Capstone Project: Technical Report

Title: Predicting Likelihood and Severity of Road Accidents

GENERAL ASSEMBLY

CLASS: DAI06

NAME: SAID JOMAA

DATE: 13/06/2023

Executive Summary

Overview

The objective of this study is to build a high-precision model that predicts the severity of accidents and identifies the key factors that contribute to fatal accidents. By analysing road accident data from Victoria, Australia, spanning from 2006 to 2020, I aim to determine the most influential variables associated with severe accidents. This research utilizes multiple logistic regression to establish a predictive approach for accident severity. The findings from this study can assist authorities in prioritizing corrective measures and implementing targeted interventions to reduce the occurrence of fatal accidents.

Methodology

The dataset was initially processed to handle outliers, missing values, and skewness. To further transform the data, one-hot encoding and label encoding techniques were applied. To address the issue of data imbalance, two techniques were used: under sampling and SMOTE. This resulted in the creation of two different datasets.

The regression analysis consisted of several iterations following three main steps:

1. Feature selection: The features to be included in the model were chosen.
2. Model development and adjustment: A regression model was developed and fine-tuned.
3. Model assessment and validation: The performance of the model was evaluated and validated.

Our final model uses the under-sampling data set. The data is further split into training and testing sets using an 80:20 ratio. With a prediction threshold set as 0.65.

Overview of the result

Based on the results obtained from the logistic regression using the forward stepwise method, the accuracy of the predictions is found to be 81.4% The analysis highlights the significant role of several variables in determining the severity of accidents. Factors such as speed zone, urbanisation, the number of people involved (leading to more distractions), street light availability, unfavourable weather conditions, and the influence of unsafe and poor-quality vehicles are found to contribute to an increase in accident severity.

Limitations

One of the limitations of this study was the limited availability of complete data, which affected the analysis. Although the current analysis did not yield significant results, it is expected that with a larger and more comprehensive dataset, the models could generate more robust and conclusive findings.

Nevertheless, the study provided valuable insights into the relationships between causal factors and accident outcomes. By understanding these relationships, it becomes possible to identify the factors that have the potential to result in catastrophic events as well as those that may lead to less severe incidents. This knowledge can contribute to improving preventive measures and enhancing overall safety measures.

Introduction

Victoria has been a leading advocate for road safety in Australia for many years, boasting a lower fatality rate per 100,000 population compared to the national average. However, despite these achievements, the progress in reducing road fatalities in Victoria has stalled, and motor vehicle accidents continue to cause significant loss of life and injuries, with profound health and financial implications.

The economic impact of road casualties in Victoria is projected to exceed $1 billion annually, encompassing property damage and associated costs. Back in 1993-94, the total lifetime medical expenses and related costs for motor vehicle traffic-related injuries amounted to $570 million. According to Minister for Road Safety Luke Donnellan, it is estimated that approximately 2,500 people will lose their lives in car accidents in Victoria over the next ten years, while 50,000 individuals will suffer severe injuries requiring hospitalization and life-altering changes.

Despite the proactive measures taken by the Victoria management authority to address this issue through rule enforcement and safety initiatives, it is evident that a significant proportion of severe road accidents result from driver carelessness and irresponsibility. Therefore, it is essential to conduct an in-depth analysis to determine the contributing factors to the severity of accidents in Victoria and develop effective strategies to mitigate the frequency and severity of such incidents.

Consequently, the primary objectives of this study are as follows:

1. Analyse data on accident severity in Australia, considering various factors such as age, causes of accidents, types of road users, and locations. The aim is to identify the key factors associated with severe outcomes in accidents, conditional on an accident occurring.

2. Develop a logistic regression-based predictive model for accident severity in Australia, utilizing data spanning from 2006 to 2020. This model will enhance our understanding and prediction capabilities regarding the factors that influence the severity of accidents, thereby facilitating the formulation of targeted interventions to reduce the occurrence of severe outcomes.

Data

The primary source of data for this study was the Road Crash Information System (RCIS), which provides a comprehensive national road crash database. The database encompasses over 500,000 crash records and includes 100 different features, covering a period of four years from 2006 to 2020. Detailed information about the database can be found in Appendix A-1 and A-2.

Given the extensive number of features available, our analysis aimed to explore various methods for predicting the probability of fatalities associated with different factors. The dataset consisted of a diverse range of features, including both numerical and categorical data. To fulfill the research objective of identifying potential factors contributing to fatal road injuries, we performed data refinement and parsing in the Data Pre-processing section. This ensured that the data was suitably prepared for further analysis and investigation.

Exploratory Data Analysis

To facilitate analysis and derive meaningful insights from the road accident dataset, I categorizes the findings into three key areas of inquiry:

- The "When" - Period

- The "Where" - Geographical Location

- The "How" - Additional Factors

Furthermore, to enhance our understanding of the factors contributing to severe outcomes in road accidents, we will filter the dataset to include only records where the "SEATING POSITION" is specified as the driver. Focusing on data related to one person in the driver's seat will provide us with better clarity regarding the factors influencing the severity of outcomes in road accidents.

The “When” – Time

From 2006 until 2020, the number of deaths is depicted in Figure 1, with each month represented separately. The months of March to May are the deadliest. By contrast, during the year, August and September are the months of have the fewest severe accidents

