Jingcheng Zhou ID: 82204407

Maritess Pagaduan ID: 23659191

Christian Llave ID: 85662382

Huaipu Wang ID: 28629644

Pan Zheng

INFO 634

I. Abstract

INTRODUCTION: Road safety is a paramount concern for New Zealand, and the government's "Road to Zero" policy underscores its commitment to reducing accidents and their associated human and economic costs.

OBJECTIVE: This study employs a multifaceted approach to analyse and address road safety issues, encompassing temporal trends, regional comparisons, influences of various factors, and predictors of accident severity. Temporal trends analysis examines changes in car crash frequency over the past three years to identify emerging patterns. Regional comparisons assess car crash rates and severity across different parts of New Zealand, pinpointing geographical patterns and hotspots. Analysis of factors, such as weather conditions and road infrastructure, investigates their contributions to accident severity.

METHODS: The study leverages various data sources, including New Zealand Police data and the Crash Analysis System (CAS) from Waka Kotahi, to derive insights. To identify the most influential factors contributing to road accidents, the study employs the Random Forest algorithm. This powerful statistical tool assesses the relative importance of various factors and their impact on accident occurrence. The study aims to pinpoint key contributors to road accidents in New Zealand.

RESULTS: In our analysis of random forest models, certain variables consistently emerged as highly important across various types of crash outcomes. These key variables include street conditions like Street Light, Speed Limit, Number of Lanes, and Road Configuration, as well as environmental factors such as Region, Light, and Cliffs/Banks. Notably, when focusing on severe crashes, injuries, and fatalities, we observed that these same variables maintained their significance.

CONCLUSION: In New Zealand, there has been a progressive decline in car accidents from 2021 to early 2023. It's worth noting that car crash frequencies vary among different regions, with a significant increase during holiday periods, particularly around Christmas. Additionally, there is a higher occurrence of accidents between 15:00 PM and 18:00 PM, indicating a peak in accident rates during that time frame. In addition, random forest models have identified specific variables, such as Street Light, Region, Lanes, and Speed Limit, which consistently exhibit predictive power across multiple types of crash outcomes. Prioritising inspections and interventions related to these variables can assist in implementing more targeted and effective road safety measures.

II. Introduction

Road safety is a critical concern for any society, affecting both individuals and the community as a whole. New Zealand, like many countries, faces the challenge of reducing road accidents and their associated impacts. This introduction provides an overview of the road safety situation in New Zealand and the significance of analysing the factors influencing road accidents. New Zealand's "Road to Zero" policy, spearheaded by the government, reflects a strong commitment to minimizing road fatalities and serious injuries (NZ Transport Agency, 2019). Road accidents not only result in personal tragedies but also have far-reaching consequences for families and society at large. Understanding the factors that contribute to road accidents is essential for devising effective prevention strategies.

This study approaches the issue of road safety from multiple angles. It encompasses regional analysis, examining the variations in accident rates across different parts of the country. It also delves into the influence of holidays on accident rates, recognizing that these periods often see heightened traffic and potential risks. Hourly analysis will shed light on how accident rates fluctuate throughout the day, helping identify high-risk periods. To identify the most influential factors contributing to road accidents, this study employs the Random Forest algorithm. Random Forest is a powerful statistical tool that can assess the relative importance of various factors and their impact on accident occurrence. By employing this method, we aim to pinpoint the key factors that drive road accidents in New Zealand.

Road safety remains a paramount concern in New Zealand, with the government's "Road to Zero" policy underscoring its commitment to reducing accidents and their associated human and economic costs. By conducting a comprehensive analysis of regional, holiday, and hourly factors using the Random Forest algorithm, this study seeks to provide valuable insights into the primary contributors to road accidents. This research aims to serve as a foundation for the development of effective strategies and policies to prevent accidents, thus making New Zealand's roadways safer for all its citizens.

