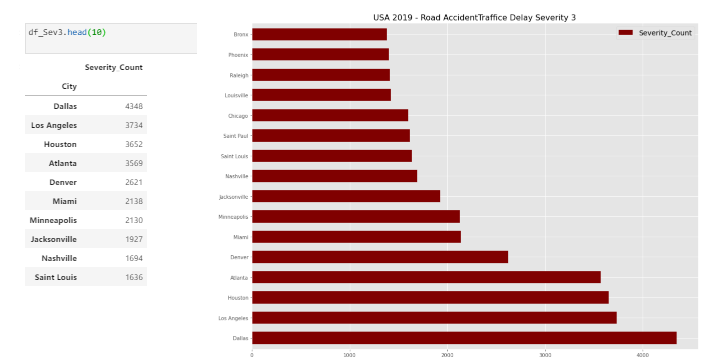
* 1. **Below bar chart indicates Accident count in state wise**

****

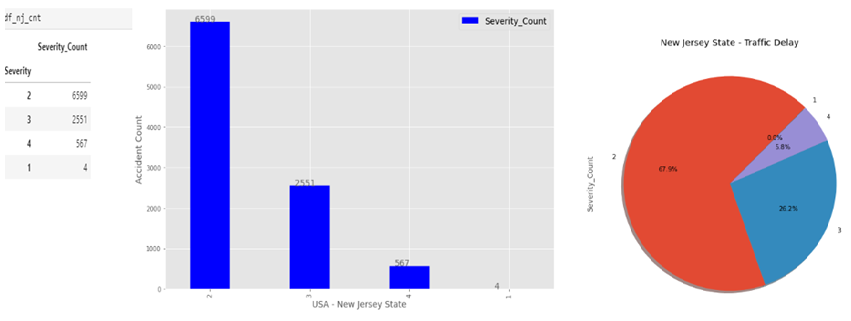
* 1. **Below Bar chart indicates only Severity4 to City wise**

****

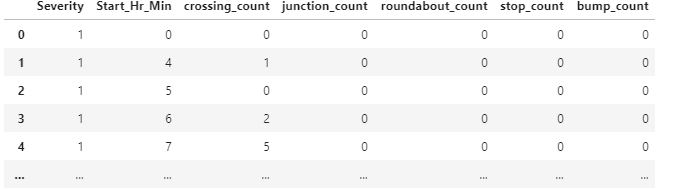
* 1. **Below Bar chart indicates only Severity3 to City wise**

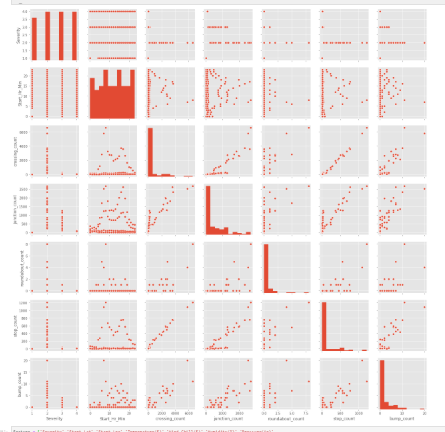
****

* 1. **Below bar/pie chart shows Severity count for State: New Jersey count**

****

* 1. **Below plot indicates other column impact analysis to Severity**





1. **Discussions:**

Mentioned below plots are using in this project.

* Folium map
* Pie Chart
* Bar chart (vertical and horizontal)
* Pairplots

This project has below dependencies

* pandas
* numpy
* bs4
* matplotlib
* requests
* folium
* sklearn

After build the Train and Test data, prediction gives Severity 1 to 4.

1. **Conclusion**:

Built useful models to predict traffic delays and plan for travel in.

Accuracy of the models has room for improvement.

Below is the final score in Logical Regression model.

