Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

CrossRef DOI of original article: 10.34257/GJCSTDVOL21IS1PG1

Detailed Analysis and Identification of Key Factors Resulting in Motor Accidents across the UK

Harshita Garg[1]

[1] Birkbeck University, London

Received: 12 June 2021 Accepted: 1 July 2021 Published: 15 July 2021

Abstract

Motor accidents across the globe amount to a large number of deaths every year. The collisions result in not just the personal injury to people involved but also in the loss of money to the motor insurance companies, trauma to the people involved, and added pressure on the emergency services. With the help of data analytics techniques, this project aims to identify critical factors that might contribute to the accidents. Upon investigating the temporal features and geo-spatial features of the motor accident locations, we tried to establish a correlation between the accident intensity and its key factors. For this exploratory analysis, we also considered weather conditions and daily average traffic flow data. We then trained Supervised learning models on the data to find out the best performing multi-label classification model.

Index terms— supervised learning; accident analysis; multilabel classification.

1 Introduction

round the world, every year, more than 1.25 million people are killed and 50 million are injured in road traffic accidents. (Source -Express, road safety facts [1]) The source claims that "Every day, on average, five people are killed and 64 seriously injured on UK roads." Driving is considered the most dangerous activity we do every day.

Several factors contribute to road accidents. Some of these are -severe weather conditions, the distraction of driver, failure to give or understand appropriate signals, reduced motor skills due to old age, or alcohol consumption.

If there was a way to find the key factors responsible for motor accidents happening on the roads, lots of these effects could be minimized. If the hotspots for accidents could be identified, emergency services could be put on high alert in those areas, increasing the response time and potentially reducing the loss of life. If we can predict the likelihood of a crash in real-time, the driver could be warned of potential danger. The government can issue advisory to all the motorists on the accident's hotspots or put signboards to notify the road users.

2 II.

3 Literature Review

Accidents dataset for the UK region, which is available at the government of UK website [2], is an immensely popular dataset and many academicians have based their research on this, with some variations.

Jinning You et al. attempted to calculate the crash likelihood in [3]. They used web crawling techniques to obtain live weather data and oversampling to solve the problem of inherent imbalance in the dataset and applied random forest and SVM classifier algorithms on the training dataset. SVM classifier performed better for them when used with the web crawling techniques.

The relationship between road accidents and traffic on the roads has got a lot of attention in recent years. **??**alifu [4] used a similar approach for the accident prediction for unsignaled urban junctions in Ghana. He combined accident data with the Annual Average Data Flow and analyzed the effect on different kinds of junctions like signaled junctions, unsignaled junctions, T-junctions, X-junctions etc.

Traffic data visualization is another approach that researchers have studied extensively to discover patterns and make clusters amongst traffic accidents. In the research paper [5], authors Chen et al. state that "Data visualization is an efficient means to represent distributions and structures of datasets and reveal hidden patterns in the data. "

This project builds upon many of the approaches described above and draws a parallel with the model developed by You et al [4] but is different in the sense that it involves not only the accident, and traffic data but also the detailed demographics of the driver and the vehicle involved.

4 III.

5 Secondary Data

I obtained the data for accidents from the government of the UK website [2]. Statistics on road safety in Great Britain are based on accidents reported to the police in a form submitted by the attending officer.

To quantize the accident severity, many factors were considered. One of the significant variables for this model was the volume of traffic flowing on the road at the accident time. Taner J.C. [6] explained that the traffic volume and crash data follows the model Y = ?F?,

6 A

Author: Birkbeck University, London. e-mail: Harshita.garg@hotmail.com where Y is crash count, F is traffic volume, and ? and ? are calibration coefficients. In other words, the crash count is directly proportional to the amount of traffic on the road. Annual average daily flow(AADF) data is available on the government of the UK website [7]. This dataset gives the estimated annual average of the flow of traffic on most of the major and minor UK roads.

The data for the vehicles involved in road accidents is from the same source as the accidents dataset [2]. Vehicles dataset includes the details of the vehicles involved in accidents.

IV.

7 Methodology a) Data Preparation

Many columns in the dataset had missing values. Columns with more than 20% missing values were dropped. We also decided to drop the features that were not considered important in the classification problem at hand. After combining the Accident dataset with vehicles and the AADF table, many records for AADF were found to have missing values. The missing data was because not all the accident spots had AADF values available. This trend was more common in the minor roads, mainly B, C, and U roads. The final data frame had nearly 50% of values missing.

We created a linear regression model to calculate the value of traffic based on the variable's latitude, longitude, road class, and road type. All the records with a valid AADF value in the combined data frame were used as the training dataset, and all the records with missing AADF values were used as a test dataset. Performance of the model was about 70%, which was okay.