III. Literature Review

Numerous factors contribute to car crashes, including road conditions, weather, driver behavior, and more (Kerner,2018). However, it is well-established that the state of traffic networks plays a pivotal role in people's lives. Research investigating the connection between specific factors and car crash outcomes has been conducted in various countries worldwide, particularly in regions prone to significant car accidents or catastrophes (Haynes,2008). In Australia, for instance, studies have revealed compelling evidence that the aftermath of car crashes significantly impacts mental health, leading to conditions such as depression and post-traumatic stress disorder. Furthermore, a separate study has shown that elevated car crash rates in Japan and China are associated with higher levels of individual health issues. Additionally,

recent research suggests that car crashes are not solely linked to individual characteristics and life events but are also intertwined with our society as a whole.

New Zealand has witnessed a substantial 49% increase in crash statistics compared to the figures achieved in 2013. A comprehensive study modelled twenty-one factors over a period of four years before and four years after 2013, and the findings have identified three significant contributors to this surge: alcohol, learner licenses, and the Auckland region (Lewis-Evans,2010). These results suggest a systemic failure in the detection and prosecution of offending related to road safety (Poulsen,2012).

Similar to Australia, New Zealand has a rich body of research that suggests a profound connection between car crashes outcomes. Studies in Australia have uncovered compelling evidence that the aftermath of car crashes significantly impacts individual, leading to conditions such as depression and post-traumatic stress disorder (Tay,2001). As New Zealand shares many similarities with its Australian neighbour, it is imperative to investigate whether these findings are replicated in the unique Kiwi context.

As we explore the intricate relationship between car crashes and society as a whole, it is crucial to examine how New Zealand's cultural and societal fabric interweaves with this phenomenon. Recent research has illuminated the notion that car crashes are not solely linked to individual characteristics and life events but are also intertwined with the broader society. New Zealand, with its unique blend of indigenous Māori culture and multicultural diversity, presents an intriguing case study to understand the broader implications of car crashes on society and the potential interventions needed to safeguard the nation's well-being (Walton,2020).

IV. Research Question

1. Temporal Trends: Examining changes in car crash frequency over the past three years to identify any emerging trends or patterns.
2. Regional Comparison: Evaluating car crash rates and severity across different regions in New Zealand, with a focus on geographical patterns and hotspots.
3. Influence of Factors: Investigating how weather conditions, time of day, road infrastructure, etc. contribute to the severity of car crashes in New Zealand.

V. Data Sources

The New Zealand Police Data’s Demand and Activity section provides counts of recorded vehicle collisions in the last three years at most (2020 - 2023). This dataset contains over 205,000 rows and 36 columns including territorial authority, region, time of day, and month of year, which are relevant to the study. There is an opportunity to group territorial authorities by urban context to observe differences in time-of-day patterns between city and non-city incidents.

Meanwhile, the Crash Analysis System (CAS) from Waka Kotahi provides more field details of over 821,000 crashes. This dataset contains 72 columns, including the variables such as the presence of roadworks, coordinates, speed limits, debris, road conditions, weather, and light among other characteristics. These contain variables that can be counted as predictors to determine important factors that contribute to crash outcomes. The interpretation of crash variables are available in their data dictionary (Waka Kotahi NZ Transport Agency), and a definition of severity is found in their guide to treatment of crash locations (Waka Kotahi NZ Transport Agency) to get a better definition of the levels of severity.

VI. Methods

We aim to derive insights from the complex landscape of car crashes in New Zealand by applying a variety of data analytics techniques. To provide a multidimensional view of the phenomenon, our methodological framework includes both machine learning statistics and powerful data visualisation techniques.

a. Crashes over time

Analysing temporal trends in car crash rates is a critical first step towards improving road safety. Authorities can proactively address potential safety hazards by identifying emerging patterns and shifts in collision occurrences during the last three years. With this knowledge, specific interventions can be strategically planned and carried out. For example, if an increase in accidents is observed during specified time periods, law enforcement authorities can increase their presence on the roadways during those hours, discouraging hazardous behaviour and assuring fast response to crises.

b. Crashes by region

Understanding regional disparities in car crash frequencies and severity assists authorities to better allocate resources and respond to incidents. Using a detailed analysis of crash data, geographic hotspots or places with greater accident rates can be precisely determined. As a result, law enforcement, emergency responders, and municipal governments can strategically distribute resources where they are most needed. This tailored approach improves their effectiveness and response times, potentially reducing the impact of accidents and saving lives. Furthermore, concentrating efforts in areas with a greater crash frequency can result in more efficient use of financial and people resources, ensuring that limited resources are used where they can have the most influence on road safety.

