Accident Severity

1. Business problem

Accidents in traffic lead to associated fatalities and economic losses every year worldwide and thus is an area of primary concern to society from loss prevention point of view. Modeling accident severity prediction and improving the model are critical to the effective performance of road traffic systems for improved safety. In accident severity modeling, the input vectors are the characteristics of the accident and attributes of vehicle while the output vector is the corresponding class of accident severity.

There are two main engineering approaches for dealing with traffic safety problems: the reactive approach and the proactive approach. The reactive approach, or retrofit approach, consists of making the necessary improvements to variable, for instance, existing hazardous sites in order to reduce collision frequency and severity at these sites. The proactive approach, on the other hand, includes a collision prevention approach, like, preventing a potential unsafe road conditions from occurring in the first place. We focus on proactive approach which involves prediction of accident severity and working backwards, the concerned entity implements appropriate remedial measures to improve road safety. By recognizing the key factors that influence accident severity, the solution may be of great utility to various Government Departments/Authorities like Police, R&B and Transport from public policy point of view. The results of analysis and modeling can be used by these Departments to take appropriate measures to reduce accident impact and thereby improve traffic safety. It is also useful to the Insurers in terms of reduced claims and better underwriting as well as rate making.

2. Data

The UK government collects and publishes (usually on an annual basis) detailed information about traffic accidents across the country. This information includes, but is not limited to, geographical locations, weather conditions, type of vehicles, number of casualties and vehicle manoeuvres, making this a very interesting and comprehensive dataset for analysis and research.

The data come from the Open Data website of the UK government, where they have been published by the Department of Transport.

The dataset comprises of two csv files:

- Accident_Information.csv: every line in the file represents a unique traffic accident (identified by the Accident_Index column), featuring various properties related to the accident as columns. Date range: 2005-2017
- Vehicle_Information.csv: every line in the file represents the involvement of a unique vehicle in a unique traffic accident, featuring various vehicle and passenger properties as columns. Date range: 2004-2016

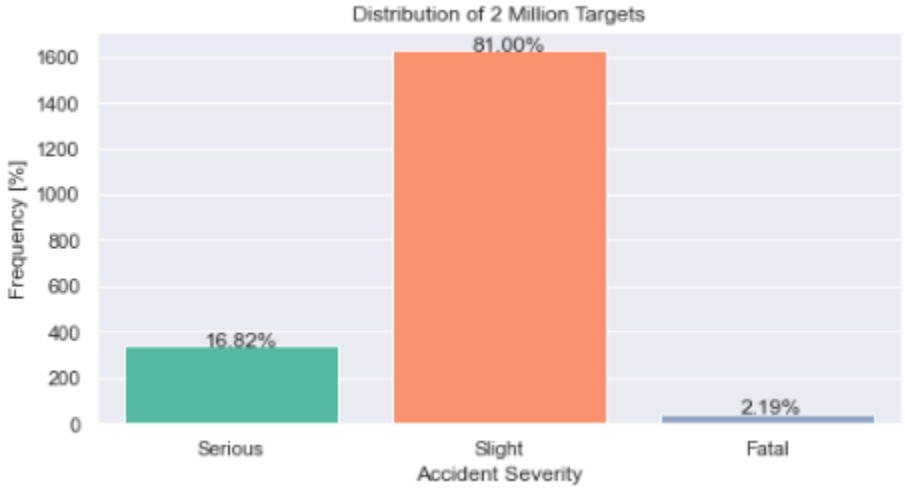
Our target is to predict the accident severity. The severity is devided to two catagories; severe and slight.

	Accident_Index	Age_Band_of_Driver	Age_of_Vehicle	Driver_Home_Area_Type	Driver_IMD_Decile	Engine_Capacity_.CC.	Hit_Object_in_Carriageway	Hit
0	201091NM02142	26 - 35	13.0	Small town	NaN	1389.0	Other object	
1	201091NM01964	36 - 45	11.0	Data missing or out of range	NaN	955.0	None	
2	201091NM01964	46 - 55	NaN	Rural	NaN	NaN	None	
3	201091NM01935	26 - 35	2.0	Urban area	NaN	1997.0	None	
4	201091NM01935	46 - 55	NaN	Rural	NaN	NaN	None	

