ST309 Group Project Report

# The Severity of Road Accidents:

*An inside look into the attributes of fatal accidents in the UK*


**Abstract**

There have been multiple analyses done on road accidents in the UK, but we have them to be generally insufficient in providing a robust analysis of the rich dataset at hand. In our report, we attempt to identify the main causes of severe road accidents and casualties through a range of supervised and unsupervised learning techniques. we have found that there is no single variable that determines the severity of an accident/casualty, but rather several characteristics that may influence the likelihood of a severe accident/casualty. Some of these characteristics are: a vehicle getting rear-ended is less likely to result in a severe accident, vehicles driving in bad road conditions are no more likely to get into severe accidents, and the probability of a casualty getting severely hurt increases almost linearly with age. Despite these findings, our models' low explanatory power suggests that there may be other tangible (e.g. vehicle's travelling speed) and intangible (e.g. human error) factors that contribute to the variation in accident/casualty severity



***Contribution:*** *33.33%, 33.33%, 33.33%*

# Table of Contents

# 1. Introduction

In the UK, there are more than 170,000 road casualties a year due to road accidents. Additionally, more than 1% of these casualties are fatal with about 1,800 deaths in an average year (UK Department for Transport, 2017), and an even greater number suffer from serious injuries.

In our study, we attempt to predict the likelihood of a given accident being severe[1], given certain attributes concerning the driver, their vehicle, and other external factors. This is important because it can enable relevant stakeholders to take both reactive and preventative measures to mitigate the seriousness of accidents, such as through weather warnings, regulatory requirements on vehicles, insurance payouts, and age-related regulations.

Due to the multi-dimensionality of our data set, we have chosen to employ data analytics procedures to generate a predictive model that would give us the likelihood of a severe accident, given that an accident has occurred. Ultimately, we intend to generate a model that can be applied to many situations.

This study would ideally allow us to achieve several goals:

- Drivers could pick the vehicle or mode of transport that is the safest for them and be warier of the road conditions under which accidents are more likely to be severe, avoiding them if possible
- Life insurers could consider relevant characteristics of the driver and the environmental conditions before charging the fair insurance premium
- Policymakers could enact or amend the present policies for the benefit of all road users

There have been some attempts made by different people to explore the data - majority of the analyses have been centred around data visualisation and a few have tried to determine the severity of an accident using a heatmap. However, we have found the existing analyses to be generally insufficient in providing a robust analysis of the rich dataset at hand.

Our data was obtained entirely through the UK Department of Transport's Road Safety Data and spans the years 2005 to 2017. The original dataset comprised three different documents, which were focused around **Accidents**, **Vehicles**, and **Casualties** involved in the accidents respectively. Due to the way the data was sorted across the three documents, we had to perform the analysis separately on each set of data. This proved wise as we obtained insightful results across the datasets and had the opportunity to corroborate our results with one another.

[1] A severe accident is one in which there is at least one casualty who suffers from severe or fatal injuries, which will be further defined in Section 3: Description of Variables.

In bringing together our results, we have uncovered important implications that we believe can be productively utilised by different stakeholders for varying purposes. Having tried out several formats, we have found the most effective way of organising the report to be by the type of analysis conducted rather than by dataset. For each section, we begin first by explaining the analysis tool employed before exploring the results for each dataset. In this way, we avoid repetition and focus our attention on the results drawn from each dataset.

Despite the similarity in methodology, there are distinct purposes and goals of each dataset. With the **Accidents** dataset, we are concerned with the characteristics of accident sites that have the propensity to be hotspots for severe accidents. In doing so, we hope that the results will be used by urban road planners and other relevant stakeholders to design safer roads for road users. While accidents may sometimes be unavoidable, road designs can and should be made to reduce the severity of such accidents when they inevitably do occur. With the **Vehicles** dataset, we look at variables such as which part of the vehicle was hit during first impact and the age of the vehicle. It is our hope that these results would help car manufacturers design sturdier vehicles and offer scheduled maintenance to improve their vehicle's ability to withstand accidents should they happen. Finally, the **Casualties** dataset would hopefully allow first responders and emergency wards in hospitals to effectively attend to casualties based on data centred around the victims' physical characteristics (such as age and gender) and the circumstance of the accident as reported by bystanders (such as whether the victim was a passenger or driver).

## 2. Description of Variables

Due to the breadth of our data, there are many variables at use in our report (Department for Transport, 2011). For ease of readability, we have explained the variables and their meanings meticulously in *Appendix A1: Full List of Variable Names & Descriptions.*

Furthermore, owing to the overly-granular classification made by the UK Department for Transport, we have taken the liberty to reclassify some variables (eg. instead of having ten different classifications for motorcycles of varying engine capacity, we have grouped them into a single class).

The two main predictor variables in our datasets are *Accident Severity* and *Casualty Severity*[2]. The datasets obtained from the UK Department for Transport classified accident/casualty severity into three groups: Fatal, Serious and Slight. For our entire analysis, we reclassified Fatal and Serious accidents/casualties as *Severe* accidents/casualties, while Slight accidents/casualties had been reclassified as *Non-Severe* accidents/casualties.

[2] *Accident severity* was used in the **Accidents** and **Vehicles** analysis, while *casualty severity* was used in the **Casualties** analysis.

A *Severe* casualty refers to any casualty who had died or suffered any other serious injury such as a broken neck, internal injuries, concussions, or burns, amongst others. A *Non-Severe* casualty refers to any casualty who suffered slight injuries such as shallow cuts, bruising, or slight shock. Finally, a *Severe* accident is one in which at least one *Severe* casualty was recorded, while a *Non-Severe* accident implies that there were no *Severe* casualties at all. This reclassification allows us to handle *Accident Severity* and *Casualty Severity* as a binary variable, as opposed to having three different indicators.

## 3. General Data Analysis

### Methodology

In this section, we describe the methodology we employed in the analysis of our data. Due to the nature of our dataset, we performed three separate analysis in parallel (**Accidents, Vehicles, Casualties**) with the accident index being the bridge across the datasets. The table below explains our data in greater detail:

| **Name of Dataset** | **Description of Dataset** |
|---|---|
| **Accidents** | The **Accidents** dataset was focused primarily on the conditions surrounding the accident itself, rather than the individual vehicles/casualties involved. There are 1,990,447 observations from 2005 to 2017. *Note: accident is classified as severe if there is at least one severe casualty.* |
| **Vehicles** | The **Vehicles** dataset looks at the condition of the vehicles involved in the accident, with each observation (row) corresponding to a vehicle. There are 2,616,061 observations from 2005 to 2017. *Note: a vehicle is classed as 'Severe' if it was involved in an accident that was classed as 'Severe'. The goal of the Vehicles dataset is to determine the likelihood a vehicle gets into a severe accident, rather than the likelihood of the vehicle being severely damaged.*[3] |
| **Casualties** | The **Casualties** dataset is centred around each individual casualty involved in the accident. There are are 2,699,389 observations from 2005 to 2017. *Note: casualties who were not injured were not recorded at all.* |

**Step 1: Cleaning the Data**

We removed variables (columns) that (1) we deemed redundant, (2) had high collinearity with other variables or (3) had too many missing values. This allowed us to avoid removing too many rows. For variables with a small number of missing values, we attempted to use K nearest neighbours (K-NN) to impute these values. K-NN imputation showed that the assignment of the missing values was as good as random i.e. the missing values imputed did not systematically belong to any specific type of accident, vehicle or casualty. Furthermore, K-NN imputation also resulted in inaccuracies; for example, some car occupants ended up being classified as

[3] There was no information on vehicular damage and the accident severity was extracted from the accidents dataset and matched by accident index. This is a potential disadvantage that we will address in the Shortfall section of our report.

boarding/alighting passengers (which is applicable to only bus or coach passengers). For such cases, we decided to omit the observations (rows). Further elaboration is given for each dataset below:
**Accidents:** No special consideration needed, just omitted NAs.

**Vehicles:** A notable observation made for the **Vehicles** dataset was that the vehicle ages of bicycles were missing for most observations classed as Bicycles. To remedy this, we replaced all missing values of bicycles' age with the mean age of the complete bicycle cases, which turned out to be 2.24. The purpose of doing this is primarily because we want to retain our bicycle observations.

**Casualties:** Two variables, (1) *Pedestrian Road Maintenance Worker* and (2) *Casualty Home Area Type* were removed as they were deemed largely redundant in affecting casualty severity. Furthermore, 52% and 14.6% of their values were missing respectively. The variables *Accident Index, Vehicle Reference and Casualty Reference* were removed because they are simply indices that have no predicted power. After cleaning the data, we were left with nine variables in total, including our dependent variable: *Casualty Severity.*

**Step 2: Overview of Data**

We began each analysis by providing an overview and summary of the data in order to get a big picture of the data that we are working with. This was mainly done through data visualisation packages on R.

**Step 3: Creating Training & Testing Sets**

The next step we took was to partition out training and test sets from the larger dataset. While each of the three datasets contained over two million observations, our training and testing sets contained 100,000 and 150,000 observations respectively. The reason for doing this is purely for practical purposes pertaining to computational efficiency. We created our training and testing data sets through the following steps:

1. Separate the data set into two groups: severe and non-severe observations
2. For the **training set**: extract randomly from the severe and non-severe observations to make up 100,000 observations. We do this in a ratio of severe-to-non-severe observations that is higher than that of our full dataset (highlighted below). This is done to to train our model on more severe accidents so that it would be better adapted at identifying such accidents.
3. For the **testing set**: extract randomly from the severe and non-severe observations to make up 150'000 observations. We do this in a ratio of severe-to-non-severe observations that is the same as the full dataset, to create a representative sample of our data.

**Accidents & Vehicles**: For both the **Accidents** and **Vehicles** datasets, the training set comprised 25% severe accidents and 75% non-severe accidents, while the testing set comprised about 14% severe accidents to replicate the full dataset's proportion. We found a 25-75 ratio for the training set to be optimal in producing models with the highest accuracy rates, whilst maintaining a reasonable false positive and false negative rate.

**Casualties:** The training set used comprised 50% severe accidents, and 50% non-severe accidents, which was the optimal in raising the accuracy rates and lowering the false positive and false negative rates. The testing set comprised about 12.8% severe accidents - the same proportion as the full dataset.

Two other categorical datasets were also isolated from the main dataset to analyse pedestrian and car casualties in more detail (refer to file CasualtiesAnalysis.R):

1. Dataset with only pedestrian casualties
2. Dataset with only car casualties (made up of car occupants and taxi/private hire car occupants).

**Step 4: Logistic Regression**

Rather than using a linear regression model, we have chosen to use the logistic regression model instead as it ensures that the predictor variable takes a value between 0 and 1 - a useful property for interpreting probabilities. This is done using a logistic function to model a binary dependent variable like accident/casualty severity.

To improve upon the model, we used a stepwise regression, which selects explanatory variables through minimising the Akaike Information Criteria[4] (AIC) by removing each explanatory variable in a stepwise procedure and measuring the resulting model's AIC. The disadvantage of the stepwise regression model is noteworthy: because we drop variables one at a time, we may inadvertently miss the 'optimal' model since the order in which variables are dropped matter. Despite this flaw, the method still allows for a systematic removal of variables to determine their importance in the model rather than just looking at the individual p-values of each variable to determine its statistical significance to the model.

Using the logistic regression model, we can make predictions about how likely an accident or causality is to be severe, given certain attributes linked to that casualty/vehicle/accident. In predicting the severity status of new observations, there are four possible outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). As the names suggest, a TP (TN) is the outcome where we predict an accident/casualty to be severe (non-severe) when it is severe (non-severe). A FP occurs when we predict an accident/casualty to be severe when it was not severe, while a FN occurs when we predict an accident/casualty to be non-severe when it was severe. Throughout much of our analysis, the TP, TN, FP, FN rate will be a measure of the appropriateness of our

[4]AIC = -(2/n)*(Maximised log likelihood) + (2/n)*(No. of estimated parameters)

model, but we will derive a Cost-Benefit (CB) matrix in *Section 5: Performance Measurement* to make comparison across measures. The formulas for the TP, TN, FP, FN are shown below:

$$True\ Positive\ Rate\ (TPR) = TP/(TP + FP) \qquad False\ Positive\ Rate\ (FPR) = FP/(FP + TN)$$
$$True\ Negative\ Rate\ (TNR) = TN/(TN + FN) \qquad False\ Negative\ Rate\ (FNR) = FN/(FN + TP)$$

**Step 5a: Classification Tree**

Classification is a supervised learning method that uses the training data to identify the characteristics of different classes and predict the class of a new observation based on the observation's characteristics (Yao, 2018). Using recursive binary splitting, each feature is partitioned to the point where there are multiples classes or groups, each with a label of either "Severe" or "Non-Severe". This is a simple, yet effective method to categorise our data into classes and visualise the relationship between variables in determining an accident/casualty's severity.

**Step 5b: K-Fold Cross Validation**

To improve the results of our classification tree, we performed a 10-fold Cross Validation (CV), repeated three times to train our model whilst varying the complexity parameters[5]. The k-fold CV involved partitioning our training set into 10 smaller sets, training our data on nine sets and testing it on the last set, and repeating this for all ten sets so that every set was trained with nine times and tested on once. Finally, this procedure is repeated three times by randomly partitioning our dataset for each 10-fold CV and the accuracy rate of our resultant model tested. This process is then repeated 15 times for varying levels of the complexity parameters to determine the optimal complexity parameter for our model. The benefit of the k-fold technique was that it produced a more specific classification tree which allowed us to more finely classify each observation. However, the obvious drawback is that it can lead to overfitting if not properly controlled.

**Step 6: Random Forest**

Taking things, a step further, we used the random forest technique to reduce variance of our model. This method is like but superior to bagging, as it involves randomly selecting p number of variables (out of all variables in our model) to consider at each split. In our case, we selected p to be four so that at each split, four variables are randomly considered for the split. The main advantage of this over bagging is that it further increases the randomness between classifiers by decorrelating the classifiers, thereby reducing the variance of our model.

A key visualisation of the random forest is the Variance Importance Plot which ranks variables according to their importance, as measured by the mean decrease in accuracy and mean decrease in Gini. The mean decrease in accuracy can be interpreted as the decrease in accuracy due to the exclusion of a single variable. This is done through the random assignment of said variable to all observations (whilst preserving the variable's distribution)

[5] Defined as the cost of adding an additional parameter to the model

and measuring the decrease in accuracy of the resulting tree. Generally, the bigger the decrease in accuracy, the more important that variable is for the classification of the data. On the other hand, the Gini index measures the homogeneity of the resulting nodes after a variable is used to split that node (i.e. the purer the resulting nodes after the split, the lower the Gini index). At the end, the changes in Gini are summed for each variable and normalised at the end for comparison (Metagenomics Statistics, 2011).

**Step 7: K-Means Clustering**

We concluded the analysis portion of our report with some unsupervised learning techniques such as K-means clustering, which works by partitioning the whole dataset into K distinct, non-overlapping clusters. The algorithm works by first randomly assigning all objects into one of K clusters, calculating the centroid for each cluster, and assigning each object to the cluster whose centroid is closest using a distance measure like Euclidean distance. This iterates until the cluster assignments stop changing and K distinct clusters are formed.

To determine the optimal number of clusters for our data, we used three methods (UC Business Analytics R Programming Guide, 2018).: the Elbow method, the Silhouette method, and the Gap Statistic method. The Elbow method works by measuring and the compactness of the clustering and minimising the within sum of squares; the Silhouette[6] method works by measuring the quality of a clustering (i.e. how well each object lies within its cluster); the Gap Statistic compares the total intracluster variation for different values of k with their expected values under null reference distribution of the data (i.e. a distribution with no obvious clustering)

After obtaining our K clusters, we used the Euclidean distance[7] measure to find the point closest to each cluster's centroid to be our cluster representative (intra-group differences). To identify inter-group differences, we created a new column in our dataset to indicate whether each observation was in our cluster of interest and used this cluster identity variable as our dependent variable in the formation of a classification tree. This produced a classification tree which helped us identify common characteristics between and within clusters.

Step 8: Performance Measurement

Finally, we finished our report by comparing the performance of our different models through calculation of the expected benefit of each model and the Area Under Curve (AUC) as suggested by the ROC curve.

[6] The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters (Rousseeuw, 1987)

[7] The Euclidean distance takes the straight-line distance between two points in a Euclidean space

# 4. Main Results

## I. Preliminary Analysis

### Accidents

The **Accidents** dataset is mainly concerned with characteristics surrounding the accident. The graphs below show selected variables worth exploring in this section.

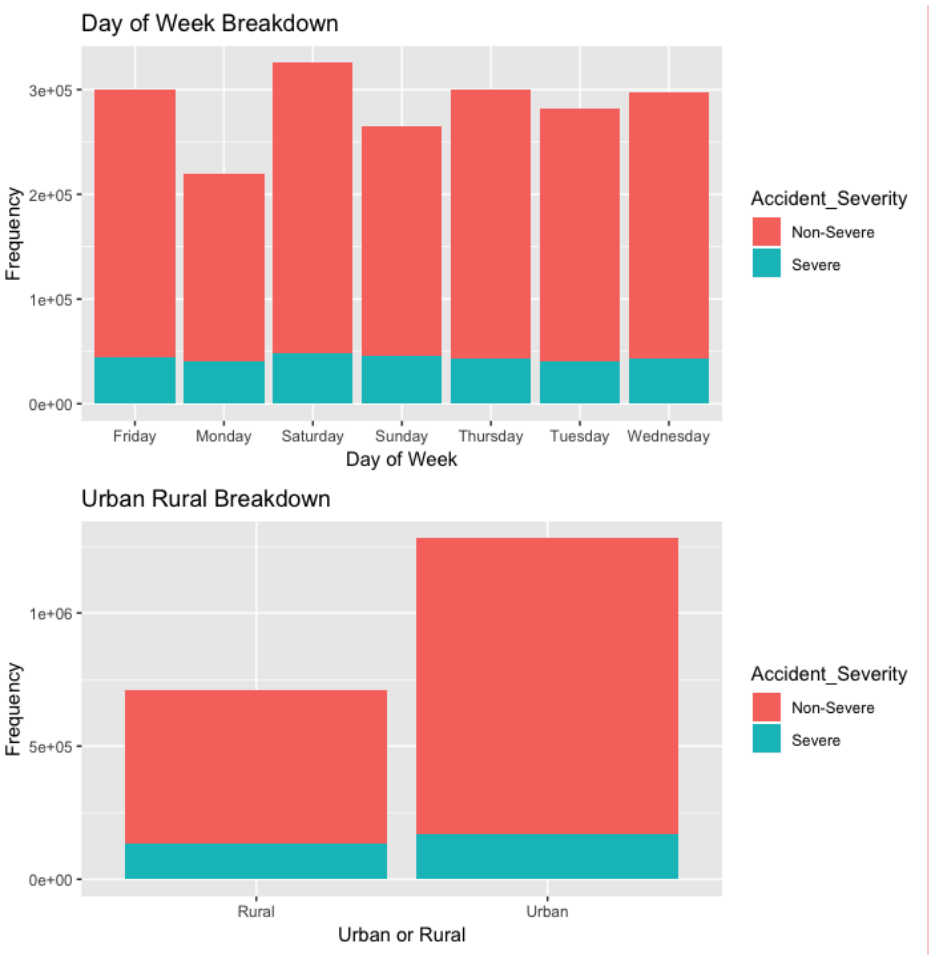



*Figure A.2 Graph describing Light Conditions and Weather Conditions*

Figure A.1 shows that accidents appear across the week seem to occur at relatively equal frequencies and equal severe and non-severe splits, with Monday being the most 'dangerous' day with the highest percentage (18.6%) of accidents being severe versus ~14% over the entire population. The figure also shows how about 90% of accidents occur on Single Carriageways - a two-way road with no physical separation.

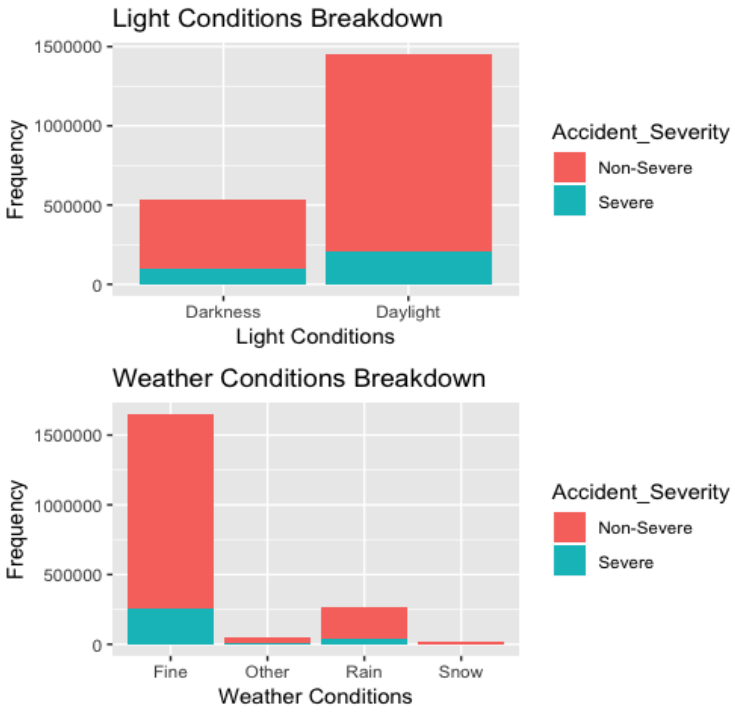



*Figure A.2 Graph describing Light Conditions and Weather Conditions*

Figure A.2 explores the variables of Light Conditions and Weather conditions in greater detail. Majority of accidents occur in conditions deemed to be ideal conditions (Daylight and Fine weather) as expected.

## Vehicles

The **Vehicles** dataset is mainly concerned with characteristics surrounding the vehicles involved in an accident. The graphs below show selected variables worth exploring in this section.

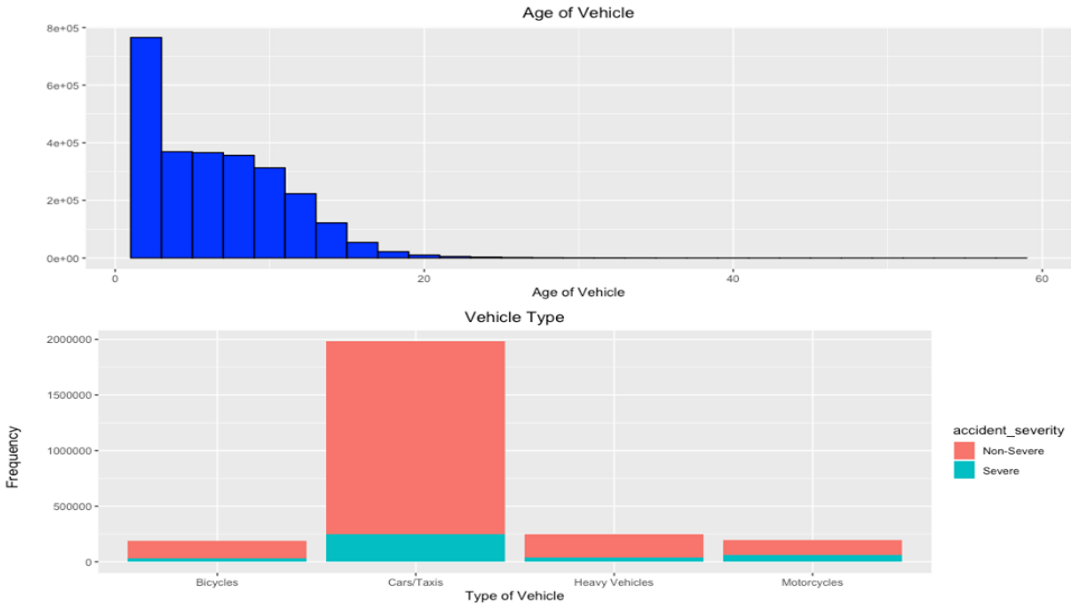



*Figure B.1: Graphs describing the frequencies of vehicle age and type*

From figure B.1, we can see that most of the vehicles involved in accidents are relatively young - most are in fact under 10 years old. This is not surprising given that older cars tend to be less commonly driven than younger cars. The main vehicle type involved is predominantly Cars/Taxis, with similar numbers across the other types of vehicles. Motorcycles suffer a higher incidence of severe accidents than the Bicycles and Heavy Vehicles.

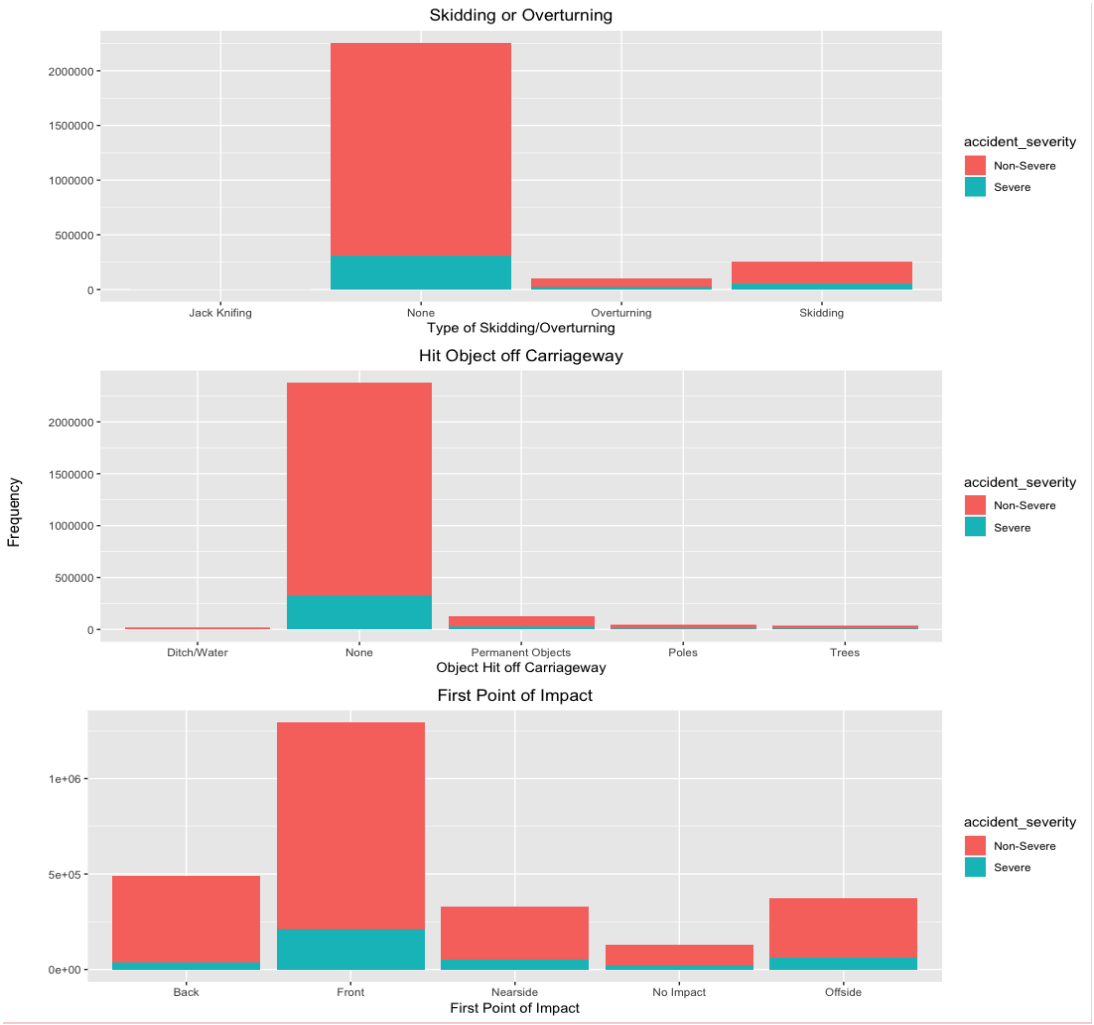



*Figure B.2: Graphs describing the vehicle's characteristics during and after the accident*

Figure B.2 shows further information on the circumstances surrounding the accident. Notably, we see that most vehicles did not skid, overturn, or jackknife. However, amongst those which did, overturned vehicles appear to more likely be severe accidents than skidding vehicles. For the graph showing first point of impact, we see that majority of vehicles suffer a front point of impact compared to other types of impact. This may imply that the vehicle collided into an object ahead of it. Additionally, if we compare the number of severe accidents across all types of impacts, impacts to the back of the car tend to be proportionally safer (smaller % of severe accidents).

*Note: No impact means that the vehicle stopped suddenly to avoid impact with another object /pedestrian/vehicle.*

## Casualties

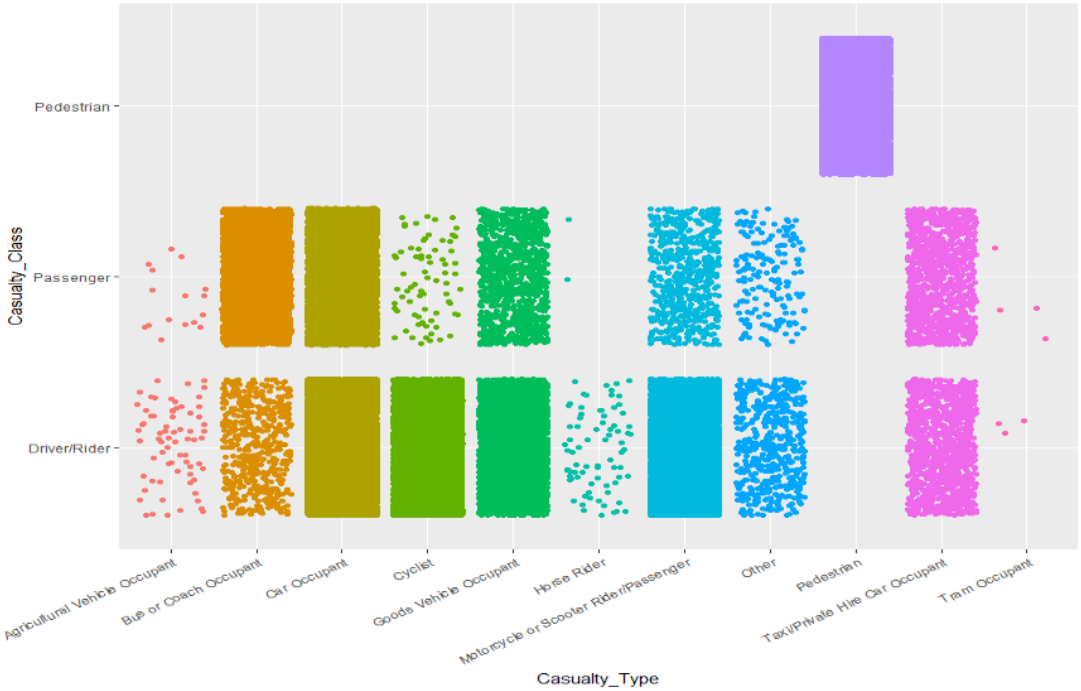



*Figure C.1: Plot showing the casualty class of various casualty types*

The figure above shows that while drivers/riders make up the majority of casualties among most casualty types, passengers make up the majority of casualties amongst bus or coach occupants. This is not surprising given that there tend to be many more passengers in buses/coaches, compared to one driver.

