ACTL3142: Statistical Machine Learning for Risk and Actuarial Applications

z5417266

July 30, 2023

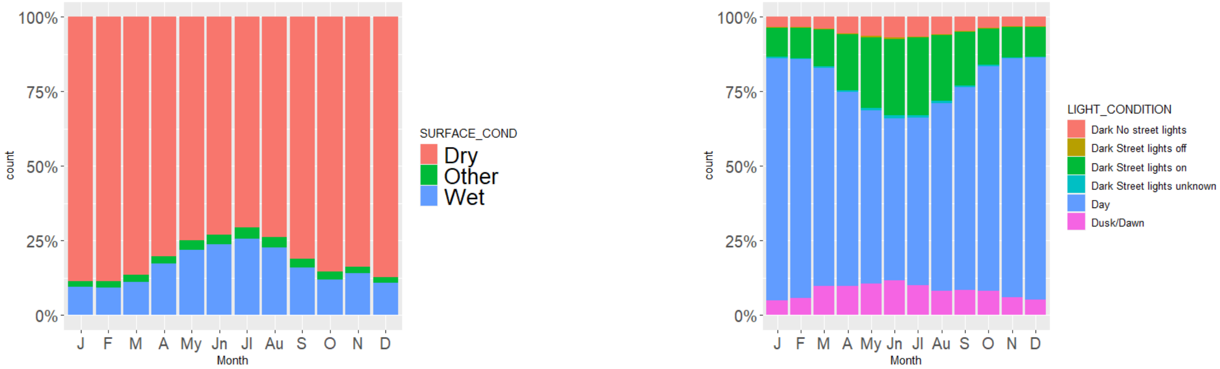
Contents

Executive Summary

This report is written with the intent of relaying the nature of fatal accidents on Victorian roads and possible proactive measures that can be campaigned to foster a safer and more careful conscience towards road safety. In the preliminary stage of this report, an exploratory data analysis (EDA) will be performed on road accident data from 'VicRoads', the Victorian government agency responsible for road safety and management. The dataset provides details on the characteristics of the driver, the vehicle, and the conditions of the crash, along with a flag indicating a fatal crash. A Generalised Linear Model (GLM) will be fit to the data to help determine relationships between the covariates given and the chance of an accident causing fatality. Finally, a predictive model will be constructed based on driver and vehicle characteristics that can aid in guiding focus areas of preventative measures to specific demographics at risk.

Part I: Exploratory Data Analysis

The goal of exploring this data is to uncover whether the facets surrounding the accidents follow trends or patterns and to explain these trends. Several new variables were created and many unnecessary ones were removed prior to this analysis. The new variables had the purpose of tidying up existing variables, for example, time of day was converted to 'hour', and year of manufacturing with accident year were used to create vehicle age. Variables removed included un-descriptive data such as identification variables, and postcodes were refactored by region [1].

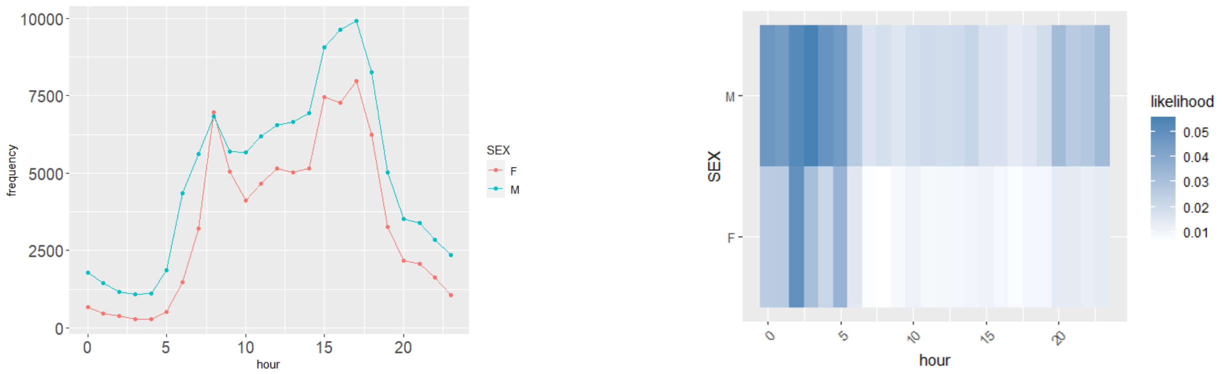


(a) Proportion of Surface Condition

(b) Proportion of Light Condition

Figure 1: Stacked bar charts depicting the proportion of accidents with respect to month

An idea that should be established in this EDA is that the data sourced is heavily reliant on the rarity of events. This includes anything from the age of commuters to the traffic size on weekends. For example, out of all accidents, 16.2% occur under wet surface conditions and 16.8% in 'Dark street lights on' conditions. It cannot be concluded from this data that these conditions are more or less likely to cause accidents relative to other conditions. This is dependent on the sampling size and chance of these conditions occurring for any single person in a population, which is not given. This is illustrated in **Figure 1** which shows that each of the aforementioned conditions are more likely during accidents in Winter months when there are fewer daylight hours and higher precipitation rates. This is not to say these conditions do not contribute to accidents. Rather, this points towards the fact one-dimensional analysis of frequency is not sufficient to deduce the causes of accidents but can relay the probability of such conditions given an accident has occurred. This relates to all variables in the data set, as the population sizes are not shared. However, assuming the motorist population is evenly split between males and females, men are over-represented in crash statistics accounting for 58% of the data, and further given a crash has occurred, males account for 78.6% of fatalities. Statistics showing over-representation, especially in fatality rates, can signify key demographics for targeted prevention.



(**a**) Line graph of accident by time and sex

(**b**) Heatmap by time and sex

Figure 2

The next insight relates to nighttime drivers who are prone to fatigue. A stark pattern is that in **Figure 2 (a)**, the proportion of drivers who experience accidents in the early morning hours are overwhelmingly male. Further, despite having one of the lowest incident rates, **Figure 2 (b)** shows this group has the highest fatality rate with respect to time of day and sex, possibly due to fatigue, a low-light level, or drug-fuelled driving. Another insight on accidents during dawn is that weekends are overrepresented, where according to **Figure 3**, about half of the dawn accidents are on Saturday and Sunday mornings. This could be related to higher usage of roads after Friday and Saturday night events and could be inflated by the prospect of driving under the influence of drugs such as alcohol.

