Introduction

Every year, there is a record of 1.25 million people's death due to road crashes on the average of 3,287 deaths per day as stated by the World Health Organization. Road accident fatalities and disabilities is by degree being identified as a crucial public health concern for many countries (Almohimeed, 2019).
In this report, United Kingdom 2019 data recorded and reported from all the road traffic crashes involving casualties and a greater number of other road traffic accidents that are non-fatal was used. These datasets used for this analysis are Accidents 2019 road safety data that contains the information of the accidents recorded in 2019, Casualties 2019 road safety data that contains the information of accidents casualties logged, Vehicles 2019 road safety data that contains information of vehicles involved in accidents and the Adjustment files which contains information of the government model used for the probabilities of dissimilar injuries that occurred for each accident.

This dataset contains date, time, location, number of casualties by severity and details of the vehicles and pedestrians reported to be involved in the road traffic accidents. Casualty and Vehicle dataset details can be connected to suitable accident by the field of Accident Index.

The aim of this project is to develop a model that would predicts accidents and the injuries that may be incurred, also to advise the government agencies about how the road safety can become better.

Data Loading and Pre-processing

To begin the analysis, United Kingdom 2019 data which consist of three files Accidents 2019 road safety data, Casualties 2019 road safety data and Vehicles 2019 road safety data was loaded, followed by gaining insights into the data and data cleaning.

Data insights: Data structure for the three dataset was obtained and it was observed that accidents dataset has 117536 observations with 32 columns, casualties dataset has 153158 observations with 16 columns and vehicles dataset has 216381 observations with 23 columns. Also, accidents dataset has missing values in Location_Easting_OSGR, Location_Northing_OSGR, Longitude, Latitude, Time, LSOA_of_Accident_Location columns. There are no missing values in casualties dataset and vehicles dataset.

Data Pre-processing: features such as time and date was transformed from object datatype to daytime format. Also, all the variables with object datatype in accident dataset, casualties dataset and vehicles dataset were converted to string.

Feature Engineering: To enhance the performance of the model, new features such as hour, minute, decimal time, day, sunrise, and sunset were created using timestamp. Furthermore, vehicle dataset feature (Accident Index) and casualties dataset feature (Vehicle Reference) was merged, after which the merged data was then linked to the accident dataset feature (Accident Index) for the purpose of building a predictive model.

Data Cleaning: Features that are not useful (such as Location_Easting_OSGR, Location_Northing_OSGR, and LSOA_of_Accident_Location) for the analysis was dropped because they have high cardinality. It was observed that columns with the same local authority district have the same local authority highway values, so the mean of the longitude and latitude of those columns with the same local authority district and local authority highway was pulled and used to fix the null values in longitude and latitude columns. The null values in Time column were replaced with the mode of all the values in the Time column. Also, the outlier detected in age of vehicles, number of vehicles and engine capacity was fixed with percentile.

Analysis

Questions related to the nature and characteristics of the accidents are answered below:

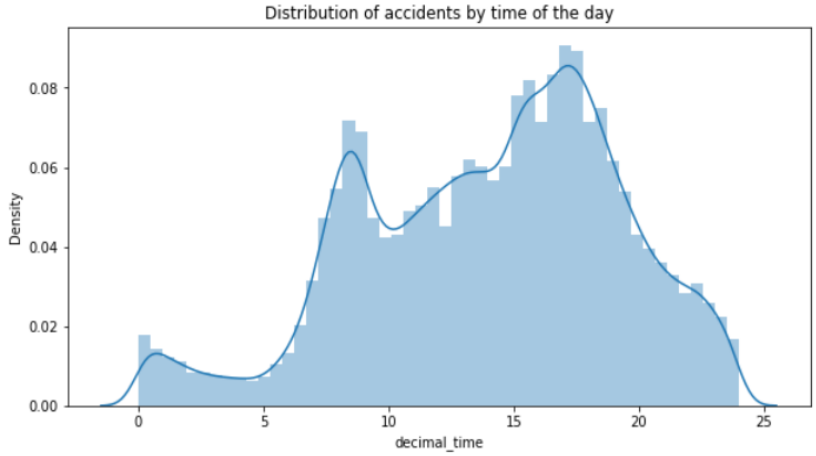


Fig 1

The graph above shows that the count of decimal time of the day which accidents occur mostly is during the rush hour period in the morning at 8:00 when people are going out. Also, there is significant portion in a day that accident do happen which is from 15:00 to 20:00 that can be classified under rush hour period in the afternoon when people are going back home.

