IBM Capstone Project Report

-Amit Pathak

INDEX

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.

In a world of increasingly populated cities with underdeveloped public transportation systems, it's not surprising that road accidents are a leading cause of death worldwide. According to a World Health Organization report (February 2020), roughly 20-50 million people endure non-fatal injuries from road traffic accidents every year with and additional 1.35 million people who die from these crashes.

Considering this yearly impact of these accidents, it is important that we are able to determine severity of crashes given some data (in this case, the shared dataset) so that we can then explore the variables that may be factored into preventive measures. As such, the objective of this project is to build and train a supervised machine learning model that is able to detect relationships between certain variables in order to tag a reported accident based on severity.

This will have implications for those in the field of traffic control and safety, like those handling the medical emergencies that result from road accidents, particularly those occuring in the Seattle area. So, we're looking for insights that could aid mainly in post-crash care, but hopefully also in preventive measures.

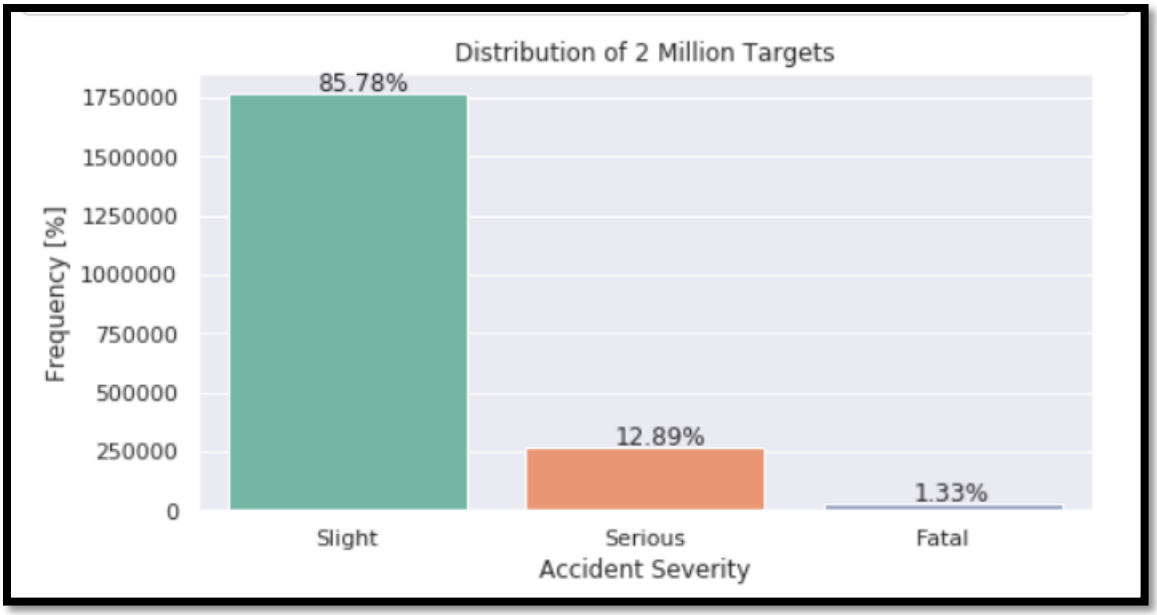
Since this is essentially a classification model that will train over pre-determined severity ratings given Seattle Police Department and Seattle Transport Records, it is also an objective to ensure that this project can be built upon so that anyone interested in developing it further, perhaps with more data or more ideas about potential applications, may do so.

DATA PREPARATION

Load Data

The data contains 57 columns and 2058408 rows. Following are the features of the data.

1. Accident_Index
2. Location_Easting_OSGR
3. Location_Northing_OSGR
4. Longitude
5. Latitude
6. Police_Force
7. Accident_Severity
8. Number_of_Vehicles
9. Number_of_Casualties
10. Date
11. Day_of_Week
12. Time
13. Local_Authority_(District)
14. Local_Authority_(Highway)
15. 1st_Road_Class
16. 1st_Road_Number
17. Road_Type
18. Speed_limit
19. Junction_Detail
20. Junction_Control
21. 2nd_Road_Class
22. 2nd_Road_Number
23. Pedestrian_Crossing-Human_Control
24. Pedestrian_Crossing-Physical_Facilities
25. Light_Conditions
26. Weather_Conditions
27. Road_Surface_Conditions
28. Special_Conditions_at_Site
29. Carriageway_Hazards
30. Urban_or_Rural_Area
31. Did_Police_Officer_Attend_Scene_of_Accident
32. LSOA_of_Accident_Location