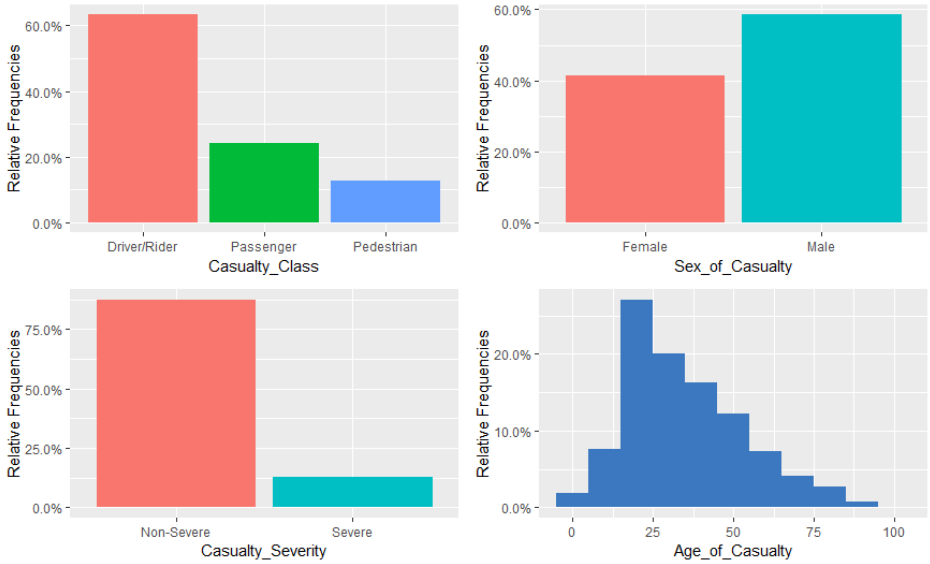



*Figure C.2: Graph showing various attributes of the casualty*

From Figure C.2 above, we observe that most casualties are drivers (over 60%). Furthermore, the majority sex (almost 60%) seems to be Male. Most casualties fall between the ages of 16-50. The number of casualties tend to jump at age 16, possibly because that is the minimum driving age in the UK. These descriptions, however, do not

give us any relationship between *Casualty_Severity* and the rest of the variables. The total dataset comprises ~12.8% of casualties who were severely injured.

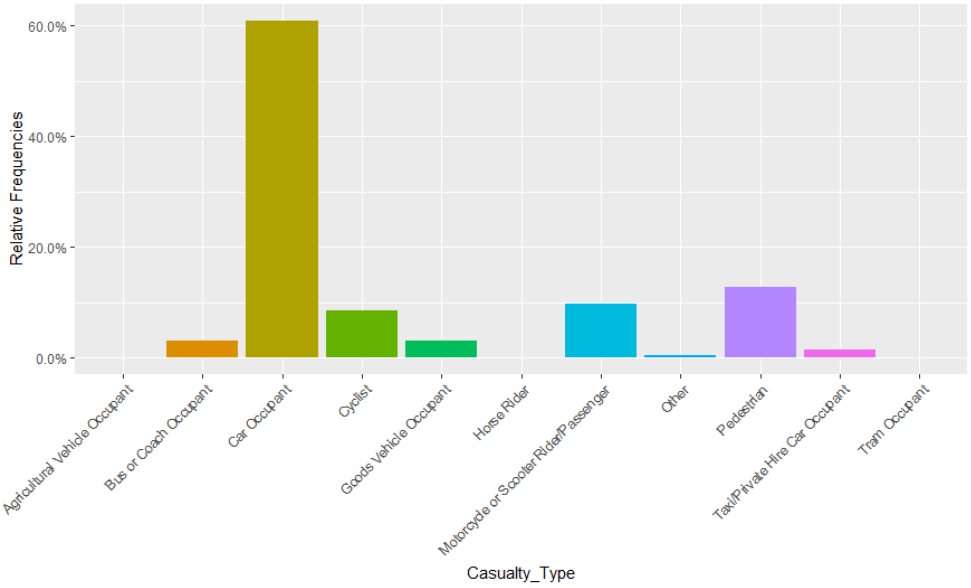



*Figure C.3: Graph showing relative frequencies of casualty types*

Figure C.3 above also shows that *Car Occupants* (both drivers and passengers) comprise the majority (just over 60%) of our dataset.

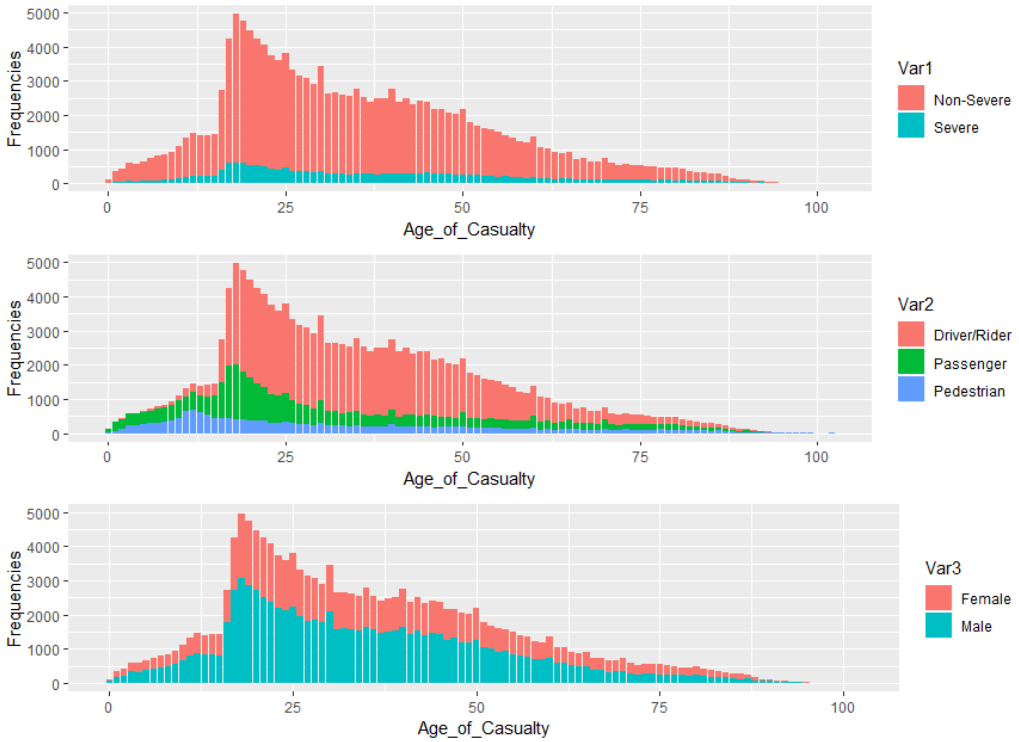



*Figure C.4: Graphs showing the attributes of casualties at every age*

Figure C.4 shows the relative frequencies of the *Age of Casualty* split according to *Casualty Severity*, *Casualty Class* and *Sex of Casualty Respectively*. They demonstrate that the relationships derived from figure C.2 above hold across all ages.

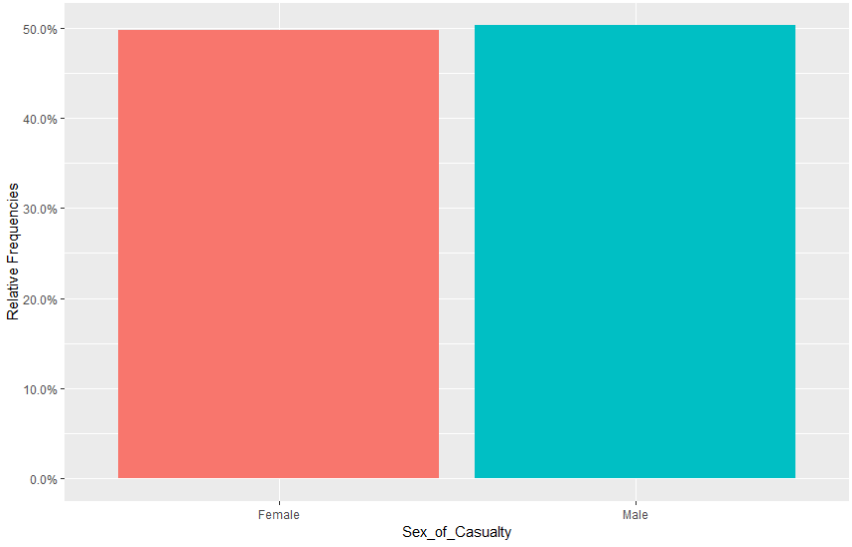



*Figure C.5: Graph showing the relative frequency of the casualty's sex*

An interesting observation is that, while there seems to be about 60% male and 40% female in the full dataset (as seen in Figure C.2), we observe roughly equal proportion of males and females when we take a subset of only

casualties who were car occupants and taxi/private hire occupants (henceforth referred to as Car casualties). Next, we move on to explore *Casualty Severity* against key variables:

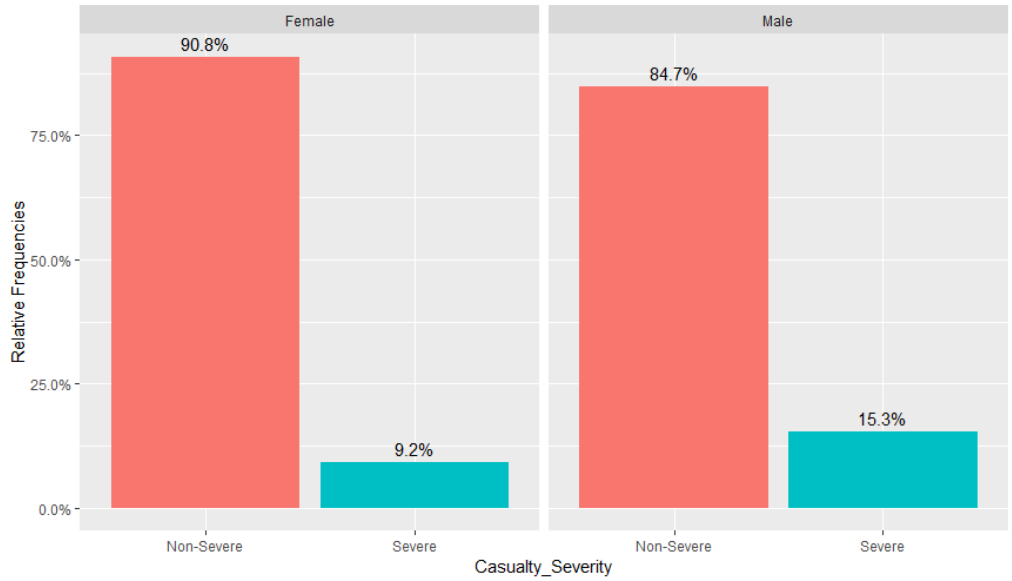



*Figure C.6: Graph showing relative frequencies of casualty severity for males and females*

Examining *Casualty Severity* by *Sex of Casualty*, we discover that males comprise a larger proportion of severe casualties than females - 15.3% and 9.2% respectively. This disparity can also be observed when we only focus on the Pedestrian casualties or Car casualties (see *Figure D1(i)* and *Figure D1(ii)* in Appendix D1).

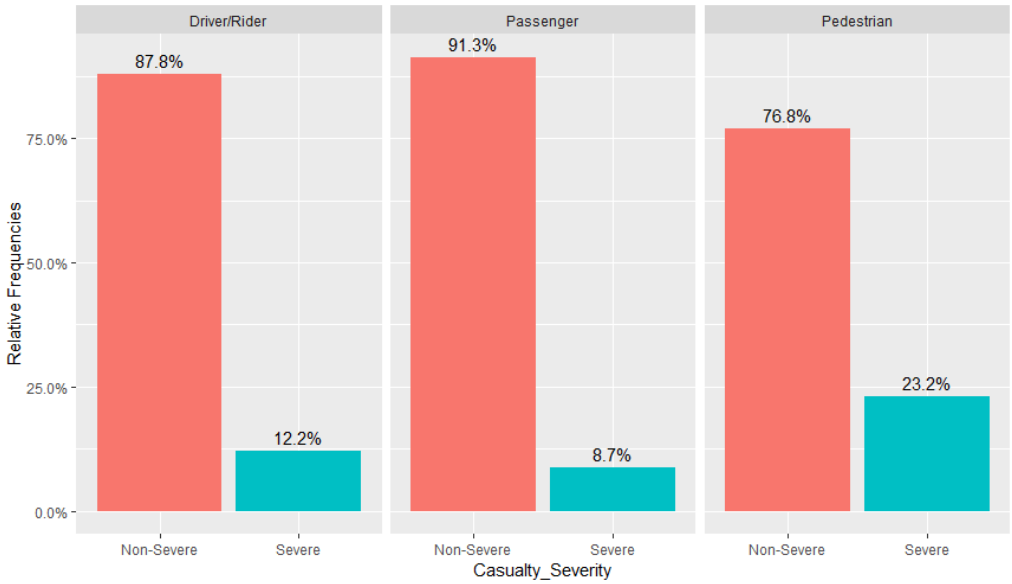



*Figure C.7: Graphs showing relative frequencies of casualty severity for different types of road users (cars)*

Examining *Casualty Severity* by *Casualty Class*, we discover that Pedestrians comprise the largest proportion of severe casualties, roughly double that of Drivers/Riders and triple that of Passengers (23.2%, 12.2% and 8.7%

respectively). When we focus only on Car casualties however, there is little disparity between the proportion of severe casualties between Drivers/Riders and Passengers (see *Appendix D1(iii)*).

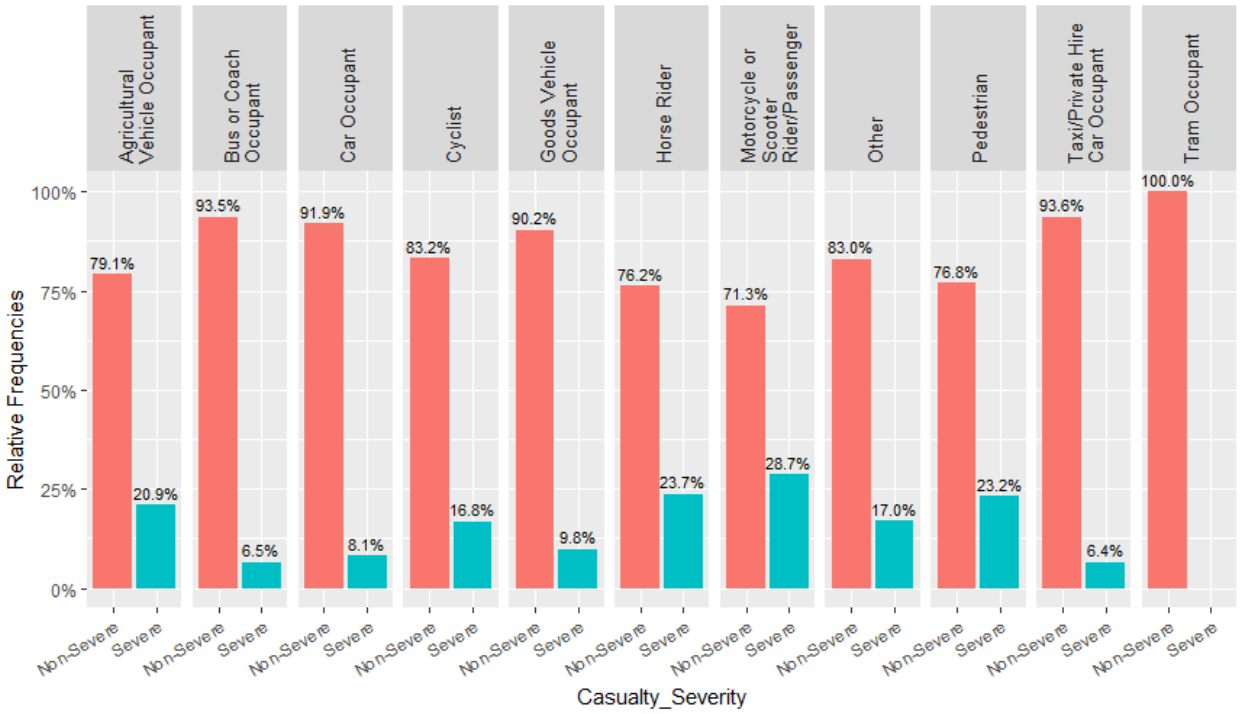



*Figure C.8: Graph showing the relative frequency of casualty severity of different types of road users (all)*

Examining *Casualty Severity* by *Casualty Type*, we discover that Motorcycle or Scooter Riders/Passengers, about 5.5% more than Pedestrians and more than triple that of Car Occupants (28.7%, 23.2% and 8.1% respectively).

## II. Logistic Regression

# Accidents

We begin by running accident severity against all 12 other explanatory variables and obtained the following:

```
> summary(glm1.Accidents)

Call:
glm(formula = Accident_Severity ~ ., family = binomial, data = Atrain)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.49357  -0.78023  -0.65960  -0.09879   2.78698

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.9895810  0.0620598 -15.946  < 2e-16 ***
Number_of_Vehicles            -0.3957502  0.0117260 -33.750  < 2e-16 ***
Number_of_Casualties           0.2469023  0.0091113  27.098  < 2e-16 ***
Day_of_WeekMonday              0.1479539  0.0288360   5.131 2.88e-07 ***
Day_of_WeekSaturday           -0.0204170  0.0268399  -0.761 0.446840
Day_of_WeekSunday              0.0921451  0.0275986   3.339 0.000842 ***
Day_of_WeekThursday           -0.0575306  0.0277522  -2.073 0.038171 *
Day_of_WeekTuesday            -0.0344901  0.0279694  -1.233 0.217524
Day_of_WeekWednesday          -0.0549246  0.0276945  -1.983 0.047341 *
Road_TypeOne Way Street        0.1913787  0.0578635   3.307 0.000942 ***
Road_TypeRoundabout            0.2492746  0.0600383   4.152 3.30e-05 ***
Road_TypeSingle Carriageway    0.3119849  0.0242263  12.878  < 2e-16 ***
Road_TypeSlip Road            -0.4046414  0.0892202  -4.535 5.75e-06 ***
Speed_limit                    0.0086416  0.0007999  10.803  < 2e-16 ***
Junction_DetailNone            0.1542014  0.0167580   9.202  < 2e-16 ***
Junction_DetailRoundabout     -0.4229628  0.0493857  -8.564  < 2e-16 ***
Light_ConditionsDaylight      -0.2607383  0.0167980 -15.522  < 2e-16 ***
Weather_ConditionsOther       -0.2975041  0.0506625  -5.872 4.30e-09 ***
Weather_ConditionsRain        -0.2707471  0.0287418  -9.420  < 2e-16 ***
Weather_ConditionsSnow        -0.3062484  0.1049507  -2.918 0.003523 **
Road_Surface_ConditionsSnow   -0.5586113  0.0573444  -9.741  < 2e-16 ***
Road_Surface_ConditionsWet    -0.0399327  0.0216587  -1.844 0.065223 .
Special_Conditions_at_SitePresent -0.1527861  0.0499134  -3.061 0.002206 **
Carriageway_HazardsPresent    -0.0599998  0.0556103  -1.079 0.280618
Urban_or_Rural_AreaUrban      -0.2184308  0.0219184  -9.966  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figure A.3 R output showing the significance of different explanatory variables using all variables*

From figure A.3, we can see that there is only one explanatory variable which is statistically insignificant - *Carriageway Hazards*. It is interesting to note that all the other 11 variables are statistically significant. Using this model (with a 0.5 cut-off probability[8] in order for accident severity to be classed as severe) to predict the accident severity in the testing set produced an AR of 84.5%, FPR of 0.64%, and FNR of 97.5%. The extremely high FNR is a cause for concern. This is extremely worrying as a FN means being unable to correct predict when an accident is a severe one or not. We try to mitigate this by improving upon the model and varying our cut-off probability.

[8] We will vary the cut-off probability in later sections to determine the optimal cut-off probability

We employed a stepwise regression to select variables through using the AIC criterion to improve on the model. Doing this, only one variable was removed: *Carriageway Hazard.* Running a new logistics regression with without the aforementioned variable, we obtain the following results:

```
> summary(glm2.Accidents)

Call:
glm(formula = Accident_Severity ~ . - Carriageway_Hazards, family = binomial,
    data = Atrain)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.49280  -0.77996  -0.65944  -0.09887   2.78670

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -0.9909601  0.0620464 -15.971  < 2e-16 ***
Number_of_Vehicles             -0.3952983  0.0117162 -33.739  < 2e-16 ***
Number_of_Casualties            0.2469147  0.0091113  27.100  < 2e-16 ***
Day_of_WeekMonday               0.1477706  0.0288353   5.125 2.98e-07 ***
Day_of_WeekSaturday            -0.0204376  0.0268398  -0.761 0.446378
Day_of_WeekSunday               0.0921110  0.0275983   3.338 0.000845 ***
Day_of_WeekThursday            -0.0574969  0.0277521  -2.072 0.038283 *
Day_of_WeekTuesday             -0.0345307  0.0279696  -1.235 0.216986
Day_of_WeekWednesday           -0.0549230  0.0276943  -1.983 0.047346 *
Road_TypeOne Way Street         0.1911474  0.0578624   3.303 0.000955 ***
Road_TypeRoundabout             0.2495147  0.0600369   4.156 3.24e-05 ***
Road_TypeSingle Carriageway     0.3121525  0.0242259  12.885  < 2e-16 ***
Road_TypeSlip Road             -0.4038541  0.0892163  -4.527 5.99e-06 ***
Speed_limit                     0.0086179  0.0007996  10.777  < 2e-16 ***
Junction_DetailNone             0.1536945  0.0167517   9.175  < 2e-16 ***
Junction_DetailRoundabout      -0.4226031  0.0493833  -8.558  < 2e-16 ***
Light_ConditionsDaylight       -0.2602332  0.0167916 -15.498  < 2e-16 ***
Weather_ConditionsOther        -0.2974437  0.0506614  -5.871 4.33e-09 ***
Weather_ConditionsRain         -0.2707806  0.0287417  -9.421  < 2e-16 ***
Weather_ConditionsSnow         -0.3060836  0.1049471  -2.917 0.003539 **
Road_Surface_ConditionsSnow    -0.5580969  0.0573381  -9.733  < 2e-16 ***
Road_Surface_ConditionsWet     -0.0399610  0.0216584  -1.845 0.065030 .
Special_Conditions_at_SitePresent -0.1550810  0.0498677  -3.110 0.001872 **
Urban_or_Rural_AreaUrban       -0.2180639  0.0219161  -9.950  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Figure A.4 R output of glm after removing insignificant variables*

The removal of only one variable meant that the result of this skimmed logistic regression was exactly the same as our initial model. We obtained AR, FPR, and FNR of 84.5%, 0.64%, and 97.5% respectively. These results do seem to preliminarily suggest that a logistic regression may not be the best model to describe accidents.

Plotting the probability of a severe accident against our exploratory variables did yield some interesting results worth looking at:

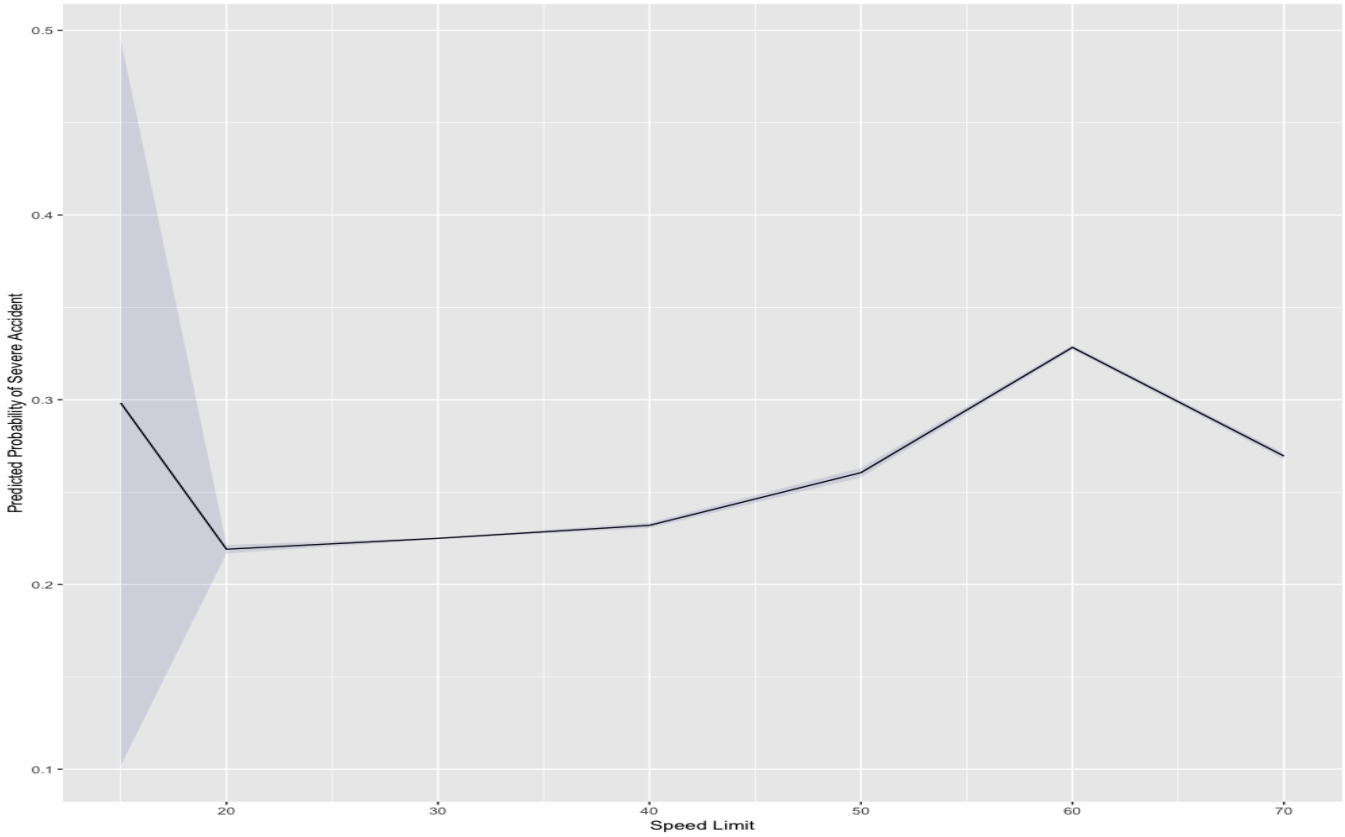



*Figure A.5: Predicted probability of Severe Accident against Speed Limit*

It is interesting that roads with speed limits of <20mph are more likely to produce severe accidents than roads with a speed limit of 70mph. The speed limit which was most likely to produce a severe accident was 60mph.

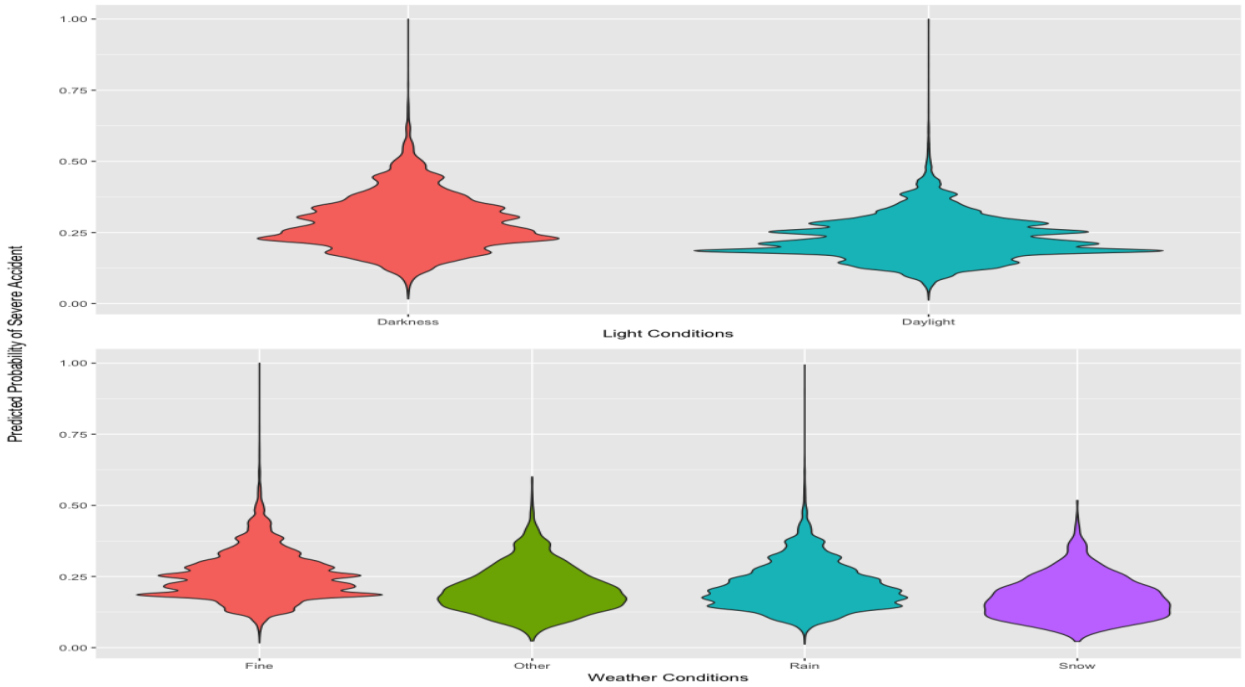



*Figure A.6: Predicted probability of Severe Accident against Explanatory Vehicles*

Notably, the predicted probability of a Severe accidents in 'bad' conditions are largely similar and in line with conditions that are perceived to be ideal (Darkness vs Daylight and Fine vs Rain, Snow and Other).

