****

**NANYANG TECHNOLOGICAL UNIVERSITY**

**ACADEMIC YEAR 2022/2023 SEMESTER 2**

**BC2407 - Analytics II: Advanced Predictive Techniques**

**Project Title:**

The Quest for Zero Fatalities: Utilising Machine Learning To Discover Key Factors To Reduce Singapore Traffic Accidents

**Final Report**

**Submitted By:**

Seminar Group Number: S03 Team Number: 05

|  |  |
| --- | --- |
| Chong Wei Kang | U2121461B |
| Loy Sheng | U2010347G |
| Ng Shang Yau | U2121960G |
| Nguyen Thi Thanh Mai | U2010484A |
| Toh Yi Cheng | U2010698A |

# **Table Of Contents**

[**Table Of Contents 2**](#_o5nhu7fnmv0m)

[**1. Executive Summary 4**](#_z5qrlsffmoub)

[**2. Business Problem 5**](#_g2thxmopwc7)

[2.1 Overview of Land Transport Authority (LTA) Singapore 5](#_nlnpjwhusimn)

[2.2 Defining the Business Problem 5](#_vb49ut50hawy)

[2.3 Current Measures 6](#_52g52vl6jd1e)

[**3. Objectives 6**](#_xqc0grsbib99)

[3.1 Business Objectives 6](#_b464dug8s0y9)

[**4. Analytics Solution 7**](#_o6tbyre2vg9a)

[4.1 Key Business Questions 7](#_88apwbo0xen4)

[4.2 Predicting Accident Severity with Machine Learning 7](#_bkqsk94v9lo7)

[**5. Preparation of Datasets 9**](#_3inqi7qif83a)

[**6. Data Exploration and Visualisation 11**](#_yh4zig3dae45)

[6.1 Human Factors 11](#_2plf8awkfq8t)

[6.2 Road and Environment Factors 12](#_7mc03jnp6mql)

[6.3 Driver Behaviour Factors 13](#_ifznnnege8gu)

[**7. Prescriptive Techniques and Basis for Comparison 13**](#_sbk001cxi0wl)

[**8. Model Analysis 14**](#_datu5qpcduku)

[8.1. Logistic Regression 14](#_sdxnaretlkf3)

[8.2. Multivariate Adaptive Regression Splines (MARS) 15](#_nubphawouzty)

[8.3. Neural Network 16](#_egrlydustcxf)

[8.4. Random Forest 17](#_uq5jnl6uxfb9)

[8.5. Comparison 18](#_cv6djj8cnjwq)

[**9. Identification of Key Predictors 18**](#_n99je98as1sc)

[9.1 Use of Restraint System (Restraint\_Equipment\_Usage) 18](#_e3bwwvtqmid3)

[9.2 Number of Occupants (Num\_Of\_Occupant) 19](#_tzua28cdwj64)

[9.3 First Harmful Event of Crash (Harmful\_Event) 19](#_b6l2rxlc862v)

[9.4 Driver Related Factors (Driver\_Factors) 19](#_m01tp5plj9np)

[9.5 Instances of Hit and Run (Hit\_And\_Run) 19](#_gg76b3nfbddf)

[9.6 Age of the Person Involved (Age) 19](#_5qxrk7m4t7yk)

[9.7 Constructing a Minimal Dataset 19](#_x7pnyzwuzc6)

[**10. Recommendations 20**](#_y78z71tdhcca)

[10.1 Use of Restraint System 20](#_q5spi280bro2)

[10.2 Instances of Hit and Run 20](#_manuovif5ybx)

[10.3 Driver factors 20](#_xd9cslihksxk)

[10.4 Harmful events 21](#_v2e1c3cu2m9d)

[10.5 Curated Prediction Tools 22](#_gvihl6m25njq)

[**11. Discussion 22**](#_n17myqlhkvlh)

[11.1 Limitations 22](#_sx4osdahp59l)

[11.2 Future Work 23](#_p3lt22wy6muo)

[**12. Conclusion 24**](#_qw3yn0fo5ssn)

[**13. References 25**](#_lx5qw7izl45m)

[**14. Appendix 34**](#_fik0gucplla2)

[Appendix A: Background 34](#_kd6he6p97i2n)

[Appendix B: Overall Data Pipeline 35](#_jaikd2afr4r)

[Appendix C: Data Preparation 36](#_ff31x27iyesc)

[Appendix D: Data Exploration & Visualization 37](#_f05jc8fovy09)

[Appendix E: Logistic Regression 41](#_rm2u3kpmxmq5)

[Appendix F: MARS 58](#_xfqp9scfph18)

[Appendix G: Neural Network 60](#_6xnsgvzgiike)

[Appendix H: Random Forest 64](#_rw0pegad39kg)

[Appendix I: Variable Importance of The 4 Models 67](#_vb4bohy8f8me)

[Appendix J: Random Forest on Minimised Dataset 70](#_8bbr9mczu1io)

[Appendix K: Curated Prediction Tool 71](#_52c5544i40q6)

# 

# **1. Executive Summary**

This report investigates the potential of Analytics and Machine Learning to enhance road safety in Singapore. As Singapore adapts to embracing Covid-19 as the new norm, the resumption of daily activities has led to an uptick in traffic accidents and the associated fatalities (Lim, 2023). Despite dedicated efforts, traffic accidents and fatalities persist, posing a **multifaceted** challenge.

With the latest breakthroughs in machine learning, this report explores the potential of using machine learning models to develop data-driven recommendations for mitigating the occurrence of such devastating incidents, ultimately **reducing the significant costs** of traffic accidents and enabling the Singapore government and Land Transport Authority (LTA) to **allocate their resources more effectively**. To address this crucial issue, our team employed the Fatality Analysis Reporting System (FARS) dataset from the National Highway Traffic Safety Administration (NHTSA) in the United States as the foundation for our analysis.

To ensure that our recommendations and analysis remain as applicable to Singapore as possible, we have pre-processed the datasets with an initial round of feature selection, focusing on aspects pertinent to Singapore, such as excluding rural areas, four season weather. This was achieved through meticulous **cross-analysis with online research**. The flow of our analytical approach will be as follows:

1. Initially, this report seeks to pinpoint the key factors associated with fatal accidents. This is done with the help of Tableau’s data visualisation and interactive dashboarding to examine the complex interplay of multiple factors, including human factors, road design, and weather conditions.
2. Following this, we looked at solidifying the **correlation** between these factors and accident fatalities by assessing and quantifying their significance. To achieve this, we employed four supervised machine learning models: Logistic Regression, Neural Network, Multivariate Adaptive Regression Splines (MARS), and Random Forest. By using key performance metrics such as accuracy, False Negative Rates (FNR) and F2 Score to evaluate our models, we concluded on the use of Random Forest as the optimal predictive model for our response variable, “Fatality” of a traffic accident as it exhibits the highest accuracy, F2 score and lowest FNR.
3. By weighing the respective variable importance from the four trained predictive models, the most significant features resulting in road accident fatalities are identified. The top 6 features are - “Use of Restraint Systems”, “Number of Occupants”, “Harmful Event”, “Driver Behaviours”, “Instances of Hit-and-Run” and “Age of Person Involved”. Based on the identified significant factors, our team then presents **five** targeted recommendations which are derived from analytics insights, designed to complement existing strategies by LTA and Traffic Police.

Nevertheless, we acknowledge the limitations of applying US data to Singapore's unique road context. Additionally, the "black box" nature of our models, coupled with minimal feature selection techniques, could potentially lead to suboptimal interpretation of model results. Despite these challenges, this report emphasises the potential of our proposed recommendations to **leverage the wealth of traffic data** collected by the LTA, ultimately driving significant improvements in road safety.

# **2. Business Problem**

## **2.1 Overview of Land Transport Authority (LTA) Singapore**

The Land Transport Authority (LTA) is the statutory board of the Singapore government responsible for planning, operating, and maintaining Singapore's land transport infrastructure and systems. The LTA's mission is to provide a safe, efficient, and sustainable land transport system that meets the evolving needs of Singapore's residents and businesses.

Given the LTA's commitment to ensuring road safety, this report seeks to investigate the potential of analytics and machine learning in enhancing Singapore's road safety measures. Specifically, this report will explore how analytics can be used to predict and prevent traffic accidents, with the ultimate goal of reducing the number of accidents and improving overall road safety in Singapore.

## **2.2 Defining the Business Problem**

Although Singapore is generally considered a safe place, it was found to have the second highest road fatality rate when compared to four other cities of similar size (Budget Direct Insurance, 2023). Specifically in Singapore, the Traffic Police reported that the number of road traffic accidents rose in 2021, resulting in an increase in both fatalities and injuries. It was reported that 100 people lost their lives on the roads in 2021, compared to 80 in 2020 [*(Figure A-1)*](#r9wcl1o0r23o). The total number of fatal and injury accidents also rose, from 5,556 in 2020 to 6,039 in 2021 [*(Figure A-2)*](#7zgvt78dvq98) (Singapore Police Force, 2021). The sharp decrease in the number of fatal and injury road accidents in 2020 was primarily attributed to the stringent COVID-19 measures implemented in Singapore the previous year, including the nationwide lockdown. However, as the country gradually lifted these restrictions and saw an increase in commuter traffic, the number of road accidents also increased. This highlights the need for continual efforts to prioritise road safety, even in the face of changing circumstances. The reason behind this is that traffic accidents incur significant expenses, with conservative estimates suggesting that the yearly costs approach nearly $610.3 million (Chin et al., 2006).

Reducing road traffic accidents remains a challenge for the Singapore government despite its efforts to encourage safe driving habits. The problem is multifaceted and arises from multiple root causes, making it difficult to address effectively. For example, factors such as human error, vehicle malfunction, road design, and weather conditions can all contribute to accidents. The main cause of pedestrian fatalities in Singapore was identified as jaywalking in 2022, suggesting that drivers are not the only cause of road traffic accidents (Shiying, 2022). Pedestrians may have a role to play as well.

Additionally, the Heinrich Domino Theory suggests that even a small error in one of several interrelated elements can lead to a chain reaction that results in a major accident. This highlights the challenge in addressing road traffic accidents. Thus, in a bid to reduce the number of accidents and fatalities, we propose a holistic approach using Machine Learning to consider the interplay of these various factors.

## **2.3 Current Measures**

Firstly, Singapore’s Intelligent Transport System (ITS) in Singapore is a network of cutting-edge gadgets, sensors, and cameras that operate around the clock to collect and disseminate data on traffic flow, journey times, and road demand, thereby enhancing road safety (LTA, 2023). Additionally, the ITS leverages real-time dynamic data and advanced analytics to optimise transportation networks, reduce traffic accidents, and boost safety in Singapore. With these systems, drivers can access real-time information about road conditions, enabling them to make informed travel plans and avoid accident-prone areas. Furthermore, The Expressway Monitoring and Advisory System (EMAS), a component of the ITS, detects accidents and incidents in real-time, automatically adjusting traffic signals, alerting emergency services, and redirecting traffic flow as needed (LTA, 2023). EMAS enables swift response from recovery crews to respond to accidents, an impressive average 24 minutes faster (Centre for Liveable Cities, 2018), and curbs the risk of secondary accidents occurring.

Secondly, the Traffic Police has also mandated simulation training for all learner motorists as of 16 December 2019 as part of ongoing efforts to enhance road safety and decrease traffic accidents involving inexperienced drivers. . The training scenarios are based on the top 10 causes of traffic accidents that frequently cause injuries. Through simulation training, learner drivers can practise defensive driving or riding in a safe and controlled environment. Thus, this teaches new drivers basic and safe driving practices and prepares them for a variety of road circumstances (Singapore Police Force, n.d.).

Despite the LTA and Traffic Police's current efforts to reduce traffic accidents, accident frequency remains high. Therefore, in this paper, we will present additional targeted and comprehensive recommendations in Section 10, designed to complement existing strategies and ultimately reduce traffic accidents in Singapore.

# **3. Objectives**

## **3.1 Business Objectives**

Our group's objective is to **alleviate the substantial costs** associated with traffic accidents, particularly those resulting from severe injuries and fatalities. Specifically, traffic fatalities alone are estimated to cost S$1.273 million each (Chin et al., 2006). As such, by minimising the incidence of road accidents, we strive to assist the LTA and the Singapore Government in **allocating resources more efficiently** across various sectors of the economy and society, ultimately contributing to a brighter future for all.

# 

# **4. Analytics Solution**

## **4.1 Key Business Questions**

|  |  |  |
| --- | --- | --- |
| **Business Question** | **Explanation** | **Approach** |
| What are the factors that contribute to the **occurrence** of fatal accidents? | This question aims to explore the trends or patterns of specific human, road, and environmental factors that affect the likelihood of fatal accidents occurring.  Answers to this question will provide a comprehensive understanding of how various factors such as driver behaviour, road design, vehicle type interact and affect the likelihood of fatal accidents. | Data Visualisation |
| What are the **most significant factors** to **predict** the accident outcome whether one **dies** in a road accident? | This question aims to identify the most critical factors leading to fatalities in road accidents by constructing predictive models.  Answers to this question will enable the LTA to develop data-driven, targeted interventions that specifically address these key factors, ensuring the most efficient use of resources and the highest possible impact on road safety. This analysis will also support informed decision-making by prioritising areas with the greatest potential for reducing fatalities. | Modelling Approach using Logistic Regression, MARS, Neural Network & Random Forest |

## 

## **4.2 Predicting Accident Severity with Machine Learning**

Even though the data is obtained from the United States, it can be harnessed to generate valuable insights for Singapore’s context. Our team will clean and process the data carefully to increase its accuracy and interpretability for LTA.

Through utilising Tableau's data visualisation capabilities, we will explore trends or patterns that might increase the occurrence of severe accidents with fatal outcomes. Tableau’s user-friendly interface enables us to explore the complex interplay between various factors and assess their impact on fatality likelihood. Besides visualisations, we will conduct a cross-analysis of our findings using variable importance derived from predictive models, which is essential in ensuring the reliability of our results. This process is a crucial element for accomplishing our business objectives as it allows us to accurately identify the most important variables that contribute fatalities. Furthermore, accurately pinpointing the critical factors involved in severe accidents enables us to formulate targeted recommendations to reduce the amount of fatalities, ultimately leading to **efficient resource allocation** and minimised time wasted [*(Figure B-1)*](#x754gv6u4hwf)*.*

Moreover, reducing fatalities on Singapore's roads would lead to a **decrease in the costs associated** with traffic accidents in the country. Since human lives are at stake, it is crucial to assess the performance of the machine learning models used. To accomplish this, we aim to evaluate the models holistically across multiple performance metrics, ultimately determining the primary model that we will recommend for LTA’s adoption. This process includes adjusting the model's input parameters to account for new data in Singapore's context, ensuring that we obtain a model that can most accurately predict fatalities and identify key variables affecting fatalities on Singapore's roads.

**4.3 Evaluation of Analytics Outcome**

Our response variable, denoted as Y, represents the outcome of an accident, indicating whether it resulted in a fatality or not for drivers and occupants. This metric is of paramount importance for the LTA to monitor as it reflects the safety and well-being of individuals following an accident. Moreover, it enables the LTA to evaluate and assess the effectiveness of any initiatives aimed at improving road safety. Additionally, this variable is in line with the international standard used by the World Health Organization for comparing the road safety status of different countries, which is Deaths per 100,000 people. Understanding and analysing the trends and factors contributing to fatal accidents can aid in the development of policies and interventions aimed at reducing the number of deaths on the road.

Furthermore, our team would like to suggest 3 key performance indicators with a focus on ensuring they are verifiable, precise, and unambiguous, allowing for clear assessment of the analytics solution's impact. We also devised various performance measurements that the LTA can track over time, providing a comprehensive understanding of the effectiveness of the implemented analytics solution and facilitating future data-driven decision-making in road safety improvements.

| **Key Performance Indicators** | **Performance Measurements** |
| --- | --- |
| 1. **Reduction of road fatality rate** per 100,000 citizens by 10% **within 5 years** | Record and compare the annual number of deaths related to road accidents per year after implementing the analytics solution to the baseline figure prior to implementation |
| 1. Increase the **survival rate** of road accident victims by 15% **within 5 years** | Record and compare the number of people who died in a crash, en route to medical and those who survived per year after implementation of the analytics solution. |
| 1. **Decrease costs** associated with road accident fatalities **by 20% within 5 years** | Record the estimated economic costs due to loss of workers, welfare costs, insurance payouts, healthcare and other costs attributable to road accidents per year after implementation of the analytics solution. |

# **5. Preparation of Datasets**

**5.1 Data Acquisition & Overview**

Our group used the Fatality Analysis Reporting System (FARS) dataset, maintained by the National Highway Traffic Safety Administration (NHTSA) in the United States, for our analysis of traffic accidents. This comprehensive dataset offers a wealth of information on fatal motor vehicle crashes that occurred on public roads in the US, including nuanced details on the crash location and time, the types of vehicles involved, and individual-specific characteristics such as age, sex, and alcohol involvement.

**5.2 Data Pre-Processing**

Due to the extensive number of data files in the FARS dataset, our group opted to merge multiple CSV files to obtain the essential information needed for our analysis. This merging and initial feature selection process was guided by literature reviews that explored the impact of various factors, such as traffic violations, roadways, vehicles, and environmental conditions on traffic accidents. By eliminating insignificant factors, we aim to bolster our models’ predictive capabilities. Additionally, as we tailor our study to Singapore's context, we will only concentrate on pertinent factors that can be applied locally. Finally, recognizing that traffic accidents are often highly complex and influenced by numerous factors, we have classified these factors into five groups based on our literature review:

1. **Vehicle Characteristics**

Firstly, the type, weight, and speed of a vehicle all impact the severity and fatality of traffic accidents. Each vehicle type has varying risk factors, with cars and motorcycles having increased severe injuries in good weather, while buses experience severe injuries mostly tied to pedestrian collisions (George et al., 2017). Furthermore, heavier vehicles tend to be safer for their own occupants but more hazardous for other motorists, and accidents involving heavier vehicles result in more severe outcomes. (Anderson & Auffhammer, 2011) High vehicle speeds is also a major causative factor for road accidents with severe injuries. (Alrejjal et al. 2022).