This is complemented by an analysis of the likelihood of specific crash severities for each region. Likelihood metrics account for the disproportionate counts relative to the area's population, while displaying the most characteristic crash severity for each region.

Heatmaps display geographic patterns and hotspot concentrations of car crashes. We aim to generate interactive and informative heat maps using Tableau. These heat maps will visually depict locations with high crash frequencies, highlighting zones that require targeted

interventions. This is further enhanced by finding the likelihood of certain crash severities for each region, to even out the effect of high crash volumes in population-dense areas like Auckland. Stakeholders can go into specific areas of interest for more in-depth explorations thanks to Tableau's interactive nature.

c. Crashes by time of day

Exploring the trends in the time of day allows for authorities to allocate resources at peak times, such as restricting night-time driving for particular vehicle types or increasing surveillance during peak risk hours. Variations in peak times are potentially affected by the urban context and day of the week. These variations are then accounted for in the graphical interpretation.

d. Predictors of severity

Investigating the impact of many factors such as weather and road infrastructure provides significant insights into the underlying causes of severe crashes. This knowledge can be used to create educated policies, rules, and guidelines targeted at reducing accidents and risk. For example, if data show that a specific weather condition routinely leads to accidents, policymakers should create weather-specific driving advisories and safety campaigns to educate drivers on safe driving practices in such conditions.

Our predictions come from the use of tree-based methods, which is regarded as generally more robust than linear regression models, especially when handling categorical data (Varghese,2018), non-linear relationships, and variable interactions. We aim to use random forests to deduce complex connections between multiple contributing elements and the level of severity. Random forests can illuminate the decision-making process that leads to specific crash scenarios by recursively splitting the data based on the most important variables. This, in turn, provides a more detailed understanding of the factors that determine crash severity and occurrence.

Because crash severity is described as levels, such as Fatal, Serious, Minor, and Non-Injury, the random forest in this case is a classification model and will be measured against its recall and precision. Recall is used as the main metric because we want the model to capture the actual severe crashes to be able to properly determine their factors. Precision is the secondary metric to ensure the model can have predictions that are truly severe, as misclassifying minor crashes as severe will have an impact on resources to be allocated.

e. Predictors of the number of casualties and damage

Similarly, random forest models also allow us to connect the number of casualties or damage dealt by the crash with its important factors. Understanding the predictors point to factors that are associated with higher casualties and damage, and how to reduce them. Because the model output is numeric, the random forest performs a regression task, and will be measured against its mean-square error (MSE).

VII. Data Cleaning, Wrangling, and Pre-Processing

Tableau

a. Urban Context

Using the Police Data, we used Regular Expressions to create a column based on the Territorial Authority, enabling us to categorise it into two distinct groups: Urban and Non-Urban.

b. Weekday Context

Using the Police Data, we used Regular Expressions to create a column based on the "Occurrence Day of Week", enabling us to categorise it into two distinct groups: Weekday and Weekend.

R Studio

c. Coordinates

In accordance with the CAS data field descriptions, the CAS data originally provided coordinates in terms of "northing" and "easting." These are typically used in specific map projections or coordinate systems. However, for better compatibility and visual representation in Tableau, we converted these coordinates into the more commonly used "longitude" and "latitude" by using R.

d. Model Variables

We sorted out which columns were viable predictor and response variables. We first removed metadata columns and derived variables. We then grouped together variables by the kind of casualty they were, such as vehicle damage, object damage, and property damage. We then segmented outcomes (count of casualties, severity type, count of injuries) as responses, and remaining indicators (ex. environmental factors, road infrastructure) as predictors. Our pipeline ensured the exclusion of response variables from other models onto the model in question, to avoid them from being used as a predictor.

e. Null Values

We filled in null values for speedLimit based on the speed limits mandated by Waka Kotahi. Other null values were part of the data options as they were not applicable or were properly interpreted as nil or having no value. This process was to ensure the variables were usable in the random forest and regression modelling.

f. Class balance for the Random Forest classifier

There are four main classes in Crash Severity: Fatal Crash, Serious Crash, Minor Injury, and Non-Injury. Most of the crashes in the CAS are in Minor and Non-Injury, which have lower

stakes than Fatal and Serious Crashes, as the latter would require more action to prevent increased social cost. Additional pre-processing was done to account for this.

