1 Exploratory Data Analysis

Exploratory Data Analysis was conducted on accidents that occurred on Victoria roads to gain basic insight on the conditions at the time of the accident, profiles of the drivers and the characteristics of the vehicles.

1.1 Conditions at the time of Accidents

For conditions at the time of accidents, we look at the following three conditions: light conditions, surface conditions and atmospheric conditions. For more details on each type of condition, please visit [Appendix Section 4.1.1].

Most fatal accidents occurred in poor light conditions where streets were dark with no street lights or street lights turned off (6.05%). Most fatalities occurred in dry surface conditions (1.77%). Finally, out of all atmospheric conditions, foggy atmospheric conditions resulted in the most fatalities (3%). Dark or foggy conditions may make drivers more prone to mistakes due to poor visibility, which also reduces that likelihood of taking evasive action successfully. As for surface conditions, drivers are more likely to travel at higher speeds in drier conditions, which increases the odds of fatalities in an accident as we will see in section 2 under speed zones.

1.2 Profile of Drivers involved in Accidents

During our exploratory data analysis, we focused on the following profiles of drivers: sex, age group and whether seat belts were worn by the driver. For more information about the data for these categories, please visit [Appendix Section 4.1.2].

Majority of accidents involved male drivers (58.52%), as opposed to female and unknown. Furthermore, most fatalities and accidents also involved male drivers (2.16%), this was more than double that of female drivers (1.04%).

Secondly, drivers within the ages of 22-29 (20.74%) and 30 to 39 (20.50%) were involved in accidents the most on Victoria roads, far greater than any other age group. However, it was older drivers who were above the age of 50 (4.14%) that were involved in most fatalities.

Finally, accidents where drivers did not wear seatbelts were 4 times more likely to be fatal (6.89%) than drivers that wore seatbelts.

1.3 Characteristics of Vehicles involved in Accidents

During our exploratory data analysis, we focused on the following characteristics of vehicles: body styler, make and type. For more information about the data for these categories, please visit [Appendix Section 4.1.3].

The data suggests that the vehicles that were most involved in accidents had a SEDAN body style (55.52%). In addition, the make of the vehicle that were most involved in accidents are Toyota (18.57%) closely followed by Holden (17.87%). Finally, the type of vehicle that was most involved in accidents were cars. These may be due to the fact that most drivers in the population drive cars with these characteristics.

Vehicles that resulted in most fatal accidents had P MVR body style (6.67%). Vehicles made by Kenworth (8.6%) also resulted in the highest fatalities when compared to other makes, which is more than 4 times higher than the second highest percentage fatality. Finally, the vehicle type that resulted in the highest percentage of fatalities were prime movers (7.1%) closely followed by Heavy vehicles (6.1%). These statistics are very interesting as all these characteristics relate to heavy-duty truck-like vehicles, suggesting that accidents involving trucks result in higher likelihood of fatalities. We can speculate that this is because trucks are very heavy and can generate high amounts of force during a collision, making the chances of survival for those involved in the accident to be very slim.

2 Modelling the Relationship of Fatal Accidents and Different Variables

2.0 Construction of the Model

The relationship between the fatality of the accidents and the different variables were modelled using binary logistic regression. For more information about how the model was constructed, the assumptions that were identified and how the variables for the model were selected, please visit [Appendix Section 4.2.0]. For the numerical results of the model which includes the value of coefficients, please visit [Appendix Section 4.2.1].

2.1 Evaluation of the GLM Model

After constructing the model using the down-sampled, balanced train set data, we now evaluate the model using a balanced test data set. For evaluation, a confusion matrix was constructed.

Confusion Matrix of logistic regression model against the downsampled test set:

	Reference	
Prediction	Non-Fatal	Fatal
Non-Fatal	470	170
Fatal	176	469

This logistic regression model has an accuracy of 0.7307 with a 95% confidence interval (0.7056, 0.7548). Therefore, the error rate is 0.2693.

We have a false-positive rate of 0.272 and a true-positive rate of 0.734. Since we are focused on modelling fatalities, we are more concerned with predicting fatalities correctly. Therefore, a true positive rate of 0.734 suggests that we correctly predict that a fatality will occur given the covariates of the model, 73.4% of the time.