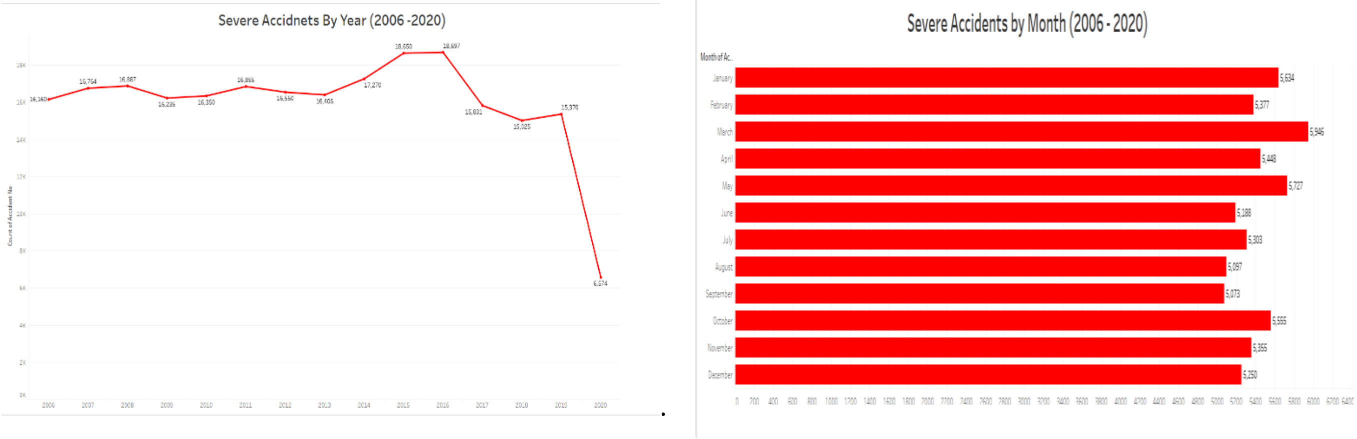
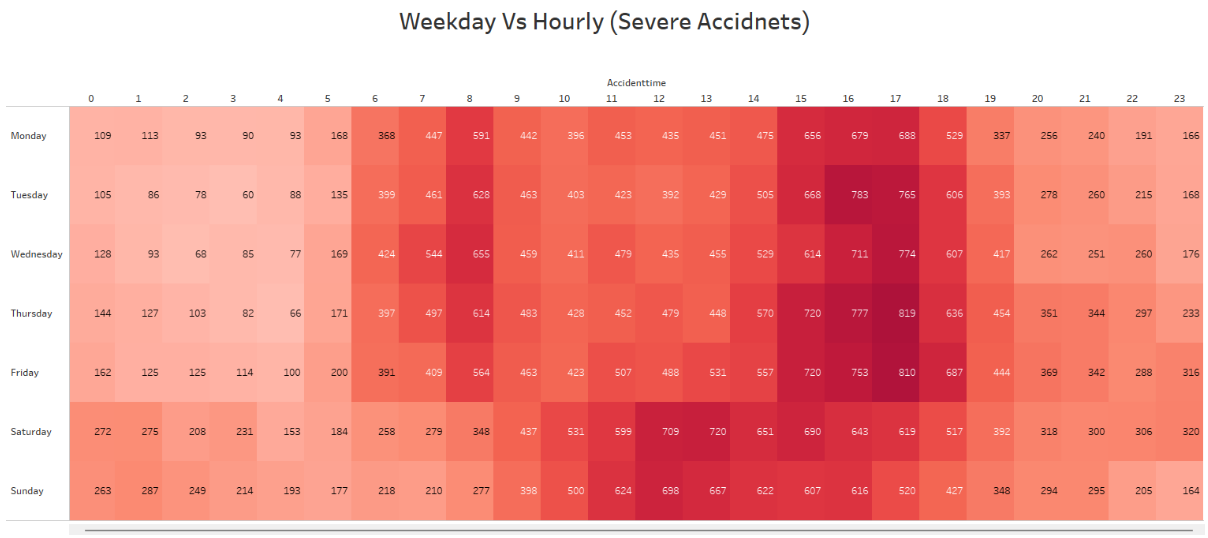


Figure 1: Severe Accidents by Years and Months

The heatmap on the right (Figure 2) illustrates the intensity of severe accidents at each hour of the day. It reveals that the highest number of severe accidents often occurs at 15:00, coinciding with a time when people are particularly vulnerable. It is worth noting that this peak could potentially be attributed to the dismissal time of schools when stricter safety measures are typically enforced in school zones.

Analysing the severe accidents by time, it is evident that during weekdays, the number of severe accidents is notably high at 15:00 and 16:00. However, on weekends, a different trend emerges, with a gradual increase in severe accidents starting from 12:00 and peaking at 15:00. After this peak, the number of severe accidents generally declines as the day progresses.



Weekday Vs Hourly (Severe Accidnets)

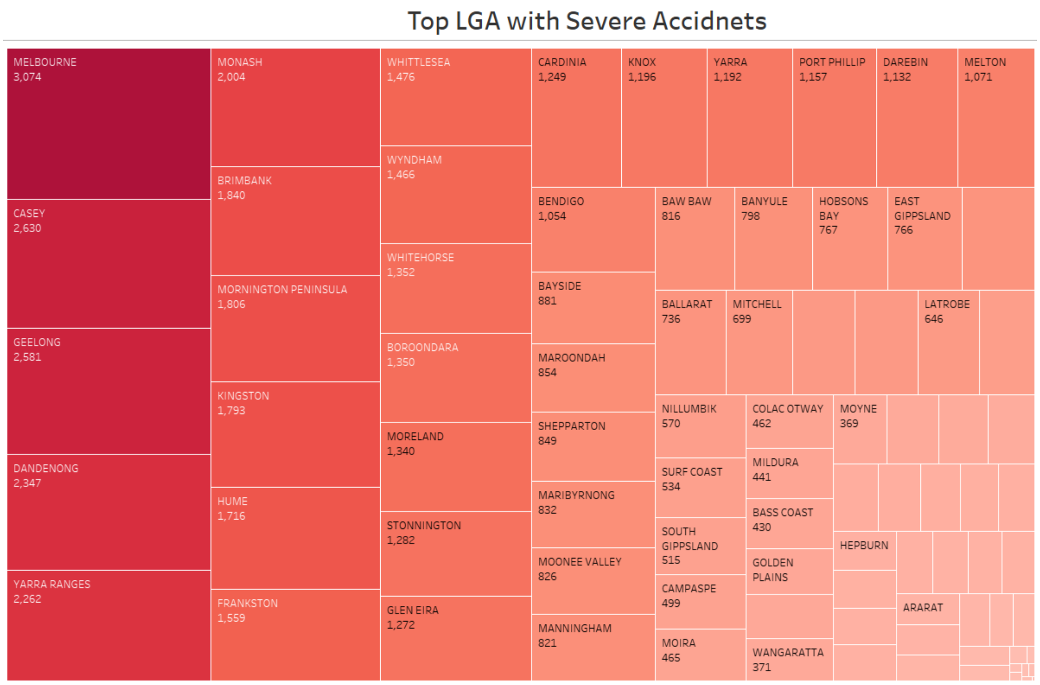
	Accidenttime																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Monday	109	113	93	90	93	168	368	447	591	442	396	453	435	451	475	656	679	688	529	337	256	240	191	166
Tuesday	105	86	78	60	88	135	399	461	628	463	403	423	392	429	505	668	783	765	606	393	278	260	215	168
Wednesday	128	93	68	85	77	169	424	544	655	459	411	479	435	455	529	614	711	774	607	417	262	251	260	176
Thursday	144	127	103	82	66	171	397	497	614	483	428	452	479	448	570	720	777	819	636	454	351	344	297	233
Friday	162	125	125	114	100	200	391	409	564	463	423	507	488	531	557	720	753	810	687	444	369	342	288	316
Saturday	272	275	208	231	153	184	258	279	348	437	531	599	709	720	651	690	643	619	517	392	318	300	306	320
Sunday	263	287	249	214	193	177	218	210	277	398	500	624	698	667	622	607	616	520	427	348	294	295	205	164

figure 2: Weekday vs Hourly severity accidents (2006-2020)

The “Where” – Location

In Figure 3, the top Local Government Authority (LGA) locations in Victoria with the highest number of severe accidents are depicted. The LGAs "MELBOURNE," "GEELONG," and "CASEY" emerge as the areas with the highest number of severe accidents. However, it is important to note that the number of severe accidents alone does not necessarily indicate the level of road safety in these areas. It is possible that the data may be influenced by population density, as areas with a higher population may have a higher likelihood of experiencing more severe accidents.

