

UNSW

Accident Data Analysis

Dhwanish Kshatriya

30/07/2023

## Table of Contents

Section 1 – Exploratory Data Analysis (EDA).....	2
1.1 – Driver Age (Driver Characteristics) .....	2
1.2 – Driver Sex (Driver Characteristics).....	2
1.3 – Vehicle Age (Vehicle Characteristics) .....	2
1.4 – Vehicle Type (Vehicle Characteristics).....	2
1.5 – Helmet/ Seatbelt Worn (Accident Conditions).....	3
1.6 – Time of Day (Accident Conditions) .....	3
Section 2 – Statistical Modelling.....	3
2.1 – Impact of SEX .....	3
2.2 – Impact of AGE .....	3
2.3 – Impact of HELMET_BELT_WORN .....	3
2.4 – Impact of VEHICLE_TYPE.....	4
2.5 – Impact of TOTAL_NO_OCCUPANTS .....	4
2.6 – Impact of ACCIDENT_TIME.....	4
2.7 – Impact of ACCIDENT_TYPE.....	4
2.8 – Impact of LIGHT_CONDITION.....	4
2.9 – Impact of ROAD_GEOMETRY .....	5
2.10 – Impact of SPEED_ZONE.....	5
2.11 – Impact of SURFACE_COND .....	5
Section 3 – Predictive Modelling .....	5
3.1 – Background Information.....	5
3.2 – Gradient Boosting Machine (GBM) Model.....	5
Appendix 1 – Figures, Tables, Models & Explanations .....	7
Statistical Learning Model Selection: Logistic Regression.....	9
Statistical Learning Model Creation.....	9
Predictive Analysis Data Cleaning, Balancing & Split .....	12
Predictive Analysis Model Breakdowns.....	13
Appendix 2 – Generative AI Use .....	16

## **Section 1 – Exploratory Data Analysis (EDA)**

The following is an Exploratory Data Analysis (EDA) of the VicRoads Accident Data. The dataset includes data on 200,000 drivers involved in road crashes in Victoria between 2006 and 2020. The dataset tracks 27 variables describing 3 categories; the driver, vehicle, and accident conditions. In this EDA, we aim to gain insights and into 2 “interesting” variables from each category (6 in total) and their impact on accident fatality. It was identified that less than 2% of the accidents reported in VicRoads Accident Data are fatal, meaning the dataset is imbalanced and thus will require balancing when modelling.

### **1.1 – Driver Age (Driver Characteristics)**

In reference to *Figure 1 - Proportion of Fatal Accidents by Age Group*, we can see a clear link between the driver's age group (as defined in figure 1) and the proportion of accidents that were fatal. Individuals under 18 years of age exhibited a far larger proportion of fatal accidents than older individuals. Then, a significantly lower proportion can be observed for individual's aged 18-21. Age groups containing the ages 22-49 showcase even lower proportions of fatal accidents, with the proportions varying relatively less. From here, there is a noticeable increase in the proportion of fatal accidents for age groups containing ages 50-69, and a large spike in the proportion for ages above 70. While no conclusions can be drawn from this data, a clear relationship is shown between age group and proportion of fatal accidents. We can speculate the high fatality rate for ages below 17 may be associated with their inexperience in driving, and the relatively higher fatality rate for older age groups (50 and above) may be linked to the decreased health associated with aging.

### **1.2 – Driver Sex (Driver Characteristics)**

In reference to *Figure 2 - Accidents by Driver Sex*, we can observe that men exhibit a much larger number of accidents, fatal accidents and over double the proportion of fatal accidents than women (2.16% vs 1.04%). Whilst the VicRoads Accident Data also defines sex as either male (M), female (F) and undefined (U), it is difficult to make any inference on the relation between an undefined sex and the proportion of accidents that are data, since it is difficult to interpret what is considered an “undefined” sex in this scenario. Once again, conclusions cannot be made on this data alone, as factors such as average distance driven by drivers of each sex or number of drivers of each sex are not considered.

### **1.3 – Vehicle Age (Vehicle Characteristics)**

Since the data ranges from observations between 2006 and 2020, we found it more informative to analyse vehicle age instead of manufacturing year. This conversion was done within the data using the formula; *Vehicle Age = Manufacturing Year – Accident Year*. In reference to *Figure 3 - Proportion of Fatal Accidents by Vehicle Age*, we obtained a clear, proportional relationship between vehicle age and the proportion of accidents that were fatal. Whilst this showcases that older vehicles have higher proportions of accidents that are fatal, these statistics also raise questions about what aspects about older cars causes an increase in this proportion.

### **1.4 – Vehicle Type (Vehicle Characteristics)**

In reference to *Figure 4 - Proportion of Fatal Accidents by Vehicle Type*, we can see that for most vehicle types (cars, panel vehicles, station wagons and taxis), the proportion of accidents that are fatal is similar and relatively low (below 2%). However, other vehicles exhibit a significantly higher proportion indicating that vehicle type may in fact impact the proportion of accidents that are fatal. It can be observed that the vehicle types with higher proportions are relatively heavier vehicles. Thus, whilst the VicRoads Accident data doesn't provide data on vehicle weights (which could provide us further insights), it still draws a connection between fatality and vehicle type.