Finally, when computing the AUC of the ROC curve, we obtain 0.7308. This is well above 0.5, suggesting that the model is efficient and performs soundly in distinguishing between fatalities and non-fatalities during accidents on Victoria Roads.

AIC and Deviance of the logistic regression GLM model:

AIC	Null Deviance	Residual Deviance
5991	7609.4 (5488 df)	5907.0 (5447 df)

In our logistic model, we have a null deviance of 7609.4 and a residual deviance of 5907.0. This means that incorporating the covariates in our model as opposed to not reduces the deviance by 1702.4. This is a very big difference, suggesting that the model's covariates were a good choice and that the model fits the data well.

Therefore, this model was chosen to highlight the relationship between fatality of accidents on Victoria roads and different variables.

2.2 Interpretation of the GLM Model

Now that we understand that the logistic regression GLM model is sound in distinguishing between fatalities and non-fatalities during accidents on Victoria Roads from section 2.1, we now focus on interpreting the relationships prevalent in the model to gain additional insight. To understand how these relationships were derived from the model, please see [Appendix Section 4.2.0.4].
In this main report, we include interesting relationships that the model claims. For all relationships that can be derived from this model, please visit [Appendix Section 4.2.0.5].

SEX: According to our model, accidents with male drivers have 1.604 ($e^{0.4727}$) times the fatal odds of accidents with female drivers in Victorian road accidents.

SEATBELTS: Accidents where drivers do not wear seatbelts have 4.173 ($\frac{e^{1.504}}{e^{0.075}}$) times the fatal odds of accidents where drivers do wear seatbelts. Seatbelts are a mandatory feature that keep people safe in vehicles, as a result drivers that do not wear them have a higher chance of death.

VEHICLE TYPE: The model suggests that accidents are more likely to be fatal if they involve heavy vehicles. Heavy vehicles over 4.5 tonnes have 4.2 times the fatal odds of accidents that involve cars. On the other hand, the safest vehicle type according to the model is the taxi, which only has 0.83 times the fatal odds of accidents that involve cars. These results match with the exploratory data analysis conducted.

ACCIDENT TYPE: According to the model, the most dangerous accident type is when a pedestrian was struck. In this type of accident, the odds of fatality was 5.66 times that of accidents that involved collision with another vehicle. This is most probably because pedestrians are very vulnerable during a traffic accident and do not have the safety that is provided within a vehicle.

LIGHT CONDITIONS: The model suggests that accidents had a higher likelihood of being fatal in dark streets. Mainly when there were no street lights present or when they were off. The odds of fatality when street lights were off in the dark is 1.5 times that of the odds of an accident being fatal when the street lights were on in the dark.This may be because drivers were unable to take evasive action in time or had to react late due to poor visibility which resulted in higher odds of fatality.

ROAD GEOMETRY: According to the model, accidents that were not at an intersection resulted in the highest likelihood of fatalities, i.e. 1.6 times the odds of fatality when it occurred in a cross intersection. This may be because drivers may be generally more cautious near intersections or more safety is imposed near intersections such as roundabouts, road signs or traffic lights.

SPEED ZONES: The model suggests that fatality in accidents increases with the maximum allowed speed in the area (speed zone). This can be seen by an increase in the coefficient value in conjunction to the maximum allowed speed in the model. Looking at the extremes, accidents that occur at a speed zone of 110 have 21.56 times the odds of being fatal than accidents that occur at a speed zone of 40. This increasing trend in higher odds of fatality as the speed zone increases makes sense as accidents that occur at higher speeds are more dangerous and cause more damage.

2.3 Contextual Issues and Possible Solutions .

From the analysis of our interpretable GLM model, we understand that two of the biggest factors in increasing the odds of fatality in accidents are the vehicle weight and speed. The higher the speed and weight of a vehicle, the bigger the odds of fatality in the accident. These can be seen through our analysis of vehicle type and speed zones. Therefore, one solution is to erect speed signs specific to heavy vehicles in Victorian roads where these accidents are common. On the other hand, a policy setting the maximum allowable speed of a vehicle to be the minimum of the speed zone or a limit based on the vehicle's weight could be put in place.