1. **Driver Characteristics and Behaviours**

Secondly, numerous studies have demonstrated that distracted driving heightens the driver's reaction time, which in turn elevates the likelihood of accidents. Liang's summary underscores the influence of distinct distractions and their specific types on driving performance (Liang, 2020). Moreover, alcohol consumption fosters impulsive and risk-prone behaviour, impaired cognition, slower reaction time, and lower hazard recognition, all of which heighten the probability of traffic accidents (Zhao et al., 2014). Drunk driving has also been pinpointed as a prevalent cause of road accidents in Singapore (John, 2018). Other driver characteristics, such as age (Regev et al., 2018) and seating arrangement in the motor vehicle, can impact the likelihood and severity of traffic accidents (Mayrose, 2008).

1. **Non-driver Characteristics and Behaviours**

Third, traffic accidents also encompass pedestrian-related incidents which pose a significant concern, particularly in areas with heavy pedestrian traffic. Pedestrians intoxicated by alcohol are more prone to be involved in fatal motor accidents (Lasota et al., 2020). In Singapore, elderly pedestrians constitute the majority of fatal pedestrian accidents, with a considerable proportion resulting from jaywalking (Channel NewsAsia, 2023).

1. **Road Characteristics**

Fourth, road conditions and features have a substantial impact on road accidents, with wider roads and highways generally experiencing a higher number of incidents (Malin et al., 2018). Hazardous roads can arise from factors such as poor lighting, steep downhill slopes, and surface damage (Pembuain et al., 2019), while weather conditions like heavy precipitation and snowfall also contribute to creating hazardous road conditions (Malin et al., 2019).

1. **Environmental Conditions & Timing factors**

Lastly, adverse weather significantly affects road conditions. Factors such as visibility, and the likelihood of single- versus multiple-vehicle crashes influence the overall impact of traffic accidents, highlighting the complexity of weather-related effects (Becker et al., 2022). Certain months show higher road accident rates, potentially due to seasonal weather changes (Alrejjal et al. 2022). Accidents are also more common during nighttime, especially in urban areas (National Center for Statistics and Analysis, 2022).

**5.3 Data Merging**

Considering the five categories above, we will merge data from Accident.csv, Person.csv, Vehicle.csv, Distract.csv, and Driverrf.csv for our analyses. First, Driverrf.csv and Distract.csv will be combined with Vehicle.csv. Next, this merged Vehicle.csv will be joined with Person.csv using ST\_CASE and VEH\_No as identifiers. Finally, the merged Person.csv will be combined with Accident.csv using ST\_CASE as the identifier, consolidating all relevant information for our analysis [*(Figure C-1)*](#xmbqd2dz37s8). The final dataset contains 20,203 rows and 42 columns, including 33 categorical, 6 continuous, and 3 identifier columns.

**5.4 Initial Data Cleaning**

During the initial data cleaning process, we eliminated the ST\_CASE, VEH\_NO, and PER\_NO columns, as they function solely as identifiers and do not contribute to our predictions. Additionally, we assigned appropriate labels to categorical variables to facilitate interpretation, with our data dictionary providing detailed explanations for each label. When encountering unknown data, previously encoded as '99', '998', or '999', we substituted it with the mode or median of the respective column, based on the data type. However, if the proportion of unknown data surpassed 10% of the total rows, we opted to create an 'Unknown' column instead of attempting to impute the data, thus preventing potential class imbalance issues.

**5.5 Regrouping Variables**

While cleaning the data, we found many columns that had multiple categories and factors, including our target Y variable, Injury\_Severity, which had eight levels. To simplify it for analysis, we renamed this column as 'Death' and combined the categories where a person had light injuries into one group called 'No' and the categories where a person had died due to fatal injuries into another group called 'Yes'. This resulted in our dataset having 54% of rows with Deaths and 46% of rows with no Deaths [*(Figure D-1)*](#_f05jc8fovy09)*.*

We adopted a similar approach for other features, as having high cardinality can adversely affect the performance of applied machine learning models. For instance, HARM\_EV had a very high cardinality with a total of 57 categories. Having too many categories increases the risk of the model becoming too specialised to the training data and not generalising well to future test data. (Sangani, 2021)

To avoid the issue of high cardinality, we also explored regrouping several variables to reduce the number of categories. We chose not to use one-hot encoding as it could result in multicollinearity and the curse of dimensionality, both of which can negatively affect machine learning model accuracy. (Yiu, 2019) Our data dictionary also provides more information on our methodology for combining the several variables.

# **6. Data Exploration and Visualisation**

In this section, we conduct data exploration to gain insights into how road and environmental factors, human factors, and driver behaviour contribute to fatal accidents. Through this process, we were able to identify initial insights and trends that helped inform our subsequent analysis.

## **6.1 Human Factors**

Firstly, the dashboard [(*Figure D-3*)](#teatj5fic861) revealed interesting insights regarding human-related factors on how they impact the fatalities, as below:

1. ***Younger individuals were more involved in fatal accidents than their older counterparts.*** This trend may be attributed to several factors, such as lack of experience, risk-taking behaviour, and overconfidence in their driving abilities. This finding is also consistent with the Singapore Police Force's report, which highlights the age group of 20 to 29 years old as the most susceptible to fatal accidents, accounting for 23.9% of all fatal accidents from January to September 2021. In contrast, the age group with the lowest number of fatal accidents were those aged 70 years and above, accounting for 2.9% of all fatal accidents (Singapore Police Force, 2021).
2. ***Men had a higher rate of fatalities from road accidents compared to women***. This may be due to men having a higher tendency to engage in risky driving behaviours such as driving under the influence of drugs or alcohol, speeding, or not wearing a seatbelt. Additionally, men generally drive more miles than women, which increases their exposure to the risk of accidents. Globally, males have been found to be at a higher risk of road traffic crashes than females from a young age, with young males being almost three times more likely to be killed in a crash than young females (World Health Organization, 2022).
3. ***Light vehicles were more likely than other types of vehicles to be involved in fatal accidents***. Light vehicles, such as motorcycles, cars, vans, and pick-ups, are more commonly used for everyday transportation and are more susceptible to accidents due to their size and speed. They also have an increased risk of rollovers or collisions with larger vehicles, making them more likely to be involved in fatal accidents. In the 2021 Annual Traffic Statistics Report by the Singapore Police Force, light vehicles had the highest number of fatal accidents, with 113 fatalities out of a total of 172 (Singapore Police Force, 2021).
4. Interestingly, ***higher fatalities were observed for drivers travelling at average speeds on dry roads.*** One possible explanation is that vehicles travelling at the speed range of 70-73 km/hr on dry road conditions may be in a "danger zone" where drivers feel comfortable driving at this speed range but may not be able to react quickly enough to unexpected situations. In contrast, drivers may naturally slow down even more on wet road conditions due to decreased traction, which can reduce the likelihood of a fatal accident.
5. ***Lack of restraint equipment usage increases risk of fatalities in road accidents***. Fatalities were higher among vehicle occupants who did not use restraint equipment, despite the nearly equal proportion of fatal and non-fatal casualties. This suggests that utilising seat-belts can lower the likelihood of fatalities by 45-50% for drivers and front seat passengers, and by 25% for rear seat passengers (World Health Organization, 2022).
6. ***Vehicles with fewer occupants are more vulnerable to fatal accidents***. One plausible explanation for this trend is that the number of occupants in a vehicle affects the overall weight, and a heavier vehicle tends to provide superior crash protection. Heavier vehicles are more likely to continue moving forward after a collision, thereby reducing the force experienced by the occupants (Insurance Institute for Highway Safety, n.d.). As a result, a vehicle with fewer passengers will be exposed to a larger crash force, increasing the likelihood of serious injuries and possibly death, assuming all other factors remain constant.

## **6.2 Road and Environment Factors**

Secondly, the dashboard [(*Figure D-4)*](#q01z4p24jylp)also uncovered several intriguing findings related to road and environmental factors:

1. ***Fatal accidents are more likely to occur in non-junction/non-intersection locations and roadways***. These areas may have a higher incidence of accidents due to the potential for conflicting movements of different road users, such as vehicles, pedestrians, and cyclists. Additionally, these areas often lack the infrastructure to manage the flow of traffic and ensure safety, such as the absence of crosswalks, pedestrian crossings, or traffic signals. In addition, non-compliance with traffic rules and regulations by road users in these areas can further increase the risk of accidents.
2. Surprisingly, ***light conditions have little effect on the risk of fatality in traffic accidents***. This may be due to the fact that drivers tend to adjust their speed and driving behaviour in response to varying light conditions, which could reduce the likelihood of accidents.
3. ***Fatal accidents occur more frequently in two-way divided traffic flow than in other types of traffic flow***. Two-way divided traffic flow is the most common type of traffic flow in urban areas and often involves higher speed limits, which can increase the severity of accidents. This type of traffic flow also has a higher likelihood of accidents occurring due to the increased interaction between vehicles and pedestrians. The complexity of navigating intersections in two-way divided traffic flow can also lead to confusion and mistakes by drivers, further increasing the risk of accidents.

## **6.3 Driver Behaviour Factors**

Lastly, the dashboard [*(Figure D-5)*](#59bpix5s0jq) also offered some insights on driver-behaviour factors on how they impact the fatalities.

1. The data suggests that ***certain driver behaviours increase the likelihood of fatal accidents***. For example, driving under the influence of alcohol or at high speeds can greatly increase the risk of a fatal accident. Additionally, non-compliance with traffic laws, carelessness, improper lane usage, and negligence were identified as prevalent factors in fatal accidents. These findings highlight the importance of promoting responsible driving behaviour and enforcing traffic laws to reduce the number of fatalities on the road.

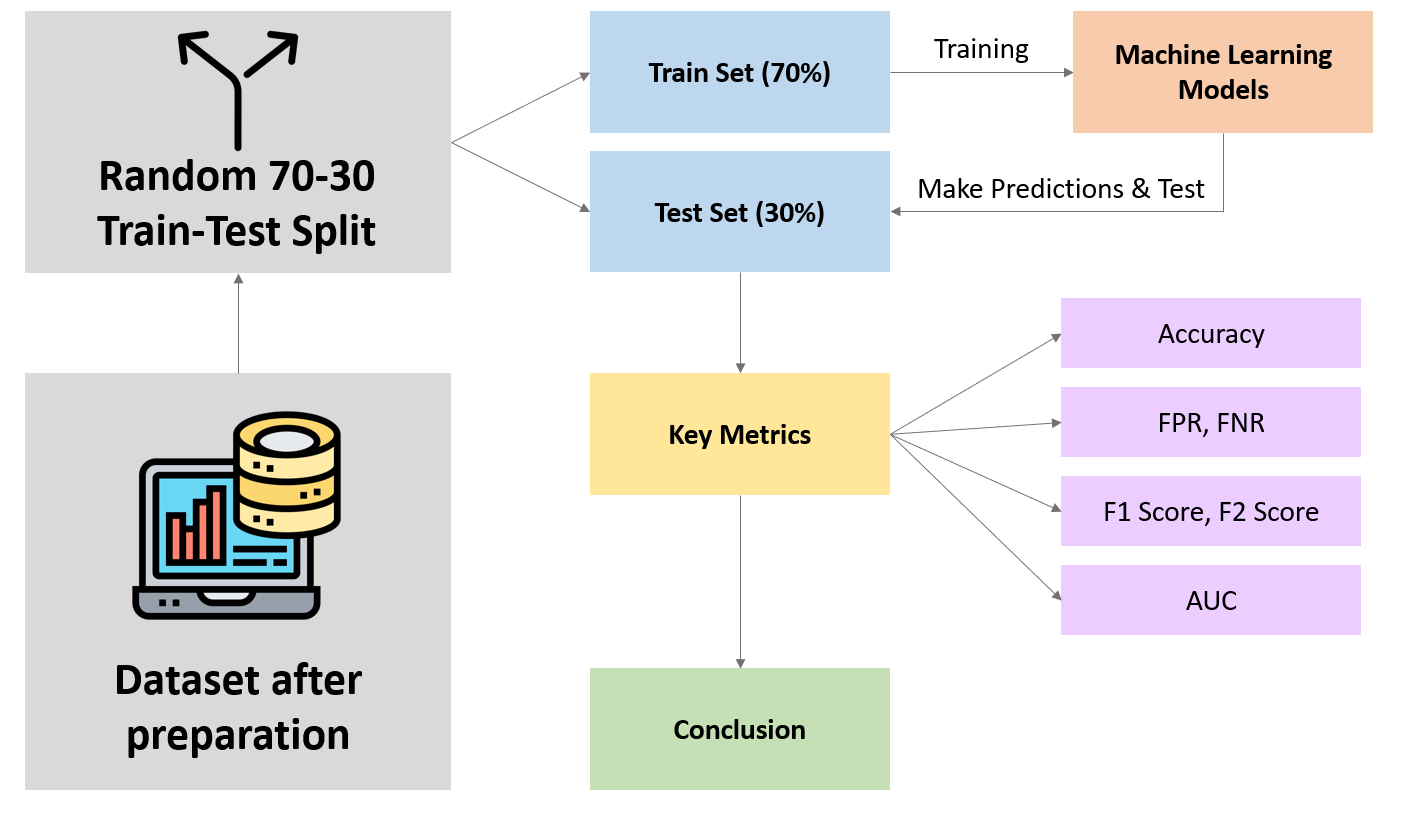
# **7. Prescriptive Techniques and Basis for Comparison**

Our project delves into exploring the factors that magnify the chances of dying in accidents. Thus, we aim to predict the likelihood of death with the 'Death' column as our target Y variable using four classification models, namely Logistic Regression, Multivariate Adaptive Regression Splines (MARS), Neural Network, and Random Forest.

To compare and assess the performance of each model, we first split our dataset into training and testing sets, with a ratio of 70-30, after conducting the necessary data preparation steps as outlined in Section 5 of our report. This was to ensure fair evaluation and comparison of the models. Given that the proportion of fatalities and non-fatalities is approximately equal [*(Figure D-1)*](#_f05jc8fovy09), we also deemed conducting sampling techniques on the data unnecessary.

To gauge the predictive capabilities of our models, we chose several specific metrics to evaluate them.

|  |  |
| --- | --- |
| **Metric Used** | **Definition** |
| Accuracy | Percentage of correct predictions compared to all predictions |
| False Positive Rate (FPR) | ***False Positive (FP) = Predicted a non-fatal accident as fatal***  Percentage of FP compared to all non-fatal accident predictions |
| False Negative Rate (FNR) | ***False Negative (FN) = Predicted a fatal accident as non-fatal***  Percentage of FN compared to all fatal accident predictions |
| Area Under ROC Curve (AUC) | Measure of classifiers’ ability to differentiate between classes. (Higher scores relates to better performing classification) |
| F1 Score[[1]](#footnote-1) |  |
| F2 Score1 |  |

Accuracy by itself, while essential, provides limited insight given the context of our research. As such, the F2 score was also considered since it gives more weight to the False Negatives compared to the F1 score. Our primary objective is to minimise False Negatives as they could lead to loss of lives and considerable costs. Specifically, a higher weight of 0.8 is used for false negatives. A **higher false negative count is penalised** with a lower F2 score, as the denominator is larger. 

With these metrics in mind, our group will then draw a conclusive evaluation of which model is most suitable for our use case, and assist LTA in identifying the critical predictors for forecasting accident fatalities using that model. *(Figure 1)* We will also dive into a more detailed discussion of the various models' predictive capabilities in the upcoming section.

*Figure 1. Summary of methodology to compare models’ predictive capabilities*

# **8. Model Analysis**

## **8.1. Logistic Regression**

Logistic regression is a supervised machine learning algorithm that predicts the likelihood of an event happening based on one or more independent variables (Bonthu, 2021). The algorithm then determines the relationships between the independent variables and the dependent variable. For our project, we utilised the binary logistic regression because our dependent variable is binary in nature, i.e., whether there is an instance of a death or not in the traffic accident (Statistics Solutions, 2021).

The logistic function, also known as a sigmoid function, is used to model the relationship between the factors and the outcome (Bonthu, 2021). It produces a value between 0 and 1, representing the probability of the event occurring. The logistic regression model will then return a binary outcome based on a specific threshold value that is predetermined by us.

We first started our first model, lm1, by including all the independent variables [(*Figure E-1*)](#ygys8pggqyeh). Then, we used the step() function to remove the insignificant independent variables in a stepwise manner. Backward elimination is a crucial step in our model building process as it helps us remove irrelevant independent variables that do not contribute to the prediction of the outcome. By removing these variables, we can simplify the model and make it more interpretable. (Javatpoint, n.d.). We removed variables that had a significance level of at least 1%, which was determined using the p-values provided in the summary statistics of each model using a stepwise approach. This process was repeated until the 4th step where our model, lm4, had all significant independent variables. [(*Figure E-6*)](#sfc7wc7fg4lf)

There was also an issue of multicollinearity. Multicollinearity is a statistical concept in which there is an occurrence of high correlations amongst several independent variables. This was an issue because it will undermine the statistical significance of the dependent variable (Hayes, 2023). Furthermore, this violated the assumption of the logistic regression, which assumes an absence of multicollinearity (Statistics Solutions, 2021). To tackle the issue of multicollinearity, we utilised the variance inflation factor (VIF) function, which produced the generalised VIF (GVIF) score for each independent variable (Agarwal, 2022). A high GVIF score suggests that there is high multicollinearity between the independent variables, which can affect the validity of the regression model. Hence, we took a similar step-wise approach and removed variables with the highest GVIF. Thereafter, Critical\_Activity, and Vehicle\_Classification was removed, which resulted in our final model, lm6. As shown in [*Figure E-10*](#esuaza8r2gpj), the lm6 model consisted only of significant variables, and there was an absence of multicollinearity between independent variables (GVIF score of less than or equal to 10).

The final performance of the logistic regression on the test set is detailed below.



*Figure 2. Logistic Regression Test Set Metrics*

## **8.2. Multivariate Adaptive Regression Splines (MARS)**

MARS is a form of regression analysis that allows for flexible modelling of high dimensional data (Friedman, 1991). MARS involves creating a flexible regression model by dividing the data into smaller segments and fitting a simple linear regression model to each segment. MARS can be interpreted as an ensemble of linear functions joined by hinge functions (Dobilas, 2020). These basis functions are generated through a stepwise search over all possible univariate knot locations and non-linear interactions among all variables using an adaptive regression algorithm (Zhang & Goh, 2016).