The machine learning models try to derive a meaningful relationship between the features present and the target variable. The ability of a model to predict the outcome successfully depends mainly on the types of features present in the dataset. This is where feature engineering comes into the picture. We engineered different features from the existing ones to increase the predictive powers of the models. We converted Hour of the day into a cyclic feature such that hour 0 is closer to hour 24. Data distribution after conversion of time into cyclic feature is plotted below. Mean encoding is encoding categorical features based on the ratio of occurrence of positive class in the target variable. For the problem at hand, the target variable is Accident Severity, and the positive class is the 'fatal' class. Thus, we converted the categorical variable 'road name' to mean encoded value which better represented the target variable accident severity. Two problems were solved here in one go -categorical variable with an unmanageable number of levels was converted to a quantitative one, and the target values were embodied into the feature, thus increasing the predictive power of the model.

8 b) Exploratory Data Analysis

A layered analysis was done for the exploratory variables to fully understand the dataset and the impact every variable had on the severity of accidents. We plotted the distribution of the number of accidents concerning some predictor variables and accident severity. The first graphs show that more accidents tend to happen on the weekdays rather than on the weekends. A maximum number of accidents seem to take place on day 6(Fridays). The third plot shows that most accidents happen on road type 6, which stands for single carriageway. According to the last graph, most accidents happen at junction 0(not a junction) and 3(T or staggered junction).

The second plot indicates that maximum accidents occur on A roads, followed by the unclassified 'U' category roads. Also, the maximum number of fatal accidents happen on the A roads.

 10.34257/GJCSTDVOL21IS1PG1

We can identify accident hotspots by doing the geospatial analysis of accidents data. A number of accidents was plotted on the UK map based on their location information, and we obtained the following plot. The above plot shows that the ratio of pedal cycle, two-wheelers and buses and coaches is centered more towards 0, suggesting that there are fewer roads that have a high distribution of these vehicles on average. The distribution of ratios for large goods vehicles(LGVs) is between 0 to 0.3, and that for heavy goods vehicles (HGV) is between 0 and 0.2. We get the maximum ratio for the cars and taxis(between 0.6 to 1.0), which is the trend that one would typically expect on any UK road.

We visualized the distribution of accidentseverity concerning the age and sex of the driver available in the vehicle dataset. Here 1(green) represents male drivers, 2(blue) represents female drivers, and 3 is unknown gender. This graph clearly shows that male drivers are more likely to be involved in motor accidents than female drivers. The graph peaks at the age band 6, which represents the age range 26-35 years, showing that this age range is more likely to be involved in accidents than the other age bands.

9 c) Modeling

After the exploratory analysis of the dataset, some models were created in Python and evaluated for their performance. Before starting the modeling process though, some important decisions were taken.

10 Choice of Metric:

A classifier is only as good as the metric used to evaluate it. A wrong metric misleads into believing that the classifier is working fine. Standard performance metrics treat all the classes in a multiclass problem as equally important. Whereas, in imbalanced classification problems, minority classes are often more important than the majority classes.

Following the general guidelines given by Jason Brown in his book [8], we decided to use F2-measure as a metric for model evaluation. In this case, False Negatives are more costly than False Positives. Meaning that if there is a likelihood of an accident happening at some location and we reported as negative (False Negative), it could be dangerous. On the other hand, if there is less probability of accidents happening and is flagged as an accident (False Positive), it was okay because it would warn the driver to be more cautious while driving. We calculated F2 measure as Generalization of F-beta score is calculated with the value of beta being equal to 2. Beta value 2 means that more emphasis is given on Recall than Precision.

Spot Checking the Algorithms: Spot-checking machine learning algorithms means evaluating a suite of different machine learning algorithms with minimum hyper tuning. Thus, giving each algorithm a fair chance to perform under comparable conditions. Spot-checking helps us decide which algorithms to use for the final model. We used the following framework for spotchecking: -Linear Algorithms: We checked the following linear algorithms.

11 ? Logistic Regression

? Linear Discriminant Analysis ? Naïve Bayes Non-Linear Algorithms: Nonlinear algorithms tend to perform better when the problem is inherently non-linear.

12 ? Decision Trees

? Support Vector Classifier Ensembles: Ensembles are the group of algorithms, whose predictions are combined to give a better performance. Models tested here were:-? Random Forest ? Bagging ? Adaboost Sampling: Sampling is the process that attempts to reduce the class imbalance by decreasing the number of samples in the majority class(also called undersampling) or by increasing the number of samples in the minority class(also known as over-sampling). Cost-sensitive learning: Normal algorithms treat all the classes as equal. We can change this trend by enforcing cost-sensitive learning, in which we applied a cost to penalize the model if it does not predict the minority class correctly.

V.

13 Model Evaluation and Testing a) Linear and Non-linear Models

All the linear, non-linear, and ensemble models were trained on the training set using the 10-fold crossvalidation method. The models were beyond the computing capacity of the laptop they were training on. Hence we decided to do the training on the cloud "floydhub". Floydhub is an extremely easy to use and intuitive platform for running python scripts on the cloud. We then recorded the average of the F2-scores and standard deviation. After comparing the Precision and Recall values and the overall weighted f2-score,we decided to investigate the final four models further -Linear Regression, Naïve Bayes, Random Forest, and Adaboost.