Secondly, after further analysis of how road geometry affects fatality of accidents, we understand that accidents are more fatal in non-intersection areas. Therefore, it would be in the best interest to place speed bumps and other safety measures in non-intersection areas, particularly in roads that allow for high speed and have poor lighting.

Finally, the model suggests that accidents where drivers do not wear seat belts significantly increase the odds of fatalities. Thus, it would be best to place seat belt cameras in common accident areas and charge heavy fines to incentivise drivers to abide by the law.

3 Predictive Modelling of Drivers and Fatal Accidents

Initially, 3 models were constructed to predict fatality of accidents with respect to driver demographic and vehicle information. These models involve an elastic net logistic regression, random forest and k-nearest neighbours model. To view more information about the initial set up required to construct these models, please visit [Appendix Section 4.3.0].

3.1 Training Predictive Models and Choosing Tuning Parameters

After the initial set up, we have a balanced training set for our predictive models. Note that the elastic net logistic regression model and knn models have balanced and standardised training sets. For each predictive model, we conduct stratified k-fold cross validation in order to train the model and choose the optimal tuning parameters such that the error rate is minimised. Stratified k-fold cross validation was used to ensure that the proportion of fatal and non fatal accidents are appropriate in each fold to eliminate the issue of imbalanced data and ensure models are more suitable for prediction.

Based on the criteria of minimising the error rate, the stratified 10-fold cross validation suggests that the best tuning parameters for the elastic net logistic regression is $\lambda = 0.0001795657$ with $\alpha = 0.55$, with a minimum error rate of 0.3737. The stratified 5-fold cross validation on the random forest model suggests that the best value for the number of random variables to be considered at each split is 2 with a minimum error rate of 0.3977. Finally the stratified 10-fold cross validation on the knn model suggests that the best value for k to minimise the error rate is k = 9 with an error rate of 0.4114.

3.2 Evaluation of the Predictive Models

To evaluate the predictive models, we analyse the confusion matrix generated by the models based on the balance testing set. Note that the testing set for the elastic net logistic model and knn model were standardised.

Confusion Matrix of the elastic net logistic regression model against the downsampled test set:

	Reference	
Prediction	Non-Fatal	Fatal
Non-Fatal	426	286
Fatal	220	353

The accuracy of the elastic net logistic regression model based on the test data is 0.6062 with a 95% confidence interval of (0.5789, 0.6331). This model has a true positive rate of 0.5524 and a false positive rate of 0.3406. The confusion matrix's ROC has an AUC of 0.6059.

Confusion Matrix of the random forest model against the downsampled test set:

	Reference	
Prediction	Non-Fatal	Fatal
Non-Fatal	490	341
Fatal	156	298

The accuracy of the random forest model based on the test data is 0.6132 with a 95% confidence interval of (0.586, 0.64). This model has a true positive rate of 0.4664 and a false positive rate of 0.2415. The confusion matrix's ROC has an AUC of 0.6124.

Confusion Matrix of the k-nearest neighbours model against the downsampled test set:

	Reference	
Prediction	Non-Fatal	Fatal
Non-Fatal	377	300
Fatal	156	298

The accuracy of the KNN model based on the test data is 0.5572 with a 95% confidence interval of (0.5295, 0.5846). This model has a true positive rate of 0.5305 and a false positive rate of 0.3406. The confusion matrix's ROC has an AUC of 0.5571.

Based on the test data confusion matrices of the three predictive models, it is clear that the KNN model is the least effective as its accuracy (0.5572) is far lower than the random forest (0.6132) and elastic net logistic regression (0.6062) models. If we solely look at the accuracy of the models as a metric for the final two predictive models, it is clear that the random forest is a slightly better predictive model than the elastic net logistic regression model. However, based on the context of our problem, we are required to create a model that is able to predict when fatalities occur in accidents correctly. This implies that a high true positive rate is important in our predictive models. Therefore, although the random forest model is more accurate by 0.007, the true positive rate of the elastic net logistic regression model is 0.086 higher than the random forest model. Therefore, our selected model for predictive modelling of drivers and fatal accidents is the elastic net logistic regression model. For more information about the selected model such as its coefficients, please visit [Appendix Section 4.3.1].