Therefore, when interpreting the data, it is crucial to consider factors beyond the number of severe accidents, such as population density, traffic volume, road infrastructure, and other variables that contribute to road safety. A comprehensive analysis should consider multiple factors to assess the true level of road safety in different LGAs.



Top LGA with Severe Accidnets

Figure 3: Top LGA areas with highest Severe Accidents (2006-2020)

Furthermore, when examining the relationship between location and fatality rate, it is important to consider the nature of the routes where these accidents occur. It is notable that the majority of fatalities have occurred on non-intersection routes, primarily on highways. This suggests that a significant proportion of accidents leading to severe outcomes are happening on highways. Shown in fig4.

Additionally, it is worth mentioning that approximately 31% of accidents take place in rural areas of Victoria are severe. This statistic highlights the significance of considering the rural regions when analysing road safety and severe accident occurrences. Illustrated in fig7.

It is crucial to understand the specific characteristics of different road types and locations in order to develop targeted road safety measures and interventions that address the unique challenges and risks associated with each type of road and location.

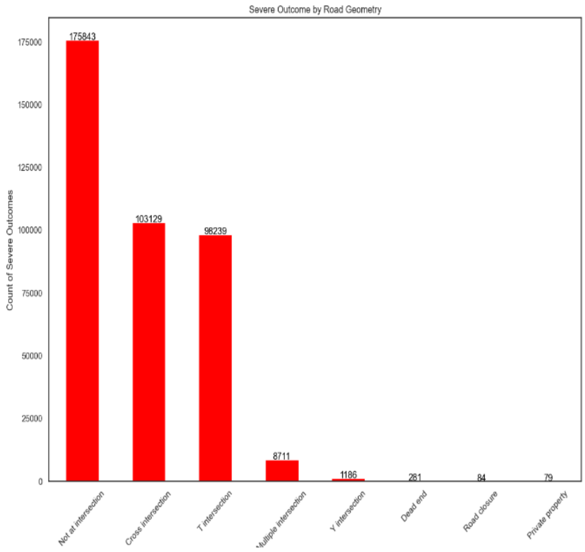


Figure 4: Severe Accidents by Road Geometry (2006-2020)

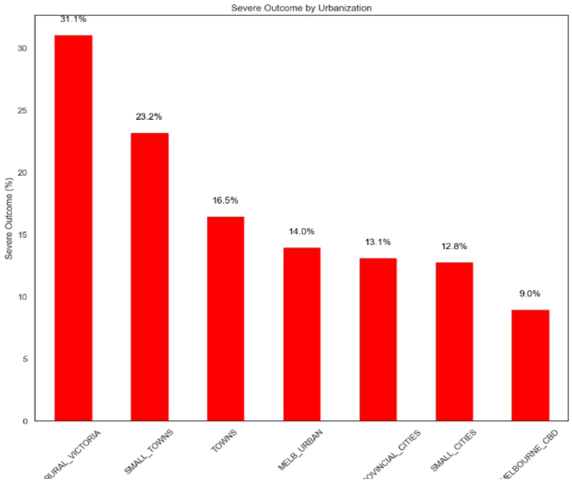


Figure 5: Severe Accidents by Urbanization (2006-2020)

The "Why" – Other factors

The analysis reveals a clear pattern where a significant number of severe accidents resulting in fatalities are concentrated on highways with speed limits of 100km/h or higher (Figure 6). This finding aligns with the understanding that the chances of survival in a high-speed collision are significantly reduced. The second highest number of fatalities is observed on highways with speed limits of 80km/h and 60km/h, respectively.

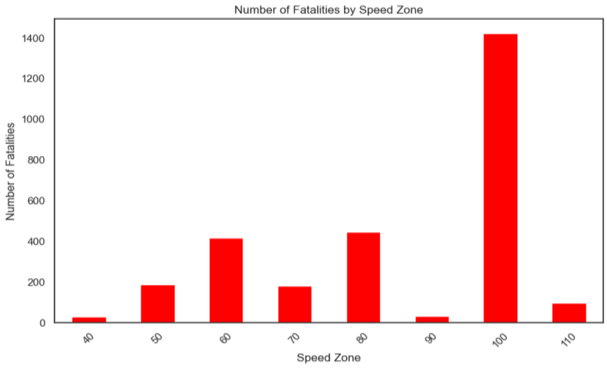


Figure 6: Fatalities by Speed Limits (2006-2020)

In terms of death by road users, motor vehicle drivers accounted for most severe accidents (68.9 percent), followed by motorcycle riders (21.3 percent) and bicycles (9.9 percent). A further concern is that nearly half of the fatalities were caused by a collision with another motor vehicle. In the second largest number of fatalities, a collision with a stationary object on the road was the cause.

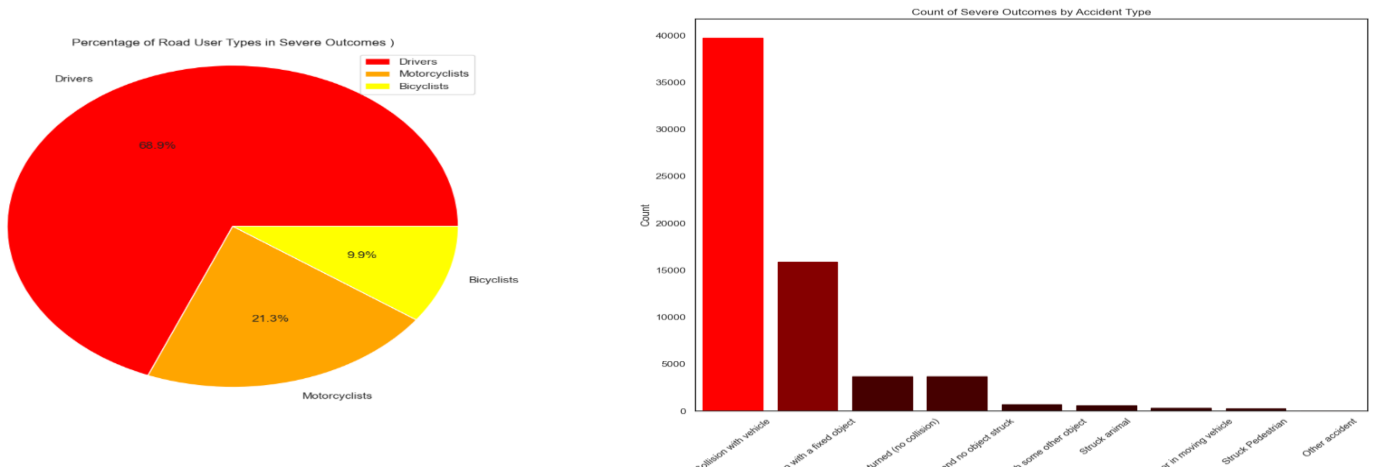


Figure 7: Fatalities by Road Users and Accident Types (2006-2020)

When analysing road surface conditions, it was found that wet roads posed the highest risk, followed by dry and muddy roads, accounting for most severe accidents. However, it should be noted that there were no significant differences in the number of fatalities based on weather conditions. Although rain and cloudy conditions showed a slightly higher number of car fatalities, the difference was not substantial.