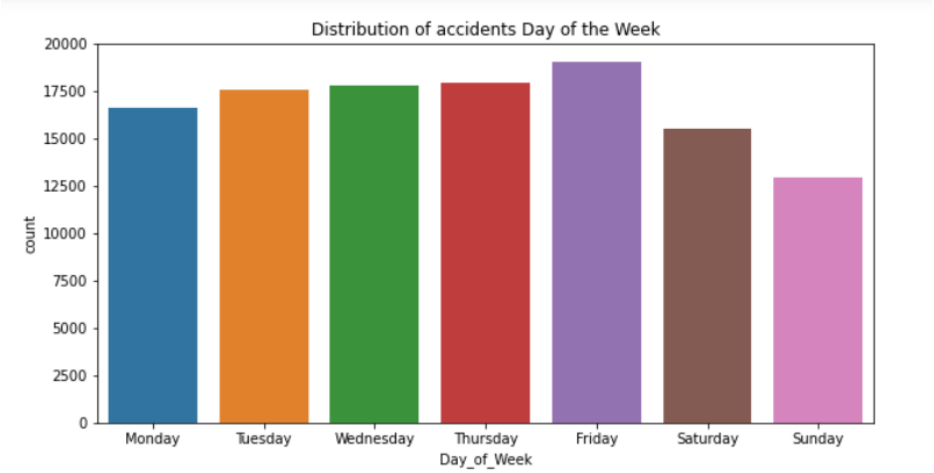


Fig 2

The graph above shows that the significant day of the week which accident occur mostly is on Friday.

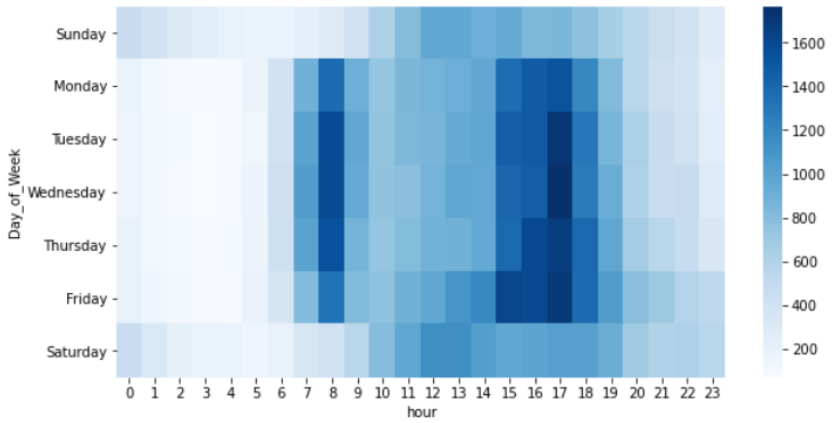


Fig 3: Heatmap showing days of the week and hours of the day which accident occur

The Heatmap above shows how dense the condition is and, in this case, accident occurs during rush hour period in the morning by 8:00 and during rush hour period in the afternoon between 15:00 and 18:00 with accidents occurring mostly at 17:00 from Mondays to Fridays. It can also be observed that there is high concentration of accidents happening on Fridays from 15:00 to 17:00.

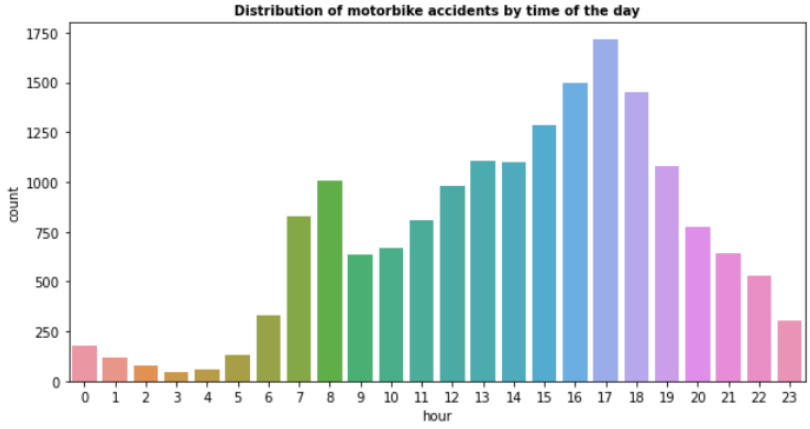


Fig 4

The graph in Fig 4 shows the hours of the day in which motorbikes accidents occurs mostly. This falls within the range of 15:00 to 18:00 in afternoon rush hour period with 17:00 being the most time motorbikes accidents occurs.