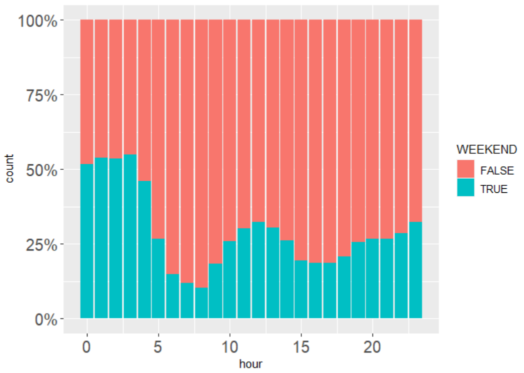
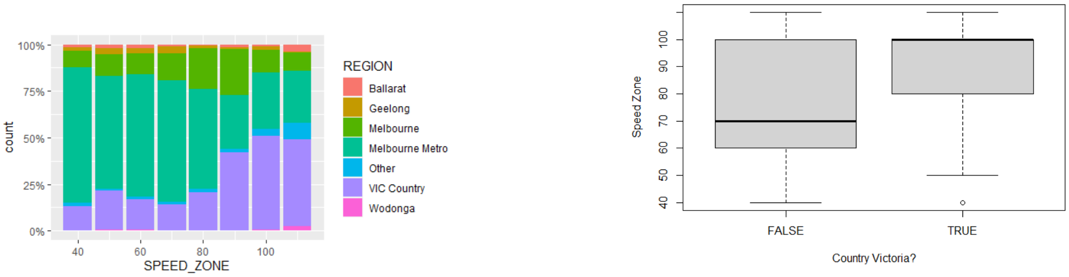


Figure 3: Proportion of weekend by hour

This last insight relates to the region the owner of the vehicle resides in and the speed zone where the accident occurs. **Figure 4 (a)** shows that the 'VIC country' region accounts for more accidents in faster speed zones, and **Figure 4 (b)** depicts that fatalities of rural motorists occur in faster speed zones (100km/h median) compared to the rest of Victoria (80km/h median). This combination of increased frequency and fatalities of rural vehicle accidents in fast speed zones can be attributed to the proportion of those speed limits in country areas [2], as well as casual and excessive speeding on rural roads. This could be a catalyst for the higher observed overall fatality rates for rural commuters. Additional insight on the definition of fatality is shared in the **Exploratory Data Analysis Appendix**, where it is postulated fatalities may not only refer to drivers and passengers.



(a) Stacked bar of Speedzone and region

(b) Boxplot of fatalities by rural status

Figure 4

Part II: Creating a Relational Model

This section will detail the result of using a generalized linear model (GLM) for the purpose of inference. Specifically, the relationship between the fatality of an accident given its occurrence will be explained by several predictors, and their insights will be shared. The main goal of this section is to balance both statistical validity and ease of interpretation to highlight the links between fatal accidents and their circumstances. Despite the effectiveness and predictive prowess having a less significant role in inference, its presence is still important as a model that cannot produce accurate and desirable outcomes is one that should not be interpreted due to a lack of intrinsic credibility. For details of the construction and selection of the final relational model, refer to the **Technical Appendix: Relational Model**.

In summary, a Logistic Regression GLM was constructed based on the fact the errors of the response are binary and a canonical logit link function was implemented. Techniques used to find the model of best fit were employed such as stepwise selection and the Wald Test. However, the imbalance in the response variable led to the model becoming hesitant in predicting any accidents as fatal. Despite the strong accuracy and ROC-AUC due to the overwhelming true negatives, alternative methods to model and assess were developed. **Table 4** outlines all the models and imbalanced learning algorithms considered and their metrics when applied to a validation set.

Through careful examination of this menu of models, considering complexity, optimal F_1 and precision-recall, two cost-sensitive models were selected and their deviances were compared. The deviance of the more parsimonious model was smaller, so the complex model was rejected. This model was further shrunk through Lasso regression and the 'one standard error rule' to ensure a simple yet credible model. Overall, the final model came to 27 predictors in **Table 1**. Benchmark factors and coefficients for all categorical variables are found in the appendix **Table 5**.

To begin analysis, it should be established that these coefficients indicate how the log odds of a fatality change based on each predictor. Since the logarithmic operation is monotonic, the direction in which the probability and odds of a fatality change is the same direction as the log odds. In the case of dummy variables, such as gender or accident type, the coefficients are with respect to the baseline factor. For example, the coefficient for males is 0.337, so the expected log odds of having a fatality increase by 0.337 if the driver is a male over the alternative of a female driver (baseline). This is backed up by the **EDA (Part I)** which found that males are overrepresented in fatalities (78.6%).

(Intercept)	Male	Aged 70+	Seatbelt not worn	No. Occupants	Speed Zone
-5.4180	0.3377	0.3261	1.1062	0.0884	0.0367
Coll. other object	Coll. vehicle	No coll.	Struck animal	Struck Pedestrian	Overturned
-0.3625	-0.2917	-0.0317	-0.5035	1.2728	-0.4221
Other Surface	Other Atmosphere	Raining	Paved Road	Melbourne Metro	VIC Country
-0.2194	-0.1516	-0.1003	0.0046	-0.0704	0.2394
Prime Mover	Kenworth Make	Heavy > 4.5T	Single Trailer	Petrol	Weekend
0.2885	0.2239	0.8650	0.2401	-0.1346	0.0078
Vehicle Age	Day	Dusk/Dawn	Not at intersection	-	-
0.0012	-0.2436	-0.1814	0.1910	-	-

Table 1: Predictors of Final Relational model with coefficients to 4dp

The intercept for this GLM is -5.418, meaning by default, the probability of a driver having an accident is $\frac{\exp(-5.418)}{1+\exp(-5.418)} = 0.0044$. This serves as an anchor for the probabilities predicted by the logistic regression and the coefficient is highly negative due to the majority class of non-fatals skewing overall predictions.

To report some of the strongest influences contributing to fatalities, not wearing a seatbelt compared to wearing a seatbelt (1.106) and striking a pedestrian over colliding with a fixed object (1.273) have the most positive influences on the log odds of fatality. The former highlights the importance of in-built vehicle safety features that should be promoted in usage and properly enforced, while the latter confirms suspicions in the **EDA appendix** that the fatality variable applies to all parties, not only drivers.

The importance of baseline factors is elucidated by the accident type variable. The baseline as aforementioned is colliding with a fixed object, the coefficients are negative for all other types of accidents bar striking a pedestrian. This does not mean these types of accidents decrease the log odds of a fatality in an absolute sense, rather the log odds fall when compared to the benchmark factor.