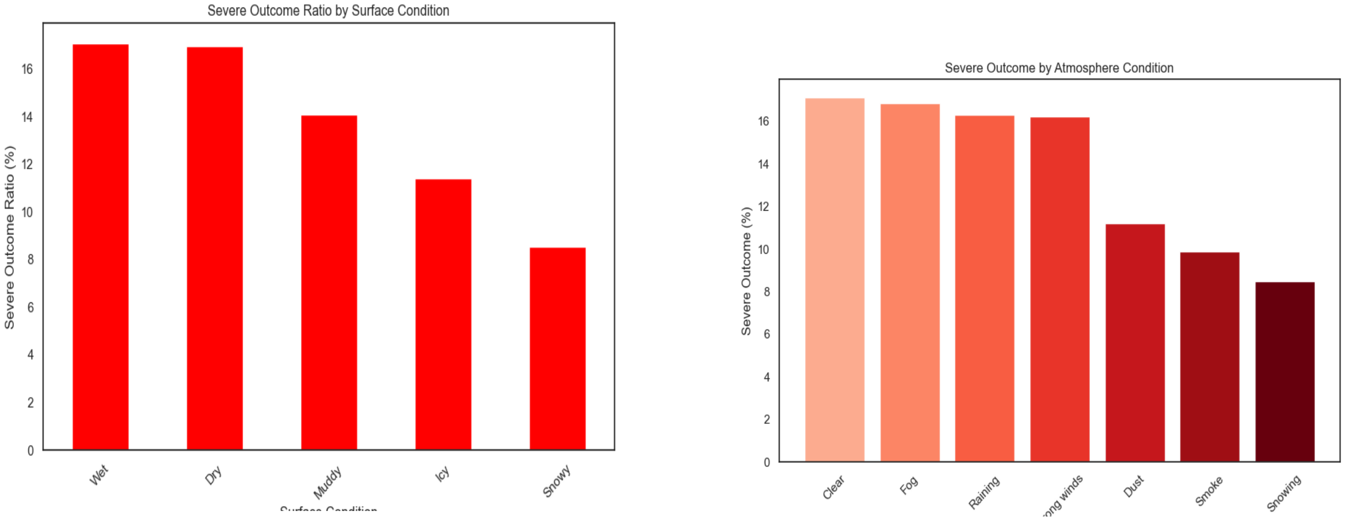


Figure 8: Fatalities by Surface Conditions and Weather Conditions (2006-2020)

Method

Regression Model

In my study, I opted to use the Logistic Regression Model to address our research questions. Logistic regression models are commonly employed to model the probability of a specific event based on independent predictor variables.

Despite the emergence of more advanced machine learning techniques in accident investigations, logistic regression offers several advantages that make it suitable for my project. Firstly, the findings of regression models are easily interpretable. The model can be represented by a single formula, incorporating the independent variables and their coefficients. These coefficients provide valuable insights into the critical variables and the magnitude and direction of their association with the dependent variable, such as the severity outcome.

Secondly, logistic regression models are relatively straightforward to create once the risk factors are identified. They do not require tuning multiple hyperparameters as some machine learning methods do. This simplicity makes logistic regression a commonly used initial classifier in predictive research and serves as a reliable baseline for more advanced classifiers.

Lastly, while association rules can be effective in identifying hidden patterns within large datasets, they are fundamentally different from classification methods like logistic regression modelling. Logistic regression focuses on establishing associations between independent variables and the target variable, whereas association rules focus on discovering associations between different variables in the dataset.

Therefore, considering these advantages, logistic regression was deemed suitable for my project objectives.

Data Preparation

Data Filtering

In the process of building statistical models, it is common practice to select variables that contribute to a concise and effective model, aiming for simplicity rather than excessive complexity (referred to as parsimony). This involves carefully evaluating each variable to ensure its usefulness, as larger numbers of variables do not necessarily lead to better models.

To achieve a parsimonious model, it is crucial to have a clear understanding of the specific problems we intend to address using our dataset. In this project, our goal was to predict the factors that contribute to severe road accidents. To focus our analysis and avoid potential biases, we filtered the data points to include only the variables related to 'Drivers' and 'Motorcyclists' within the Road User Types category. By applying this filter, the dataset was reduced to 311,199 records, enabling us to concentrate specifically on the relevant variables for our analysis.

Data Pre-processing

In the project, I followed a systematic approach to statistical modelling. I analysed the dataset to ensure its tidiness and suitability for the chosen modelling technique. The merged dataset was found to be unstructured and contained irrelevant data that did not contribute significantly to the prediction process. To address these issues, I performed data pre-processing steps to enhance the dataset's quality and improve modelling efficiency.

The pre-processing tasks included dealing with missing values, addressing skewness in the data, encoding categorical variables, and managing data imbalance. I undertook these steps as part of the data preparation process, aiming to create a refined dataset for analysis and modelling.

By applying these pre-processing techniques, I ensured that the dataset was properly organized and met the requirements of the selected statistical technique, facilitating accurate and meaningful analysis of the data.

Dealing with Missing Values

Addressing missing values in a dataset can be a challenging task as it requires careful consideration to maintain data integrity. Different methods exist for handling missing values, but they must be chosen with caution to ensure the reliability of the results.

In my study, managing missing values for numerical variables was relatively straightforward, but dealing with missing categorical data presented some challenges. The dataset I used contained a total of 135,303 missing values, which accounts for approximately 0.83% of the entire dataset.

I took appropriate measures to handle these missing values effectively, employing methods such as imputation or excluding incomplete cases based on the specific context and requirements of the analysis. By addressing missing values, I aimed to minimize any potential bias or distortions in the results and maintain the integrity of the dataset during the modelling process. the majority of the important variables have little or no missing values while most of the missing values are associated with various vehicle information.

ased on the research problem and technical requirements of my study, I evaluated several approaches for dealing with missing values and made decisions based on their limitations and suitability. The table below outlines the rejected techniques and the corresponding problems associated with each approach.

Table 1. Rejected Missing Data Dealing Techniques

Techniques	**Problems**
Drop / Remove all missing values	Dropping missing values would result in a significant loss of data, making this method impractical.
Imputation Using Mode Values	Replacing missing values with the mode for categorical variables increased skewness, rendering this method unsuitable.
Imputation Using (Mean/Median) Values	Imputing missing values with mean or median values can significantly reduce model accuracy and introduce bias.
Amelia predictive model (Multiple Imputation)	The computational resource requirements of the Amelia predictive model were too high for a home computer.
K-nearest neighbour	K-nearest neighbour imputation method is computationally expensive and requires storing the entire training dataset in memory.

