Mind Benders

Abhinav Bhatnagar*
BE16B016

Rahul Chakwate*
AE16B005

Data Analytics Laboratory, Indian Institute of Technology Madras

*equal contributions

Problem Statements Identified

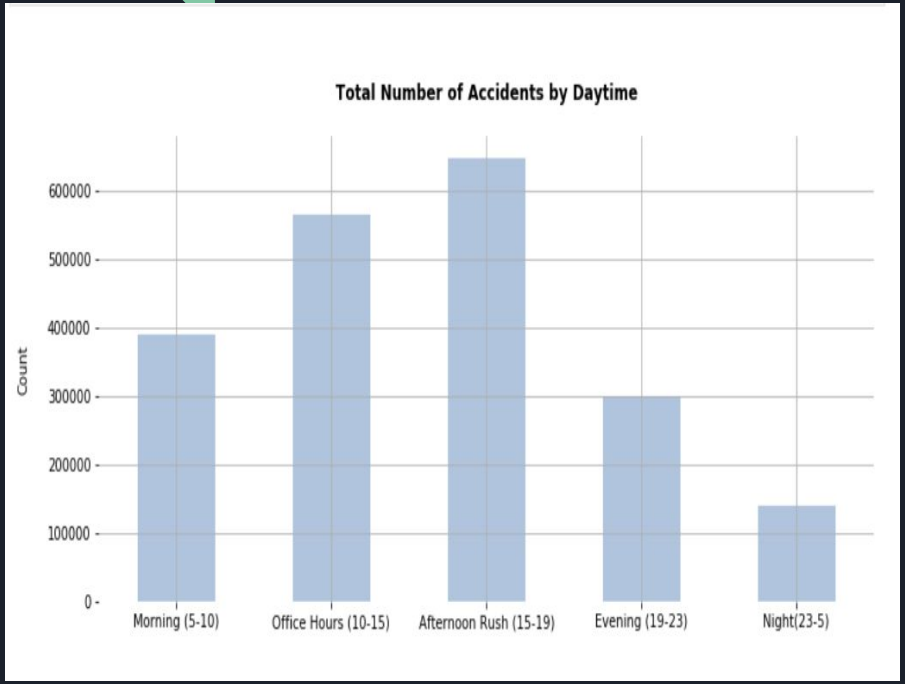
1. What are the important features in predicting (i) casualty and (ii) severity using Lasso, Random Forest, etc.?
2. Group different locations (districts) into various categories based on safety level (Safe, Moderate, Risky, etc.) using clustering techniques.
3. Predict the safest age group and gender for the driver for different given vehicle types.