3.3 Drivers to Target for the Prevention Campaign

For additional information of how the characteristics of drivers to target for the prevention campaign were decided and derived from our GLM and predictive model, please visit [Section 2.2] and [Appendix Section 4.3.2] respectively.
Based on the interpretation of the predictive and interpretable models, the prevention campaign should be targeted towards drivers with at least one of the following characteristics:

- Drivers of ages above the retirement age. This is because fatal odds increase with age.
- Drivers of heavy vehicles such as trucks and buses. This is due to Heavy vehicles having the highest fatal odds of any other vehicle type.
- Drivers that do not wear seatbelts, or have been found not wearing seatbelts at least once by camera. Not wearing seat belts increases the odds of fatality in an accident 5 fold.
- Drivers with multiple speeding offences. Our GLM model from section 2 shows that the fatal odds in an accident is positively correlated with speed zones, implying that speed makes accidents more dangerous.

4 Appendix

4.1.0 Calculations

$(\%\ Accidents) = \frac{fatal + non\ fatal}{200000} * 100$

% Accidents are defined to be the percentage of accidents that the category involves in the whole data set.

$(\%\ Fatal) = \frac{fatal}{fatal + non\ fatal} * 100$

% Fatal are defined to be the percentage of accidents that were fatal in that given category.

4.1.1 Data of Conditions at the time of Accidents

Light Condition	Dark no lights	Dark with lights	Dark, unknown lights	Day	Dusk/Dawn
Non-Fatal	9393	32622	1304	136999	16295
Fatal	605	561	15	1990	216
% Accidents	5	16.6	0.66	69.49	8.26
% Fatal	6.05	1.69	1.14	1.43	1.31

Surface Condition	Dry	Wet	Other
Non-Fatal	159,368	31,925	5,320
Fatal	2,868	497	22
% Accidents	81.11	16.21	2.67
% Fatal	1.77	1.53	0.41

Atmos. Cond	Clear	Fog	Raining	Other
Non-Fatal	163329	1649	21966	9669
Fatal	2963	51	305	68
% Accidents	83.15	0.85	11.14	4.87
% Fatal	1.78	3	1.37	0.7

4.1.2 Data of Profile of Drivers involved in Accidents

Sex	Male	Female	Unknown
Non-Fatal	114,508	81757	348
Fatal	2527	858	2
% Accidents	58.52%	41.31%	0.18%
% Fatal	2.16%	1.04%	0.57%

Age Group	16-21	22-29	30-39	40-49	50-59	60+
Non-Fatal	24369	40864	40350	35019	26340	29671
Fatal	427	624	649	528	514	645
% Accidents	12.40%	20.74%	20.50%	17.77%	13.43%	15.16%
% Fatal	1.72%	1.50%	1.58%	1.49%	1.91%	2.23%

Seatbelt	Seat Belt Worn	Seat Belt Not Worn	Other
Non-Fatal	143144	3943	49526
Fatal	2327	292	768
% Accidents	72.74%	2.12%	25.15%
% Fatal	1.60%	6.89%	1.53%

4.1.3 Data of Characteristics of Vehicles involved in Accidents

Body Style	Coupe	DC UTE	Other	P MVR	S WAG	SED
Non-Fatal	7558	5697	8231	2545	13799	5648
Fatal	98	170	282	189	269	98
% Accidents	3.83%	2.93%	4.26%	1.37%	7.03%	2.87%
% Fatal	1.28%	2.90%	3.31%	6.67%	1.91%	1.71%

Body Style	SEDAN	TRAY	UTIL	VAN	WAGON
Non-Fatal	109569	3626	9491	6155	24294
Fatal	1465	118	224	140	334
% Accidents	55.52%	1.87%	4.86%	3.15%	12.31%
% Fatal	1.32%	3.15%	2.31%	2.22%	1.36%

