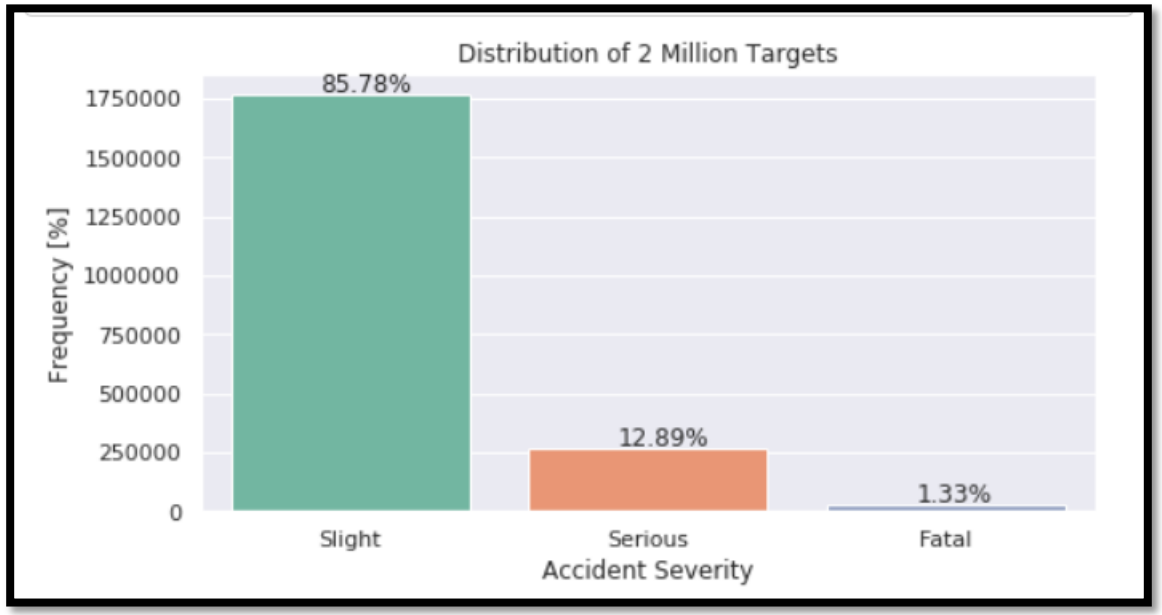
IBM Capstone Project Report

INTRODUCTION

We will be working on UK Accidents Dataset which contains 10 years of data from 2005 to 2014. These files provide detailed road safety data about the circumstances of personal injury road accidents. We are going to perform Exploratory Data Analysis after Data Cleaning and Data Pre-processing. Once the data is ready for analytics phase, we will try to identify the best algorithm that can predict the accident severity and give the highest accuracy on our model.

TARGET VARIABLE



As we can see in the above graph, there are very few fatal accidents that have occurred, most of them have been slight accidents. Since the target variable has very high variablility it will definitely cause a problem while predicting. So, we will do group by splitting.