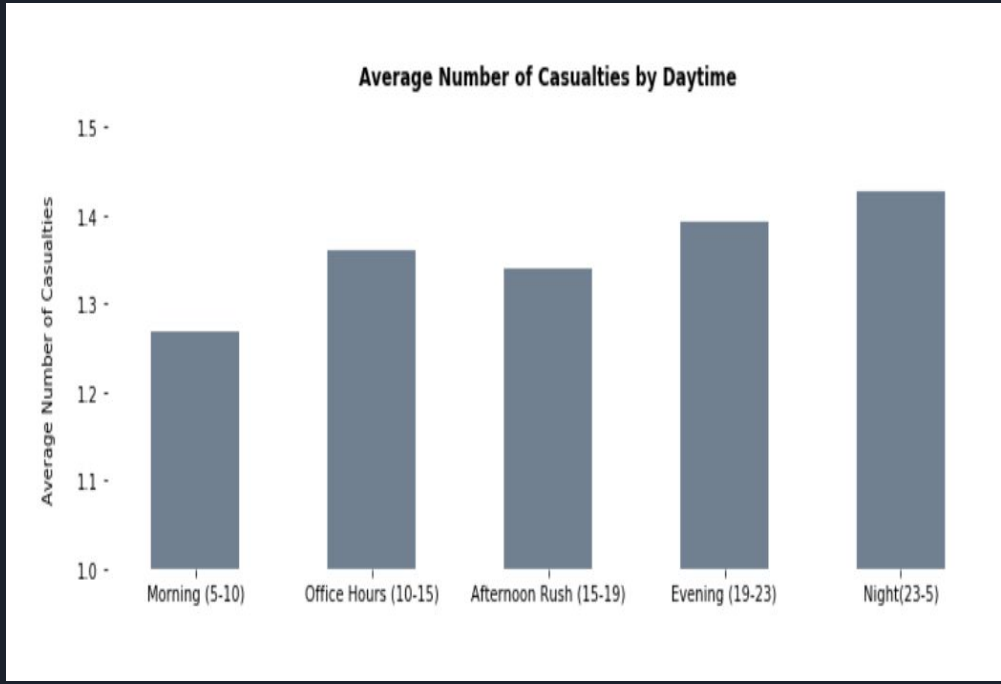
Q1. Predicting Importance of Features

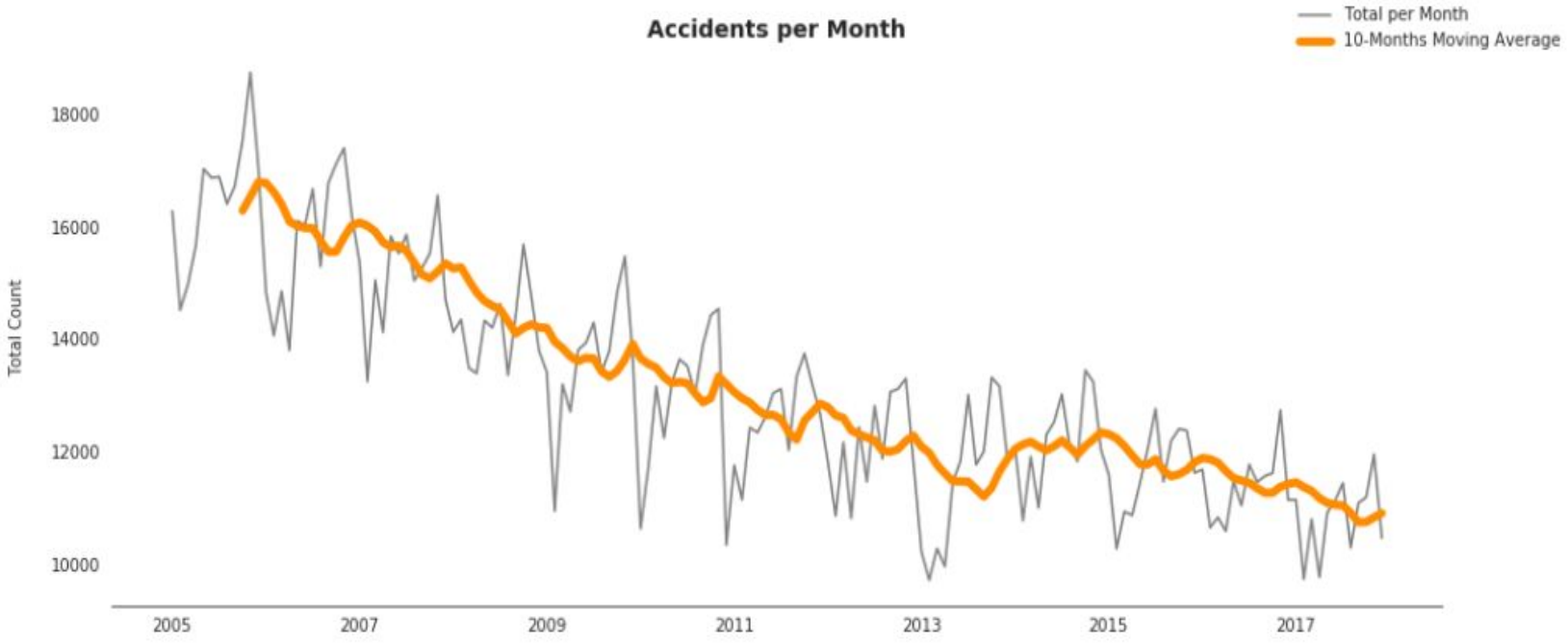
Data Preprocessing:

