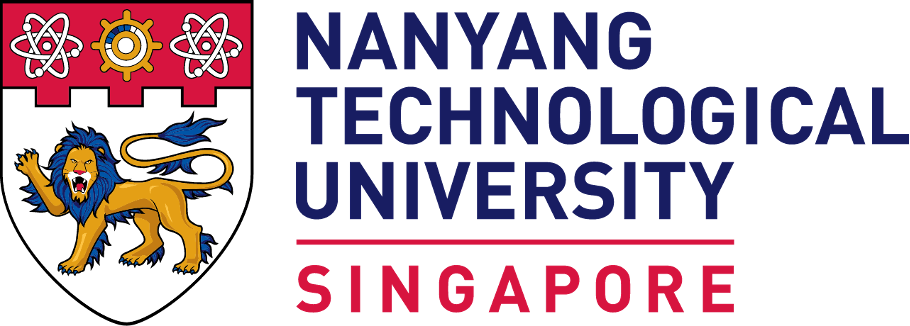
****

**BC2407 – ANALYTICS ll: ADVANCED PREDICTIVE TECHNIQUES**

**GROUP PROJECT REPORT**

**A Holistic Approach to Predictive Road Accident Severity using Machine Learning Paradigms**

**Seminar Class 2, Team 6 (Monday - 2.30PM)**

*Prepared for: Prof. Neumann Chew*

|  |  |
| --- | --- |
| **Name** | **Matriculation Number** |
| Ethan Ong Sze Rey | U2010183H |
| Joel Mui Kit | U2010877H |
| Khoo Teng Khing, Joshua | U2021421C |
| Brandon Quek | U2011292D |
| Lim He Ying | U2011832J |

## 

**Table of Contents**

[Executive Summary 4](#_Toc99915938)

[1. Introduction 5](#_Toc99915939)

[1.1 Overview of Department of Transport (DfT) 5](#_Toc99915940)

[1.2 Defining the Business Problem 5](#_Toc99915941)

[1.3 Current Measures 5](#_Toc99915942)

[2. Objectives 6](#_Toc99915943)

[2.1 Business Objectives 6](#_Toc99915944)

[3. Analytics Solution 6](#_Toc99915945)

[3.1 Key Business Questions 6](#_Toc99915946)

[3.2 The Analytics Solution: Predicting Accident Severity with Machine Learning 7](#_Toc99915947)

[3.3 Evaluation of Analytics Outcome 7](#_Toc99915948)

[4. Preparation of Datasets 8](#_Toc99915949)

[4.1 Data Acquisition and Overview 8](#_Toc99915950)

[4.2 Data Cleaning and Pre-processing 8](#_Toc99915951)

[4.2.1 Initial Cleaning 8](#_Toc99915952)

[4.2.2 Regrouping Variables 8](#_Toc99915953)

[4.3 Data Exploration and Visualisation 9](#_Toc99915954)

[5. Identification of key predictors 11](#_Toc99915955)

[5.1 Methodology 11](#_Toc99915956)

[5.1.1 Logistic Regression 11](#_Toc99915957)

[5.1.2 MARS 11](#_Toc99915958)

[5.2 Model Evaluation 12](#_Toc99915959)

[5.3 Causality Analysis of Key Predictors 12](#_Toc99915960)

[5.3.1 Geographical Location (Latitude and Longitude) 12](#_Toc99915961)

[5.3.2 Urban or Rural Locality 12](#_Toc99915962)

[5.3.3 Casualty Age Group 13](#_Toc99915963)

[5.3.4 Speed Limit 13](#_Toc99915964)

[5.3.5 Engine Capacity (cc) 13](#_Toc99915965)

[5.3.6 Time of the Day 13](#_Toc99915966)

[5.3.7 Casualty Type 14](#_Toc99915967)

[5.3.8 Vehicle Manoeuvre 14](#_Toc99915968)

[5.3.9 Journey Purpose 14](#_Toc99915969)

[6. Prescriptive Techniques and Basis for Comparison 14](#_Toc99915970)

[7. Prediction model for Factors resulting in High Risk of Serious Accidents 16](#_Toc99915971)

[7.1 Logistic Regression with Backward Elimination 16](#_Toc99915972)

[7.2 Multivariate Adaptive Regression Splines (MARS) 16](#_Toc99915973)

[7.3 Classification and Regression Trees (CART) 17](#_Toc99915974)

[7.4 Random Forest Classifier (RF) 18](#_Toc99915975)

[7.5 Selecting the best predictive model 19](#_Toc99915976)

[8. Curated Recommendations to tackle Root Causes 20](#_Toc99915977)

[8.1 Community Centric Hazard Sharing Application 20](#_Toc99915978)

[8.2 Upgrading Traffic Fixtures at Accident Hotspots 21](#_Toc99915979)

[8.3 Legislative Amendments 22](#_Toc99915980)

[8.4 Increasing frequency of scheduled road inspections 22](#_Toc99915981)

[8.5 Efficient emergency response service standby scheduling 22](#_Toc99915982)

[9. Discussion 23](#_Toc99915983)

[9.1 Benefit of the Analytics Solution 23](#_Toc99915984)

[9.2 Limitations of Analysis 24](#_Toc99915985)

[10. Conclusion 24](#_Toc99915986)

[Appendices 25](#_Toc99915987)

[Appendix 1: Logistic Regression Results 25](#_Toc99915988)

[Appendix 2: MARS results 32](#_Toc99915989)

[Appendix 3: CART Results 34](#_Toc99915990)

[Appendix 4: Random Forest results 36](#_Toc99915991)

[Appendix 5: Sample Mock Ups for Hazard Reporting App 37](#_Toc99915992)

[References 40](#_Toc99915993)

## 

## Executive Summary

Over the years, road accidents have become more **prevalent** due to the increase in car usage in the United Kingdom (UK), resulting in high human costs and detrimental effects to the UK economy. Reducing accident severity can potentially reduce the cost of road traffic crashes which is estimated to be **3% of a country’s GDP** (WHO, 2021). With relaxation of measures, many are returning to offices and road usage is expected to **surge** (Auto, 2022). Hence, there is an **urgent** need to predict potentially fatal traffic accidents **before they occur**.

With advancements in technology and machine learning, the Department for Transport (DfT) can make use of **big data** to predict the occurrence of severe accidents. This report seeks to explore the use of machine learning models to develop recommendations to reduce the occurrence of severe accidents. The team utilized the datasets provided by data.gov.uk to aid us in addressing the issue.

This report explores the combined analytical application of two approaches to tackle this issue. The first approach integrates Logistic Regression and Multivariate Adaptive Regression Splines (MARS), to identify and understand the **root causes** of severe accidents. This is achieved by identifying key predictors of accident severity for both models and focusing on **common predictors**. To verify the significance of our model-identified predictors, the team conducted **cross-analysis** with online research. Using these key predictors, our team developed five **targeted** recommendations to reduce the occurrence of severe accidents.

1. Development of a community-centric hazard reporting app
2. Upgrading traffic fixtures
3. Legislative amendments
4. Increasing frequency of scheduled road inspections
5. Efficient emergency response service standby scheduling

These recommendations are derived from **analytics insights** with the aim of **addressing** the **limitations** of existing measures.

The second approach involves comparing four different machine learning models - Logistic Regression, MARS, Classification And Regression Tree (CART) and Random Forest to predict accident severity in areas within the UK. Our team then utilizes key performance metrics such as F2 score and False Negative Rates(FNR) to evaluate the machine learning model. Our team recommends the usage of the **Random Forest model** trained on the **under sampled** train set to predict accident severity as it exhibits the **highest F2 score** and **lowest FNR**.

This report also assessed the limitations of our accident severity model. The first limitation is that the team used **naive sampling techniques** which reduces the model accuracy. The second limitation is due to the **dynamic** nature of accident data. Datasets that are from three to five years ago might not be representative of today’s situation.

The report concludes by highlighting the benefits of adopting a **combination** of machine learning methods in reducing the occurrence of severe accidents in a **cost-efficient way** and to build safer roads in the future.

## 1. Introduction

### 1.1 Overview of Department of Transport (DfT)

The Department for Transport (DfT) is the department of the United Kingdom Government responsible for planning and investing in the UK's transport infrastructure. Comprising over 18,000 staff, the DfT has maintained a **strong interest** in ensuring high standards of safety and security in the UK's transport system since its conception. Given the UK government’s focus on maximising road safety, this report centralises on exploring the use of analytics and machine learning to **enhance road safety** in the UK.

This report seeks to explore the use of analytics in reducing the occurrence of serious road accidents by **predicting them before they happen**.

### 1.2 Defining the Business Problem

Road networks are a vital component of the UK’s economy and way of life. Two-thirds of all journeys to work in the UK are by car, and even those who do not drive **still rely on roads** for the delivery of food to supermarkets and goods to shops. Additionally, roads support job creation and unlock new development through greater access to labour markets and new opportunities for factories and businesses (Department for Transport, 2020).

Prior to the emergence of the COVID-19 pandemic, road traffic in the UK had steadily increased over the years. In 2019 alone, **356.5 billion** vehicle miles were driven on UK roads - accounting for **90%** of passenger journeys and two thirds of freight (Department for Transport, 2020). This highlights the **sheer volume** of road usage in the UK.

However, this increase in traffic use has also led to a rise in traffic accidents in recent years - some of which may lead to severe injuries, or even death. In 2019, there were a total of **153,158 casualties** in reported road traffic accidents in the UK - of which **25,945** were **serious** injuries, and **1,752** resulted in **fatalities** (Department for Transport, 2020). The total human costs for casualties in road accidents were estimated to have amounted to a massive $8.7b British Pounds - with **severe injuries** and **fatalities** accounting for **$7.1b** (Statista, 2022). Despite these startling statistics, the trend in the number of fatalities and serious injuries have shown **little improvement** over the past few years.

### 1.3 Current Measures

Currently, the DfT has taken several measures that aim to reduce road accidents. These measures can be split into two key themes - promoting safe behaviour and condemning dangerous behaviour.

To promote safe behaviour, the DfT has rolled out extensive education programmes across the UK. Most notably, the DfT successfully rolled out the “Green Cross Code'' initiative which primarily targeted children aged 3-16 years old. Recognising that young adults are overrepresented in road accident statistics, the DfT launched the Driver2020 Project, which aims to explore different ways to make young drivers safer, and more confident in their first year of driving. Additionally, the DfT has revised its driving tests to be more capable of assessing a candidate’s ability to drive safely and responsibly in a range of road and traffic conditions (Department for Transport, 2019).

Additionally, recognising police enforcement as a key factor in tackling drink-driving, the DfT recently began research and development to manufacture Mobile Evidential Breath Test Instruments (MEBTI) - a device that enables the Police to obtain breath samples of suitable accuracy to be used in court. Furthermore, the DfT has also rolled out the “Operation Snap” initiative, which provides road users with a platform to upload dashcam media or personal videos as evidence of traffic offences (Department for Transport, 2019).

These measures have been **limited** in their **effectiveness** as they have been rolled out on a **country-wide basis**. Instead, **targeted** solutions are required to reduce severe accidents in accident hotspots in the UK, which will be discussed in Section 8.

## 2. Objectives

### 2.1 Business Objectives

Road accidents lead to significant human costs, economic welfare costs, and also have a serious detrimental impact on the UK’s economy. The UK government has a vested interest in upkeeping the UK's economy and protecting citizens from unnecessary harm. Therefore, given that higher traffic use could likely lead to more traffic accidents, **the DFT would be able to better reduce costs attributable to road accidents by reducing the occurrence of severe road accidents.** This is because a large proportion of costs attributable to road accidents are due to accidents resulting in severe injuries or fatalities.

Furthermore, governments strive to attain **allocative efficiency** of limited resources. Hence, a reduction in road accident costs would enable the UK government to better invest in other facets of its economy and society.

## 3. Analytics Solution

### 3.1 Key Business Questions

|  |  |
| --- | --- |
| Business Question | Explanation |
| **When** and **where** do most serious accidents happen? | This question aims to identify high-risk cities, days, and timings as areas of concern, so fatalities can be reduced by improving emergency response times. |
| Can there be **better traffic amenities** to reduce the risk of serious accidents? | This question aims to explore the potential to improve road amenities to reduce accident severity caused by poor infrastructure. |
| Can we identify **high risk** drivers and high-risk casualties? | This question aims to classify high-risk and low-risk drivers and casualties, so that customised recommendations can be made for the more vulnerable group. |

### 3.2 The Analytics Solution: Predicting Accident Severity with Machine Learning

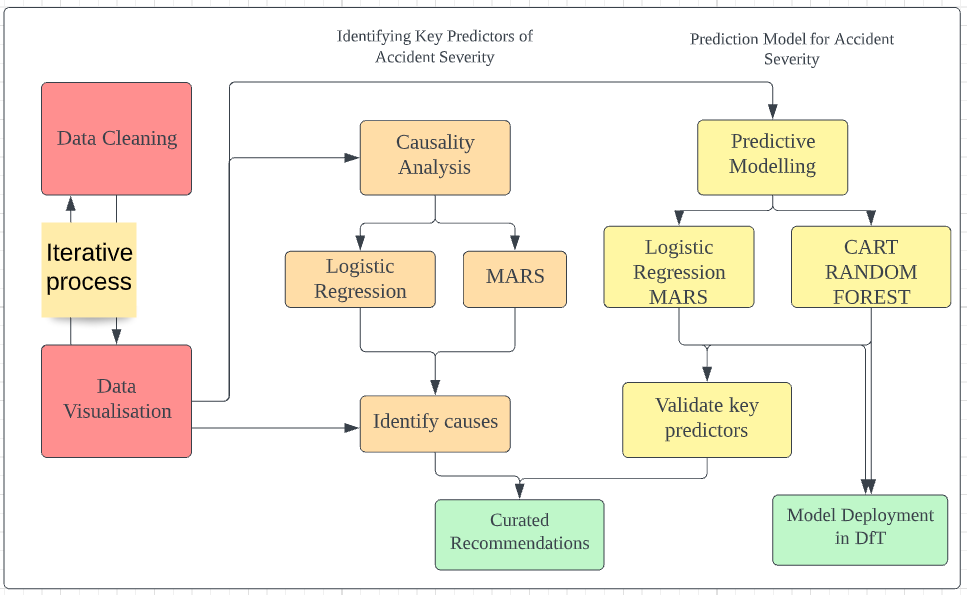


Figure 1: High Level Overview of Analytics Solution

Due to advancements in accident reporting and data collection procedures, the DfT possesses large amounts of data that it can harness. Through the use of analytics, the DfT can generate key insights from data to make better-informed decisions.