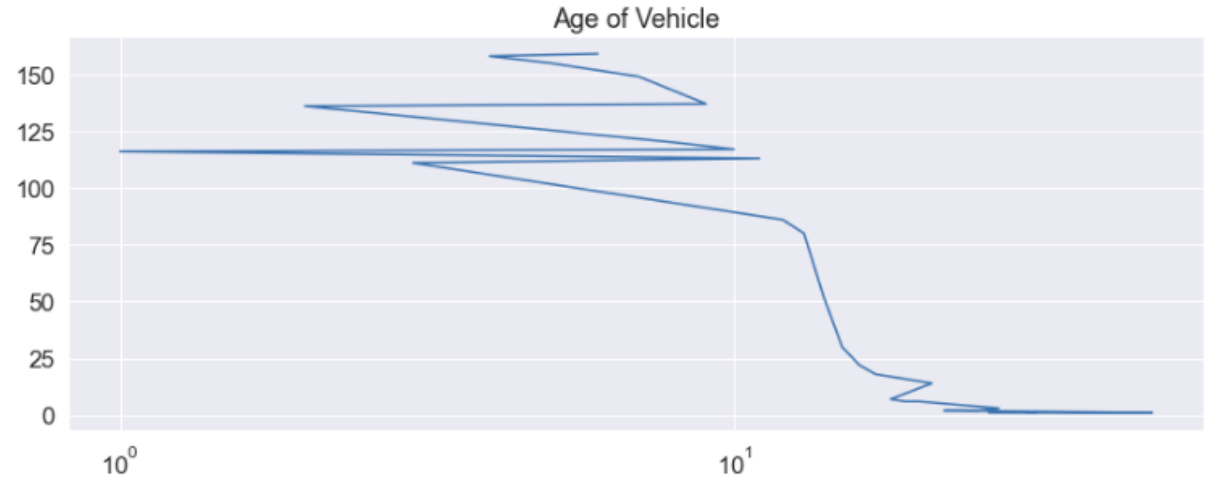
	Accident_Index	Age_Band_of_Driver	Age_of_Vehicle	Driver_Home_Area_Type	Driver_IMD_Decile	Engine_Capacity_.CC.	Hit_Object_in_Carriageway
count	2010	2010	1805.000000	2010	198.000000	1844.000000	2010
unique	1396	9	NaN	4	NaN	NaN	10
top	201063CP26210	36 - 45	NaN	Rural	NaN	NaN	None
freq	5	406	NaN	964	NaN	NaN	1912
mean	NaN	NaN	7.561773	NaN	6.176768	1978.297180	NaN
std	NaN	NaN	4.650678	NaN	2.666595	1853.042052	NaN
min	NaN	NaN	1.000000	NaN	1.000000	49.000000	NaN
25%	NaN	NaN	4.000000	NaN	4.000000	1299.000000	NaN
50%	NaN	NaN	7.000000	NaN	6.000000	1598.000000	NaN
75%	NaN	NaN	10.000000	NaN	9.000000	1997.000000	NaN
max	NaN	NaN	48.000000	NaN	10.000000	16120.000000	NaN