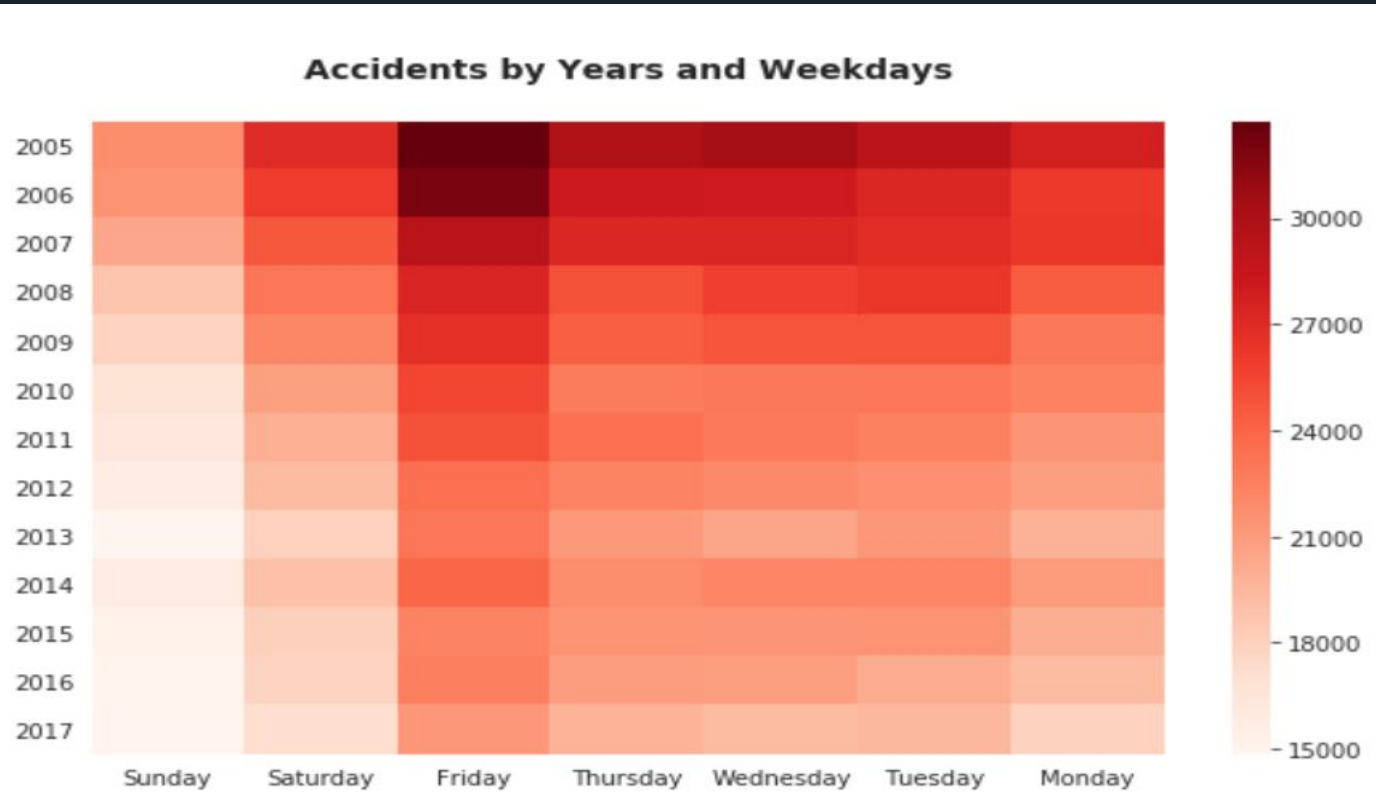
- Finding Missing data and removing or imputing it
- Finding a new feature called number of days since accident occured
- Extracting the hour when the accident occured
- Finding features like daytime groups in which accidents occured
- Identifying Outliers and removing them for numerical data
- Converting Categorical data to category type
- Dropping off unnecessary columns: Driver_IMD_Decile Accident_Index,Location_Easting_OSGR","Location_Northing_OSGR" etc

Feature Engineering

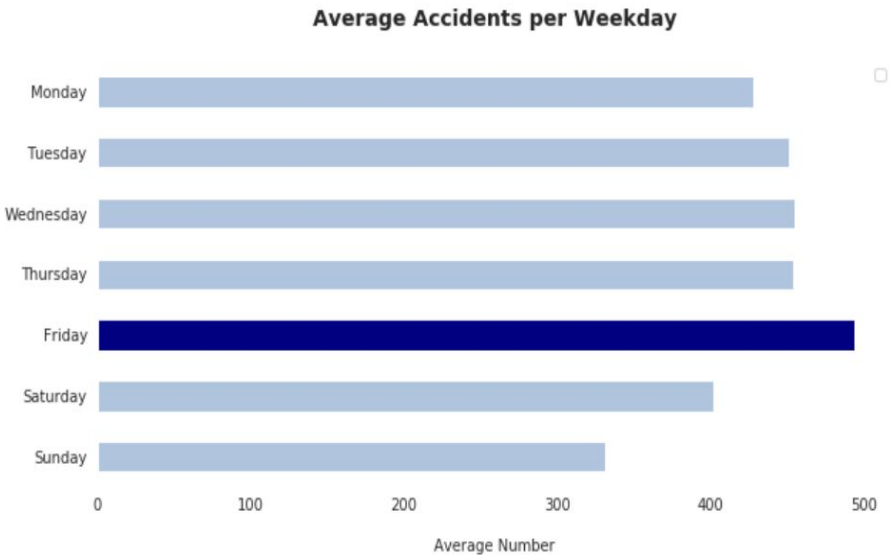


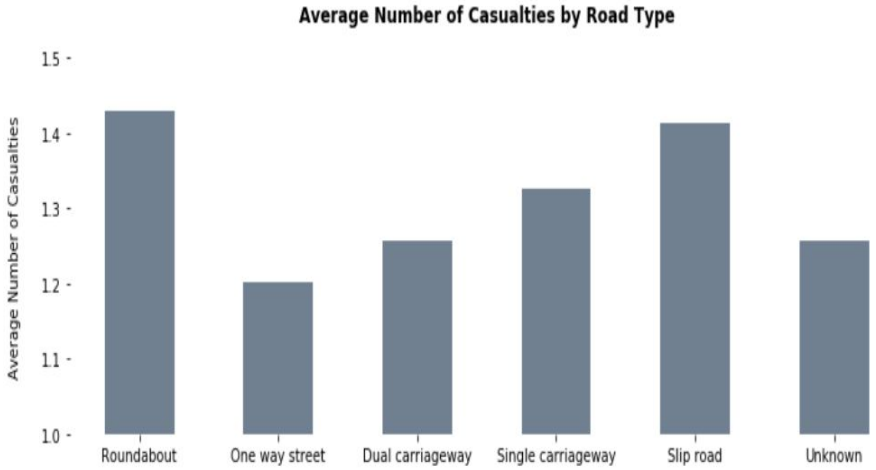



Accidents per Month
Total per Month
10-Months Moving Average
Total Count
18000
16000
14000
12000
10000
2005
2007
2009
2011
2013
2015
2017