Other notable influences include driving a rigid vehicle heavier than 4.5 tonnes compared to a car (0.865), the speed zone (0.0367 per km/h), the number of occupants (0.088 per extra occupant excl. driver), and a rural driver compared to a Melbourne driver (0.239). These all relate to possible intensifiers in the severity of crashes. Heavier vehicles and faster speeds mean a greater impact on collision, more occupants means a greater chance of at least one fatality, and driving on rural roads is correlated with faster speed zones due to greater distance between locations as found in the **EDA** [2], pressuring fatalities upwards.

Lastly, some other noteworthy coefficients are vehicle age (0.0012 increase per year) and being aged seventy years or older compared to the benchmark of sixteen to eighteen years (0.326). This could signal that elderly drivers are the most vulnerable age group on roads, with vehicle age playing a part due to older cars featuring fewer safety considerations, higher chances of faults, and weaker protective fortitude.

Limitations of this model may include the fact that the model is relatively under-dispersed so it may not account for a wide enough variation in the mean response, and the optimal decision threshold should ideally be closer to the natural threshold of 50%, however it is closer to 30%. Furthermore, this model involved a Lasso, so traditional GLM Wald Test p-values cannot be considered, however, it is assumed that using the 'one standard error' rule would ensure that all significant predictors are accounted for.

Part III: Creating a Predictive Model

For this stage of the report, a prediction model will be developed and applied to `Drivers_Eval.csv` to determine which groups of people are predisposed to fatal accidents. Compared to **Part II**, the given variables are purely based on driver and vehicle characteristics, and there is less emphasis on interpretability. Since the goal of this endevour is to target the $2,500$ most likely to experience a crash out of $10,000$ the main concern when evaluating models should be how well the model avoids misclassifying actual fatalities when considering the $2,500$ most likely predictions by the model. In other words, when the models are used for validation, what matters most is how well the model captures true fatalities and avoids false non-fatalities when the threshold is set such that it predicts 25% of the set as having a fatal accident.

For details on the construction of alternative models, refer to the **Technical Appendix: Predictive Model**. In summary, using the decision criteria used above in conjunction with other measures like precision and recall, classification and rebalancing techniques were used in combination to produce several models, including Logistic Regression with Lasso and Ridge, KNN, Tree Pruning, Random Forest and Boosting. Overall, using a validation set approach, it was found that oversampling the data and applying forward stepwise GLM selection, using stratified k-fold cross-validation maximising PR-AUC as selection criteria, produced the best recall (TPR, priority metric) for the purpose of this prediction. Then, a ridge regression was applied with the optimal λ for misclassification, in order to offset the effect of overfitting as a result of oversampling, then the final model was trained on the entire data set (oversampled train, validation, and test), and then applied to the unseen data. Assuming the measured validation and test recall are relatively accurate estimates, around half of the fatalities in the dataset will be captured in the set of $2,500$.

Using these predictions produced by the model, certain target demographics can be determined as necessary candidates for road safety and fatality prevention campaigns. While it is appreciated that this model does not capture all fatalities, and most selected accidents will not be fatal, it has elucidated trends corroborated by both the **EDA** and **Part II**. Furthermore, vital predictors in **Part II** such as speed zone and type of accident are not available and their absence has likely reduced the potency of the model.

To uncover these trends, the composition of categorical variables in the total dataset will be compared to the composition of the 25% most likely to have an accident. **Figure 5** highlights three of the most significant deviances in composition. The most drastic one is men comprising 58.65% of the dataset, yet 84.52% of predicted fatalities. Recall from **Part I** that a very similar occurrence happened where 58% of the original dataset were males while taking up 78.6% of fatalities. This indicates a possible bias towards male drivers in the final model used or may be circumstantial on the data points in `Drivers_Eval.csv`. Furthermore, VIC Country accidents grew in proportion from 23.17% to 62.4%. The danger of rural driver accidents was also echoed in **Part I** with the analysis in relation to speed zone and **Part II** through the fitted coefficients. Moreover, the proportion of drivers not wearing seatbelts also increases from 2.06% to 7.68% as reinforced in **Part II** in the analysis of coefficient estimates.

Analysing the drivers that the final model predicted would have fatal accidents with near certainty could reveal patterns in both the underlying decision-making of the model and indicate priorities when determining methods of prevention. **Table 2** shows a few of the most extreme predictions of fatalities. The commonality between all these observations is that they are rural drivers and involve the lack of seatbelt wearing. Further, a large portion of the vehicles are 20th century makes possibly indicating poorer safety and structural fortitude. The accident with the highest chance of fatality could be an instance of the model requiring a refactor, as the number of occupants, 27, seems to drag the probability higher which is a positive coefficient. Furthermore, many of the highly predicted accidents involve P MVR vehicle types typically of Kenworth make, with one example included in **Table 2**, supported by **Part II** as features of heavy vehicles share highly positive coefficients. Lastly, diesel cars occupy many of the top positions. As stated in the **Technical Appendix: Relational Model**, this may be a confounding variable associated with heavier vehicle types that have diesel engines and are involved in more fatalities. As a result, it may overestimate the probability of fatality for diesel engines in general.