Make	Ford	Holden	Honda	Hyundai	Kenworth	Mazda
Non-Fatal	30336	35064	8591	8612	1126	11518
Fatal	539	679	84	101	106	138
% Accidents	15.44%	17.87%	4.34%	4.37%	0.62%	5.83%
% Fatal	1.75%	1.90%	0.97%	1.16%	8.6%	1.18%

Make	Merc B	Mitsub	Nissan	Other	Subaru	Toyota	Volks
Non-Fatal	3998	13083	11628	27392	4482	36563	4220
Fatal	68	208	203	581	49	580	51
% Accidents	2.03%	6.65%	5.91%	13.99%	2.27%	18.57%	2.14%
% Fatal	1.67%	1.56%	1.72%	2.08%	1.08%	1.56%	1.19%

Type	Car	Heavy	Other	Panel Van	Prime Mover	Station Wagon	Taxi	Utility
Non-Fatal	123297	2939	5841	4913	1623	37723	2672	17605
Fatal	1679	191	211	86	124	599	29	468
% Accidents	62.49%	1.57%	3.03%	2.5%	0.87%	19.16%	1.35	9.04%
% Fatal	1.34%	6.10%	3.49%	1.72%	7.10%	1.56%	1.07%	2.59%

4.2.0 Construction of GLM Modelling Relationship of Fatal Accidents and Different Variables

4.2.0.1 Identification and Explanation of Assumptions for the Construction of the GLM

We are tasked to understand the relationship between fatal accidents and the different variables available in the data set**.** Since the fatal accidents are a binary outcome, the relationship can be modelled using binary logistic regression as logistic regression assumes that the dependent variable (fatal accidents) are binary.
Since our chosen generalised linear model is constructed using logistic regression, we have the following assumptions:

1. The observations in the data set are independent.
2. A linear relationship is assumed between the independent variables and the log odds.
3. There is little to no collinearity between the independent variables.

4.2.0.2 Dealing with Imbalanced Data

It is important to notice that the data set contains imbalanced data. The number of non-fatalities (196,613) far outweigh the number of fatalities (3387). Therefore, it is highly probable that if we sample from the entire data set to create a test and training data set, the model will tend to be incorrect in predicting fatalities as there are very few observations of them. This means that the true-positive rate will be very close to 0. This can be solved using either upsampling the fatal accidents, or downsampling the non-fatal accidents. Since we are using logistic regression that assumes that the observations in the data set are independent, we must downsample the non-fatal accidents. Therefore, we create a new dataset *FatalData.down* that contains equal amounts of fatalities as non-fatal accidents. We then create a training and test set using an 80/20 split respectively.

4.2.0.3 Selecting Variables for the GLM Model

Initially, the logistic regression GLM model included all covariates of the data set excluding: DRIVER_ID, AGE GROUP, LICENSE_STATE, VEHICLE_ID, OWNER_POSTCODE and ACCIDENT_NO. This is because many of these covariates such as DRIVER_ID, VEHICLE_ID and ACCIDENT_NO had no causation in regards to fatal outcomes. Furthermore, AGE GROUP was not included as the covariate AGE was included. Having both adds collinearity between the covariates AGE and AGE_GROUP which goes against the assumption of a logistic regression model. In addition, the model is more interpretable by containing a numerical covariate as opposed to categorical variables in the situation that both represent the same concept (i.e. age). Finally, OWNER_POSTCODE was not included as it involved too many categories making the entire model very difficult to interpret.

This logistic regression GLM model was trained using the down-sampled, balanced training set. Arguably, the model remained too difficult to interpret due to the large amount of covariates that were involved. Therefore, a new nested logistic regression model was created using a subset of the covariates in the old model using the notion of significance.

Consider the following hypothesis test: $H_0: \beta_k = 0$ whereas $H_1: \beta_k \neq 0$. We have the test statistic Z such that:

$$Z = \frac{\widehat{\beta_k} - 0}{SE(\beta_k)} \sim N(0, 1).$$

For large absolute value of the z-statistic above, there is evidence against $H_0: \beta_k = 0$. Therefore, using a significance level $\alpha = 0.0001$ to improve the interpretability of the model, we create a new logistic regression model containing only the variables that are significant. Thus, we create a new logistic regression model containing the following variables: fatal ~ Sex + AGE + HELMET_BELT_WORN + VEHICLE TYPE + TOTAL_NO_OCCUPANTS + ACCIDENTTIME + ACCIDENT_TYPE + LIGHT_CONDITION + ROAD_GEOMETRY + SPEED_ZONE + SURFACE_COND + ATMOSPH_COND.