### 1.5 – Helmet/ Seatbelt Worn (Accident Conditions)

In reference to *Figure 5 - Proportion of Fatal Accidents by Helmet/Belt Worn*, we can observe a clear relationship between the proportion of accidents that were fatal and whether a helmet/belt was worn. We can see that the proportion of accidents that were fatal is significantly higher when the seatbelt wasn't worn (almost 7%), whereas the proportion is significantly lower when it is worn (less than half).

### 1.6 – Time of Day (Accident Conditions)

From *Figure 6 - Proportion of Fatal Accidents by Hour of Day*, we can see that the proportion of accidents that are fatal begins increasing from 4:00 pm and essentially continues increasing until it peaks at 1:00 am. It then begins to decrease from 1:00 am to 5:00 am, in which the proportion is still relatively high until it eventually plummets at 6:00 am and remains relatively low from 6:00 am to 3 pm. This data showcases that the proportion remains above 4% from 12:00 am to 4:00 am, providing valuable insights into what times of day drivers may be most prone to fatal accidents. As discussed previously, though we cannot draw any conclusions from these statistics, we can speculate that this might have to do with darkness levels outside, temperature and lack of traffic/ authority to enforce road safety laws.

## Section 2 – Statistical Modelling

For details on Model Selection and Creation, refer to Explanation 1 – Statistical Learning Model Selection & Creation.

At a p-value of 0.01, there were 24 predictors deemed significant. These are the predictors with “\*\*\*” or “\*\*\*\*” beside them in *Figure 7 - Logistic Regression Model Summary*. Below, we will analyse how each predictor impacts the odds of an accident being fatal. These are calculated by using the *logit* canonical link, equivalent to a change in odds by a factor of  $e^{x^T \cdot \beta}$ . Summaries for the change in odds for each significant predictor can be found within *Table 1* to *Table 11*. These results are explained below:

### 2.1 – Impact of SEX

The SEX variable is categorical, partaking 3 values, of which only M (male) and F (female) are significant. The odds of an accident being fatal are 1.4947 times higher for males than females (reference category) given all other variables are constant. The data is not statistically significant at a level of 5% for the category of U (undefined sex). This result follows our insights from the EDA, showcasing that men are significantly more likely to be involved in a fatal accident than women.

### 2.2 – Impact of AGE

The AGE variable is numerical. Our model showcases that the odds of an accident being fatal are 1.0108 times higher for each unit increase in age. This result contradicts our insights from the EDA, as the relationship identified was more complex than a linear relationship. This is a case in which the assumption falls through, as our logistic regression fails to model the significantly higher proportion of fatal accidents for ages 0-17. Regardless, the model indicates that the fatality rate (slowly) increases with age, meaning older drivers are significantly more prone to fatal accidents than younger drivers. As previously discussed, we can speculate that this relationship can be attributed to the decrease in health that often comes with ageing.

### 2.3 – Impact of HELMET\_BELT\_WORN

HELMET\_BELT\_WORN is another categorical variable partaking 3 values, of which 1 predictor is significant. This is SEATBELT\_NOT\_WORN, which increases the odds of an accident being fatal by a factor of 3.6697. This is also in accordance with our investigation in the EDA, which showed that individuals are significantly more prone to a fatal accident if they are not wearing a seatbelt as

compared to if they are. This data is heavily in favour with Victorian laws regarding seatbelts being a requirement in a moving vehicle.

#### **2.4 – Impact of VEHICLE\_TYPE**

VEHICLE\_TYPE is a categorical variable with various predictors, of which 4 are significant. The most influential of these is HEAVY\_VEHICLE, with a factor of increase in odds of 3.2457. This is followed by PRIME\_MOVER with a factor of 2.544, then OTHER with a factor of 1.4865 and finally a UTILITY with a factor of 1.239. These values are in reference to the reference predictor of a CAR. As discussed in the EDA, we can observe that the significant values are all heavier vehicles, whilst the insignificant predictors are all lighter vehicles. Though we cannot necessarily draw conclusions about the impact of a weight of a vehicle on the probability of an accident being fatal, our logistic regression showcases that specific vehicle types have a larger impact on fatality than others.

#### **2.5 – Impact of TOTAL\_NO\_OCCUPANTS**

TOTAL\_NO\_OCCUPANTS is a numerical variable. Our logistic regression implies that the odds of an accident being fatal increases by a factor of 1.1225 for each unit increase in TOTAL\_NO\_OCCUPANTS given all other predictors kept constant. Thus, our logistic regression suggests that individuals are significantly more prone to being involved in an accident with a fatality when they have more passengers. This may be attributed towards the fact that as number of passengers increases beyond a certain point, the driver has a various number of potential distractions which may impact their focus when driving.

#### **2.6 – Impact of ACCIDENT\_TIME**

ACCIDENT\_TIME is another numerical variable. Our regression model outputs a decrease in odds of a fatal accident by a factor of 0.9838 for each increase in unit ACCIDENT\_TIME given all other predictors are kept constant. This result also contradicts the insights provided by our EDA, which showcased a non-linear relationship between proportion of accidents that were fatal and ACCIDENT\_TIME. Once again, the assumption of linearity breaks, with the model failing to capture the increasing proportion of fatal accidents between 4:00 pm and 11:00 pm (*Figure 6*). However, the model does accurately capture the decreasing proportion of fatal accidents from 1:00 am to 6:00 am (*Figure 6*). Regardless, the model fails to accurately depict the correct relationship between the proportion of fatal accidents and ACCIDENT\_TIME.