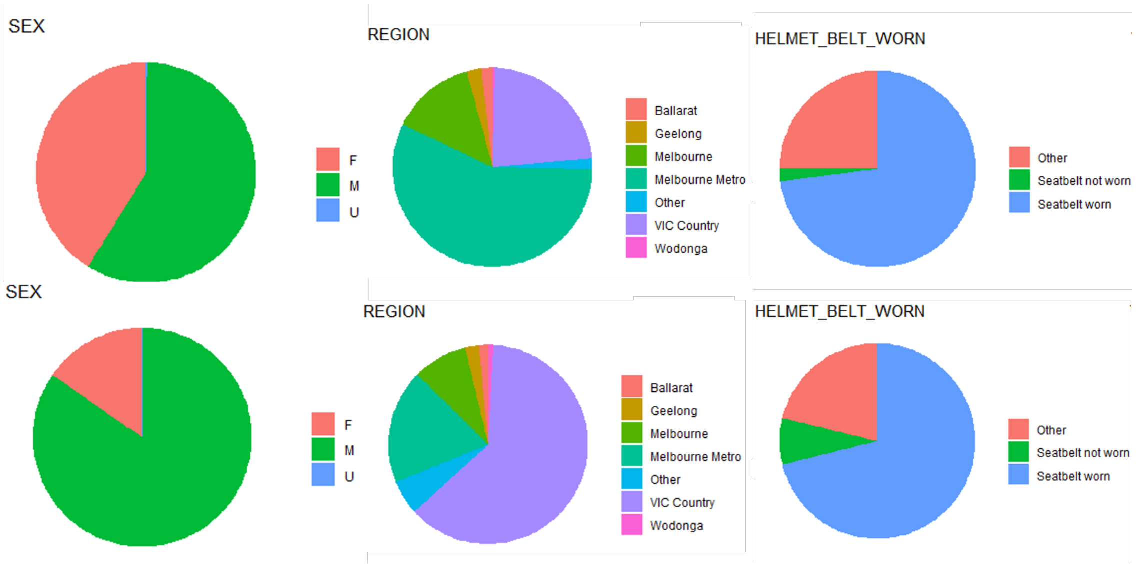


Figure 5: Composition of `Drivers_Eval.csv` (top row) vs. 25% most likely to be fatal (bottom row)

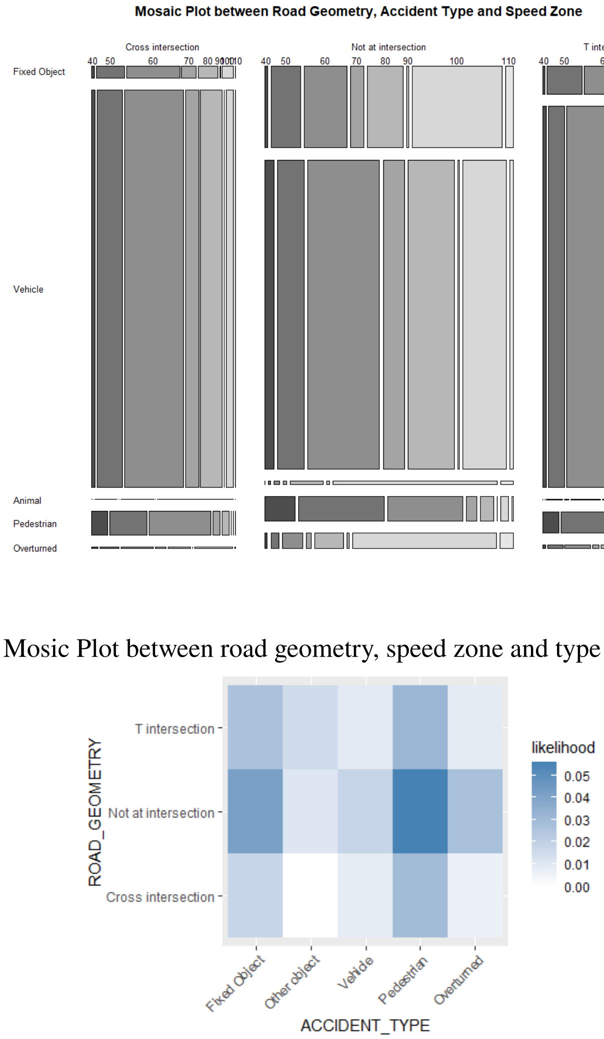
Sex	Age	Vehicle	Make	Fuel	Seatbelt	Year Manuf.	Region	No. Occupants	Prob.
M	39	Other	Mistub.	Diesel	Other	2006	Geelong	27	0.998
M	59	Other	Merc.	Diesel	Other	1996	Melbourne	13	0.964
M	66	Van	Other	Diesel	No	1996	Country	1	0.958
M	21	Car	Holden	Diesel	No	1997	Country	6	0.955
M	66	P MVR	Kenworth	Diesel	Yes	1999	Country	1	0.946
M	77	Utility	Nissan	Petrol	No	1991	Country	1	0.9098
F	25	Car	Holden	Petrol	No	1999	Country	4	0.862
F	26	Utility	Holden	Diesel	No	2017	Country	1	0.833

Table 2: Cross-section of most likely candidates for fatalities

Conclusion

In summation, through extensive exploratory and modeling ventures, drivers who are male, are from rural areas, operate older heavy-duty vehicles, and have a habit of not wearing a seat belt face the highest chance of being involved in a fatality. Drivers of these demographics are highly advised to exercise caution on Victorian roads and should be subject to increased road safety awareness. Common safety protocols such as seatbelt wearing should be promoted to prevent excessive fatalities and commuters of rural areas are recommended as the primary focus group for a new road safety campaign.

Exploratory Data Analysis Appendix



(a) Mosic Plot between road geometry, speed zone and type of accident

(b) Heatmap by geometry and type of accident

Figure 6

An additional insight regards road geometry and its relationship with the present speed zone, and type of accident. The **Figure 6 (a)** mosaic plot shares proportions from each group. Something interesting is that 'not at intersection' accidents have higher concentrations across the faster speed zones compared to intersection accidents since the latter occur in speed zones that are appropriate for safety protocols that enforce stoppage such as traffic lights. Further, accidents such as pedestrian and vehicle collisions occur in slower speed zones than fixed object collisions and overturned vehicles. As for fatalities (**Figure 6 (b)**), 'pedestrian' accidents produce the highest rates of fatalities across the accident types. This most likely verifies that the 'fatal' variable applies to all involved parties (including pedestrians) and is not reserved for the driver and passengers only. The high fatalities from 'not at intersection' and 'fixed object' collisions could be due to the higher speeds as witnessed in **Figure 6 (a)** for these types of accidents, and that fixed object collisions most likely required the vehicle to be maneuvered towards the object in a 'head-on' setting which causes the greatest change in speed and impulse.

Technical Appendix: Relational Model

To construct a GLM, it must first be determined what the shape of the response variable is. In this case, the target is whether the accident was fatal or not. This is a binary outcome, hence this falls under the classification/logistic form of GLM. Other GLMs such as Regression (Gaussian) and Poisson shall not be considered due to the possibility of negative responses from the former, and the fact that variance and mean are not proportional from the latter. There are 3 main components to the GLM:

1. The Systematic Component: Let η be a linear predictor of the form

   $$\eta_i = x_i'\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = g(\mu_i),$$

   where $x_{i1}, \ldots x_{ip}$ are the ith observation of the explanatory predictors, g is the link function, and μ_i is the mean of the response of the ith observation. This function g is the connection between the linear predictors and the location of the response, in other words, the mean observation can be expressed as a function of the linear predictors $\mu_i = g^{-1}(\eta_i) = g^{-1}(x_i'\beta)$ in order to model the relationship between predictors and expected response.