In our project, multiple MARS models were built to identify the optimal number of degrees of interaction, represented as “Degrees”, to create the model with minimal prediction error. “Degrees'' in MARS modelling refers to the maximum degree of input terms in the regression function, which refers to the maximum number of features involved in each interaction. It is one of the two main hyperparameters that can be tuned to improve the predictive strength of MARS. The other hyperparameter being “nprune”, which is the maximum number of terms in the final pruned model, was excluded in the tuning process since preliminary testing shows that not limiting it results in the most accurate test results. We varied the number of “Degrees” from the default 1 degree, which refers to no interaction terms, to 4 degrees. Apart from “Degrees”, all other parameters were left at default settings.

Based on the results of the 4 models, as shown in [*Figure F-1*](#_xfqp9scfph18), the most optimal model is created when Degrees = 1. While the MARS model with 3 degree of interaction results in the best predictive accuracy of 86.68%, it falls short in terms of the false negative rate of 13.39%. On the other hand, the MARS model with 1 degree of interaction has the least false negative rate of 12.95%, and thus F2 score, with a comparable accuracy of 86.37%. A subsequent MARS model was then created with Degrees = 1, together with 10-fold cross validation. The cross-validated model had observed no change in test results [*(Figure F-2)*](#pgj5hgrsiydu). This will be our optimal MARS model.

For MARS, the variable importance can be ranked either using General Cross Validation (GCV), which ranks features based on its extent of reducing the GCV estimate of error, or Residual Sum of Squares (RSS). Using its internal feature importance function, we were then able to identify the most significant features that result in traffic accident fatalities[*(Figure F-4)*](#mpm5rw79alhc).

## **8.3. Neural Network**

Neural Network is another machine learning algorithm that is modelled after the human brain. It consists of many interconnected processing nodes called neurons arranged in layers, and each layer performs a specific function. (Goodfellow et al., 2017). Using a process called "backpropagation," neural networks are capable of learning complex, non-linear correlations between input and output variables (Rumelhart et al., 1986). During training, the network modifies the weights of the connections between neurons to minimise the difference between expected and actual output. This process is repeated over multiple iterations until the network can accurately predict the output for new inputs.

Specifically, for our project, the Neural Network was trained with hyperparameters set to 100 epochs & batch size of 16. In other words, the model runs through the entire dataset a total of 100 iterations and processes 16 samples before the model’s parameters are updated during training. The neural network we used had two hidden layers, and one output layer for binary classification. Hidden layers process the input data and extract features, while the output layer produces the final prediction. The number of neurons in each layer determines the complexity and learning ability of the network.

After building our neural network, we compiled it using binary cross entropy as the method to measure the loss between the predicted and actual output. We also utilised the Adam optimizer with momentum, which allowed for efficient updating of the neural network weights during training. For the hidden layers, we used the Rectified Linear Unit (ReLU) activation function as it only activates neurons with positive inputs and makes the computation process faster. The output layer used the sigmoid activation function, which ensured that the output of the model was between 0 and 1 and was suitable for binary classification [*(Figure G-1)*](#9d2hlcldlekg)*.*

The neural network showed significant signs of overfitting as seen from the validation accuracy ([*Figure G-2*](#_6xnsgvzgiike)) and loss graph ([*Figure G-3*](#_6xnsgvzgiike)). The model had stopped training at 40 epochs due to the early callback function that had monitored the validation loss. This overfitting means that the neural network will perform worse on new data that the model has not seen before, decreasing its practical value for LTA. Some reasons for this could be too little input data and the model being too complex. Further tuning needs to be done to counteract this overfitting issue.

We utilised SHapley Additive exPlanation (SHAP) values to estimate feature importance in the neural network. SHAP values are a statistical approach based on game theory that can explain the output of any machine learning model. It allows us to understand the individual contribution of each feature in predicting the outcome of the neural network model.

The final performance of the neural network on the test set can be seen at [*Figure G-5*](#cm1lvde0bp77).

## **8.4. Random Forest**

Random Forest is a popular ensemble learning method that is based on Bootstrap Aggregation (Bagging) and utilises decision trees as its base learners. Each decision tree is built using different random samples with replacements from the original data and trained independently (i.e., Bootstrapping). Additionally, at each split of the tree, the model randomly selects a subset of features to consider for splitting the data, which is known as random subset feature selection (Breiman, 2001). In Random Forest, the model aggregates the predictions of all the decision trees and returns the majority vote as the final prediction for classification problems (Liaw & Wiener, 2002).

For our project, the Random Forest model has hyperparameters set to mtry=9 and ntrees=500. Mtry, refers to the number of variables randomly sampled as candidates at each split while ntrees refers to the number of trees in the forest. For the remaining hyperparameters, the default settings were used. We discovered that mtry=9 [*(Figure H-1)*](#kix.uze4n36teinj) resulted in the highest accuracies and that the out-of-bag error stabilised after 100 trees, which led us to determine that 500 trees were sufficient for training the model [*(Figure H-2)*](#kix.uze4n36teinj).

Additionally, the Random Forest model utilises the bagging technique to mitigate biases and reduce the influence of a strong predictor variable. It also aggregates the majority vote of 500 trees to overcome single-tree bias and prevent overfitting, thereby generating improved predictive accuracy on a new test set. The outstanding performance of Random Forest is evident in *Figure 12*, where it attains a remarkable 86.8% accuracy on the test set, exhibiting a lower false negative rate of 7.3% and a relatively higher F2 Score of 90.1% compared to the other 3 models, which is particularly significant in our given context.



*Figure 3. Random Forest Testset Metrics*

The randomForest package does offer a built-in solution to this issue through permutation feature importance. This technique allows us to gain insight into which features possess predictive power. The idea behind permutation feature importance is that if a feature contains valuable information and is randomly shuffled, the model's prediction accuracy will decrease. The degree of decrease in accuracy is indicative of the feature's impact on the model's predictions. (Billiau, 2021)

Our investigation reveals that several variables are highly significant in predicting whether an individual will die in a traffic accident. This is evidenced by the fact that removing any of these variables leads to a significant decrease in accuracy, as measured by the Mean Decrease in the accuracy of the predictions. Therefore, these variables should be given careful consideration in our analysis. [*(Figure H-3)*](#kix.yp0i9a7k2kl8)

## **8.5. Comparison**

We assessed the accuracy of the 4 models based on a set of categorical model evaluation metrics, namely: FPR, FNR, ACC, AUC, F1 Score, and F2 Score.

Figure 13 summarises the results of the test set performance for each model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **FPR** | **FNR** | **ACC** | **AUC** | **F1 Score** | **F2 Score** |
| **Logistic Regression** | 18.5% | 11.4% | 84.7% | 0.930 | 0.842 | 0.868 |
| **Neural Network** | 17.4% | 14.0% | 84.4% | 0.843 | 0.855 | 0.857 |
| **MARS** | **14.2%** | 13.0% | 86.4% | 0.947 | 0.854 | 0.864 |
| **Random Forest** | 18.2% | **7.3%** | **86.8%** | **0.955** | **0.865** | **0.901** |

*Figure 4. Summarised performance metrics for every model*

Based on observation, the Random Forest model performed significantly better than the other three models, especially in achieving a low False Negative Rate (FNR) of 7.3% and a high F2 Score of 0.901, as well as the highest predictive accuracy of 86.8%. This is particularly beneficial for our study, as we aimed to have a model with minimal FNR, as mentioned in [*Section 7*](#_sbk001cxi0wl)of our report. Although the MARS model had the lowest False Positive Rate (FPR), its overall performance was not as strong as the Random Forest model. As such, it can be concluded that the **Random Forest model** is the **most** suitable for our project.

# **9. Identification of Key Predictors**

To identify the key predictors for each model, we utilised the variable importance function which helps to establish the usefulness of a certain variable in predicting the dependent variable (Zvornicanin, 2022). To mitigate the limitations of individual variable importance methods and to account for differences in the feature sets used in each model, we selected features that were ranked in the top 6 by at least 3 out of the 4 models [*(Appendix I)*](#n9jbdq852py9). This approach enabled us to accurately identify the most important predictors. In this section, we will explore the reasons why these predictors were deemed most important.

## **9.1 Use of Restraint System (Restraint\_Equipment\_Usage)**

Occupants who did not use any form of restraint system have a higher likelihood of dying in an accident. Specifically, seat belts prevent drivers and passengers from being ejected in case of a collision, lowering the risk of mortality rate and serious injury by 45% and 50%, respectively (Centers for Disease Control and Prevention, n.d.). The use of seat belts is not only beneficial to front-seat passengers but also to rear-seat occupants, reducing fatality rates by 25-75% for the latter group (Nester, 2015).

## **9.2 Number of Occupants (Num\_Of\_Occupant)**

The number of occupants in a vehicle is also an important predictor. Possible explanations have been discussed in [*Section 6*](#j902ja1wsfu0) of our report.

## **9.3 First Harmful Event of Crash (Harmful\_Event)**

The first harmful event, denoting the first injury or damage-producing event of a crash, is an important predictor. Specifically, the **pedestrian** and **pedalcyclist** subcategories are the most important based on the output of our four models. According to the WHO, pedestrians, cyclists, and motorcyclists account for over 50% of all road traffic fatalities (World Health Organisation, 2022), mainly due to their lack of protection such as airbags (Yannis et al., 2020) and lower visibility on the road. These vulnerable road users are thus more likely to sustain injuries and fatalities in a collision compared to vehicle occupants.

## **9.4 Driver Related Factors (Driver\_Factors)**

Driver-related factors describe the conditions and situations of drivers involved in the crash. Our models ranked the **violence** subcategory as the most important predictor, which indicates whether the driver has committed a violent offense such as robbery or rape. (National Institute of Justice, n.d.) Studies have shown that individuals who previously committed violent crimes struggle to regulate their negative emotions, leading to anger and impulsiveness. (McNeill, 2017) This often results in aggressive driving behaviour, which can put themselves and other road users at risk of a fatal accident. (Flood Law, 2021)

## **9.5 Instances of Hit and Run (Hit\_And\_Run)**

Hit and run accidents, which refer to instances where the contact vehicle fails to stop and render aid to the victims, are important predictors. Such accidents have dire consequences, as delays in treating injured victims worsen the severity of injuries and increase the likelihood of death. Victims of hit and run accidents often do not receive prompt medical attention, which can worsen their injuries and result in traffic fatalities (World Health Organisation, 2022).

## **9.6 Age of the Person Involved (Age)**

The age of the occupants in a vehicle is also an important predictor. Possible explanations have been discussed in [*Section 6*](#tenkdsu8xrcq) of our report.

## **9.7 Constructing a Minimal Dataset**

We used the above 6 key features to construct a potential minimal dataset for predictions, reducing the need for extensive data collection for the LTA. Utilising random forest, the model achieved similar performance to the full model while using only ⅕ of the features. The results are shown in [*Appendix J*.](#_8bbr9mczu1io)

# **10. Recommendations**

Our recommendations focus on four key factors: use of restraint systems, first harmful event, driver-related factors, and instances of hit and run. Age and number of occupants were deemed beyond LTA's control and are not included in our recommendations. We

## **10.1 Use of Restraint System**

The Singapore government has made wearing rear seat belts mandatory for all automobiles registered in the nation since 1993. However, many continue to disregard these regulations. According to (Singapore General Hospital 2017), over 60% of back-seat passengers fail to wear seat belts, and they are 45% more likely than front-seat passengers and drivers to experience serious injuries.

To address this issue, we recommend two possible approaches. Firstly, a **nationwide road safety campaign** should be launched to educate and encourage the backseat passengers to wear their seat belts. Furthermore, the LTA should **work with automakers** to integrate extra safety measures to encourage backseat passengers to use their seat belts such as extending the beeping sound to remind drivers and all passengers to buckle up. Car manufacturers may also install extra sensors, auditory and visual warnings, and seat belt reminders to encourage safer driving and boost compliance with rear seat belt laws. The government could also provide incentives to automobile makers to emphasise and promote certain features. By taking these measures, the LTA can ensure that the public exercises proper road safety.

## **10.2 Instances of Hit and Run**

According to existing legislation, drivers who cause motor vehicle accidents must immediately stop their vehicles, even if the collision occurs with a stationary or parked vehicle with the owner nowhere in sight (Singapore Legal Advice, 2022). Failure to comply include penalties such as the at-fault driver banned from possessing or getting a driver’s licence for up to 12 months. Despite these penalties, hit-and-run accidents continue to occur in Singapore, causing injuries and fatalities. For example, just last month, there was an accident at Pasir Ris involving an elderly woman who died from injuries, with the van driver fleeing the scene (Mujibah, 2023).

Therefore, the LTA should take a multi-pronged approach to address this issue. Firstly, the government should enforce even **tougher** penalties, such as revocation of driving licences and denial of driving privileges to offenders. This would convey a clear message to potential offenders and serve as a deterrence to future hit-and-run cases. Secondly, the LTA should **increase the deployment of surveillance cameras** to monitor traffic and detect hit-and-run drivers. This would not only aid in catching the culprits, but the cameras’ footage can also be used as evidence in court to hold the offenders accountable for their misconduct.

## **10.3 Driver factors**

In Singapore, drivers who have been convicted of non-traffic violence charges may face licence suspension or disqualification if the offence is related to their driving conduct or ability to drive safely. To strengthen the current regulations, we propose a set of recommendations aimed at enhancing restraint, behaviour monitoring, and counselling for individuals with a violence charge.

Firstly, we suggest that the driver's licence of individuals with a violence charge be **restricted**, such as limiting their driving hours or routes. Additionally, they may be required to **install an ignition interlock device** that requires passing a breathalyser test before starting the car. Secondly, **mandatory counselling** or therapy sessions should be introduced to address the underlying issues that led to the violent behaviour, instead of being determined on a case-by-case basis. Lastly, **increased monitoring measures** could be implemented, such as requiring the reporting of driving activities to a probation officer or other authority.

These initiatives have been successfully enforced in various countries. For example, some states in the United States have mandatory licence restrictions for drivers with a violence charge and require the use of ignition interlock devices, which have prevented over 2.3 million drunk driving attempts since 2006. In Canada, Ontario has introduced a law requiring ignition interlock devices for first-time drunk drivers.

## **10.4 Harmful events**

According to WHO (2018), Singapore's fatal pedestrian ratio per 100k people was higher than countries with similar incomes such as the UK, Australia, and New Zealand. However, the ratio improved in 2021 to 0.4 (Singapore Police, 2021), but there has been a surge in traffic accidents involving elderly pedestrians, with 22 out of 62 pedestrian fatalities in 2021 being aged 65 and above. The Road Safety for Seniors 2021 campaign was launched to promote safe road behaviour among elderly pedestrians and drivers. The campaign includes initiatives such as roadshows, talks, and educational materials to raise awareness of the risks faced by seniors on the roads and provide practical tips for staying safe. Infrastructure and technology upgrades, such as implementing smart traffic systems and upgrading pedestrian crossings, are also in the works to enhance road safety for seniors (Singapore Police Force, 2021).

Following the spirit of the campaign, we recommend implementing enhanced measures, including the use of blind spot radar and traffic sensing devices.

* ***Blind spot radar:*** Firstly, blind spot radar uses sensors, such as radar or ultrasonic sensors, to detect other vehicles or objects located in the blind spot areas on either side of the vehicle. If an object is detected within the driver's blind spot area, the system alerts the driver through a visual or audible signal, such as a light on the side mirror or an audible warning. LTA can encourage the installation of blind spot radar on vehicles through subsidies or mandatory enforcement to help drivers avoid collisions when changing lanes or turning. Additionally, blind spot radar systems may use other technologies, such as cameras or lidar sensors, to provide a more comprehensive view of the car's surroundings, further enhancing road safety.
* ***Traffic sensing device:*** Secondly, the traffic sensing device which use sensors to detect the presence and movement of vehicles and other road users. Once the sensors detect the presence and movement of vehicles the information is processed and transmitted to the user through visual or audible signals. For example, a light on the device may flash or an audible warning may sound to indicate the presence of a vehicle behind. These sensors can be integrated into devices such as traffic lights or street lamps or installed on road infrastructure, providing valuable data on traffic patterns and volume for better infrastructure planning and design.

These initiates also benefit other vulnerable road user groups, including pedal cyclists and motorcyclists. In fact, motorcyclists accounted for 46.7% of total fatality in 2021, whereas cars and vans users only accounted for 14% (Singapore Police Force, 2021). Thus, all road users would benefit from these safety measures.

## **10.5 Curated Prediction Tools**

Our team has developed a prediction tool using a random forest model built on the top 6 most important features, to assist LTA in predicting the likelihood of fatalities for occupants or drivers involved in accidents. The model has the potential to identify high-risk drivers in real-time and simulate contexts, enabling LTA and hospitals to allocate emergency and crash care resources more efficiently. By using only six variables in the prediction model, the tool can save resources and produce timely results while achieving the same level of accuracy as using more than 30 variables. Additionally, the tool can be simulated for any scenario by changing the inputs of one or many factors. This can help develop targeted interventions for high-risk drivers and vulnerable individuals, such as driver education programs or stricter enforcement of traffic rules. A mockup of the prediction tool can be seen at [*Figure K-1*](#arv7vrgzo34m).

# **11. Discussion**

## **11.1 Limitations**

1. **Lack of Singapore-specific traffic data**

As our machine learning models were trained on US traffic data, they may not accurately reflect Singapore's unique traffic characteristics, such as the influence of weather patterns (such as monsoon seasons) and road infrastructure. This could limit the accuracy and reliability of our models in predicting accident patterns and making effective recommendations for the LTA. To address this limitation, we suggest that the LTA train the model on Singapore-specific data, which would provide a more accurate reflection of Singapore's traffic context and improve the model's effectiveness in this setting. Additionally, incorporating more data sources and improving the model's parameters could also enhance the accuracy of the predictions and recommendations.

1. **Missing Data**

Our analysis was affected by a significant amount of missing data in the US traffic dataset, with some entries reported as unknown or not reported. While imputation techniques could have been used to address this issue, we chose not to do so in order to avoid introducing class imbalance in the dataset. As a result, our analysis may not be fully representative of the actual traffic patterns and risk factors in Singapore. However, we acknowledge that imputation could be a valuable technique for handling missing data in future studies. Additionally, efforts should be made in the future to reduce or eliminate missing data in order to improve overall data quality and promote traffic safety.