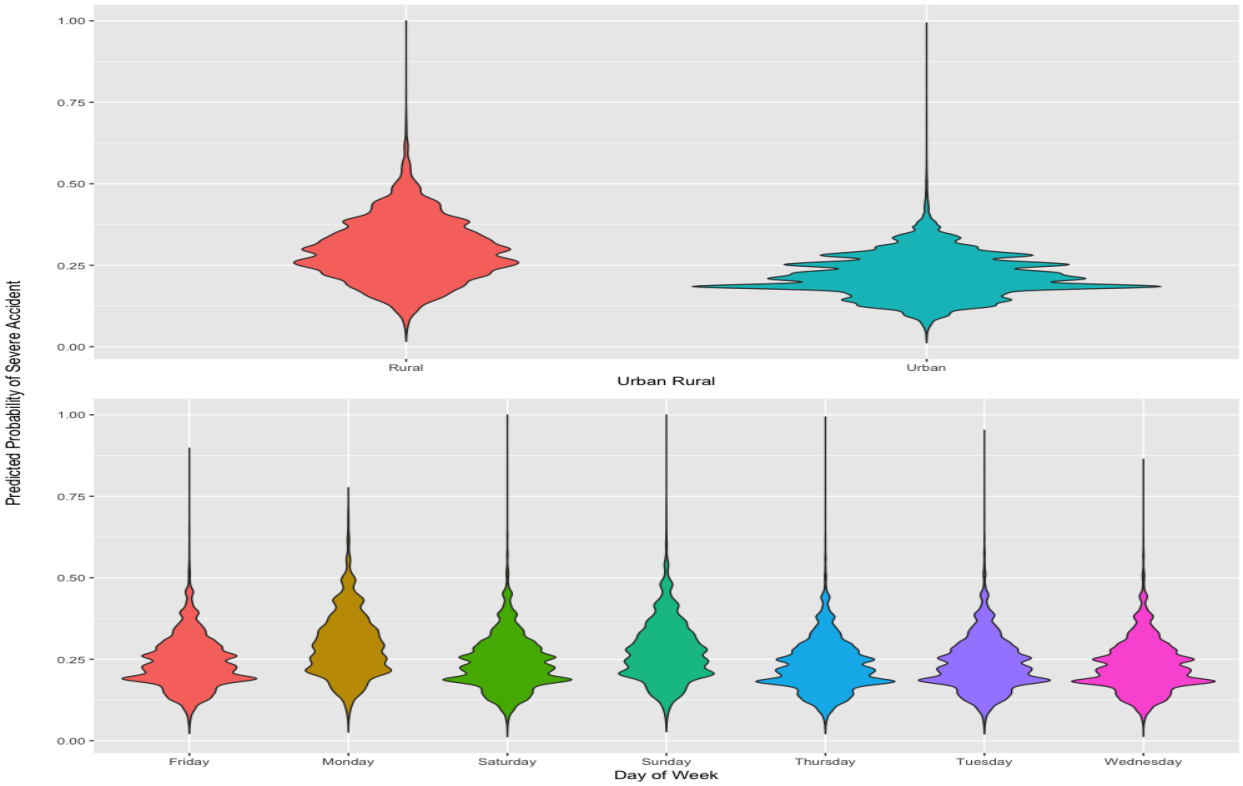



*Figure A.6: predicted probabilities of a severe accident against selected explanatory variables*

Monday and Sunday seem to have a higher probability of a severe accident. A potential explanation of this could be due to the fact that people are looking to go to work or return home. Additionally, accidents in rural areas tend to be slightly more likely to be severe, this could be due to rural areas having higher speed limits on average as opposed to urban areas (e.g. cross-country highways)

## Vehicles

We begin by running accident severity against all 10 other explanatory variables and obtained some interesting results:

```
glm(formula = accident_severity ~ ., family = binomial, data = Vtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6991  -0.7732  -0.6610  -0.1086   2.2656

Coefficients:
                                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                                 -1.8005445  0.3482897  -5.170 2.35e-07 ***
vehicle_typeCars/Taxis                      -0.4959651  0.0290566 -17.069  < 2e-16 ***
vehicle_typeHeavy Vehicles                  -0.2438581  0.0351585  -6.936 4.03e-12 ***
vehicle_typeMotorcycles                      0.4986309  0.0352782  14.134  < 2e-16 ***
skidding_and_overturningNone                -0.4003931  0.2734176  -1.464  0.14308
skidding_and_overturningOverturning         -0.1575812  0.2750994  -0.573  0.56677
skidding_and_overturningSkidding            -0.3002207  0.2740389  -1.096  0.27328
hit_object_in_carriagewayNone               -0.1217673  0.1979708  -0.615  0.53850
hit_object_in_carriagewayOther Vehicles     -0.1750603  0.2067682  -0.847  0.39719
hit_object_in_carriagewayPhysical Infrastructure -0.0833390  0.2027065  -0.411  0.68098
vehicle_leaving_carriagewayYes               0.6467195  0.0326431  19.812  < 2e-16 ***
hit_object_off_carriagewayNone               0.2358639  0.0743724   3.171  0.00152 **
hit_object_off_carriagewayPermanent Objects  0.0779746  0.0740722   1.053  0.29249
hit_object_off_carriagewayPoles              0.0799909  0.0830875   0.963  0.33568
hit_object_off_carriagewayTrees              0.6727475  0.0842857   7.982 1.44e-15 ***
first_point_of_impactFront                   0.7382414  0.0249122  29.634  < 2e-16 ***
first_point_of_impactNearside                0.6670126  0.0306488  21.763  < 2e-16 ***
first_point_of_impactNo Impact               0.7948001  0.0389842  20.388  < 2e-16 ***
first_point_of_impactOffside                 0.7360824  0.0294582  24.987  < 2e-16 ***
was_vehicle_left_hand_driveYes               0.2434532  0.2229581   1.092  0.27487
sex_of_driverMale                            0.2300736  0.0176816  13.012  < 2e-16 ***
age_of_driver                                0.0094420  0.0004724  19.988  < 2e-16 ***
age_of_vehicle                               0.0049104  0.0016638   2.951  0.00316 **
```

*Figure B.3: R output showing the significance of different explanatory variables using all variables*

From figure B.3, we can see that there are only seven explanatory variables which are statistically significant. Using this model to predict the accident severity in the testing set produced an AR of 85.3%, a FPR of 1.1%, and a FNR of 94.3%! To improve this model, we employed a stepwise regression to select variables through using the AIC criterion. Doing this, only eight variables were kept: *Age of Vehicle, Hit Object off Carriageway, Sex of Driver, Age of Driver, Vehicle Leaving Carriageway, First Point of Impact,* and *Vehicle Type.* This result is largely similar to the significance tests we did in the generalised logistic regression above. Running a new logistics regression with the variables mentioned above, we obtain the following results:

```
glm(formula = accident_severity ~ . - hit_object_in_carriageway -
    was_vehicle_left_hand_drive - skidding_and_overturning, family = binomial,
    data = Vtrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7450  -0.7740  -0.6633  -0.1195   2.2836

Coefficients:
                                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -2.2641438  0.0849841 -26.642  < 2e-16 ***
vehicle_typeCars/Taxis                    -0.4927890  0.0290144 -16.984  < 2e-16 ***
vehicle_typeHeavy Vehicles                -0.2394526  0.0350958  -6.823 8.93e-12 ***
vehicle_typeMotorcycles                    0.5177949  0.0350651  14.767  < 2e-16 ***
vehicle_leaving_carriagewayYes             0.6946597  0.0319038  21.774  < 2e-16 ***
hit_object_off_carriagewayNone             0.1844335  0.0740489   2.491  0.01275 *
hit_object_off_carriagewayPermanent Objects 0.0566403  0.0739239   0.766  0.44356
hit_object_off_carriagewayPoles            0.0417553  0.0825474   0.506  0.61297
hit_object_off_carriagewayTrees            0.6526146  0.0841749   7.753 8.97e-15 ***
first_point_of_impactFront                 0.7488297  0.0248085  30.184  < 2e-16 ***
first_point_of_impactNearside              0.6793468  0.0305797  22.216  < 2e-16 ***
first_point_of_impactNo Impact             0.8073474  0.0389063  20.751  < 2e-16 ***
first_point_of_impactOffside               0.7442534  0.0294263  25.292  < 2e-16 ***
sex_of_driverMale                          0.2325369  0.0176733  13.157  < 2e-16 ***
age_of_driver                              0.0091466  0.0004699  19.465  < 2e-16 ***
age_of_vehicle                             0.0052568  0.0016619   3.163  0.00156 **
---
```

*Figure B.4: R output for glm after removing insignificant variables*

Testing our new model on the testing data, we obtain AR, FPR, and FNR of 85.1%, 1.0%, and 94.9% respectively, which is slightly worse than our initial full model. These results do seem to preliminarily suggest that a logistic regression may not be the best model to describe accidents. However, plotting predicted probabilities of a severe accident against our explanatory variables provided some interesting graphs (see below).

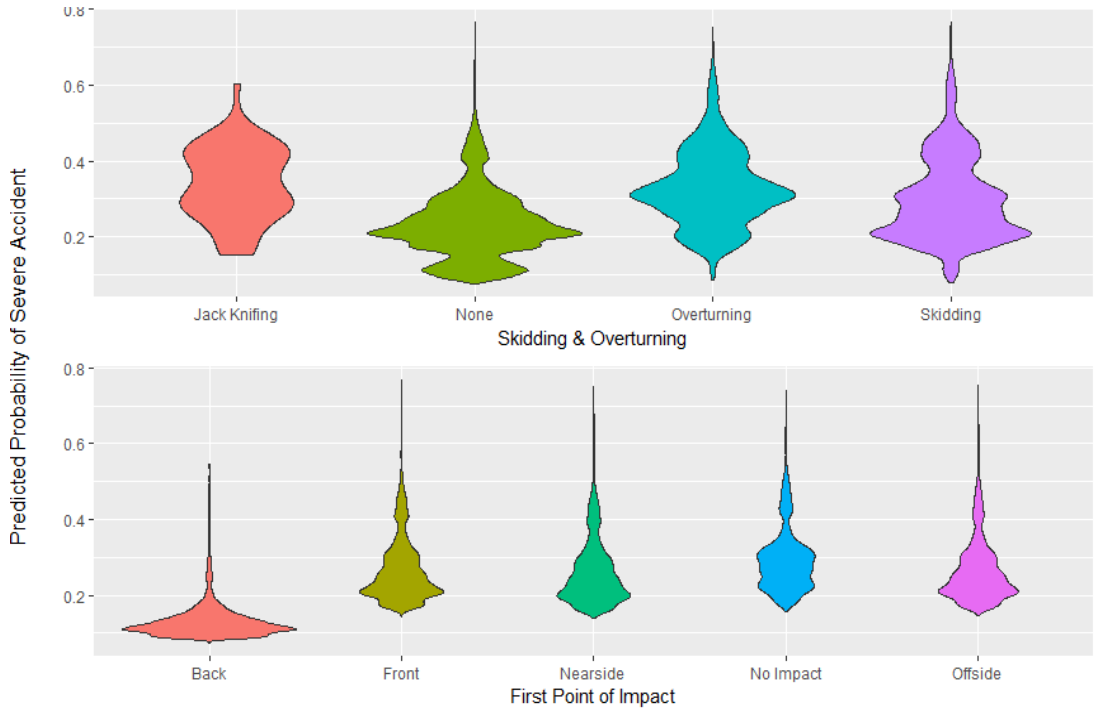



*Figure B.5: Predicted probabilities of a severe accident against selected explanatory variables*

Looking at figure B.5, a vehicle which did not skid, jackknife, or overturned appears to produce the lowest predicted probability of a severe accident (~20% v ~30%). More interestingly, the second graph shows that vehicles which had been hit at the back have the lowest probability of being damaged (~10% v 20+%). The large width of the plot for back further implies that most of these rear-ended accidents have relatively low probabilities of being severe compared to impacts at other points of the car. Another notable point of observation is that vehicles with no impact suffered were just as likely to get into severe accidents as vehicles which got impacted at the front, nearside, and offside. One possible explanation is that drivers swerved or braked suddenly to avoid collision with another object, leading to injuries caused by the car grinding to a stop quickly.

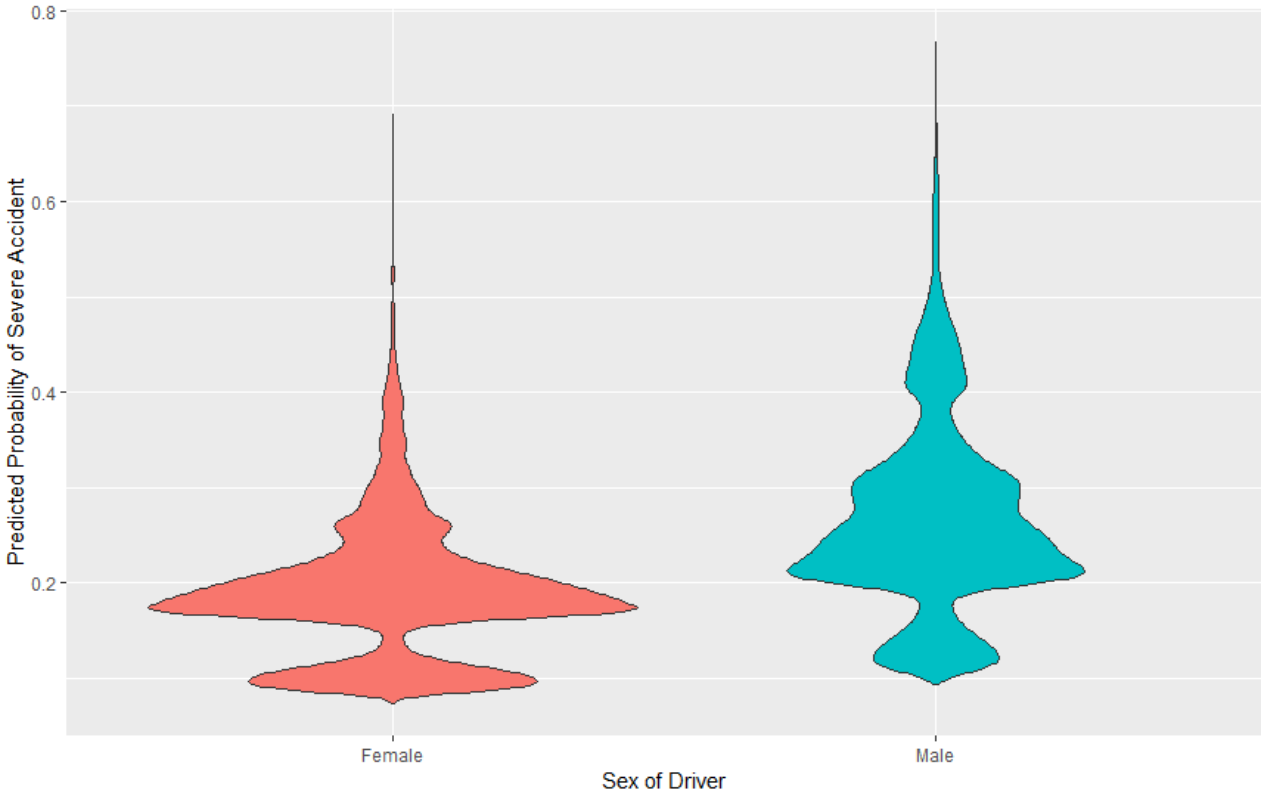



*Figure B.6: Graph showing the predicted probabilities of a severe accident V sex of driver*


*Figure B.7: Graphs showing the age of driver and age of vehicle against the predicted probabilities of a severe accident; area shaded in blue shows the 95% confidence interval*

The two graphs above (figures B.6 and B.7) show several fascinating insights: male drivers, older vehicles, and older drivers are more likely to get into a severe accident. Inspecting the latter observation, the probability of a severe accident decreases as a driver moves between the ages of 16-25 but gradually increases thereafter. In short, this result is saying that young drivers (who currently pay more for vehicular insurance) are as likely to get into an accident as a 50+ year-old!

# Casualties

We begin by running *Casualty Severity* against all 8 other explanatory variables and obtained some interesting results:

```
glm(formula = Casualty_Severity ~ ., family = binomial, data = cDtrain1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95978  -1.01534  -0.00282   1.06527   1.95674

Coefficients: (3 not defined because of singularities)
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    0.1734410  0.3381213   0.513 0.607983
Casualty_ClassPassenger                        0.3322528  0.0650710   5.106 3.29e-07 ***
Casualty_ClassPedestrian                       0.4370245  0.2667250   1.638 0.101321
Sex_of_CasualtyMale                            0.3929633  0.0148884  26.394  < 2e-16 ***
Age_of_Casualty                                0.0123820  0.0003599  34.400  < 2e-16 ***
Pedestrian_LocationCrossing Elsewhere         -0.0148030  0.0621090  -0.238 0.811618
Pedestrian_LocationNot Pedestrian                     NA         NA      NA       NA
Pedestrian_LocationOn/Near Pedestrian Crossing -0.0045182  0.0507276  -0.089 0.929028
Pedestrian_LocationOther                       -0.2259408  0.0746387  -3.027 0.002469 **
Pedestrian_LocationPavement/Sidewalk           -0.2834704  0.0689871  -4.109 3.97e-05 ***
Pedestrian_MovementNot Pedestrian                     NA         NA      NA       NA
Pedestrian_MovementOther                       -0.1657807  0.0556862  -2.977 0.002910 **
Pedestrian_MovementStationary in Carriageway   -0.5025370  0.0729025  -6.893 5.45e-12 ***
Pedestrian_MovementWalking Along in Carriageway -0.2396620  0.0908122  -2.639 0.008313 **
Car_PassengerNot Car Passenger                 0.1719990  0.0682607   2.520 0.011744 *
Car_PassengerRear Seat Passenger               0.1517411  0.0332808   4.559 5.13e-06 ***
Bus_or_Coach_PassengerNot a Bus or Coach Passenger -0.6219096  0.1754334  -3.545 0.000393 ***
Bus_or_Coach_PassengerSeated Passenger         -1.0257580  0.1502495  -6.827 8.67e-12 ***
Bus_or_Coach_PassengerStanding Passenger       -0.5040253  0.1596732  -3.157 0.001596 **
Casualty_TypeBus or Coach Occupant             -1.4192312  0.2854358  -4.972 6.62e-07 ***
Casualty_TypeCar Occupant                      -0.9678455  0.2657957  -3.641 0.000271 ***
Casualty_TypeCyclist                           -0.1688195  0.2663636  -0.634 0.526216
Casualty_TypeGoods Vehicle Occupant            -0.8937614  0.2681126  -3.334 0.000858 ***
Casualty_TypeHorse Rider                       0.1956570  0.3469361   0.564 0.572784
Casualty_TypeMotorcycle or Scooter Rider/Passenger 0.5186406  0.2661753   1.948 0.051356 .
Casualty_TypeOther                             -0.3046392  0.2792203  -1.091 0.275257
Casualty_TypePedestrian                               NA         NA      NA       NA
Casualty_TypeTaxi/Private Hire Car Occupant    -1.4882780  0.2743301  -5.425 5.79e-08 ***
Casualty_TypeTram Occupant                     -0.2238013  0.9715686  -0.230 0.817819
```

*Figure C.9: R output showing the significance of different explanatory variables using all variables*

From Figure C.9, we can see that there are only four explanatory variables which are statistically significant for all indicators. They are: *Car Passenger, Bus or Coach Passenger, Sex of Casualty and Age of Casualty*. Furthermore, we also observe 'NA' values above, due to the multicollinearity between variables e.g. casualty labelled as 'Pedestrian' under *Casualty Type* is always labelled as 'Not a Bus or Coach Passenger' under *Bus or Coach Passenger.* This leads to perfect multicollinearity, resulting in very inaccurate estimates. Furthermore, not all variables in Figure C.9 are applicable to all casualties. For example, *Car Passenger* is not applicable to pedestrian casualties.

To attempt to improve this model, we employed a stepwise regression to select variables. However, doing this returned same model as the full model that was initially attempted above i.e. the step function did not remove any variable at all. Paying closer attention to the eight variables, we noticed that only four of them (namely *Casualty Class, Age of Casualty, Sex of Casualty* and *Casualty Type)* were applicable to all our casualties. The remaining variables were only applicable to a subset of them.

To determine the variables for our logistic regression, we used the Variable Importance Plot (see Figure C.14 in 4(b)(IV) Random Forests). The variables *Age of Casualty*, *Sex of Casualty* and *Casualty Type* occurred within the top four variables for both 'Mean Decrease Accuracy' and 'Mean Decrease Gini' plots. We tried different interactions of these three variables and settled on the model below, based on the accuracy rate.

```
glm(formula = Casualty_Severity ~ Casualty_Type + Sex_of_Casualty +
    Age_of_Casualty + Casualty_Type:Age_of_Casualty, family = binomial,
    data = cDtrain1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.09106  -1.01143  -0.00074   1.07313   2.04303

Coefficients:
                                                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                    -3.190e-01  6.551e-01  -0.487   0.6263
Casualty_TypeBus or Coach Occupant                             -1.658e+00  6.645e-01  -2.495   0.0126 *
Casualty_TypeCar Occupant                                      -7.298e-01  6.553e-01  -1.114   0.2654
Casualty_TypeCyclist                                           -2.240e-01  6.568e-01  -0.341   0.7330
Casualty_TypeGoods Vehicle Occupant                            -7.180e-01  6.649e-01  -1.080   0.2802
Casualty_TypeHorse Rider                                       -1.059e-01  8.627e-01  -0.123   0.9023
Casualty_TypeMotorcycle or Scooter Rider/Passenger              4.391e-01  6.567e-01   0.669   0.5038
Casualty_TypeOther                                              3.479e-01  6.824e-01   0.510   0.6103
Casualty_TypePedestrian                                         4.066e-01  6.556e-01   0.620   0.5351
Casualty_TypeTaxi/Private Hire Car Occupant                    -1.308e+00  6.785e-01  -1.929   0.0538 .
Casualty_TypeTram Occupant                                     -5.257e+01  2.329e+02  -0.226   0.8214
Sex_of_CasualtyMale                                             3.625e-01  1.462e-02  24.799   <2e-16 ***
Age_of_Casualty                                                 1.526e-02  1.475e-02   1.034   0.3009
Casualty_TypeBus or Coach Occupant:Age_of_Casualty              7.087e-03  1.489e-02   0.476   0.6340
Casualty_TypeCar Occupant:Age_of_Casualty                      -5.845e-03  1.476e-02  -0.396   0.6921
Casualty_TypeCyclist:Age_of_Casualty                            6.946e-04  1.481e-02   0.047   0.9626
Casualty_TypeGoods Vehicle Occupant:Age_of_Casualty            -4.101e-03  1.500e-02  -0.273   0.7845
Casualty_TypeHorse Rider:Age_of_Casualty                        6.689e-03  2.056e-02   0.325   0.7449
Casualty_TypeMotorcycle or Scooter Rider/Passenger:Age_of_Casualty  2.150e-03  1.482e-02   0.145   0.8846
Casualty_TypeOther:Age_of_Casualty                             -1.462e-02  1.525e-02  -0.959   0.3377
Casualty_TypePedestrian:Age_of_Casualty                        -3.670e-03  1.477e-02  -0.248   0.8038
Casualty_TypeTaxi/Private Hire Car Occupant:Age_of_Casualty    -2.280e-03  1.530e-02  -0.149   0.8816
Casualty_TypeTram Occupant:Age_of_Casualty                      1.236e+00  5.350e+00   0.231   0.8173
```

*Figure C.10: R output showing the significance of different explanatory variables using selected variables*

We immediately observe that most of our variables are not statistically significant, likely due to the fact that we may be using too many indicators to explain the variation *Casualty Severity*. This is a large drawback of this model which we further evaluate in the later part of this report. Testing our new model on the testing data, we obtain AR, FPR, and FNR of 70.2%, 30.0%, and 42.6%.

Now, for the two subsets of the casualties dataset (Pedestrian casualties and Car casualties) which we extracted earlier, we tried to fit logistic models through the same procedure (using Variable Importance Plots - see *Figure D5(i) and Figure D5(ii)* in Appendix D5). From *Figure D2(i)* and *Figure D2(ii)* in Appendix D2, we notice that we obtain more statistically significant variables. However, for the logistic model with only Pedestrian casualties, we obtained an AR < 30%. Furthermore, for the logistic model with only Car casualties, we obtain a FNR of about 97%.

These results do seem to preliminarily suggest that a logistic regression may not be the best model to describe accidents. Nonetheless, plotting predicted probabilities of a severe accident against our explanatory variables provided some interesting graphs (see next page).

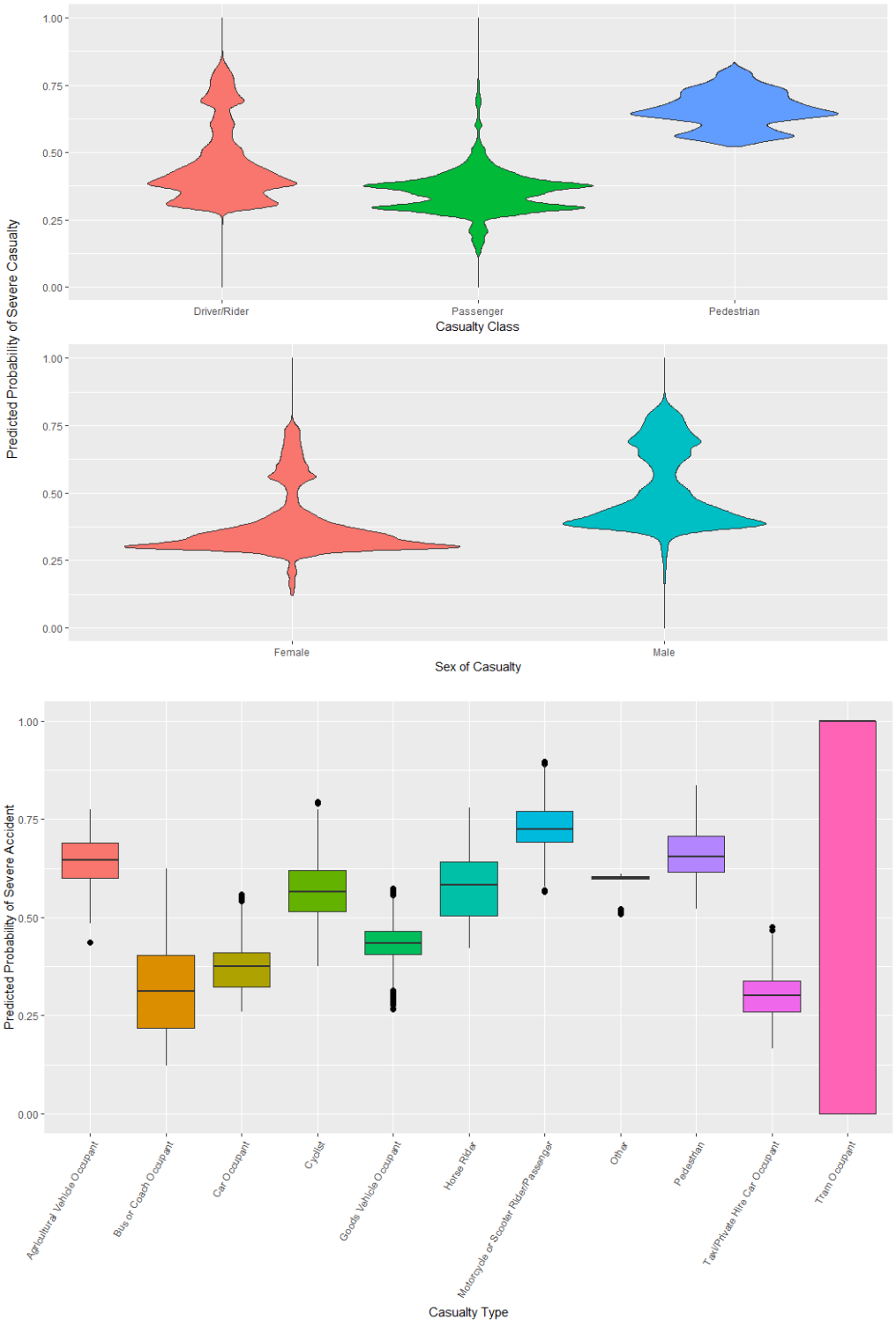



*Figure C.11: Charts showing predicted probabilities of a severe accident against explanatory variables*

Looking at the charts above (which apply to all casualties, without sub-setting), pedestrians tend to have a higher probability of severe injury compared to passengers and drivers (approximately 65%, 30% and 40% respectively). This is not surprising, given than pedestrians are not protected by mechanisms like airbags at the time of accident. Furthermore, males have a higher probability of severe injury compared to females (approximately 50% and 30%

respectively). From the boxplot, we observe that motorcycle or scooter casualties more likely to have a severe injury, with bus/coach occupant, car occupant and taxi/private hire car occupant being the least likely. This may likely be because motorcycles/scooters are less shielded from the elements unlike larger covered vehicles.

When we conducted the analysis for only Pedestrian casualties (Figure D2(iii) in *Appendix D2*), an interesting result was that pedestrians who were crossing the road or walking along the carriageway at the time of accident were generally more likely to have a severe injury than pedestrians who were stationary. Furthermore, when we conducted the analysis for only Car casualties (Figure D2(iv) in *Appendix D2*), we noticed that car occupants were more likely to have a severe injury than taxi/private hire car occupants.

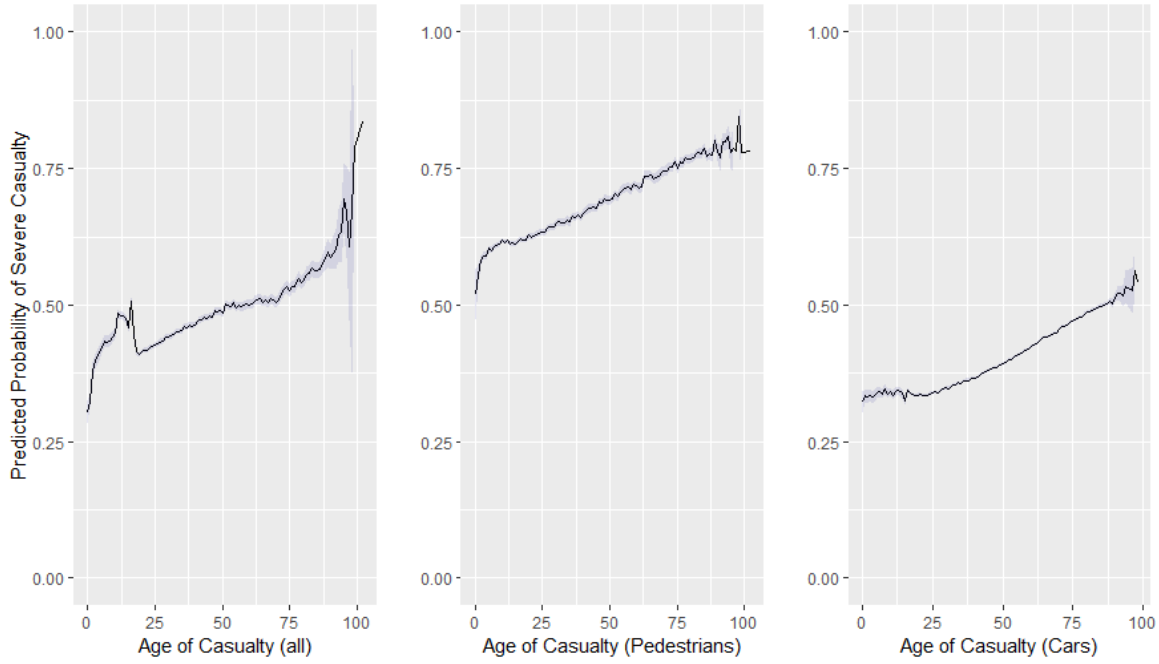



*Figure C.12: Graphs showing the age of casualty (all), age of casualty (pedestrians) and age of casualty (cars)*

The graphs above show several fascinating insights: on the left, we notice that the probability of a severe accident decreases as the casualty moves between the ages of 16-20 but gradually increases thereafter. Generally, we notice a positive relationship between the *Age of Casualty* and predicted probability of severe. However, pedestrians tend to have a higher probability of severe injury compared to car occupants across all age groups.