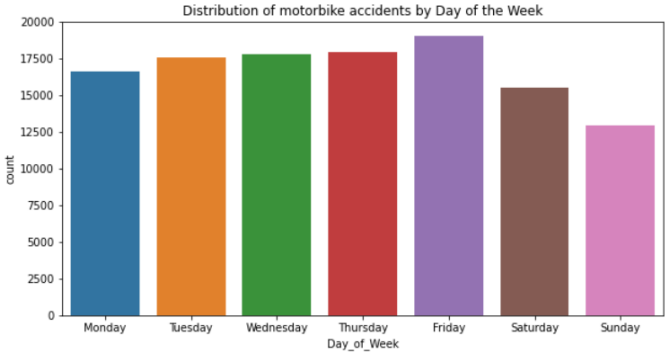


Fig 5

Day_of_Week	count	of %
Friday	2838	0.163395
Monday	2275	0.130980
Saturday	2312	0.133111
Sunday	2213	0.127411
Thursday	2609	0.150210
Tuesday	2524	0.145316
Wednesday	2598	0.149577

Table 1

In Fig 5, It can be observed that motorbike accidents occurs mostly on Fridays, also when the percentage of the motorbike accidents was calculated, The day with the highest percentage (0.163) of motorbikes accidents is Friday. The significance of motorbikes accidents in week days using Shapiro-Wilk Normality Test shows that the pvalue is 0.578 which is greater than 0.05 meaning that motorbikes accidents in days of the week is probably gaussian which is not Significant. However when the significance of motorbikes accidents in time of the day using Shapiro-Wilk Normality Test was done, the result shows that the pvalue is 0.00 which is less than 0.05 meaning that motorbikes accidents in hours of the day is probably not gaussian which means that there is significant hours of the day in which motorbike accident occurs.

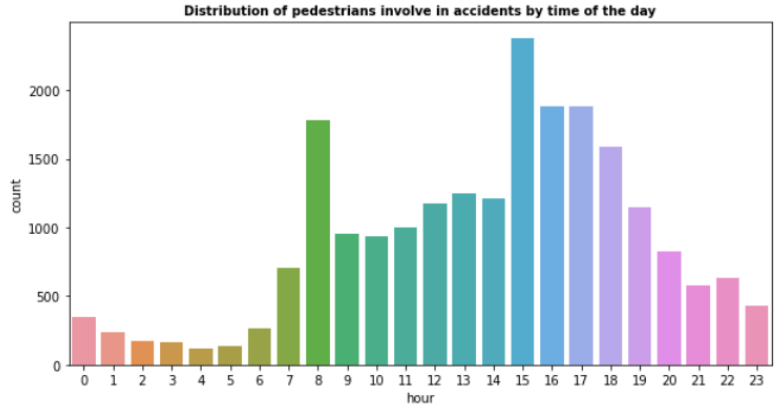


Fig 6

Day_of_Week	count	of %
Friday	3649	0.167616
Monday	3132	0.143868
Saturday	2839	0.130409
Sunday	1964	0.090216
Thursday	3482	0.159945
Tuesday	3332	0.153055
Wednesday	3372	0.154892