14 b) Sampling Methods

The distribution of observations across different classes (accidents severity):-1 2 3 3382 41947 231842 The above table showed that the data distribution was highly skewed amongst the three classes with the 83% of the total accidents belonging to class 3(mild), nearly 16% belonging to class 2(serious) only 1% of the accidents belonging to class 3 that represents the fatal accidents.

Most machine learning algorithms are designed such that they perform the best if trained on the problems with equal class distribution throughout the dataset. When this is not the case, models learn to conclude that very few minority class instances exist. Hence, they are not critical and can be ignored. But this is far from true.

For this project, we investigated under-sampling methods and the combination of under-sampling and over-sampling method. In the combination method, we oversampled Class 1 using SMOTE(Synthetic Minority Over Sampling Technique) by the ratio of 4. The other two classes were under-sampled using random undersampling. The ratio of the three classes 1:2:3 was 4:0.8:0.4. In this technique, the number of samples of minority class was increased, and that of majority class was decreased, while maintaining the imbalance, thus training the model on more realistic data.

After the sampling, we trained the four bestperforming models on this data and tabulated their results. Sometimes the training error gives optimistic results, but the model does not perform well on the test dataset. Hence we also tested these models on the test set, and included their F2 scores in the table.

15 c) Merging two classes

The classification problem that we were trying to solve here is a multiclass classification with three classes-Fatal, Serious, and Mild. For the accident dataset, accidents that involved deaths were defined as fatal accidents and accidents that involved a serious injury to the driver or passengers were classified as severe accidents. From a driver's point of view, whether he gets a red warning for a fatal accident or an amber alert for a severe accident, it should not make much of a difference. Moreover, in the multiclass classification models trained above, we saw that most of the classifiers ignored class 2. And for class 2, the recall value was relatively low in most of the models examined.

One way to deal with this problem was by combining the classes fatal and severe. The multiclassification problem now became a binary classification problem with two classes minority(serious + fatal) and majority(mild).

We then trained the best performing algorithms, chosen earlier on the data with two classes and we plotted their performance in following box and whisker plot.

16 Conclusion

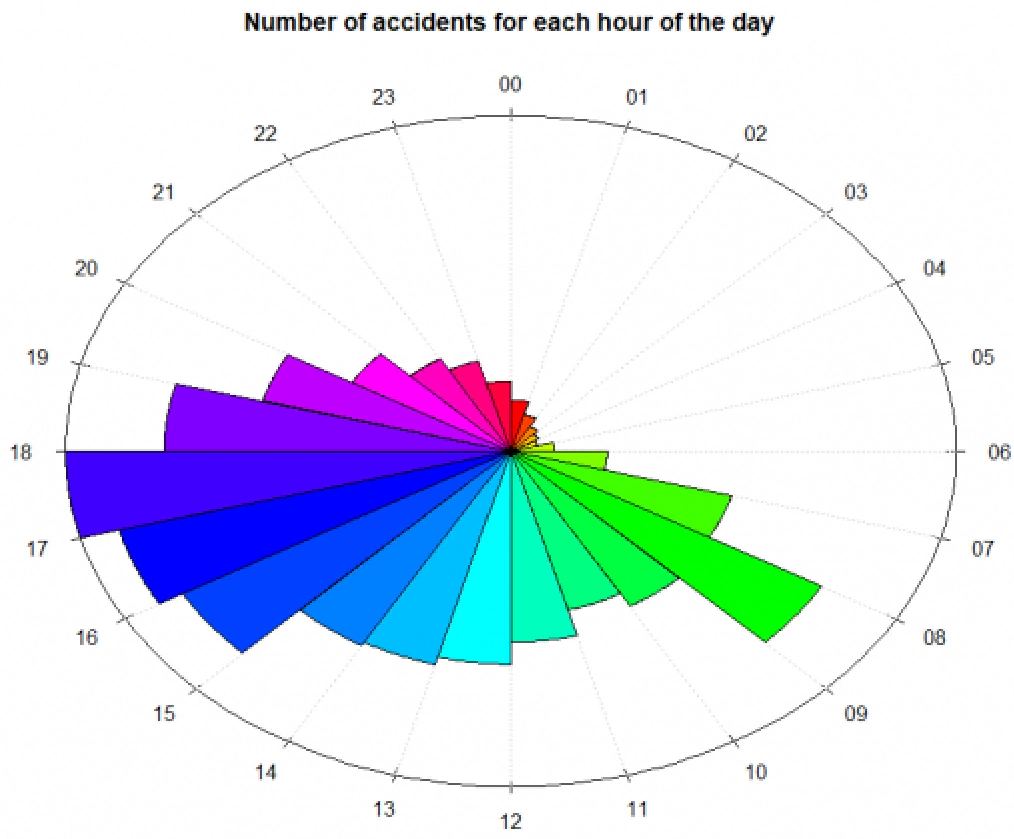
This project attempted to identify the key factors responsible for motor accidents happening across the UK and created models to correctly classify the accidents by their severity level. The historical records of accidents datasets were analyzed to understand the trends and to see if any critical factors could be identified while classifying accidents into 3 different classes-Mild, Serious, and Fatal. Different types of predictor variables were analyzed concerning the frequency of accidents. The variables included temporal variables like time of the day, month etc. A strong correlation was found between the time of the day and the number of accidents.