## III. Classification Tree

# Accidents

The minimum split - the minimum number of observations that must exist in a node for a split to be attempted (Therneau and Beth Atkinson, 2018) - chosen was 2,500. Other larger numbers (5,000, 10,000) were used initially for the minimum split, but this yielded no results as there were not enough observations in each node. Additionally, the complexity parameter was set at 0.001, such that any split that does not decrease the overall lack of fit by a factor of 0.001 was not attempted.

The tree below (see figure A.8) suggests that the most probable characteristics of a severe accident are as follows:

- Two or more vehicles involved
- Rural area
- Less than three casualties
- Takes place on a slip road or one-way street
- Speed limit over 55mph

Based of this tree, 55.6% of the accidents in this group are severe, which is not a very high and convincing number. Overall, this classification tree had an AR, FPR, and FNR of 84.4%, 0.92%, and 96.5% when tested on the testing data.

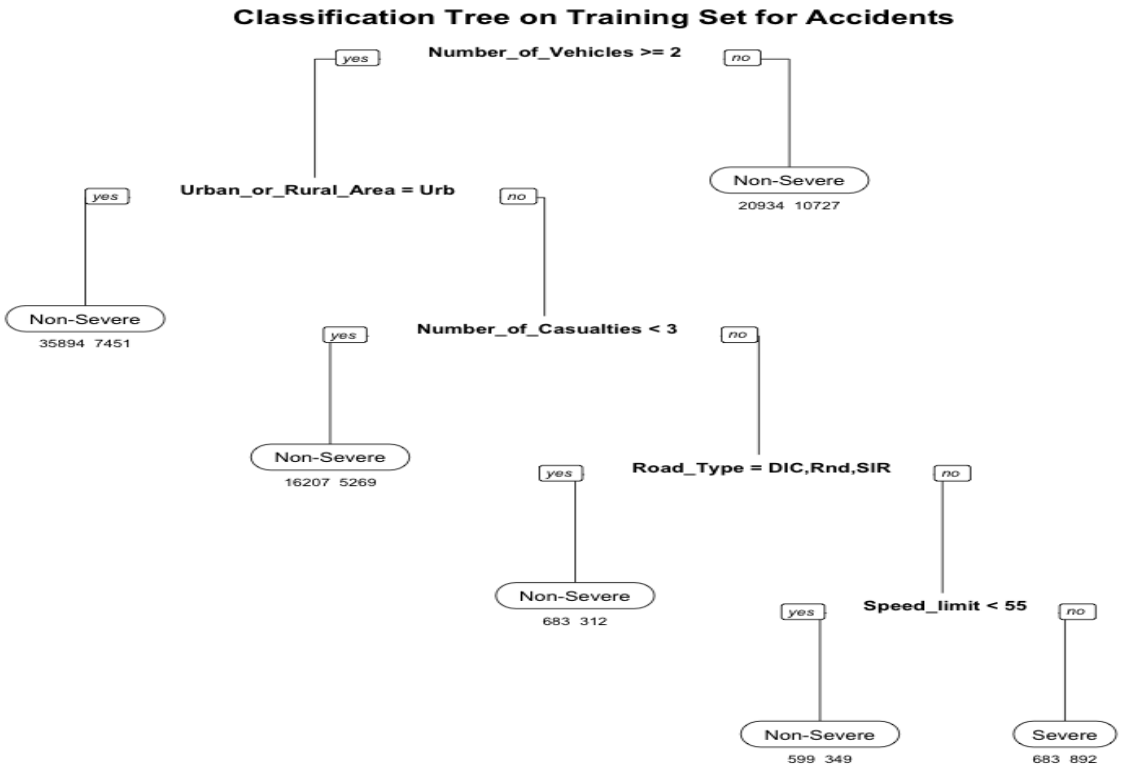



*Figure A.8: Classification Tree for Accidents*

To improve this classification tree model further, we used a 10-fold cross validation method which was repeated three times using 15 different values for the complexity parameter to create a bootstrap aggregation classification tree. The model produced results that were slightly better, producing AR, FPR and FNR of 84.5%, 0.74% and 96.9%.

## Vehicles

All 10 explanatory variables were used in the analysis. The minimum split chosen was 2,500 and complexity parameter was set at 0.001. Overall, this classification tree had an AR, FPR, and FNR of 85.4%, 0.9%, and 94.5% when tested on the testing data.

Interestingly, bicycles were grouped together with cars/taxis and heavy vehicles in terms of accident severity, while motorcycles were classed separately *(see Figure B.8)*. The tree obtained suggests that the groups of people who were more likely to sustain severe injuries are:

- Motorcycles which had not left the carriageway (70.1% severity rates).
- Motorcycle riders above the age of 36 who had left the carriageway and had skidded or overturned were the next most likely group to suffer serious injuries (55.6% severity rates)

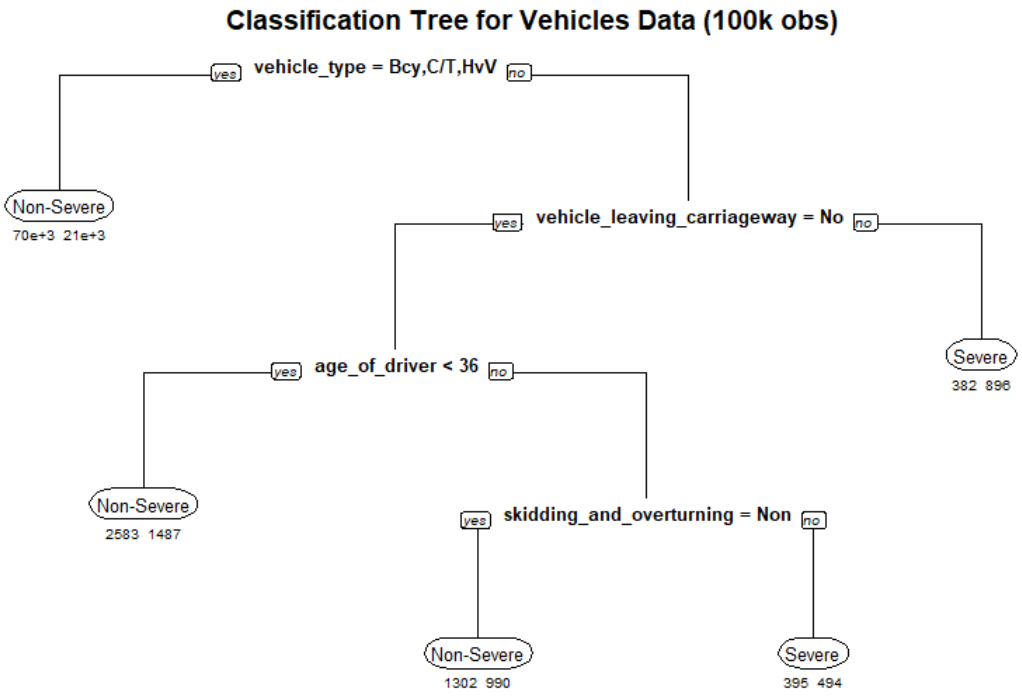



*Figure B.8: Classification Tree for Vehicles*

To improve this classification tree model further, we used a 10-fold cross validation method which was repeated three times using 15 different values for the complexity parameter to create a classification tree with optimal

parameters (see *Appendix C4*). The model produced similar results when tested against the testing set, producing AR, FPR, and FNR of 85.4%, 0.9%, and 94.5%.

## Casualties

The model used for the tree was *Casualty_Severity ~ Casualty_Type + Sex_of_Casualty + Age_of_Casualty*. The minimum split chosen for the classification tree on the casualties training set was 10,000. Additionally, the complexity parameter was set at 0.0001.

The tree obtained below suggests that the groups of people who were more likely to sustain severe injuries are:

- Pedestrians, cyclists, motorcycle or scooter riders/passengers, agricultural vehicle occupants, horse riders and tram occupants
- Bus or coach occupants, car occupants, goods vehicle occupants, taxi/private car hire occupants who are aged 67 or above

This seems to suggest that the type of vehicle plays an important role in affecting accident severity. However, beyond the age of 67, the probability of severe injury increases despite the type of vehicle. We should, however, note that a node below is classified as severe if there at more than 50% severe casualties in that node. The middle and right severe nodes below have 55% and 70% severe casualties, which is not very convincing. Overall, this classification tree had an AR, FPR, and FNR of 66.6%, 32.8%, and 37.0% when tested on the testing data.

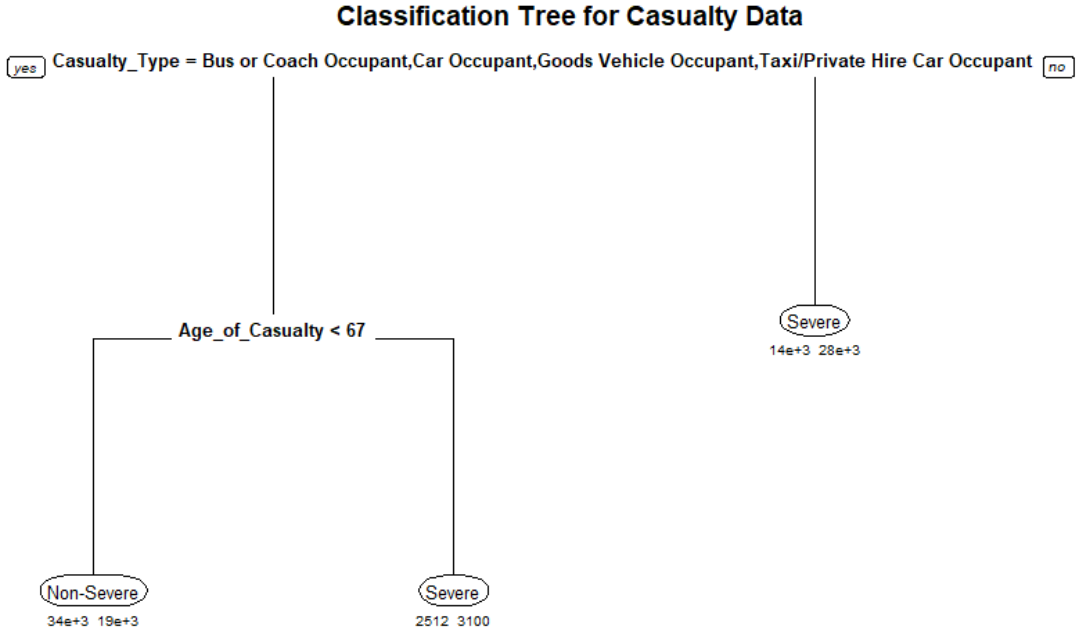



*Figure C.13: Classification Tree for Casualties (all)*

To improve this classification tree model further, we used the same 10-fold cross validation method as in the **Accidents** and **Casualties**. The resulting model produced very similar results when tested against the testing set, producing AR, FPR, and FNR of 67.6%, 31.5%, and 38.3%.

We conducted the analysis for only Pedestrian casualties (*Figure D3(i)* in *Appendix D3* for simple classification tree and *Figure D4(ii)* in *Appendix D4* for 10-fold CV tree respectively). The simple classification tree suggests that groups of pedestrians who are more likely to sustain severe injuries are:

- Any pedestrian aged 57 and over
- Male pedestrians under 57 years old
- Female pedestrians under 57 years old who are crossing the road
- Female pedestrians between 44 and 57 years old who are walking or stationary along the carriageway

The 10-fold CV model returned very similar results in terms of AR, FNR and FPR. However, the tree returned very many nodes, with very close ratios of severe to non-severe casualties in most of those nodes (which is not very convincing)

We also conducted the analysis for only Car casualties (*Figure D3(ii)* in Appendix D3 for simple classification tree and *Figure D4(iii)* in Appendix D4 for 10-fold CV tree respectively). The simple classification tree suggests that groups of Car occupants who are more likely to sustain severe injuries are:

- Any Car occupant aged 70 and over
- Male Car occupants between 14 and 22 years old sitting in the rear seat

Once again, the 10-fold CV model returned very similar results in terms of AR, FNR and FPR. However, the tree returned very many nodes, with very close ratios of severe to non-severe casualties in most of those nodes.

## IV. Random Forest

We performed random forest for all three datasets but only show the results for the **Vehicles** and **Casualties** dataset here as the technique did not yield interesting results for the **Accidents** dataset.

# Vehicles

For our random forests the AR, FPR, FNR of the model are 75.4%, 2.6%, and 90.8% respectively, which is a very bad performance considering the percentage of severe cases in the testing model is ~14% (i.e. if we sweepingly predicted every accident to be non-severe, we would obtain accuracy rates of ~86% minimally). Nonetheless, these results further validate the importance of certain variables.

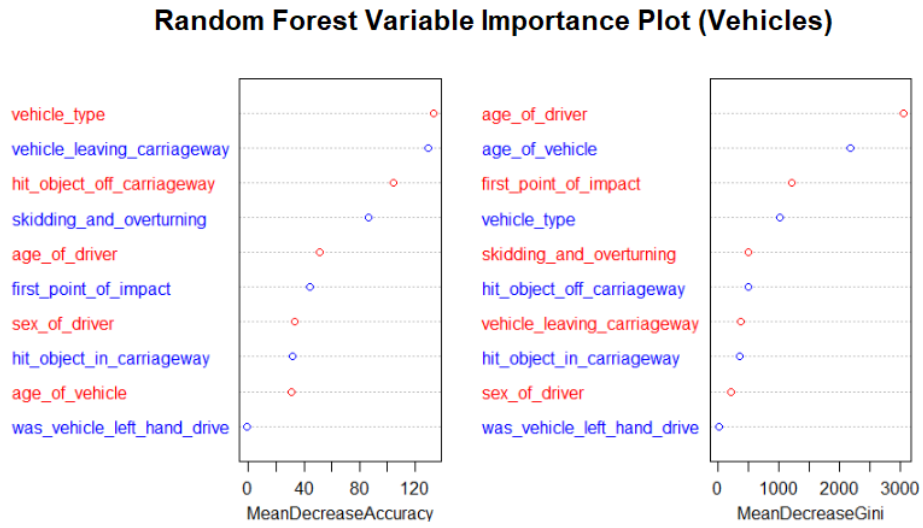



*Figure B.9: Variable Importance Plot from Random Forest technique for Vehicles*

The large mean decreases in accuracy arising from the omission of *Vehicle Type* and *Vehicle Leaving Carriageway* is an indicator that these two variables are relatively useful variables for the classification of severe accidents. Meanwhile, the high mean decrease in Gini coefficient signifies that *age of driver* and *age of vehicle* contribute to the purity of the resulting nodes when these variables are used for classification.

## Casualties

For our random forests the AR, FPR, FNR of the model are 68.1%, 30.9%, and 38.9% respectively, which is a very bad performance considering the percentage of severe cases in the testing model is ~12.8% (i.e. if we sweepingly predicted every accident to be non-severe, we would obtain accuracy rates of ~87.2% minimally).

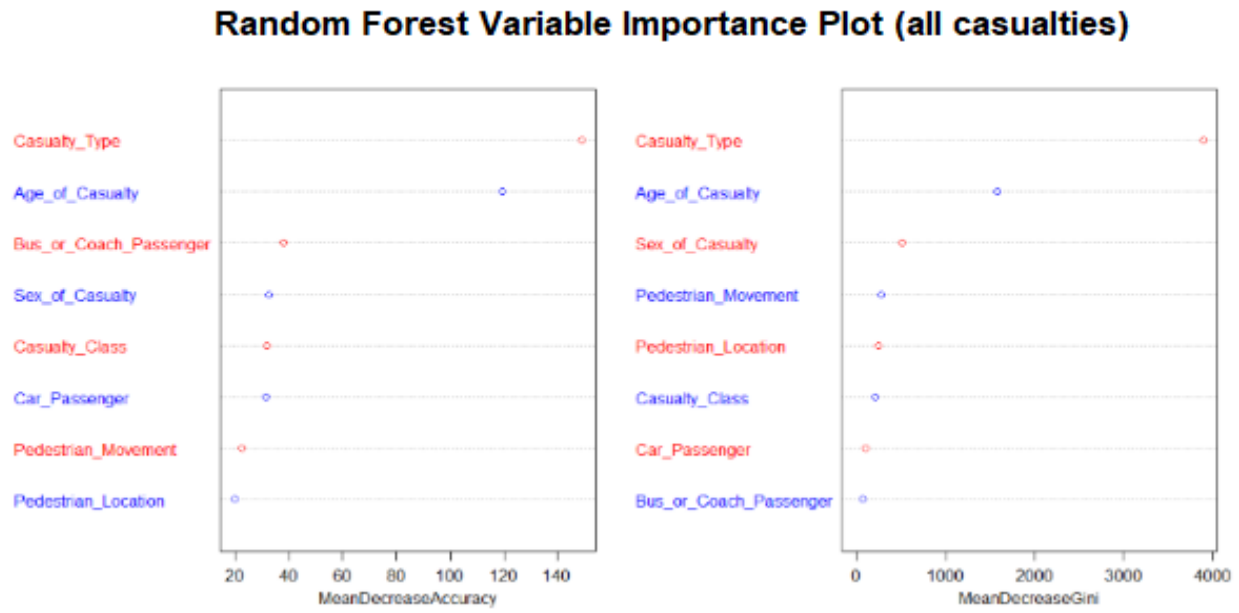



*Figure C.14: Variable Importance Plot from Random Forest technique for Casualties (all)*

The large mean decreases in accuracy arising from the omission of *Casualty Type* and *Age of Casualty* is an indicator that these two variables are relatively useful variables for the classification of severe accidents. Meanwhile, the high mean decrease in Gini coefficient signifies that *Casualty Type* and *Age of Casualty* contribute to the purity of the resulting nodes when these variables are used for classification.

When we conducted the random forests for only Pedestrian casualties or only Car casualties (see *Figure D5(i) and Figure D5(ii)* in Appendix D5), we obtained AR, FPR and FNR that were as inconclusive as the ones above (for all casualties).

For only Pedestrian casualties, it is worth noting that the variables *Pedestrian Movement* and *Pedestrian Location* were listed as two out of the three most important variables for both measures (mean decrease in accuracy and mean decrease in Gini), which is arguably expected since these variables very closely describe the pedestrians right before the accident.

However, for only Car casualties, *Age of Casualty* and *Sex of Casualty* were deemed more important than *Car Passenger*, which is unexpected. This may make us more inclined to believe that position in the car (front seat, rear seat or driver) does not affect the severity of a casualty.

## V. Clustering

# Accidents

Based on the three methods, we chose our optimal number of clusters to be four and three. The Gap statistic method was ignored as it required too many clusters (from the graph above, the gap statistic is increasing in k).

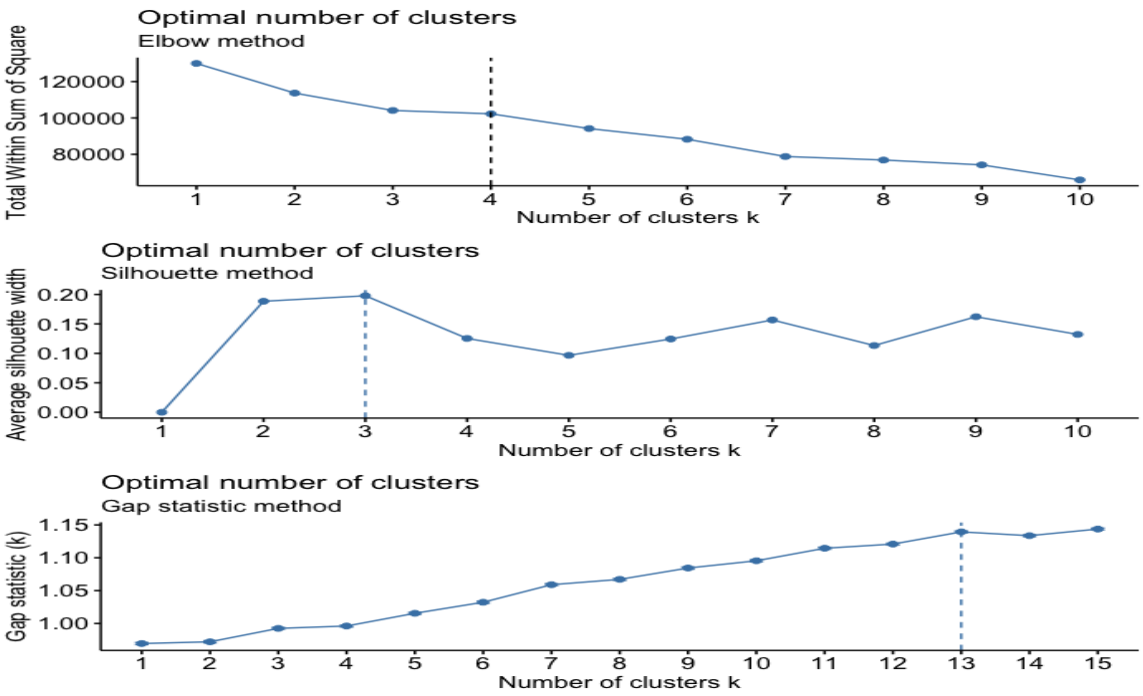



*Figure A.9: Graphs showing the optimal number of clusters for Accidents*

Using k=4, we obtained four clusters of accident severity rates of 19.9%, 16.5%, 13.7% and 13.8%. In order to visualise each cluster, we use decision trees as seen below. In the final nodes, "Yes" refers to a casualty belonging to said cluster, and "No" otherwise.

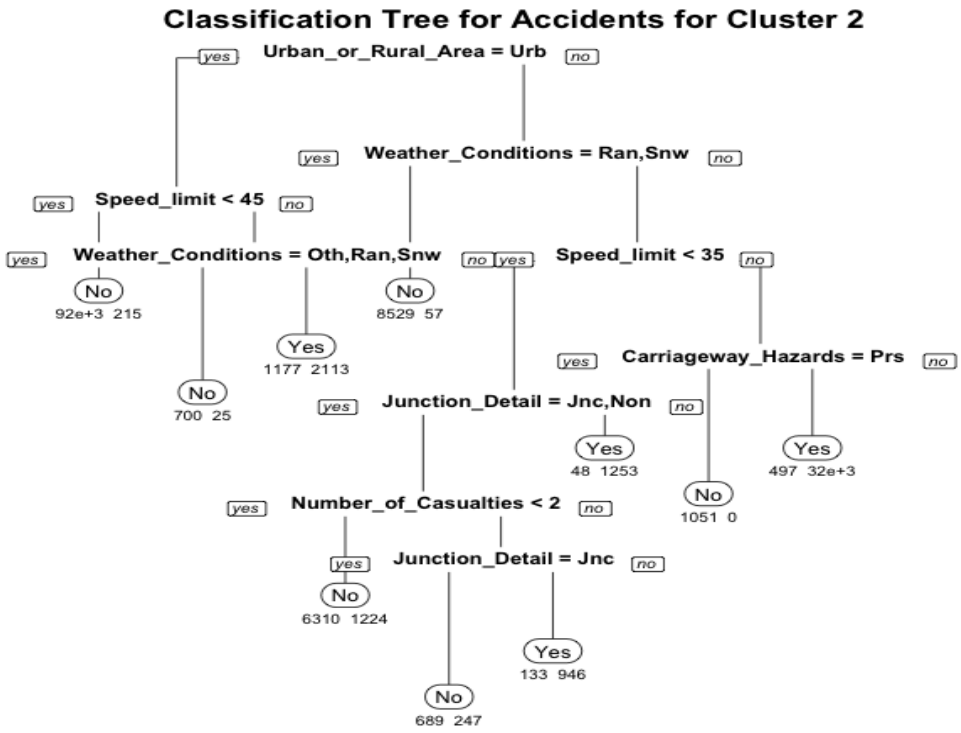



*Figure A.10: Classification tree showing the classification of cluster 2 (for k=4) for Accidents*

The accidents in cluster 2 (19.9% severity rate) are characterised by accidents that involve two vehicles, have one casualty, occurred in during the day in good weather conditions on a Sunday on roundabout with a speed limit of 60mph which was dry in a rural area.

It is interesting to note that both clusters 2 and 3 (highest and lowest severity rates) have some of the same common characteristics - *Two Vehicles involved, One Casualty, occurs on a Sunday in Daylight.* The difference between the two lies in the difference in Road Type, Speed Limit, Weather Conditions, whether the accident was in an Urban area, Road Surface conditions and Junction detail

To further illustrate the differences across clusters, we created a classification tree based on cluster numbers, showing the case for which k=4 comparing clusters 3 and 2.

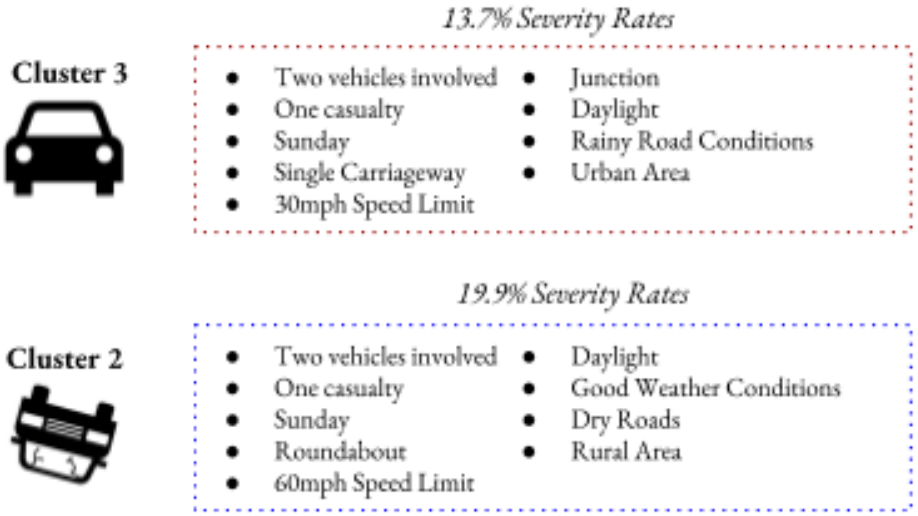



*Figure A.11: Graphic showing cluster representatives for Accidents dataset*

## Vehicles

Based on the three methods, we chose our optimal number of clusters to be two and four. The Gap statistic method was ignored as it required too many clusters (from figure B.10, the gap statistic is increasing in k).

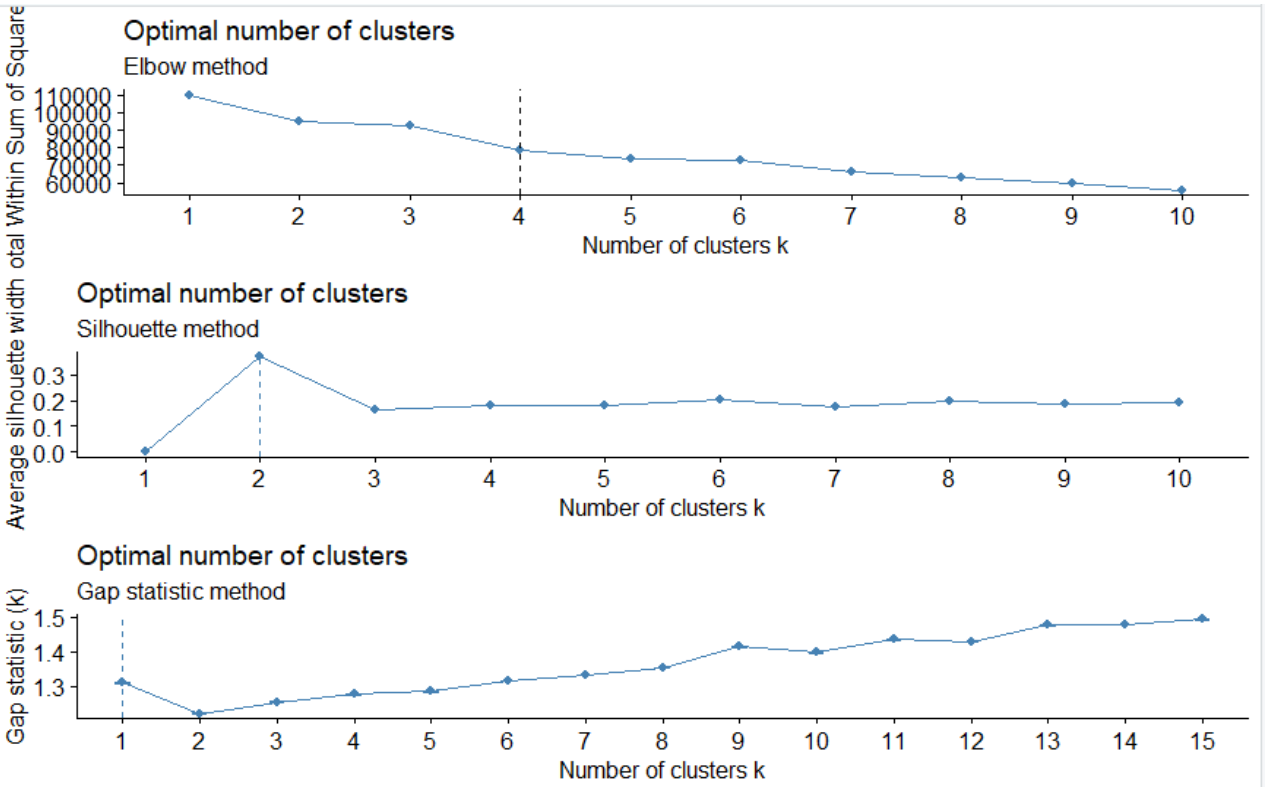



*Figure B.10: Graphs showing the optimal number of clusters for Vehicles*

Using k=2, we obtained two clusters of accident severity rates of 22.8% and 13.4%. While the percentage point difference appears deceptively small, the percentage difference in accident severity rates between these two clusters actually represent a 70% difference in rates between cluster 1 to 2. To visually inspect these differences, we created a classification tree for this case *(see figure B.11).*

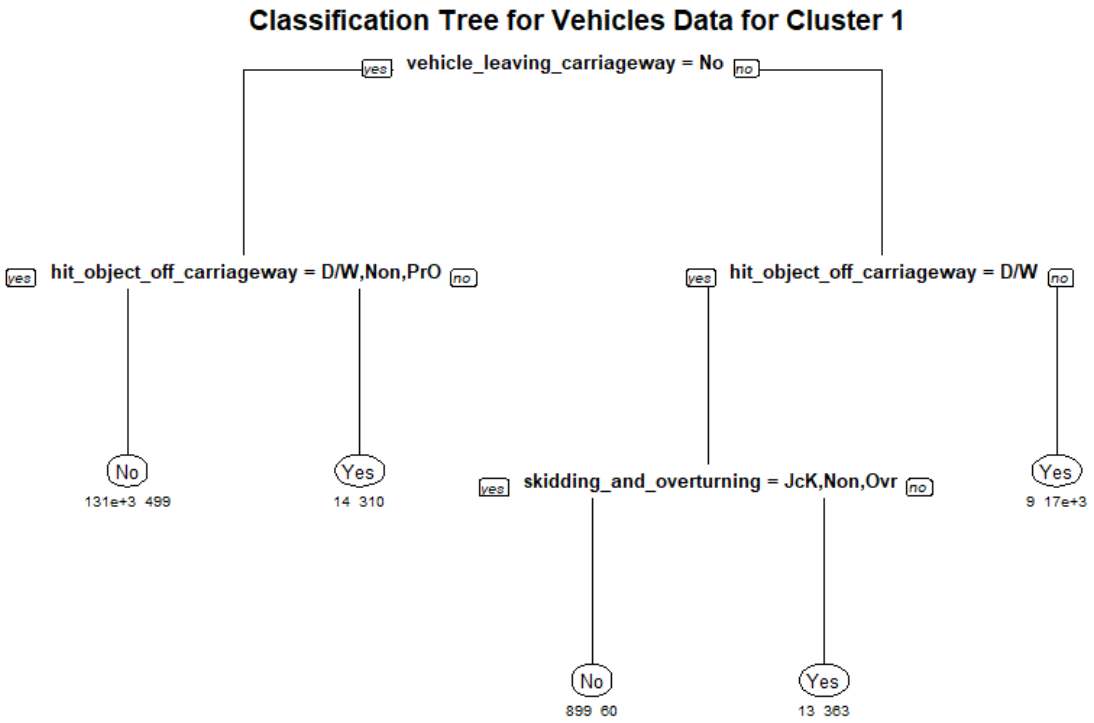



*Figure B.11: Classification tree showing the classification of cluster 1 (for k=2) for Vehicles*

We observe that accidents in cluster 1 (22.8% severity rate) are characterised by vehicles which had left the carriageway and had collided into a pole or tree. In addition, cluster 1 accidents can also be classed as vehicles which had left the carriageway but did not land into a ditch or waterbody. In the other cluster (13.4% severity rate), accidents are mainly vehicles which had left the carriageway and had hit nothing, a ditch or waterbody or a permanent object such as a bus stop or crash barrier.