#### **2.7 – Impact of ACCIDENT\_TYPE**

ACCIDENT\_TYPE is a categorical variable with 6 significant predictors. Our regression model showcases a decrease in the odds of a fatal accident for most predictors when compared to the reference predictor – COLLISION\_WITH\_FIXED\_OBJECT. The exact values for the change in odds can be found in *Error! Reference source not found.*. However, there is a particularly large increase in the odds of a fatality (by a factor of 4.042) for the predictor PEDESTRIAN\_STRUCK. Whilst this result is seemingly obvious, we can conclude the fact that collisions with fixed objects and pedestrians significantly increase the fatality rate in an accident. We can speculate that these can be attributed towards both situations applying a significant force on an individual, which could be the cause of death in these scenarios. On the contrary, it is interesting that the change in odds decreases by a factor of 0.6514 for vehicle-on-vehicle collisions (when compared to the reference predictor). This may be attributed towards the safety mechanisms within vehicles, such as crumple zones and airbags.

#### **2.8 – Impact of LIGHT\_CONDITION**

LIGHT\_CONDITION is another categorical variable with a reference predictor of DARK\_NO\_STREET\_LIGHTS and 3 significant predictors, STREET\_LIGHTS\_ON, DAY, and DUSK/DAWN. We obtain a decrease in odds by a factor of 0.8245, 0.5065 and 0.4535 for each of these predictors respectively. Using these, we can draw insights from our regression model regarding

the lower risk of fatal accidents during accident conditions with more light. These results are relatively intuitive and may be simply attributed towards increased driver visibility when there is lighter, and decreased driver focus typically in darker, late-night situations.

## 2.9 – Impact of ROAD\_GEOMETRY

ROAD\_GEOMETRY is a categorical variable with 2 significant predictors. These are NOT\_AT\_INTERSECTION and T\_INTERSECTION, which have an increase in odds of a fatal accident by a factor of 1.4333 and 1.1928 respectively in comparison to their reference predictor CROSS\_INTERSECTION. It is difficult to draw many relationships from this data, besides the fact that you are less likely to have a fatal accident at an intersection as compared to an alternate road geometry. We can speculate that this is due to the design of intersections, which focuses on accident prevention and may also focus on accident severity minimisation.

## 2.10 – Impact of SPEED\_ZONE

SPEED\_ZONE is our last numerical variable. Our regression model showcases an increase in odds of a fatal accident by a factor of 1.0428 for each unit increase in speed given all other predictors are kept constant. This result is relatively intuitive and by the consensus, that faster speed zones lead to more severe accidents. Our model aligns with the idea that individuals can drive slower (in slower speed zones) whenever possible to minimise the probability of an accident being fatal.

## 2.11 – Impact of SURFACE\_COND

Our final variable of SURFACE\_COND showcases an interesting result. It has a reference predictor of a DRY surface condition, with a decrease in odds of an accident being fatal by a factor of 0.7211 for WET surface conditions, and 0.2676 for OTHER surface conditions. This result is particularly interesting as many individuals are under the impression that driving on wet surfaces results in more severe accidents. However, this result implies otherwise. Though we cannot draw any conclusions, we can speculate that this result may be attributed towards the increased driver focus and safer driving behaviours applied during riskier, wet surfaces.

# Section 3 – Predictive Modelling

## 3.1 – Background Information

Information regarding the processes involved in data cleaning, balancing, and splitting can be found in *Explanation 2*. Further, model breakdowns on underperforming models can be found from *Model 1* to *Model 3*.

## 3.2 – Gradient Boosting Machine (GBM) Model

The Gradient Boosting Machine (GBM) is a prediction algorithm which used a group of weaker, less accurate models (decision trees, in this case). The algorithm starts by making a single leaf. From here, it begins a cycle of creating new, larger trees based on the successes/ failures of previous trees. It continues to do this until several iterations are reached or the error is below a certain threshold.

The entire model was input into the `gbm()` function in R. The following model decisions were made:

Variable	Value	Explanation
n.trees	1000	Set to a reasonably large size that wouldn't cause processing issues
interaction.depth	4	Set to allow a reasonable amount of flexibility without risking overfitting the training data.
n.minobsinnode	10	Prevent overfitting on the training data (especially since the data was oversampled)
Shrinkage	0.01	Set to a good balance between computational time and model performance
Verbose	FALSE	We primarily care about model precision

After training, the GBM model was evaluated on a validation set and obtained an AUC of 0.6873709. Though it is not a significant improvement on lasso regression when considering AUC (which only has a 0.007969 difference), GBM is the most accurate model in the end.

This model was then used to make predictions on driver data to select 2500 drivers who were most likely to have a fatal accident. Using this set of predicted drivers, we can draw some observations:

- Despite men accounting for 5865 observations (58.65%), they were accountable for 1535 (61.40%) of the predicted fatalities.
- Despite the VEHICLE\_TYPES considered as significant predictors accounting for 1450 observations (14.50%); they were accountable for 238 (23.80%) of the predicted fatalities.
- Despite individuals ages 0-17 accounting for only 57 (0.57%) of observations, they were accountable for 52 (2.08%) of predicted fatalities.

Such results from our model align with the consensus formed by our EDA and statistical analysis performed in sections 1 and 2. Thus, we can see that drivers with a higher likelihood of fatal accidents hold the characteristics previously discussed.