Geo-spatial factors were studied to see if they contribute to the severity of the accidents. A graph between the road class and accident severity revealed that the maximum number of accidents happen on Aclass roads and not on the motorways, where the speed limit is usually more. Weather data, which was initially thought to be an important contributor in accidents, surprisingly did not emerge as a critical factor. More than 80% of the accidents happen on bright days with no heavy rains/ snow.

Coming to the question -' What are the key factors responsible for accidents?' On examining the feature importance of our chosen random forest model, the following plot was obtained. From the features point of view, the geographical position is an essential feature in determining the accident probability. The latitude and longitude values were used to find the accident hotspots across the UK. The traffic flow data was the third most crucial feature in classifying accidents. Some of the engineered features proved to be particularly important from a classification point of view. Time features like month; day of the week proved to be important as well.

Accident prediction is inherently a difficult problem to solve and this project is a small step forward in facilitating progress on the same. With the systematic approach presented here, we introduced a model that gave promising results and classified accidents with an excellent f2 score estimate and a good recall score for the accident class. With more time, it would be a good idea to explore other possibilities for the ensembles in modeling the data. Ensembles generally tend to perform better than the individual classifiers. A warning system could be developed to warn the drivers in real-time using the model developed here.

 10.34257/GJCSTDVOL21IS1PG1



1

Figure 1: Figure 1 :

2

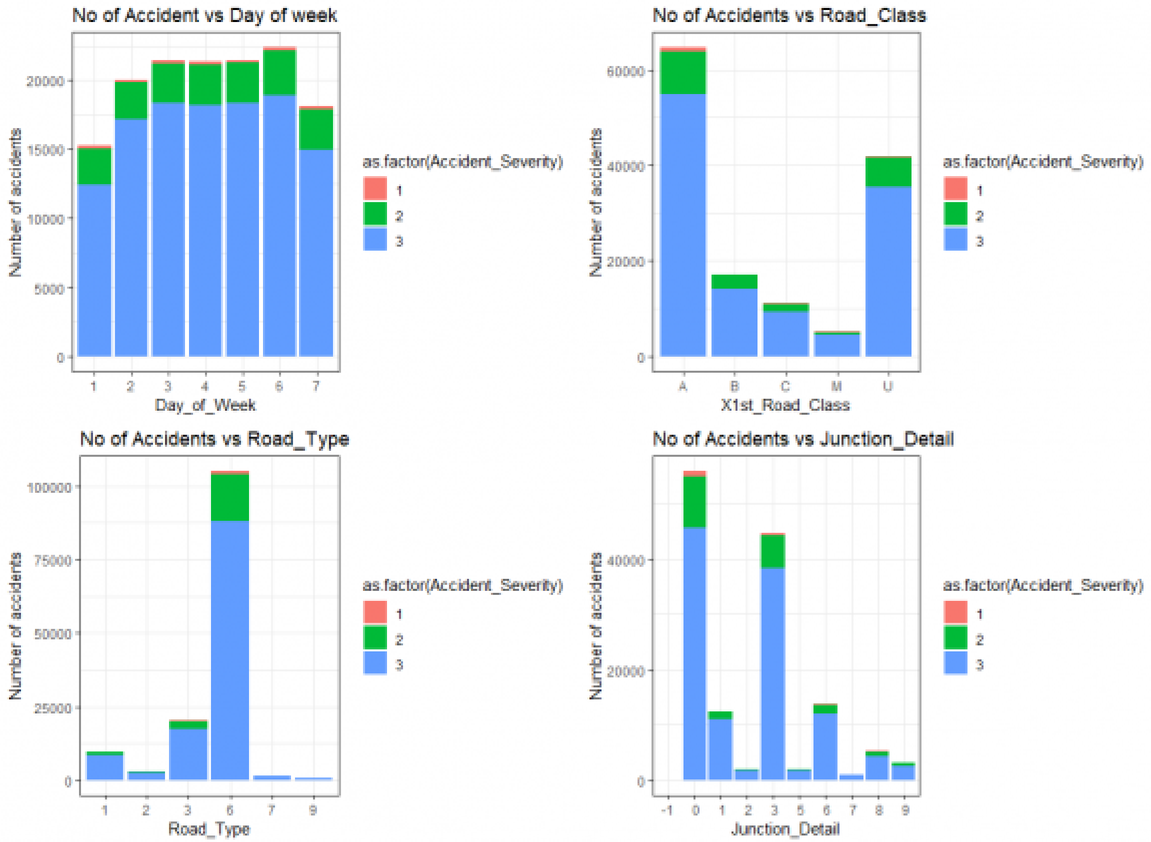


Figure 2: Figure 2 :