For the case k=4, the clusters had severity rates of 23.3%, 10.6%, 21.0%, and 14.7%. To further illustrate the differences across clusters, we created a visual representation of the centroid's nearest neighbour. Looking at these cluster representatives, it is apparent that there are many similar characteristics (e.g. age of vehicle, vehicle type, etc.). However, some notable differences are the age and sex of the driver, the first point of impact on the vehicle, and whether the vehicle had left the carriageway.

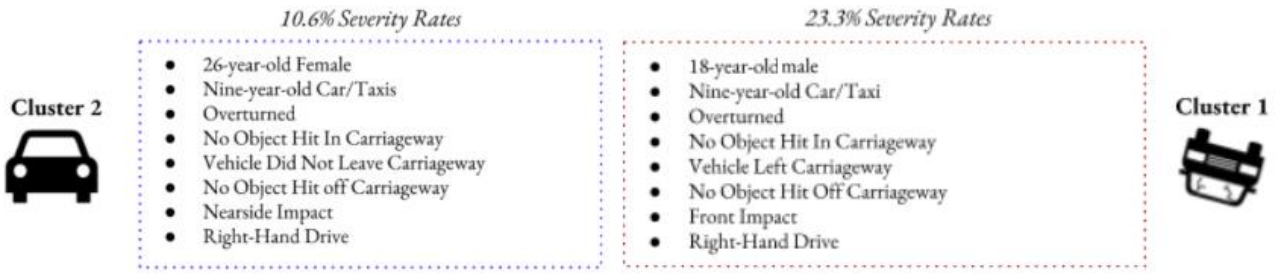



*Figure B.12: Graphic showing cluster representatives for Vehicles dataset (for k=4)*

## Casualties

Using k=7 (the median of the 3 methods), clusters 1 and 5 had the smallest and largest severity rates (5.9% and 30.8%) respectively. To find the differences between each of these clusters and all other clusters (intergroup differences), we use decision trees derived below.

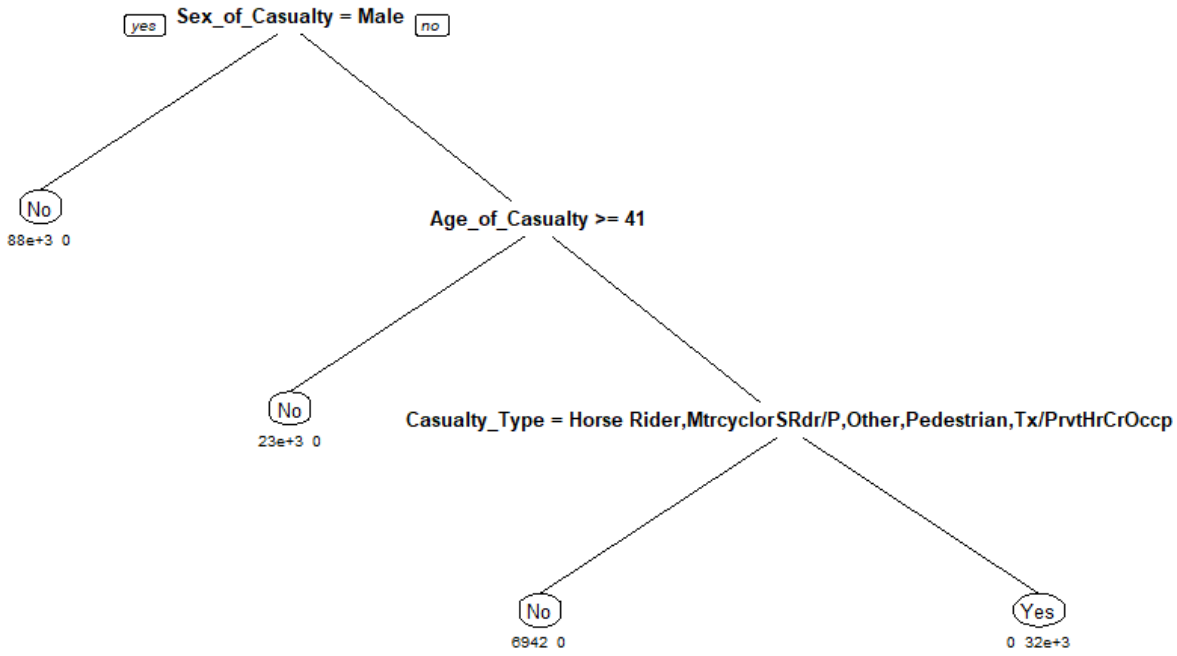



*Figure C.16: Classification tree showing the classification of cluster 1 (for k=7) for Casualties*

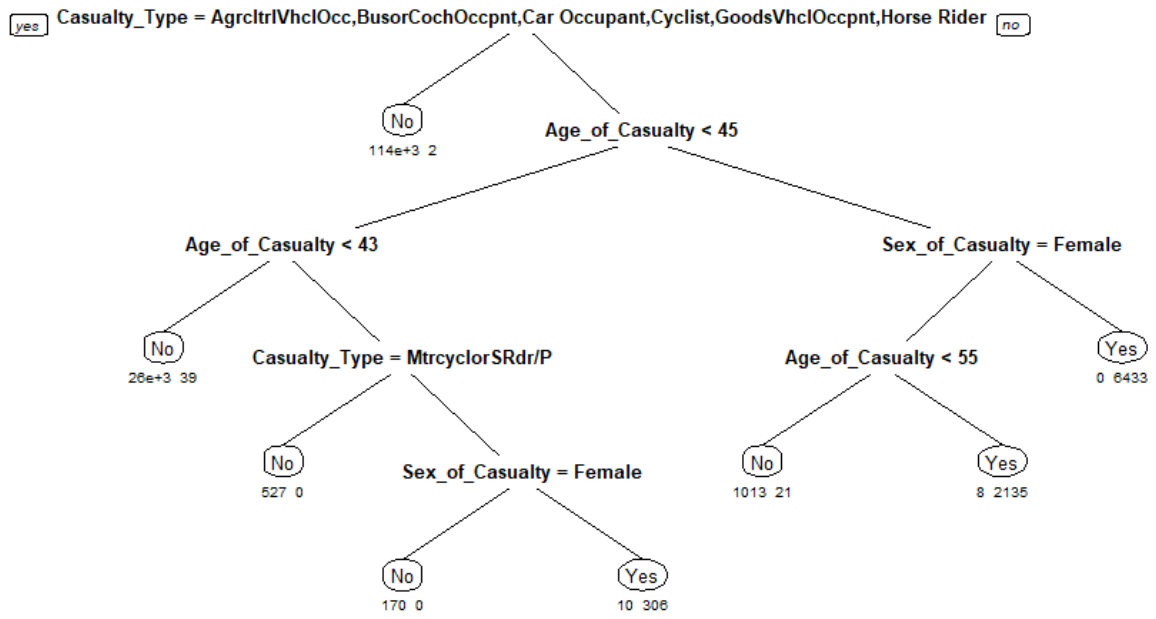



*Figure C.17: Classification tree showing the classification of cluster 5 (for k=7) for Casualties*

We observe that accidents in cluster 1 are characterised by Females under 41 years old who are either cyclists, car occupants, bus or coach occupants, goods vehicle occupant, tram occupant or agriculture vehicle occupant.
Accidents in cluster 5 are characterised by:

- Males aged 45 and over, who are motorcycle or scooter riders/passengers, taxi/private car hire occupants or tram occupants
- Females aged 55 and over, who are motorcycle or scooter riders/passengers, taxi/private car hire occupants or tram occupants
- Males aged between 43 to 45, who are taxi/private car hire occupants or tram occupants

To illustrate the similarities across clusters, we created a visual representation of the cluster representative, which is the observation closest to the cluster centroid (measured using Euclidean distance). Table C.3 shows the clusters with lowest and highest severity rate for all casualties.

In addition, we have also included Figures C.19 and C.20, which show the clusters with lowest and highest severity rate obtained using data with only pedestrian casualties and only car casualties respectively.

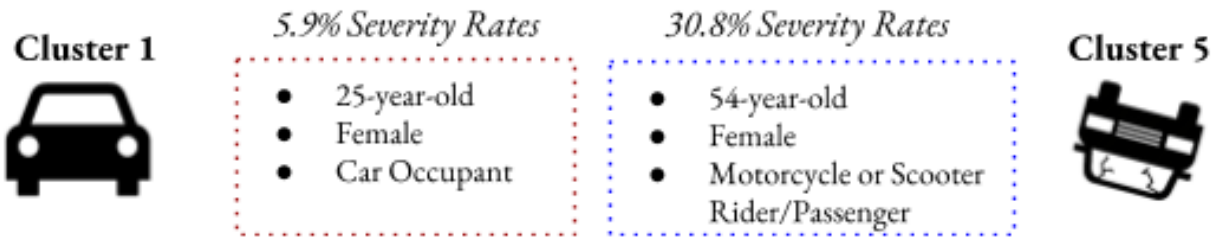



*Figure C.18: Graphic showing cluster representatives for Casualties s dataset all)*

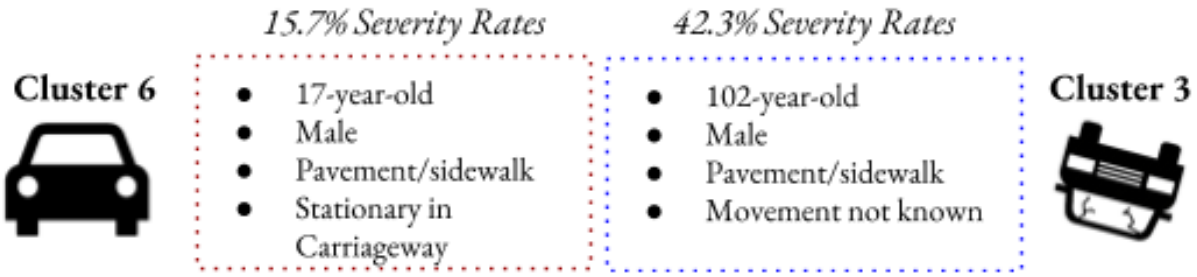



*Figure C.19: Graphic showing cluster representatives for Casualties s dataset (Pedestrians)*


*Figure C.20: Graphic showing cluster representatives for Casualties s dataset (Cars)*

From the information above, we realise that the only systematic difference between clusters with low severity rates and clusters with high severity rates is that the latter tends to have older casualties. Later in this report, we argue why the details of the cluster representatives above may not be a good representation of the casualties in each cluster.

# 4. Implications of Results

| No | Point | Implication |
|---|---|---|
| | **Accidents** | |
| 1 | Accidents on Sundays and Mondays had a higher probability of being severe accident | Rental car companies can look to introduce steeper premiums for rentals that include those two days. Additionally, first responders can also look to staff themselves appropriately on those day to ensure that they are able to respond to the accidents effectively and in an appropriate manner |
| 2 | Accidents on roads with a speed limit of 60 have a higher probability of being severe | Authorities could introduce new or revised regulations to try and promote safer driving. This could be through new speed cameras or potentially revising the speed limit in certain areas |
| 3 | The probability of an accident being severe in perceived safe conditions is surprisingly like perceived adverse conditions | Drivers perceive certain conditions (Darkness/Snow/Wet road) as more dangerous than others (Dry/Daylight) and as a result, take more caution and drive safer in those conditions |
| | **Vehicles** | |
| 1. | Rear-end impact on vehicles produced the lowest predicted probabilities of a severe accident | Vehicle manufacturers can identify structural weaknesses in the front and side body of their vehicles if the relatively high severity rates are caused by weak durability in those areas |
| 2. | Probability of a severe accident increases proportionally with age of vehicle & driver[9] | Authorities can consider increased regulation of old vehicles and require more frequent checks on drivers as they get older |
| 3. | Motorcycles which got into an accident on the carriageway were the most susceptible to a severe injury | Regulating the speed limit for motorcycles on the carriageway and/or designating special lanes for motorcycles may better safeguard motorcycle riders[10] |

[9] Specifically, probability of a severe accident decreases from 16-25 years old but increases thereafter..

[10] In fact, this has been recommended before by the World Health Organisation in 2010 (WHO, 2010)

| Casualties | | |
|---|---|---|
| 1. | The probability of severe casualty increases almost linearly with age after the age of 20 | Car insurers may wish to increase insurance premium based on the age of driver. Health insurers may want to take into account whether the applicant owns a car. Furthermore, the transport authority may wish to re-examine drivers beyond a certain age, when they are deemed more likely to get into an accident causing severe injury |
| 2. | For car passengers, being in the front seat, rear seat or driver, is insufficient to pin down severity of a car casualty | First responders could collect more data in this regard, such as the number and position of airbags and where in the rear seat the passenger was seated (e.g. side or middle) |
| 3. | Pedestrians and Motorcycle/Scooter riders/passengers have a higher probability of sustaining a severe injury than occupants of covered vehicles (e.g cars and buses) | Urban planners could find a way to isolate walkways for pedestrians and lanes for 2-wheeled vehicles. Furthermore, walkways and sidewalks could be made safer for senior people (e.g. better visual aids) since they are the ones who are most likely to suffer a severe injury |

# 5. Performance Measurement

While the use of AR, FPR, and FNR are good preliminary measures of the performance measure of our models, they do not consider the nature of the data and thus the relative consequences of the different mistakes (i.e. getting a false positive does not incur the same cost as a false negative).

As such, we have designed a cost-benefit matrix to quantify the true benefits and costs of a TP, TN, FP, FN. The values have been fitted into the table below and the subsequent paragraphs will expound on how these figures were obtained.

**Cost-Benefit Matrix for Accidents**

| | Non-Severe | Severe |
|---|---|---|
| Non-Severe | TRUE NEGATIVE +£25.8k | FALSE NEGATIVE -£235k |
| Severe | FALSE POSITIVE -£70.2k | TRUE POSITIVE +£400k |

**Cost-Benefit Matrix for Casualties**

| | Non-Severe | Severe |
|---|---|---|
| Non-Severe | TRUE NEGATIVE +£15.0k | FALSE NEGATIVE -£208k |
| Severe | FALSE POSITIVE -£62.3k | TRUE POSITIVE +£181k |

*Figure D.1: Cost-Benefit matrix for accidents and casualties*[11]

Much of this section is based on the Department for Transport's Accident and casualty costs and Reported Road Casualties in Great Britain: 2012 Annual Report (UK Department for Transport, 2018). The RAS60 provided the average cost of a casualty and accident over the period 2010-17, which was instrumental in providing us the benefit of identifying a TP and TN. On the other hand, the 2012 annual report provided a breakdown of the different types of costs associated with an accident/casualty, allowing for us to quantify the true cost of a FP and FN.

**Benefits of True Positive and True Negative**

We begin by retaining the original accident severity classifications of 1,2,3 (representing fatal, severe, and slight respectively). Using the cost per accident for the years 2010-17 (for which data was available), we calculated the growth rate of the nominal increase in cost of each severity type (i.e. slight, severe, and fatal). Then, we extrapolated the missing values for cost of accident[12] for the years 2005-09 based on the growth rate calculated:

[11] Both the **Accidents** and **Vehicles** dataset use the accident CB matrix, while **Casualties** uses the CB matrix for casualties.

[12] Same calculation applies to cost of casualty.

$$Nominal\ Cost\ of\ Accident\ in\ 2005\ =\ Nominal\ Cost\ of\ Accident\ in\ 2017 * (1 - Growth\ Rate)^{12}$$

For like to be compared with like, we adjusted each year's cost to 2017 prices on the assumption that the year-on-year inflation rate was 2.0%. For example, for the year 2005, the real value of a cost of accident would be:

$$Real\ Cost\ of\ Accident\ in\ 2005\ =\ Nominal\ Cost\ of\ Accident\ in\ 2005 * (1 + Inflation\ Rate)^{12}$$

The next step is to take the mean value of each type of accident across the years (i.e average cost of fatal accidents from 2005 to 2017) and assign each observation with the average cost depending on their severities.

At this point, we begin classifying fatal and severe accidents as *severe,* and slight accidents as *non-severe*. Taking the average of the cost of *severe* and *non-severe* accidents across all observations, we obtain our values of a TP and TN respectively (i.e. the benefit of identifying a TP is the average cost that a severe accident would incur and the benefit of identifying a TN is the average cost that a non-severe accident would incur).

**Costs of False Positive & False Negative**

**RAS60003: Total value of prevention[1] of reported accidents by severity[2] and cost element: GB 2012**

£ million

| | Cost Elements | | | | | | |
|---|---|---|---|---|---|---|---|
| | Casualty related costs | | | Accident related costs | | | |
| Accident severity | Lost output | Medical and Ambulance | Human costs | Police costs | Insurance and admin | Damage to property | Total |
| Fatal | 1,040 | 9 | 2,042 | 29 | 1 | 19 | 3,139 |
| Serious | 526 | 315 | 3,582 | 44 | 4 | 108 | 4,578 |
| Slight | 389 | 165 | 1,854 | 67 | 15 | 381 | 2,871 |
| All injury accidents | 1,955 | 490 | 7,478 | 139 | 19 | 508 | 10,589 |
| Damage only accidents | 0 | 0 | 0 | 77 | 124 | 4,332 | 4,533 |
| **All accidents** | **1,955** | **490** | **7,478** | **217** | **143** | **4,840** | **15,122** |

1 The number of reported road accidents were based on 2012 data
2 The costs were based on 2012 prices and values

*Figure D.2: Table on total value of prevention of accidents by severity*

To obtain the costs of FP and FN, several assumptions were made on what these figures comprised. Most, if not all, of these assumptions stemmed from the RAS60003[13] (*see table above)*. To calculate the cost of an accident,

[13] Obtained from the Reported Road Casualties in Great Britain: 2012 Annual Report.

we have extrapolated the information under "Casualty related costs" on the basis that a typical accident involves 1.79 casualties[14], rather than from "Accident related costs".

We have done this for two reasons: first, the accident-related costs are significantly smaller than the casualty-related costs and focusing on the latter will sufficiently capture the relative differences in cost between an FP and FN. Secondly, and more importantly, we believe that accident-related costs do not accurately reflect the cost of an FP or FN. Thinking about what truly constitutes the cost of an FP or FN, we believe firmly that it is driven mainly by the costs incurred through the worsening of a casualty's or a stakeholder's position, rather than through the prevention of accidents per se - a central tenet to the formulation of the accident-related costs in the table. For these reasons, it would be logical and practical to focus on the computation of the CB matrix through casualty-related costs.

For the FP cost, we assumed it to be the sum of Medical Costs (MC) and Psychological Costs (PC). The MC arises when a hospital or emergency care unit prepares to treat a severely hurt patient when the patient is in fact not. This figure was obtained by taking the difference in MC between treating patients from severe and non-severe accidents. Meanwhile, the PC is best illustrated by the age-old classic: "The Boy Who Cried Wolf". When a hospital receives a FP, it believes that the next severe accident is also a FP. In our model, we assume that this incurs a cost of the negative of the benefit of a TP (i.e. the next patient that comes in is actually severely hurt, but the hospital believed otherwise and did not respond appropriately).

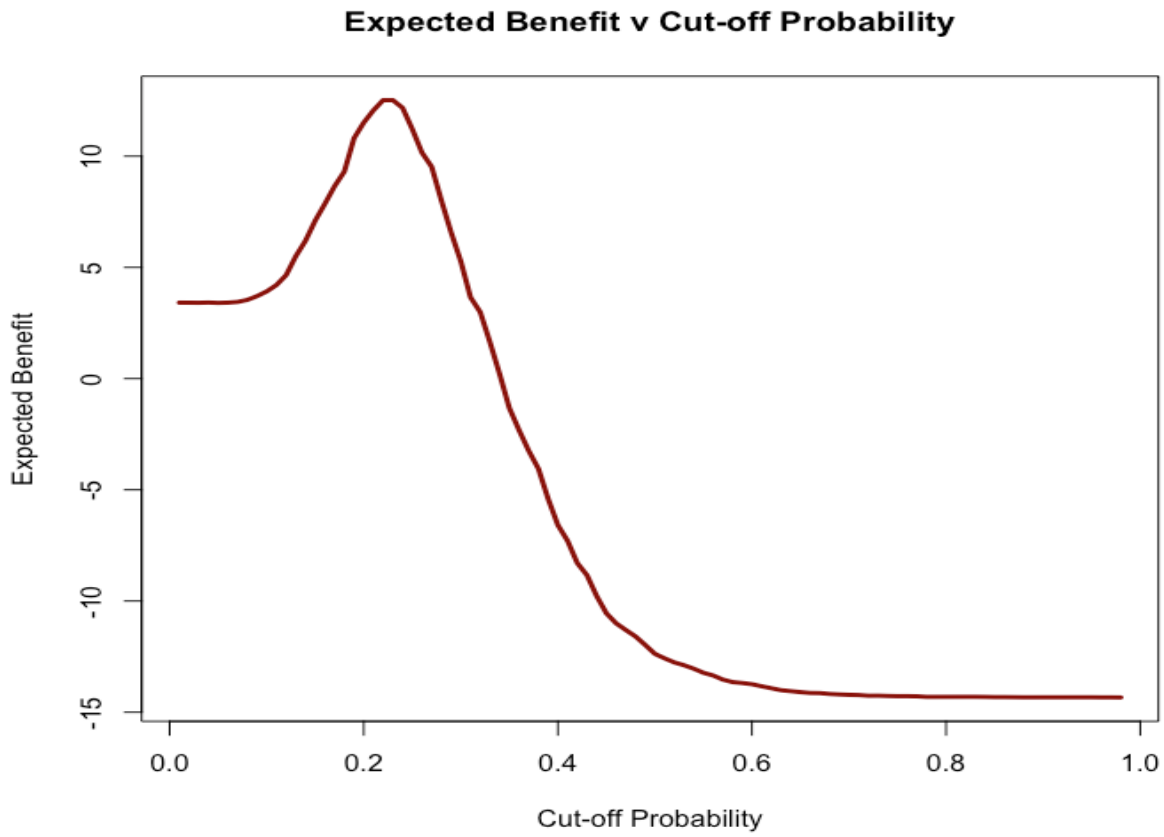
The FN is quantified by the Human Cost (HC) incurred when we classify an accident as non-severe, when it is in fact severe. This comes out to be the difference in HC incurred between a severe and non-severe accident. The full calculations can be found in *Appendix A2: Calculation of False Positive & False Negative Benefit*, but all calculated values of TP, FN, FP, TN can be found in the table above.

## a. Accidents Performance

The expected benefit of the non-adjusted (i.e. using a 0.5 cut-off probability) full logistic regression, non-adjusted skimmed logistic regression, random forest, full classification tree, and classification tree with k-fold CV are -12.36, -12.38, -8.97, -11.62, and -11.87 respectively.

However, through a cursory glance of the predictions of the logistic regression models, a cut-off probability of 0.5 appears arbitrary and unsuitable for our model. Plotting a graph of expected benefit against the cut-off probability for both logistic regression models, we obtained an optimal cut-off probability of 0.22 and 0.23 respectively.

[14] In the case of a fatal accident, 1.07 casualties are fatal, 0.29 serious, and 0.43 are slight based on the 2012 annual report; for the case of a slight accident, we assume that there are 1.79 casualties involved.




*Graph A.11: Expected benefit vs cut-off probability for the logistic regression model in the* ***Accidents*** *dataset*[15]

Using the optimal cut-off that we just got, we reran the calculations for the expected benefits of our 2 logistic regression models, obtaining expected benefits of +12.56 (Full) and +12.52 (Skimmed). These results highlight the superiority of using logistic regression models over the other methods employed.

To ascertain this conclusion, we generated ROC curves and calculated the AUC of each method. Unsurprisingly, we obtained almost identical results: the AUC for the full logistic regression, skimmed logistic regression, RF, Class Tree, and K-fold Class Tree were 0.59, 0.59, 0.53, 0.51, and 0.51 respectively (see graph below). Given these results, it appears that our full and skimmed logistic regression models performed the best - a similar result to our expected benefit test.

[15] We obtained the same results and graph for both logistic regression models. As such, we have omitted the skimmed logistic regression model's graph and have only displayed the full logistic regression model.

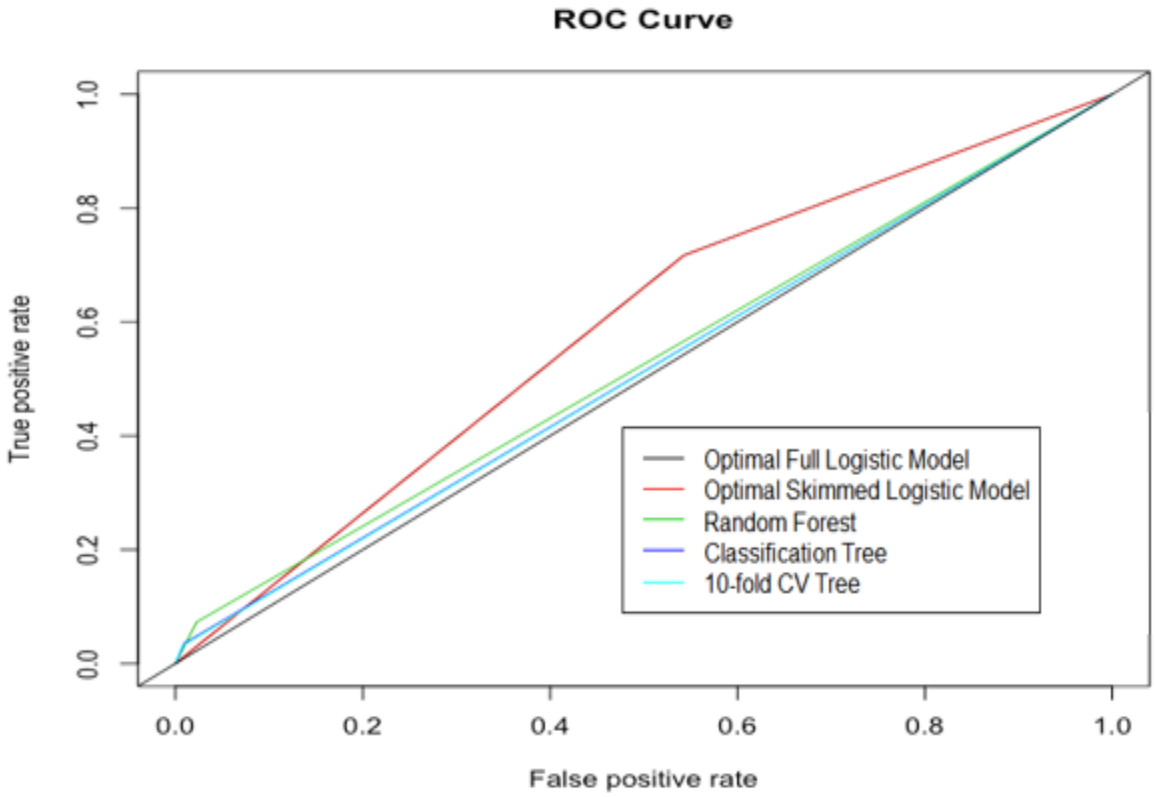



*Figure A.12: ROC graph for the various methods for the Accidents dataset*

However, when plotting the ROC curve, it became clear that none of these methods were particularly well suited to predict an accident severity - shown through the curves' proximity to the diagonal.

## b. Vehicles Performance

The expected benefit of the non-adjusted (i.e. using a 0.5 cut-off probability) full logistic regression, non-adjusted skimmed logistic regression, random forest, full classification tree, and classification tree with k-fold CV are -7.93, -8.31, -5.75, -7.92, and -7.92 respectively.

Plotting a graph of expected benefit against the cut-off probability for both logistic regression models, we obtained an optimal cut-off probability of 0.21 for the full model and 0.23 for the skimmed model.