1. **Feature Selection & Hyperparameter Tuning**

Lack of feature selection techniques and hyperparameter tuning may have limited our models' performance potential. Feature selection identifies crucial traits, preventing unnecessary features that degrade model accuracy (Guyon & Elisseeff, 2003) while hyperparameter tuning optimises parameter settings to enhance model performance. These techniques are vital for maximising machine learning models' potential. Therefore, to fully utilise the potential of machine learning models, it would be essential for LTA to employ feature selection techniques and hyperparameter tuning.

1. **Blackbox nature of model**

The Random Forest model is a "blackbox" that does not provide an explanation of how it made its predictions or decisions. While these models can perform well, they also have drawbacks, such as a lack of transparency that can make it challenging to identify errors or biases. Additionally, because the model is not interpretable, it may be difficult to understand how it operates and which features it considers in making predictions. Finally, due to their complexity, blackbox models can be difficult to fine-tune for specific business needs.

## **11.2 Future Work**

1. **Model Deployment and Retraining**

To ensure continued accuracy of the predictive models, it is important to approach model deployment as an iterative process, as road traffic data is dynamic and can change over time (Kavikondala et al., 2019). Retraining of predictive models should be done if new input data significantly differs from the initial training set. However, manual retraining can be time-consuming due to the numerous data points involved (Patruno, 2019). Therefore, automated retraining using **automated MLOps pipelines** can be a more effective solution. MLOps pipelines can collect, evaluate, and select data for training, as well as automate the process of creating and training models (Symeonidis et al., 2022).

1. **Location Analytics to Allocate Resources**

Our model's training on US data limits our ability to leverage location analytics to identify and target high-risk areas within Singapore. By analysing historical accident patterns and identifying areas with a high frequency and severity of accidents, location analytics can assist in allocating resources and prioritising infrastructure improvements. This can improve emergency response times and reduce the likelihood of future accidents. In the future, incorporating Singapore-specific data can help in pinpointing areas of concern and assist LTA in making data-driven decisions to promote traffic safety.

1. **Deployment of interactive dashboard and model for LTA to use**

An interactive dashboard, user-friendly UI, and traffic accident prediction model can enhance the LTA’s operations. The predictive model can forecast the likelihood and severity of accidents based on real-time traffic data and other relevant factors. This can be incorporated into an interactive dashboard, allowing LTA staff to monitor the situation in real-time, identify potential hotspots, and allocate resources effectively. This enables them to interact with the data intuitively, perform what-if analysis, and assess the effectiveness of various accident-reduction measures. This combination of data-driven insights and user-friendly tools can help the LTA mitigate the impact of traffic accidents, and improve road safety.

# **12. Conclusion**

LTA’s extensive data collection through ITS poses a large potential for machine learning and analytics to minimise traffic accidents and fatalities. By tapping on Singapore's vast traffic data, our proposed analytics solution can provide actionable insights to LTA and the Singapore government, such as significant predictors of fatal accidents, enabling them to take preemptive and corrective actions to reduce the likelihood and severity of traffic accidents.

To ensure long-term viability, our proposed models must be regularly retrained with updated traffic and road conditions data to tackle model drift. While this report has highlighted the role of machine learning in promoting road safety, there are likely to be further opportunities for innovative solutions as technology continues to evolve. Thus, LTA must remain vigilant in identifying and implementing new measures to enhance road safety. By adopting a proactive and forward-thinking approach, we can continue to leverage the power of technology and make significant strides in improving the safety of our roads for all users.

# **13. References**

Agarwal, A. (2022, December 16). How to find VIF on a data in R . ProjectPro. Retrieved from <https://www.projectpro.io/recipes/find-vif-on-data-r#:~:text=Variance%20inflation%20factor%20(VIF)%20is,variables%20in%20a%20regression%20model>.

Alrejjal, A., Moomen, M., &amp; Ksaibati, K. (2022, May 14). Evaluating the impact of traffic violations on Crash injury severity on wyoming interstates: An investigation with a random parameters model with heterogeneity in means approach. Journal of Traffic and Transportation Engineering (English Edition). Retrieved February 13, 2023, from https://www.sciencedirect.com/science/article/pii/S2095756422000393

Anderson, M., &amp; Auffhammer, M. (2011, June). POUNDS THAT KILL: THE EXTERNAL COSTS OF VEHICLE WEIGHT. NBER WORKING PAPER SERIES. Retrieved February 12, 2023, from https://www.nber.org/system/files/working\_papers/w17170/w17170.pdf

Becker, N., Rust, H. W., &amp; Ulbrich, U. (2022, August 17). Weather impacts on various types of road crashes: A quantitative analysis using generalized additive models - European Transport Research Review. SpringerOpen. Retrieved February 13, 2023, from https://etrr.springeropen.com/articles/10.1186/s12544-022-00561-2

Billiau, S. (2021, June 13). From Scratch: Permutation Feature Importance for ML Interpretability. Towards Data Science. Retrieved April 3, 2022, from <https://towardsdatascience.com/from-scratch-permutation-feature-importance-for-ml-interpretability-b60f7d5d1fe9>

Bonthu, H. (2021, July 11). An introduction to logistic regression. Analytics Vidhya. Retrieved from <https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/>

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

Budget Direct Insurance. (2023, February 5). Road Accident Statistics in Singapore 2023. Budgetdirect.com.sg; Budget Direct Insurance. <https://www.budgetdirect.com.sg/car-insurance/research/road-accident-statistics-in-singapore>

Centers for Disease Control and Prevention. (n.d.). Policy impact: Seat belts. Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/transportationsafety/seatbeltbrief/index.html#:~:text=Among%20drivers%20and%20front%2Dseat,of%20serious%20injury%20by%2050%25.&amp;text=Seat%20belts%20prevent%20drivers%20and%20passengers%20from%20being%20ejected%20during%20a%20crash.

Centre for Liveable Cities. (2018). Tapping Tech for Smoother Traffic. Centre for Liveable Cities. Retrieved March 21, 2023, from <https://www.clc.gov.sg/docs/default-source/urban-solutions/urb-sol-iss-13-pdfs/11_case_study-singapore-intelligent-transport-systems.pdf>

Channel NewsAsia. (2023, February 14). More traffic accidents involving elderly pedestrians in 2022: SPF. CNA. Retrieved February 18, 2023, from https://www.channelnewsasia.com/singapore/more-traffic-accidents-involving-elderly-pedestrians-2022-fatalities-spf-police-traffic-police-3276866

Chin, H. C., Haque, M. M., & Jean, Y. H. (2006). *An estimate of road accident costs in Singapore* (M. Hoque, Ed.). Www.researchgate.net; Research Gate. <https://www.researchgate.net/publication/268152486_An_estimate_of_road_accident_costs_in_Singapore>

CNA. (2022, January 6). More traffic accidents involving elderly pedestrians in 2022; 2 fatalities in 5 days: SPF. Channel NewsAsia.<https://www.channelnewsasia.com/singapore/more-traffic-accidents-involving-elderly-pedestrians-2022-fatalities-spf-police-traffic-police-3276866>

Direct Asia. (2022, August 30). What are the seat belt rules in Singapore?: Directasia Insurance. Direct Asia. Retrieved April 1, 2023, from <https://www.directasia.com/blog/seat-belt-rules-in-singapore>

Dobilas, S. (2020, November 29). Mars: Multivariate adaptive regression splines - how to improve on linear regression? Medium. Retrieved April 1, 2023, from <https://towardsdatascience.com/mars-multivariate-adaptive-regression-splines-how-to-improve-on-linear-regression-e1e7a63c5eae>

Flood Law. (2021, June 21). The risks of aggressive driving. The Flood Law Firm. Retrieved April 1, 2023, from <https://www.thefloodlawfirm.com/blog/why-aggressive-driving-causes-crashes/>

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 19(1), 1–67. <http://www.jstor.org/stable/2241837>

García-Herrero, S., Febres, J. D., Boulagouas, W., Gutiérrez, J. M., &amp; Mariscal Saldaña, M. Á. (2021, July 4). Assessment of the influence of technology-based distracted driving on Drivers' infractions and their subsequent impact on traffic accidents severity. International journal of environmental research and public health. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8297255/>

George, Y., Athanasios, T., &amp; George, P. (2017, June 8). Investigation of road accident severity per vehicle type. Transportation Research Procedia. Retrieved February 13, 2023, from <https://www.sciencedirect.com/science/article/pii/S2352146517307081>

Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. The MIT Press.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection - journal of machine ... Journal of Machine Learning Research. Retrieved March 30, 2023, from <https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>

Hayes, A. (2023, February 25). Multicollinearity: Meaning, Examples, and FAQs. Investopedia. Retrieved from <https://www.investopedia.com/terms/m/multicollinearity.asp#:~:text=Multicollinearity%20is%20a%20statistical%20concept,in%20less%20reliable%20statistical%20inferences>.

Insurance Institute for Highway Safety. (n.d.). Vehicle size and weight. IIHS. Retrieved from <https://www.iihs.org/topics/vehicle-size-and-weight>

Javatpoint. (n.d.). Backward Elimination in Machine Learning . www.javatpoint.com. Retrieved March from <https://www.javatpoint.com/backward-elimination-in-machine-learnin>

John. (2020, May 18). The 5 most common causes for road accidents in Singapore: Articles: Motorist singapore. Motorist.sg. Retrieved February 13, 2023, from https://www.motorist.sg/article/173/the-5-most-common-causes-for-road-accidents-in-singapore

Lasota, D., Al-Wathinani, A., Krajewski, P., Goniewicz, K., &amp; Pawłowski, W. (2020, December 2). Alcohol and road accidents involving pedestrians as unprotected road users. International journal of environmental research and public health. Retrieved February 18, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7729452/

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3). <https://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf>

Lim, J. (2023, February 14). More elderly pedestrians injured, killed in traffic accidents in 2022: Traffic police. The Straits Times. Retrieved April 1, 2023, from <https://www.straitstimes.com/singapore/courts-crime/more-elderly-pedestrians-died-and-got-hurt-in-traffic-accidents-in-2022-traffic-police#:~:text=Overall%2C%20the%20number%20of%20people,before%20the%20Covid%2D19%20pandemic>.

Lin, H.-A., Chan, C.-W., Wiratama, B. S., Chen, P.-L., Wang, M.-H., Chao, C.-J., Saleh, W., Huang, H.-C., &amp; Pai, C.-W. (2022, November 10). Evaluating the effect of drunk driving on fatal injuries among vulnerable road users in Taiwan: A population-based study - BMC Public Health. BioMed Central. Retrieved from <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-022-14402-3>

LTA. (2023, March 2). Intelligent Transport Systems. LTA. Retrieved March 21, 2023, from https://www.lta.gov.sg/content/ltagov/en/getting\_around/driving\_in\_singapore/intelligent\_transport\_systems.html

Malin, F., Norros, I., &amp; Innamaa, S. (2018, October 29). Accident risk of road and weather conditions on different road types. Accident Analysis &amp; Prevention. Retrieved February 13, 2023, from https://www.sciencedirect.com/science/article/pii/S0001457518308455

Mayrose, J., &amp; Priya, A. (2008, August 5). The safest seat: Effect of seating position on occupant mortality. Journal of safety research. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/18786431/>

Mekonnen, T. H., Tesfaye, Y. A., Moges, H. G., &amp; Gebremedin, R. B. (2019, March 9). Factors associated with risky driving behaviors for road traffic crashes among professional car drivers in Bahirdar City, Northwest Ethiopia, 2016: A cross-sectional study - environmental health and Preventive Medicine. BioMed Central. Retrieved from <https://environhealthprevmed.biomedcentral.com/articles/10.1186/s12199-019-0772-1>

Mujibah, F. (2023, February 5). Elderly woman in hit-and-run accident in Pasir Ris dies from injuries. The Straits Times. Retrieved April 1, 2023, from <https://www.straitstimes.com/singapore/elderly-woman-in-hit-and-run-accident-in-pasir-ris-dies-from-injuries>

National Center for Statistics and Analysis. (2022, July). Rural/urban comparison of traffic fatalities: 2020 data (Traffic Safety Facts. Report No. DOT HS 813 336). Traffic Safety Facts. Retrieved February 12, 2023, from <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813336>

Nester, J. (2015, April 28). The more you know: How countries are combating road fatalities. Institute for Transportation. Retrieved April 1, 2023, from <https://intrans.iastate.edu/news/the-more-you-know-how-countries-are-combating-road-fatalities/#:~:text=It%20is%20pretty%20clear%20that,for%20rear%2Dseat%20car%20occupants>

Pembuain, A., Priyanto, S., &amp; Suparma, L. (2019, October 1). The effect of road infrastructure on traffic accidents. The Effect of Road Infrastructure on Traffic Accidents | Atlantis Press. Retrieved February 13, 2023, from https://www.atlantis-press.com/proceedings/apte-18/125918743

Regev, S., Rolison, J. J., &amp; Moutari, S. (2018, July 7). Crash risk by driver age, gender, and time of day using a new exposure methodology. Journal of Safety Research. Retrieved February 13, 2023, from <https://www.sciencedirect.com/science/article/pii/S0022437517307600>

Rudin, C. (2019, May 13). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature News. Retrieved March 30, 2023, from <https://www.nature.com/articles/s42256-019-0048-x>

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. <https://doi.org/10.1038/323533a0>

Shiying, W. (2022, February 14). Fatal road accidents in S’pore rose 25% in 2021 as more activities resumed | The Straits Times. Www.straitstimes.com. <https://www.straitstimes.com/singapore/courts-crime/more-people-died-or-got-hurt-on-spore-roads-in-2021-as-more-activities-resumed>

Singapore General Hospital. (2017, June 2). Led study reinforces importance of wearing seatbelts. SGH. Retrieved April 1, 2023, from <https://www.sgh.com.sg/news/others/sgh-led-study-reinforces-importance-wearing-seatbelts#:~:text=The%20findings%20also%20showed%20that,number%20of%20seat%2Dbelt%20violations>

Singapore Legal Advice. (2022, March 29). Hit-and-run victims: Next steps and claiming compensation. SingaporeLegalAdvice.com. Retrieved April 1, 2023, from <https://singaporelegaladvice.com/law-articles/what-if-you-get-hit-by-a-drink-driving-diplomat/>

Singapore Police Force. (2021). *ANNUAL TRAFFIC STATISTICS 2021*. SPF. <https://www.police.gov.sg/-/media/1F7F9460FD8F48928B6DEFE096414975.ashx>

Singapore Police Force. (2021, July 16). Launch of Road Safety for Seniors 2021 Campaign. Retrieved from<https://www.police.gov.sg/Media-Room/News/20210716_launch_of_road_safety_for_seniors_2021_campaign>

Singapore Police Force. (n.d.). Mandatory simulator training for Learner Motorists. Singapore Police Force. Retrieved from [https://www.police.gov.sg/Media-Room/News/20191207\_OTHERS\_Mandatory\_Simulator\_Training\_for\_Learner\_Motorist](https://www.police.gov.sg/Media-Room/News/20191207_OTHERS_Mandatory_Simulator_Training_for_Learner_Motorists)

Statistics Solutions. (2021, June 23). Binary Logistic Regression. Statistics Solutions. Retrieved from <https://www.statisticssolutions.com/binary-logistic-regression/>

Symeonidis, G., Nerantzis, E., & Papakostas, G. (2022, January). (PDF) MLOps -- definitions, tools and challenges - researchgate. ResearchGate. Retrieved March 26, 2023, from     [https://www.researchgate.net/publication/357552787\_MLOps\_--\_Definitions\_Tools\_and\_Challanges](https://www.researchgate.net/publication/357552787_MLOps_--_Definitions_Tools_and_Challenges)

World Health Organization. (2022, June 20). Road traffic injuries. World Health Organization. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>

World Health Organization. (2018). Deaths in the Road. Retrieved from <https://extranet.who.int/roadsafety/death-on-the-roads/#country_or_area/SGP>

Xingcan, L. (2020, November). A brief review of the impact of distracted driving on Traffic Safety. IOP Conference Series: Earth and Environmental Science. Retrieved February 12, 2023, from <https://www.researchgate.net/publication/346745342_A_brief_review_of_the_impact_of_distracted_driving_on_traffic_safety>

Yannis, G., Nikolaou, D., Laiou, A., Sturmer, Y. A., Buttler, I., & Karpa, D. J. (2020). Vulnerable road users: Cross-cultural perspectives on performance and attitudes. <https://www.sciencedirect.com/science/article/pii/S0386111220300716#:~:text=Pedestrians%2C%20pedal%20cyclists%20and%20motorcyclists,impact%20of%20a%20road%20crash>

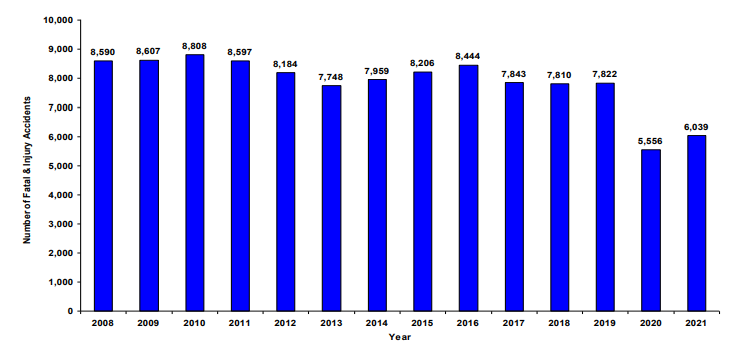
Zhang, W., & Goh, A. T. C. (2016). Evaluating seismic liquefaction potential using multivariate adaptive regression splines and logistic regression. Geomechanics and Engineering, 10(3), 269–284. <https://doi.org/10.12989/gae.2016.10.3.269>

Zhao, X., Zhang, X., &amp; Rong, J. (2014, February 23). Study of the effects of alcohol on drivers and driving performance on Straight Road. Mathematical Problems in Engineering. Retrieved February 13, 2023, from <https://www.hindawi.com/journals/mpe/2014/607652/>

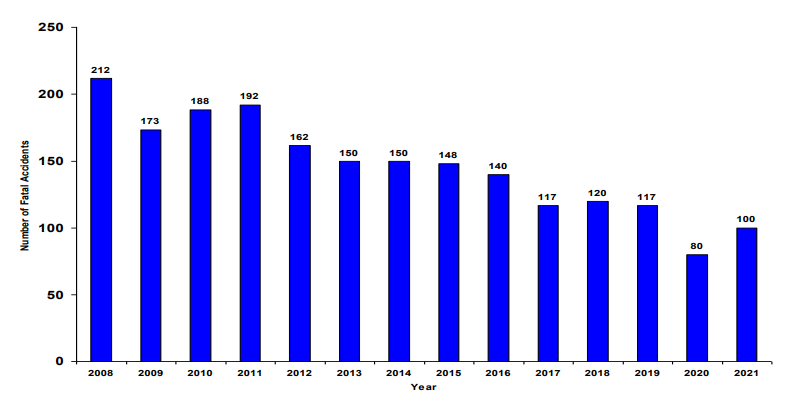
Zvornicanin, E. (2022, November 4). What is feature importance in machine learning? Baeldung on Computer Science. Retrieved from <https://www.baeldung.com/cs/ml-feature-importance#:~:text=3.-,3.,a%20current%20model%20and%20prediction>.