For instance, analytics could be used to identify certain conditions that are more likely to lead to severe accidents. This would allow DfT to identify roads that are at higher risk of causing severe accidents and take necessary corrective actions to reduce these risks. Furthermore, being cognisant of factors that would increase risk of severe accidents would empower DfT to conduct better urban planning by building fewer roads that exhibit such characteristics. This would lead to **better allocative efficiency** of resources and less time wasted.

A decrease in risks of severe accidents on UK’s roads would lead to fewer severe injuries and/or fatalities, thereby **reducing the costs attributable to traffic accidents** in the UK. Hence, this report seeks to evaluate the performance machine learning models and determine which model can most accurately forecast accident severity and identify factors that are important for such predictions. Based on these factors, the team would provide recommendations to the DfT on possible countermeasures to reduce risk of severe accidents.

### 3.3 Evaluation of Analytics Outcome

|  |  |
| --- | --- |
| Performance Targets | Performance Measures |
| 1. Reducing number of severe accidents per year by **3%** | 1. Record the number of severe accidents per year after implementation of analytics solution |
| 1. Reduce costs attributable to road accidents | 1. Record the estimated human costs, economic welfare costs, and value of insurance pay-out attributable to road accidents per year after implementation of analytics solution |

## 4. Preparation of Datasets

### 4.1 Data Acquisition and Overview

Data used for analysis was sourced from data.gov.uk. Attributes are sampled in the figure below.

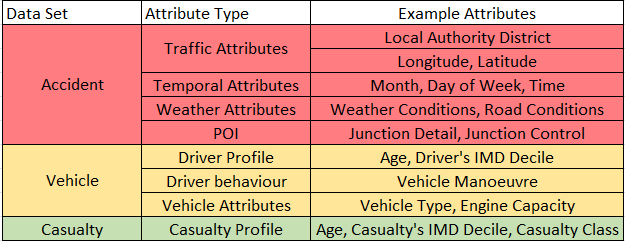


Figure 2: Example attributes for each dataset

Common to all three datasets is the **accident index** - a unique identifier that combines the year in which the accident took place with the accident reference number. The datasets are joined with this field for further analysis and exploration.

### 4.2 Data Cleaning and Pre-processing

#### 4.2.1 Initial Cleaning

As the three datasets are related, the first few steps of data cleaning are performed in common. We subset the data to only include 2017-2019, which is the scope of our analysis. Initially, categorical variables are represented by numbers while the corresponding labels are included in the data dictionary. Categorical variables were manually renamed with their labels to increase ease of interpretation.

According to the data dictionary, data that is missing or out of range is encoded with a factor of “-1”. Unknown data from self-reported accidents is encoded as either “9” or “99”. As this **missing data** is not helpful for analysis, we have **removed** them from the datasets.

#### 4.2.2 Regrouping Variables

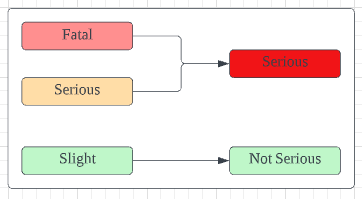
Before data cleaning, accident severity was a categorical variable with three levels: “Fatal”, “Serious” and “Slight”. We found that costs related to fatal and serious accidents are much greater than that of slight injuries (Catalano et al., 2020). The figure on the right shows how the categories were **reclassified**. This classification is helpful since our **key objective** is to reduce the occurrence of severe traffic accidents, represented by the group **“Serious”.**

Figure : Variable regrouping

When analysing the structure of the datasets, certain features exhibited high **cardinality (**many categories**)**, which will affect the results of applied machine learning methods. For example, the derived column “month” in the Accident dataset has a cardinality of 12. Having high cardinality greatly increases the number of possibilities for combining features. Without sufficient training data, this could result in the model being overfitted to the training set and unable to generalize well (Sangani, 2021).

To work around this, we explored regrouping several variables to reduce cardinality. One-hot encoding was not an option as it would likely result in the curse of dimensionality, leading to multicollinearity and affecting machine learning model accuracy.

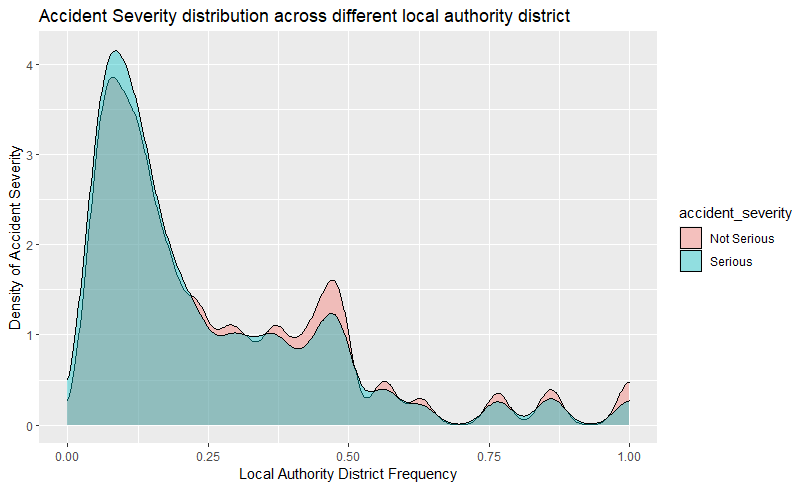
Geospatial data poses a challenge as it has extremely high cardinality. For example, there are over 300 Local Authority Districts in the UK. To deal with this problem, geospatial categories were tested to explore their impact on accident severity. If the impact on accident severity was minimal, the columns were removed to reduce cardinality. From Figure 4, we can see that **Serious accidents have higher likelihood at local authority districts with lower frequencies**, therefore, local authority is an important geospatial data. Frequency encoding [[1]](#footnote-1)was used to assign weights to different local authority districts based on frequency of occurrence, reducing the dimensionality of the dataset while ensuring that important variables were not excluded from analysis (Kumar, 2021). However, one downside of frequency encoding is that it may result in loss of information if there are categories with the same number of occurrences.

Figure : Accident severity distribution

### 4.3 Data Exploration and Visualisation

After data cleaning, the following data visualizations were created to highlight initial insights and trends from each dataset. Data was further cleaned **iteratively** as we identify trends during data exploration.

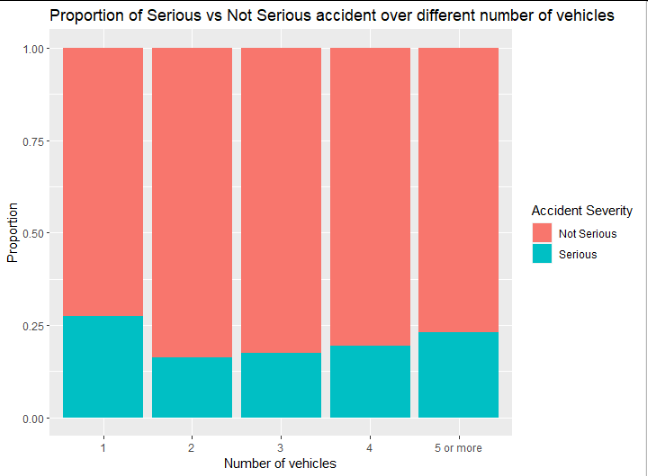
This plot compares the proportion of serious and not serious accidents between accidents involving different numbers of vehicles. As expected, accident severity tends to increase as the number of vehicles in the accident increases. Multi-vehicle pile ups tend to take place on highways or roads with higher speed limits where visibility is low, contributing to more serious injuries or even death. Interestingly, this plot reveals that accidents involving just one vehicle have the highest proportion of serious accidents. This could possibly be explained by the fact that accidents involving only one vehicle are between a driver and pedestrian. As pedestrians are not protected by a car chassis or protective riding equipment, they have a much higher chance of getting seriously injured in a road accident.

Figure : Proportion of accident severity based on number of vehicles

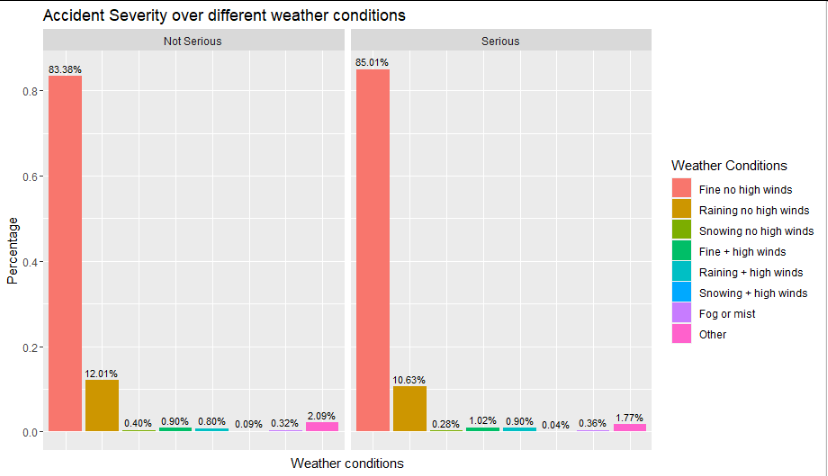


Figure 6: Accident severity over different weather conditions

Contrary to popular belief, this plot shows that the **vast majority of accidents occur during perfect weather conditions** rather than during inclement weather. While it is logical to assume that more accidents would occur during rainy or snowy weather where road conditions are worse and visibility is lower, the observations from our dataset revealed that a whopping 83.69% of road accidents occurred during fine weather. A possible explanation for this observation is that drivers either **drive more cautiously** or completely avoid driving during inclement weather as they are aware of the increased risk of getting into an accident. However, another explanation for this disproportionate observation is that the UK’s climate mostly consists of fine weather conditions with no high winds, therefore the majority of accidents would occur during such conditions.

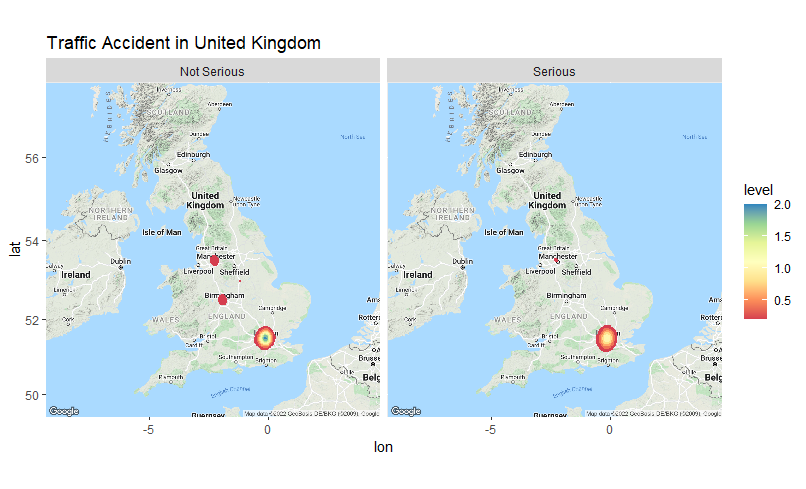


Figure 7: Accident heatmap

Exploration of the geographical variables revealed that there are three main **hotspots** where road accidents occur. There is a giant hotspot centred on **London** which is to be expected, given it is the capital of the UK and is the largest metropolitan hub in the country. The other hotspots are **Manchester and Birmingham**, the second and third most populous cities in the UK respectively. Thus, there is an observable relationship between population density and accident frequency.

## 5. Identification of key predictors

### 5.1 Methodology

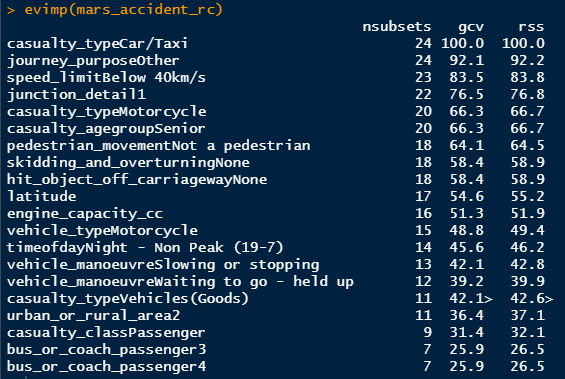
Given such a large variety of predictors, there is a need to identify predictors of high significance towards better understanding the root causes of accident severity to develop effective and targeted recommendations. A hybrid evaluation model through a combination of models and expert opinion would enable us to identify significant variables with the greatest confidence. Logistic regression and MARS were the chosen techniques due to their high degree of **explainability**.

#### 5.1.1 Logistic Regression

Having identified the binary outcome variable, logistic regression was performed on the cleaned dataset. Statistical significance of predictor variables is represented by p-values. In general, a predictor variable is considered to be significant if its **p-value is below 0.05**. Initial logistic regression results can be viewed in Appendix B. However, using p-value on its own is insufficient evidence to suggest that a variable is important. We further verified significance by calculating the odds ratio confidence interval for each variable. If the confidence interval includes 1, that means the predictor does not affect accident severity.

Due to the high volume of predictor variables in the combined dataset, a **backward elimination** algorithm was applied to the first logistic regression model to narrow down the number of variables to be included. The function of choice is ‘step’ from the base stats library in R. It optimizes the model by choosing variables to include based on minimizing the Akaike Information Criterion (AIC)[[2]](#footnote-2). Results of logistic regression with backward elimination can be found in Appendix B.

#### 5.1.2 MARS

In addition to logistic regression, MARS can be used to corroborate our findings and give them more credibility. When initializing **MARS, degree 2** was chosen instead of the default degree 1 as we were interested in exploring interaction effects between the variables. With the “evimp” function, MARS returns three metrics used to evaluate variable importance:

1. Number of model subsets including the variable. **(nsubsets)**
2. Generalized Cross-Validation **(GCV)**, a measure of residual sum of squares that incorporates a model complexity penalty.

