COURSERA DATASCIENCE CAPSTONE

Anshuman H

INTRODUCTION

- Predicting Severity of Accident is important
  - To understand the factors leading to severity
  - Understand measure which can be implemeneted
  - Testing the impact of measures

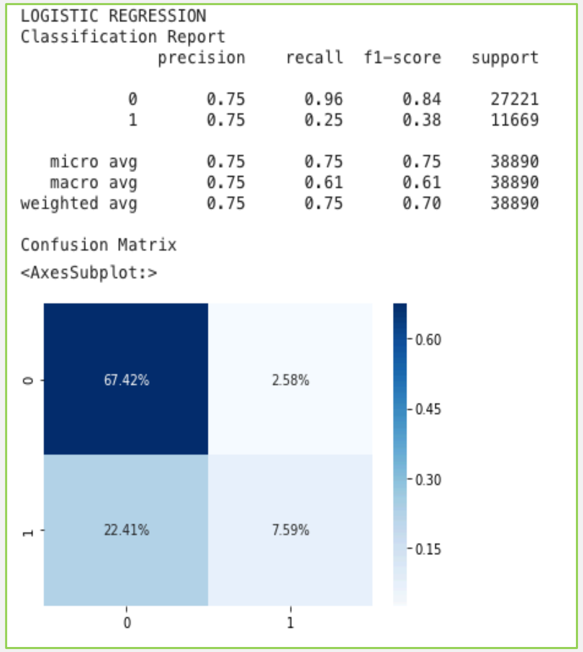
DATA ACQUISITION AND CLEANING

- Data file used is "Collision - All Years" which has been sourced from SDOT Traffic Management Division.
- The data has 38 attributes and covers the time period from 2004 to present date.
- The data is split into 2 kinds of severity ”Injury collision” and “Property Damage Only Collision”.

METHODOLOGY

- Recursive Feature Elimination (RFE) has been used to further refine the selected features
  - Predictor_Matrix_N=['PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT',' SDOT_COLCODE', 'INTKEY','LIGHTCOND_N','ROADCOND_N','WEATHER_N', 'INATTENTIONIND_N','UNDERINFL_N','HITPARKEDCAR_N','SPEEDING_N', 'COLLISIONTYPE_N','JUNCTIONTYPE_N','PEDROWNOTGRNT_N']
- The performance of the below models would be compared
  - Logistic Regression
  - Gaussian Naïve Bayes
  - Random Forest Classifier

RESULT

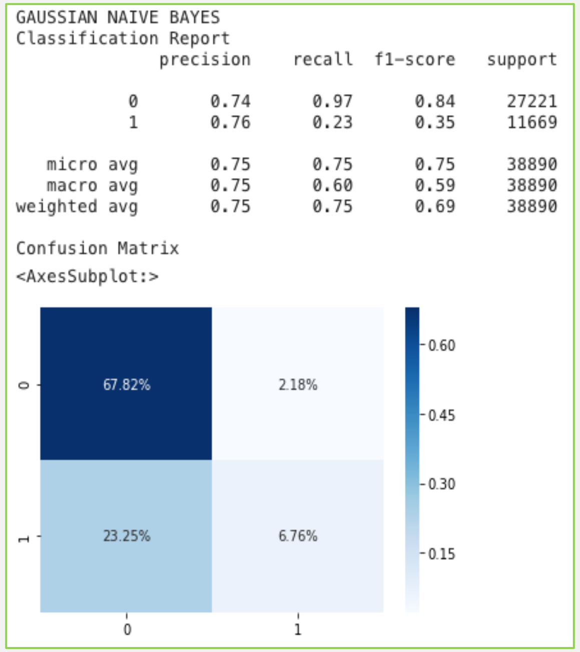


```
LOGISTIC REGRESSION
Classification Report
              precision    recall  f1-score   support

           0       0.75      0.96      0.84     27221
           1       0.75      0.25      0.38     11669

   micro avg       0.75      0.75      0.75     38890
   macro avg       0.75      0.61      0.61     38890
weighted avg       0.75      0.75      0.70     38890

Confusion Matrix
<AxesSubplot:>
```