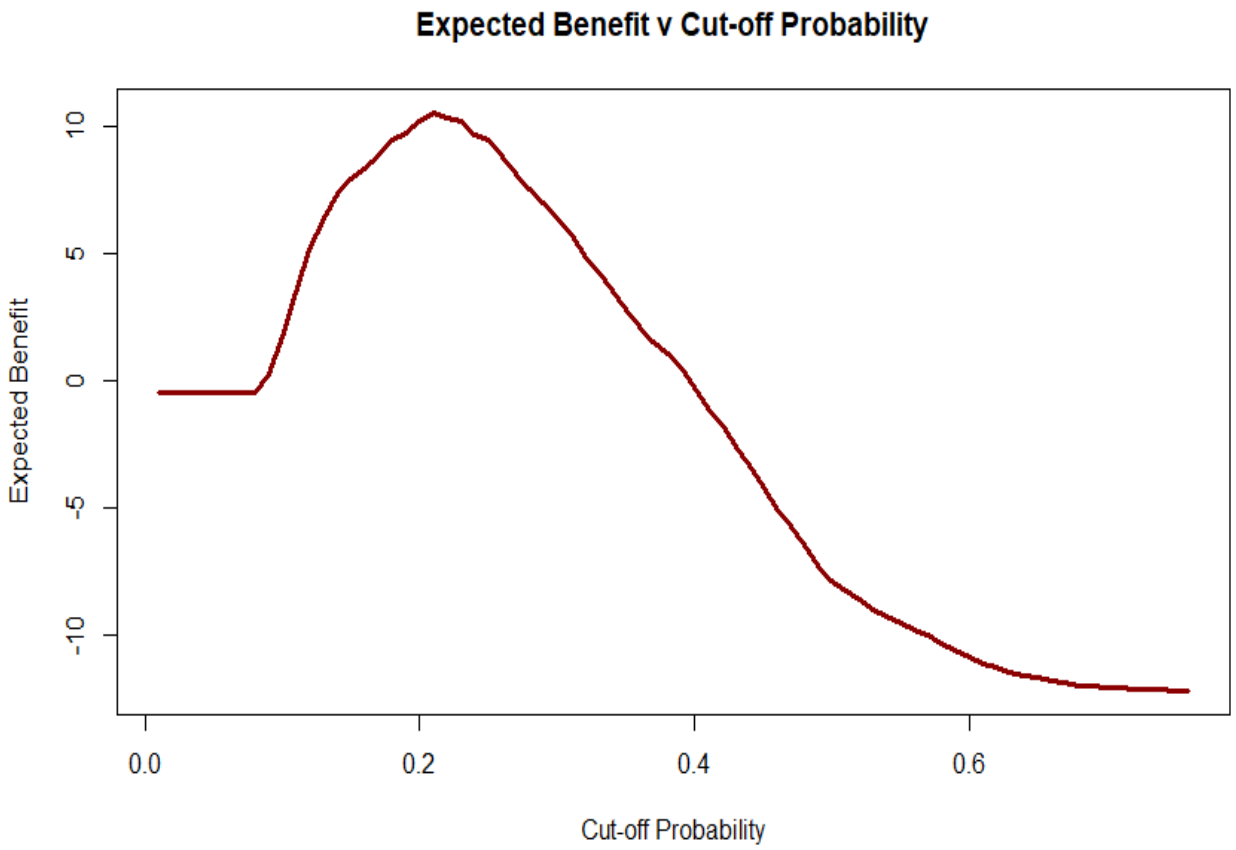



*Figure B.13: Expected benefit vs cut-off probability for the logistic regression model in the **Vehicles** dataset*[16]

Recalculating the expected benefit of our logistic regression models with the new optimal cut-off probability, we obtained expected benefits of +9.66 (full) and +9.43 (skimmed). These results affirm the advantage of the logistic regression models over the other methods, which had expected benefits of -5.75 (RF), -7.92 (Class Tree), and -7.92 (K-fold Class Tree)[17].

Our ROC curves and AUC produced identical results: the AUC for the full logistic regression, skimmed logistic regression, RF, Class Tree, and K-fold Class Tree were 0.585, 0.591, 0.533, 0.522, and 0.522 respectively (see figure B.14). Given these results, it appears that our skimmed logistic regression model performed the best - a similar result to our expected benefit test.

[16] Only the graph for the full logistic regression case is shown.

[17] Rehashed from the previous page.

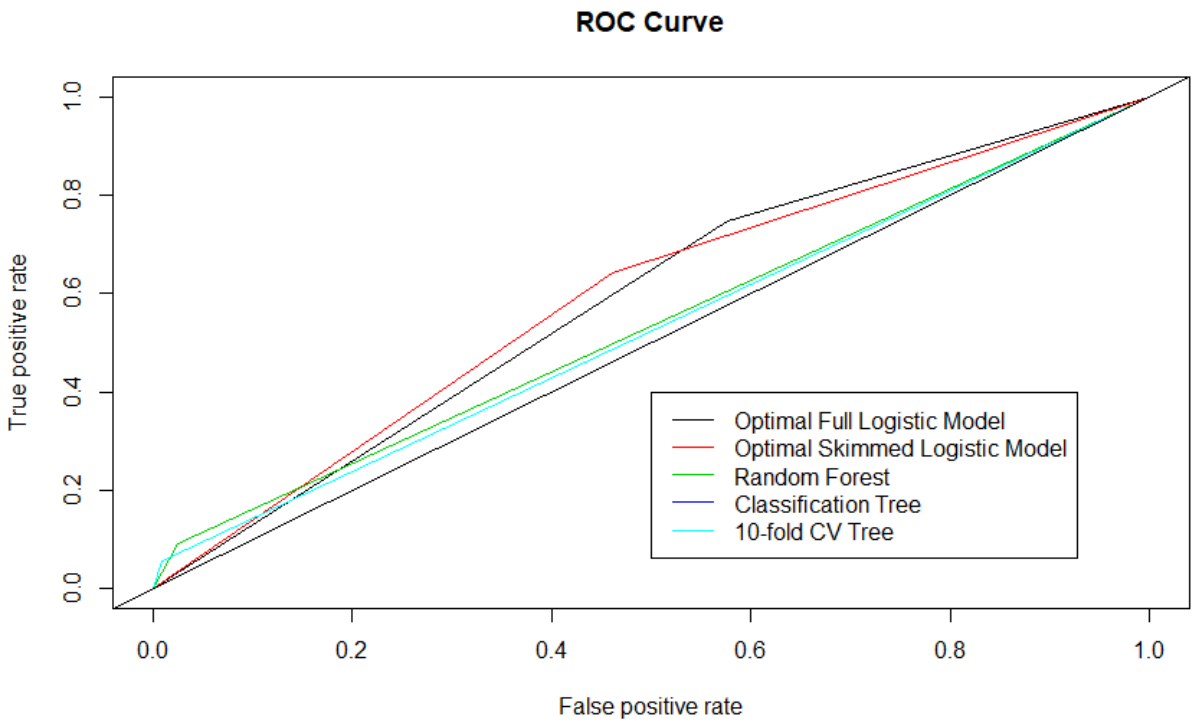



*Figure B.14: ROC graph for the various methods for the Vehicles dataset*

Unfortunately, the proximity of our ROC curves with respect to the diagonal is a signal that neither of these methods are particularly adept at predicting an accident's severity and are as good as random.

## c. Casualties Performance

The expected benefit of the (i.e. using a 0.5 cut-off probability) logistic regression (final model specified in Section 4(b)(II), random forest, full classification tree, and classification tree with 10-fold CV are -3.86, -4.02, -4.37 and -4.11 respectively. Hence, the logistic regression appears superior at first glance.

Next, we plot a graph of expected benefit against the cut-off probability for the logistic regression models. We obtained an optimal cut-off probability of 0.58 in this case. Figure C.21 below shows a marginal difference in expected benefit when we compare the original cut-off probability of 0.50 and the optimal cut-off probability of 0.58.

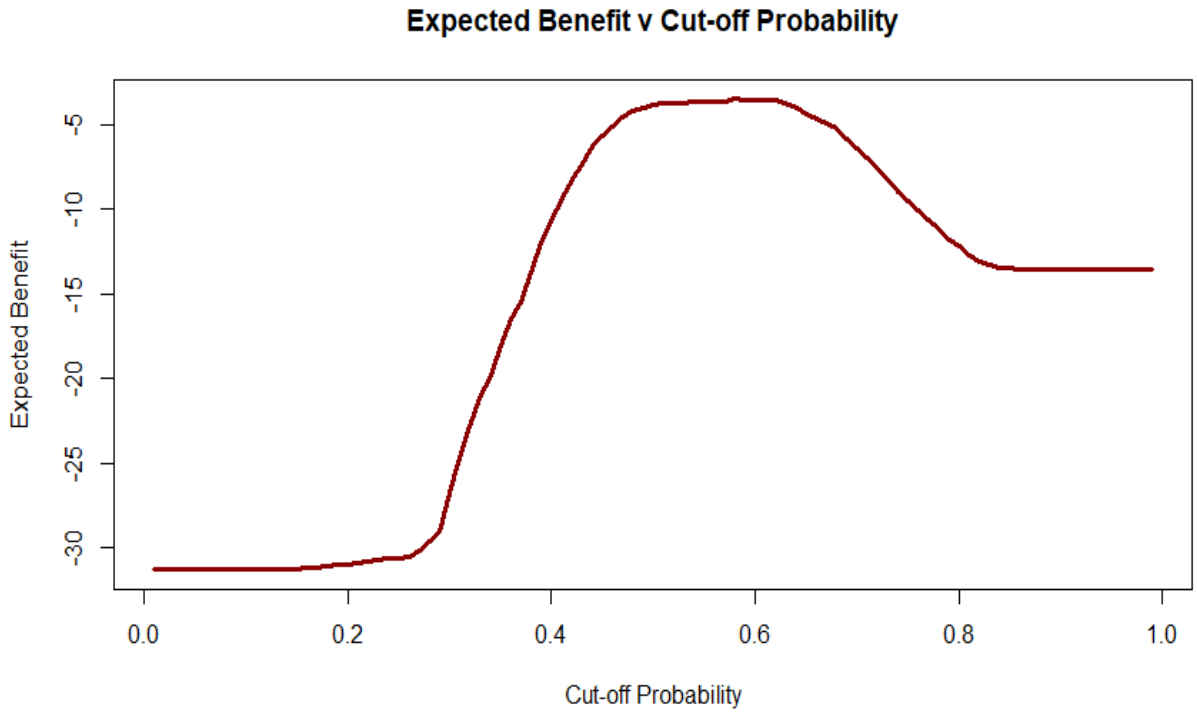



*Figure C.21: Expected benefit vs cut-off probability for the logistic regression model in the Casualties dataset*

Recalculating the expected benefit of our logistic regression model with the new optimal cut-off probability, we obtained expected benefits of -3.48. This result affirms the advantage of the logistic regression models over the other methods, which had the aforementioned expected benefits.

For the subsets comprising only Pedestrian casualties and only Car casualties, similar results were gleaned when the same method was applied i.e. the logistic regression with optimised cut-off probability always performed the best based on our cost benefit analysis

Alternatively, the AUC for the optimised logistic regression, Random Forest, Classification Tree, and 10-fold Classification Tree were 0.638, 0.651, 0.651 and 0.651 respectively (see figure C.22). Given these results, it appears that our skimmed 10-fold Classification Tree performed the best. However, the Classification Tree and Random Forest models performed very nearly as good.

For the subsets comprising only Pedestrian casualties and only Car casualties, the ROC curves were inconclusive as the AUC values were very close to 0.5, rendering the models as good as random.

Aggregating the results from the cost-benefit analysis and ROC curves, it is difficult to pin down a model that is best for explaining how the variables in our Casualty dataset affect *Casualty Severity*.

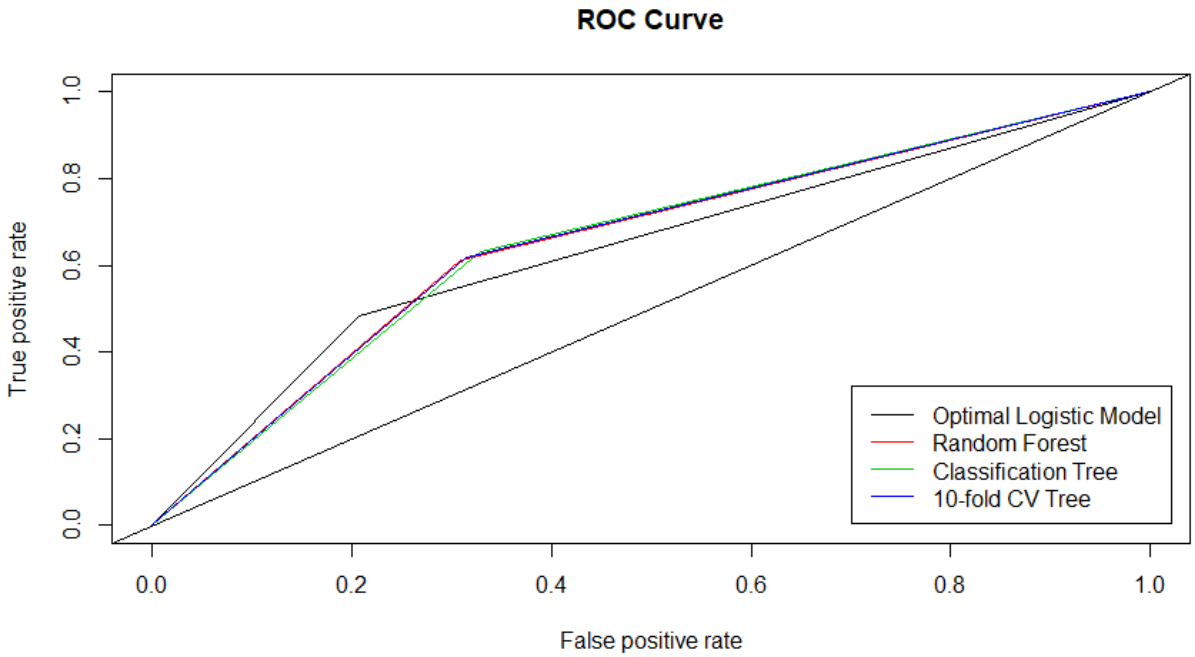



*Figure C.22: ROC graph for the various methods for the **Vehicles** dataset*

## 6. Shortfall in Methodology

Despite our best attempts in ensuring the robustness of our methodology, we acknowledge that there are certain shortfalls and assumptions of our methods which may be erroneous.

Firstly, for the Vehicles dataset, we assumed that all vehicles involved in a severe accident (as defined by the Accidents dataset) were also severe by matching the accident indexes of accidents to vehicles (i.e. two vehicles may have the same accident index if they were involved in the same accident). According to the STATS20 form (UK Department of Transport, 2011), an accident is classified as severe if there is at least one severe casualty involved. However, we can think of several instances in which the above assumption may be rendered false. For example, if two cars collided but all the casualties involved were confined to a single car, then we would have a case where one car would be classed as a severe accident while the other car would be a non-severe accident. In such an example, our assumption of both cars being severe accidents would be false.

The second, potentially erroneous shortfall of our methodology is the low predictive power for all our prediction models (low accuracy rates, high FNR/FPR). Despite our attempt to improve upon our models, we generally could not produce sufficiently robust prediction models. As a result, some of the implications and graphs derived from, say, the logistic regression models may not be wholly accurate.

The third shortfall in our methodology is the potential inaccuracies in our CB analysis, where we may not have precisely predicted the costs associated with TP, TN, FN, and FP. Whilst we assumed that medical and

psychological costs were the sole contributor to the cost of a FN, we may have inadvertently left out some other costs (eg. reputational loss to an insurance firm due to unforeseen insurance payout costs in case of death that we did not predict).

The fourth problem that we could identify has to do with identifying a good representative of each cluster using Mahalanobis distance. Ultimately, the closest point to the centroid may not necessarily be a good representation of the cluster. In an extreme scenario, we could imagine a case where one distinct point is marginally closer to the centroid than every other point, but every other point in the cluster is identical in every regard. In this situation, that one distinct point, which was chosen as the cluster representative, would not have been a good representation of the cluster.

# 7. Conclusion

Moving forward, further analysis can be done beyond the scope covered in the project to uncover greater insights. Given time, an additional step we would have taken would be to subset our Accidents and Vehicles datasets even further, as in the case of the Casualties dataset (where pedestrian and car casualties were extracted from the full dataset). One such way to do this for the Vehicles dataset would be to extract *Cars/Taxis* out separately for analysis as this class of *Vehicle Type* dominates the rest in number (75.7% of observations are *Cars/Taxis*).

In addition to this, we have come to realise that the variables used in our data analysis have proven insufficient in explaining most of the variation in severity rates across accidents and casualties. The low predictive power of our models does suggest the omission of important variables that have strong explanatory power. Variables such as the the speed of the vehicle when it crashed, the presence of construction going on on the road at the time of the accident, and the presence of speed cameras in the proximity are some such examples.

While practical consideration may limit our ability to gather these information (eg. drivers who crash their car are unlikely to give accurate accounts of the speed they were driving at right before the crash), we do suggest proxies for these measures. In the example given, we could use data from speed cameras in the vicinity of the accident to determine the car's travelling speed leading up to the accident.

However, even with these measurable data in hand, it is very likely that there are other key contributors to an accident/casualty's severity that cannot be quantified. Put simply, these determinants can range anywhere from human error (i.e. how the driver reacted) to luck (i.e. the severity of an accident/casualty may be as good as random). Although our report does not claim to hold the answer to the question: "*what is the cause of severe accidents and casualties?",* we do believe that we are one step closer to solving this problem.

# 8. Bibliography

*Anon, Metagenomics. Statistics. Dinsdale et al. Supplemental Material. Available at: https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html [Accessed February 14, 2019].*

*Peter J.Rousseeuw, 2002. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. NeuroImage. Available at: https://www.sciencedirect.com/science/article/pii/0377042787901257 [Accessed February 13, 2019].*

*Terry Therneau and Beth Atkinson (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart*

*Uc-r.github.io. (2019). K-means Cluster Analysis · UC Business Analytics R Programming Guide. [online] Available at: https://uc-r.github.io/kmeans_clustering#gap [Accessed 12 Feb. 2019].*

*UK's Department for Transport. (2011). STATS20: Instructions for the Completion of Road Accidents Reports from Non-CRASH Sources.*

*UK's Department for Transport. (2017). Reported road casualties in Great Britain: 2017 annual report.*

*UK's Department for Transport, (2018). Accident and casualty costs (RAS60). GOV.UK. Available at: https://www.gov.uk/government/statistical-data-sets/ras60-average-value-of-preventing-road-accidents [Accessed February 13, 2019].*

*World Health Organisation. (2010). Decade of Action for Road Safety 2011-2020. Retrieved from: https://www.who.int/violence_injury_prevention/publications/road_traffic/saving_millions_lives_en.pdf?ua=1*

*Yao, 2018, Chapter 3: Classification, lecture notes, ST309: Elementary Data Analytics, London School of Economics & Political Science, delivered October 2018.*

# 9. Appendix

## Appendix A: General

**A1: Full List of Variable Names and Description**

| Variable Name[18] | Description |
|---|---|
| Accident Severity<br>(outcome variable) | Binary variable with 1 representing a severe accident and 0 representing a non-severe accident<br>• **Severe** refers to death of the casualty or any other serious injury such as a broken neck, internal injuries, concussions, or burns, amongst others<br>• **Non-Severe** refers to any slight injury such as shallow cuts, bruising, or slight shock |
| Accident Index | The identification number of the accident with the first four digits representing the year in which the accident took place |
| Vehicle Type | The type of vehicle involved in the accident.<br>• **Bicycles** represent pedal cycles and mobility scooters;<br>• **Motorcycles** represent motorcycles of all engine capacities;<br>• **Cars/Taxis** represent passenger cars, taxis, and private hire cars;<br>• **Heavy Vehicles** - minibuses, coaches, agriculture vehicles, vans, and other goods vehicles. |
| Skidding and Overturning | • **Skidding** describes a vehicle sliding as a result of stopping/turning too quickly<br>• **Jackknifing** occurs when a vehicle bends into a V-shape following a skidding motion<br>• **Overturning** occurs when a vehicle was at any time on its roof, side, front, or rear |
| Hit Object in Carriageway | Describes the first object hit when the vehicle was travelling on a carriageway.<br>• **None** occurs when no object was hit by the vehicle<br>• **Other Vehicles** refers to open door of vehicles, a parked vehicle, or a previous road accident<br>• **Animals** refer to any animal except any animal pulling a vehicle<br>• **Physical Infrastructure** refers to roadworks, bridges, bollards, the central island of a roundabout, and kerbs |
| Vehicle Leaving Carriageway | Binary variable describing whether or not the vehicle involved in the accident had left the carriageway at any point in its journey |
| Hit Object off Carriageway | First object hit off carriageway<br>• **None** occurs when no object was hit by the vehicle |

[18] The variable names have been colour coded to represent the dataset they pertain to. Red is for Vehicles, Green is for Casualties, and Black is for Accidents.

| | |
|---|---|
| | • **Permanent Objects** include bus stops, crash barriers, and other permanent objects<br>• **Poles** refer to telegraph poles/electricity poles, lamp posts, and traffic signals<br>• **Trees** as described<br>• **Ditch/Water** refer to any submergence of the vehicle under water or its entry into a ditch |
| First Point of Impact | • **No Impact** occurs when the vehicle stops suddenly to avoid another vehicle/pedestrian/object in road, but there is no impact<br>• **Front** refers to an impact on the front of the vehicle<br>• **Back** refers to an impact on the rear end<br>• **Offside** is an impact on the side of a vehicle furthest from the kerb (right in the UK)<br>• **Nearside** is an impact on the side of a vehicle nearest to the kerb (left in the UK) |
| Was Vehicle Left-Hand Drive | • **Yes -** if vehicle is left-hand drive<br>• **No** if vehicle is right-hand drive |
| Sex of Driver | • **Male**<br>• **Female** |
| Age of Driver | Describes the age of the driver |
| Age of Vehicle | Describes the age of the vehicle |
| Accident Index | The identification number of the accident with the first four digits representing the year in which the accident took place |
| Vehicle Reference | Index of vehicle involved in the accident |
| Casualty Reference | Index of casualty involved in the accident |
| Casualty Class | • **Driver/Rider** refers to the driver or rider of the vehicle<br>• **Passenger** refers to vehicle or pillion passenger<br>• **Pedestrian** |
| Sex of Casualty | • **Male**<br>• **Female** |
| Age of Casualty | Age of casualty involved in the accident |
| Age Band of Casualty Involved in the Accident | Age bands are: 0-5, 6-10, 11-15, 16-20, 21-25, 26-35, 36-45, 46-55, 56-65, 66-75 and Over 75 |

| | |
|---|---|
| Casualty Severity<br><br>(outcome variable) | • **Severe** refers to death of the casualty or any other serious injury such as a broken neck, internal injuries, concussions, or burns, amongst others<br><br>• **Non-Severe** refers to any slight injury such as shallow cuts, bruising, or slight shock |
| Pedestrian Location | Describes the location of the casualty, given that they were classified as pedestrian under "Casualty Class":<br><br>• **Not Pedestrian**<br><br>• **On/Near Pedestrian Crossing** refers to pedestrians who were crossing on a pedestrian crossing facility or zig-zag approach/exit lines<br><br>• **Crossing Elsewhere** refers to pedestrians who were crossing elsewhere within 50 metres of a pedestrian crossing<br><br>• **Carriageway** refers to pedestrians who were either stationary on a carriageway or crossing across a carriageway<br><br>• **Pavement/Sidewalk** refers to pedestrians who were on a footway, verge, refuge or central island |
| Pedestrian Movement | Describes the movement of the casualty, given that they were classified as pedestrian under "Casualty Class":<br><br>• **Not Pedestrian**<br><br>• **Crossing the Road** refers to pedestrians who were crossing the road at the time of accident<br><br>• **Stationary in Carriageway** refers to pedestrians who were stationary on a carriageway at the time of accident<br><br>• **Walking Along in Carriageway** refers to pedestrians who were walking along in a carriageway at the time of accident |
| Car Passenger | Describes the seat that the casualty (excluding driver) occupied immediately prior to the accident. This is only intended to show whether car and taxi/private hire passenger casualties were in the front or rear seat:<br><br>• **Not Car Passenger:** this term also includes drivers and pedestrians<br><br>• **Front Seat Passenger** refers to passengers who were seated in the front<br><br>• **Rear Seat Passenger** refers to passengers who were not seated in the front, including middle row seats |
| Bus or Coach Passenger | Describes the seat and movement of the casualty (excluding driver) immediately prior to the accident. "Bus or Coach" includes buses, coaches and minibuses equipped to carry 17 or more seated passengers:<br><br>• **Not Car Passenger**: this term also includes drivers, pedestrians, and passengers in other types of vehicles<br><br>• **Boarding or Alighting** refers to passengers who were boarding or alighting the bus or coach vehicle at the time of accident<br><br>• **Standing Passenger** refers to passengers who were standing on the bus at the time of accident<br><br>• **Sitting Passenger** refers to passengers who were sitting on the bus at the time of accident |
| Casualty Type | Describes more specific information about the type of casualty::<br><br>• **Pedestrian**<br><br>• **Cyclist** |

| | |
|---|---|
| | • **Motorcycle or Scooter Rider/Passenger**: includes riders and passengers of all motorcycles, electric motorcycles, scooters and mobility scooters<br>• **Taxi/Private Hire Car Occupant**<br>• **Car Occupant**<br>• **Bus or Coach Occupant**: includes all buses, coaches and minibuses<br>• **Horse Rider**<br>• **Agricultural Vehicle Occupant**<br>• **Good Vehicle Occupant** |
| Number of Vehicles | Number of vehicles involved in accident |
| Number of Casualties | Number of casualties involved in accident |
| Day of Week | Day of week which accident occurred on |
| Road Type | Describes the type of road on which the accident occurred:<br>• **Roundabout** - includes both large and mini-roundabouts<br>• **One-way street**<br>• **Dual Carriageway** - road where the opposing carriageways are physically separated<br>• **Single Carriageway** - roads separated by only markings or none at all are considered as single<br>• **Slip Road** - roads that are that help to direct traffic from one road to another |
| Speed Limit | Describes the speed limit of the road the accident occurred on (miles per hour) |
| Junction Detail | • **Junction** - where two or more roads meet<br>• **Roundabout**<br>• **None** |
| Light Conditions | • **Darkness** - the half hour following sunset and preceding sunrise<br>• **Daylight** - all other times |
| Weather Conditions | Refers to conditions at location and time of accident<br>• **Fine** - no condition that makes driving dangerous<br>• **Rain** - drizzle, hail and sleet that has not built deposit<br>• **Snow** - sleet that builds deposit<br>• **Other** - Fog or mist or other conditions not covered |

| | |
|---|---|
| Road Surface Conditions | Refers to road surface conditions at location and time of accident<br>● **Dry**<br>● **Wet**<br>● **Snow** |
| Special Conditions at Site | Refers to if any special conditions at site were present<br>● **Present**<br>● **None** |
| Carriageway Hazards | Refers to if any carriageway hazards were present<br>● **Present**<br>● **None** |
| Urban or Rural Area | Refers to whether accident occurred in an **Urban** or **Rural** area |

**A2: Calculation of False Positive & False Negative Benefit**

Based on RAS60003, the total human costs (HC) of fatal, severe, and slight casualties are £2,042m, £3,572m, and £1,854m respectively. In 2012 (year in which the above data is from), there were 1,637 fatal accidents, 20,901 severe accidents, and 123,033 slight accidents. Taking the averages, we obtain HC of £2,042m/1637 = £1.25m; £3,572m/20901 = £171k, and £1,854m/123,033 = £15k for a fatal, severe, and slight accident respectively. Furthermore, we know that the proportion of accidents which are fatal in our classification of severe is (1637/ (1637+20901) = 7.3% (in our model, we classed fatal and severe accidents as 'Severe', while slight accidents are classed as 'Non-Severe'). This averages the HC of a 'Severe' accident to £1.25m*(7.3%) + £171k*(92.7%) = £250k/severe accident. Meanwhile, the average HC of a 'Non-Severe' accident is unchanged at £15k.

Based on RAS60003, the total medical cost (MC) of fatal, severe, and slight casualties are £9m, £315m, £165m respectively. Using the same reasoning, the MC averages to £9m/1637 = £5.5k, £315m/20901 = £15,1k, £165m/123,033 = £1.3k for a fatal, severe, and slight accident respectively. This averages the MC of a 'Severe' accident to £5.5k*(7.3%) + £15.1k*(92.7%) = £14.4k/severe accident and £1.3k for a 'Non-Severe'.

False Positive occurs when we classify an accident as severe when it was in fact non-severe. This generates a MC and a PC. From above, we know that the MC is -£13.1k (MC difference between Severe and Non-Severe accidents (£14.4k - £1.3k)), while the PC is cost arising from the phenomenon that hospitals ignore a True Positive the next time a 'Severe' is flagged'. With a population incident rate of 14% (probability of an accident being severe), the cost of this PC is the probability that the next accident is severe multiplied by the lost cost-savings from potentially being able to save that person (assume 100% probability of survival if hospital reacts accordingly). Therefore, PC is 14%*(-£408k) = -£57.12k. Therefore, total benefit of a FP is -£70.2k (MC + PC).

False Negative occurs when we classify an accident as non-severe when it was in fact severe. The benefit (or cost) of a FN is the HC difference between 'Severe' and 'Non-Severe accidents' (£250k - £15k). The reasoning behind this is that stakeholders do not react appropriately in the case of a severe accident, thereby incurring the HC associated with a lost-life. This figure turns out to be -£235k.