3

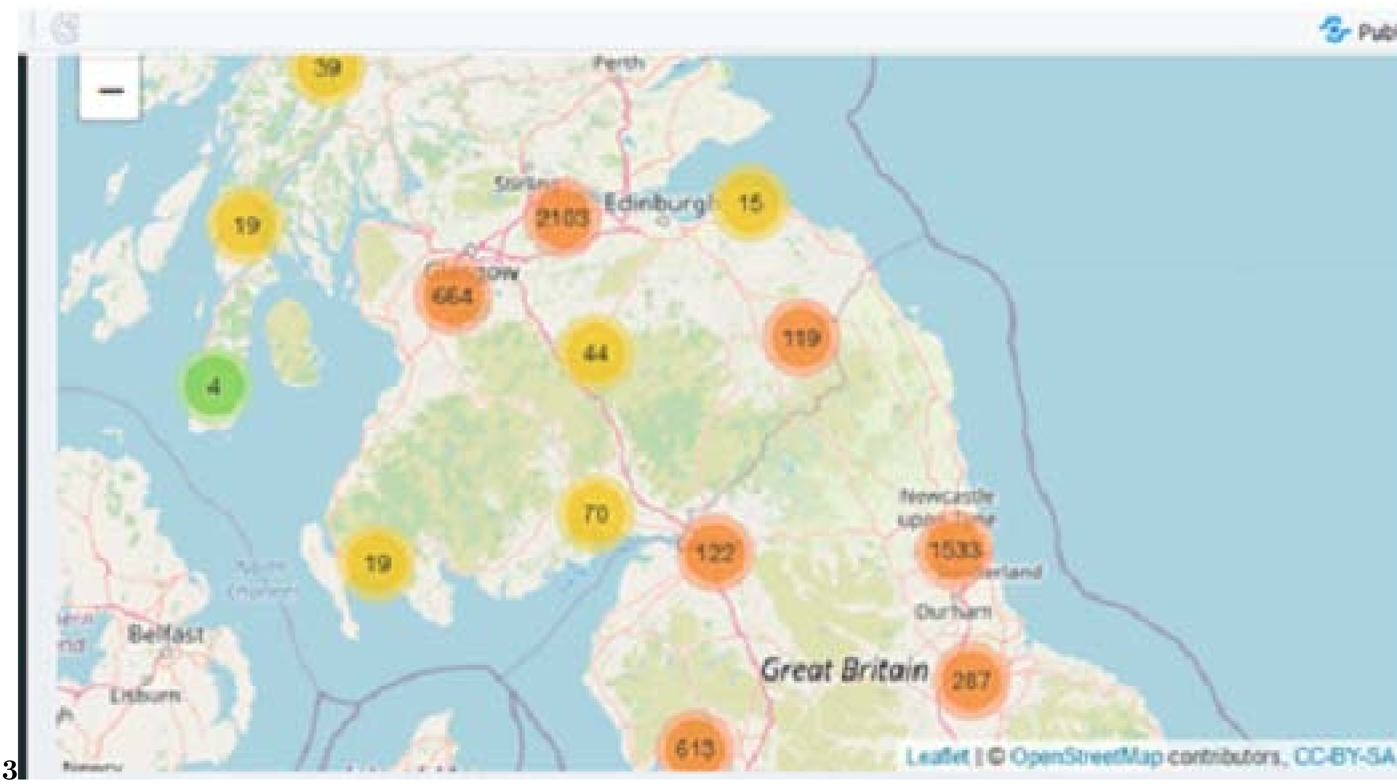


Figure 3: Figure 3 :