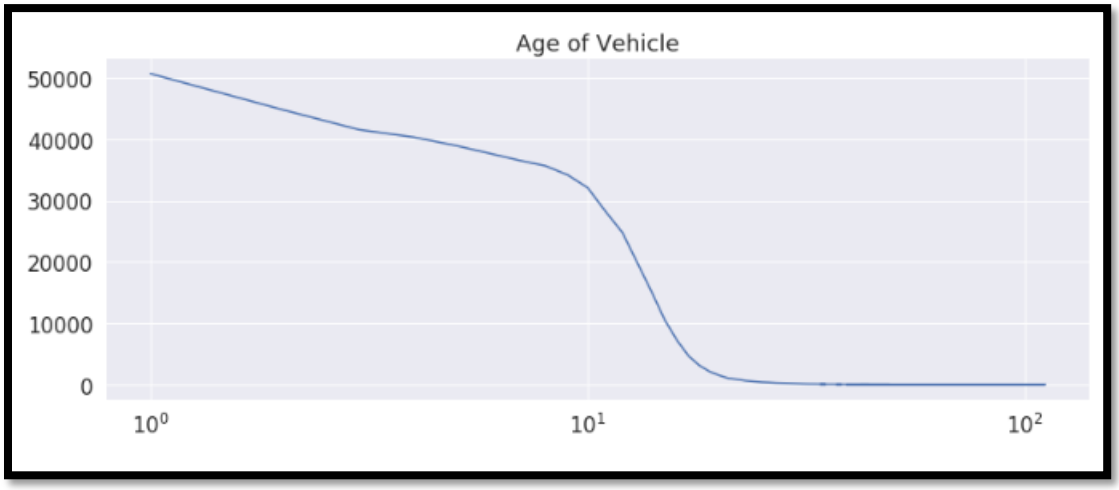
As we can see in the above graph, there are very few fatal accidents that have occurred, most of them have been slight accidents. Since the target variable has very high variablility it will definitely cause a problem while predicting. So, we will do group by splitting.