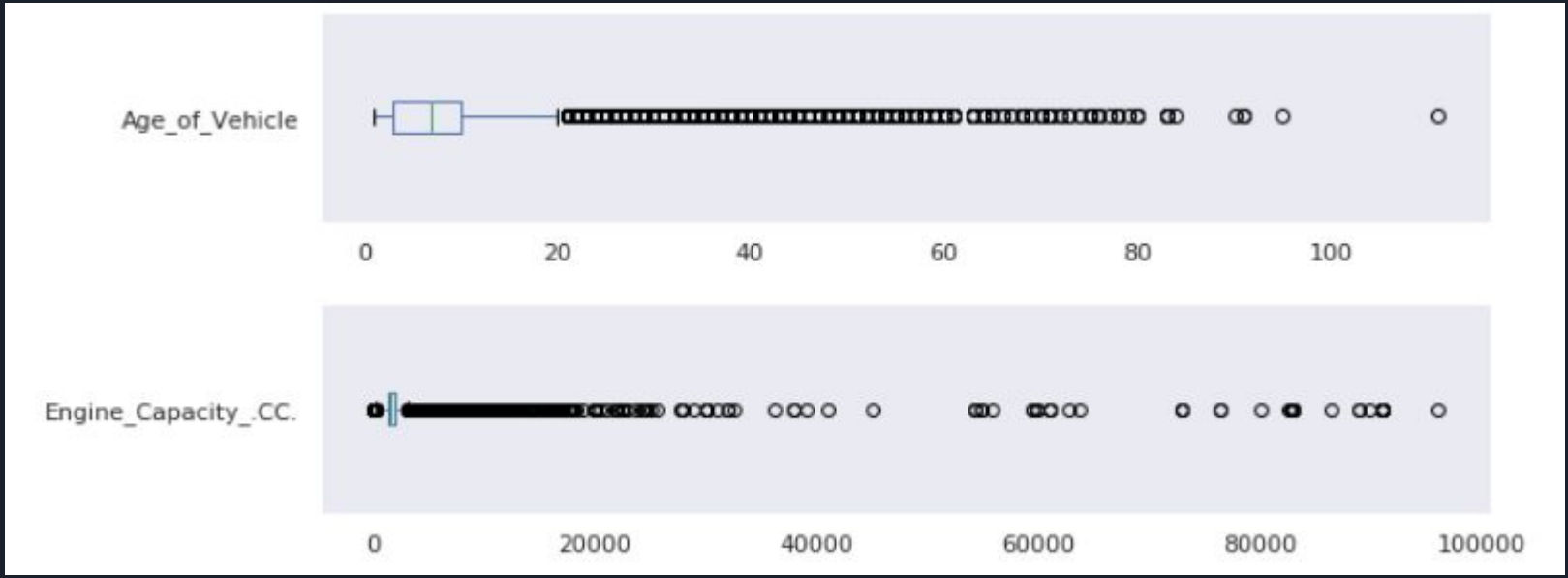

Accidents by Years and Weekdays
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Sunday
Saturday
Friday
Thursday
Wednesday
Tuesday
Monday
30000
27000
24000
21000
18000
15000