4

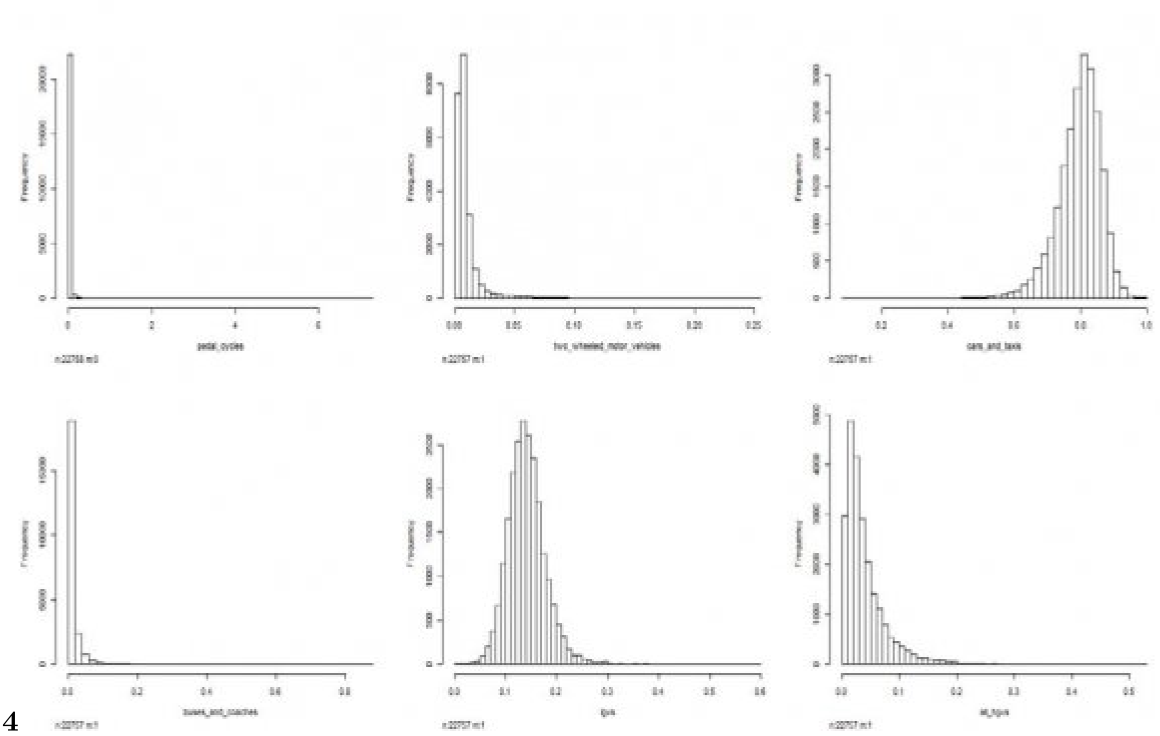


Figure 4: Figure 4 :

5

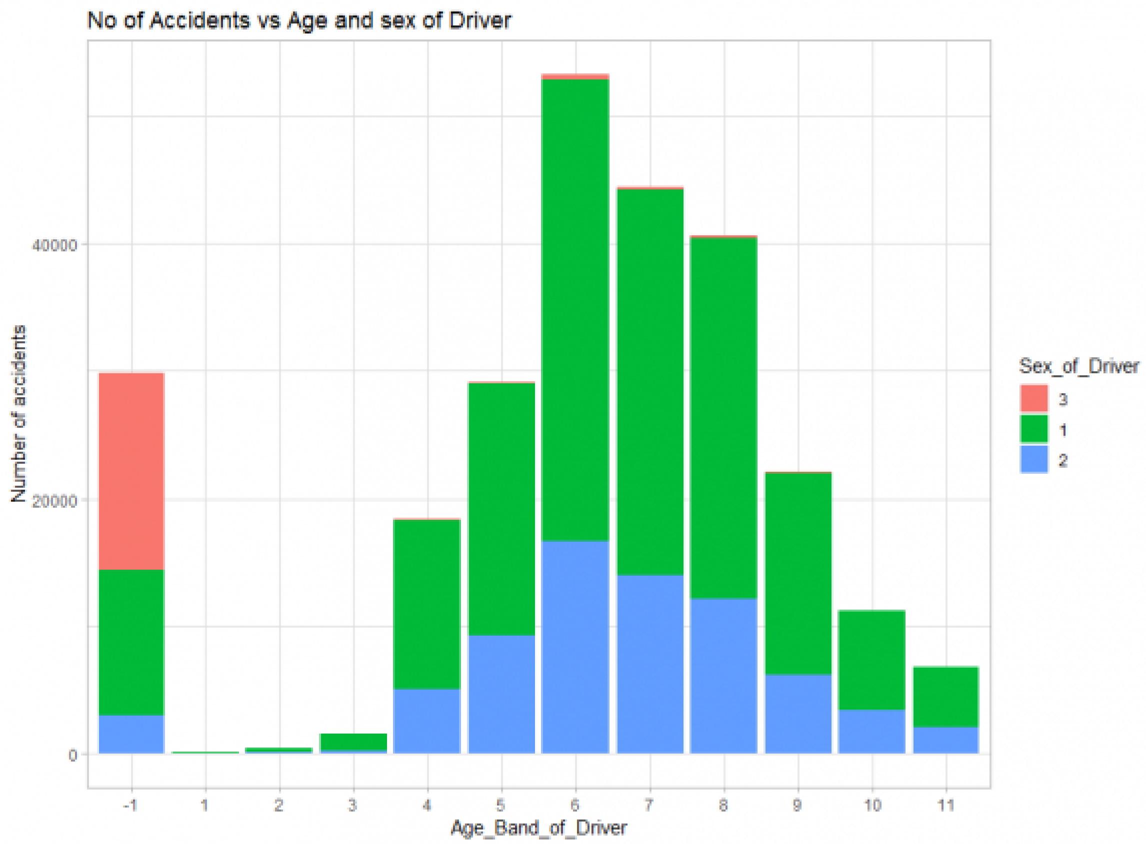


Figure 5: Figure 5 :