Please see [Appendix Section 4.2.1] for the results of the new and selected logistic regression GLM modelling the relationship between fatality of accidents and different variables.

4.2.0.4 Way of Interpreting Relationships from the Model

To understand the relationship from the GLM model, we use the following principle. Holding all other covariates constant, a one unit increase in a numeric covariate k is expected to change the fatal odds by a factor of e^{β_k}. As for categorical covariates m such as speed zone (which has the reference as SPEED_ZONE040), accidents in speed zone denoted by m have an e^{β_m} times the fatal odds of accidents in the 40 speed zone. Using this, we are able to gather many insights from our logistic regression model.

4.2.0.5 More Relationships and Interpretations derived from the Model

AGE : With each increase in year of age of the driver involved in the accident, the change in odds of fatality of the accident are by a factor of 1.01. This implies that as drivers get older, the odds of fatality in the event of an accident increase slightly. This may be because people become more feeble as they age, which heavily reduces the chance of survival for the driver.

TOTAL NUMBER OF OCCUPANTS: With one additional person in the vehicle involved in an accident in Victorian roads, the model suggests that the odds of that accident being fatal increases by a factor of 1.2. We can speculate this to be the case because, with more people involved in the accidents, the likelihood of that accident being fatal is higher as it only requires one person to die for an accident to be deemed fatal.

ACCIDENT TIME: The model suggests that time of accident changes the odds of fatality by very little. With each unit of time, the odds of fatality in the accident decrease by 0.00056% ($1 - e^{0.65*10^{-6}}$).

SURFACE CONDITION: The model implies that the odds of fatality in accidents are higher in dry surfaces. This may be due to drivers being able to drive at faster speeds in dry surfaces that makes accidents and collisions more dangerous. Odds of an accident being fatal in dry surface conditions is 1.18 times that of the odds in wet surfaces.

ATMOSPHERIC CONDITION: The model suggests that accidents are more likely to be fatal in clear atmospheric conditions. This, once again, may be because clear atmospheric conditions allow for drivers to travel at higher speeds, which makes accidents more fatal when they occur. Odds of an accident being fatal in clear atmospheric conditions are 1.33 times the odds when it is raining and 1.22 times the odds when it is foggy.

4.2.1 Results of the GLM Model involving the coefficients

	β Estimate	Std. Error	z-value	Pr(>|z|)
Intercept	-1.972e+00	3.249e-01	-6.070	1.28e-09
Male Sex	4.727e-01	7.200e-02	6.565	5.19e-11
Unknown Sex	-5.456e-01	9.510e-01	-0.574	0.566149
Age	1.006e-02	1.854e-03	5.424	5.81e-08
Seat Belt not worn	1.504e+00	1.992e-01	7.551	4.33e-14
Seat Belt worn	7.526e-02	7.609e-02	0.989	0.322645
Heavy Vehicle Type	1.437e+00	2.031e-01	7.075	1.50e-12
Other Vehicle Type	8.656e-01	1.731e-01	5.001	5.71e-07
Panel Van Vehicle Type	2.226e-01	2.041e-01	1.091	0.275345
Prime Mover - Single Trailer Vehicle Type	1.186e+00	2.738e-01	4.331	1.48e-05
Station Wagon Vehicle Type	2.938e-02	8.489e-02	0.346	0.729281
Taxi Vehicle Type	-1.896e-01	2.950e-01	-0.642	0.520568
Utility Vehicle Type	2.659e-01	1.075e-01	2.473	0.013387