Visualization of Certain Features





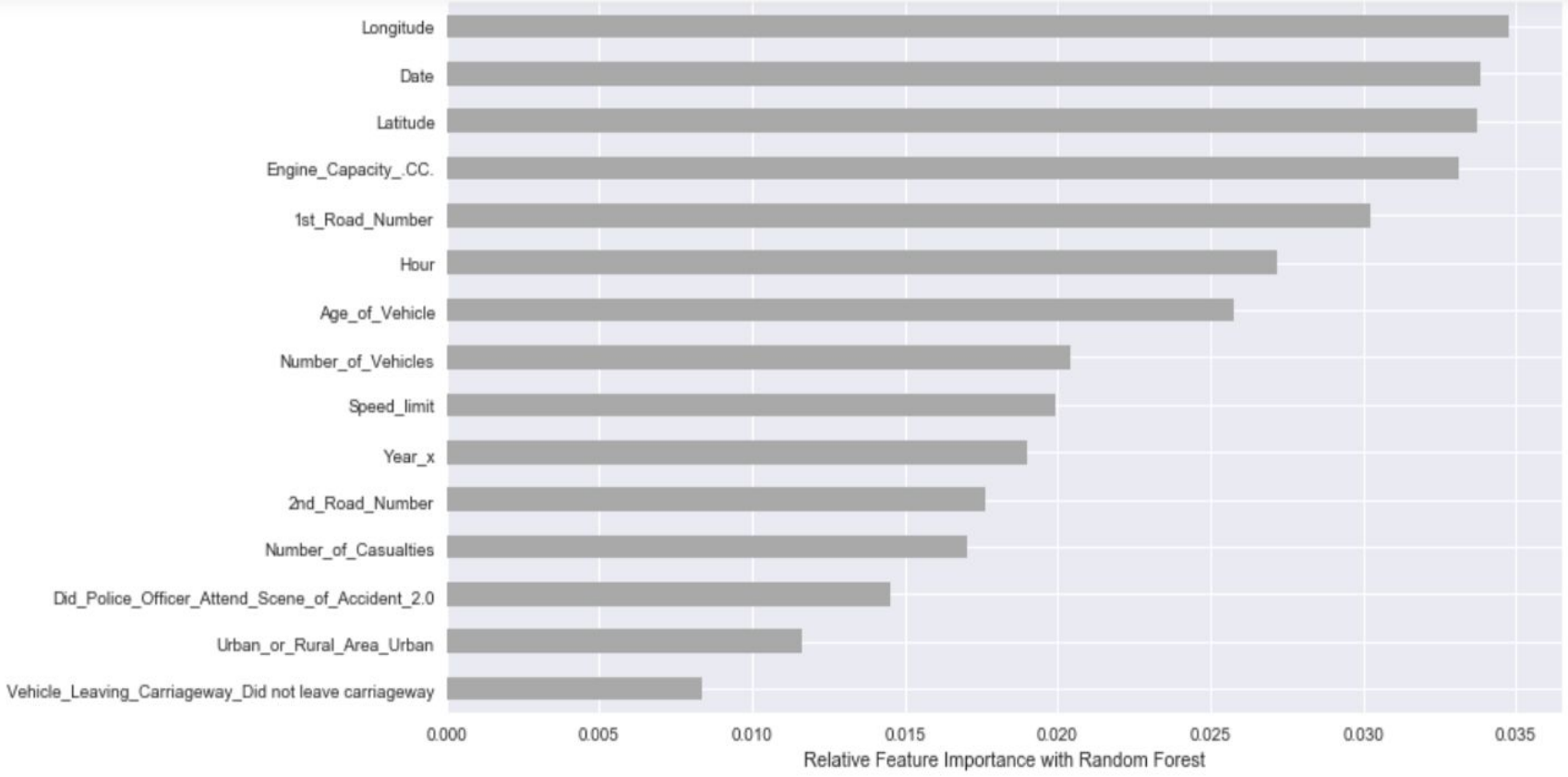
Box Plot Depicting Outliers



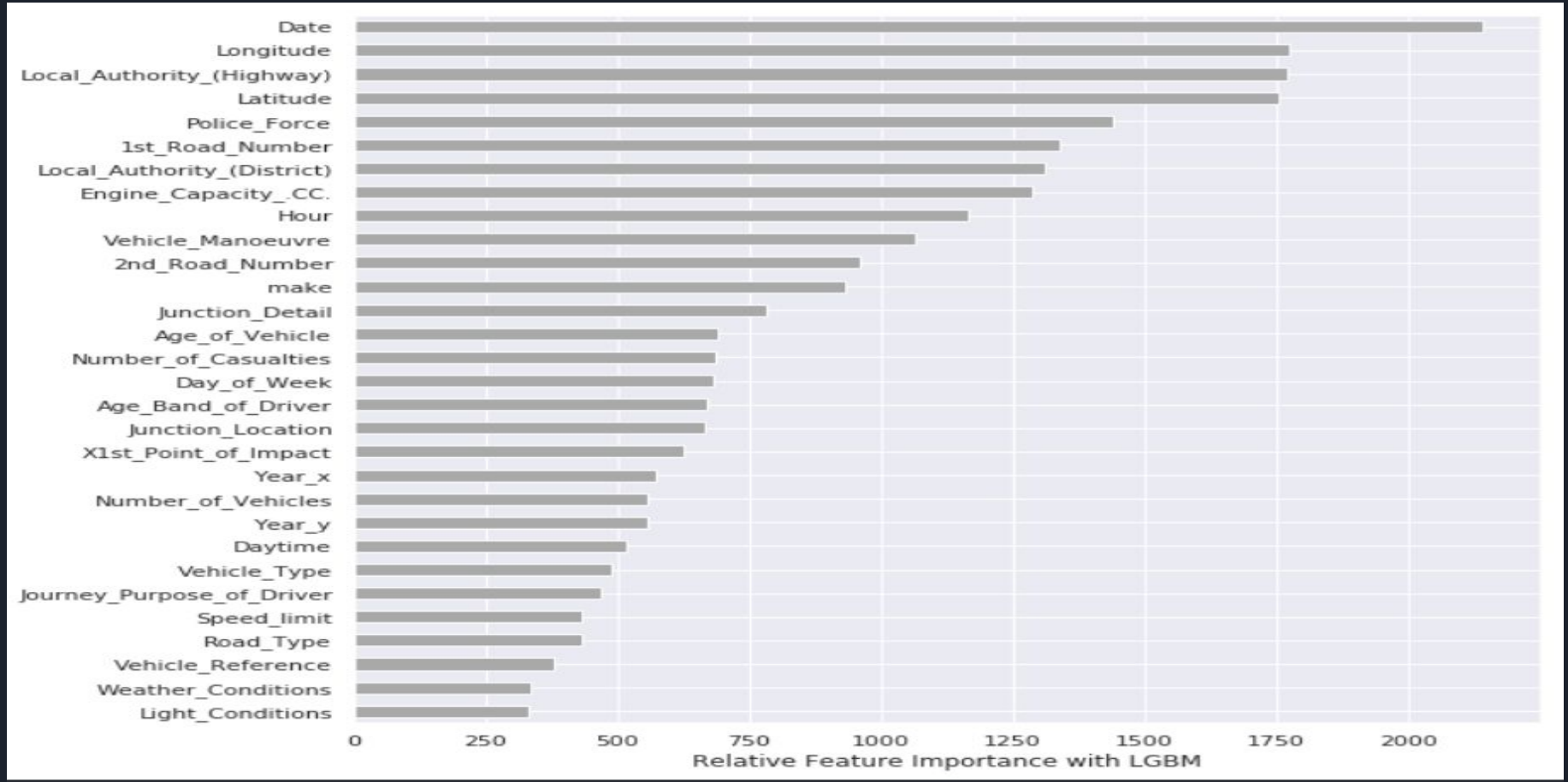
Q1. Building the Model for Feature Importance