Appendix 1 – Figures, Tables, Models & Explanations

Figure 1 - Proportion of Fatal Accidents by Age Group

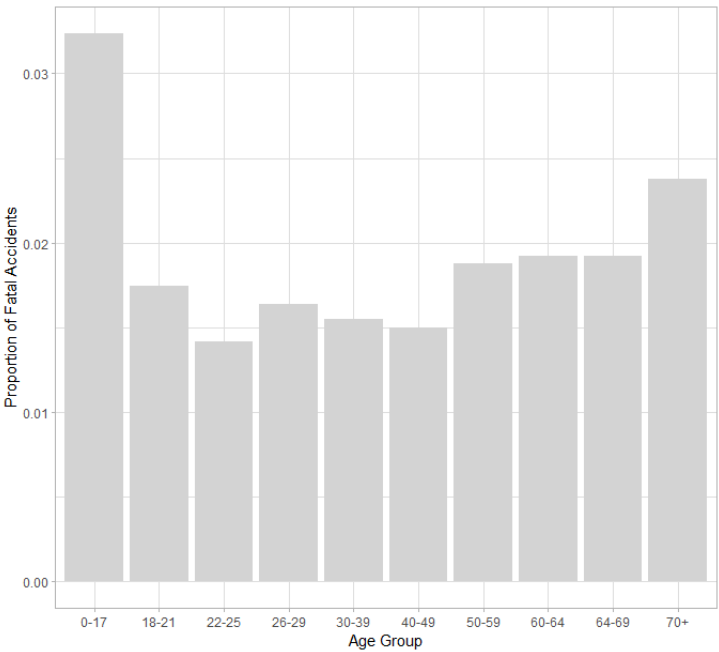


Figure 2 - Accidents by Driver Sex

	SEX	totalAccidents	fatalAccidents	proportionFatal
1	F	82615	858	0.010385523
2	M	117035	2527	0.021591832
3	U	350	2	0.005714286

Figure 3 - Proportion of Fatal Accidents by Vehicle Age

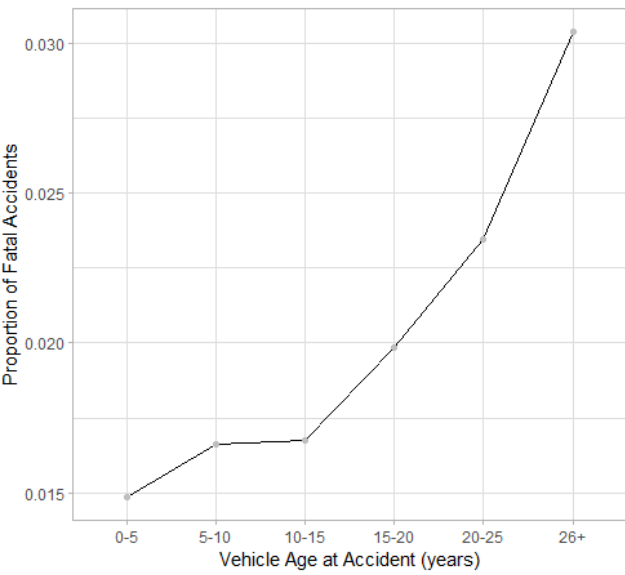




Figure 4 - Proportion of Fatal Accidents by Vehicle Type

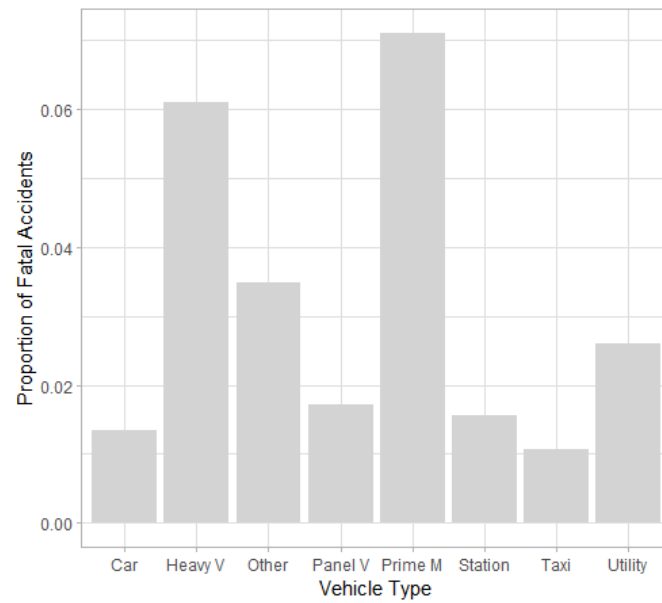
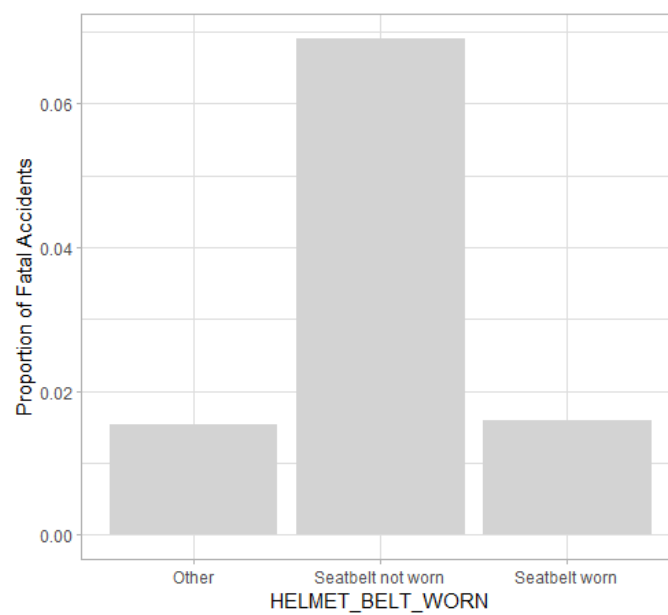
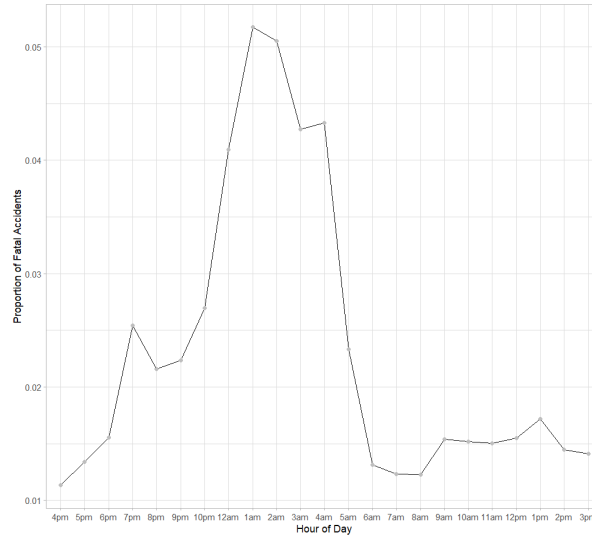