Total Number of Occupants	1.926e-01	3.623e-02	5.316	1.06e-07
Accident Time	-5.650e-06	1.662e-06	-3.400	0.000674
Collision with other object Accident Type	-8.601e-01	5.513e-01	-1.560	0.118762
Collision with Vehicle Accident Type	-5.626e-01	9.447e-02	-5.955	2.60e-09
Fall from or in moving vehicle Accident Type	-6.974e-02	5.982e-01	-0.117	0.907182
No collision and no object struck Accident Type	-4.862e-01	5.002e-01	-0.972	0.331067
Struck animal Accident Type	-1.733e+00	4.987e-01	-3.476	0.000509
Struck Pedestrian Accident Type	1.171e+00	1.407e-01	8.317	< 2e-16
Vehicle overturned (no collision) Accident Type	-9.629e-01	2.347e-01	-4.102	4.10e-05
Dark Street lights off Light Condition	1.352e-01	4.951e-01	0.273	0.784792
Dark Street lights on Light Condition	-2.816e-01	1.455e-01	-1.936	0.052923
Street lights unknown Light Condition	-4.250e-01	4.168e-01	-1.020	0.307875
Day Light	-7.927e-01	1.285e-01	-6.171	6.80e-10

Condition				
Dusk/Dawn Light Condition	-9.679e-01	1.672e-01	-5.788	7.11e-09
Not at intersection Road Geometry	4.779e-01	8.644e-02	5.528	3.24e-08
Other Road Geometry	-3.635e-01	2.436e-01	-1.492	0.135600
T intersection Road Geometry	3.426e-01	9.842e-02	3.481	0.000499
050 Speed Zone	5.760e-01	2.562e-01	2.248	0.024545
060 Speed Zone	1.040e+00	2.458e-01	4.229	2.35e-05
070 Speed Zone	1.567e+00	2.600e-01	6.025	1.69e-09
080 Speed Zone	1.786e+00	2.503e-01	7.135	9.70e-13
090 Speed Zone	1.999e+00	3.957e-01	5.052	4.38e-07
100 Speed Zone	2.758e+00	2.504e-01	11.015	< 2e-16
110 Speed Zone	3.071e+00	3.601e-01	8.529	< 2e-16
Other Surface Condition	-1.252e+00	3.452e-01	-3.626	0.000288
Wet Surface Condition	-1.695e-01	1.468e-01	-1.154	0.248444
Fog Atmospheric Condition	-2.005e-01	3.364e-01	-0.596	0.551168
Other Atmospheric Condition	-7.463e-01	2.175e-01	-3.431	0.000601
Raining Atmospheric Condition	-2.874e-01	1.732e-01	-1.659	0.097051

AIC	Null Deviance	Residual Deviance
5991	7609.4 (5488 df)	5907.0 (5447 df)

4.3.0 Initial Setup to Construct Predictive Models

In order to construct the 3 predictive models (i.e. elastic net logistic regression, random forest and knn), we require a balanced data set. Therefore, the observations that had no fatality were downsampled until the data set contained an equal number of fatal and non fatal observations. This is identical to the process in section 2. After the data set was balanced, an 80/20 split was conducted on the data set to create a training and test set.

For the elastic net logistic regression and knn models, the numerical predictors in the training and test data set were standardised so that their sample variance was 1.
Standardisation is required in elastic net logistic regression because the size of the coefficients of the predictors are restricted based on the values of the variables.
As for knn, standardisation is required because the algorithm calculates the distance between data points. If there is no standardisation, predictors with higher value ranges dominate when distances are calculated.

4.3.1 Results of the Elastic Net Logistic Regression Model (Selected Predictive Model)

Covariates	β estimates
Intercept	-0.638591499
SEXM	0.488622271
SEXU	-1.091795710
AGE	0.102699404
LICENCE_STATEVictoria	-0.094134639
Seat Belt not worn	1.800690649
Seat Belt worn	0.176662360
VEHICLE_YEAR_MANUF	-0.095766865
DC UTE Body Style	-0.327909652
Other Body Style	0.181531598
P MVR Body Style	0.573539480