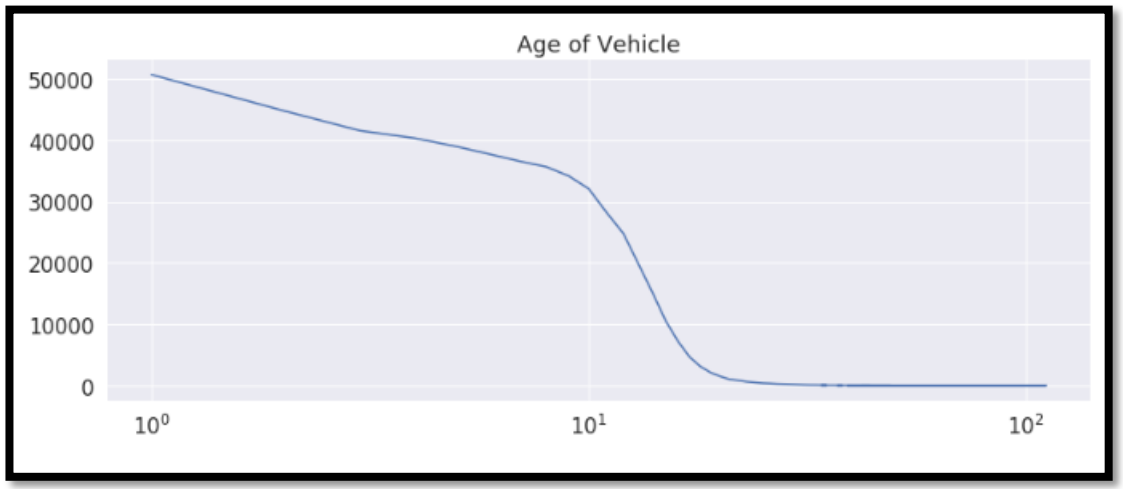
MISSING VALUES

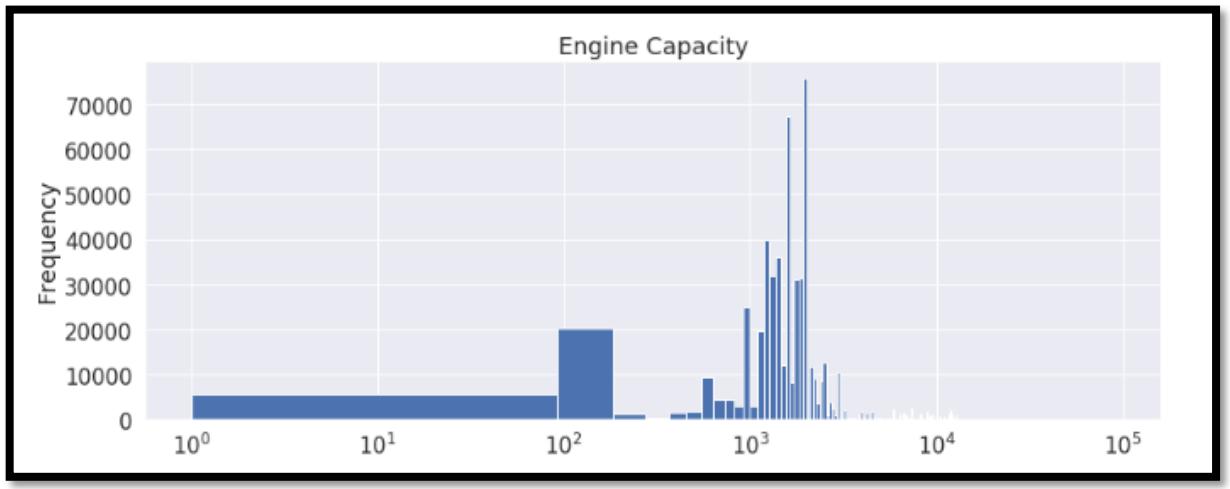
Age_of_Vehicle	102920
Driver_IMD_Decile	206616
Engine_Capacity_.CC.	76213
make	34503
model	96628
Propulsion_Code	71130
Vehicle_Location.Restricted_Lane	300
2nd_Road_Class	263982
2nd_Road_Number	5998
Did_Police_Officer_Attend_Scene_of_Accident	43
Latitude	36
Location_Easting_OSGR	36
Location_Northing_OSGR	36
Longitude	36
LSOA_of_Accident_Location	44560
Pedestrian_Crossing-Human_Control	210
Pedestrian_Crossing-Physical_Facilities	395
Speed_limit	23
Time	51
InScotland	19