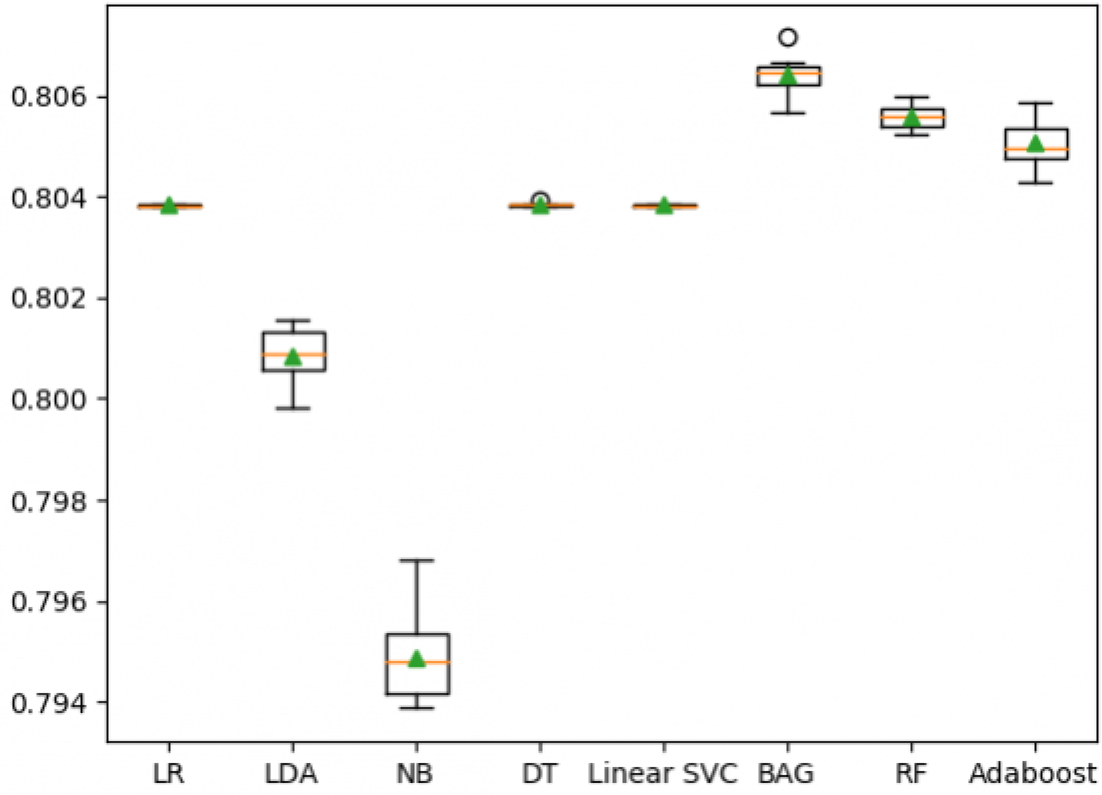
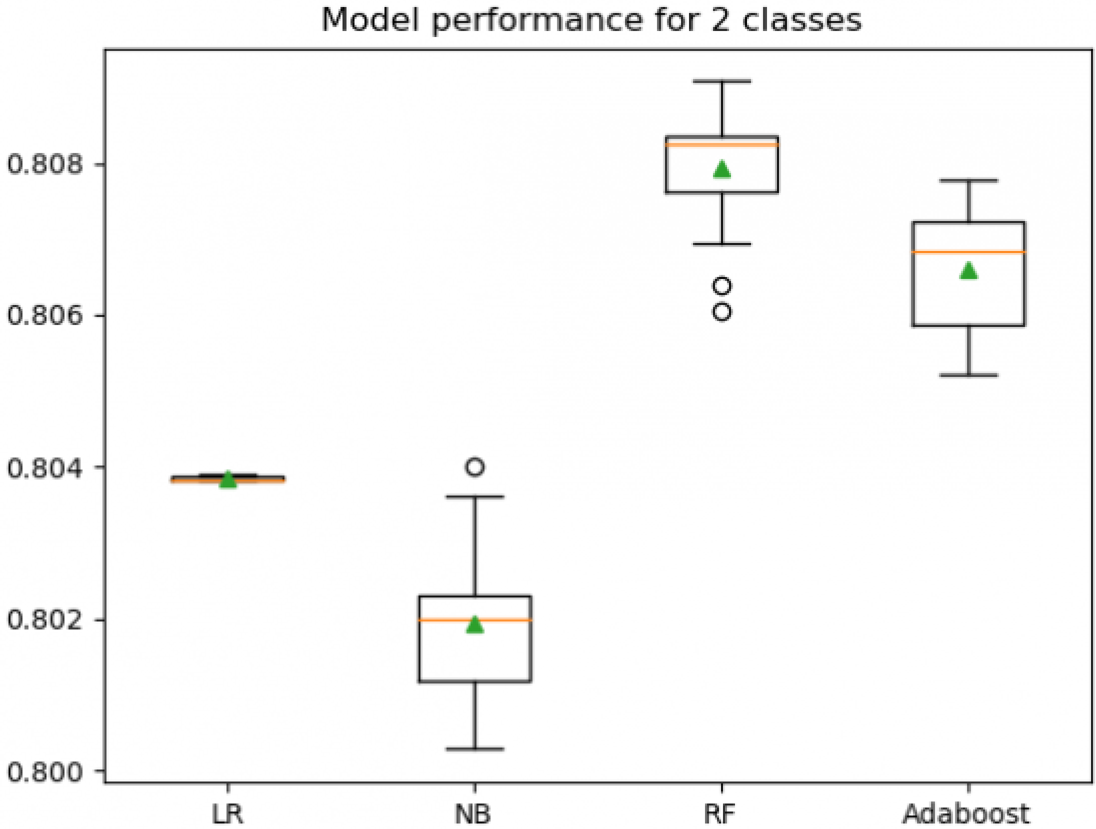