Figure : MARS Variable Importance Outputs

1. Residual Sum of Squares **(RSS)** which does not include model complexity.

### 5.2 Model Evaluation

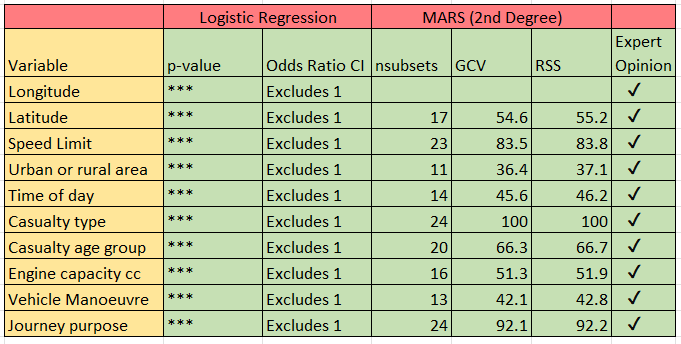


Figure 9: Table of significant variables as selected by criteria

Based on logistic regression with backward elimination and MARS degree = 2, our team narrowed down the key predictors as displayed in Figure 9. The variables identified must meet the following **criterion**.

1. Have significant p-values (p-value <0.05 in logistic regression)
2. Odds Ratio Confidence Interval excluding 1
3. They must be identified as important by the 2nd degree MARS Model.
4. Supported by domain knowledge of experts

### 5.3 Causality Analysis of Key Predictors

#### 5.3.1 Geographical Location (Latitude and Longitude)

Geographical location is considered an important predictor for severity of road traffic accidents. It consists of the Latitude and Longitude of the accident site. As identified during data exploration, `local\_authority\_district` was found to have an effect on accident severity. Districts with lower accident frequency tend to have more serious accidents. These districts are usually rural areas further away from the city centre, London. Some rural communities have to wait up to 20 minutes for emergency response services (BBC, 2019). In serious accidents, every second counts, hence we could attribute the greater probability of serious accidents to the **poorer emergency response times in these areas**.

#### 5.3.2 Urban or Rural Locality

Urban or Rural Locality of the accident site is identified as an important predictor for road accidents severity. Based on our data exploration, we found that there was a higher likelihood of severe accidents occurring in Rural areas as compared to Urban areas. This finding is supported by a study which found that while accident frequency is higher in Urban areas, the proportion of severe accidents is higher in Rural areas (Cabrera-Arnau, 2020). Rural areas have a larger proportion of older drivers who have slower reaction times than young drivers. Additionally, their increased fragility means they have a higher chance of dying or getting seriously injured from accidents than young drivers. Drivers also have a higher tendency to speed in rural areas due to the lighter traffic flow, increasing the severity of accidents.

#### 5.3.3 Casualty Age Group

Casualty Age Group represents the age segments of the casualties and is also an important predictor of road accident severity. Our data exploration found that **seniors are more prone** to severe accidents than non-severe accidents, while adults are less prone to severe accidents. Generally, seniors are more prone to accidents and are therefore more likely to sustain severe injuries compared to adults (Brand et al., 2012). Additionally, teenagers are at higher risk of fatal accidents due to their inexperience and reckless driving habits (Stevens, 2018). Every year, over 1,750 people are killed in traffic accidents and over 150,000 people are injured - many of whom are young or senior citizens from deprived communities (Towards Zero Foundation, 2021). Therefore, this supports our model prediction that casualty age group is an important predictor of accident severity.

#### 5.3.4 Speed Limit

Speed limit on the highway is identified as an important predictor of road accidents severity. Through data exploration, a **clear distinction** can be seen, the probability of severe accidents is much higher at speed limits over 40km/h, compared to non-severe accidents. This indicates that **higher speed limits** will result in a higher probability of severe accidents. A study done by Insurance Institute for Highway Safety (IIHS), concluded that every 8 km/h increase in maximum state speed limit was associated with an 8% increase in mortality rates on highways. This is because drivers have the tendency to drive faster with a higher speed limit, and higher speed will lead to greater impact in the event of a collision (IIHS, 2021).

#### 5.3.5 Engine Capacity (cc)

Engine capacity of a vehicle is identified as an important predictor of road accidents severity. Based on research conducted by NHTSA, vehicles with a higher engine capacity is more **powerful** and causes a higher percentage of fatal crashes (NHTSA, NA). More powerful vehicles are able to accelerate faster which can result in **higher travelling speeds** (CarBikeTech Team, 2014) which was earlier explained to have caused more serious accidents.

#### 5.3.6 Time of the Day

Time of the day when the accident occurs is identified as an important predictor of road accidents severity. From our data exploration, it is shown that a higher percentage of severe accidents occur during non-peak hours at night compared to non-severe accidents. According to a study done by the National Highway Traffic Safety Administration (NHTSA), the risk of a fatal accident at night is three times higher than if they are driving during the day. This could be due to **lower seat belt usage at night**, and the **higher occurrence of speeding and drink-driving at night**. In addition, **low visibility** and drowsiness of the driver can also contribute to more severe accidents (NHTSA, 2007).

#### 5.3.7 Casualty Type

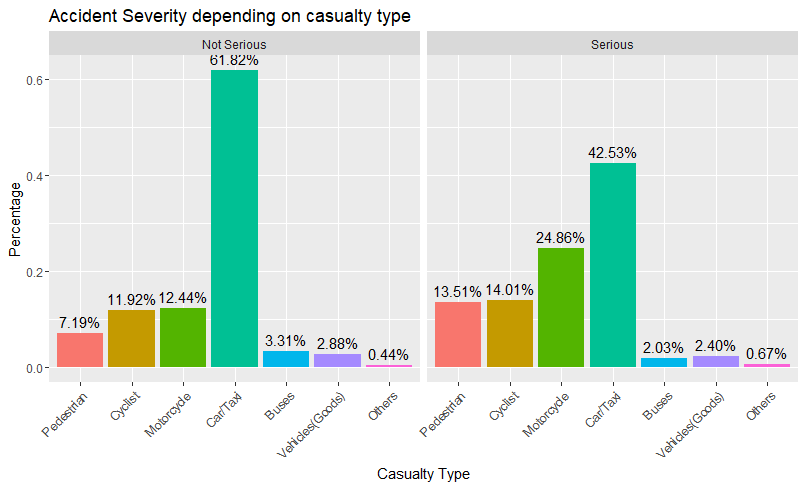
Casualty Type, which refers to the mode of transport of the injured person during an accident, was identified as an important predictor. Through data exploration, we found that **cyclists, motor cyclists or pedestrians** are more likely to be involved in severe accidents. This can be attributed to the fact that there is little to no external protective device that could have absorbed the impact of the road crash (Yannis et al., 2020), making these population groups **more vulnerable**. Furthermore, motorcycles are more susceptible to dangerous road conditions as they have lesser stability compared to other vehicles (The Brown Firm, 2020).

Figure : Accident severity distribution by casualty type

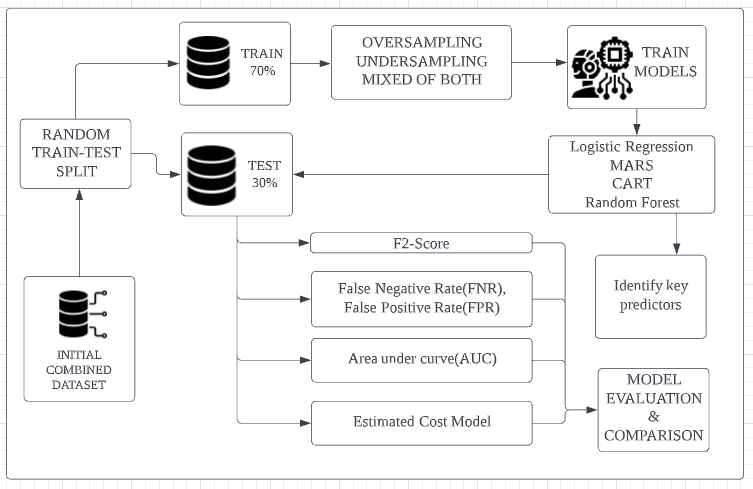
#### 5.3.8 Vehicle Manoeuvre

Vehicle Manoeuvre represents the change in travelling direction or movement of the vehicle and is identified as an important predictor of road accidents severity. From our data exploration, we found that a vehicle that is turning right and going ahead of others (**overtaking**) has a higher percentage of severe accidents compared to non-severe accidents. We also found that most **right turn manoeuvre** accidents involve larger sized vehicle (e.g. vans, trucks), thus resulting in more severe accidents (Thomas & Janina, 2016). One reason why larger sized vehicles are more prone to such accidents could be due to more prominent blind spots for drivers of larger vehicles. This is further supported by evidence that suggests that overtaking is one of the most dangerous manoeuvres for both drivers and riders since it can force the overtaking vehicle into oncoming traffic at high speeds (ROSPA, 2021).

#### 5.3.9 Journey Purpose

Journey purpose represents the purpose of driving and is identified as an important predictor of road accidents severity. Based on our data exploration, we found that school or work-related commuting have lower tendencies of serious accidents compared to the “Others” category. “Others” typically represent **leisure travel**. Leisure travellers have a higher likelihood of being involved in severe accidents. This is likely due to the fact that people are more likely to exceed speed limits and/or be easily distracted - which would lead to higher incidence of fatal accidents (Febres et al., 2019).

## 6. Prescriptive Techniques and Basis for Comparison

Since we are investigating the factors that affect accident severity, there is a clear categorical outcome variable: “**accident\_severity**”. With an abundance of predictors, classification models can be used to find the variables that have the biggest impact on the outcome of an accident being severe. The methods chosen for analysis are: Logistic Regression (LR), Classification and Regression Trees (CART), Multivariate Adaptive Regression Splines (MARS) and Random Forest (RF).

To ensure reproducibility of machine learning results, we set the seed as 2407. A **70-30 train-test split** was used to ensure that a consistent test-set can be used to fairly compare and evaluate the results. As the proportion of Serious Accidents is much lower, we explored **over sampling, under sampling and a mix of both** to create a balanced training set for model building. This is to indirectly increase the weights placed on the minority class(“Serious”).

Figure : High level overview of model training and evaluation

**False negatives**: Model predicts a Severe accident as Not Severe

**False Positives**: Model predicts a Not Severe accident as Severe

A few key metrics were identified for comparing the results of different models. Model accuracy refers to the ability of the model to correctly classify accident severity of accidents in the test set. Most importantly, a model predicting accident severity should have a **low False Negative Rate (FNR)**. As false negatives could result in loss of life and costly damages to infrastructure, the cost of such accidents is very high and minimizing these costs should be a priority.

Due to a class imbalance issue, we determine that accuracy is a poor measure of evaluation for the business problem. This is because, even if the model fails to predict any severe crashes, it can still achieve very higher accuracies and not serious accidents take up the majority of the data. Hence, F2 scores and **Area Under Curve (AUC)** which are more appropriate metrics were used. AUC is the measure of the ability of a classifier to distinguish between classes, the higher the better.

As we are dealing with a classification problem where FNR is prioritized, **F2 score** is a useful metric to evaluate model accuracy. While F1 score weighs false positives and false negatives equally, this does not match the use case. Since minimizing FNR is more important, a higher weight of 0.8 is used for false negatives. A **higher false negative count is penalized** with a lower F2 score, as the denominator is larger.

After creating models with each training set and comparing the key metrics on the test-set, we found that using the under sampled training set resulted in the test-set with the best F2 score. Thus, result interpretation in the next section focuses on models trained with the under sampled training set.

## 7. Prediction model for Factors resulting in High Risk of Serious Accidents

### 7.1 Logistic Regression with Backward Elimination

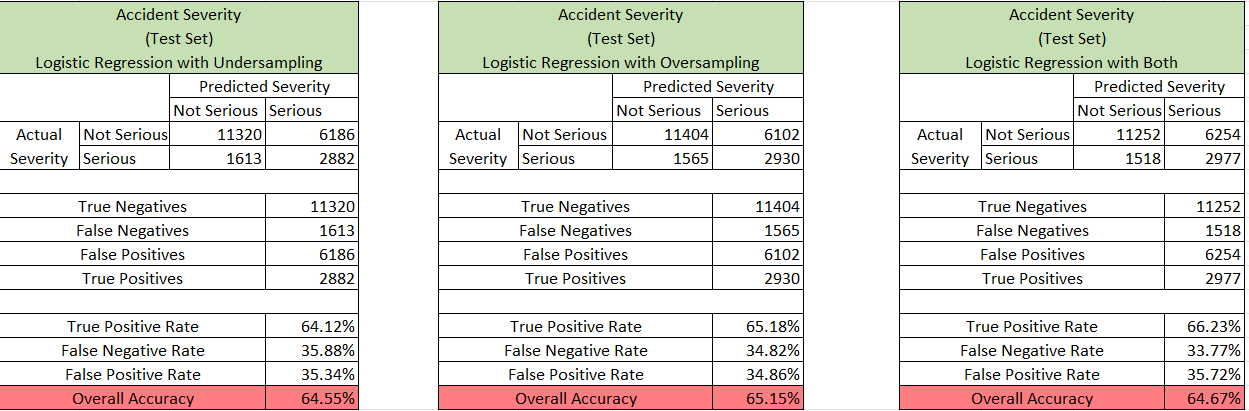
Logistic Regression is a model that classifies categorical outcome variables with a logit function. The output is log odds[[3]](#footnote-3), which can be used to calculate the probability. If probability is greater > 0.5, the model will classify the case as “Serious”. It holds value as a **highly explainable** and interpretable model that shows the statistical significance of all variables along with model coefficients which can be used to calculate the outcome. As explained in section 5.1.1, stepwise backward elimination was used to optimize the Logistic Regression model by minimizing AIC. A summary of results is shown below.

Figure : Summary of Logistic Regression results

Overall test set accuracies were 64.55% for under sampling, 65.15% for over sampling and 64.67% for a hybrid of both. FNR decreased across the three training sets, with the hybrid test set having the lowest FNR at 33.77%.