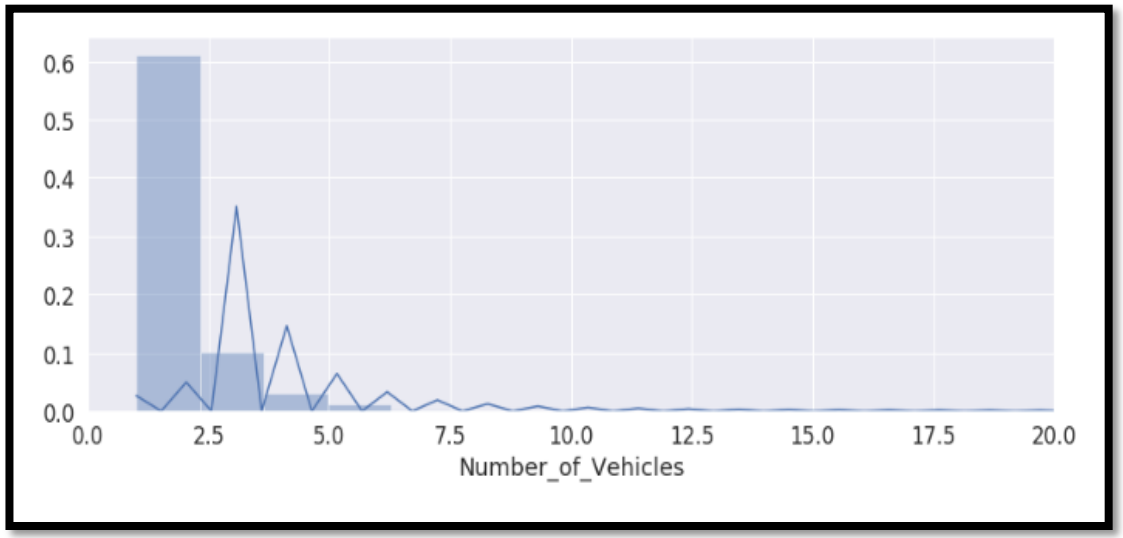
dtype: int64

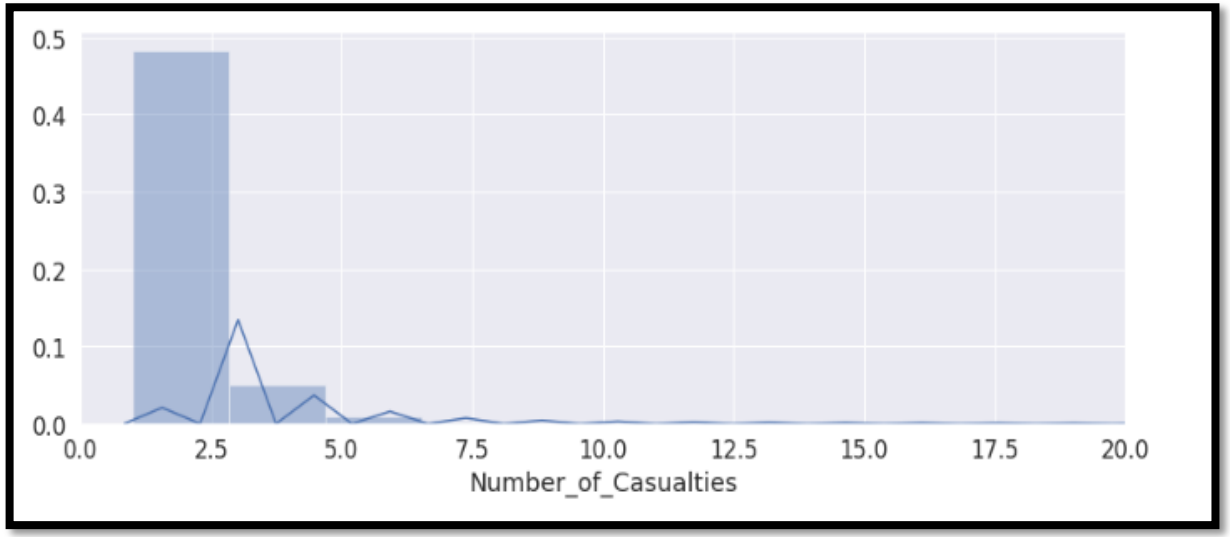
There are a lot of missing values in the dataset. We can impute these values only after performing exploratory analysis.