After evaluating various techniques, I selected Random Imputation as the chosen approach for handling missing values before constructing the regression model (Figure 11). This method effectively eliminates imputation variance while preserving the distribution of item values.

In the case of Random Imputation, I utilized a vector containing the cumulative sum of the counts for each unique value in the categorical feature. Additionally, I generated a set of random values ranging from 0 to the maximum cumulative sum. By using this set of random numbers, I imputed each missing value based on the index of the cumulative sum list. This approach ensures that the imputed values maintain the distribution pattern observed in the existing data, effectively preserving the underlying data characteristics.

When dealing with skewed data, I encountered challenges because logistic regression assumes a normal distribution. In our dataset, certain numerical variables such as TOTAL_NO_OCCUPANTS, ,NO_PERSONS and SPEED_ZONE exhibited varying degrees of skewness, as shown in Figure 9. To ensure compatibility with our modelling algorithms, which assume a normal distribution, I needed to address this skewness. Since the skewed data closely resembled beta distributions, I decided to apply a log transformation to mitigate the skewness.

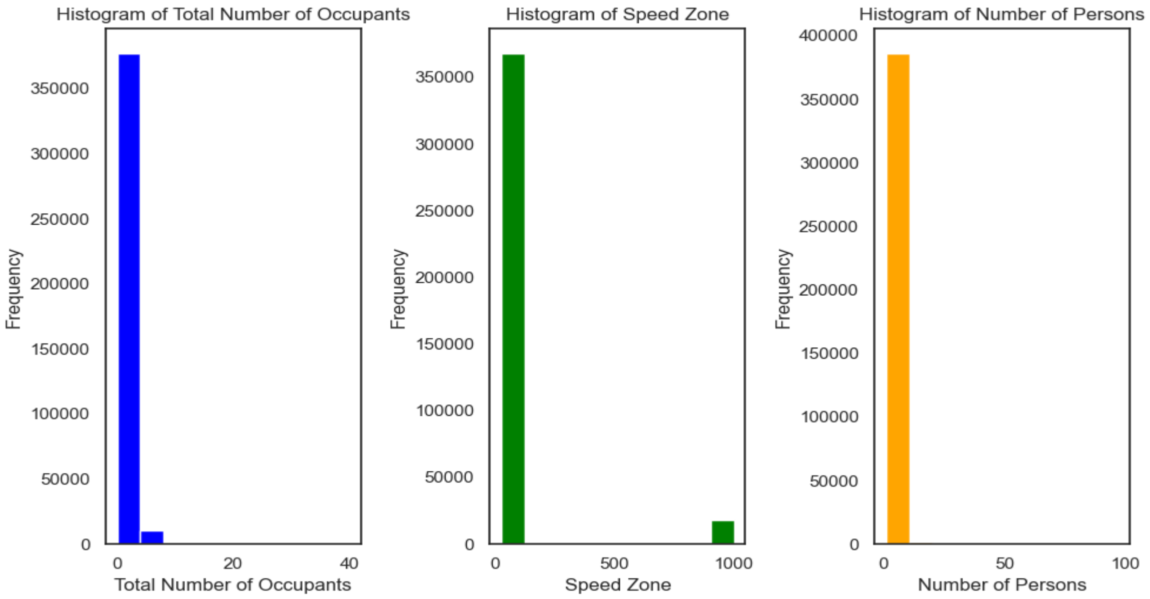


Figure 9: Data distribution of numeric columns

Categorical Data Encoding

In my study, I encountered a combination of categorical and numerical data features. Categorical variables carry valuable information that is related to the target variables. However, since I plan to use a machine learning approach like logistic regression, it requires numeric input variables. To address this, I employed two common techniques for encoding categorical data: One-Hot Encoding and Label Encoding.

Given the nature of my categorical variables, where the values represent different variables with limited relationships between them, I decided to utilize One-Hot Encoding. This technique allows me to represent categorical information effectively, especially for variables that have a significant presence in severe outcomes. By applying One-Hot Encoding, I was able to transform these categorical variables into a suitable numeric format for my logistic regression model.

Data imbalance

In our dataset, we encountered a significant class imbalance issue, particularly in the context of fatalities and non-fatalities. The count of non-severe was 322599, while the count of severe was only 64953. This class imbalance is commonly referred to as the "Rare Class Problem," where one class has significantly fewer examples than the others.

Working with imbalanced datasets poses challenges for machine learning algorithms, as they tend to overlook the minority class and perform poorly on it. However, the minority class, in this case, the fatalities, is of utmost importance to our analysis. To address this issue, I employed two methods:

1. Under sampling: I identified the total occurrences of the class of focus (fatalities) and randomly sampled a subset from the overpopulated class (non-fatalities). This approach is akin to randomly eliminating cases from the majority class, aiming to balance the class distribution.
2. SMOTE (Synthetic Minority Oversampling Technique): Instead of under sampling, which can result in data loss, I utilized an oversampling technique called SMOTE. SMOTE generates synthetic data points by bootstrapping and utilizing k-nearest neighbours. This technique helps to address the class imbalance problem by increasing the representation of the minority class.

By employing both under sampling and SMOTE, I aimed to mitigate the challenges posed by the class imbalance problem and improve the performance of the machine learning algorithms in capturing the important patterns related to fatalities.

Features Selection

In the feature selection process, I considered the patterns of correlations observed between the features and the target variable during exploratory data analysis (EDA). Additionally, I relied on domain knowledge to confirm the relevance of these features, as road traffic accidents are a universal topic. To address multicollinearity issues, I analysed the correlations among the selected features and retained only one feature from highly correlated pairs.