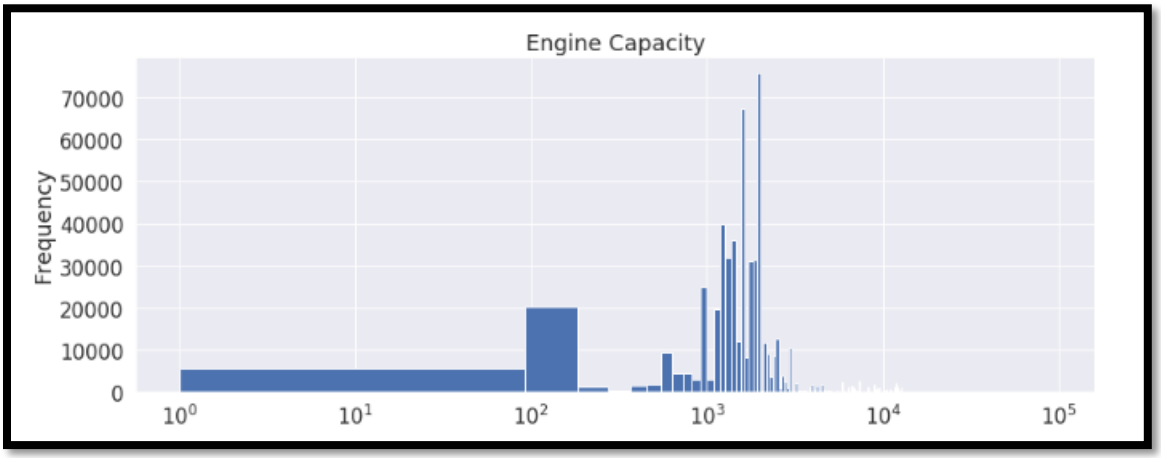
Check for Missing Values

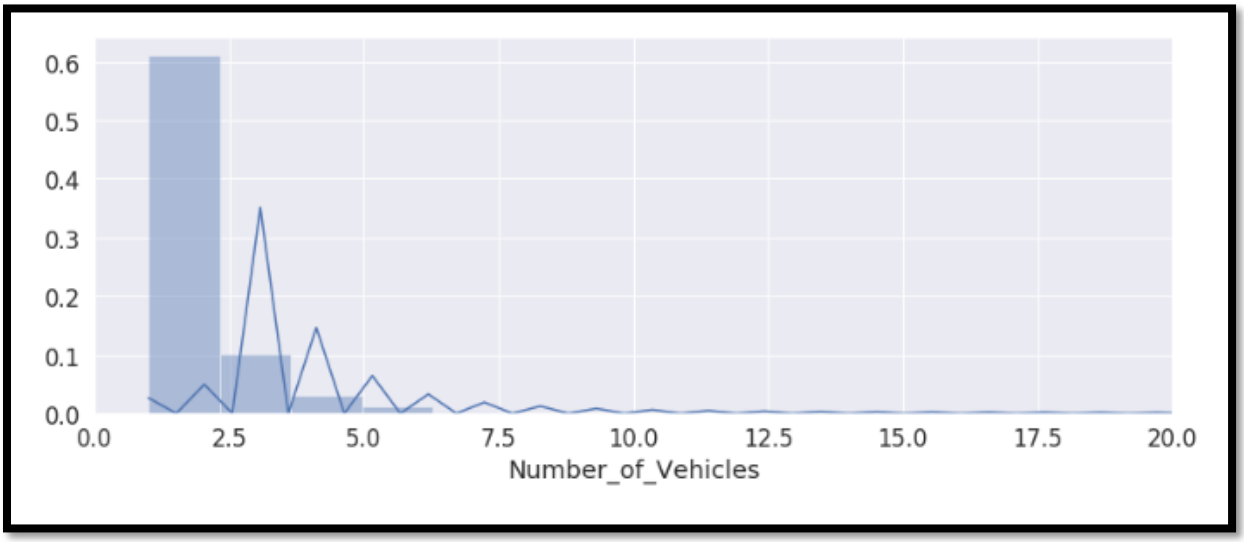
```
Age_of_Vehicle                              102920
Driver_IMD_Decile                           206616
Engine_Capacity_.CC.                         76213
make                                         34503
model                                        96628
Propulsion_Code                              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
```

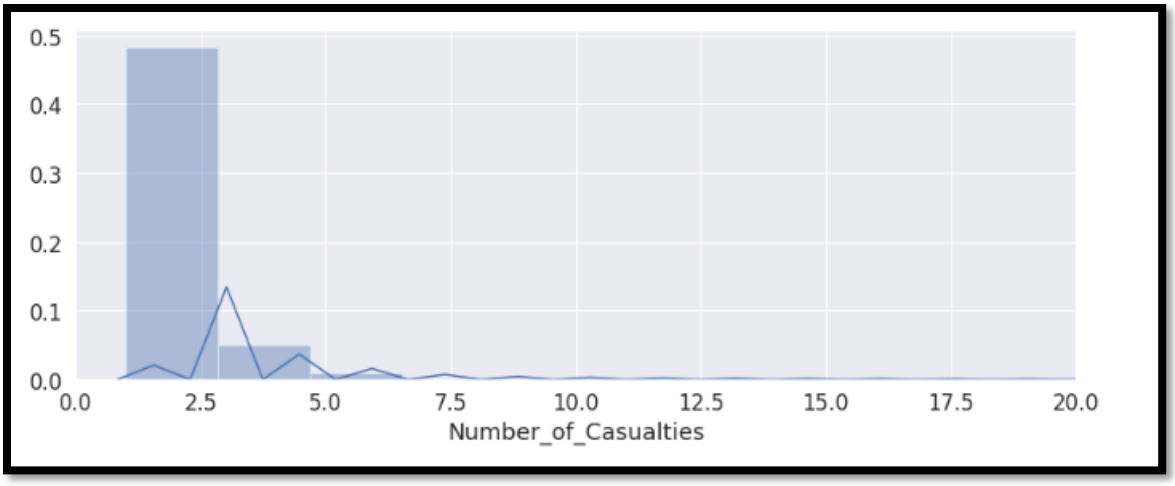
There are a lot of missing values in the dataset. We can impute these values only after performing exploratory analysis.