Table 2

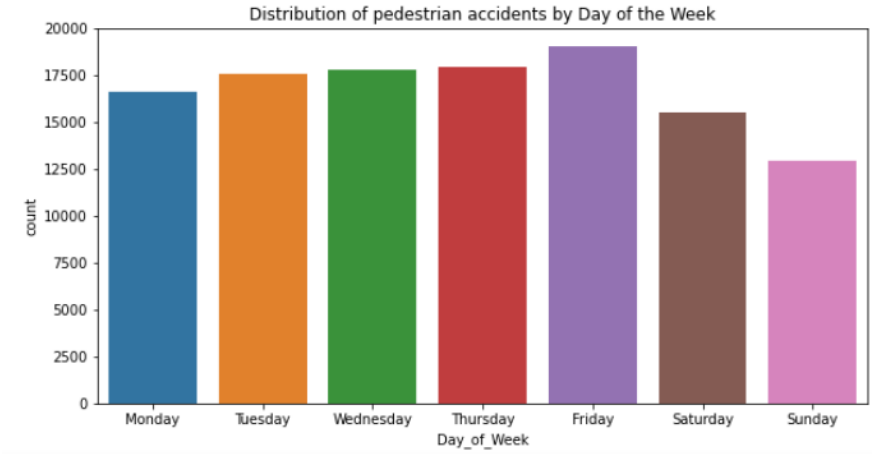


Fig 7

The graph above shows the hours of the day in which pedestrians involve in accidents mostly, it falls within the range of 15:00 to 18:00 with 15:00 (in the afternoon rush hour period) being the most time pedestrian involves in an accident. Also, in 8:00 (during the rush hour period in the morning). The day with the highest percentage (0.168) of pedestrians involve in accidents is Friday. The significance hours of the day which pedestrians involve in accidents using Shapiro-Wilk Normality Test shows that the pvalue is 0.000 meaning that hours of the day which pedestrians involved in an accident is probably not gaussian which is Significant. Also, the significance of days of the week pedestrians involves in accident using Shapiro-Wilk Normality Test shows that the pvalue is 0.381 which is greater

than 0.05 meaning that day of the week which pedestrian involve in accidents is probably gaussian which is not Significant.

To check the impact of daylight saving on road traffic accidents, student's t-test was used to compare week before and week after daylight started, to confirm if there is a significance increase of daylight saving in road traffic accidents, and it was observed that there is no significance difference with p value = 0.732, therefore it fails to reject H0. Also, student's t-test was used to compare week before and week after daylight ended to confirm if there is a significance increase of daylight saving in road traffic accidents, and it was observed that there is no significance difference with p value = 0.364, therefore it fails to reject H0.

To check the impact of sunrise and sunset times on road traffic accidents, student's t-test was used to compare the number of accidents 20 minutes to sunrise time and number of accidents 20 minutes past sunrise time to check if there is significance difference in sunrise in relation to road traffic accident and the result shows that there is no significance difference with pvalue of 0.578. Also, student's t-test was used to compare the number of accidents 20 minutes to sunset time and number of accidents 20 minutes past sunset time with pvalue of 0.810. It fails to reject HO.

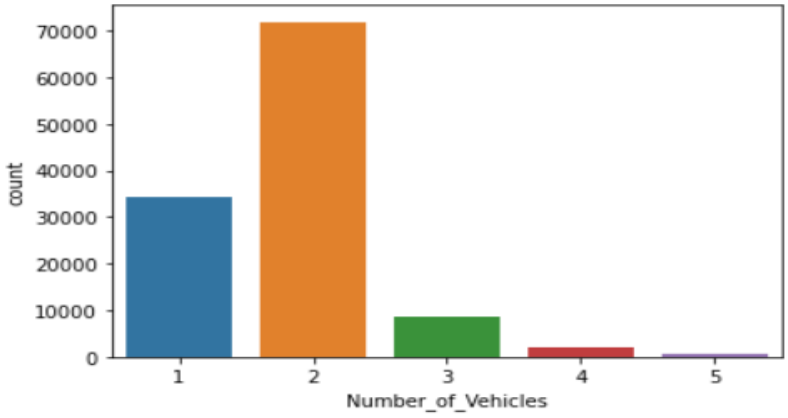


Fig 8

The graph in fig 8 shows that the highest number of accidents occurs between collision of two vehicles.

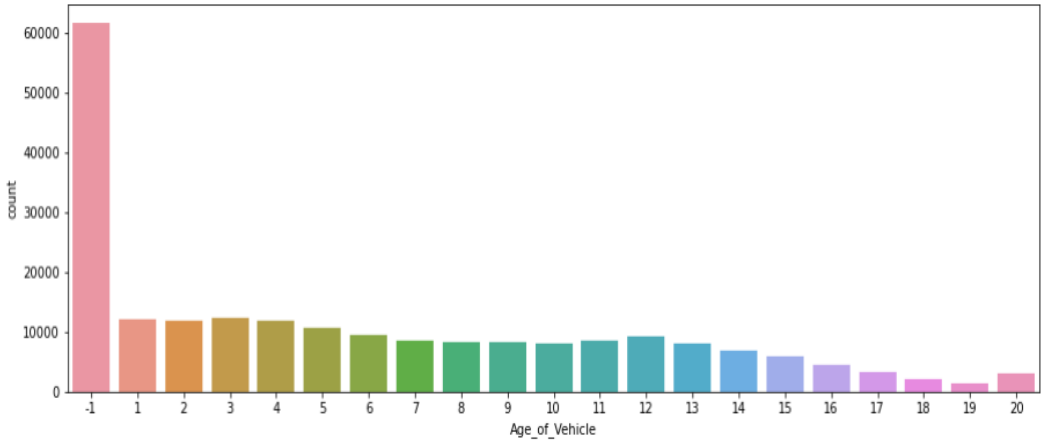


Fig 9

This shows how long each of these vehicles has been used with maximum age of vehicles reduced to 20years.