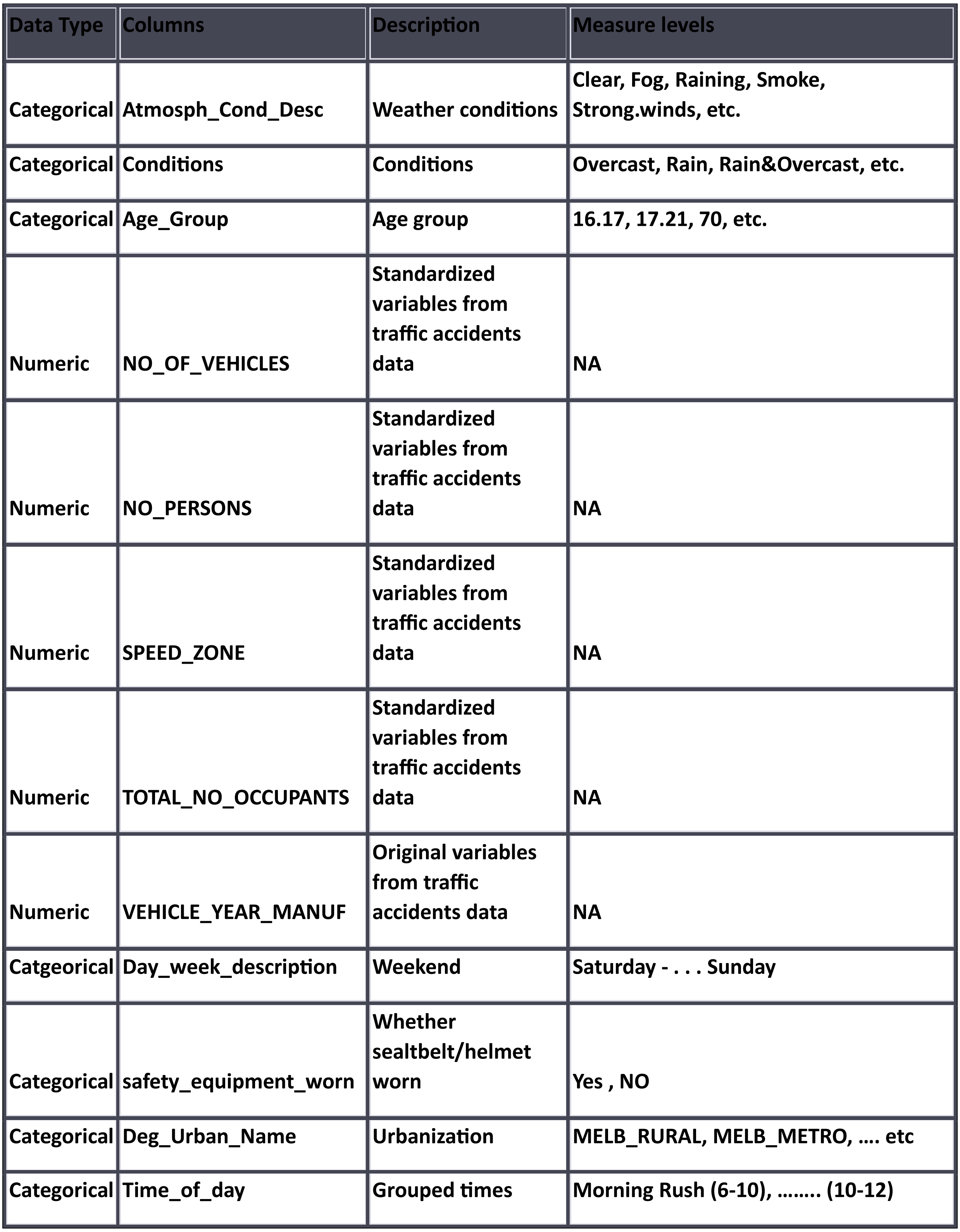
In order to enhance computational efficiency and improve the model's performance, I further narrowed down the feature set. This step was performed before applying techniques to address data imbalance. The final set of features that I selected are:

- Target variable: Severe_outcome
- User type, specifically focusing on drivers of cars or motorcyclists: Road_User_Type_Desc
- Date and time information: ACCIDENTDATE, ACCIDENTTIME

By carefully selecting these features, I aimed to optimize the model's predictive capability while considering computational constraints and the relevance of the variables to the research question.

Features that were produced from one hot encoding (Table 3):

Data Type	Columns	Description	Measure levels
Categorical	**Gender**	**Gender**	**Female - Male**
Categorical	**Accident_Type_Desc**	**Accident Type**	**Collision.with.a.fixed.object, Struck.animal, Struck.Pedestrian, Vehicle.overturned&no.collision, etc.**
Categorical	**Road_Surface_Type_Desc**	**Road surface type and condition**	**Unpaved, Dry, Icy, Muddy, Snowy, Wet, etc.**
Categorical	**Light_Condition_Desc**	**Light conditions**	**Dark.No.street.lights, Street.lights.off, etc.**



Data Type	Columns	Description	Measure levels
Categorical	Atmosph_Cond_Desc	Weather conditions	Clear, Fog, Raining, Smoke, Strong.winds, etc.
Categorical	Conditions	Conditions	Overcast, Rain, Rain&Overcast, etc.
Categorical	Age_Group	Age group	16.17, 17.21, 70, etc.
Numeric	NO_OF_VEHICLES	Standardized variables from traffic accidents data	NA
Numeric	NO_PERSONS	Standardized variables from traffic accidents data	NA
Numeric	SPEED_ZONE	Standardized variables from traffic accidents data	NA
Numeric	TOTAL_NO_OCCUPANTS	Standardized variables from traffic accidents data	NA
Numeric	VEHICLE_YEAR_MANUF	Original variables from traffic accidents data	NA
Catgeorical	Day_week_description	Weekend	Saturday - . . . Sunday
Categorical	safety_equipment_worn	Whether sealtbelt/helmet worn	Yes , NO
Categorical	Deg_Urban_Name	Urbanization	MELB_RURAL, MELB_METRO, …. etc
Categorical	Time_of_day	Grouped times	Morning Rush (6-10), …….. (10-12)

Developing Model

To further explore the factors contributing to fatal accidents in Victoria, we proceed with implementing the logistic regression model based on the insights gained from the exploratory data analysis (EDA) and the rationale discussed earlier.

The first step is to split the training data into three sets: train, test, and validation. The training dataset is used to train and fit the model, the validation dataset is used to evaluate the model's performance while tuning hyperparameters, and the test dataset is used to provide an unbiased evaluation of the final model fit on the training dataset.

The data split is performed with a ratio of 80/20, where 80% of the data is allocated to the training set and the remaining 20% is split equally between the test and validation sets. This ensures that an adequate amount of data is available for training and evaluation purposes.

By performing this data split and following the logistic regression approach, we aim to gain further insights into the variables and their impact on fatal accidents in Victoria.

However, because our dataset also comes with the rare class problem, the trained data for the logistics model is generated using different methods for dealing with this issue, notably the Under sampling and SMOTE approaches, which was discussed in the Imbalance Data section above.

Moreover, to enhance the accuracy of the model during the evaluation phase, I have decided to employ the Lift and Reduce method. This technique involves setting specific threshold ratios (0.40, 0.50, and 0.60) to determine the model's ability to predict the probability of a "FATAL ACCIDENT" occurrence. By adjusting these threshold ratios, I aim to optimize the model's performance and achieve more accurate predictions.

Evaluation

In the evaluation process of our model, I utilized the confusion matrix to assess its performance in correctly classifying cases. The confusion matrix provides insights into the true positive, false negative, false positive, and true negative predictions for each class, allowing us to evaluate the model's accuracy.

To evaluate the model, I employed several standard evaluation metrics, including accuracy, recall, and AUC. Accuracy is calculated by dividing the sum of true positives and true negatives by the total sample size. Recall is determined by dividing true positives by the sum of true positives and false negatives. These metrics provide valuable information about the model's classification performance.

In the case of predicting fatalities, I considered the cost associated with misclassifying a non-fatality as a fatality to be relatively lower compared to misclassifying a fatality as a non-fatality. Therefore, while accuracy remains a consideration in evaluating the model, I placed greater emphasis on recall.