2. The Stochastic Component: The errors in our model take the form of the Bernoulli distribution due to the binary response variable. Assuming all observations are independent, the response, Y, will have a density from the exponential dispersion family as shown below with mean μ.

   $$\begin{aligned} f_Y(y) &= \mu^y (1-\mu)^{(1-y)} \quad \text{for } y = 0, 1 \\ &= \exp(y \log(\mu) + (1-y)\log(1-\mu)) \\ &= \exp(y(\log(\mu) - \log(1-\mu)) + \log(1-\mu)) \\ &= \exp\left(y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right) \\ &= \exp\left(y\theta - \log\left(1 + e^{\theta}\right)\right) \\ &= \exp\left(\frac{y\theta - b(\theta)}{\psi} + c(y;\psi)\right) \end{aligned}$$

   where $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, $b(\theta) = \log(1+e^{\theta})$, $\psi = 1$ and $c = 0$.

3. Link Function: In the case of a Bernoulli response, the following is derived by assuming there is a canonical link between the systematic and stochastic components of the GLM, that is

   $$\theta(\mu_i) = g(\mu_i) = \eta_i = x_i'\beta.$$

   Ultimately, the GLM can be described by the equations

   $$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

   $$E[Y_i] = \mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

In terms of interpretation, a unit increase in the predictor x_i translates to an expected β_i increase in the log odds of the response or the log odds of fatality in this case. This has the same signed effect as a change in odds due to the monotonic properties of logarithms.

Now a Generalised Linear Model can be fitted to the data using the `glm` function in `R`. In order to evaluate the model, a split between training and validation sets was first established. After employing variable selection methods such as forward and backward step-wise subset selection (`regsubsets` apart of `leaps`) using metrics such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Adjusted R-squared, and the Wald Test which removes statistically insignificant covariates, a problem was encountered.

	Reference False	**Reference True**
Predict False	39246 (TN)	673 (FN)
Predict True	7 (FP)	4 (TP)

Table 3: Confusion Matrix of Forward AIC model

At a 50% probability threshold used on the validation set, the natural decision boundary, the models were hesitant at classifying any accidents as fatal, for example in **Table 3**, only 11 accidents out of $39,930$ were classified as fatal. The accuracy of the model is high (98.3%) because it favours classifying the majority class. Furthermore, the Area Under Curve (AUC) of the ROC for the model used in **Table 3** is 81.48, meaning the model has a healthy trade-off between the False Positive and True Positive rates. However, these metrics often give an optimistic view of a model's performance in an imbalanced environment [3] [4]. For the ROC, which describes the trade-off between the False positive and True Positive rates as the decision threshold varies, the False Positive rate FP/(FP + TN) is pulled down by the large number of True Negatives stemming from the imbalanced data, and a lower number of fatal predictions.

To select the best model and draw relationships in the data, the problem of imbalanced data should be discussed. This notion of imbalance in the number of fatalities to non-fatalities, which is approximately a ratio of 1 to 58, is not as impactful at this stage of the report, however, there were several methods considered and proposed by He and Garcia (2009) [4] used to reduce the effect of imbalance and appropriate evaluative metrics to assess the efficacy of models under imbalance.

1. Stratified k-fold Cross Validation: This involves k-fold cross-validation but ensures each fold retains the same proportion between the minority (fatal) and majority (non-fatal) classes. The imbalance in data is still present so the only benefit is that each fold has the same shape as the original data set which reduces the bias of each fit.

2. Adaptive Synthetic Sampling (ADASYN): This is an algorithm that interpolates artificial data points for the minority class by considering the k-nearest neighbors. This balances the dataset, however, due to the high dimensionality of the data, the Euclidean distance formula becomes too varied and may create noise. Tomek Links should be used to reduce overlapping synthetic data points and noise to better define data clusters, however, this was too memory intensive [4] (p. 6).

3. Oversampling and Undersampling: These algorithms increase samples from the majority class and reduce samples from the minority class respectively when creating balanced training sets. Each house their own problem [4], the former can often lead to overfitting while the latter removes

information. A way to combat this is to oversample only a portion of the training data and leave the remaining intact, then combine the two sets to create the full training set which could avoid overfitting.

4. Cost-Sensitive Learning: This method involves attaching costs to misclassifying classes and attempting to minimise these costs. For example, if the cost of misclassifying a fatal accident incorrectly is greater than misclassifying a non-fatality, then the GLM will be more liberal in predicting fatalities. A form of cost-sensitive learning may be dataspace weighting, which essentially bootstraps the minority class without replacement to select the best training distribution [4]. For this report n-cost means that the cost of misclassifying fatalities is n-times higher than misclassifying non-fatalities.

In collaboration with these techniques, metrics proposed by [3] and [4] were used to evaluate the models instead of accuracy and ROC-AUC. These include

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

the F_1 measure which is the harmonic mean of Precision and Recall, as well as the area under a Precision-Recall (PR) curve (**Figure 7**, a higher AUC is preferred). These quantities are focussed on analysing the minority class rather than the more frequent majority class. Furthermore, since interpretability is a major goal of this section, the optimal decision threshold should be considered. Optimal thresholds closer to 50% are ideal since this is the natural decision boundary where someone may interpret a score above as a ’true’ prediction, and a score below as a ’false’, and the coefficients should be aligned such that estimates around this boundary are close to optimal.

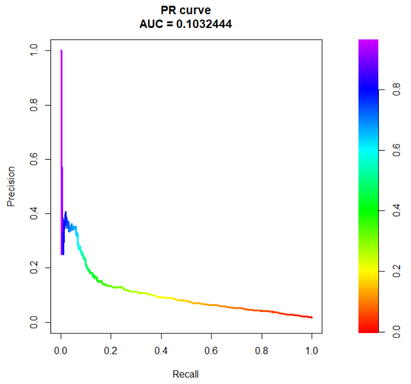


Figure 7: Precision Recall Curve of Final Model

Using all this information, **Table 4** presents a menu of models from which the most ideal can be selected for the purpose of presenting a model that accurately fits the data as well as providing correct interpretation free from imbalance. To reiterate, the purpose of this section is to balance validity and parsimony. When dissecting **Table 4**, it seems that no-cost, ADASYN, and Oversampling Techniques produce non-ideal models whether it is because of too many predictors, non-ideal boundaries or poor F_1 and PR-AUC. The 5-cost and 10-cost models provide possible optimal solutions, so it may be fair to select one type of cost each to proceed onward with. From the 5-Cost models, the Backward BIC (Model 1) model will be chosen as it houses the least predictors, highest AUC, and F_1 from that set. From the 10-Cost models, the choice is much more difficult however, the Backward BIC (Model 2) model will be selected. This model is more well-rounded than its contender which was the 10-Cost Forward BIC model.