### 7.2 Multivariate Adaptive Regression Splines (MARS)

MARS is suitable to describe nonlinear[[4]](#footnote-4) relationships. Simply explained, the MARS model is a combination of **piecewise linear functions with hinges**. The MARS model finds the best fitting hinge functions to split the data into broad sections, consisting of the forward stage to generate the best splits and the backward stage to remove splits that do not contribute significantly to improving model accuracy. This is to reduce overfitting. Apart from degree, all other parameters were left at default settings. The **degree was set to 2** to enable exploration of **interactions** between variables.

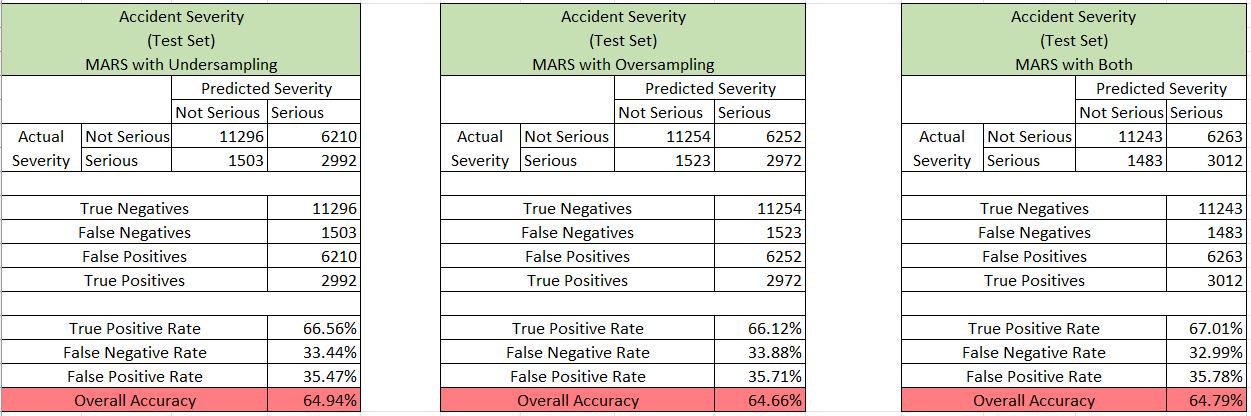


Figure 13: Summary of MARS results

Overall test set accuracies stood at 64.94% for under sampling, 64.66% for over sampling and 64.79% for a hybrid of both. FNR was 33.44% for under sampling, 33.88% for over sampling and 32.99% for the hybrid training set. Compared to logistic regression, the difference in test set accuracy is almost negligible. FNR for the MARS models was around 1-2% lower than logistic regression.

### 7.3 Classification and Regression Trees (CART)

CART is a model that uses a **rule-based approach** to build a decision tree based on Gini’s impurity index as a splitting criterion. The output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable (*Machine Learning - (CART) - Q*, 2021). Essentially, the predicted outcome is achieved by following a set of rules determined by the algorithm. For all 3 sampling techniques, minsplit = 20 was used due to a large dataset and complexity parameter(cp) = 0 to grow the tree to a maximum. After which, the tree was **pruned** using the **optimal CP** chosen by the model, which minimises the total cost of the CART model. The optimal CART tree can be seen in Appendix 3.

CART Cost = Model Prediction Error + Complexity cost (Size of Tree)

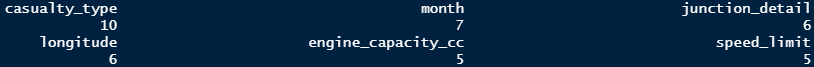


Figure 14: CART scaled variable importance

Variable importance in CART is determined by how **often** a variable is used for splitting and how close it appears to the **root** of the tree. Surrogates are irrelevant here, as there are no missing values. We identify the type of casualty (10%), month (7%), junction detail (6%), geographical location (6%), vehicle’s engine capacity (5%) and speed limit (5%) as the most important predictors. (Figure X)

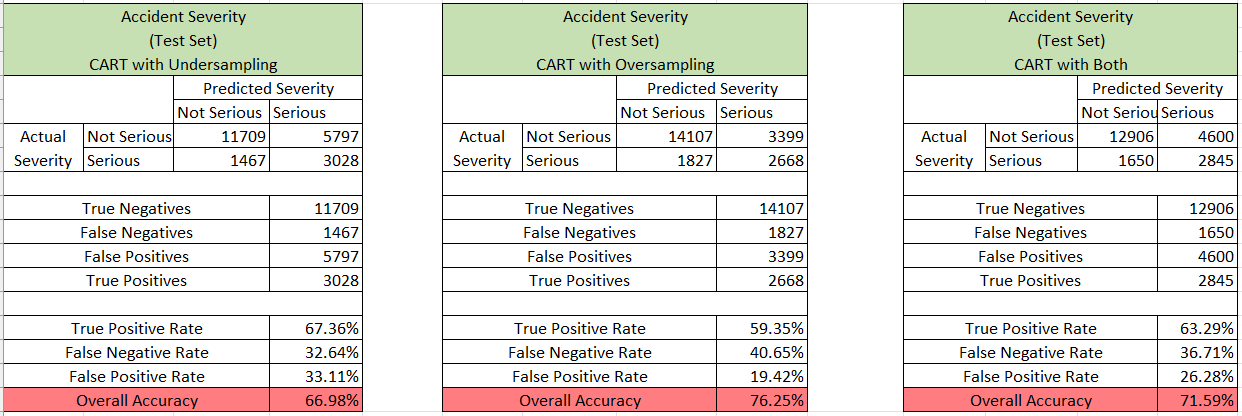
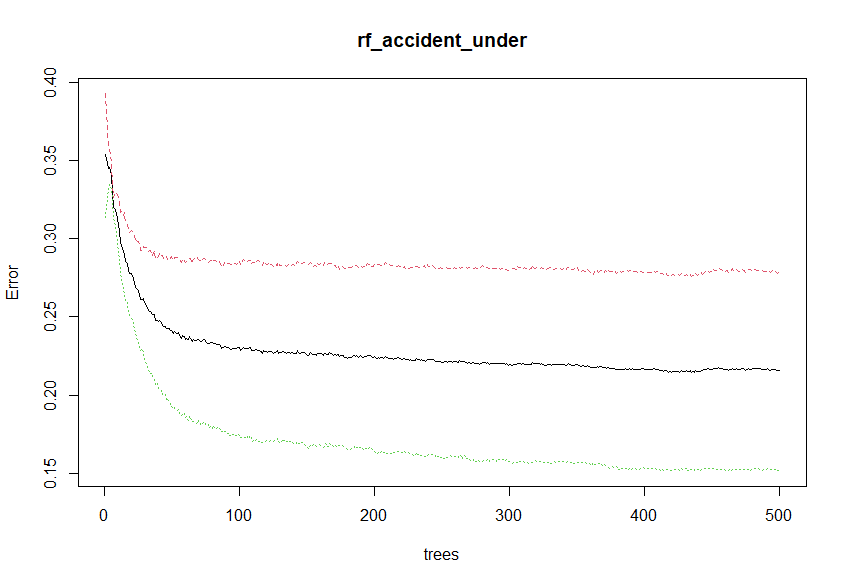


Figure 15: Summary of CART results

Overall test set accuracies stood at 66.98% for under sampling, 76.25% for over sampling and 71.59% for a hybrid of both. False negative rates fell to 32.64% for under sampling, 40.65% for over sampling and 36.71% for a mixture of both. Across all 3 sampling techniques, we can see that CART has a higher test set accuracy compared to both logistic regression and MARS. More importantly, the models all generally have lower false negative rates.

### 7.4 Random Forest Classifier (RF)

Random Forest is a machine learning model that leverages on the idea of Bootstrap Aggregation (**Bagging**). Each decision tree is built using different random samples with **replacement** from the original data and trained independently. (Bootstrap) Since this is a classification problem, the model will take the **majority vote** of all the trees as the final prediction. (Aggregation)

The Random Forest model was run in R using library(randomForest) with mtry = 9 and ntrees =500. The other hyperparameters were the default settings. Due to the large number of variables, we found that mtry = 9 gives us the best accuracies and the OOB **error stabilised** after 100 trees were grown, thus showing the sufficiency of 500 trees to train the model. (Reference Figure 16)

Figure : Out of bag error

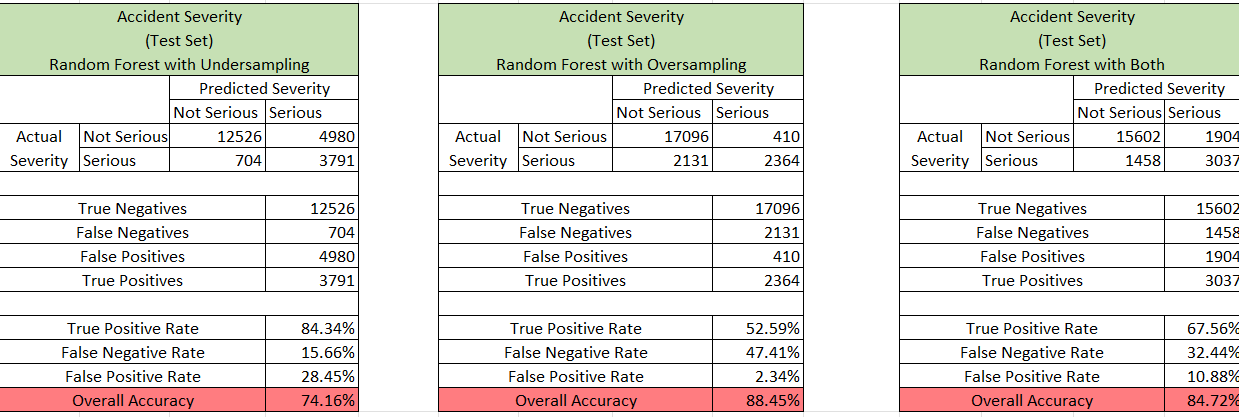


Figure 17: Summary of Random Forest results

Using bagging, RF is able to **reduce biases** and the influence of a strong predictor variable. Furthermore, by taking the majority vote of 500 trees, it is able to reduce single tree bias (as seen in CART) and thus does not overfit. Hence, it is able to produce a **superior predictive accuracy** on an unseen test set.

As seen in Figure 17, Random Forest is able to achieve a remarkable 88.45% test set accuracy for over sampling, 84.72% for a mixture of both and 74.16% for under sampling. Thus, it is superior to all the other previously mentioned models. The under sampling approach boasts a **15.66% false negative rate** which is important in this context, however, false negative rates for the oversampled train set is 47.41% and 32.44% for the hybrid sampling approach. This will be further discussed in Section 7.5 model evaluation.

However, one drawback of RF is its **black box** nature, rendering it difficult to interpret how predictors influence the predictions. RF however has an inbuilt **permutation feature importance** to allow us to gain insight into the features with predictive power. The idea is that important features contain valuable information, and if that information is destroyed by **randomly shuffling** its values, the prediction accuracy will fall. If the decrease in accuracy is large, then the predictor has a great impact on predictions. (Billiau, 2021)

The results of RF’s feature importance are shown in Figure 18. It is clear that month, casualty type, day, geographical location, vehicle manoeuvre, junction detail, vehicle engine capacity and the time of day are important. This is because removing these variables will result in the **largest Mean Decrease in accuracy** of the predictions, as summarised in the Figure on the right.

The partial dependence plot in Figure 19 further reveals the relationship between the predictors and serious traffic accidents. We set the class of interest using **which.class = “Serious”.**

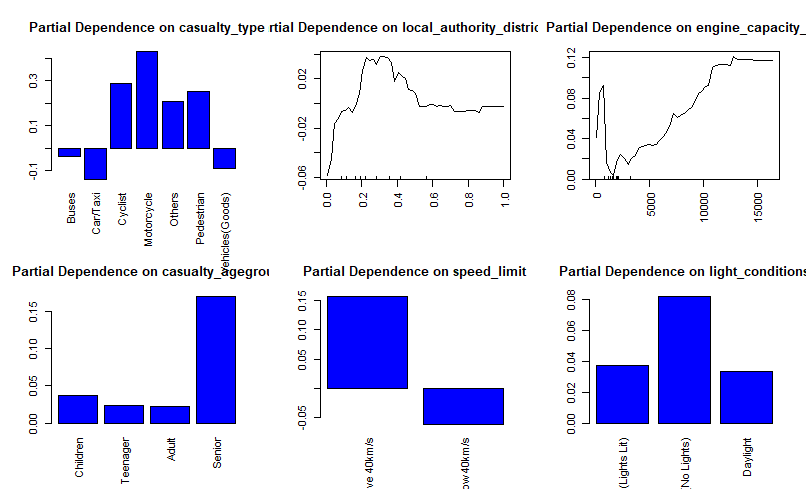
The plots **support the causality analysis** in Section 5.3. Firstly, the probability of serious accidents is much higher when the casualties are cyclists, motorcyclists, or pedestrians. Secondly, the probability of serious accidents is highest from the local authority district **frequency range of 0.2 to 0.4**. This means that the frequency of accidents in these districts are fewer but more severe. Thirdly, the probability of serious accidents increases sharply when the engine capacity of the vehicle exceeds 5000cc. Likewise, senior citizens are the most vulnerable casualties and the probability of serious accidents is much higher when the speed limit is above 40 km/s. Additionally, it was found that the probability of serious accidents are much higher at night and when the lights are unlit.