- We used the Random Forest and LGBM to predict Accident Severity and used SMOTE to prevent the class imbalance in the data.This ended up increasing our accuracies.
- We also ran group Lasso regression on predicting number casualties but it did not achieve good results.
- Random Forest does not deal with categorical features. One has to One-Hot-Encoding to feed categorical variables into RFs. How do we explain feature importance?
- However, LGBM has inbuilt functionality to handles categorical variables. So it includes categorical variables into feature importances plot.

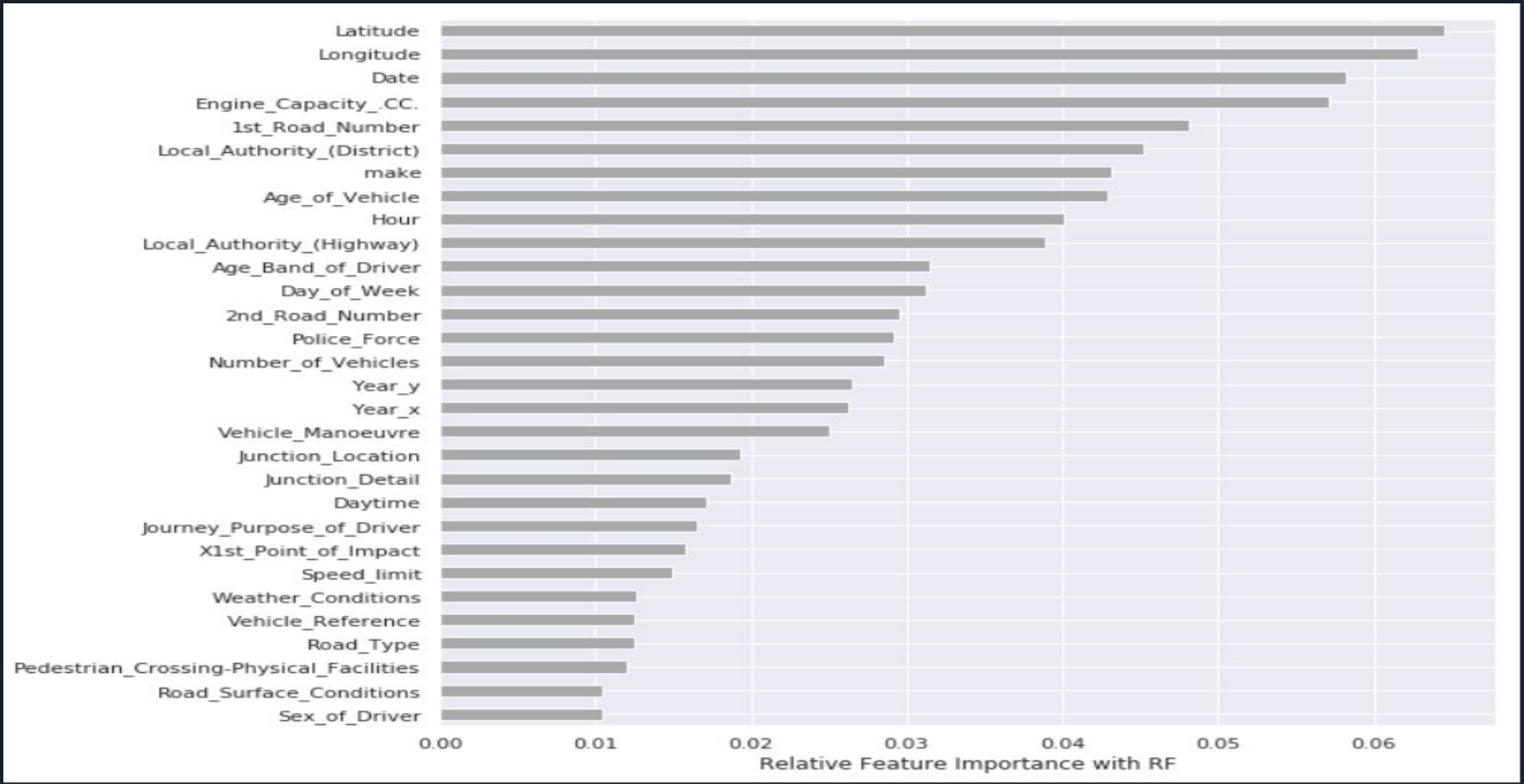
Feature Importance for predicting Severity with RF