3. Exploratory Data Analysis (EDA) :

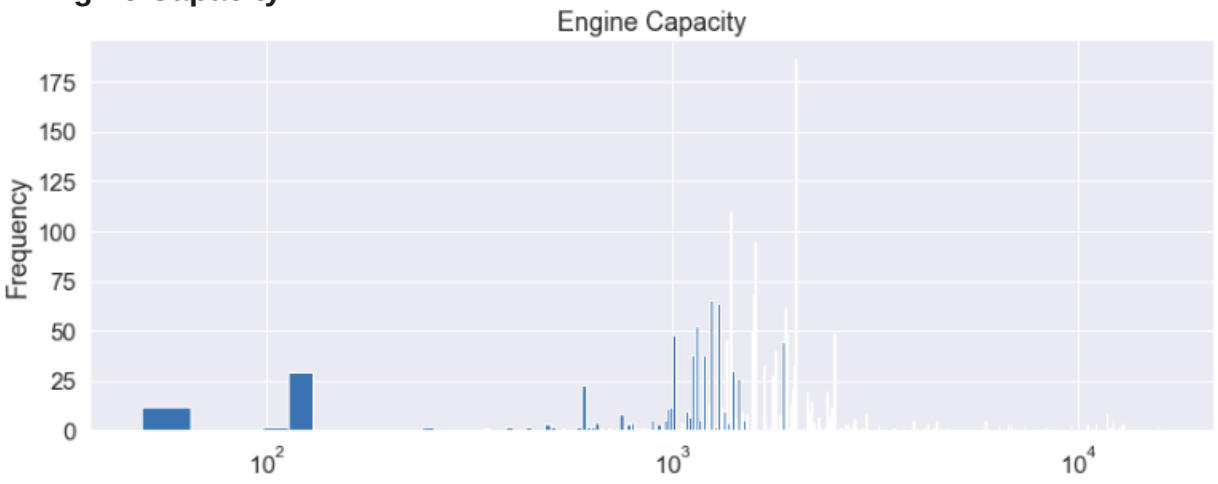
3.1 Distribution of Accident Severity



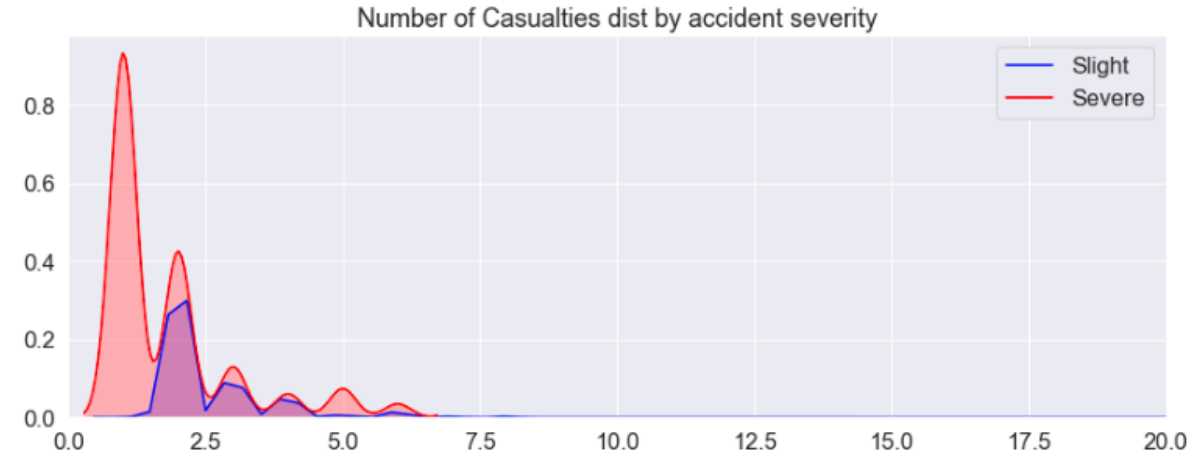
3.2 Age of Vehicle



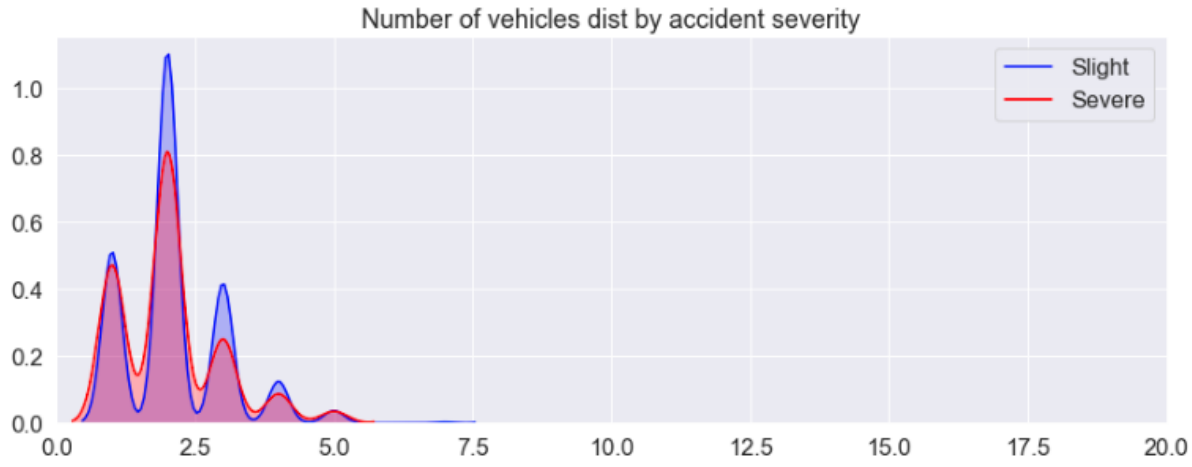
3.3 Engine Capacity



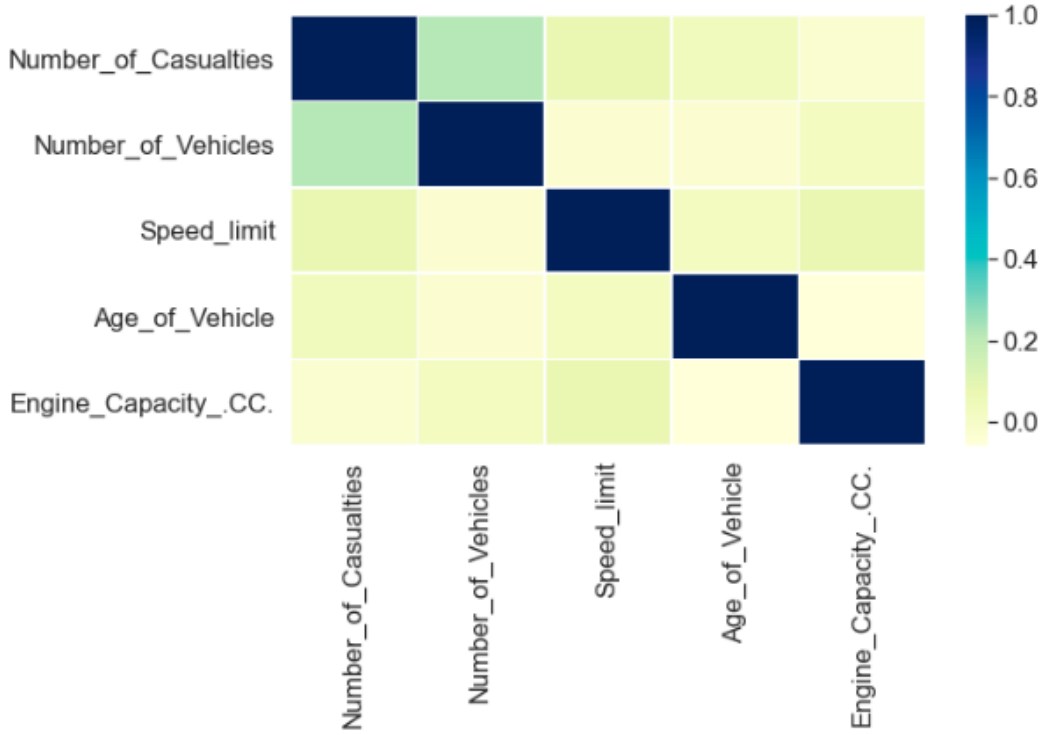
3.4 Number of Casualties distribution by accident severity



3.5 Number of vehicles distribution by accident severity



3.6 Feature correlation



4. Methodology:

Considering that most of the features in the dataset are categorical, tried to solve problem with below listed model.

1. Logistic Regression
2. Random Forest Classifier.

5. Results:

5.1 Logistic Regression

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

           0       0.31      0.64      0.42        83
           1       0.91      0.72      0.80       420

    accuracy                           0.70       503
   macro avg       0.61      0.68      0.61       503
weighted avg       0.81      0.70      0.74       503

Score: 0.7590074584050488
```

5.2 Random Forest

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

           0       0.87      0.24      0.38        83
           1       0.87      0.99      0.93       420

    accuracy                           0.87       503
   macro avg       0.87      0.62      0.65       503
weighted avg       0.87      0.87      0.84       503

Score: 0.810140562248996
```

6. Conclusion

The initial goal of this project was to train a model that would be able to predict the severity. Random Forest model gave best results.