## Appendix B: Accidents Dataset

**B1: Data Overview**

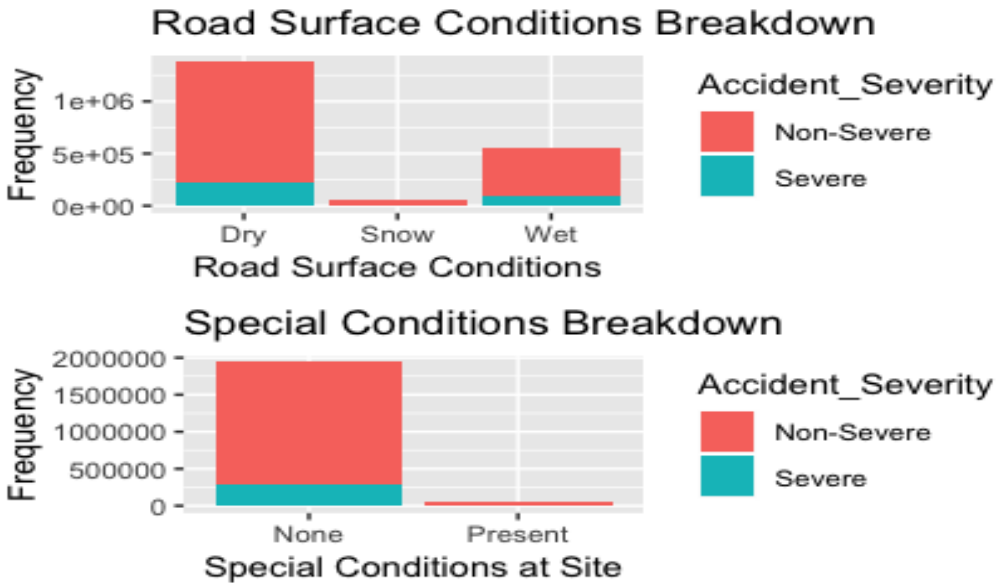



*Figure B1(i): Graph describing Road Surface Conditions and Special Conditions*

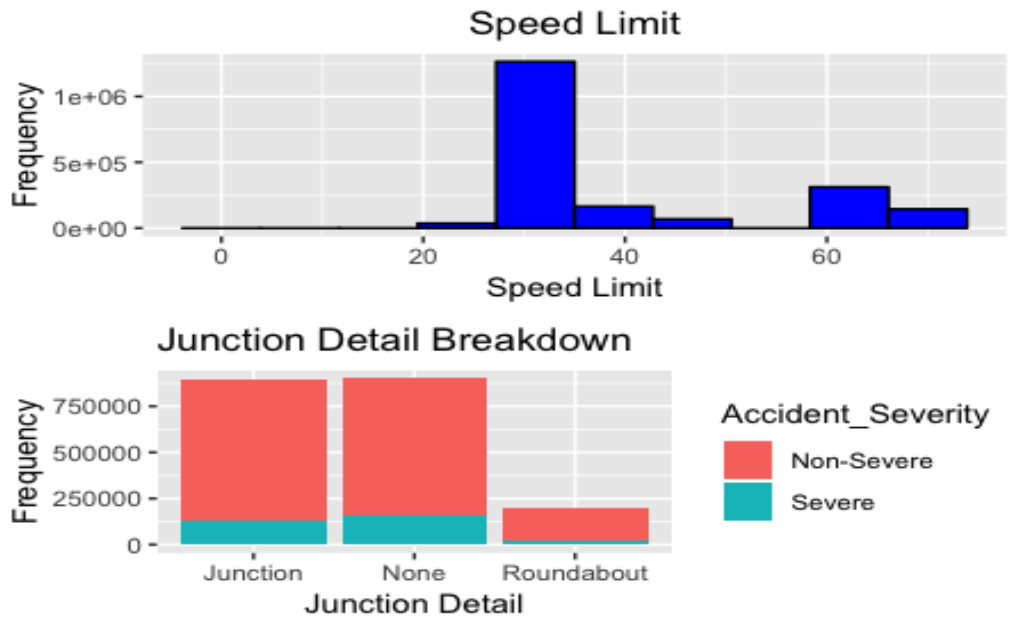


*Figure B2(ii): Graph describing Speed Limit and Junction Detail*

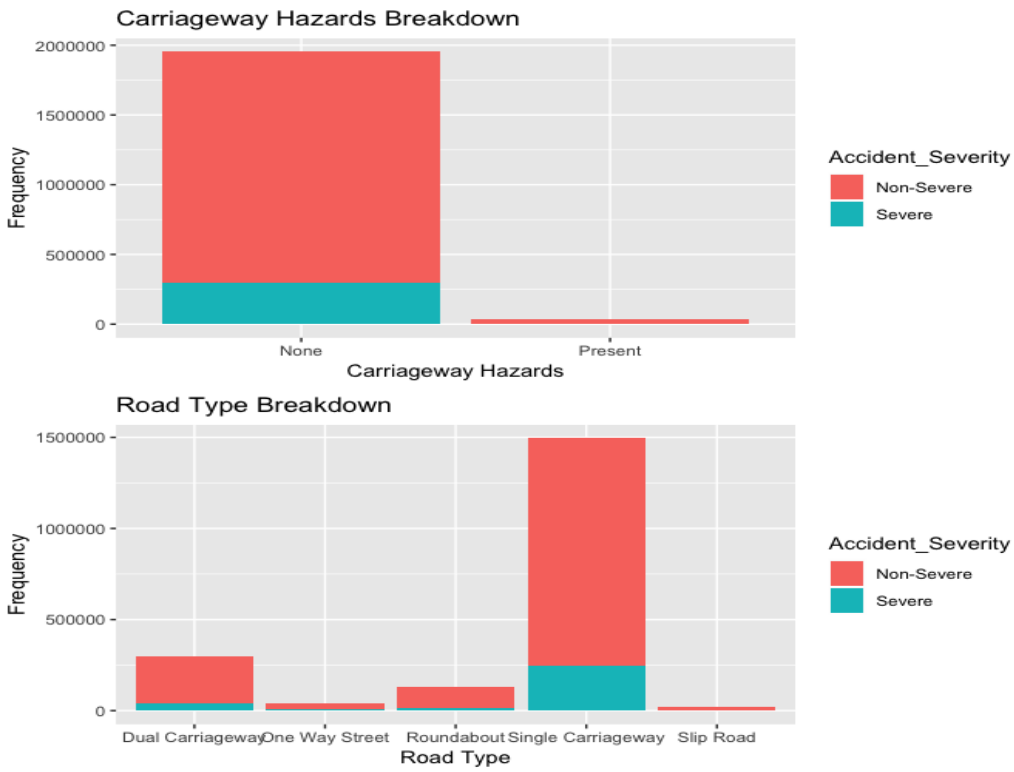



*Figure B2(iii): Graph describing Carriageway Hazards and Road Type*

**B2: Logistic Regression**

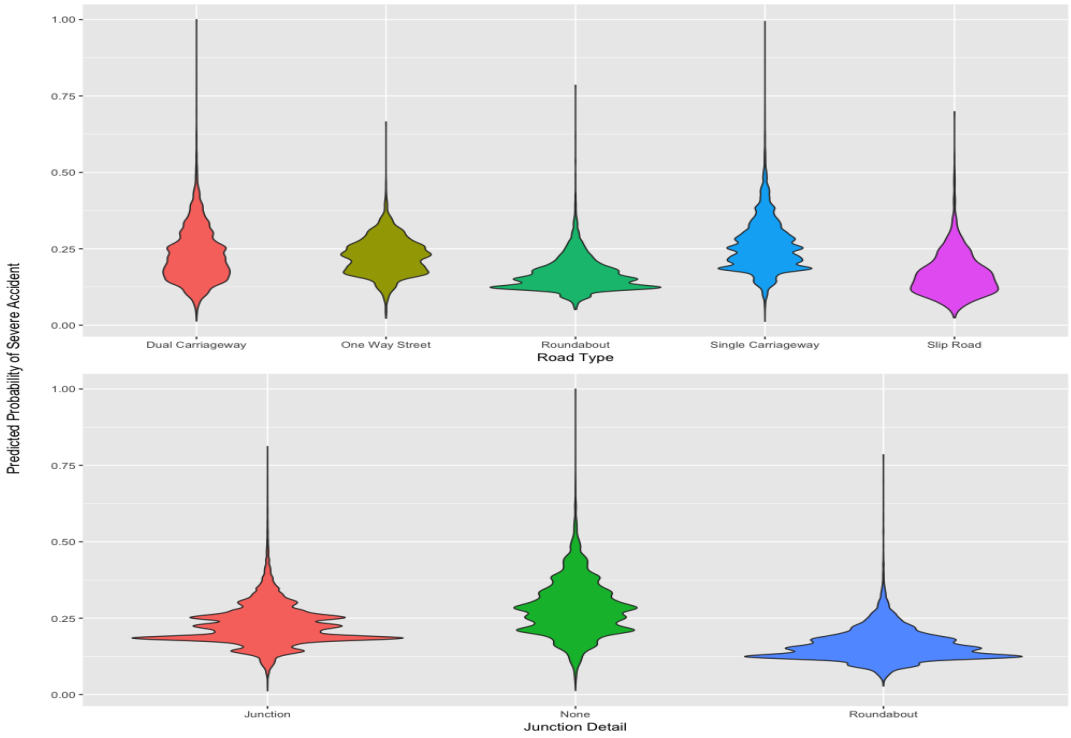



*Figure B2(i): Graphs showing the predicted probabilities of a severe accident against selected explanatory variables*

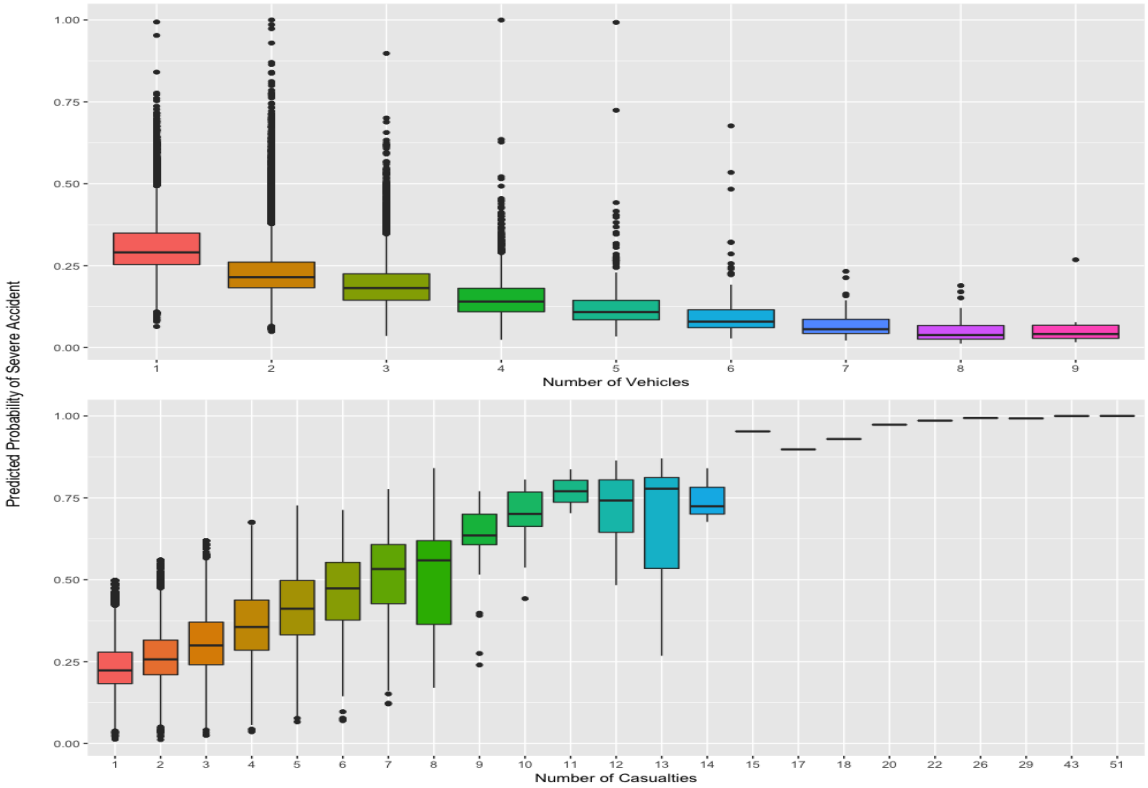



*Figure B2(ii): Graphs showing the predicted probabilities of a severe accident against selected explanatory variables*

**B4: K-fold CV**

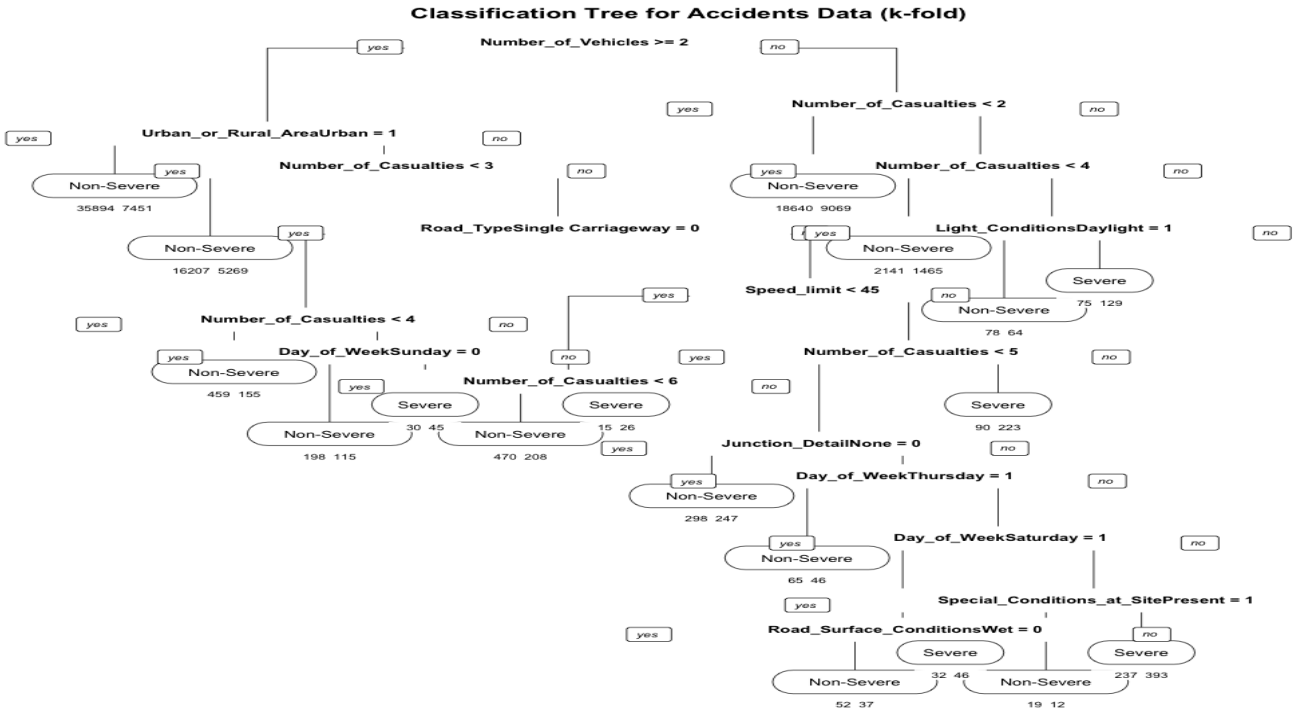



*Figure B4(i): k-fold cross validation* ***(Accidents)***

**B5: Random Forest**

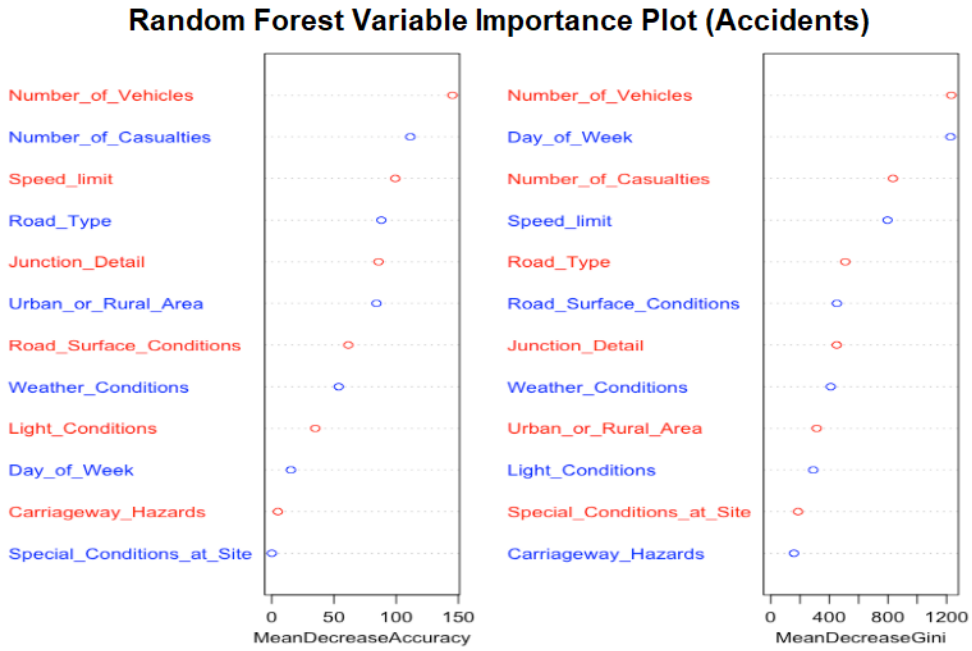



*Figure B5(i): Variable Importance Plot from Random Forest technique for Vehicles*

## Appendix C: Vehicles Dataset

### C1: Data Overview

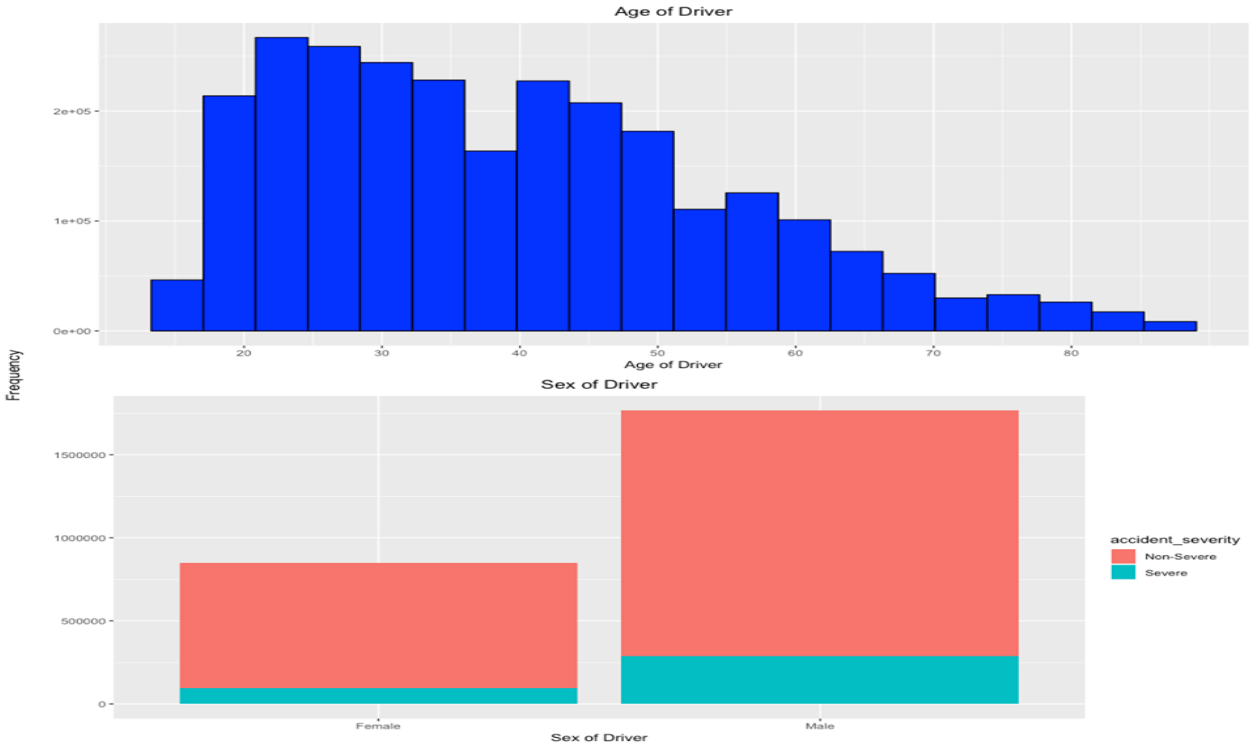



*Figure C1(i): Graph describing Age of Driver and Sex of Driver*

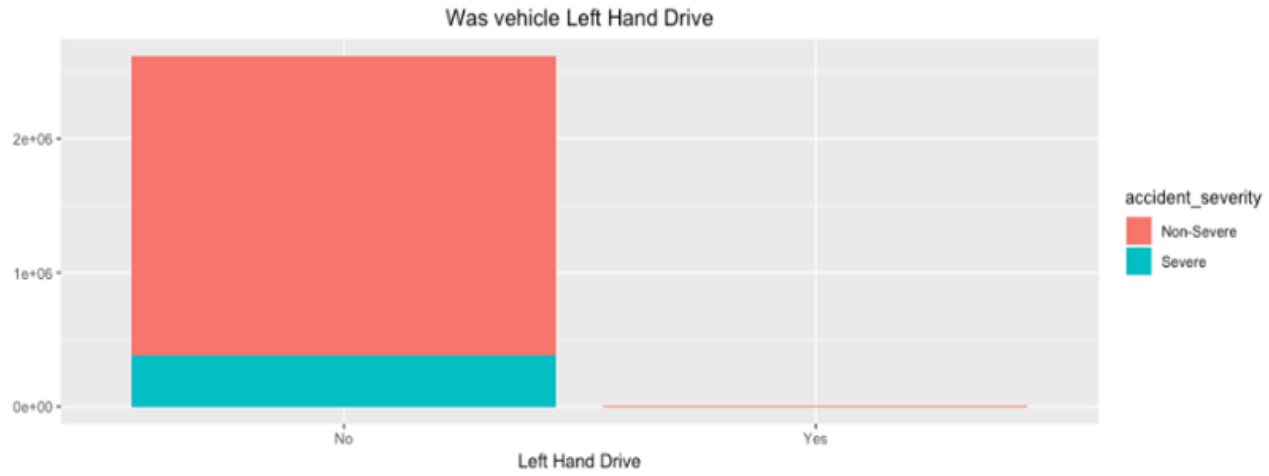



*Figure C1(ii): Graph describing whether Vehicle was Left Hand Drive*

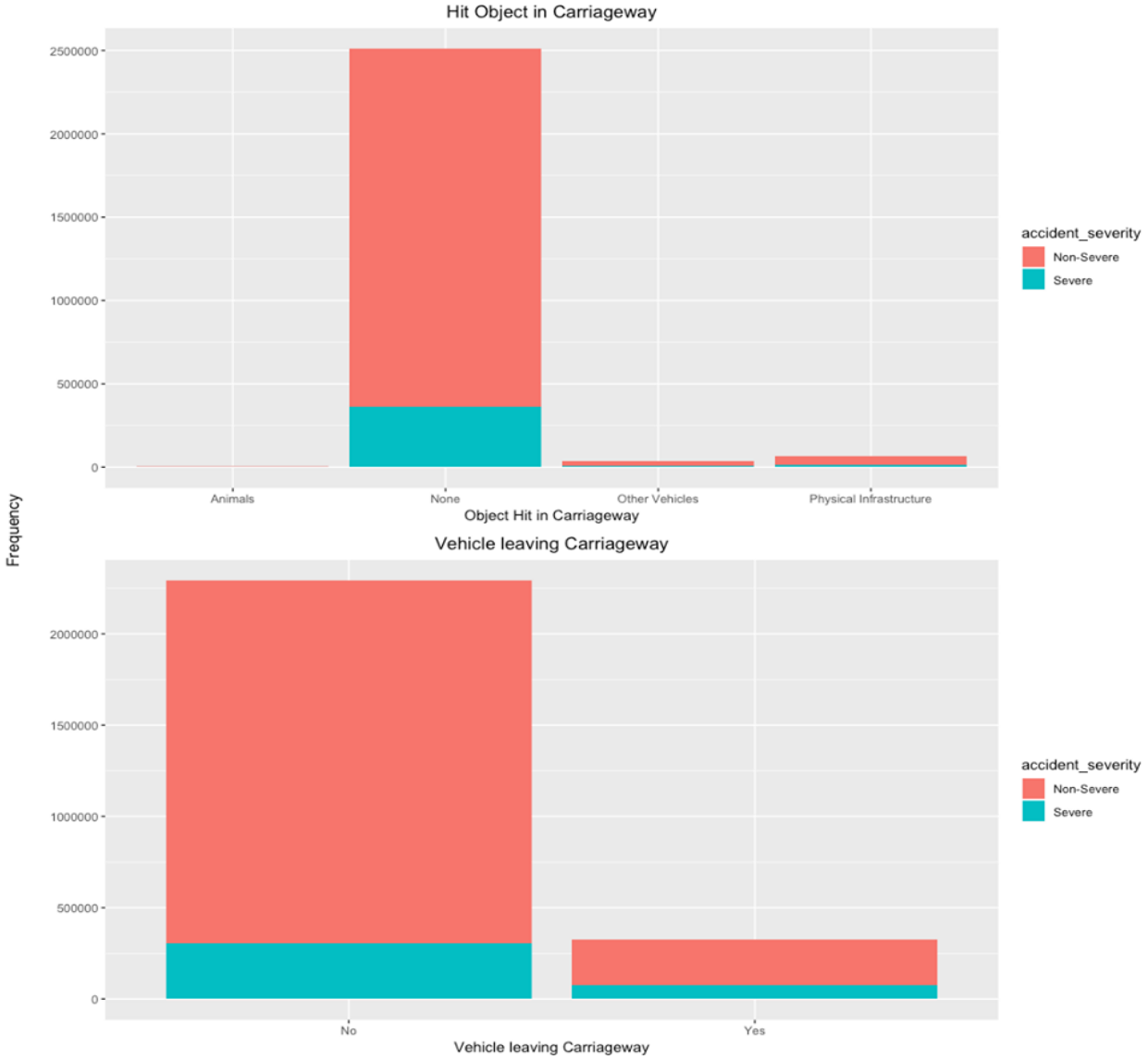



*Figure C1(iii): Graph describing if vehicle Hit Object in Carriageway and if Vehicle was leaving Carriageway*

**C4: K-fold CV**

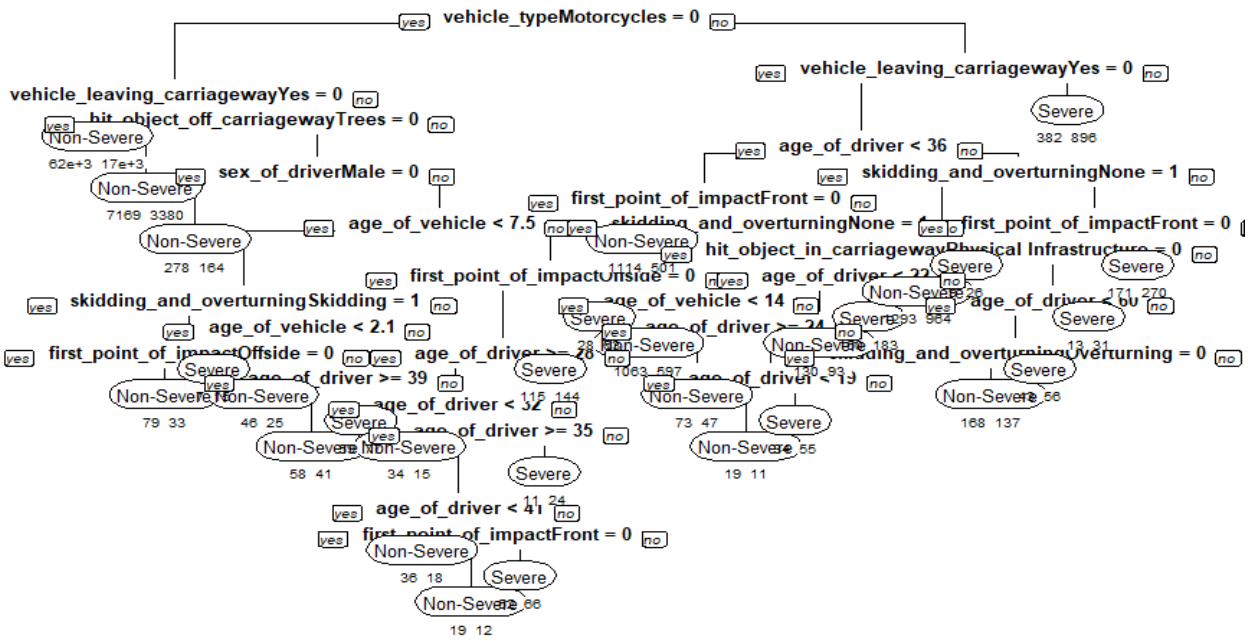



*Figure C4(i): Classification Tree derived from 10-fold CV for Vehicles*

**C6: Clustering**

| ***Cluster Representatives k=2*** | |
|---|---|
| **Cluster 1 (22.8% severity rate)** | **Cluster 2 (13.4% severity rate)** |
| Nine-year-old Car/Taxi | Nine-year-old Car/Taxi |
| Overturned | Overturned |
| Vehicle Did Not Leave Carriageway | Vehicle Left Carriageway |
| No Object Hit In Carriageway | No Object Hit In Carriageway |
| Front-side Impact | Near-Side Impact |
| RHD | RHD |
| Male | Female |
| 18-year-old | 26-year-old |

| **Cluster Representatives k=4** | |
|---|---|
| **Cluster 2 (15.1% severity rate)** | **Cluster 3 (20.9% severity rate)** |
| One-year-old Car/Taxi | Nine-year-old Car/Taxi |
| No overturned/skidded/jackknifed | No overturned/skidded/jackknifed |
| Vehicle Did Not Leave Carriageway | Vehicle Left Carriageway |
| No Object hit in Carriageway | Hit Other Vehicles in Carriageway |
| Front-side Impact | Front-Side Impact |
| RHD | RHD |
| Female | Male |
| 26-year-old | 36-year-old |

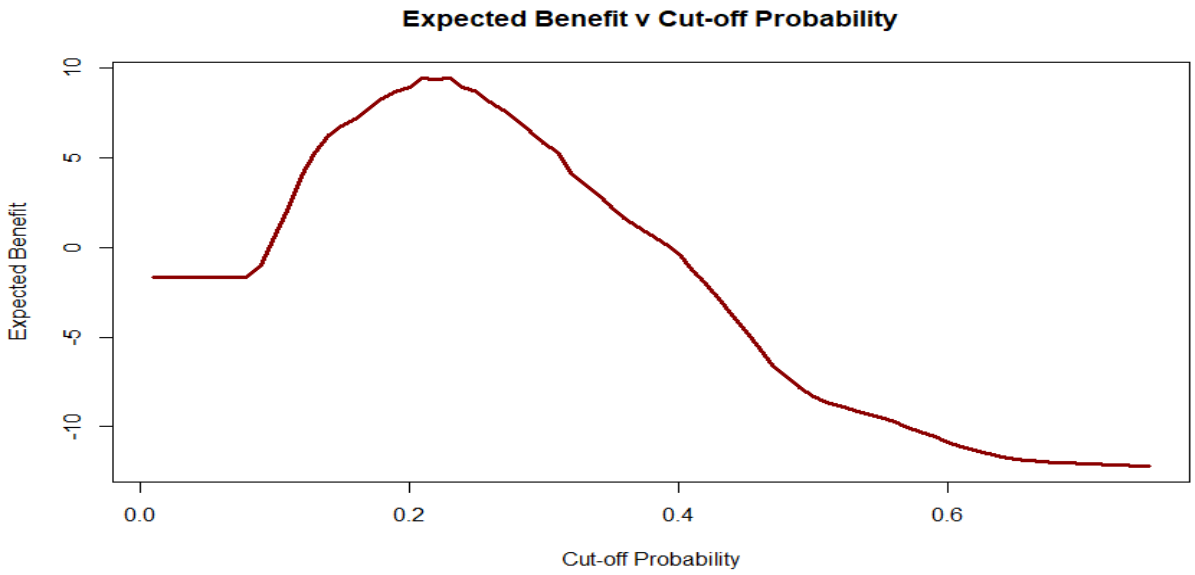



*Figure C7(i): Graph showing Expected Benefit v Cut-off Probability for skimmed logistic regression model*