Feature Importance for predicting Severity with LGBM

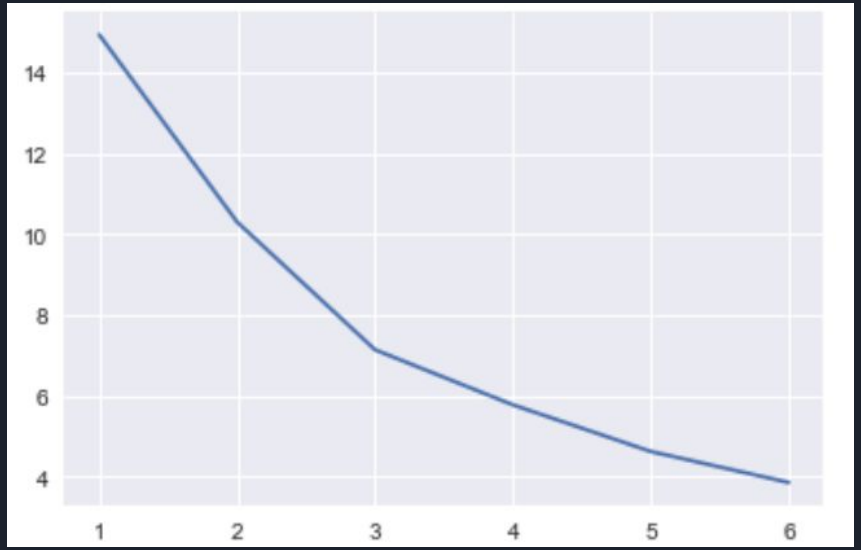


Feature Importance for predicting Casualties with RF



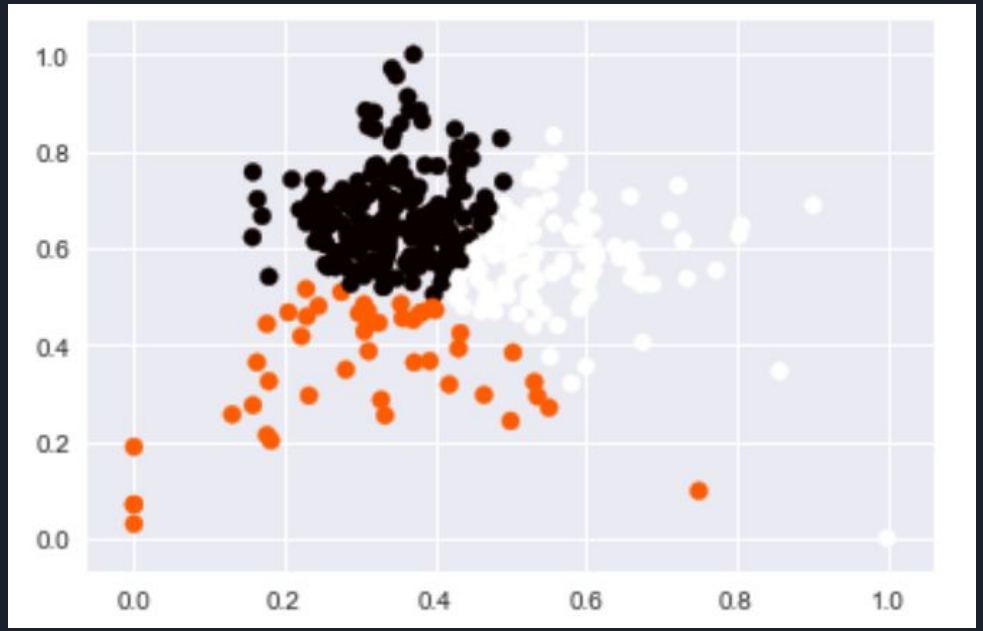
Q2. Clustering districts by the Safety Level

- We defined 2 features to be clustered:
- log(No of accidents) *Mean Number of Casualties per accident for every district
- Mean Accident Severity of every district
- We then tried to find the optimal clusters using the elbow method ie 3 here.
- We then clustered into 3 levels to find 3 different safety levels of ie Safe Moderate and Risky whose graph has been shown on the next page



Visualization of Clusters depicting safety levels of the Districts

- This gave us Safety levels of 1,0.6 and 1.2.
- Three clusters indicate that that there are 2 clusters of Low Severity but high having high and low number of accidents and average casualties and only one cluster where you have high severity and intermediate casualties and accidents.