Additionally, AUC (Area Under the Curve) was utilized to summarize the overall diagnostic accuracy of the model. AUC values between 0.5 and 0.7 indicate performance no better than random chance, values between 0.7 and 0.8 are considered acceptable, values between 0.8 and 0.9 are considered excellent, and values above 0.9 are considered outstanding. To meet the desired criteria, the model needed to achieve an AUC score greater than 0.7.

Results

When selecting arbitrary thresholds for classifying the predicted response as a fatality, the motivation was to address the possibility that the default threshold of 0.5 may result in misclassifying severe as non-severity. By lowering the threshold, we aimed to capture cases where the predicted probability of severity is lower than 0.5 but still significant. This approach allows for the correct classification of samples that might be considered false negatives in the confusion matrix. However, it is crucial to exercise caution in lowering the threshold too much, as it could lead to the model overcompensating and classifying true negatives as false positives.

Method	Accuracy	Precision	Recall	AUC
Baseline Model	0.797	0.520	0.184	0.570
Under Sample Fatality > 0.50	0.780	0.463	0.422	0.648
Under Sample Fatality > 0.65	0.766	0.441	0.514	0.673
Under Sample Fatality > 0.60	0.768	0.442	0.482	0.662
SMOTE > 0.50	0.780	0.463	0.422	0.648
Under Sample Fatality > 0.45	0.785	0.473	0.401	0.643

Table 3. Model results summary

The final selected model was "Under Sample Fatality > 0.65," which trained on the under-sampled dataset while maintaining a classification threshold of 65%. As shown in Table 6, this model demonstrated significantly improved performance in correctly classifying true positives compared to the baseline model, while still maintaining accurate classification of true negatives. Moreover, the recall of 0.51 exceeded that of the other models developed, which is particularly valuable given the nature of the classification problem. Additionally, the AUC of 0.673 fell within the acceptable range identified earlier. Considering these factors, the "Under Sample Fatality > 0.65" model was chosen as the final model.

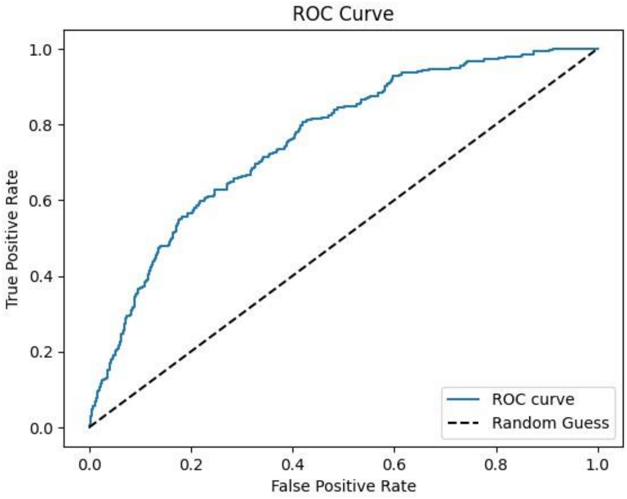


figure 10: Best model AUC chart

Limitations

Ethical Concerns

One limitation of this study is the lack of available data due to ethical considerations. Personal information such as hobbies, habits, and ethnicity, which could potentially be used to target specific groups for safer driving education, was not included. While this data is sensitive, it could have provided valuable insights for targeted preventive measures.

Lack of Traffic Data

Another limitation is the absence of traffic data for cars that did not experience accidents. The focus of this study was on predicting fatalities in the event of an accident, and the lack of comprehensive traffic data limits the scope of analysis.

Missing Data

The presence of missing data posed another challenge. Various techniques were employed to impute missing values while maintaining the proportion or mean of the data, minimizing the impact on the analysis. However, the presence of missing data may introduce bias or affect the accuracy of the models.

Overall

The primary limitation of this work was the limited availability of complete data, which restricted the depth of analysis. It is expected that with access to more comprehensive data, the models would yield more robust and meaningful findings. Nonetheless, despite these limitations, exploring the relationships between causal factors and outcomes can provide valuable insights into identifying potential factors that may lead to catastrophic events and those that may contribute to less severe incidents.

Conclusion

The analysis of road traffic accident data is crucial for authorities to make informed decisions in prioritizing and enhancing policies and infrastructure as preventive measures. Our study aimed to investigate the factors contributing to severe accidents and identify the most significant combinations of causes, providing valuable insights for authorities to guide their decision-making process. By identifying the key variables of interest, we were able to develop a robust model that yielded the best results.

This study involved several stages, including data collection, cleaning, and transformation, as well as iterative exploratory data analysis (EDA), model development, and evaluation. Understanding the factors included in the final model and interpreting its results is essential for gaining meaningful insights.

Among the variables considered, we found that the following factors had the highest predictive power for severe accidents: speed zone, involvement of pedestrians, collisions with fixed objects, gender, the number of individuals involved, and atmospheric conditions (specifically, whether they were clear or not).

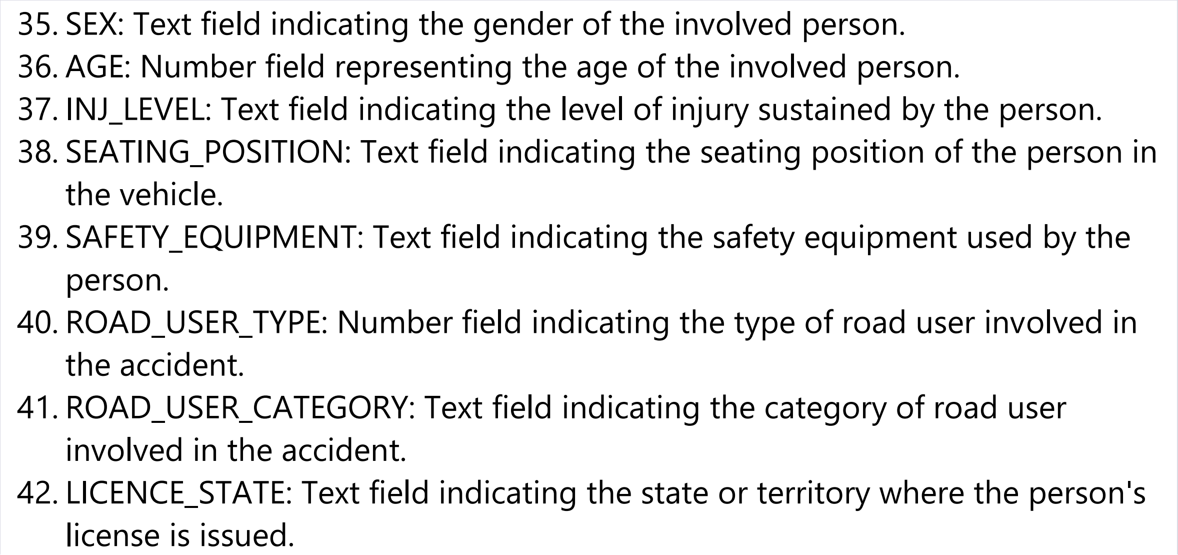
However, it is important to note that this study does not establish a causal relationship between these model features and severe accidents since we lack benchmarking data for comparison. Nonetheless, within the context of accidents that have already occurred, these are the prevalent features associated with severe outcomes. Further research could explore these areas to determine if any causal relationships exist.