S WAG Body Style	-0.020478110
SED Body Style	0.120637013
SEDAN Body Style	0.021907440
TRAY Body Style	-0.147294159
UTIL Body Style	-0.531131086
VAN Body Style	-0.003260503
WAGON Body Style	-0.055893579
HOLDEN Make	0.328071945
HONDA Make	-0.234350417
HYUNDAI Make	0.083416567
KENWORTH Make	0.996775720
MAZDA Make	-0.088718663
MERC B Make	0.008392562
MITSUBISHI Make	0
NISSAN Make	-0.058563230
Other Make	0.081289351
SUBARU Make	-0.207006118
TOYOTA Make	0.192073563
VOLKS Make	0.043034911
Heavy Vehicle > 4.5 Tonnes Vehicle Type	1.255013392
Other Vehicle Type	0.497117152
Panel Van Vehicle Type	0.249155557
Prime Mover - Single Trailer Vehicle Type	0.707209614
Station Wagon Vehicle Type	0.163504682
Taxi Vehicle Type	-0.395607277
Utility Vehicle Type	0.870541251
Gas Fuel Type	-0.058349201

Multi Fuel Type	-0.019064061
Other Fuel Type	0.022543130
Petrol Fuel Type	-0.152178118
Blue Vehicle Colour	-0.025720332
Green Vehicle Colour	0.129530545
Grey Vehicle Colour	-0.224284379
Other Vehicle Colour	0.068281998
Red Vehicle Colour	0.072447346
Silver Vehicle Colour	0.002358258
White Vehicle Colour	0.041392389
Total Number of Occupants	0.275910783

4.3.2 Interpretation of the Selected Predictive Model (Elastic Net Logistic Regression Model)

The following interpretations were derived from the elastic net logistic model in an identical way to how it was computed in [Appendix Section 4.2.0.4].

SEX: In Victorian roads, accidents with male drivers have **1.63 times** the fatal odds of accidents with female drivers. This means that accidents involving male drivers have a higher likelihood of being fatal.

AGE: An increase of one year in the age of the driver involved in an accident will increase fatal odds by **1.11 times**. This suggests that as drivers get older, the odds of fatality in an accident is higher.

LICENSE_STATE: Drivers involved in accidents with a licence state outside of Victoria have fatal odds that are **1.1 times** of drivers with a Victorian licence. This implies that drivers with licence states outside of victoria are more likely to cause fatal accidents.

SEATBELT_WORN: Drivers that do not wear seat belts have **5 times** the fatal odds of drivers that wear seat belts in the event of an accident.

VEHICLE_YEAR_MANUF: With a one year increase in the vehicle's manufacturing year of the driver involved in the accident, the fatal odds decrease by **9.1%**. Here, the model implies that older vehicles have a higher chance of resulting in fatality in the event of an accident.

VEHICLE_BODY_STYLE: The P MVR Body style has the higher fatal odds out of any other body style type by large margin. This can be observed by the size of its coefficient relative to the coefficient of other body styles. For example, the fatal odds of a P MVR body style vehicle in the event of an accident are **1.48 times** that of Other body styles (i.e. the second highest coefficient out of body styles). The P MVR body style is that of a truck, suggesting that trucks may cause more fatal accidents according to the model.

VEHICLE_MAKE: Kenworth vehicle make has the highest fatal odds of all vehicle make types. In comparison to the Mitsubishi make, the kenworth make has **2.71 times** the fatal odds in the event of an accident. Kenworth is an American truck company, therefore the model suggests for the second time that trucks result in higher fatalities in the event of accidents in Victorian roads.

VEHICLE_TYPE: Heavy vehicles above 4.5 tonnes have the highest fatal odds in accidents on Victorian roads. As opposed to cars, accidents involving heavy vehicles have **3.4 times** the fatal odds. Since heavy vehicles tend to be vehicles such as trucks, we observe a recurring theme that trucks result in higher odds of fatality in accidents. This can be more generalised to heavy vehicles.

FUEL_TYPE: Vehicles operating with fuel types other than multi, gas, petrol or diesel result in the highest fatal odds in accidents. This can be seen by other fuel types having the highest coefficient as opposed to the rest of the fuel types.

VEHICLE_COLOUR: Greens vehicles result in the highest odds of fatality in an accident to those involved, having **1.14 times** the fatal odds of black vehicles.

TOTAL_NO_OCCUPANTS: With one more occupant in a vehicle, the fatal odds in an accident change by **1.3 times**. This suggests that if more people are occupying a vehicle, the odds of fatality in an accident are higher.