Exploratory Data Analysis



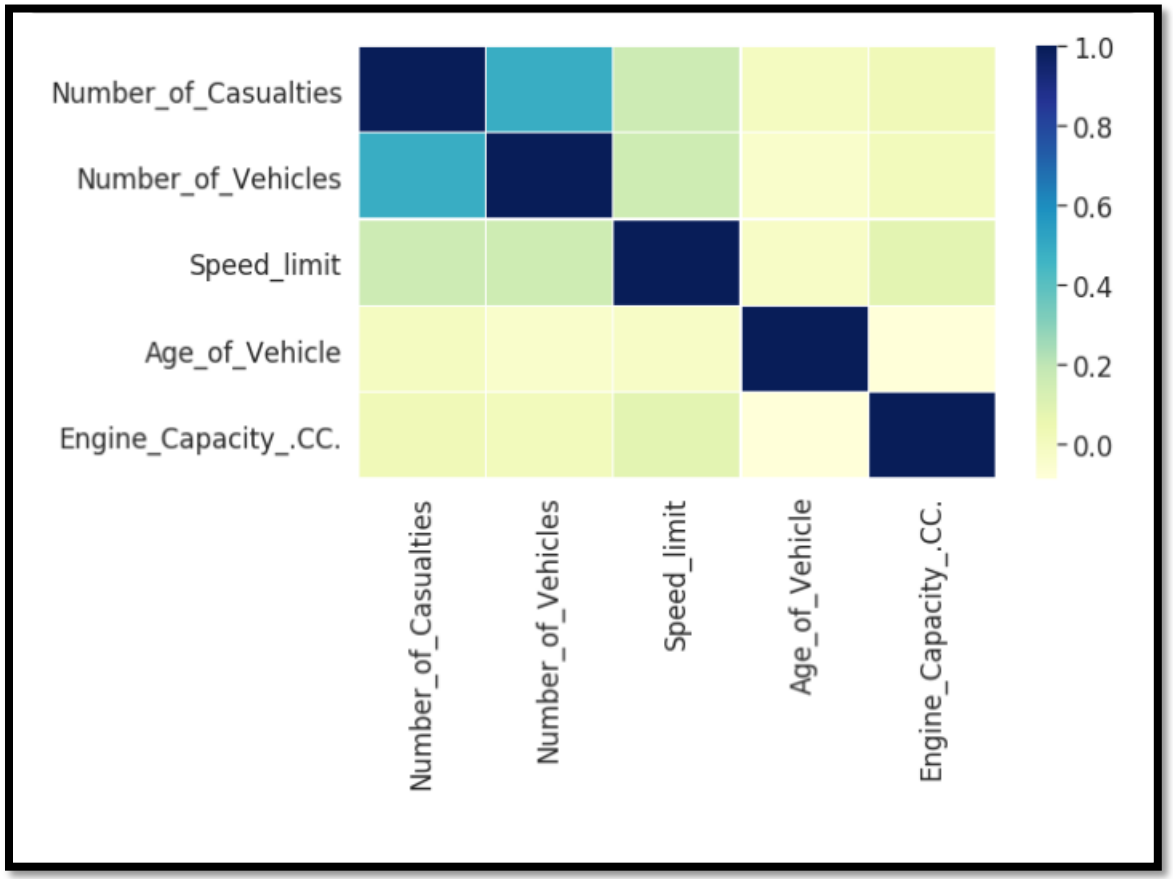






CORRELATION MATRIX

As we can see, the Number of Vehicles shows the highest correlation with the target variable, so, it will be an important feature for our model. Speed limit is also an important feature.



PREDICTION/RESULTS

LOGISTIC REGRESSION

```
Classification Report:              precision    recall

           0       0.56      0.65      0.60     61278
           1       0.74      0.66      0.70     93103

   micro avg       0.66      0.66      0.66    154381
   macro avg       0.65      0.65      0.65    154381
weighted avg       0.67      0.66      0.66    154381

Score: 0.7110141886723705
CPU times: user 2min 34s, sys: 6.93 s, total: 2min 41s
Wall time: 2min 40s
```

RANDOM FOREST

```
Classification Report:              precision    recall
re   support

           0       0.78      0.68      0.73     61278
           1       0.81      0.88      0.84     93103

   micro avg       0.80      0.80      0.80    154381
   macro avg       0.79      0.78      0.78    154381
weighted avg       0.80      0.80      0.79    154381

Score: 0.8615609002141724
CPU times: user 9min 37s, sys: 9.62 s, total: 9min 47s
Wall time: 4min 21s
```

From the above results, Random Classifier worked slightly better than the Logistic model.