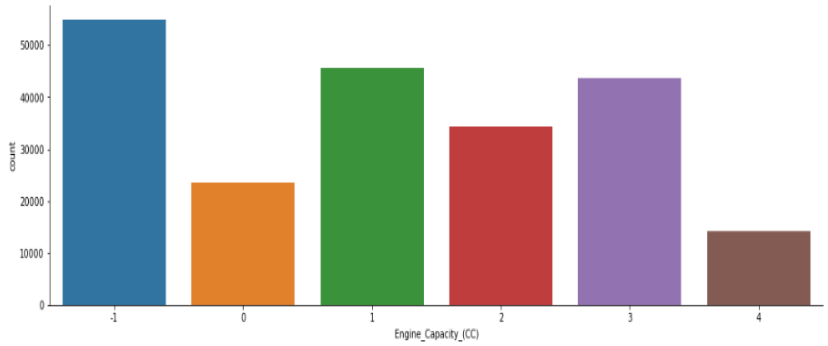


Fig 10: Engine capacity of Vehicles

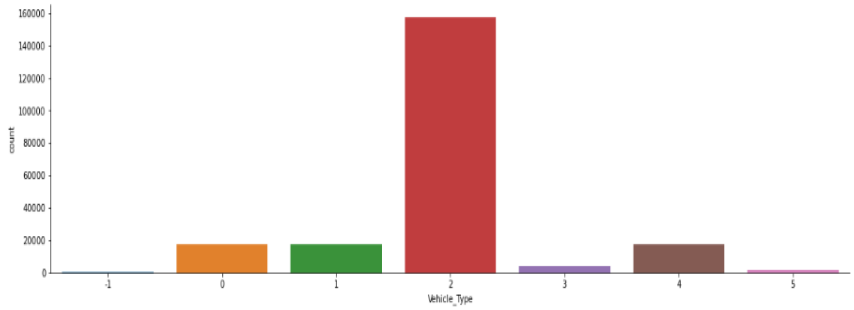


Fig 11: Vehicle types based on categorization

To determine the type of vehicles frequently involve in accident, different vehicles were categorized and the vehicle types after categorization shows that the vehicle type that most frequently involve in accident is category two which is Cars.

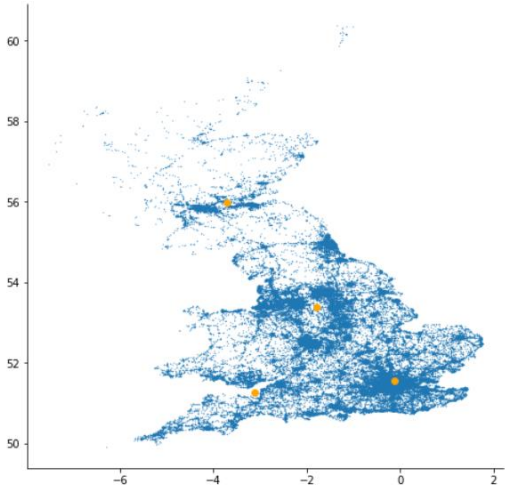


Fig 12: Dot distribution map of road traffic accident in the UK

As seen in Fig 12, Dot distribution map was used to show major locations where road traffic accident occurs in the Uk. Four cities such as London, Cardiff (Wales), Manchester, Scotland (Edinburgh, Glasgow) are areas with higher density of accident (i.e., accident hotspot) indicated with an orange dot when compared with other cities in the UK.

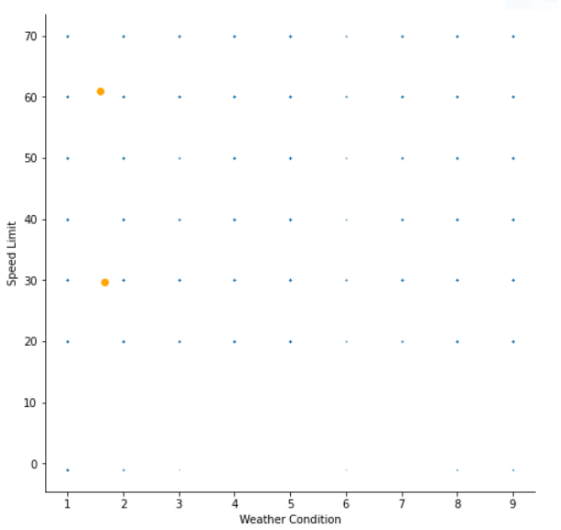


Fig 13

		Counts
Weather_Conditions	**Speed_limit**	
1	**30**	54682
	60	11226
	20	9158
2	**30**	8741
1	**40**	7990
	70	5448
	50	3756
9	**30**	2401
2	**60**	2028
8	**30**	1780
2	**20**	1399

Table 3

The graph above shows the weather conditions where speed limit is clustered. This means that there are more accidents under the weather condition 1 with speed limit of 30 and 60.
Table 3 confirms the weather conditions where speed limit is clustered and shows that the highest number of accidents occurs under weather condition 1 which denotes that the weather is fine and no high winds with speed limit 30 and 60 having a total count of 54682 and 11226 accidents respectively.