# **14. Appendix**

## **Appendix A: Background**

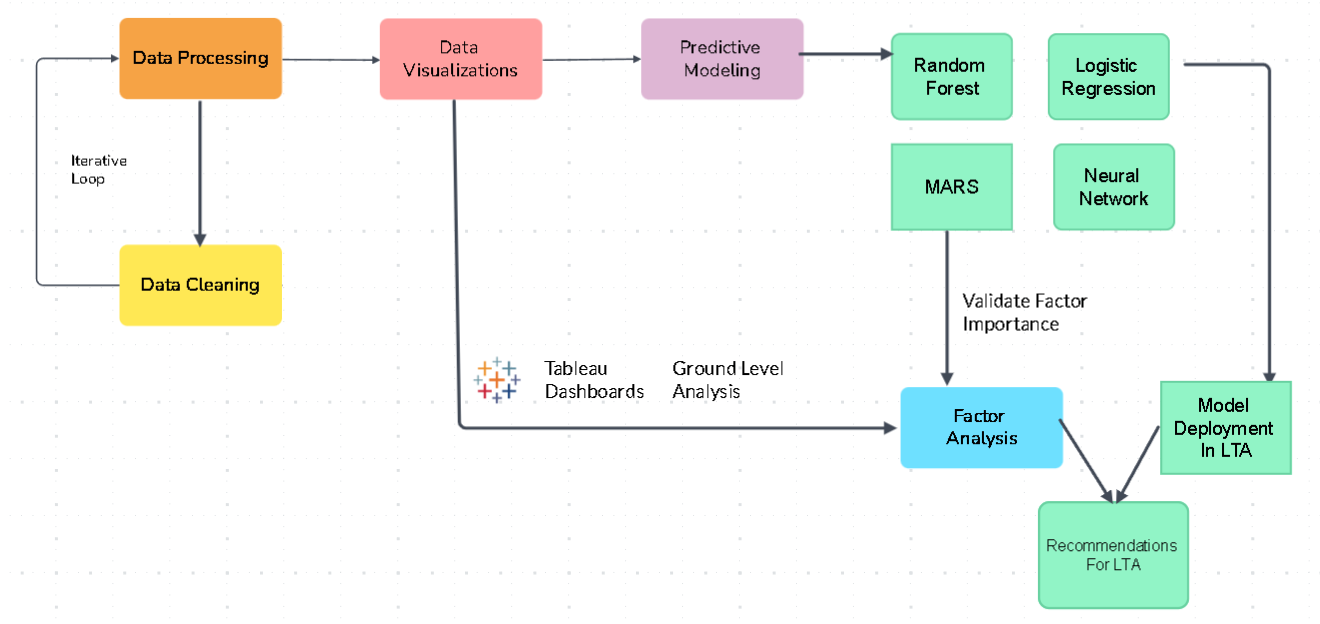


*Figure A-1. Histogram depicting the number of fatal & injury accidents from 2008 to 2021*



*Figure A-2. Histogram depicting the number of fatal accidents from 2008 to 2021*

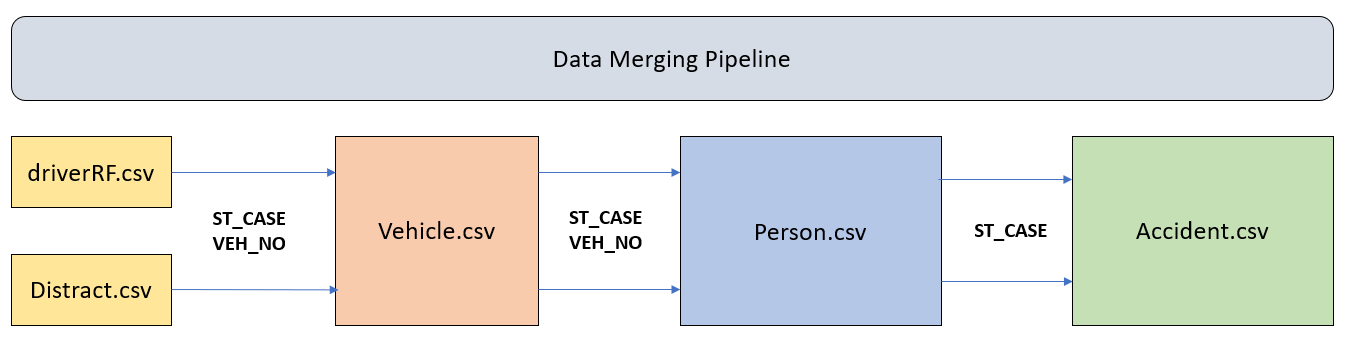
## **Appendix B: Overall Data Pipeline**



*Figure B-1. High level overview of overall data pipeline*

## 

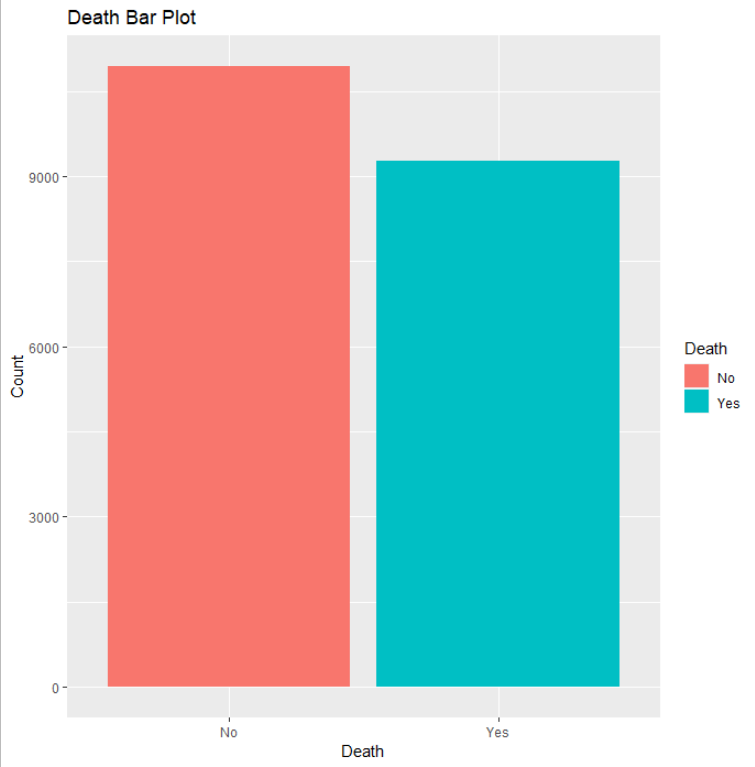
## **Appendix C: Data Preparation**



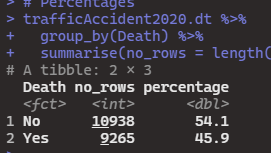
*Figure C-1. Data Merging Pipeline*

## 

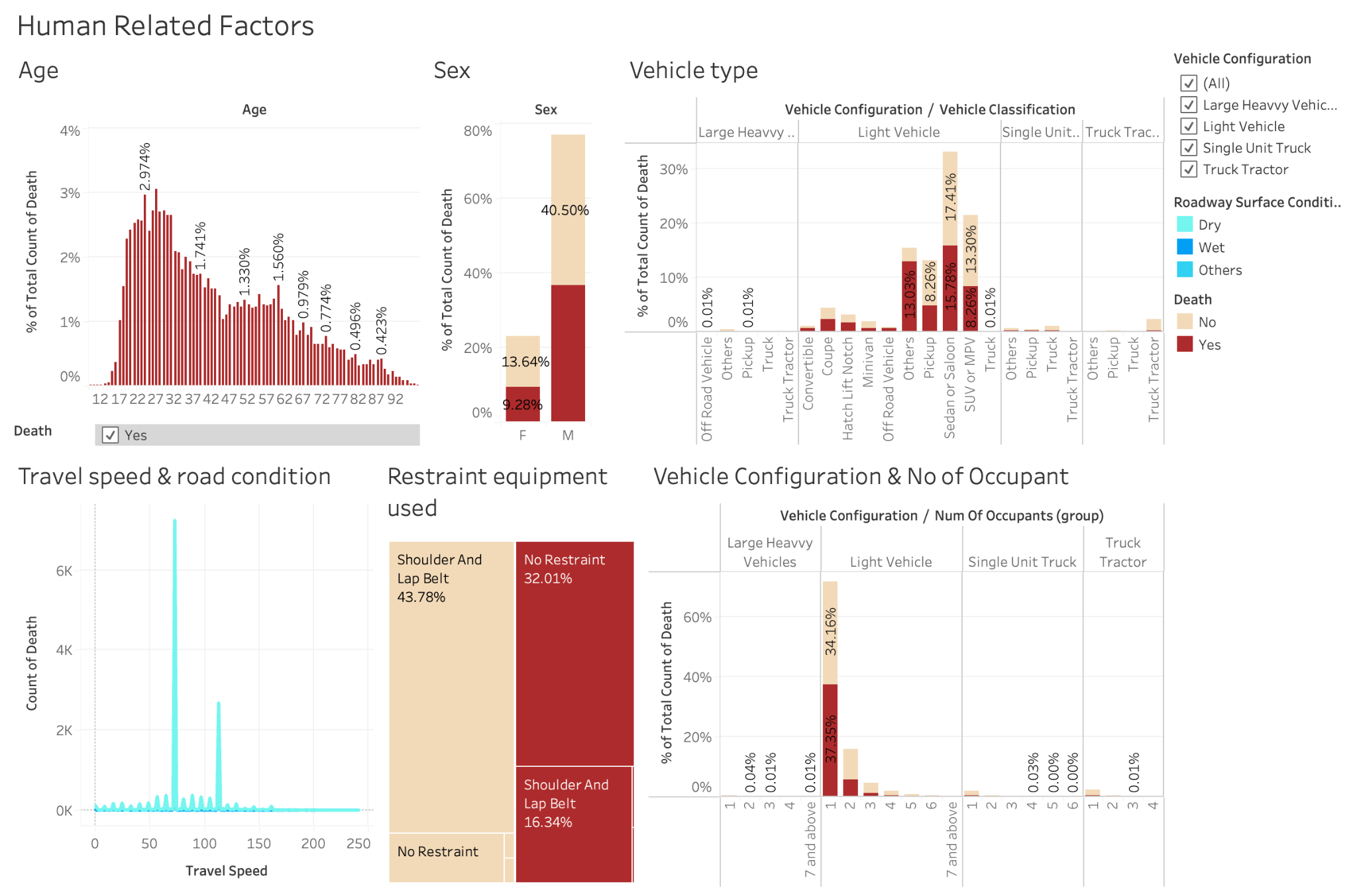
## **Appendix D: Data Exploration & Visualization**

****

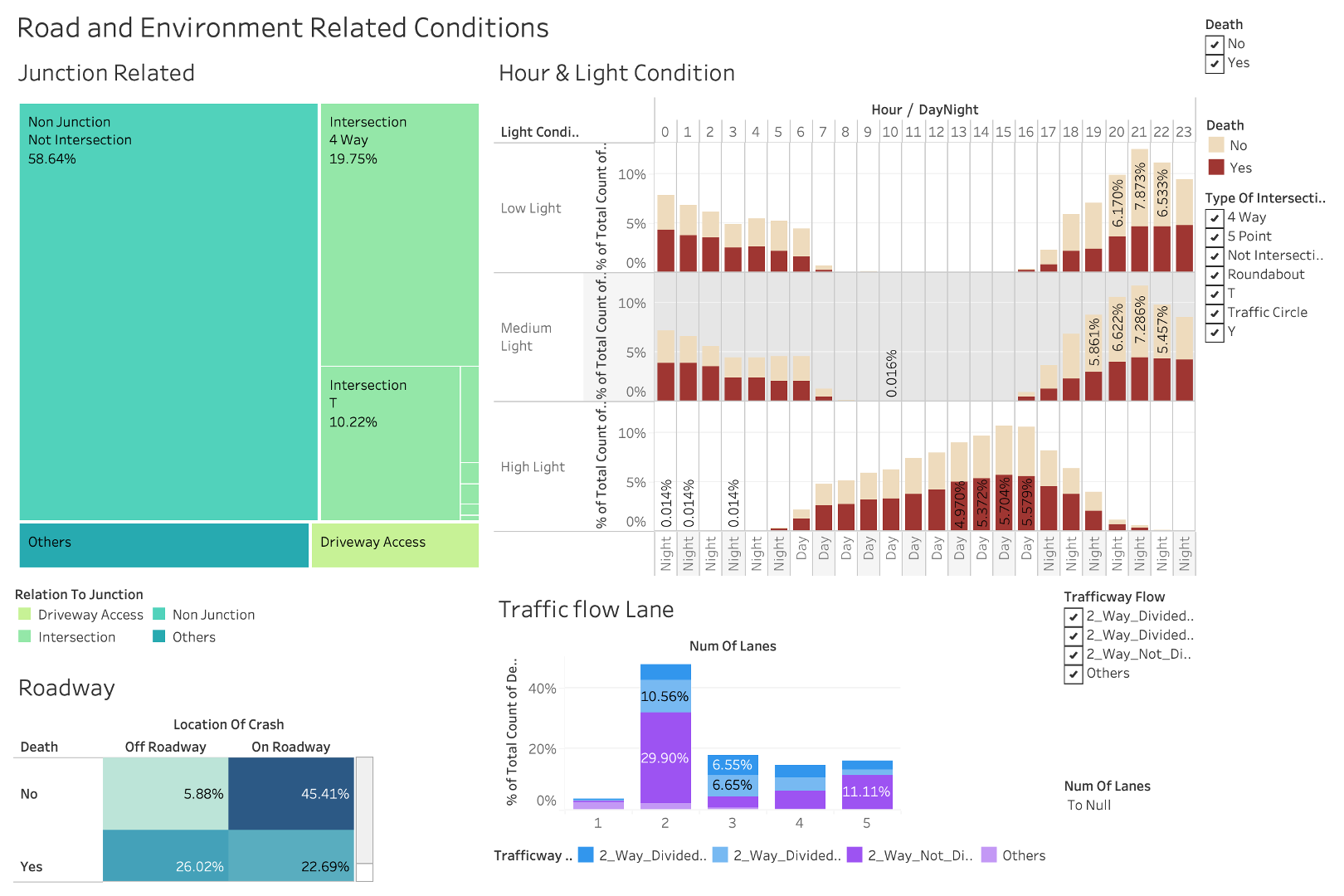
*Figure D-1. Bar Plot showing the proportion of people who died and did not die in the accident within dataset*

**

*Figure D-2. Percentages of people who died and did not die in the accident within dataset*



*Fig D-3. Dashboard for human-related factors*



*Fig D-4. Dashboard for Road and Environmental factors*

## 

*Fig D-5. Dashboard for Driver-behaviour factors*

## 

## 

## 

## 

## 

## 

## 

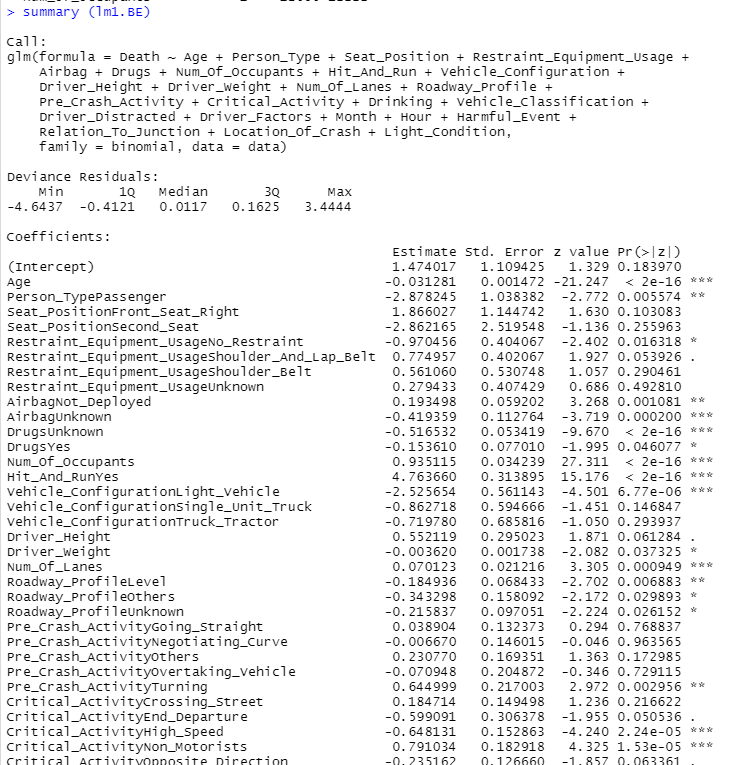
## 

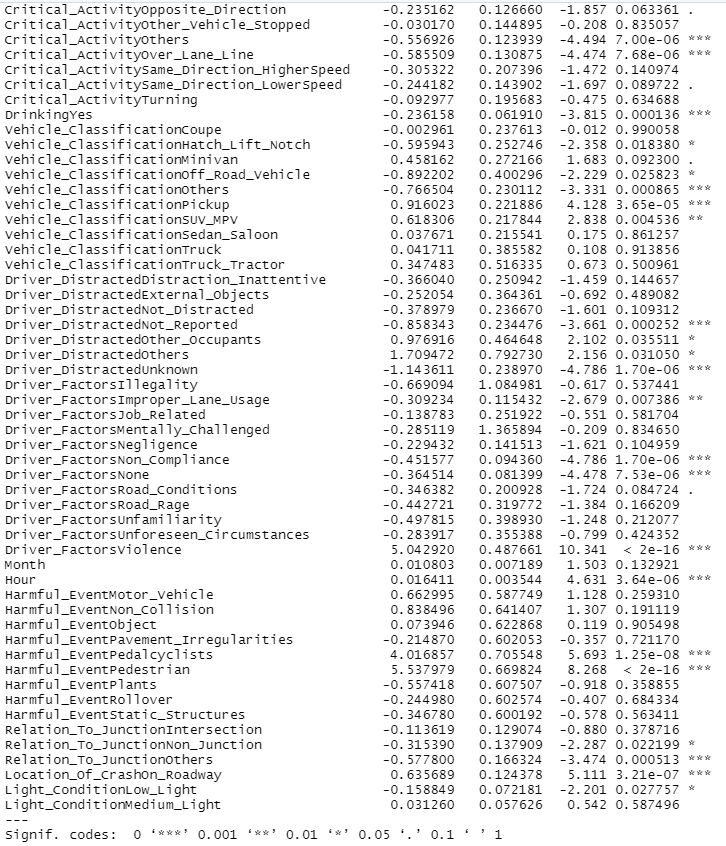
## 

## **Appendix E: Logistic Regression**

**Appendix E-1: Summary Output of lm1.BE**

We performed backward elimination using the step () function on lm1, which is the logistic regression model that considered all x variables.