Figure 5 - Proportion of Fatal Accidents by Helmet/Belt Worn



*Figure 6 - Proportion of Fatal Accidents by Hour of Day*



### *Explanation 1 – Statistical Learning Model Selection & Creation*

#### **Statistical Learning Model Selection: Logistic Regression**

It is important to note that in this section, we are concerned with the relationship between predictors and fatal accidents. Therefore, the interpretability of the model is a prime focus in this scenario. Considering that the response variable is a binary classification (equal to true or false), we can use a logistic regression. To use a logistic regression, a series of assumptions must be made:

1. Observations are independent.
2. Predictors are uncorrelated.
3. Linear relationship between predictors and  $\log(\text{odds})$

There are slight limitations here. As observations are not necessarily independent as all accidents with another vehicle in the VicRoads Accident Data account may account for at least 2 separate accidents. Thus, even though the dataset is large, a significant portion of the observations may be correlated. To improve on this, the variable of VEHICLE\_YEAR\_MANUF. This is because we are more interested in the vehicle age, not the specific year the vehicle is manufactured. Since the accidents are recorded between 14 years, the vehicle age varies despite equal VEHICLE\_YEAR\_MANUF. Further, the predictors are not necessarily uncorrelated nor is there a definite linear relationship between predictors and  $\log(\text{odds})$ , however these assumptions must be made in applying a logistic regression.

#### **Statistical Learning Model Creation**

Prior to any processing, the data was cleaned. This involved removing variables which were assumed to be unrelated to accident fatality (to decrease processing time). These variables were DRIVER\_ID, AGE\_GROUP, VEHICLE\_ID and ACCIDENT\_NO. This assumption is valid as these variables are used to merely differentiate between different accident cases.

From here, a hybrid method was used to perform subset selection, using Bayes' Information Criteria (BIC) to penalise model complexity. Whilst we would ideally like to perform an exhaustive search for the best possible method, it requires the construction of  $2^p$  models which is unfortunately not possible due to technological limitations. Instead, a hybrid method still provides a relatively strong (but not the best) model despite taking a significantly lower amount of processing time. Further, BIC was chosen as it heavily penalises model complexity, which is beneficial as we prefer a slightly simpler, more interpretable model in this scenario.

In turn, the model created contained the following 11 variables:

SEX, AGE, HELMET\_BELT\_WORN, VEHICLE\_TYPE, TOTAL\_NO\_OCCUPANTS, ACCIDENTTIME, ACCIDENT\_TYPE, LIGHT\_CONDITION, ROAD\_GEOMETRY, SPEED\_ZONE, SURFACE\_COND.