Overall, the insights gained from this study can provide valuable guidance to traffic authorities in prioritizing corrective measures and improving road safety. By leveraging these findings, authorities can make more informed decisions to enhance the effectiveness of their interventions.

Appendix provides a description of the attributes in the accident data. Here is a breakdown of the different fields and their definitions:

1. ACCIDENT_DATE: Text field representing the date of the accident in the format dd/mm/yyyy.
2. ACCIDENT_TIME: Text field representing the time of the accident in the format hh.mm.ss.
3. STAT_DIV_NAME: Text field indicating the region where the accident occurred, categorized as "Metro" or "Country."
4. ACCIDENT_NO: Text field representing the unique identifier for the accident. The format changed from November 2005 to start with the letter "T" followed by the year and a numeric sequence.
5. ACCIDENTDATE: Date field representing the date of the accident in Australian format DD/MM/YYYY.
6. ACCIDENTTIME: Text field representing the time of the accident in the format hh.mm.ss.
7. ACCIDENT_TYPE: Number field indicating the type of accident, categorized into nine basic descriptions.
8. DAY_OF_WEEK: Number field indicating the day of the week when the accident occurred, represented by values 1 to 7.
9. DCA_CODE Part 1: Text field representing the first part of the code used for classifying accidents.
10. LIGHT_CONDITION: Number field indicating the light condition or level of brightness at the time of the accident.
11. NO_PERSONS: Number field representing the total number of people involved in the accident.
12. NO_PERSONS_KILLED: Number field representing the number of people killed in the accident.
13. NO_PERSONS_INJ_2: Number field representing the number of people with a specific level of injury.
14. DCA_CODE Part 2: Text field representing the second part of the code used for classifying accidents.
15. NO_PERSONS_INJ_3: Number field representing the number of people with a specific level of injury.
16. NO_PERSONS_NOT_INJ: Number field representing the number of people who were not injured in the accident.
17. NO_OF_VEHICLES: Number field representing the number of vehicles involved in the accident.
18. POLICE_ATTEND: Number field indicating whether the police attended the scene of the accident.

19. ROAD_GEOMETRY: Number field indicating the layout of the road where the accident occurred.
20. SEVERITY: Text field representing the estimated severity or seriousness of the accident.
21. DIRECTORY: Text field indicating the name of the street directory used for providing a map reference for the accident.
22. EDITION: Text field representing the edition or version of the street directory used for providing a map reference.
23. PAGE: Text field representing the page number of the street directory used for providing a map reference.
24. GRID_REFERENCE_X: Text field representing the grid reference in the x-direction of the cell in the street directory used for providing a map reference.
25. GRID_REFERENCE_Y: Text field representing the grid reference in the y-direction of the cell in the street directory used for providing a map reference.
26. SPEED_ZONE: Text field indicating the speed zone at the location of the accident.
27. NODE_ID: Text field representing the node ID of the accident location.
28. EVENT_SEQ_NO: Number field representing the sequence number of events within the same accident.
29. EVENT_TYPE: Text field indicating the type of incident event.
30. VEHICLE_1_ID: Text field representing the first vehicle involved in the event.
31. VEHICLE_1_COLL_PT: Text field representing the collision point on the first vehicle.
32. VEHICLE_2_ID: Text field representing the second vehicle involved in the event.
33. VEHICLE_2_COLL_PT: Text field representing the collision point on the second vehicle.
34. PERSON_ID: Text field representing the person involved.



35. SEX: Text field indicating the gender of the involved person.
36. AGE: Number field representing the age of the involved person.
37. INJ_LEVEL: Text field indicating the level of injury sustained by the person.
38. SEATING_POSITION: Text field indicating the seating position of the person in the vehicle.
39. SAFETY_EQUIPMENT: Text field indicating the safety equipment used by the person.
40. ROAD_USER_TYPE: Number field indicating the type of road user involved in the accident.
41. ROAD_USER_CATEGORY: Text field indicating the category of road user involved in the accident.
42. LICENCE_STATE: Text field indicating the state or territory where the person's license is issued.

43. PEDEST_MOVEMENT: Number field indicating the movement of the pedestrian at the time of the accident.
44. PEDEST_MOVEMENT_OTHER: Text field providing additional information about the pedestrian movement.
45. VEHICLE_1_FACTOR: Text field indicating the contributing factor related to the first vehicle.
46. VEHICLE_2_FACTOR: Text field indicating the contributing factor related to the second vehicle.
47. PERSON_FAT: Number field indicating whether the person involved in the accident was killed (1) or not (0).
48. PERSON_SINJ: Number field indicating whether the person involved in the accident sustained a serious injury (1) or not (0).
49. PERSON_NINJ: Number field indicating whether the person involved in the accident sustained a non-serious injury (1) or not (0).
50. PERSON_MINJ: Number field indicating whether the person involved in the accident sustained a minor injury (1) or not (0).
51. PERSON_MAJINJ: Number field indicating whether the person involved in the accident sustained a major injury (1) or not (0).
52. PERSON_UNINJ: Number field indicating whether the person involved in the accident was uninjured (1) or not (0).
53. PERSON_FATAL: Number field indicating whether the person involved in the accident suffered a fatal injury (1) or not (0).
54. BICYCLE: Number field indicating whether a bicycle was involved in the accident (1) or not (0).
55. BUS: Number field indicating whether a bus was involved in the accident (1) or not (0).
56. CAR: Number field indicating whether a car was involved in the accident (1) or not (0).
57. MOPED: Number field indicating whether a moped was involved in the accident (1) or not (0).
58. MOTORCYCLE: Number field indicating whether a motorcycle was involved in the accident (1) or not (0).
59. OTHER_VEH: Number field indicating whether another type of vehicle was involved in the accident (1) or not (0).
60. PEDAL_CYCLIST: Number field indicating whether a pedal cyclist was involved in the accident (1) or not (0).
61. PEDESTRIAN: Number field indicating whether a pedestrian was involved in the accident (1) or not (0).
62. TRUCK: Number field indicating whether a truck was involved in the accident (1) or not (0).
63. VEHICLE_1_TYPE: Text field indicating the type of the first involved vehicle.
64. VEHICLE_2_TYPE: Text field indicating the type of the second involved vehicle.
65. VEHICLE_3_TYPE: Text field indicating the type of the third involved vehicle.

66. VEHICLE_4_TYPE: Text field indicating the type of the fourth involved vehicle.
67. VEHICLE_5_TYPE: Text field indicating the type of the fifth involved vehicle.