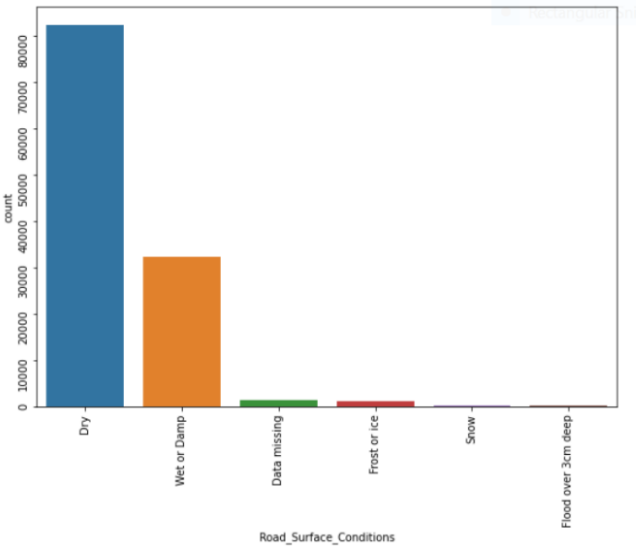


Fig 14

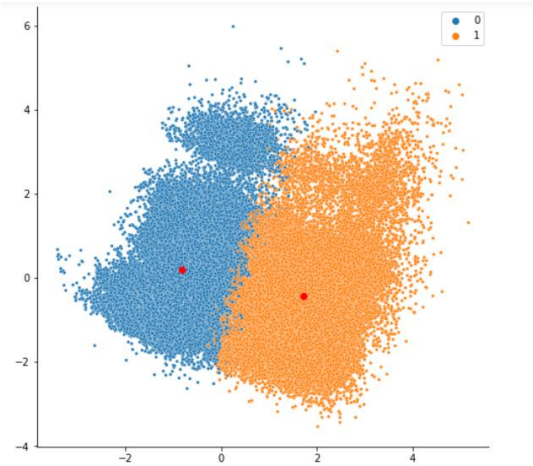


Fig 15

The graph in Fig 14 shows that the road traffic accidents occur mostly on a dry surface. In Fig 15, PCA clustering was done on some selected features such as Road surface conditions, Junction detail, day of the week, weather conditions, speed limits etc to show some conditions where accidents occur mostly.

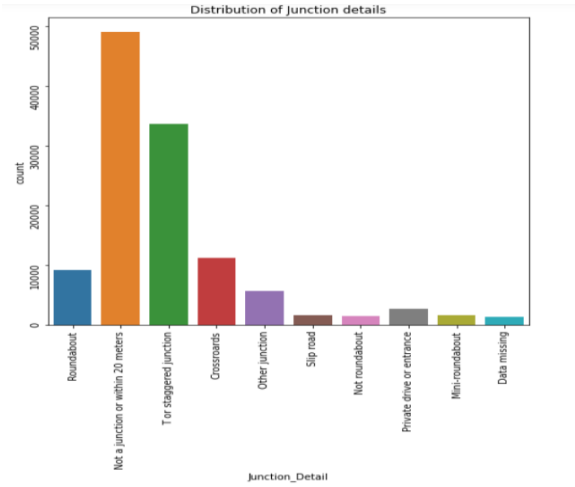


Fig 16

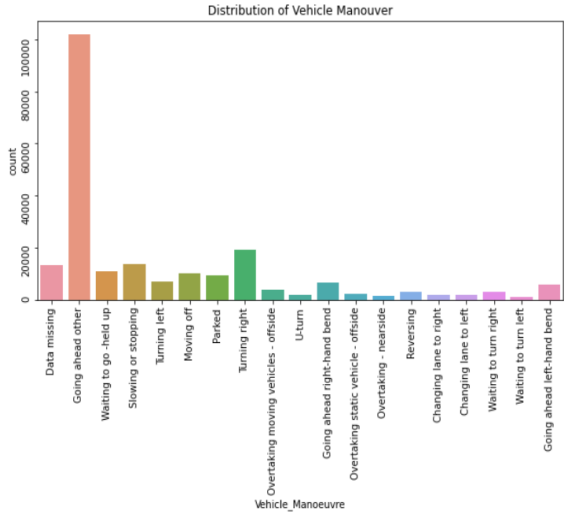


fig 17

The graph in fig 17 shows that vehicles going ahead of others have highest number road accidents.