Safety Defintion

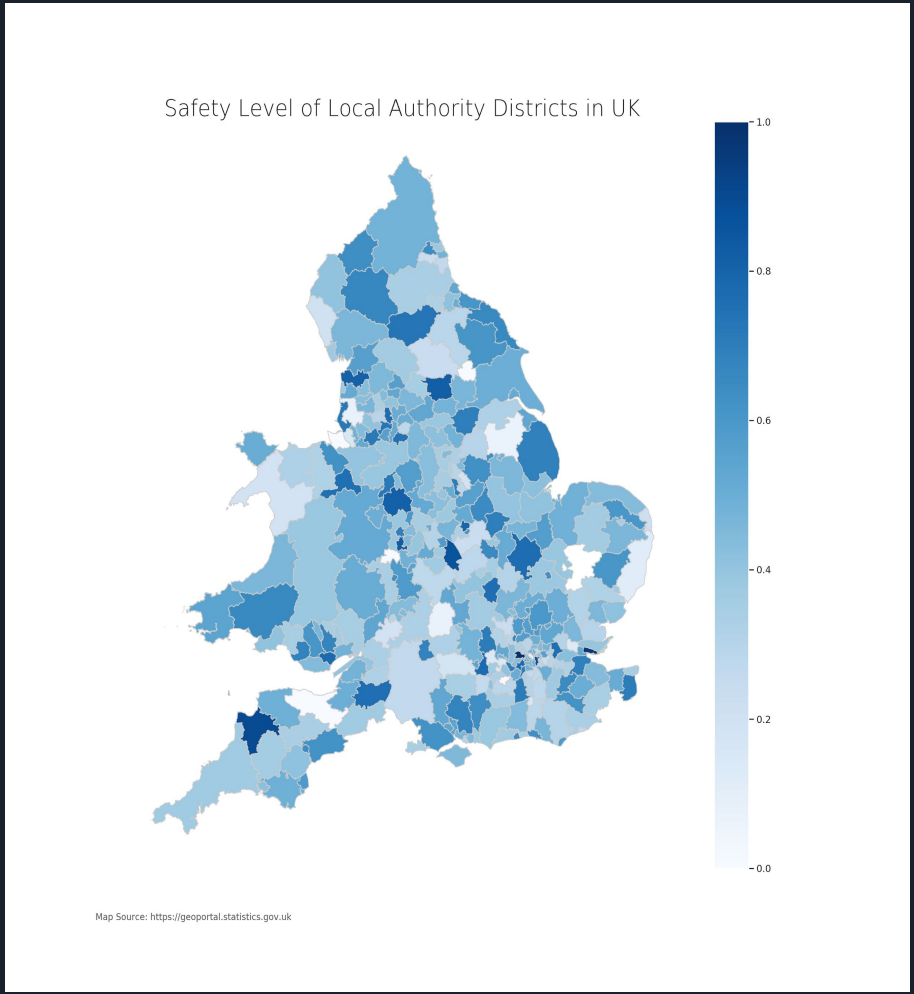
- To define Safety we ran PCA on the 3 features to get a weighted equation that will give us maximum separation among the cluster centers.
- This is the equation we got:

Safety Level= 0.550 x *(Number of Accidents per district) - 0.5968* x *(Number of casualties per accident per district) - 0.5837* x *(Accident Severity per accident per district)*

- This assigned pur clusters safety levels of -1.112, 1.082 and 0.993
- This proved to be counterintuitive so we decided to get a safety level given by the sum of the accident severity and our composite variable defined earlier

Choropleth Map:

- The Adjacent Map shows the Local Authority Districts in UK and the Safety Level associated with each district.
- Shapefile Source: https://geoportal.statistics.gov.uk
- Used Geopandas library to map the generated safety variable to the map data.
- While almost all the districts had the same names in both the files, some districts names had to be renamed in the original data file.



Q3. Predicting the safest age group and gender for models

- We used the previous safety definition to get a prediction of safe values for different models
- We also tried to find the absolute safety of each model over the entire dataset
- These were found to be the least safe
  - **RENAULT,**
  - **VAUXELL**
  - **PEUGEOT**
- These were one of the safest cars
  - **VENTURI**
  - **NORTON**
  - **SANTANA**
  - **ENFIELD**

make	Age_Band_of_Driver	Sex_of_Driver	
ABARTH	16 - 20	Female	0.000000
		Male	1.464816
	21 - 25	Female	2.164354
		Male	3.119162
	26 - 35	Female	1.098612
		Male	4.084136
	36 - 45	Female	2.555679
		Male	1.809438
	46 - 55	Female	1.098612
		Male	2.772589
	56 - 65	Male	0.693147
	66 - 75	Female	1.798150
		Male	0.000000
ACCESS	21 - 25	Male	0.000000
	36 - 45	Male	1.000000
ACURA	26 - 35	Male	0.000000
ADLY	16 - 20	Male	1.000000
	26 - 35	Male	0.693147
	36 - 45	Male	1.000000
AJS	16 - 20	Female	2.009438
ZENNCO	16 - 20	Male	1.431946
ZHONGYU	16 - 20	Male	0.000000
ZNEN	16 - 20	Female	1.193147
		Male	2.329442
	21 - 25	Female	0.000000
		Male	2.641669
	26 - 35	Female	0.000000
		Male	2.204442
	36 - 45	Female	0.693147
		Male	1.464816