To choose between each model, residual deviance will be considered where Model 1 is nested into Model 2. The deviance of Model 1 is 77205, while the deviance of Model 2 is 122394. Since the deviance of the more parsimonious model is lower, indicating a better fit to the data, Model 1 will be selected. It should be noted the scaled deviance (deviance divided by Degrees of Freedom) of Model 1 is 72205/159670 = 0.46 while the scaled deviance of Model 2 is 122284.5/159665 = 0.766. This indicates Model 1 is underdispersed relative to Model 2. Model 1 will continue but the underdispersion factor is still a limitation.

Next, regularisation and shrinkage of the chosen model will be used to select the variables which are most relevant in contributing to the fatality of an accident. Lasso regression will be used over Ridge regression as the former has the effect of variable selection due to the nature of the L1-norm penalty, which improves the interpretability of the model by eliminating the least relevant variables while Ridge only decreases the size of the coefficients which may actually worsen raw interpretability. Lasso Regression was chosen over eliminating predictors with high p-values as using cost-sensitive learning amplified the significance of all trained variables due to overfitting, hence the p-values were difficult to interpret.

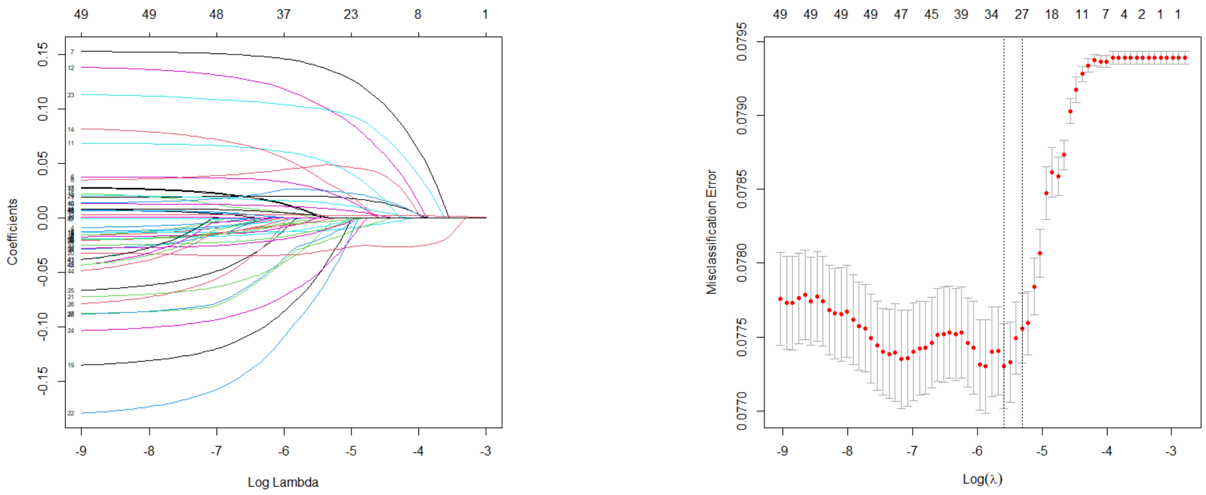
Table 4: Menu of GLMs

Model	**Max F_1**	**Optimal threshold (%)**	**AUC-PR**	**No. Pred**
Backward AIC	16.4	8	10.76	54
Backward BIC	16.7	8	10.89	31
Forward AIC	16.37	8	10.72	54
Forward BIC	16.82	8	10.72	27
5-Cost Backward AIC	16.83	28	10.34	71
5-Cost Backward BIC	16.84	30	10.32	49
5-Cost Forward AIC	16.88	30	10.31	74
5-Cost Forward BIC	16.83	30	10.27	50
10-Cost Backward AIC	16.64	51	10.20	76
10-Cost Backward BIC	16.64	57	10.21	54
10-Cost Forward AIC	16.77	51	10.31	78
10-Cost Forward BIC	16.84	50	10.16	56
100% ADASYN Backward AIC	16.93	84	10.19	98
100% ADASYN Backward BIC	16.61	84	10.16	87
50% ADASYN Backward AIC	16.31	78	9.92	89
50% ADASYN Backward BIC	16.89	81	10.10	72
50% Oversampling Backward AIC	16.93	76	10.10	97
50% Oversampling Backward BIC	16.72	76	9.7	74

To choose the optimal level of shrinkage, stratified k-fold cross-validation with the correct weighting will be performed on the entire data set for each level of the tuning parameter λ. As seen in **Figure 8**, the log-lambda value which minimises misclassification is -5.59 and reduces the model to 34 variables. To further improve parsimony, the 'one standard error rule' can be applied to find the λ which is one standard deviation above the value of lambda that produces the minimised misclassification. This value of log-lambda is -5.31 and shrinks the model to 27 variables.

Additional analyses: A discovery that may come across as unexpected is that the coefficient for the 'weekend' variable is very small with an increase in 0.00782 log odds if the accident took place on a weekend. This is surprising as in the **EDA**, it was found weekend accidents overrepresent earlier morning hour accidents which tend to be more fatal, however, this may be offset by a low number of observations in this group not significantly influencing the coefficient with leverage.

A peculiar observation is that petrol engines are less likely to be involved in fatalities than diesel engines. This may be due to a confounding effect where the majority of larger vehicles which have positive effects on fatality contain diesel engines, however, diesel engines themselves do not contribute towards fatalities.