## Appendix D: Casualties Dataset

**D1: Data Overview**

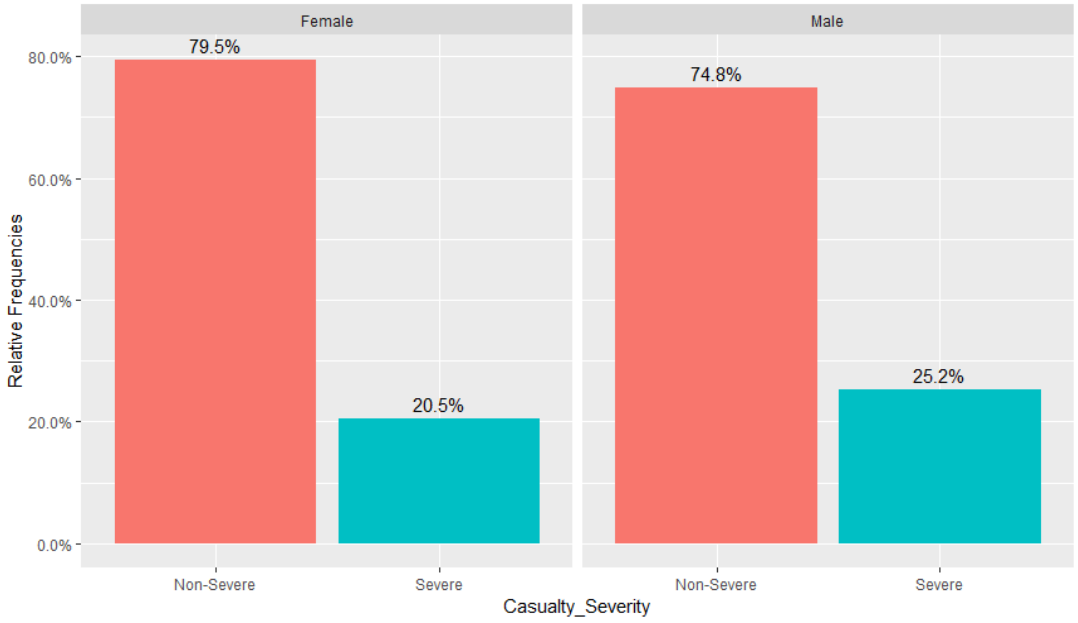



*Figure D1(i): Bar plot showing relative frequencies of Casualty Severity, split by Sex (for only Pedestrian casualties)*

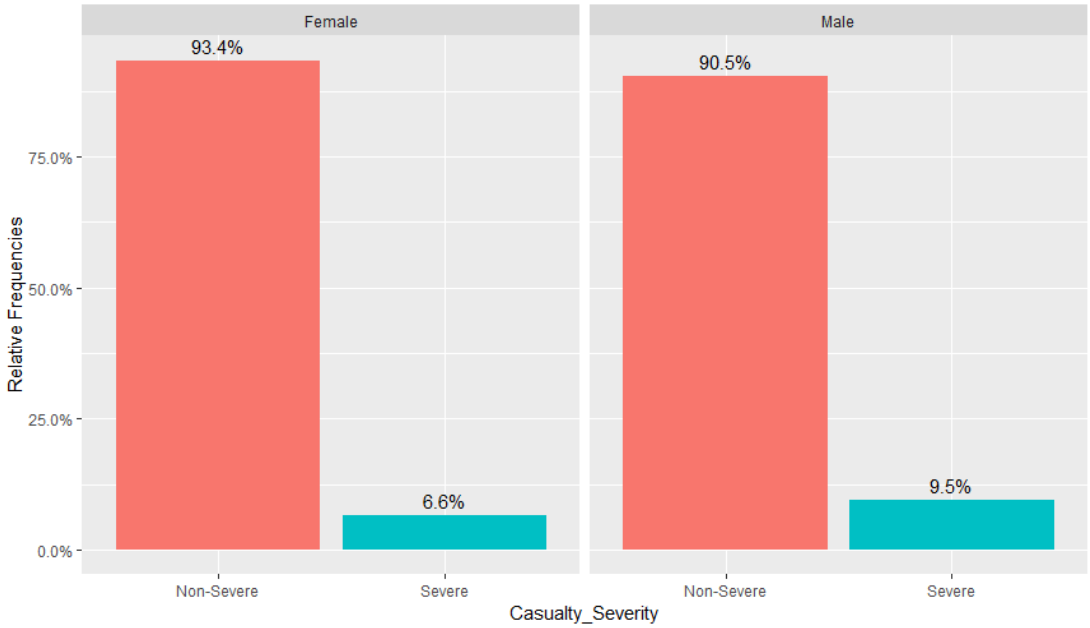



*Figure D1(ii): Bar plot showing relative frequencies of Casualty Severity, split by Sex (for only Car casualties)*

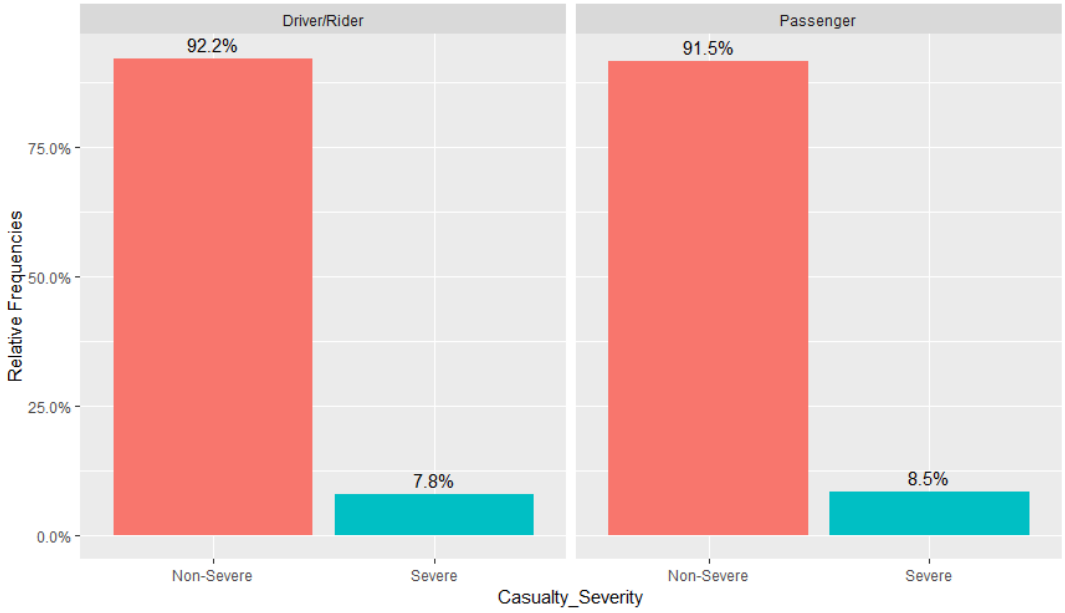



*Figure D1(iii): Bar plot showing relative frequencies of Casualty Severity, split by Casualty Class (for only Car casualties)*

**D2: Logistic Regression**

```
glm(formula = Casualty_Severity ~ Sex_of_Casualty + Age_of_Casualty +
    Pedestrian_Location + Pedestrian_Movement, family = binomial,
    data = cDPedTrain)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9132  -1.3551   0.7909   0.9299   1.3715

Coefficients:
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                    0.2448260  0.0386825   6.329 2.47e-10 ***
Sex_of_CasualtyMale                            0.2608239  0.0338927   7.696 1.41e-14 ***
Age_of_Casualty                                0.0121018  0.0007216  16.771  < 2e-16 ***
Pedestrian_LocationCrossing Elsewhere         -0.0145532  0.0620052  -0.235  0.81443
Pedestrian_LocationOn/Near Pedestrian Crossing -0.0178892  0.0507642  -0.352  0.72454
Pedestrian_LocationOther                       -0.2243351  0.0744634  -3.013  0.00259 **
Pedestrian_LocationPavement/Sidewalk           -0.2988262  0.0689441  -4.334 1.46e-05 ***
Pedestrian_MovementOther                       -0.1586483  0.0557325  -2.847  0.00442 **
Pedestrian_MovementStationary in Carriageway   -0.4883310  0.0728786  -6.701 2.08e-11 ***
Pedestrian_MovementWalking Along in Carriageway -0.2227002  0.0907713  -2.453  0.01415 *
```

*Figure D2(i): Summary of logistic regression for only Pedestrian casualty data*

```
glm(formula = Casualty_Severity ~ Casualty_Class + Sex_of_Casualty +
    Age_of_Casualty + Car_Passenger + Casualty_Type + Age_of_Casualty:Sex_of_Casualty,
    family = binomial, data = cDCarsTrain)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.355  -1.014  -0.814   1.307   1.923

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                              -1.6184813  0.3544990  -4.566 4.98e-06 ***
Casualty_ClassPassenger                   0.3138100  0.3523537   0.891    0.373
Sex_of_CasualtyMale                       0.8870192  0.0406549  21.818  < 2e-16 ***
Age_of_Casualty                           0.0167735  0.0007037  23.837  < 2e-16 ***
Car_PassengerNot Car Passenger            0.1609975  0.3527148   0.456    0.648
Car_PassengerRear Seat Passenger          0.1352434  0.0342679   3.947 7.93e-05 ***
Casualty_TypeTaxi/Private Hire Car Occupant -0.5082932  0.0691960  -7.346 2.05e-13 ***
Sex_of_CasualtyMale:Age_of_Casualty       -0.0110392  0.0009570 -11.535  < 2e-16 ***
```

*Figure D2(ii): Summary of logistic regression for only Car casualty data*

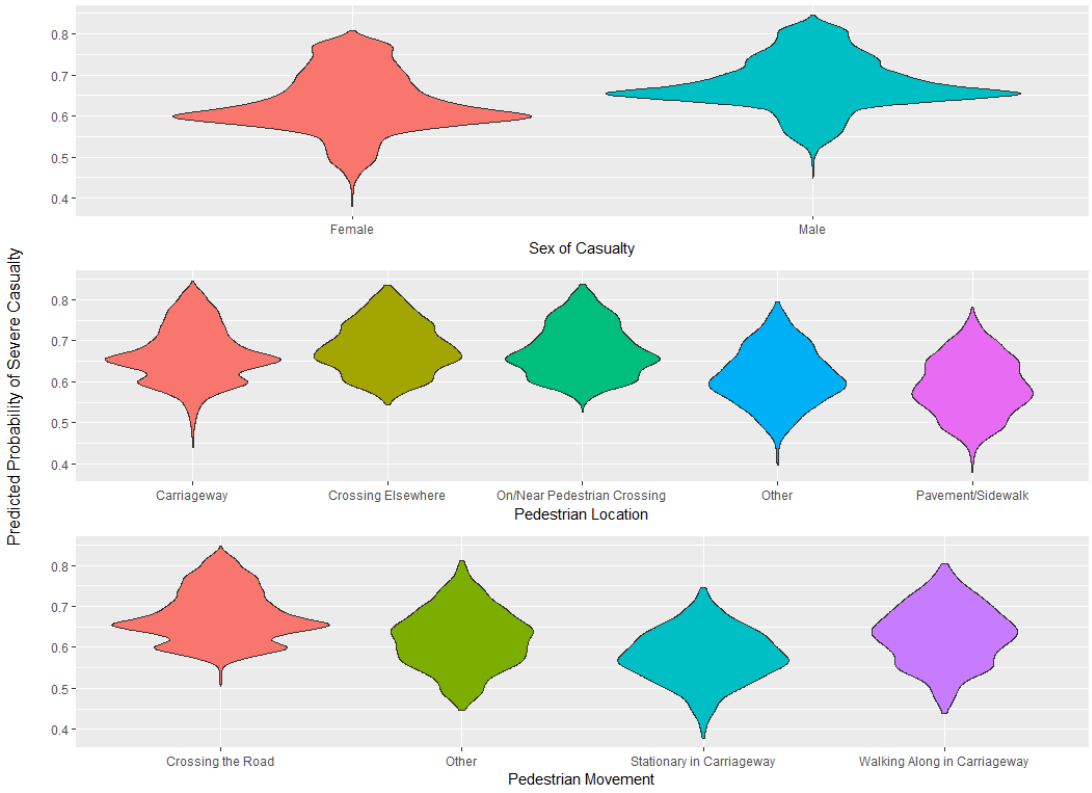



*Figure D2(iii): Charts showing the predicted probabilities of a severe accident against explanatory variables (Pedestrian casualties)*

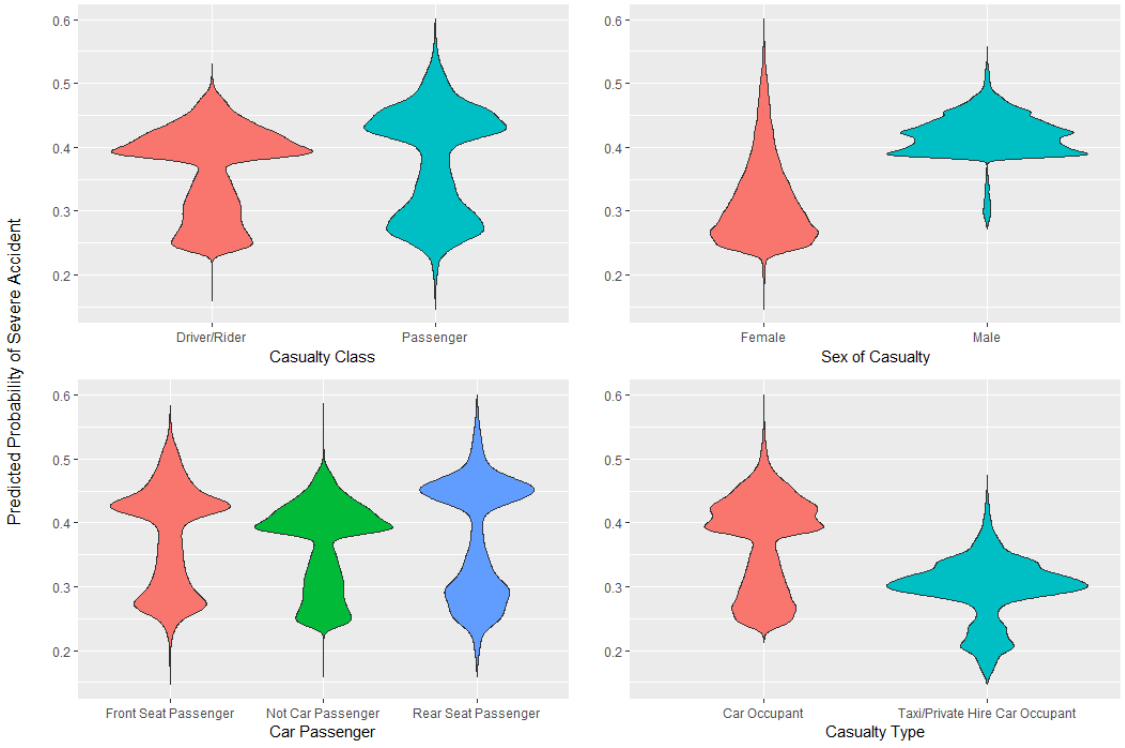



*Figure D2(iv): Charts showing the predicted probabilities of a severe accident against explanatory variables (Car casualties)*

**D3: Classification Tree**

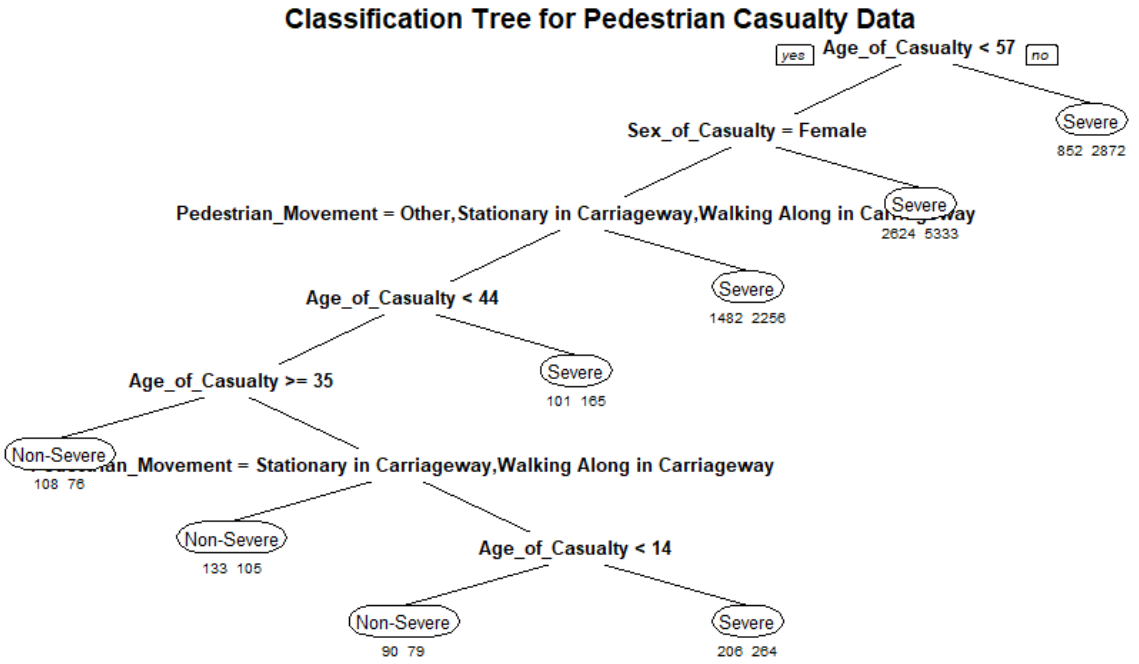



*Figure D3(i): Classification Tree derived from 10-fold CV for Casualties (Pedestrians)*

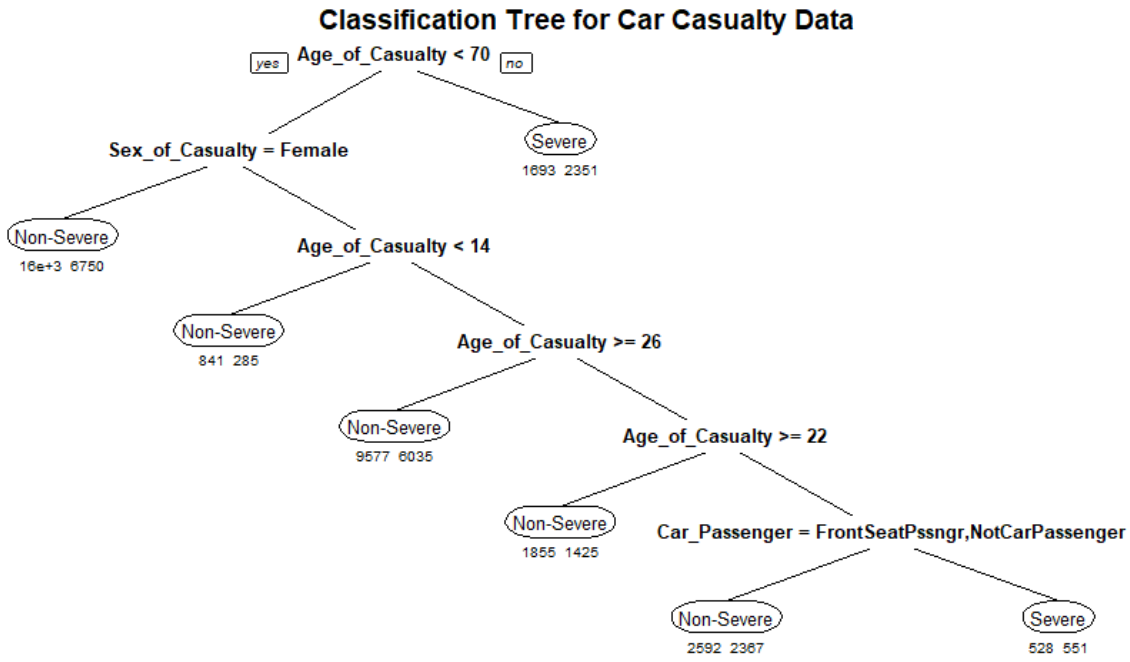



*Figure D3(ii): Classification Tree derived from 10-fold CV for Casualties (Cars)*

**D4: K-fold CV**

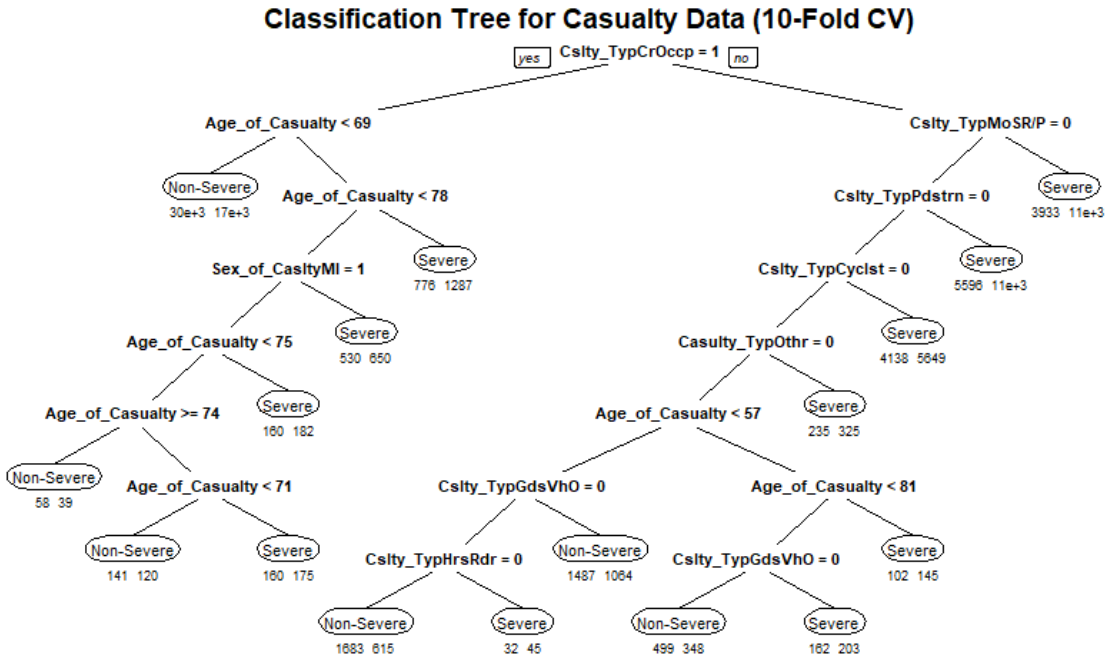



*Figure D4(i): Classification Tree derived from 10-fold CV for Casualties (all)*

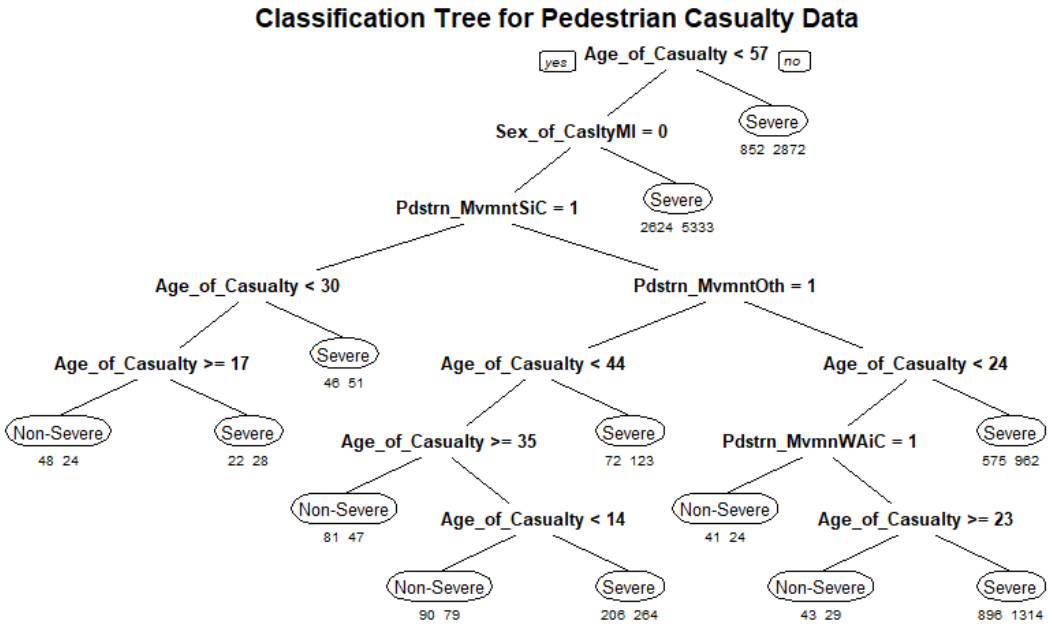



*Figure D4(ii): Classification Tree derived from 10-fold CV for Casualties (Pedestrians)*

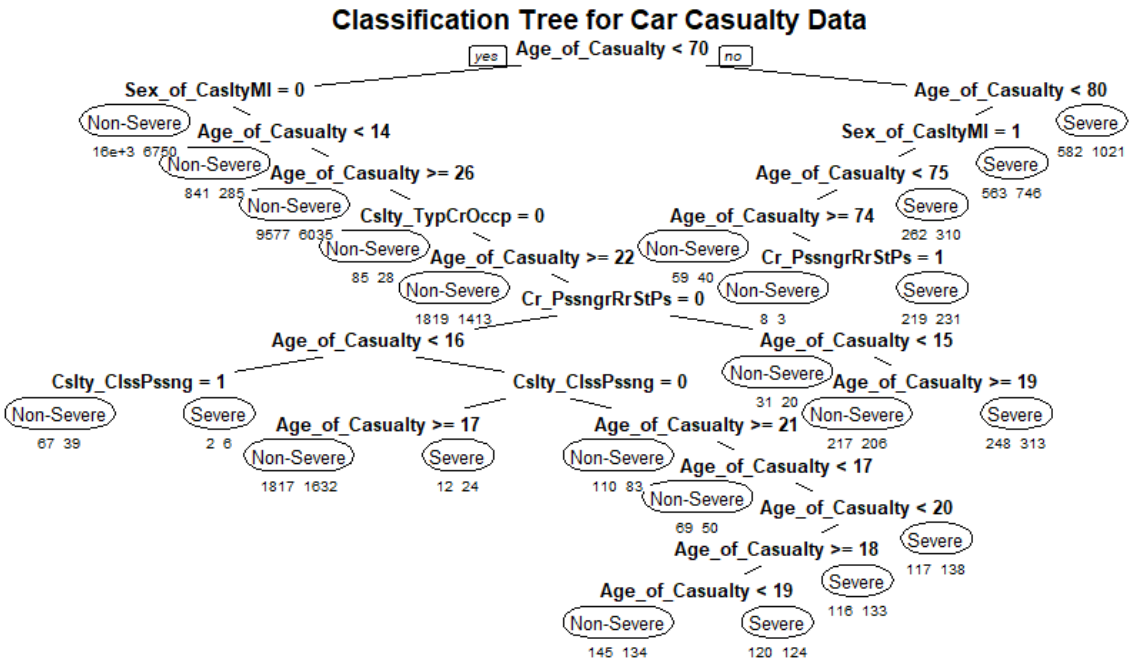



*Figure D4(iii): Classification Tree derived from 10-fold CV for Casualties (Cars)*

**D5: Random Forest**

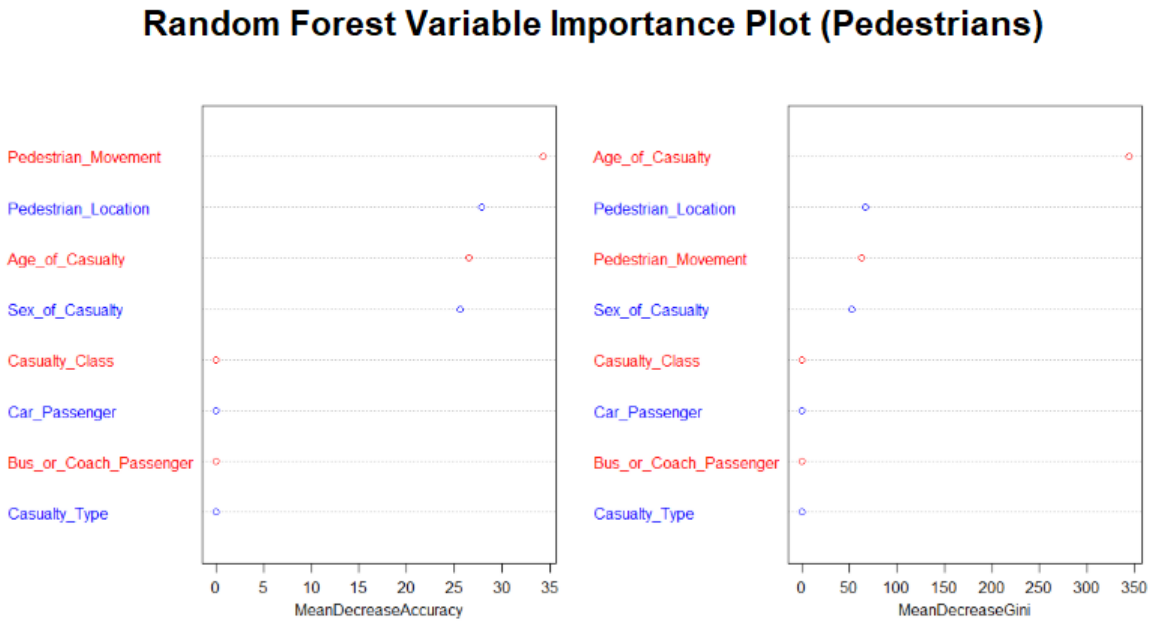



*Figure D5(i): Variable Importance Plot for Pedestrian casualties*

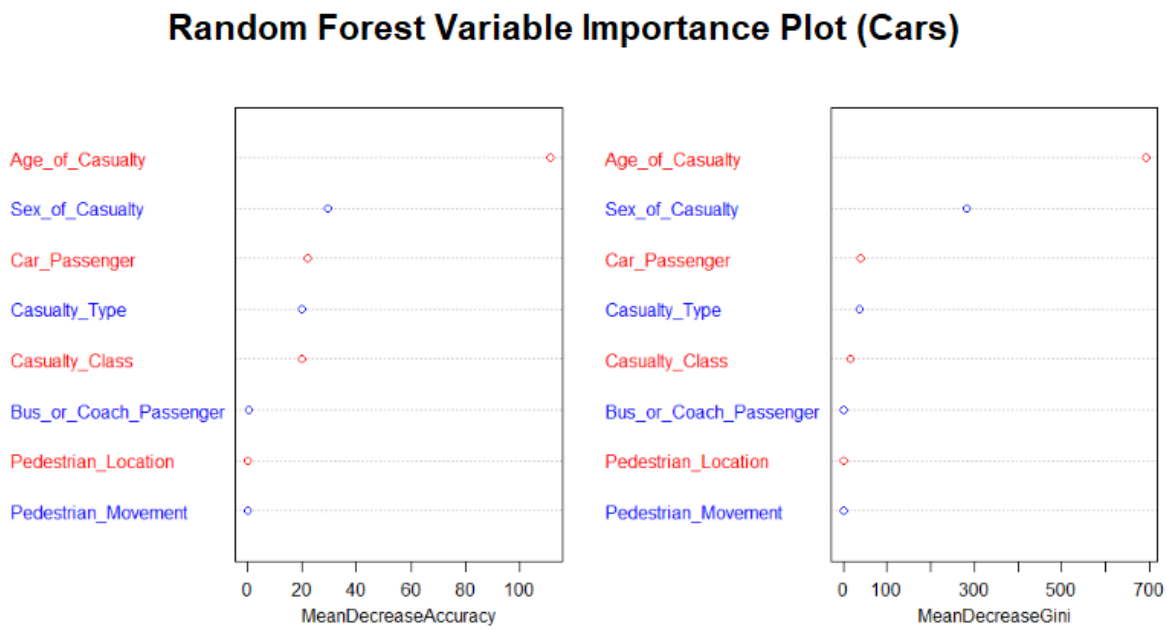



*Figure D5(ii): Variable Importance Plot for Pedestrian casualties*