Exploratory Data Analysis



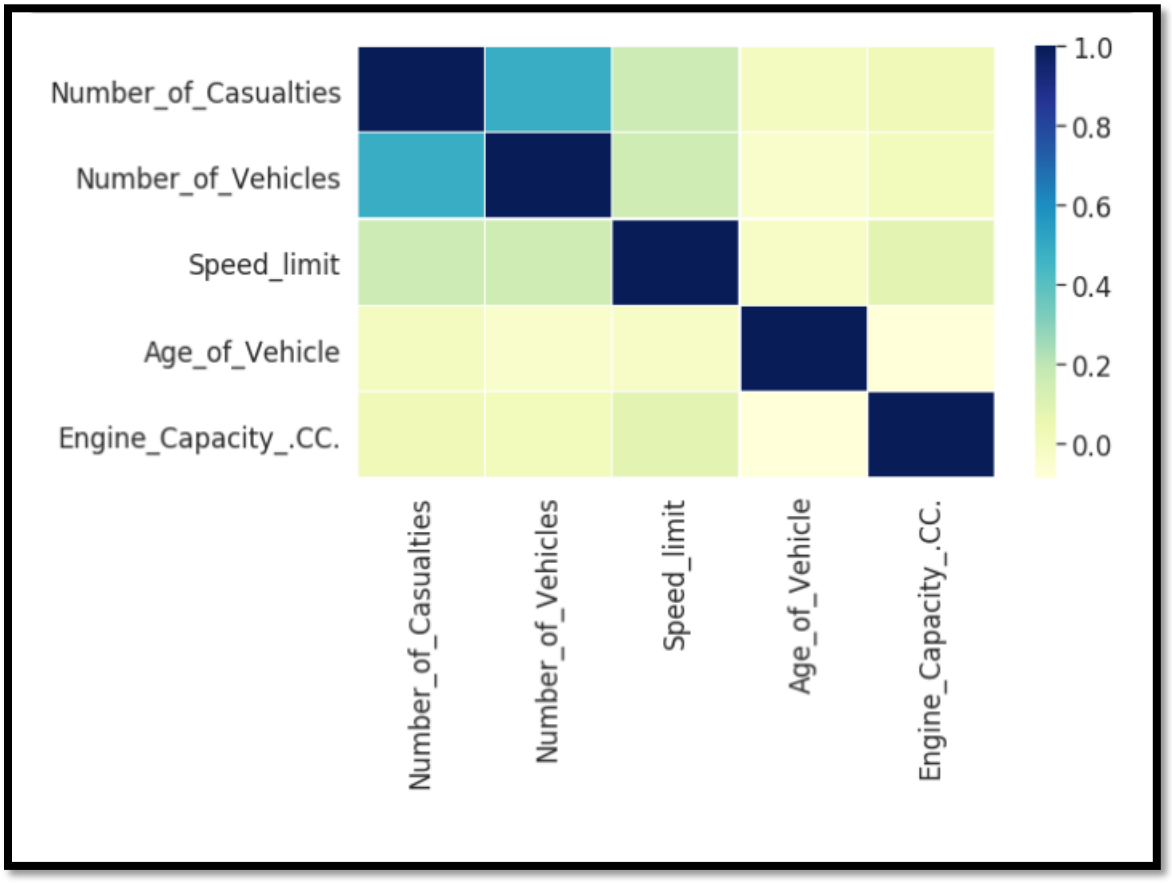






The distribution of No. of Vehicles, No. of Casualties is positively skewed.

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.



TRAINING

Before we move on to training the dataset, there are a few things that we need to look at –

We need to perform encoding of categorical and nominal data.

1. Transform Speed Limit using One Hot Encoding
2. Transform Time into date datatype
3. Impute Age of Vehicle using median values
4. Impute Make using a constant and perform One hot Encoding
5. Impute Engine Capacity with quantile and perform One Hot Encoding

PREDICTION

1. Logistic Regression

```
Classification Report:              precision    recall  f1-score   support

           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
```

2. Random Forest Classifier

```
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.