*Figure 7 - Logistic Regression Model Summary*

```
> summary(glmModel)

Call:
glm(formula = fatal ~ SEX + AGE + HELMET_BELT_WORN + VEHICLE_TYPE +
    TOTAL_NO_OCCUPANTS + ACCIDENTTIME + ACCIDENT_TYPE + LIGHT_CONDITION +
    ROAD_GEOMETRY + SPEED_ZONE + SURFACE_COND, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2435  -0.1859  -0.1240  -0.0928   3.8763

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -7.480556    0.142639  -52.444 < 2e-16 ***
SEX                  0.401896    0.043189   9.305 < 2e-16 ***
SEXU                -0.601029    0.721176  -0.833  0.40462
AGE                  0.010704    0.001069  10.013 < 2e-16 ***
HELMET_BELT_WORNSeatbelt not worn  1.308254    0.075097  17.421 < 2e-16 ***
HELMET_BELT_WORNSeatbelt worn      0.019756    0.043379   0.455  0.64880
VEHICLE_TYPEHeavy Vehicle (Rigid) > 4.5 Tonnes 1.177335    0.084536  13.927 < 2e-16 ***
VEHICLE_TYPEOther                0.396402    0.080773   4.908  9.22e-07 ***
VEHICLE_TYPEPanel Van            0.060117    0.114319   0.526  0.59898
VEHICLE_TYPEPrime Mover - Single Trailer 0.933850    0.103785   8.998 < 2e-16 ***
VEHICLE_TYPEStation Wagon        0.041889    0.049084   0.853  0.39344
VEHICLE_TYPTaxi                 -0.323049    0.192603  -1.677  0.09349 .
VEHICLE_TYPEUtility              0.249939    0.056357   4.435  9.21e-06 ***
TOTAL_NO_OCCUPANTS              0.115532    0.012721   9.082 < 2e-16 ***
ACCIDENTTIME                 -0.016282    0.003110  -5.235  1.65e-07 ***
ACCIDENT_TYPEcollision with some other object -1.676665    0.322486  -5.199  2.00e-07 ***
ACCIDENT_TYPECollision with vehicle  -0.428573    0.047486  -9.025 < 2e-16 ***
ACCIDENT_TYPEFall from or in moving vehicle -0.019230    0.318118  -0.060  0.95180
ACCIDENT_TYPERNo collision and no object struck -1.086727    0.265398  -4.095  4.23e-05 ***
ACCIDENT_TYPERstruck animal        -1.762632    0.296295  -5.949  2.70e-09 ***
ACCIDENT_TYPERstruck Pedestrian     1.396646    0.069566  20.077 < 2e-16 ***
ACCIDENT_TYPERVehicle overturned (no collision) -0.997178    0.116122  -8.587 < 2e-16 ***
LIGHT_CONDITIONDark Street lights off  0.194894    0.258562   0.754  0.45099
LIGHT_CONDITIONDark Street lights on  -0.194142    0.067940  -2.858  0.00427 *
LIGHT_CONDITIONDark Street lights unknown -0.548943    0.267473  -2.052  0.04014 *
LIGHT_CONDITIONDay                -0.680137    0.055167  -12.329 < 2e-16 ***
LIGHT_CONDITIONDusk/Dawn           -0.790839    0.084495  -9.360 < 2e-16 ***
ROAD_GEOMETRYNot at intersection    0.359973    0.054064   6.658  2.77e-11 ***
ROAD_GEOMETRYOther                -0.301236    0.156276  -1.928  0.05391 .
ROAD_GEOMETRYT intersection         0.176342    0.062947   2.801  0.00509 **
SPEED_ZONE              0.041920    0.001134  36.966 < 2e-16 ***
SURFACE_CONDOther              -1.318258    0.216910  -6.077  1.22e-09 ***
SURFACE_CONDwet                -0.327033    0.050609  -6.462  1.03e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 34343  on 199999  degrees of freedom
Residual deviance: 29751  on 199967  degrees of freedom
AIC: 29817

Number of Fisher Scoring iterations: 8
```

*Table 1 - Impact of Driver SEX on Odds of Fatality*

Variable	SEX (Categorical)
Reference Predictor	F (Female)
Significant Predictor(s)	M (Male)
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	1.4947

Table 2 - Impact of Driver AGE on Odds of Fatality

Variable	AGE (Numerical)
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	1.0108

Table 3 - Impact of Helmet/ Belt Conditions on Odds of Fatality

Variable	HELMET_BELT_WORN (Categorical)
Reference Predictor	OTHER
Significant Predictors	SEATBELT_NOT_WORN
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	3.6997

Table 4 - Impact of VEHICLE\_TYPE on Odds of Fatality

Variable	VEHICLE_TYPE (Categorical)			
Reference Predictor	CAR			
Significant Predictors	HEAVY_VEHICLE	OTHER	PRIME_MOVER	UTILITY
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	3.2457	1.4865	2.544	1.2839

Table 5 - Impact of TOTAL\_NO\_OCCUPANTS on Odds of Fatality

Variable	TOTAL_NO_OCCUPANTS (Numerical)
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	1.1225

Table 6 - Impact of ACCIDENTTIME on Odds of Fatality

Variable	ACCIDENTTIME (Numerical)
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	0.9838

Table 7 - Impact of ACCIDENT\_TYPE on Odds of Fatality

Variable	ACCIDENT_TYPE (Categorical)					
Reference Predictor	FIXED_OBJECT					
Significant Predictors	VEHICLE	NO COLLISION	ANIMAL STRUCK	PEDESTRIAN STRUCK	VEHICLE OVERTURNED	OTHER OBJECT
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	0.6514	0.3373	0.1716	4.042	0.3689	0.1870

*Table 8 - Impact of LIGHT\_CONDITION on Odds of Fatality*

Variable	LIGHT_CONDITION (Categorical)		
Reference Predictor	DARK_NO_STREET_LIGHTS		
Significant Predictors	STREET_LIGHTS_ON	DAY	DUSK/ DAWN
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	0.8235	0.5065	0.4535

*Table 9 - Impact of ROAD\_GEOMETRY on Odds of Fatality*

Variable	ROAD_GEOMETRY (Categorical)	
Reference Predictor	CROSS_INTERSECTION	
Significant Predictors	NOT_AT_INTERSECTION	T_INTERSECTION
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	1.4333	1.1928

*Table 10 - Impact of SPEED\_ZONE on Odds of Fatality*

Variable	SPEED_ZONE (Numerical)
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	1.0428

*Table 11 - Impact of SURFACE\_COND on Odds of Fatality*

Variable	SURFACE_COND	
Reference Predictor	DRY	
Significant Predictors	WET	OTHER
Odds of Fatal Accident ( $e^{x^T \cdot \beta}$ )	0.7211	0.2676

### *Explanation 2 - Predictive Analysis Data Cleaning, Balancing & Split*

#### **Predictive Analysis Data Cleaning, Balancing & Split**

First, the data was cleaned by removing all variables related to accident conditions (as required) as variables deemed insignificant in task 2. These were VEHICLE\_MAKE, VEHICLE\_COLOUR and OWNER\_POSTCODE.