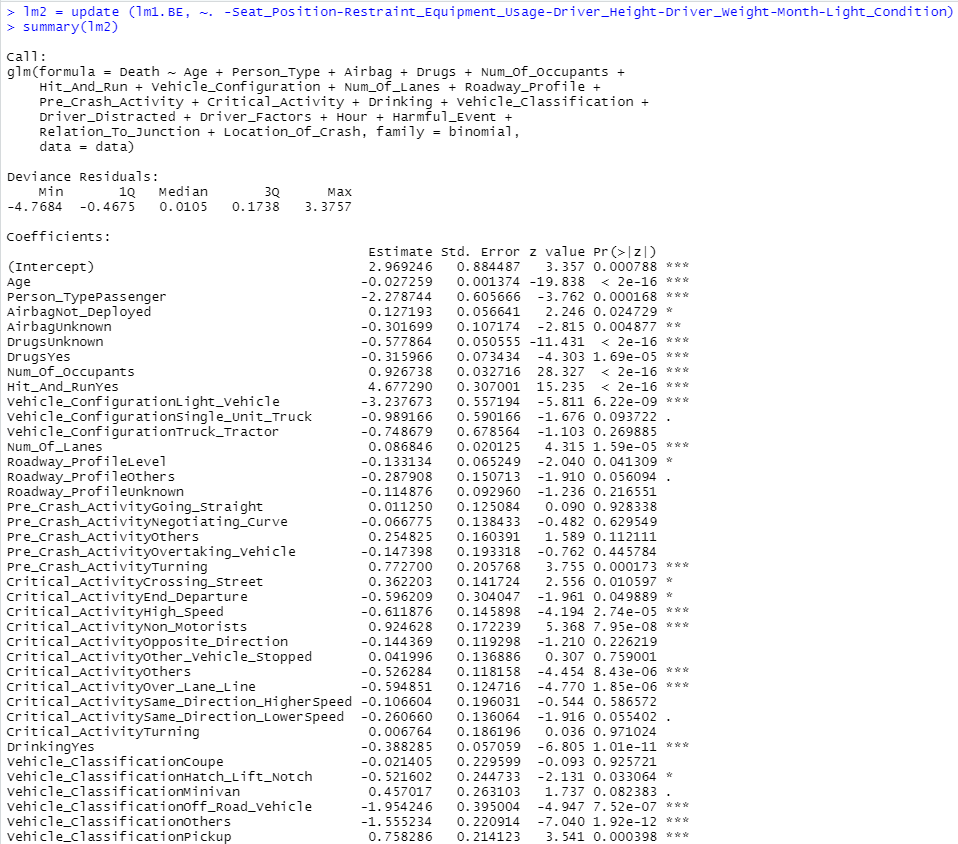
****

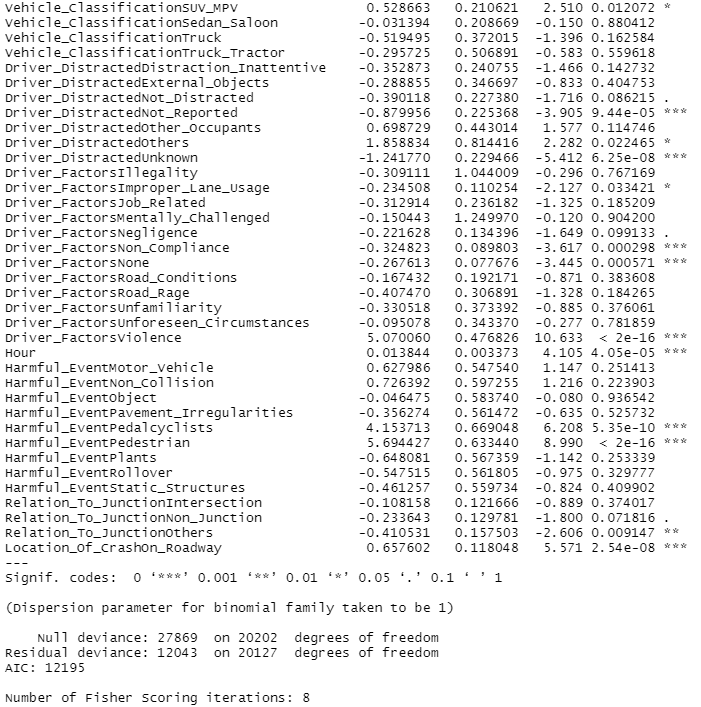
****

*Figure E-1. Summary Output of lm1.BE*

**Appendix E-2: Summary Output of lm2**

Remove insignificant variables from the lm1.BE model. The insignificant variables are Seat\_Position, Restraint\_Equipment\_Usage, Driver\_Height, Driver\_Weight, Month, and Light\_Condition. They all have less than 2 \*s.

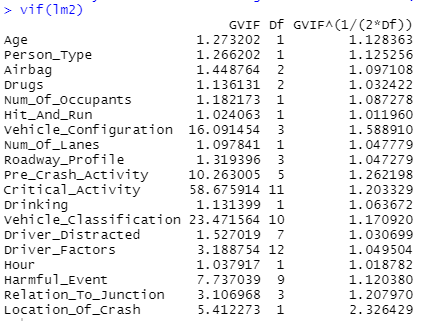




*Figure E-2. Summary Output of lm2*

**Appendix E-3: VIF output of lm2**

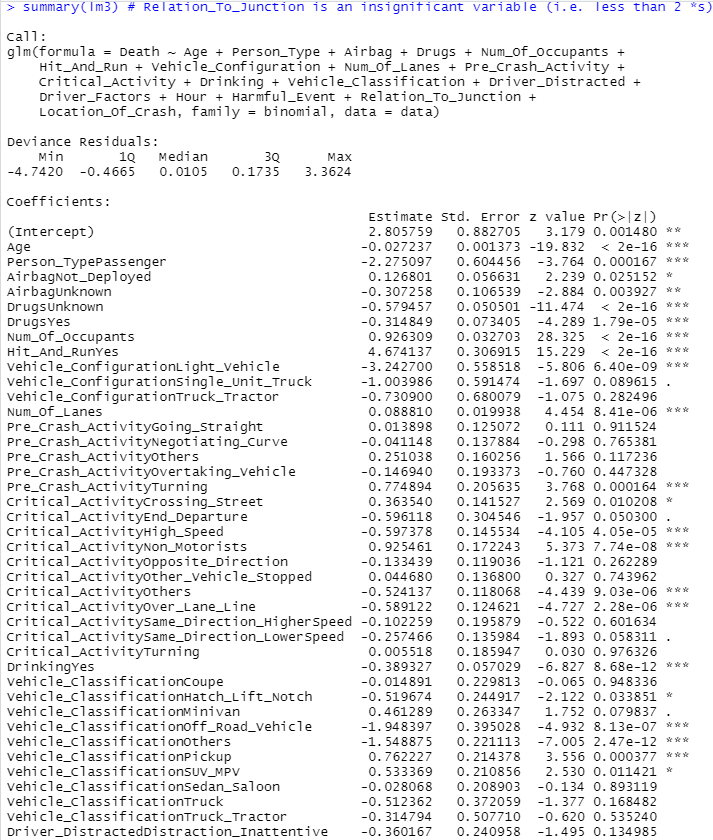
To detect multicollinearity between the independent variables in the logistic regression. A total of 4 variables with a GVIF score of > 10. These variables are Vehicle\_Configuration, Pre\_Crash\_Activity, Critical\_Activity and Vehicle\_Classification.

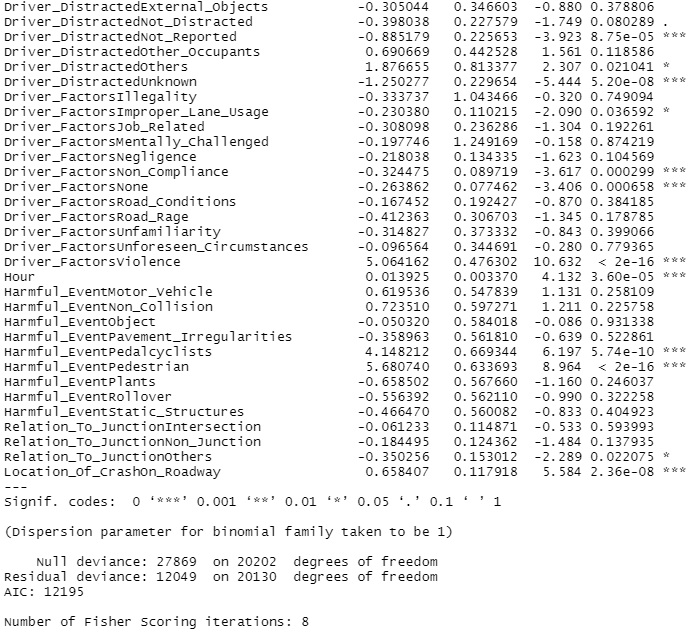


*Figure E-3. VIF output of lm2*

**Appendix E-4: Summary Output of lm3**

Remove insignificant variables from the lm2 model. The insignificant variable is Roadway\_Profile. It has less than 2 \*s.

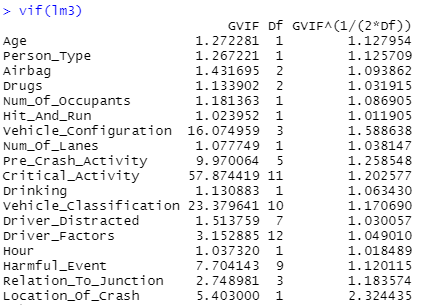




*Figure E-4. Summary Output of lm3*

**Appendix E-5: VIF output of lm3**

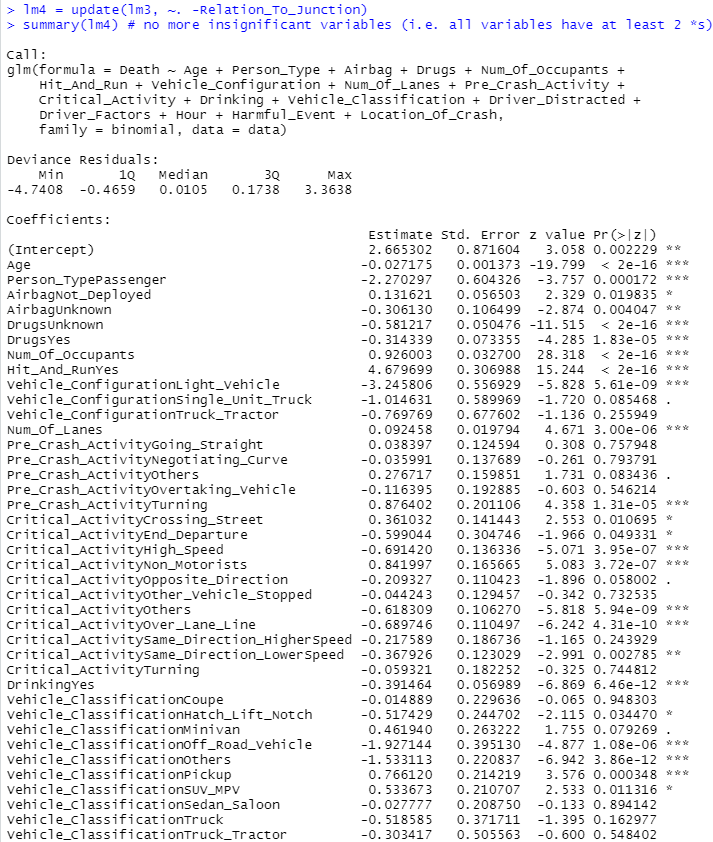
A total of 3 variables with a GVIF score of > 10. These variables are Vehicle\_Configuration, Critical\_Activity and Vehicle\_Classification.

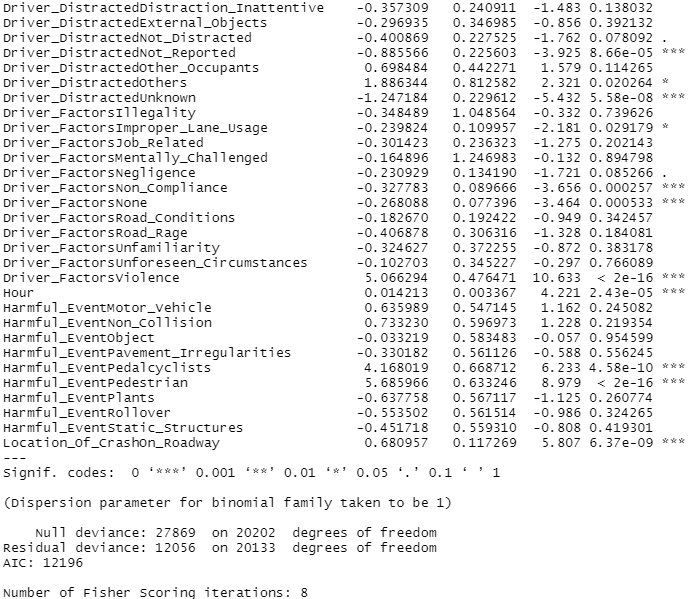


*Figure E-5. VIF Output of lm3*

**Appendix E-6: Summary Output of lm4**

Remove insignificant variables from the lm3 model. The insignificant variable is Relation\_To\_Junction. It has less than 2 \*s. From the summary output of the lm4 model, we can observe that all the independent variables present are significant. However, we are still facing an issue of multicollinearity as depicted in *Appendix E7*.

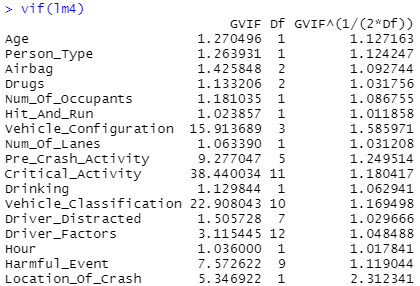




*Figure E-6. Summary Output of lm4*

**Appendix E-7: VIF Output of lm4**

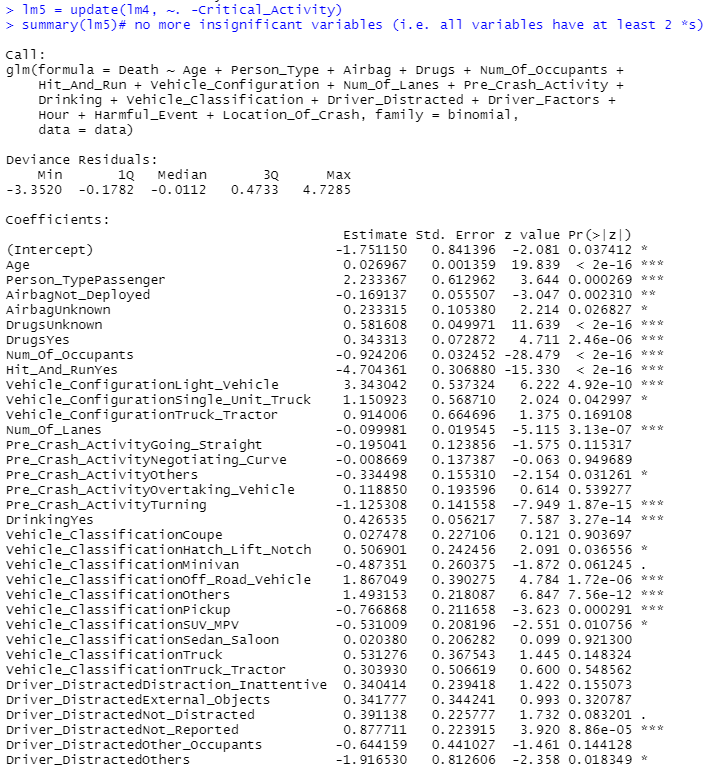
A total of 3 variables with a GVIF score of > 10. These variables are Vehicle\_Configuration, Critical\_Activity, and Vehicle\_Classification. Critical\_Activity has the highest GVIF score followed by Vehicle\_Classification, then Vehicle\_Configuration.