(a) Trajectory of coefficients by shrinkage

(b) Misclassification by shrinkage

Figure 8

(Intercept) - -5.4180	**Male** Female 0.3377	**Aged 70+** 16-18 0.3261	**Seatbelt not worn** Seatbelt on 1.1062	**No. Occupants** - 0.0884	**Speed Zone** - 0.0367
Coll. other object Fixed Object -0.3625	**Coll. vehicle** Fixed Object -0.2917	**No coll.** Fixed Object -0.0317	**Struck animal** Fixed Object -0.5035	**Struck Pedestrian** Fixed Object 1.2728	**Overturned** Fixed Object -0.4221
Other Surface Dry -0.2194	**Other Atmosphere** Clear -0.1516	**Raining** Clear -0.1003	**Paved Road** Gravel 0.0046	**Melbourne Metro** Melbourne -0.0704	**VIC Country** Melbourne 0.2394
Prime Mover Coupe 0.2885	**Kenworth Make** Ford 0.2239	**Heavy > 4.5T** Car 0.8650	**Single Trailer** Car 0.2401	**Petrol** Diesel -0.1346	**Weekend** Weekday 0.0078
Vehicle Age - 0.0012	**Day** No lights dark -0.2436	**Dusk/Dawn** No lights dark -0.1814	**Not at intersection** Cross Intersection 0.1910	- - -	- - -

Table 5: Predictors of Final Relational model with coefficients to 4dp with their baseline

Technical Appendix: Predictive Model

In terms of data preprocessing, a training-validation-test split of 80-10-10 was first made, and the range of variables that could be considered for modeling were made to be in line with `Drivers_Eval.csv`. Before delving into the models used, the selection criteria and ways to address the imbalanced data will be discussed foremost.

For the purpose of this prediction, since the goal is to target the $2,500$ most likely to have an accident out of $10,000$, judging the models based on a decision threshold where 25% of observations are classified as fatal on the validation set would best simulate the environment the final predictions would be scrutinised within. Criteria discussed in **Technical Appendix: Relational Model** such as recall, precision and PR-AUC were used, however, recall (True Positive Rate, TPR) was given priority as it measures the performance of the model in accurately capturing the minority (positive) class within the top 25%. Furthermore, the same imbalance reduction techniques were applied, however, some adjustments were made specific to each type of model which will be further discussed. To reiterate, the main imbalance reduction techniques considered were ADASYN, oversampling, and cost-sensitive weights, and variations of these techniques were also tested such as resampling only proportions of the training data and adjusting weights. Further details on these methods are elaborated in **Technical Appendix: Relational Model**.

Several classification techniques were considered in developing a predictive model:

1. Generalised Linear Model (Logistic Regression) with Regularisation: This type of classification largely remained faithful to its previous usage in **Part II**. However, new techniques to find the best predictive model were employed. Instead of restricting GLM selection to goodness-of-fit criteria such as AIC and BIC, stratified k-fold cross-validation was used in collaboration with forward and backward stepwise techniques to find the optimal predictors. However, instead of estimating the test error through cross-validated Mean Squared Error, PR-AUC was also considered. These models had the most consistent successes. Once a few optimal ones were found, they were regularised with Ridge and Lasso Regression which shrinks the sizes of the coefficients to a level that minimises errors. Ridge regression overall gave better recall and predictive power compared to Lasso, Furthermore, Ridge regression mitigates the issue of overfitting which is a symptom of most class imbalanced reduction techniques such as oversampling.

2. k-Nearest Neighbours (KNN): This technique was perhaps the most neglected. The main issue was that KNN does not have an established method to determine the probability of an accident becoming fatal, it purely performs classification and cannot be adjusted for more or less positive class predictions without changing the value of k. Furthermore, the dataset is high dimensional due to the number of predictors, so Euclidean distance is more varied and unreliable. Lastly, due to the nature of imbalanced data, higher values of k would lead to many more predictions for the majority class, while low values of k, which are better for imbalanced data, lead to severe overfitting. Overall, this technique was too risky and not worth spending too much attention on. Experimentation with KNN did not yield positive results.

3. Tree Pruning: The `rpart` function as part of the eponymous library can be used to grow decision trees and control the complexity of the tree through pruning. This method has been shown to perform poorly on imbalanced datasets as it may remove terminal nodes from the minority concept if the cost parameter is too high, and overfit if too low.[4] (p. 10). Imbalanced learning techniques

were applied to prevent this. Overall, applying improvement to data balance improved performance very slightly when tuning for the optimal complexity parameter, however, it was never as consistently performative as GLM. The choice between Gini Index and Entropy as splitting criteria is mostly insignificant [4] (p. 10).

4. Random Forest and Bagging: Both these types of tree extensions were tested. This classification technique produced the least optimal results. The tuning parameter used was the number of randomly selected variables considered at each split of the tree, those values being $\lfloor \sqrt{p} \rfloor$ and p, where p is the total number of predictors. Furthermore, a Balanced Random Forest was also applied, which creates bags that are balanced between minority and majority classes and the trees being trained on those bags, however, the results were never significant improvements. To extract probabilities for the purpose of finding the most likely fatality candidates, the ensemble vote was extracted instead of the pure classification of classes.

5. Boosting: This tree extension performed strongly on validation sets, specifically gradient descent boosting, or `gbm` package and function in `R`. Without performing any rebalance strategies, boosting when applied at the appropriate and optimal shrinkage parameter had the best metrics of recall and PR-AUC. However, after applying rebalance, the performance decreased slightly, even after tuning the shrinkage to be optimal for a specific rebalance. The number of trees was kept at 1,000 for all models so perhaps changing the number of trees depending on the rebalance would have aided a more accurate model due to changes in variance or bias.

These models have been compiled in **Table 6** using recall which measures how much of the minority class the model captures.

Model Rebalance	None	ADASYN	100% Oversampling	50% Oversampling	Cost-sensitive 5 (10)
GLM					
AIC	0.498	0.4926	0.516	0.513	0.513 (0.501)
BIC	0.507	0.4897	0.519	0.507	0.504 (0.496)
Error CV	0.507	0.5011	0.519	0.516	0.501 (0.503)
PR-AUC CV	0.507	0.5011	0.519	0.510	0.504 (0.510)
Pruning					
0.01	0.389	0.469	0.09	0.307	-
0.001	0.454	0.392	0.498	0.345	0.428 (0.053)
0.0001	-	0.392	0.322	0.389	0.365 (0.43)
Random Forest					
p	0.391	0.4123	0.36	0.371	0.374 (0.371)
$\lfloor \sqrt{p} \rfloor$	0.407	0.4144	0.372	0.379	0.381 (0.379)
Boosting					
0.1	0.501	0.460	-	-	-
0.01	0.504	0.457	0.498	0.496	-
0.001	0.484	0.457	0.483	0.492	0.41

Table 6: Condensed table of TPR at top 25% threshold

Judging from this table alone, 100% upsampling GLMs produce the best outcome on the validation set. Although there was a tie of 0.519 TPR due to the similarity in predictors, the final model selected was the cross-validated PR-AUC as it had better secondary credentials such as PR-AUC.