Next, the data was split into a training and validation set (at a 75:25 ratio). This ratio was selected as a strong balance which allows both a significant amount of training observations and robust testing.

Finally, the VicRoads Accident Data provided showed a significant imbalance, with only roughly 2% of observations being fatal. This means that the target variable is highly under-represented, and thus predictive analysis on the data is prone to, bias, poor generalisation, and misleading results (that may favour the more represented variable simply because there are more observations of it). To avoid these issues, the training data was balanced using a combination of over sampling and under sampling (using the ovun.sample function). Since we are interested rarer but highly important minority class, we can oversample fatal observations and under sample non-fatal observations to equally represent either of them within the data, thus solving the issues above. However, it is important to note that oversampling increases the risk of overfitting data, an issue that affected a set of results.

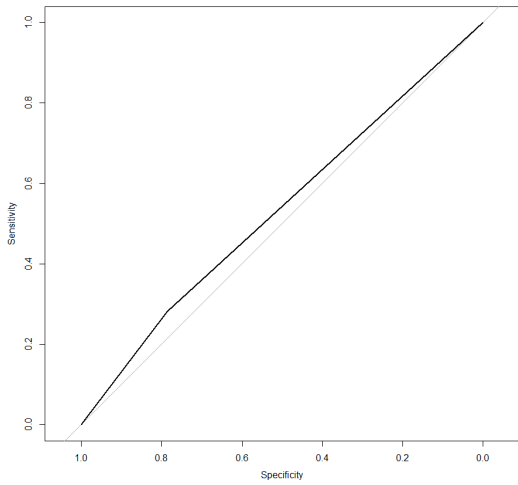
Table 12 - AUC of Different Models

Model	AUC
GLM	0.6788978
Lasso	0.6794019
Ridge	0.6790618
KNN	0.5338933
GBM	0.6873709
Random Forest	0.5877

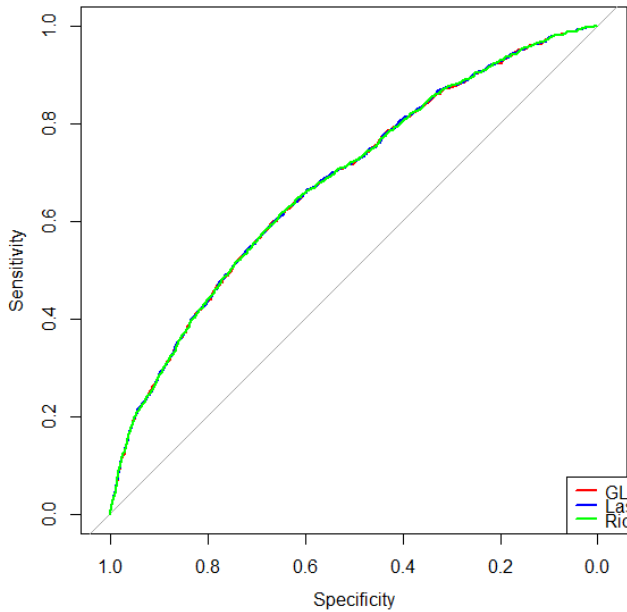
### Explanation 3 - Predictive Analysis Model Breakdowns

#### Predictive Analysis Model Breakdowns

##### Model 1 - KNN

Model	K-Nearest Neighbours (KNN)
ROC Curve	 <p>The figure is a Receiver Operating Characteristic (ROC) curve for the K-Nearest Neighbours (KNN) model. The x-axis is labeled 'Specificity' and ranges from 1.0 on the left to 0.0 on the right, with major ticks at 1.0, 0.8, 0.6, 0.4, 0.2, and 0.0. The y-axis is labeled 'Sensitivity' and ranges from 0.0 at the bottom to 1.0 at the top, with major ticks at 0.0, 0.2, 0.4, 0.6, 0.8, and 1.0. A solid black line represents the ROC curve, starting at (1.0, 0.0) and ending at (0.0, 1.0). A dashed diagonal line from (1.0, 0.0) to (0.0, 1.0) represents the performance of a random classifier. The solid curve is positioned above the dashed line, indicating better performance than random guessing. The area under the curve (AUC) is 0.5338933.</p>
Method	The balanced training data is used to train the KNN. The standard $K = 5$ is selected, and the ROC curve is plotted.
AUC	0.5338933
Analysis	KNN is an instance-based learning algorithm that can be used for both classification and regression problems. However, the method used to balance the data involved a significant magnitude of over-sampling. The technique used was <code>ovun.sample</code> , which creates duplicates of data points in which <code>fatal = true</code> . This method turned 1000 observations into 25000, essentially duplicating each observation 25 times. Thus, KNN specifically is highly prone to the training set's bias, as each individual point constitutes 25 "neighbours". Due to this, the model performs very poorly on the test set, with the lowest AUC (at 0.5338933).