Figure : RF Feature Importance

Figure : Partial Dependence Plots

### 7.5 Selecting the best predictive model

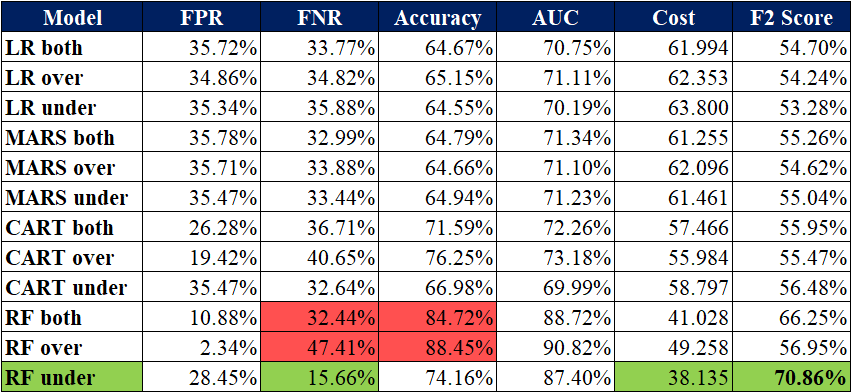
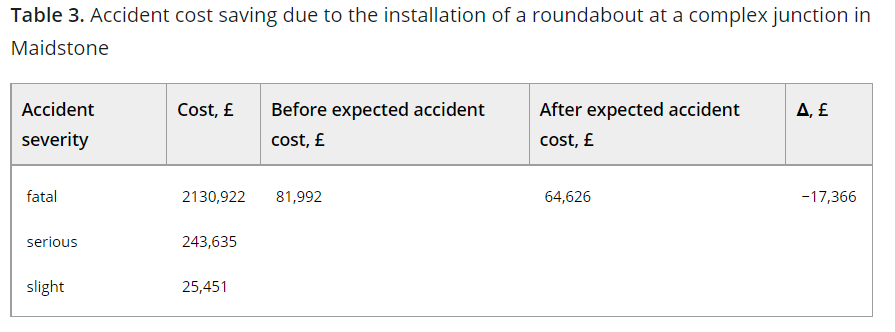
Figure 20 summarises the collated metrics for all models that were tested. From observation, the **Random Forest models** had the **highest accuracy** of the four model types. Those trained on the oversampled train set and the hybrid train set had by far the highest predictive accuracy, at 88.5% and 84.7% respectively. However, this should not be the sole measure for choosing a model. While predictive accuracy is high, the **FNR** is also very high at 47% and 32% respectively. As previously established, the use case requires minimization of FNR due to the **very high cost of false negatives**. Thus, the ideal model should have a **high F2 score**.

Figure : Model Performance Metrics

The Random Forest model trained on the under sampled train set had the desired qualities. It had by far the **highest F2 score of 0.71**, with the next closest model only achieving a score of 0.66. While predictive accuracy is slightly lower at 74.2%, this is acceptable as the FNR is less than half of the other two Random Forest models.

From economic value estimation models, it is estimated that the expected after accident cost is likely to be around **21% more expensive** than the cost of implementing preventive measures before an expected accident (Catalano et al., 2020). The cost savings per accident is shown in the table below:



Based on expected accident cost, the following equation was derived to estimate expected accident costs after factoring in weighted FNR and FPR.

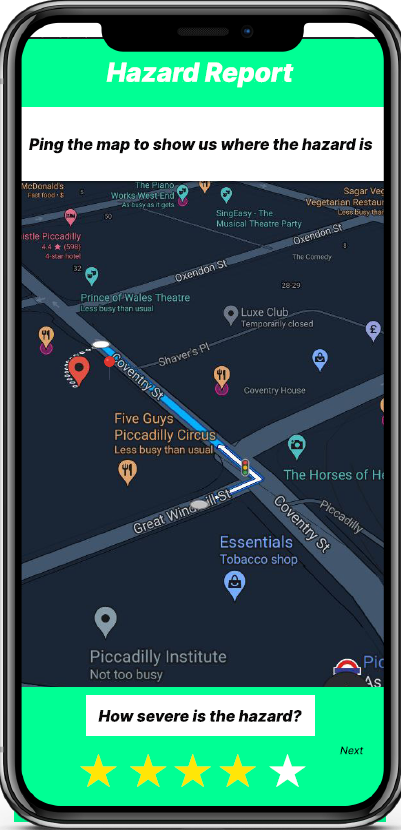
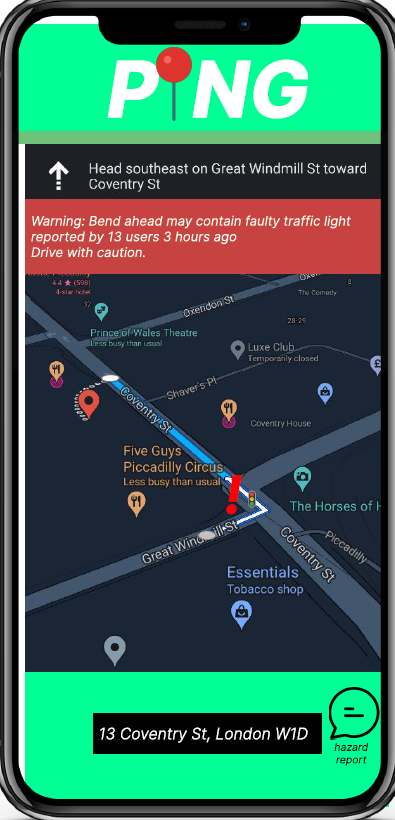
The ideal model should minimize cost in this equation; hence we chose the **Random Forest model trained on the under sampled training set, which also has the lowest false negative rate.**

## 8. Curated Recommendations to tackle Root Causes

### 8.1 Community Centric Hazard Sharing Application

Current accident reporting procedures in the UK require visiting a website and entering geographical coordinates which is **not convenient** for people who are driving to use. Furthermore, information is only routed directly to the government, which may result in a **delay in information** outreach to other road users. The Government could hence deploy a 2-in-1 navigation and hazard reporting application to streamline the open reporting process. The basic application functionalities include:

|  |  |
| --- | --- |
| **Functionality** | **Purpose** |
| Warn drivers of potential hazards/ heavy traffic in calculated routes based on other user’s reports. | * Provide **real-time information** so drivers can reroute |
| Warn drivers of serious accident hotspots and timings | * Alerting drivers to drive more cautiously at these timings, increasing **vigilance** |
| Report hazards - e.g. broken streetlights/ traffic light faults | * Allows the government to **promptly react** and provide appropriate remedial measures * Save the government cost and time of scheduled inspections |
| Near miss reporting | * Near misses have the **potential** to cause serious damage, by reporting them it allows the government to identify accident precursors so hazards can be **rectified immediately** |



Our models have allowed us to derive insights into factors that increase the probability of severe accidents. By allowing **open-reporting**, road users can gain **real-time access** to potential dangers and exercise greater caution to mitigate impacts of the root causes described earlier. In addition, a culture of open reporting promotes a driver’s **mindfulness** over safety, which will help foster a more responsible and safe driving community. (A UI Mock-up is shown in the figure on the left). Further screenshots of the UI are shown in Appendix 5.

### 8.2 Upgrading Traffic Fixtures at Accident Hotspots

One way to reduce occurrence of severe accidents is to make improvements to traffic fixtures at accident hotspots identified by the model. Certain accident hotspots may have higher rates of accident severity due to deficiencies in infrastructure. For example, traffic cameras may be insufficient and spaced too far apart, allowing for drivers to reach high speeds on the uncovered stretches of road. Improvements can be done by either upgrading existing fixtures or building new infrastructure.

Currently, both traffic lights and pedestrian crossings do not have countdown timers. The government could consider upgrading them by **adding countdown timers**. This will allow pedestrians to assess if they have enough time to cross the road. In addition, studies show that traffic lights with countdown timers would increase the likelihood of vehicles coming to a complete stop and dampen deceleration rates of vehicles (Donlon, 2017).

Our models identified that roads with higher speed limits have a higher likelihood of severe accidents occurring. To deter speeding, the government could **install more traffic cameras** along roads where the tendency to speed is higher, such as long and wide roads. Traffic cameras can significantly discourage road users from speeding, thereby reducing the likelihood of severe accident occurrences. This is corroborated by studies that show that traffic cameras have reduced excessive speeding by up to 91% and have saved 41% of people involved in fatal accidents at camera sites (Rospa, 2021).

The implementation of these traffic fixtures at accident hotspots would enhance traffic regulation and discourage dangerous road behaviour, which would reduce the frequency of severe accidents in these areas.

### 8.3 Legislative Amendments

The UK Government can revise existing traffic laws and implement new laws by **reducing speed limits** and **imposing harsher penalties** for violators of road safety. Research found that aggressive traffic enforcement reduces the number of vehicle crashes (Davis et al., 2006).

Our model found that areas with higher speed limits had more severe accidents. Reducing the speed limit at these hotspots can reduce the occurrence and severity of accidents. Research shows that reducing speed limits by 8 km/h below the optimal speed could reduce accidents by 50 percent and also reduce the frequency of fatal accidents (Chris, 2018).

Consequently, imposing higher penalties on those with higher class vehicles, for instance, Class A motorcycles (high engine capacity and power) can reduce the likelihood of these vehicle owners violating speed limits. Strict enforcement of driver penalty points is a strong deterrent as experienced drivers will risk disqualification of their driving license (Chen, 2019).

### 8.4 Increasing frequency of scheduled road inspections

The DfT could mandate local authorities to conduct scheduled road inspections at identified accident hotspots. While highway inspections are conducted frequently, smaller roads in rural areas are seldom inspected and maintained. For example, the West Sussex County Council inspects and maintains rural rights of way with volunteer help in a 15-month cycle (WSCC, 2021).

Having **semi-annual inspections** would allow for prompt detection of problems such as damaged or stolen signage and general damage to road surfaces such as potholes, which are potential hazards that could lead to severe accidents (nidirect, 2021). Road surfaces deteriorate on a daily basis, especially in urban areas where numerous vehicles have driven on them. Being able to detect road faults efficiently can reduce accidents from occurring on the road.

Our analysis found that a higher percentage of serious accidents occur during non-peak hours at night. A survey conducted by Yotta found that over a quarter of adults would avoid driving after dark due to poor lighting conditions on local roads (Yotta, 2018). One reason why street lighting is limited in rural areas is that road usage is minimal, so installing the same streetlights used in cities would be too costly and wasteful. As the lack of streetlights could lead to more severe accidents (Crabb et al., 2008), installing streetlights should be prioritised. Currently, the UK Lighting & Technology Board presides over matters relating to road lighting. The committee should strongly consider improving street lighting conditions in rural areas to reduce the number of serious traffic accidents. One possibility is to install **all-in-one solar streetlights** that use energy recycled with solar panels to power the lights at night. This solution reduces emissions and costs as the streetlights depend on solar energy rather than the power grid (Zafar, 2022).

### 8.5 Efficient emergency response service standby scheduling

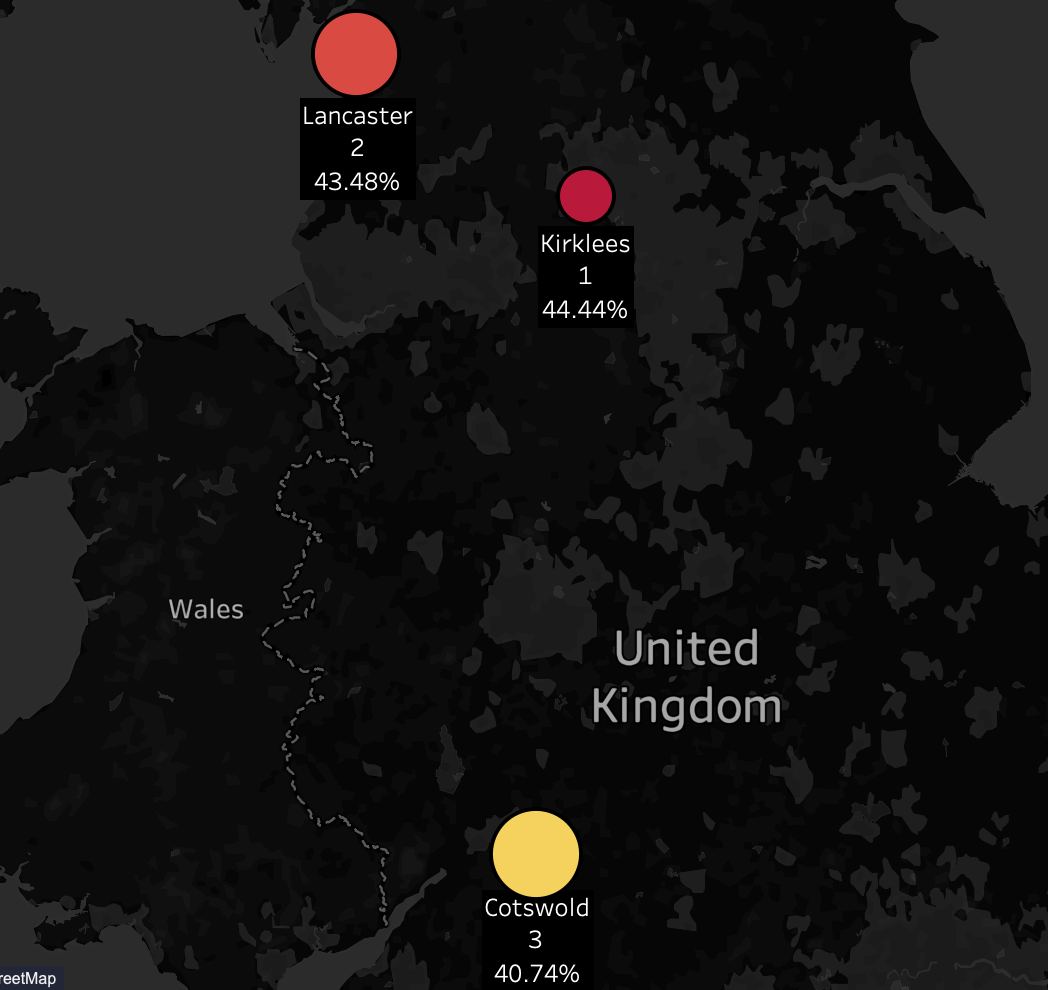
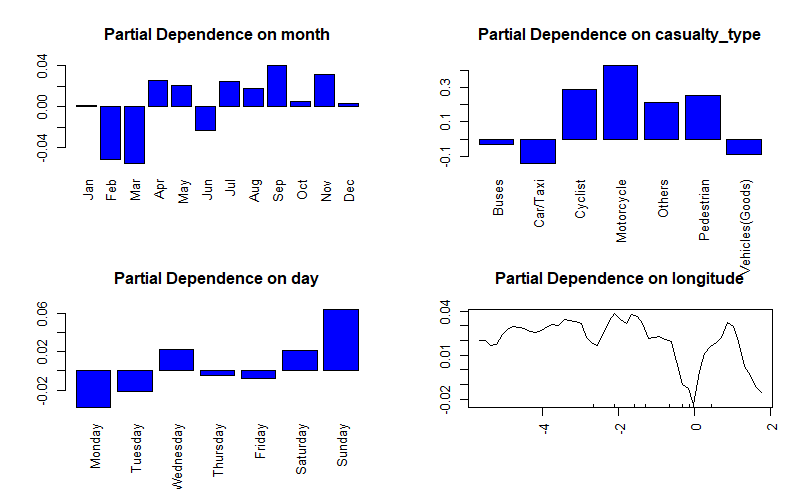
While accident prevention is ideal, accidents can still occur due to the vast majority of factors identified. As discussed in this report, the **delay in current emergency response times is alarming** (Toynbee, 2021). We recognise that **manpower allocation as an issue** due to declining birth rates in the UK. Through analytics, we are able to accurately identify severe accident hotspots. **Strategic deployment** derived from these analytics insights can help the government **reduce the arrival time** of paramedics which can reduce traffic fatalities. In fact, Zeng et. al (2019) found that every minute increase in EMS response time increased the probability of medium and severe crash injuries by 0.36% and 0.11% respectively (Zeng et., al 2019). We are able to identify the top 3 severe accident hotspots to be Kirklees, Lancaster, and Cotswold.