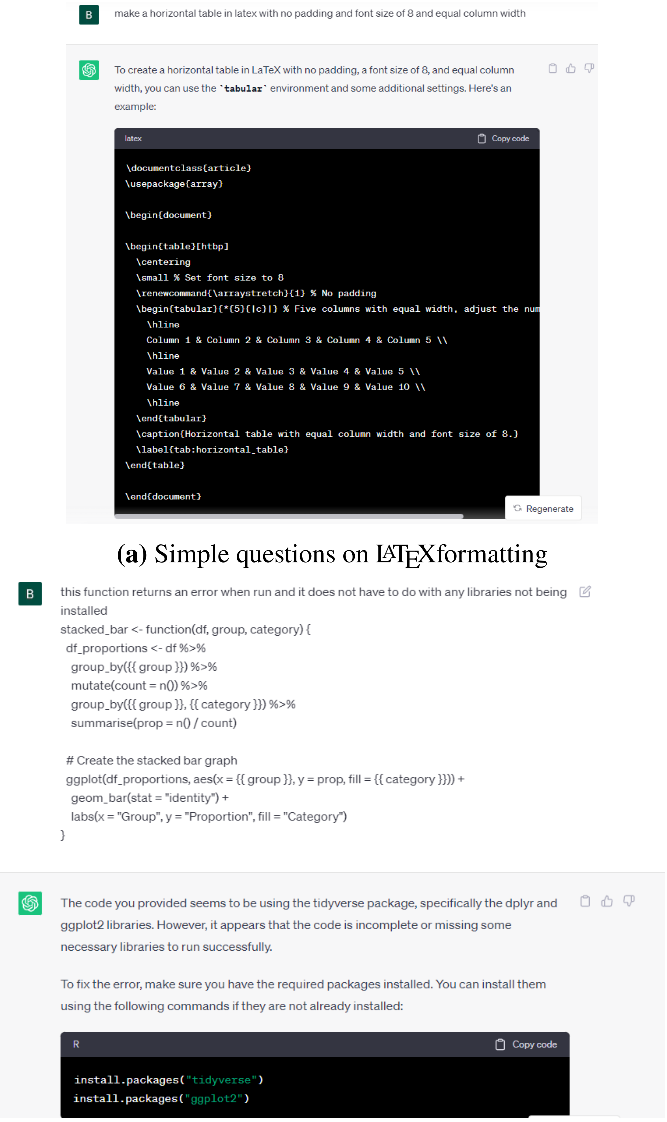
This model was further regularised through Ridge Regression tuned to the minimising value of λ as the Lasso Regression led to a significant fall in recall of 0.507 compared to Ridge which maintained it at 0.516. Even though the TPR fell, the Ridge Regression model is preferable as it reduces complexity and variance in the model as a result of noise and unsmooth data. This model was then applied to the test set where the recall rested at 0.498. This means that the model should capture roughly half of the total fatalities in the evaluative dataset.

To put this model into production, the same predictors from the chosen model were trained on the entire dataset oversampled and underwent L1 regularisation. The entire set was used to train the final model to avoid any groups or classifiers that could potentially have been absent from the oversampled training set also absent from the final models training. This was ultimately used to predict the $2,500$ most likely drivers to have an accident.

Generate AI Usage

The primary usage of generative AI, namely ChatGPT, was to aid in creating the illustrations seen throughout the document in both R and LaTeX. Prompts may include ”This piece of code is returning an error, try to debug and fix it”, ”How to parse the name of data frame variables into functions in R”, ”How to create stacked bar in ggplot2”, ”How to wrap text around a table in LaTeX”, ”How to insert multiple figures on same line in LaTeX.” Most of the code was not immediately executable, however with some human intervention the desired outcomes were found. This extends to finding which packages house certain functions, which ChatGPT is consistently poor at providing correct answers.

Photographic evidence:



make a horizontal table in latex with no padding and font size of 8 and equal column width

To create a horizontal table in LaTeX with no padding, a font size of 8, and equal column width, you can use the `tabular` environment and some additional settings. Here's an example:

```
\documentclass{article}
\usepackage{array}

\begin{document}

\begin{table}[htbp]
  \centering
  \small % Set font size to 8
  \renewcommand{\arraystretch}{1} % No padding
  \begin{tabular}{*{5}{|c}|} % Five columns with equal width, adjust the num
    \hline
    Column 1 & Column 2 & Column 3 & Column 4 & Column 5 \\
    \hline
    Value 1 & Value 2 & Value 3 & Value 4 & Value 5 \\
    Value 6 & Value 7 & Value 8 & Value 9 & Value 10 \\
    \hline
  \end{tabular}
  \caption{Horizontal table with equal column width and font size of 8.}
  \label{tab:horizontal_table}
\end{table}

\end{document}
```

(a) Simple questions on LaTeXformatting

this function returns an error when run and it does not have to do with any libraries not being installed

```
stacked_bar <- function(df, group, category) {
 df_proportions <- df %>%
  group_by({{ group }}) %>%
  mutate(count = n()) %>%
  group_by({{ group }}, {{ category }}) %>%
  summarise(prop = n() / count)

 # Create the stacked bar graph
 ggplot(df_proportions, aes(x = {{ group }}, y = prop, fill = {{ category }})) +
  geom_bar(stat = "identity") +
  labs(x = "Group", y = "Proportion", fill = "Category")
}
```

The code you provided seems to be using the tidyverse package, specifically the dplyr and ggplot2 libraries. However, it appears that the code is incomplete or missing some necessary libraries to run successfully.

To fix the error, make sure you have the required packages installed. You can install them using the following commands if they are not already installed:

```
install.packages("tidyverse")
install.packages("ggplot2")
```

(b) Generative AI not producing correct responses

Figure 9

References

[1] Impact Lists. *Useful Postcode Ranges*. `http://www.impactlists.com.au/ImpactLists/media/list-tools/Useful-Postcode-Ranges.pdf`.

[2] Austroads. *Guide to Road Safety Part 5: Road Safety for Rural and Remote Areas*. Available online: `https://austroads.com.au/publications/road-safety/agrs05-06/media/AGRS05-06_Guide_to_Road_Safety_Part_5_Road_Safety_for_Rural_and_Remote_Areas.pdf`

[3] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*, 10(3), e0118432. doi: 10.1371/journal.pone.0118432.

[4] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. doi: 10.1109/TKDE.2008.239.