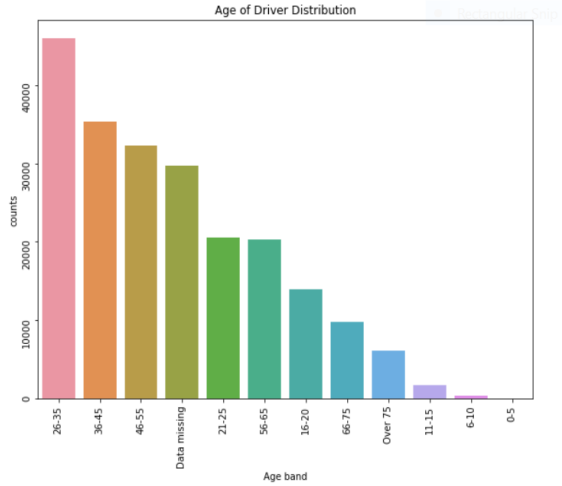


Fig 18

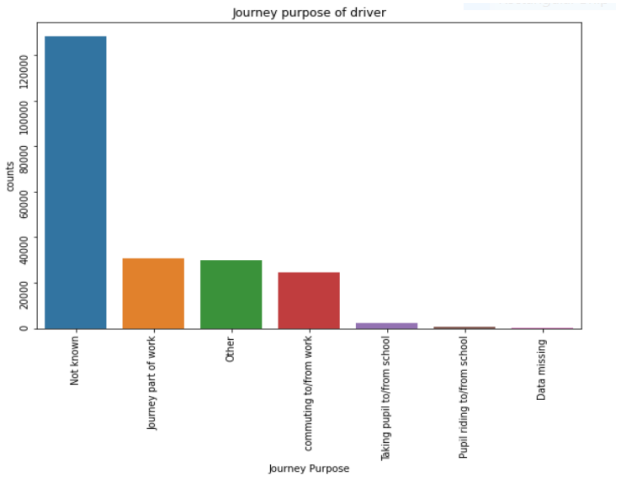


Fig 19

The graph in Fig 18 shows that the age range of drivers that is mostly involve in road traffic accidents is between 26-35 years. While the graph in Fig19 shows that the Journey purpose of the drivers that involves in a road traffic accident is not known.

Predictions

To carry out the predictions, accident data, casualties data and vehicles data were merged. Select kbest was used to select the best 20 features that are important in predicting outcomes, the merged dataframe was split into 75% train and 25% test data with x being the best features extracted and y (targeted column) being the casualty severity. Multiple classification algorithms such as KNeighbours, Decision tree, Logistic regression, Naïve bayes and Stacking classifier were trained using cross validation and balanced accuracy in which the results were displayed in box plot. Decision tree classifier have best accuracy with imbalanced dataset. The dataset was balanced using Smote. The negative numbers in the data were dropped, the data was balanced and retrained by passing it into the model to see if it will improve prediction in which the results are shown in the table below.

		Imbalanced Data accuracy	Balanced Data accuracy	Balanced Data accuracy without negative number in the merged data.
	Decision Tree	0.582	0.775	0.847
	KNeighbours	0.529	0.803	0.820
	Logistic regression	0.501	0.690	0.726
	Naïve bayes	0.563	0.625	0.682
	Stacking	0.509	0.842	0.887

The table above shows that Stacking classifier gives the best accuracy after the balanced data was retrained. So, Stacking classifier was retrained and accuracy of 0.782 was achieved. Confusion matrix and classification report for the best model is shown below.

```
[[ 535 1295]
 [1308 8801]]
              precision    recall  f1-score   support

           0       0.29      0.29      0.29      1830
           1       0.87      0.87      0.87     10109

    accuracy                           0.78     11939
   macro avg       0.58      0.58      0.58     11939
weighted avg       0.78      0.78      0.78     11939
```