Figure : Partial Dependence Plots

The partial dependence plots reveal that the probability of serious accidents is much higher on Sundays and much lower during winter months. These insights can be utilised by the DfT to **optimise manpower allocation** on days with higher probability of traffic accidents.

## 9. Discussion

### 9.1 Benefit of the Analytics Solution

Our model is assessed to be a good predictor of identifying “Serious” accidents due its **low false negative rate**, indicating it as a viable tool for the DfT to use to identify potential serious accident hotspots. By isolating important factors and identifying those that are trivial(but initially thought to be important), the DfT can **efficiently allocate its resources** to focus on factors flagged important by the model and **reduce redundancy**. Knowledge of important variables also allows us to make **curated recommendations** to address the limitations of existing measures as described in Section 8.

Based on the scope of our analysis, the overall selected model error is around 25%, which might not bring confidence to the user, and making it seem that the models cannot completely predict accident severity. However, with a low false negative rate, it does indicate the potential of using these models as **preliminary risk assessment** in identifying a mix of factors that could cause severe accident. Furthermore, model accuracy rates can be improved through using more advanced techniques like deep learning, which claims to be highly accurate.

### 9.2 Limitations of Analysis

Sampling Techniques

Our team relied on the fact that there were no data collection mistakes which could have been present. Due to a class imbalance issue, our team tried various sampling techniques to increase the weight of the minority class (“Serious accidents” are the class of concern). Our team assumed that utilising sampling techniques will increase classification accuracy **without any potential consequences**. However, these are naive resampling methods because they assume nothing about the data and no heuristics were used. We acknowledge that **better sampling techniques** can be used to improve the accuracy of the model. Even so, synthetic data manipulation has to be used carefully as the sampled data may contain values that do not make **practical sense**. Furthermore, synthetic data can never completely replace real acquired data.

Model Deployment and Retraining

Our predictive models were built using 2017-2019 data, over this time, road infrastructure could have changed. Road traffic accident data is **dynamic**; hence the predictive model may become obsolete over time due to the presence of new input data being generated (Kavikondala et al, 2019). Thus, it is imperative to **treat model deployment as an iterative process**. Predictive models need to be **retrained** if data distributions deviate significantly from those in the original training set. Manual retraining can be time-consuming due to a large number of data points (Patruno, 2019). Hence, a more effective solution would involve the use of **automated retraining**, where the model is constantly tested against new data and a combination of new and old data is automatically used to build a new model once model[[5]](#footnote-5) drift is detected (Bright et al., 2021).

## 10. Conclusion

The advancement in technology presents an opportunity for the UK government to make use of machine learning models to deal with pressing societal issues in the UK. Our proposed analytics solution provides **key insights** to the UK government on measures that can be taken to reduce the occurrence of severe accidents.

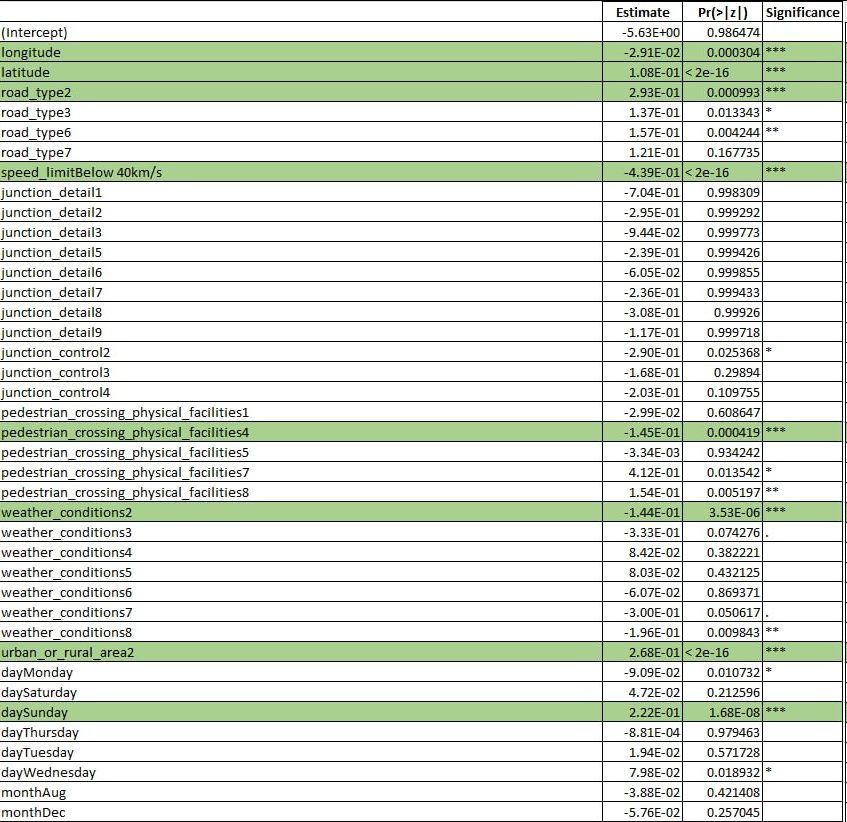
Our models identified key predictors of serious road accidents and were able to predict accident severity. The team then developed **targeted recommendations** for the DfT’s consideration. The government should consider these recommendations to take **pre-emptive and corrective actions** to reduce the risk of serious road accidents - especially at identified accident hotspots to ensure cost efficiency.

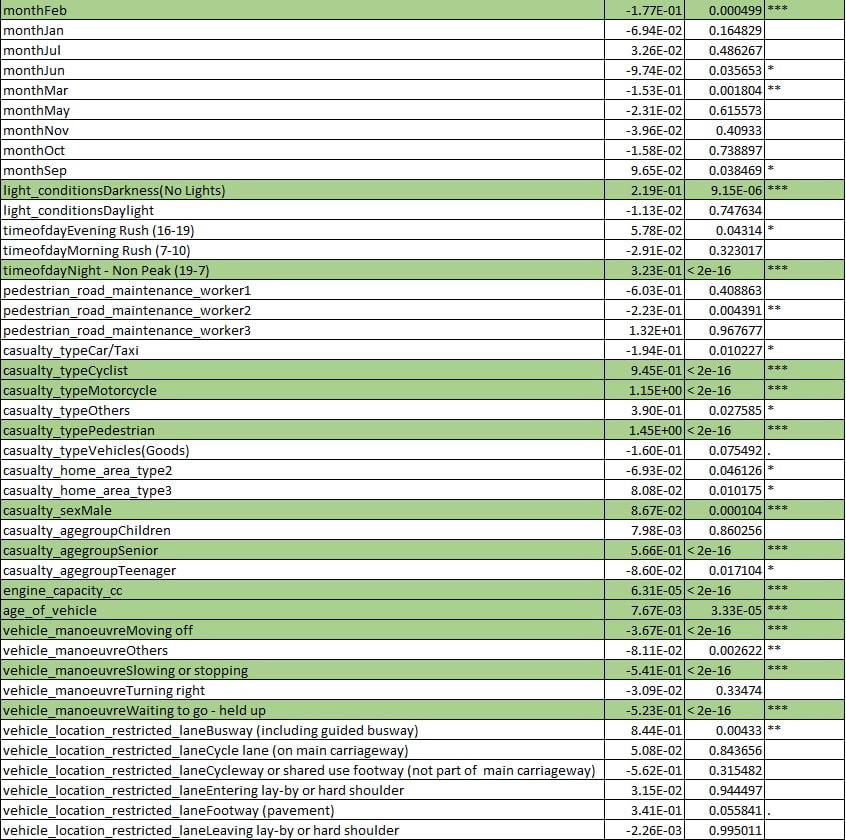
It is critical to assess and improve the long-term viability of the presented recommendations, especially since receptiveness of citizens plays an important role in determining overall success. Due **to dynamic changes** in accident data over the years, it is necessary for the DfT to **constantly update** its models to ensure predictive accuracy.

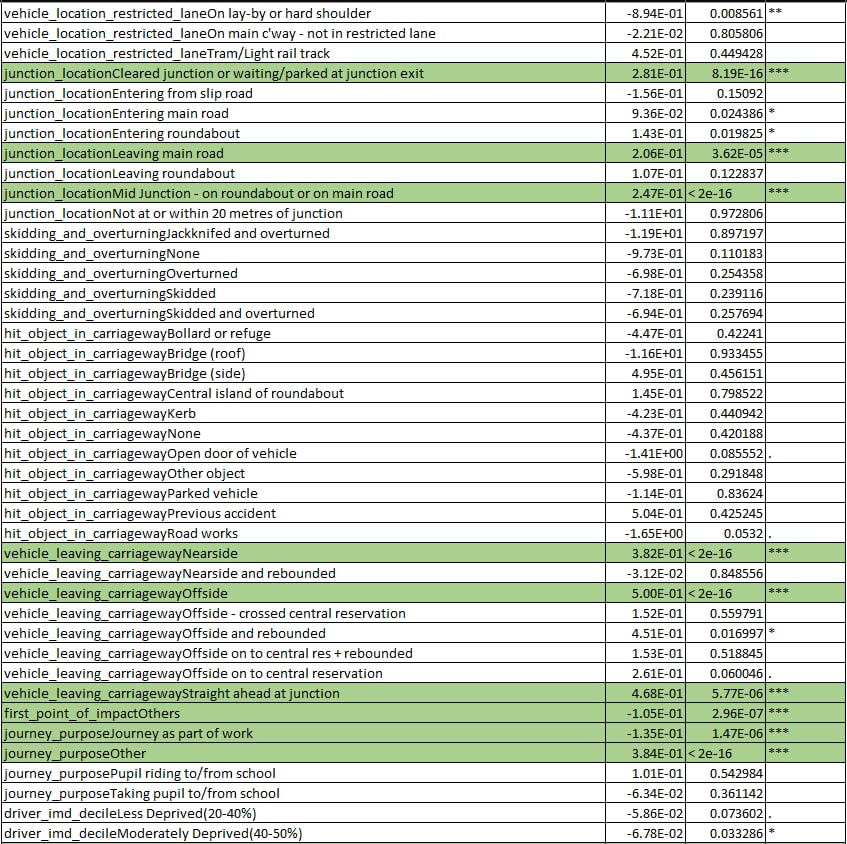
## Appendices

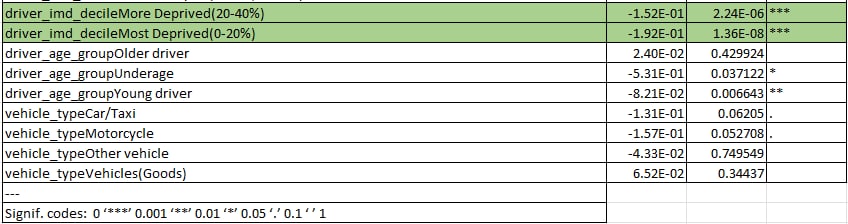
### Appendix 1: Logistic Regression Results