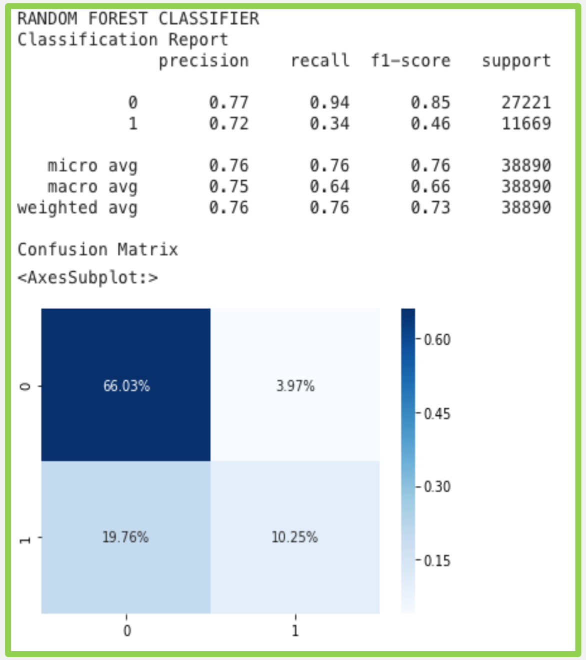


```
GAUSSIAN NAIVE BAYES
Classification Report
              precision    recall  f1-score   support

           0       0.74      0.97      0.84     27221
           1       0.76      0.23      0.35     11669

   micro avg       0.75      0.75      0.75     38890
   macro avg       0.75      0.60      0.59     38890
weighted avg       0.75      0.75      0.69     38890

Confusion Matrix
<AxesSubplot:>
```

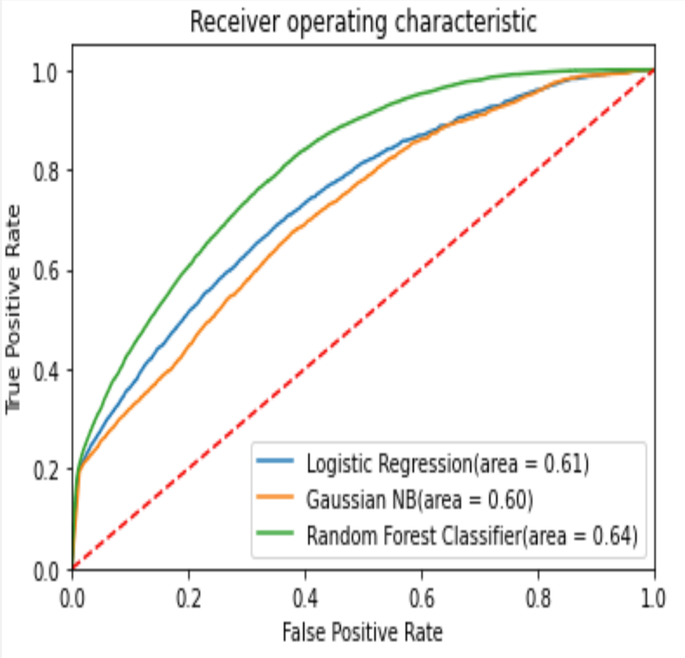


```
RANDOM FOREST CLASSIFIER
Classification Report
              precision    recall  f1-score   support

           0       0.77      0.94      0.85     27221
           1       0.72      0.34      0.46     11669

   micro avg       0.76      0.76      0.76     38890
   macro avg       0.75      0.64      0.66     38890
weighted avg       0.76      0.76      0.73     38890

Confusion Matrix
<AxesSubplot:>
```



CONCLUSION

- Based on the dataset a model to predict severity was initially developed using LogisticRegression and then further modelled using the below 3 models.
  - Logistic Regression
  - Gaussian Naïve Bayes
  - Random Forest Classifier
- **Random Forest Classifier model was able to achieve 75% success rate based on the test data.**