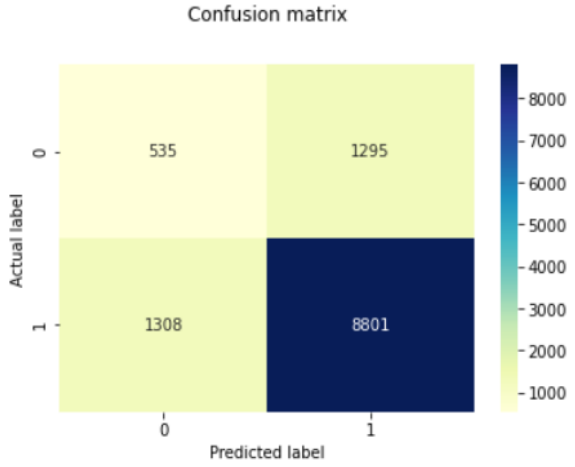


Fig 20

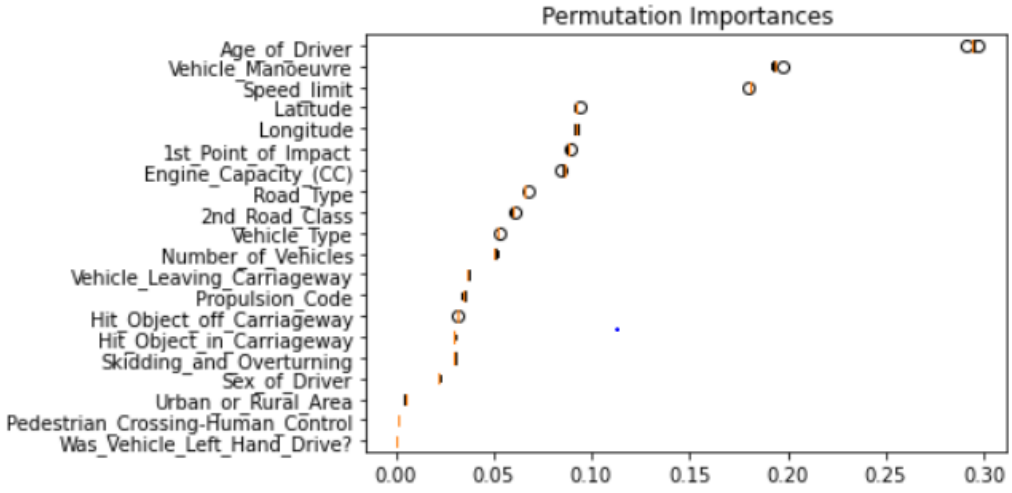
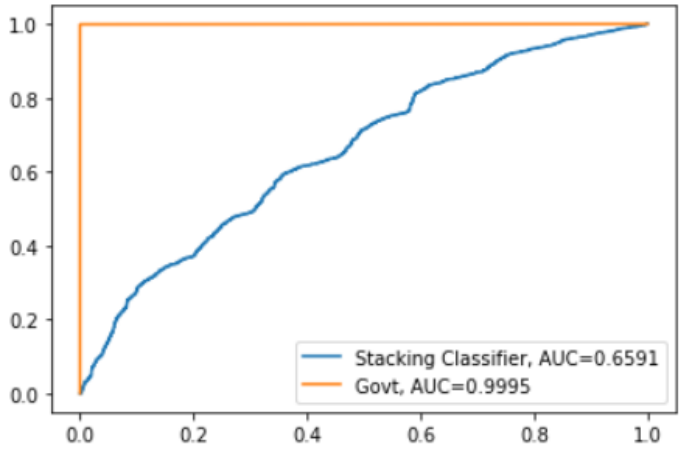


Fig 21: Feature Importance

The graph above shows that age of driver, vehicle manoeuvre and speed limit are significance features on road traffic accidents.



When the stacking classifier was compared with government model, AUC curve was created for the best model (Stacking classifier) and the government model. The Staking AUC value is lower than Government AUC Value.

Recommendations

- It can be suggested that government should construct new roads to reduce and divert traffic from congested areas and increase visibility such as road signs, markings etc.
- Some insights drawn from the type of vehicles, age of vehicles, age of drivers, involved in accidents and movement of pedestrians, locations of accidents are important information's that law makers can use to understand the causes of accidents which help them to review their policy and provide solution to tackle the problems.
- To minimise road traffic accidents, measures such as alerting road traffic users against areas i.e., dangerous junctions etc prone to accidents, and organising drivers attitude programs should be put in place.
- Cities identified on the dot map distribution as areas with higher density of accidents should be investigated to find out the construction needs.
- Government should provide education to encourage all road users. Also, Safety features in vehicles should also be promoted.