**Results of Logistic Regression with Backward Elimination**

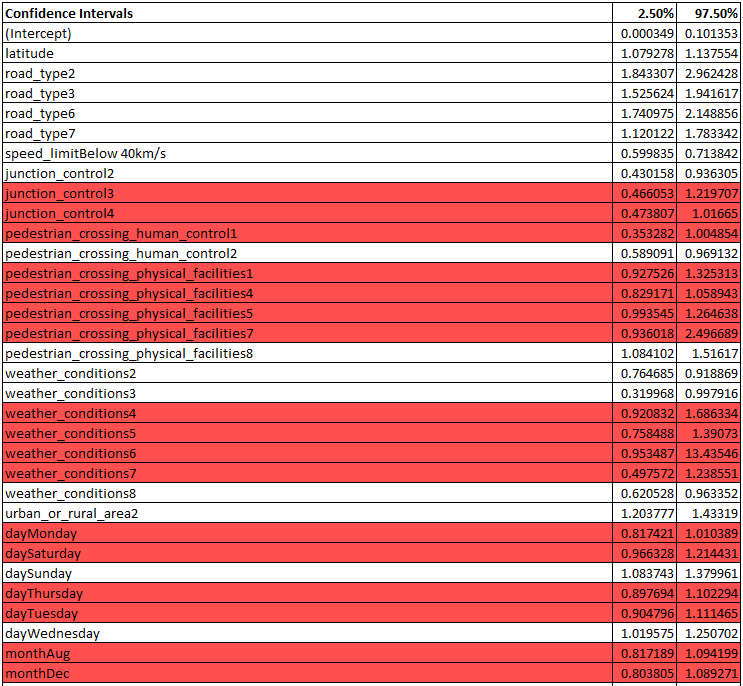


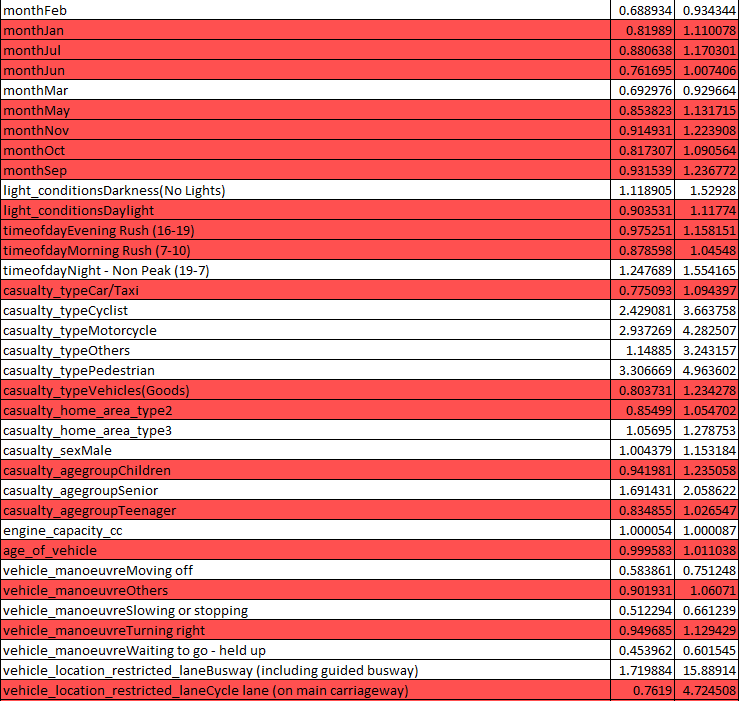


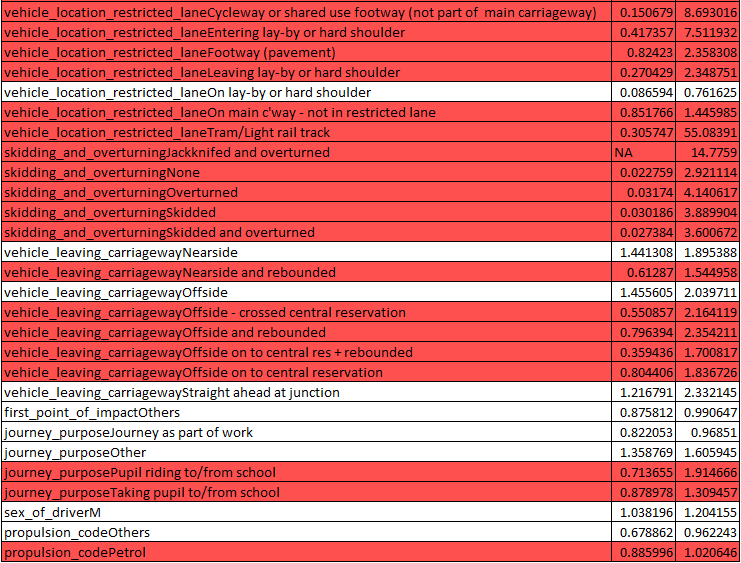




**Confidence Intervals for Logistic Regression with All Variables**





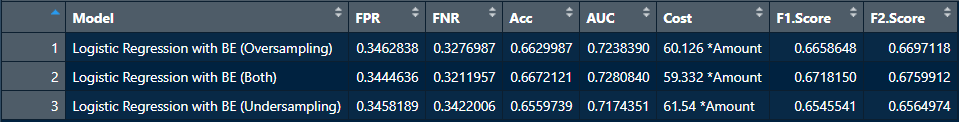


Note: Highlighted rows indicate that the CI includes 1 and the variable is not significant

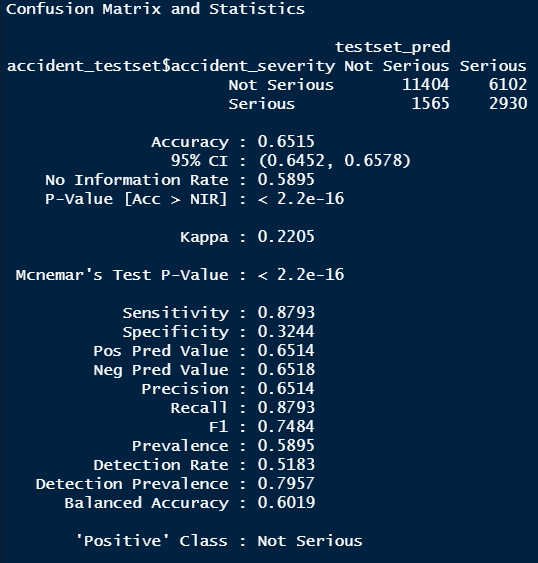
**Test set Metrics for Logistic Regression**



**Train set Metrics for Logistic Regression**



**Confusion Matrix for Logistic Regression trained on Over Sampled Train Set**

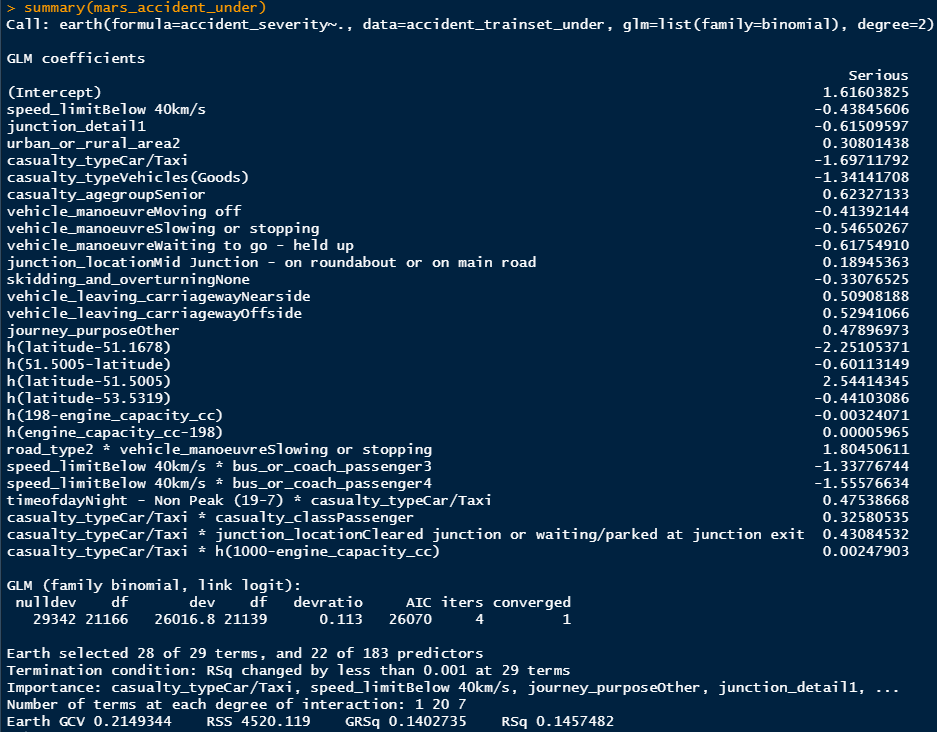


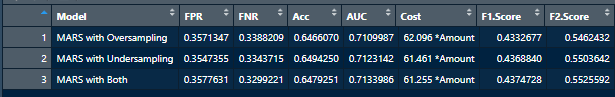
**Confusion Matrix for Logistic Regression trained on Under Sampled Train Set**

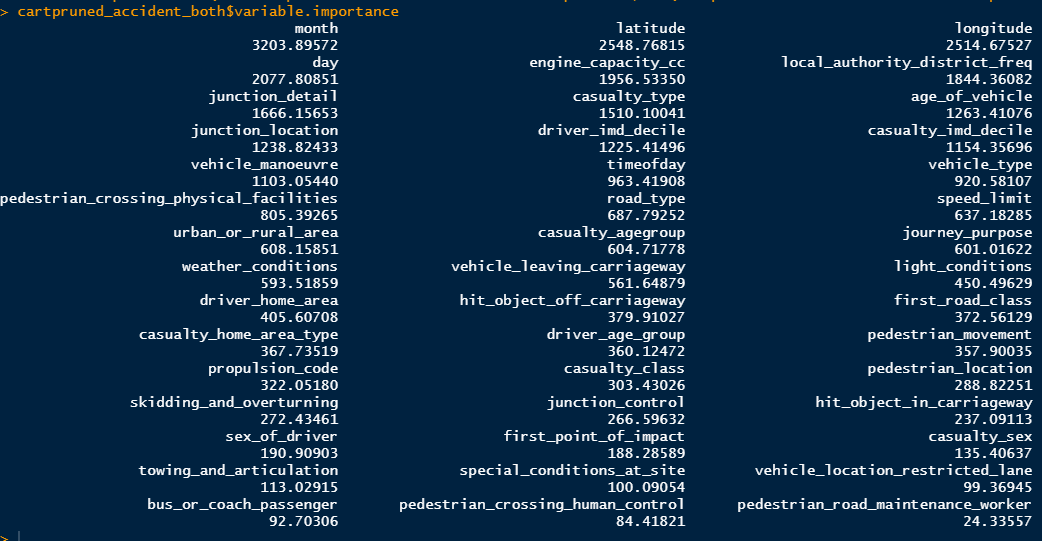
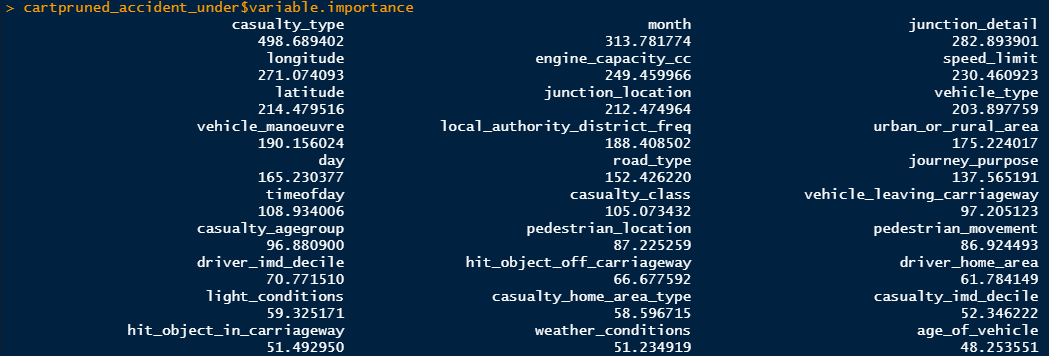
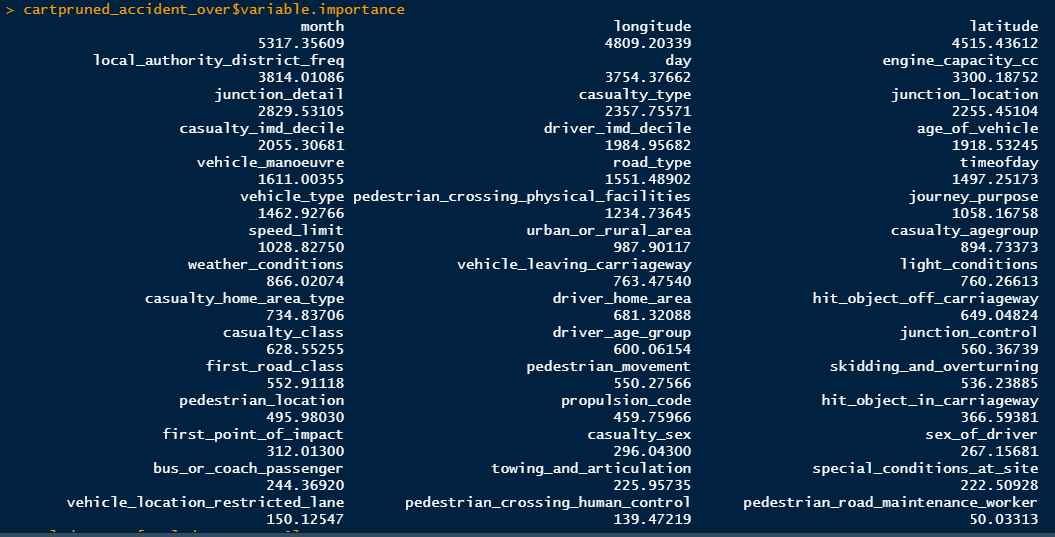
Text

Description automatically generated

### Appendix 2: MARS results





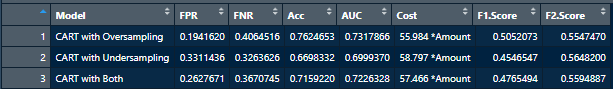
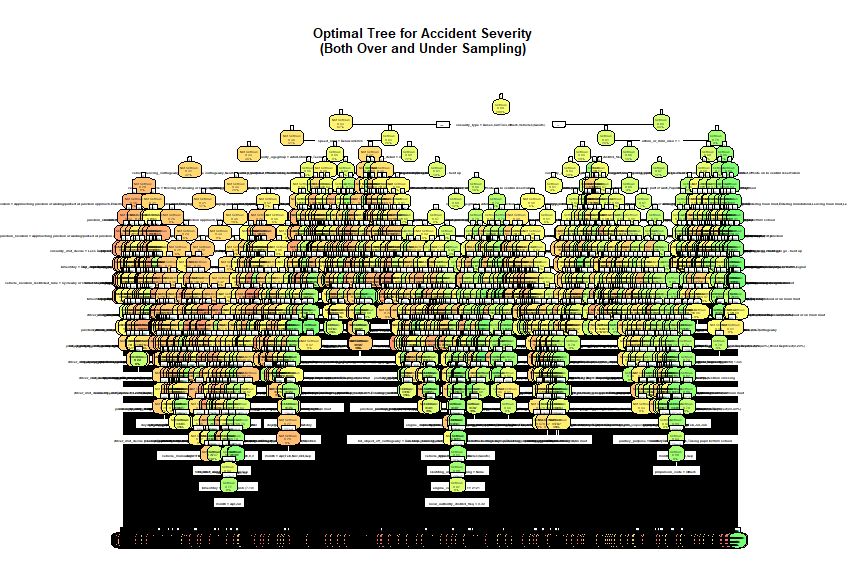