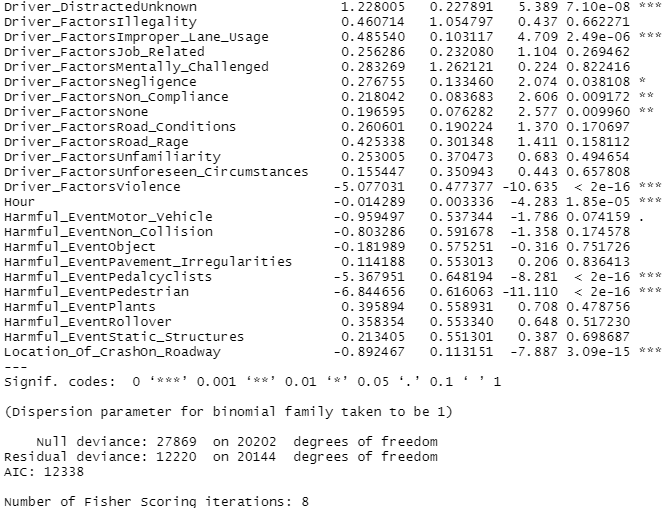


*Figure E-7. VIF Output of lm4*

**Appendix E-8: Summary Output of lm5**

Remove Critical\_Activity from the lm4 model. As stated in *Appendix E7*, Critical\_Activity has the highest GVIF score out of the 3 variables, thus will be removed first. From the summary output of the lm5 model, we can observe that all the independent variables present are significant.

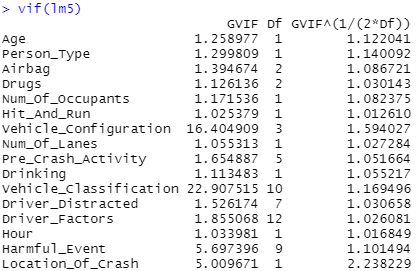




*Figure E-8. Summary Output of lm5*

**Appendix E-9: VIF output of lm5**

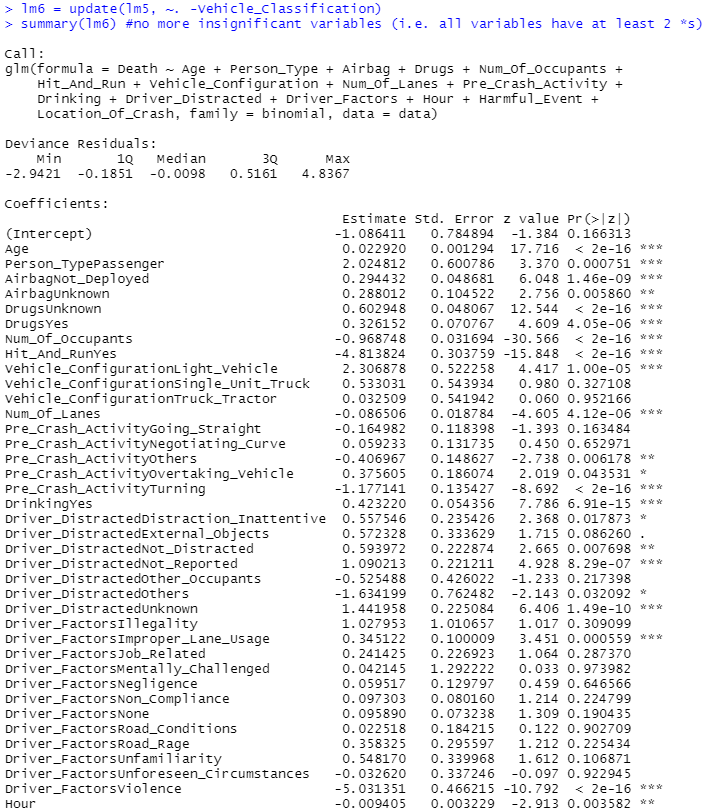
The multicollinearity issue still persists. Both Vehicle\_Configuration and Vehicle\_Classification have a GVIF score of > 10. Vehicle\_Classification has a higher GVIF score, therefore, removed it first.

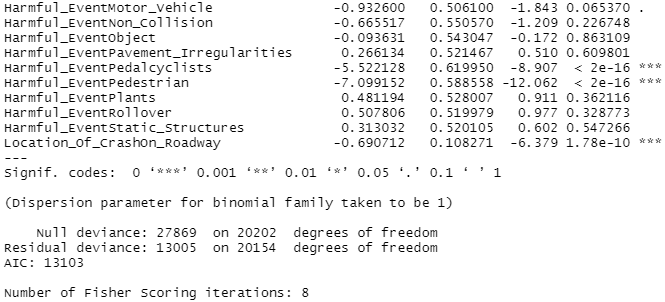


*Figure E-9. VIF Output of lm5*

**Appendix E-10: Summary Output of lm6**

Remove Vehicle\_Classification from the lm5 model because it has a higher GVIF score than Vehicle\_Configuration (as stated in *Appendix E9*). From the summary output of the lm6 model, we can observe that all the 15 independent variables present are significant.

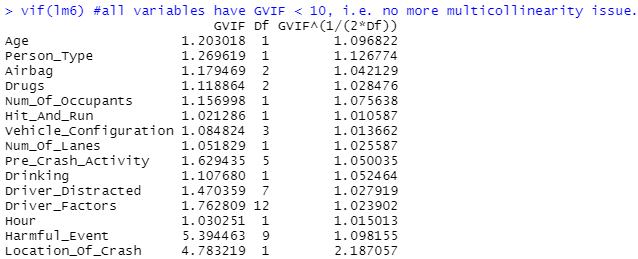




*Figure E-10. Summary Output of lm6*

**Appendix E-11: VIF output of lm6**

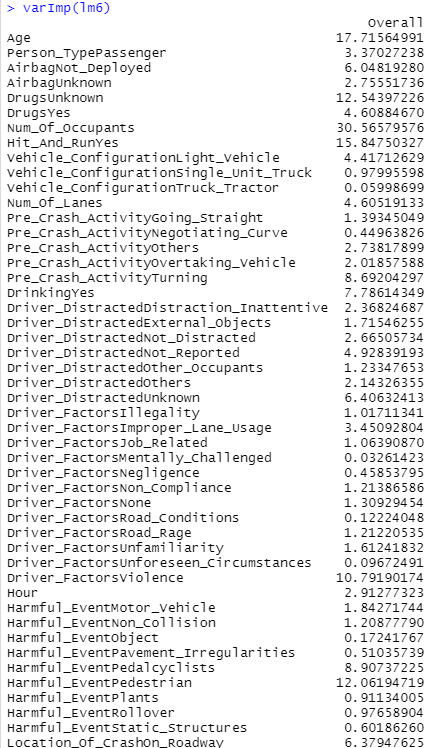
The multicollinearity issue is solved. All the independent variables in the lm6 model have a GVIF score of <= 10. We thus trained the lm6 model and used it to predict the outcome of the traffic accident, i.e., death = “Yes” or “No”.

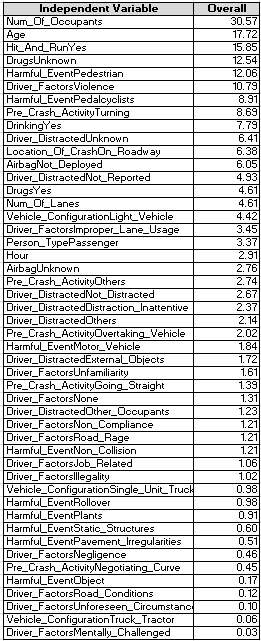


*Figure E-11. VIF Output of lm6*

**Appendix E-12: Variable Importance of Independent Variables in lm6**

The variable importance of each independent variable is ranked (from most important): Num\_Of\_Occupants, Age, Hit\_And\_Run, Drugs, Harmful\_Event, Driver\_Factors, Pre\_Crash\_Activity, Drinking, Driver\_Distracted, Location\_Of\_Crash, Airbag, Num\_of\_Lanes, Vehicle\_Configuration, Person\_Type, and Hour.





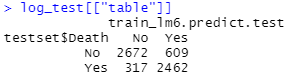
*Figure E-12. Variable Importance of Independent Variables in lm6*



*Figure E-13. Confusion Matrix Results of Predicting the Train Set With Trained lm6 Model*

## 

*Figure E-14. Model Evaluation Metrics for the Train Set*



*Figure E-15. Confusion Matrix Results of Predicting the Test Set With Trained lm6 Model*

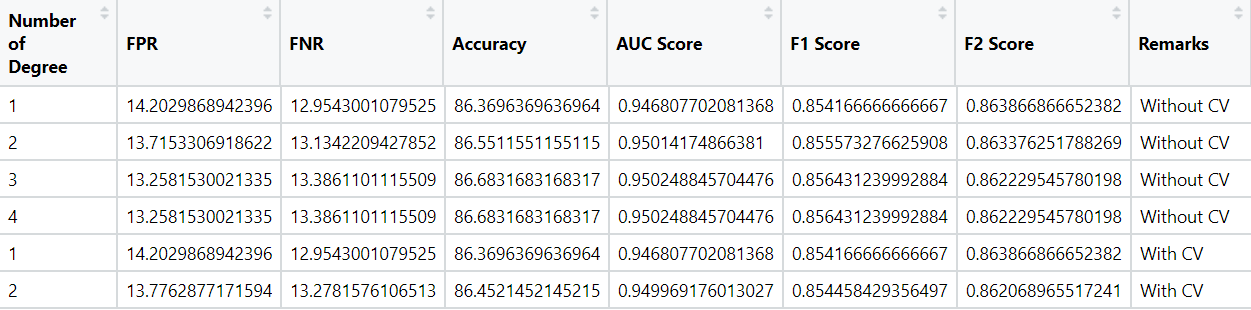
****

*Figure E-16. Model Evaluation Metrics for the Test Set*

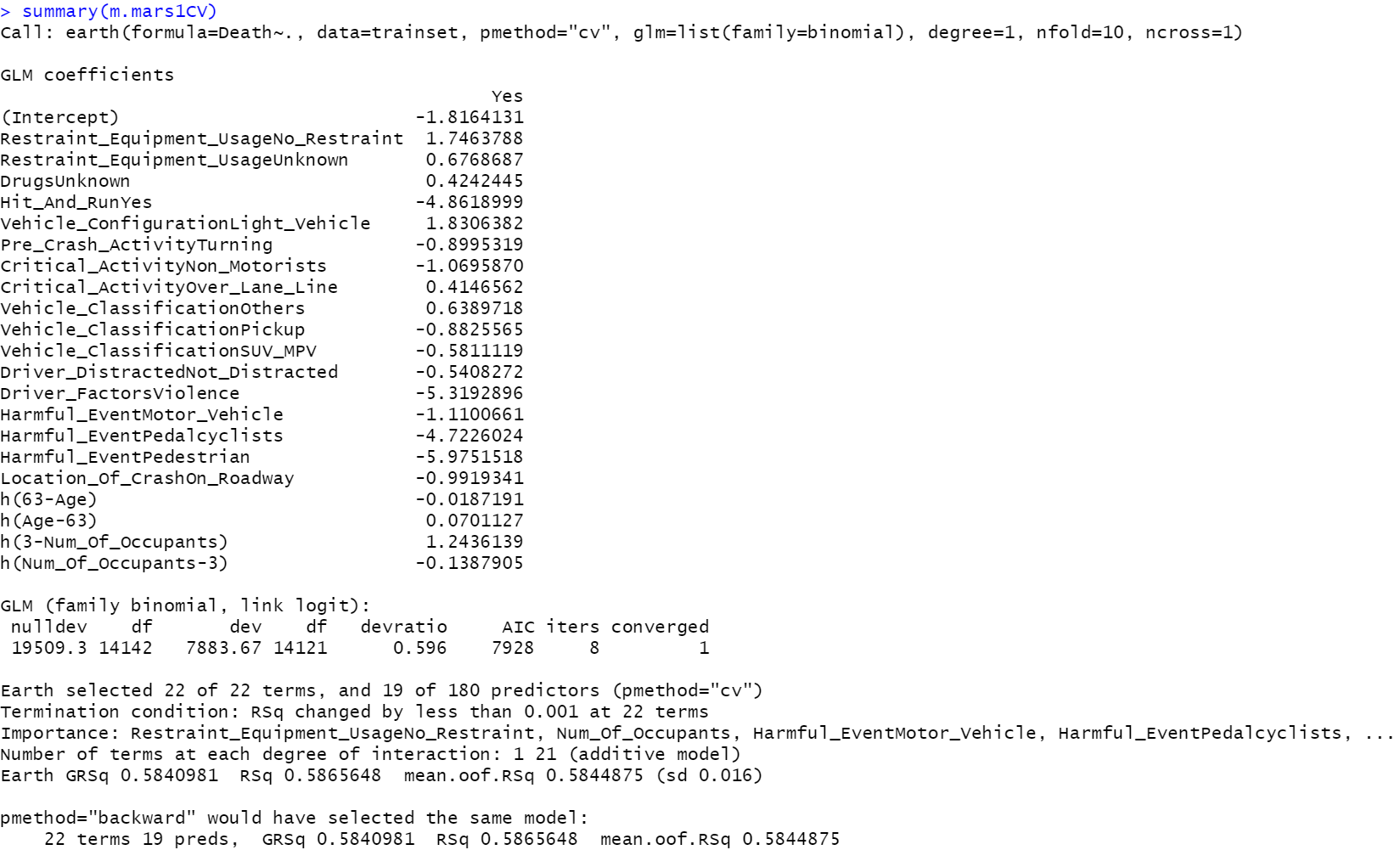
## **Appendix F: MARS**

## 

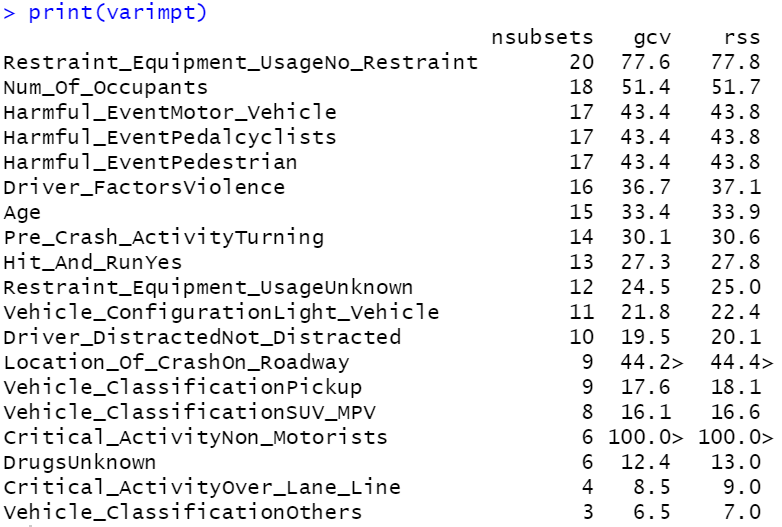
*Figure F-1. MARS Testset Metrics for varying number of degree (Without 10-fold Cross Validation)*

**

*Figure F-2. MARS Testset Metrics for varying number of degree (With 10-fold Cross Validation))*

**

*Figure F-3. MARS model with Degree = 1 and 10-fold cross validation*

**

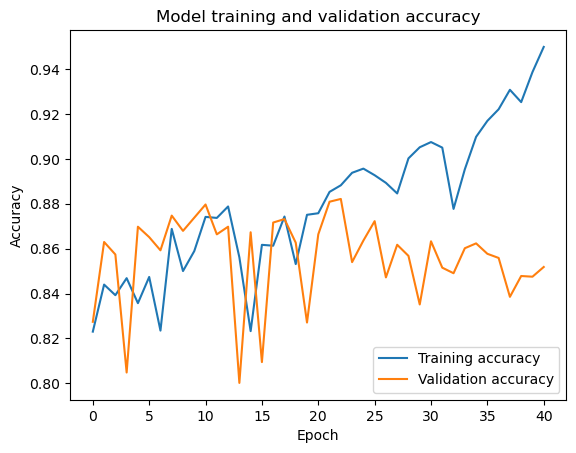
*Figure F-4. Variable Importance for MARS model with Degree = 1 and 10-fold cross validation*

## 

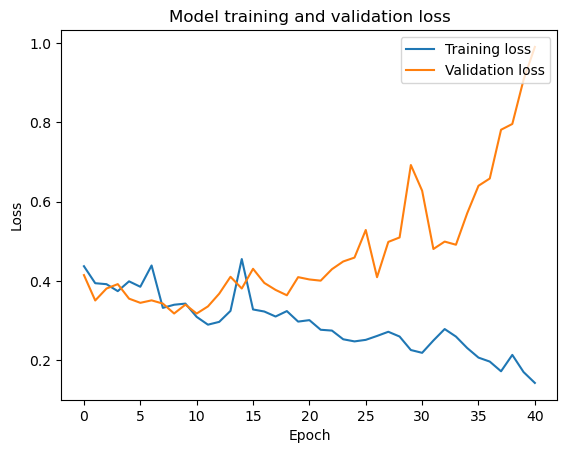
## **Appendix G: Neural Network**



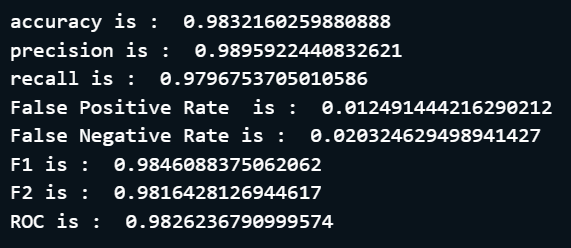
*Figure G-1. Neural Network Architecture*

**

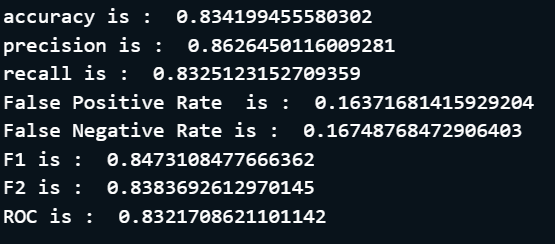
*Figure G-2. Neural Network Accuracy graph*