## Model 2 – Regression, Lasso & Ridge

Model	Logistic Regression, Lasso and Ridge		
ROC Curve	<p style="text-align: center;"><b>ROC Curve for GLM, Lasso, Ridge</b></p> 		
Method	<p>The models were trained using the balanced training set. For logistic regression, insignificant variables (identified in task 2) were excluded, resulting in the created model being equivalent to that of task 2. For Lasso and Ridge regression, cross validation was used to find a minimum <math>\lambda</math> (penalty) that best fits the model onto the training set. From here, the 3 models were evaluated on the validation set, on which they showcased very similar AUCs (listed below). These models performed very well, however could not meet the standards set by GBM.</p>		
AUC	Logistic Regression	Regression + Lasso	Regression + Ridge
	0.6788978	0.6794019	0.6790618
Analysis	<p>The reason these models all exhibit a similar AUC is because Lasso and Ridge perform shrinkage techniques to minimise the influence of specific predictors. It is likely that the predictors minimised would have been the ones excluded from the regression, thus resulting in similar modelling and performance.</p>		

Model 3 - Random Forest

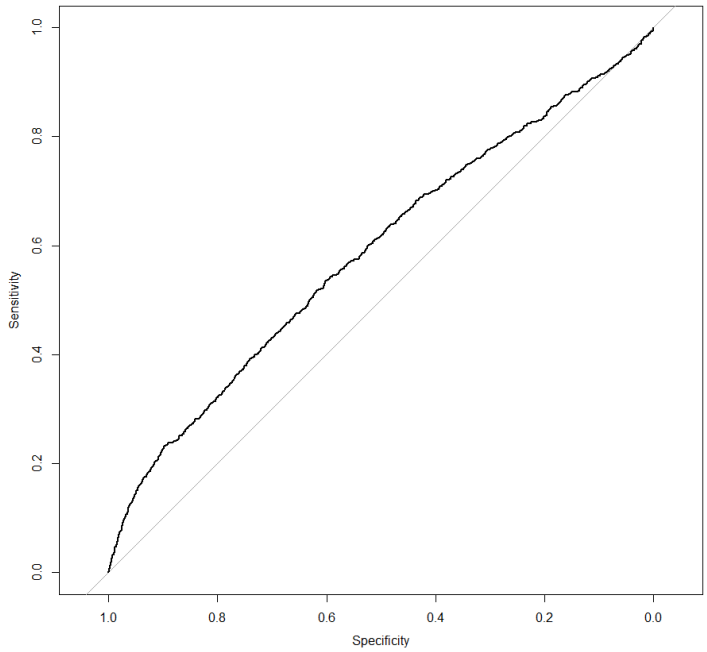
Model	Random Forest
ROC Tree	 The figure is a Receiver Operating Characteristic (ROC) curve for a Random Forest model. The title is "ROC Curve for Random Forest". The y-axis is labeled "Sensitivity" and ranges from 0.0 to 1.0 with increments of 0.2. The x-axis is labeled "Specificity" and ranges from 1.0 to 0.0 with increments of 0.2. A diagonal line from (1.0, 0.0) to (0.0, 1.0) represents a random classifier. The ROC curve is a solid black line that starts at (1.0, 0.0) and ends at (0.0, 1.0), arching above the diagonal line. The area under the curve (AUC) is approximately 0.5877.
Method	The model was fit onto the balanced training set, with 10 trees (initially). It was then reattempted with larger inputs for ntrees, however due to technological limitations, processing these larger values wasn't possible. Thus, there was a clear limitation established meaning the model was less likely to perform well on the validation set. Under the default setting with ntrees = 10, the model obtained an AUC of 0.5877, which is relatively poor compared to the other models.
AUC	0.5877

Table 13 – SEX Count in Fatal Predictions

SEX	Count
F	962
M	1535
U	3



## **Appendix 2 – Generative AI Use**

### **General Usage**

Chat GPT was used for general formatting and presentation assistance. This involved the prompts:

- “Provide me an executive summary structure”.
- “Provide me a data analytics report”.

### **Task 1 Usage**

Task 1 AI use was limited to making plots and tables look better. This involved a few prompts at after completion of my task 1 code such as:

- “Make my plot look better. Follow a black, dark grey, light grey and white theme (input code)”

Thus, Chat GPT was used to refactor my code in this way. This prompt was used a few times before realising that Chat GPT was returning me broken code, which I then took the logic from and improved the design of other plots, graphs, and curves.

### **Task 2 Usage**

Task 2 involved usage of Chat GPT to assist in researching how to perform subset selection on my dataset. The prompt utilised was:

- “How can I perform subset selection by exhaustion in R”.

This provided me the necessary information to begin subset selection on my data. After realising that certain forms of subset selection were not possible (due to technological limitations), I then performed my own research and looked through labs to identify what could be changed to fix this.

### **Task 3 Usage**

Task 3 involved more use of ChatGPT. Not only was it used as a tool to research different models and how to use them in R, but it was also used for debugging and speeding up redundant code. Some examples of this include prompts such as:

- “I have this error in R, what could be going wrong (input code)” – regarding an error about fatal needing to be converted to binary.
- “(input code of regression, lasso and ridge ROC curves) combine these 3 plots into one” – used to plot all the curves (which I had already done individually) onto one plot.
- “What is GBR and how can I use it in R?”
- “How to use the predict () function in R”