### Appendix 3: CART Results

A picture containing graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated



### Appendix 4: Random Forest results

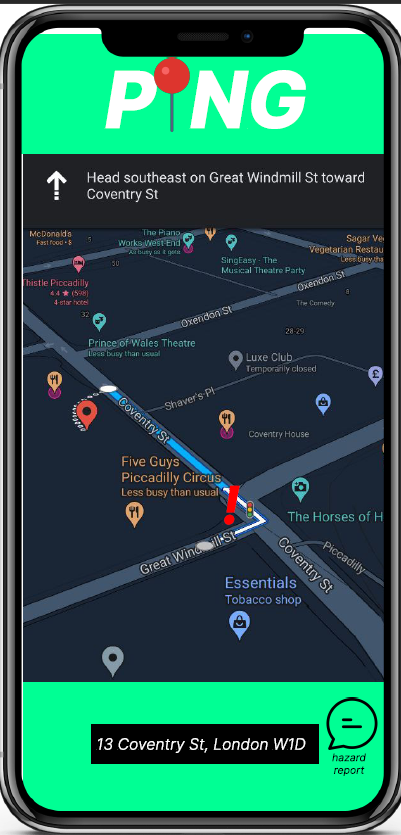
Text

Description automatically generated

### Appendix 5: Sample Mock Ups for Hazard Reporting App

Graphical user interface, application

Description automatically generated



## References

Aschalew K., Mahbub H., *Factors associated with traffic crashes on urban freeways, Transportation Engineering*, Volume 2, 2020, 100014, ISSN 2666-691X, <https://doi.org/10.1016/j.treng.2020.100014>. (<https://www.sciencedirect.com/science/article/pii/S2666691X20300154>)

Auto, H. (2022, January 16). UK points to lifting of covid-19 curbs. The Straits Times. Retrieved February 15, 2022, from <https://www.straitstimes.com/world/europe/uk-points-to-lifting-of-covid-19-curbs>

BBC. (2019, March 11). *Ambulance wait times longer in rural areas*. Rural Services Network. Retrieved April 3, 2022, from https://www.rsnonline.org.uk/ambulance-wait-times-longer-in-rural-areas

Billiau, S. (2021, June 13). *From Scratch: Permutation Feature Importance for ML Interpretability*. Towards Data Science. Retrieved April 3, 2022, from https://towardsdatascience.com/from-scratch-permutation-feature-importance-for-ml-interpretability-b60f7d5d1fe9

Brand, S., Otte, D., Mueller, C. W., Petri, M., Haas, P., Stuebig, T., Krettek, C., & Haasper, C. (2012). Injury patterns of seniors in traffic accidents: A technical and medical analysis. World journal of orthopedics, 3(9), 151–155. Retrieved April 2, 2022, from <https://doi.org/10.5312/wjo.v3.i9.151>

Bright, J., Cerè, A., Schinas, G., & Heilker, T. (2021, November 2). *Automate model retraining with Amazon SageMaker Pipelines when drift is detected | Amazon Web Services*. Amazon AWS. Retrieved April 3, 2022, from https://aws.amazon.com/blogs/machine-learning/automate-model-retraining-with-amazon-sagemaker-pipelines-when-drift-is-detected/

Cabrera-Arnau, C. (2020, April). Uncovering the behaviour of road accidents in urban areas. *Royal Society Open Science*, *7*(4). 10.1098/rsos.191739

CarBikeTech Team. (2014, September 19). *Engine Capacity (CC): Engine Volume/Displacement?* CarBikeTech. Retrieved April 3, 2022, from https://carbiketech.com/engine-capacity-cc/

Catalano, M., Galatioto, F., & Shaikh, N. (2020, October 2). Accident-related cost analysis and decision-making support through econometric modelling. *IET Intelligent Transport Systems*, *14*(12), 1573-1581. The Institution of Engineering and Technology. https://doi.org/10.1049/iet-its.2020.0093

Chen, T. (2019, November). Evaluation of Penalty and Enforcement Strategies to Combat Speeding Offences among Professional Drivers: A Hong Kong Stated Preference Experiment. *Accident: Analysis and Prevention*, *135*(1). 10.1016/j.aap.2019.105366

Cherian V, Umesh S. (2007) *Passenger Vehicle Occupant Fatalities by Day and Night – A Contrast.* Retrieved April 2, 2022, from <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810637>

Chris M. (2018, October 1). *Lowering speed limits can reduce crashes*. State Smart Transportation Initiative. Retrieved April 3, 2022, from <https://ssti.us/2018/10/01/lowering-speed-limits-can-reduce-crashes/#:~:text=Using%20regression%20analysis%20and%20considering,drop%20by%20almost%20as%20much>

Crabb G, & Crinson L.. (2008). *The Impact of Street Lighting on Night-Time Road Causalities*. Retrieved April 2, 2022, from <https://saferroadsconference.com/wp-content/uploads/2016/05/Geoff-Crabb-The-Impact-of-Street-Lighting.pdf>

Davis JW, Bennink LD, Pepper DR, Parks SN, Lemaster DM, Townsend RN. Aggressive traffic enforcement: a simple and effective injury prevention program. J Trauma. 2006 May;60(5):972-6; discussion 976-7. doi: 10.1097/01.ta.0000204031.06692.0f. PMID: 16688057. Retrieved April 3 , 2022, from <https://pubmed.ncbi.nlm.nih.gov/16688057/>

Department for Transport. (2013). *Action for Roads: A network for the 21st century*. GOV.UK. Retrieved April 3, 2022, from <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/212590/action-for-roads.pdf>

Department for Transport. (2020, September 30). *Reported road casualties annual report 2019* . Retrieved April 1, 2022, from <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/922717/reported-road-casualties-annual-report-2019.pdf>

Donlon, M. (2017, November 2). *Traffic signal countdown timers may reduce fatal car accidents*. Engineering360. Retrieved April 2, 2022, from <https://insights.globalspec.com/article/6974/traffic-signal-countdown-timers-may-reduce-fatal-car-accidents>

Febres, J. D., Mohamadi, F., Mariscal, M. A., Herrera, S., & Herrero, S. G. (2019, December 2). *The Role of Journey Purpose in Road Traffic Injuries: A Bayesian Network Approach*. Hindawi. Retrieved April 3, 2022, from https://www.hindawi.com/journals/jat/2019/6031482/

Juan D., Fatemeh M., Miguel. A., S. Herrera, & Susana G.. (2019, December 2). *AI-Powered Research Tool*. Semantic Scholar. Retrieved April 2, 2022, from <https://www.semanticscholar.org/>

Kavikondala, A. (2019, October 5). *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*. International Journal of Innovative Technology and Exploring Engineering (IJITEE). Retrieved April 3, 2022, from https://www.ijitee.org/wp-content/uploads/papers/v8i12/L33221081219.pdf

Kumar, S. (2021, March 2). *Stop One-Hot Encoding your Categorical Features — Avoid Curse of Dimensionality*. Medium. Retrieved April 1, 2022, from https://medium.com/swlh/stop-one-hot-encoding-your-categorical-features-avoid-curse-of-dimensionality-16743c32cea4

*Machine Learning - Classification And Regression Trees (CART) - Q*. (2021, August 18). Q Wiki. Retrieved April 2, 2022, from https://wiki.q-researchsoftware.com/wiki/Machine\_Learning\_-\_Classification\_And\_Regression\_Trees\_(CART)

*Motorcycle safety*. NHTSA. (n.d.). Retrieved April 2, 2022, from <https://www.nhtsa.gov/road-safety/motorcycles>

*New index reveals cost of complacency on UK road death*. Towards Zero Foundation. (2021, November 18). Retrieved April 3, 2022, from <http://www.towardszerofoundation.org/new-index-reveals-cost-complacency-uk-road-death/>

O'Neill, A. (2021, December 16). *United Kingdom - Urbanization 2020 | Statista*. Statista. Retrieved April 2, 2022, from <https://www.statista.com/statistics/270369/urbanization-in-the-united-kingdom/>

Patruno, L. (2019, December 16). *The Ultimate Guide to Model Retraining*. KDnuggets. Retrieved April 3, 2022, from https://www.kdnuggets.com/2019/12/ultimate-guide-model-retraining.html

*Pedestrian countdown signals*. Pedestrian Countdown Signals | Traffic Choices - aiding traffic scheme decisions. (n.d.). Retrieved April 2, 2022, from <https://www.trafficchoices.co.uk/traffic-schemes/countdown-signals.shtml>

*Problems with roads and streets*. nidirect. (n.d.). Retrieved April 3, 2022, from <https://www.nidirect.gov.uk/information-and-services/travel-transport-and-roads/problems-roads-and-streets>

Rabone, E. (2022, January 25). *UK motorcycle accident statistics: The Complete 2022 Guide Begin Motorcycling*. Begin Motorcycling. Retrieved April 2, 2022, from <https://begin-motorcycling.co.uk/uk-motorcycle-accident-statistics/>

*Road safety factsheet - ROSPA*. ROSPA. (2021, September). Retrieved April 2, 2022, from <https://www.rospa.com/media/documents/road-safety/factsheets/overtaking-factsheet.pdf>

Sangani, R. (2021, August 4). *Dealing with features that have high cardinality | by Raj Sangani*. Towards Data Science. Retrieved April 1, 2022, from https://towardsdatascience.com/dealing-with-features-that-have-high-cardinality-1c9212d7ff1b

SemiColonWeb. (2016, June 19). *Publications*. ETSC. Retrieved April 3, 2022, from <https://etsc.eu/how-traffic-law-enforcement-can-contribute-to-safer-roads-pin-flash-31/>

*Speed.* IIHS. (2021, May). Retrieved April 3, 2022, from <https://www.iihs.org/topics/speed#effects-of-speed-limits-on-safety>

Statista Research Department, & 10, M. (2022, March 10). *Road accident casualty costs by type in Great Britain 2019*. Statista. Retrieved April 2, 2022, from <https://www.statista.com/statistics/325832/casualty-costs-from-road-accidents-in-great-britain-uk-by-severity/>

Stevens, A. P. (2018, October 11). *Here's what puts teen drivers at greatest risk of a crash*. Science News for Students. Retrieved April 2, 2022, from <https://www.sciencenewsforstudents.org/article/heres-what-puts-teen-drivers-greatest-risk-crash>

Tamara H., Fred W., *Improving the effectiveness of road safety campaigns: Current and new practices,* IATSS Research, Volume 34, Issue 2, 2011, Pages 80-86, ISSN 0386-1112, Retrieved April 3, 2022, from <https://doi.org/10.1016/j.iatssr.2011.01.003>. (<https://www.sciencedirect.com/science/article/pii/S0386111211000045>)

Thomas R, Janina S, *Turning accidents between cars and trucks and cyclists driving straight ahead*,Transportation Research Procedia,Volume 25,2017,Pages 1946-1954,ISSN 2352-1465, Retrieved April 2, 2022, from <https://doi.org/10.1016/j.trpro.2017.05.219>. (<https://www.sciencedirect.com/science/article/pii/S2352146517305203>)

Toynbee, P. (2021, April 20). Britain's falling birthrate will damage our society – and it's not just covid to blame . The Guardian. Retrieved February 15, 2022, from <https://www.theguardian.com/commentisfree/2021/apr/20/britain-falling-birthrate-covid-pandemic-conservatives-removed-support-for-parents>

What is the number one cause of motorcycle accidents? JD Supra. (2020, October 28). Retrieved April 2, 2022, from https://www.jdsupra.com/legalnews/what-is-the-number-one-cause-of-75378/

WHO. (2021, June 21). *Road traffic injuries*. WHO | World Health Organization. Retrieved April 3, 2022, from https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

WSCC. (2021, December 3). *Inspection and maintenance schedules*. West Sussex County Council. Retrieved April 1, 2022, from https://www.westsussex.gov.uk/land-waste-and-housing/public-paths-and-the-countryside/public-rights-of-way/inspection-and-maintenance-schedules/

Yannis, G., Nikolaou, D., Laiou, A., Sturmer, Y. A., Buttler, I., & Karpa, D. J. (2020). *Vulnerable road users: Cross-cultural perspectives on performance and attitudes*. https://www.sciencedirect.com/science/article/pii/S0386111220300716#:~:text=Pedestrians%2C%20pedal%20cyclists%20and%20motorcyclists,impact%20of%20a%20road%20crash.

Yotta. (2018, October 11). *Over a quarter of UK drivers avoid driving after dark due to poor street lighting*. Open Access Government. Retrieved April 1, 2022, from https://www.openaccessgovernment.org/driving-after-dark-due-to-poor-street-lighting/53164/

Zafar, S. (2022, February 28). *Unending Benefits of Solar Street Lights*. BioEnergy Consult. Retrieved April 1, 2022, from https://www.bioenergyconsult.com/solar-street-lights/

Zeng, Q., Gu, W., & Zhang, X. (2019, June). *https://www.researchgate.net/publication/331518506\_Analyzing\_freeway\_crash\_severity\_using\_a\_Bayesian\_spatial\_generalized\_ordered\_logit\_model\_with\_conditional\_autoregressive\_priors*. https://www.researchgate.net/publication/331518506\_Analyzing\_freeway\_crash\_severity\_using\_a\_Bayesian\_spatial\_generalized\_ordered\_logit\_model\_with\_conditional\_autoregressive\_priors

1. The frequency(how often the category appears) of categories are utilised as labels. In the cases where the frequency is related somewhat with the target variable, it helps the model to understand and assign the weight in direct and inverse proportion, depending on the nature of the data. Note that this is done only on the train set as information outside the train set CANNOT be used to train the model. [↑](#footnote-ref-1)
2. A lower AIC score is desirable as it means there is less information loss. [↑](#footnote-ref-2)
3. Odds ratio is the ratio of odds of an event “Serious” in the presence of “Predictor” and the odds of event “Serious” in the absence of “Predictor”. It can be used to calculate probability of a “Serious” accident. [↑](#footnote-ref-3)
4. Relationships that cannot be described using a straight line(consists of curves or bent) [↑](#footnote-ref-4)
5. Model Drift refers to a model predictive performance degrading over time due to a change in environment that violates the model’s assumptions. [↑](#footnote-ref-5)