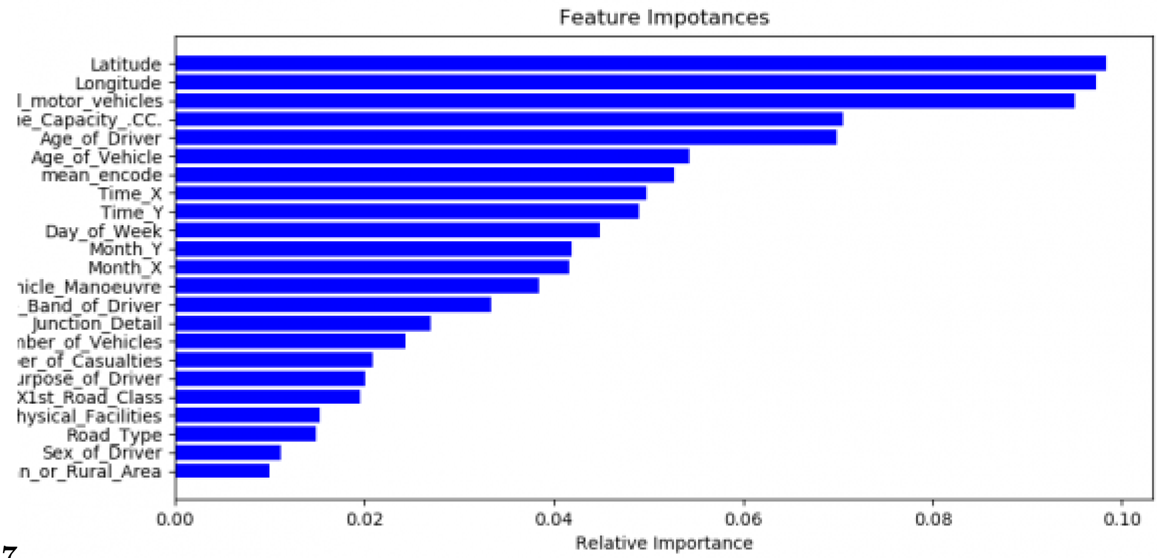
Figure 6:

 10.34257/GJCSTDVOL21IS1PG1



6

Figure 7: Figure 6 :



7

Figure 8: Figure 7 :

1

Name of the Model	Average of f2 Scores	Standard Deviation of scores	Time taken
Linear Regression	0.804	0.000	10s
Linear Discrminant Analysis	0.801	0.001	5s
Naïve Bayes	0.795	0.001	3s
Decision Tress	0.804	0.000	5s
Linear SVC	0.804	0.000	25s
Bag	0.806	0.000	16m48s
Random Forest	0.806	0.000	3m32s
Adaboost	0.805	0.000	3m41s
Performance of the above models was compared using a box and whisker plot.			

Figure 9: Table 1 :

2

Name of the model	Average of Training f2 scores	Standard deviation	Test F2 score
Under-sample LR	0.788	0.000	0.804
SMOTE LR	0.601	0.000	0.804
Under sample NB	0.779	0.001	0.794
SMOTE NB	0.608	0.002	0.778
Under sample RF	0.791	0.000	0.804
SMOTE RF	0.707	0.003	0.804
Under sample Adaboost	0.789	0.000	0.803
SMOTE Adaboost	0.683	0.002	0.802

Figure 10: Table 2 :

 10.34257/GJCSTDVOL21IS1PG1

[Chen et al. ()] *A Survey of Traffic Data Visualization*, Wei Chen , Fangzhou Guo , Fei-Yue Wang . https://ieeexplore.ieee.org/abstract/document/7120975 2010.

[Salifu ()] *Accident Prediction Models for Unsignalised Urban Junctions in Ghana*, Mohammed Salifu . https://core.ac.uk/download/pdf/82011865.pdf 2003.

[Tanner ()] *Accidents at rural three-way junctions*, J C Tanner . P -56-67. 1953. (Journal of the Institution of Highway Engineers 2)

[Department For (2019)] *GB Road Traffic Counts*, Transport Department For . https://data.gov.uk/dataset/208c0e7b-353f-4e2d-8b7a-1a7118467acc/gb-road-traffic-counts May 2019.

[Hartson (2016)] William Hartson . https://www.express.co.uk/life-style/top10facts/622407/Top-10-facts-road-safety *Top 10 facts about Road Safety*, Feb, 2016.

[Brown ()] *Imbalanced Classification with Python*, Jason Brown . 2017. p. .

[You et al. (2017)] *Realtime crash prediction on freeway and emerging techniques*, Jinning You , Junhua Wang , Jingqui Guo . https://www.researchgate.net/publication/316489169_Real-time_crash_prediction_on_freeways_using_data_mining_and_emerging_techniques April 2017. (Journal of modern transportation)

[Department For (2019)] *Road Safety Data*, Transport Department For . https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data December 2019.