**

*Figure G-3. Neural Network Loss graph*

**

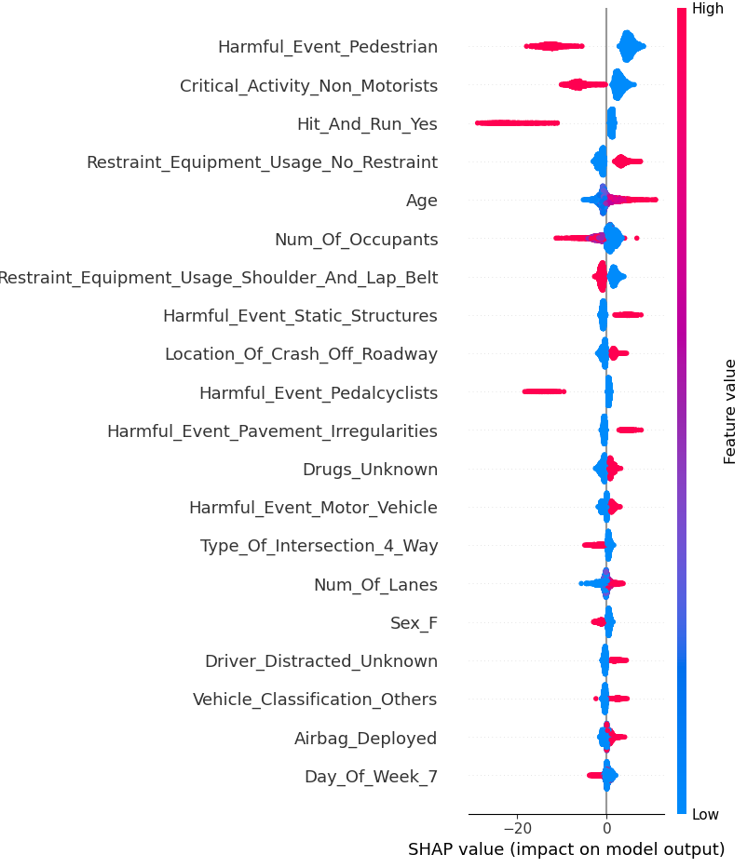
*Figure G-4. Neural Network Trainset Metrics*



*Figure G-5. Neural Network Testset Metrics*



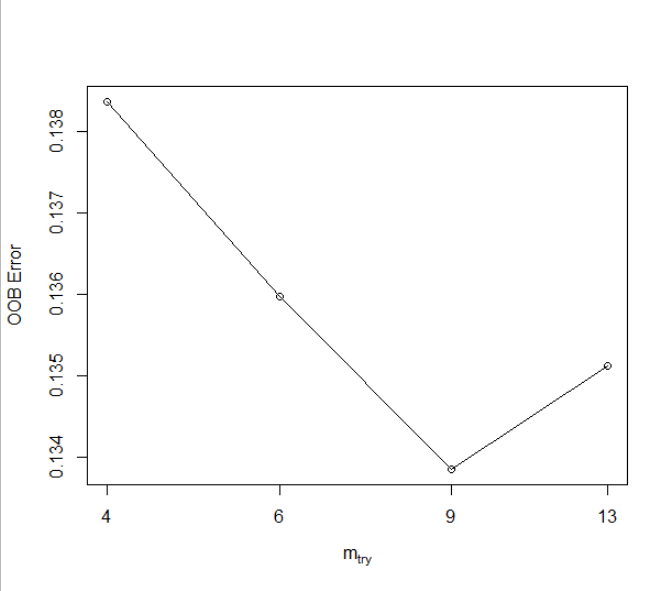
*Figure G-6. Neural Network Confusion Matrix*



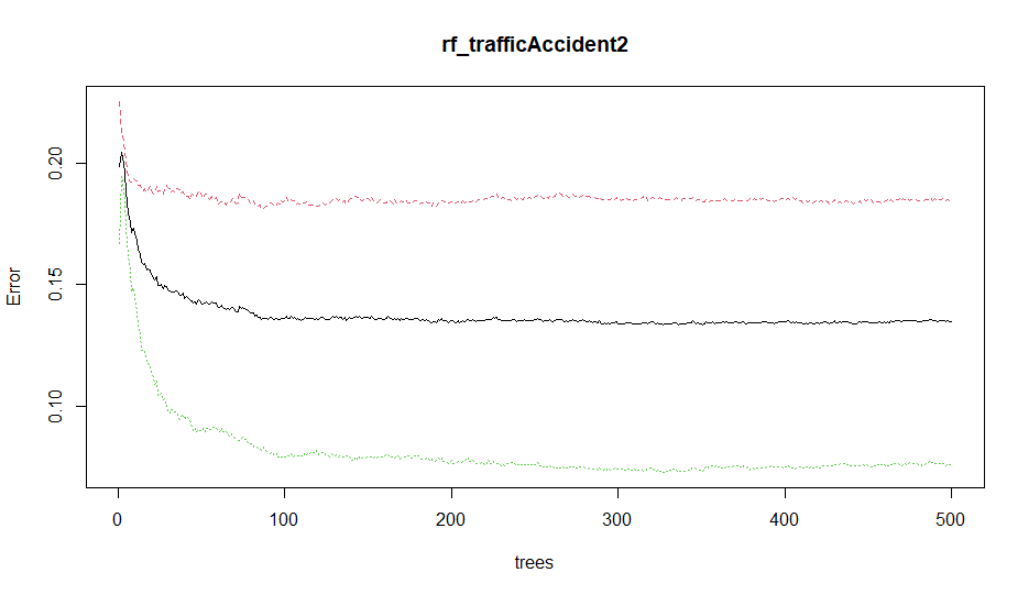
*Figure G-7. Neural Network SHAP Summary Plot*

## 

## **Appendix H: Random Forest**

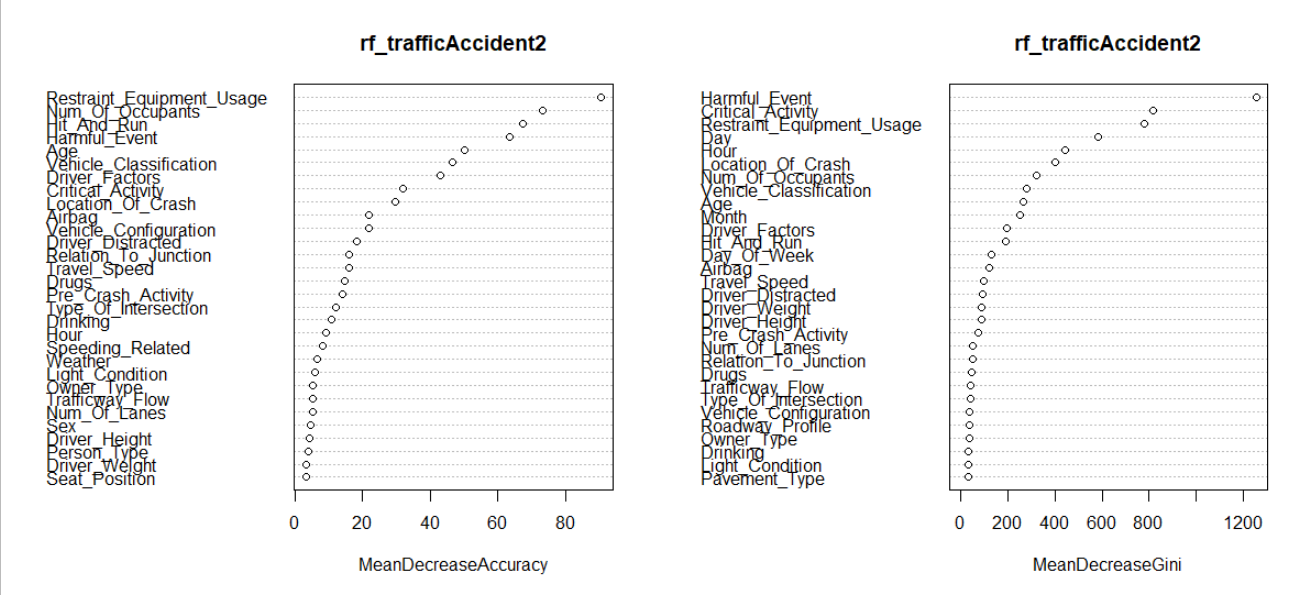
**

*Figure H-1. OOB Error against different mtry values*

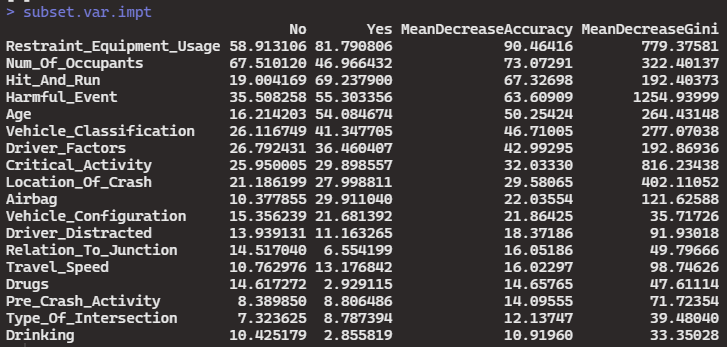


*Figure H-2. Error plot of Random Forest to show when trees stabilise*

*Random Forest Variable Importance*



*Figure H-3. Variable Importance (Random Forest) as a plot*



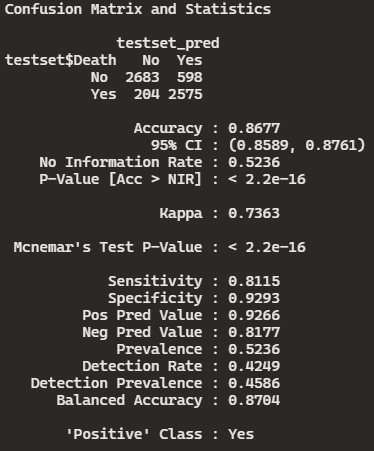
*Figure H-4. Variable Importance (Random Forest) as a table*

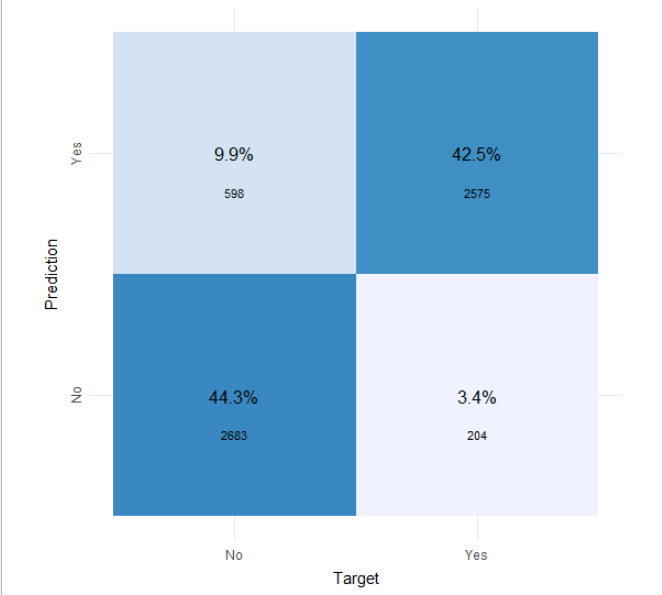


*Figure H-5. Random Forest Trainset Metrics*



*Figure H-6. Random Forest Testset Metrics*

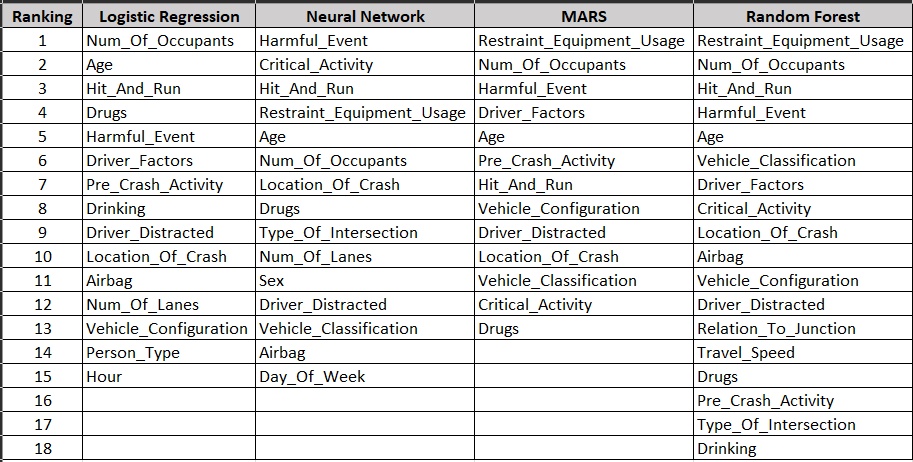




*Figure H-7. Random Forest Testset Confusion Matrix*

## 

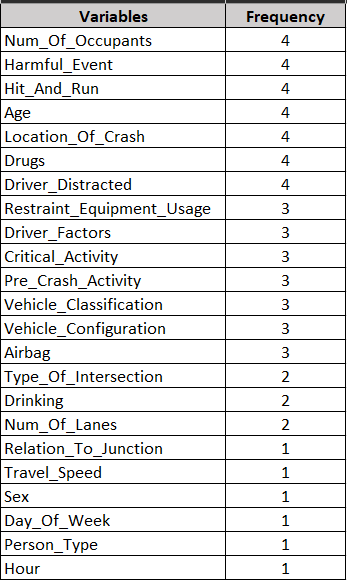
## **Appendix I: Variable Importance of The 4 Models**



*Figure I-1. Consolidated Variable Importance of the 4 Models*

**Appendix I-2: Frequency of Each Variable**

Frequency for each model is calculated by summing the number occurrence of the variable in the 4 models.



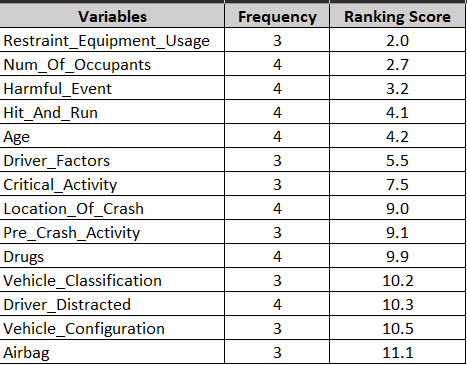
*Figure I-2. Frequency of Each Variable Occurring in the 4 models*

**Appendix I-3: Frequency & Ranking Score of Variables**

From all the variables shown in *Appendix J1*, we only consolidated the ranking score of variables that have appeared in at least 3 of the 4 models.

The ranking score shown below is calculated by dividing the total score by the frequency. The lower the ranking score, the more important is the variable in predicting the occurrence of a death in a traffic accident. The total score of each variable is calculated by summing the normalised rank of the respective variable in each of the models. We normalised the rank because each model has a different number of features that it deemed is important. Therefore, we standardised it to a count of 15 features since 2 models have 15 features in total (Logistic Regression & Neural Network), with MARS and Random Forest having 13 and 18 features respectively.

For example, Restraint\_Equipment\_Usage appeared in 3 models, and has a normalised rank of 4, 1.15 and 0.833 in Neural Network, MARS, and Random Forest respectively. Therefore, this equated to a total score of ~6. As such, the ranking score of Restraint\_Equipment\_Usage is ~2, signifying that it has the greatest importance compared to all the other features.



*Figure I-3. Frequency and Ranking Score of Variables*

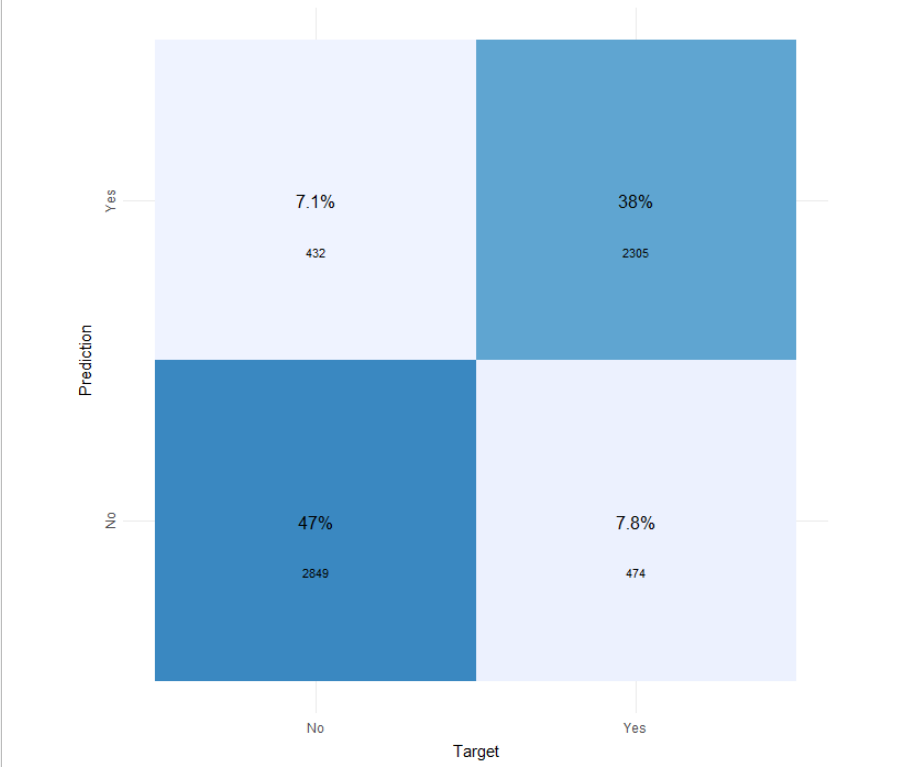
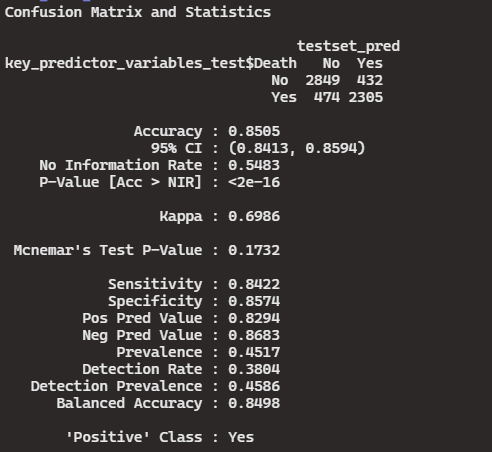
## 

## **Appendix J: Random Forest on Minimised Dataset**



*Figure J-1. Random Forest Trainset Metrics*



*Figure J-2. Random Forest Testset Metrics*

*Figure J-3. Random Forest Testset Confusion Matrix*

## **Appendix K: Curated Prediction Tool**



*Figure K-1. Mockup of the prediction tool*

1. TP = True Positive = Fatal accidents correctly predicted as fatal accidents [↑](#footnote-ref-1)