We declared initial class weights as the reciprocal of the count of each class. Using stratified random sampling, we preserved the original proportions for the test set to mimic the variations present in the original environment. Merging classes as Serious (Fatal and Serious Crashes) and Non-Serious (Minor and Non-Injury) in one of the experiments allowed us to re-frame the problem as binary instead of multi-class. De-duplicating from the majority (Non-Serious) class was done to reduce redundancies and reduce noise. Downsampling was also used in one of the experiments to better tune towards the prediction of Serious crashes.

VIII. Analysis Methods and Implementation

A. Exploratory Analysis

Regional Trend Comparison

Car accidents vary significantly across New Zealand's regions in 2021, as seen in Table 1. According to the 2018 census data, Auckland had the largest population, followed by Canterbury and Wellington. Notably, the New Zealand Transport Agency reported a high percentage of full class 1 licence holders, averaging around 80% of the total population. This indicates that a significant portion of the population is licensed to drive at various stages (full, restricted, and learner). From Table 2, during the car crash period, Auckland recorded the highest number of car accidents, followed by Waikato and Canterbury, which is expected given their larger populations. Interestingly, even regions with smaller populations experienced a considerable number of accidents, possibly due to unique road conditions and driving challenges, as well as the increase in tourism. The percentage of accidents relative to the population (per 1,000 people) provides insight into the frequency of accidents for every thousand individuals.

Region	Population Census (2018)*	Driver License Holder (2021)*	Crash Victim Count (2021)	% Pop w/ License	% Crash/Pop
Auckland	1,572	1,270	3,727	81%	0.2%
Bay Of Plenty	308	254	953	82%	0.3%
Canterbury	600	492	1,407	82%	0.2%
Gisborne	48	0	224	0%	0.5%
Hawke's Bay	166	128	587	77%	0.4%
Manawatu-Whanganui	239	186	1,039	78%	0.4%
Marlborough	47	39	115	83%	0.2%
Nelson	51	42	135	83%	0.3%
Northland	179	140	743	78%	0.4%
Otago	225	181	696	81%	0.3%
Southland	97	76	353	78%	0.4%
Taranaki	118	92	356	79%	0.3%
Tasman	52	0	182	0%	0.3%
Waikato	458	359	1,779	78%	0.4%
Wellington	507	394	1,091	78%	0.2%
West Coast	32	0	120	0%	0.4%
Grand Total	**4,668**	**3,655**	**13,387**	78%	0.3%

**Population & DL Holders are based on value per 1,000 people*
https://opendata-nzta.opendata.arcgis.com/documents/driver-licence-holders/about

(Table 1: New Zealand Population, Registered Drivers, and Crash Victim Statistics)

The Number of Crash Severities Categorized by Regions

	Crash Severity				
Region	Non-Injury Crash	Minor Crash	Serious Crash	Fatal Crash	Grand Total
Auckland Region	19,856	8,250	1,642	153	29,901
Bay of Plenty Region	4,128	1,853	442	98	6,521
Canterbury Region	5,659	3,020	762	110	9,551
Gisborne Region	915	403	116	20	1,454
Hawke's Bay Region	2,836	1,197	317	44	4,394
Manawatu-Wanganui Reg..	4,259	1,985	541	99	6,884
Marlborough Region	664	270	68	16	1,018
Nelson Region	578	318	51	6	953
Northland Region	2,714	1,476	428	100	4,718
Otago Region	3,129	1,501	333	52	5,015
Southland Region	1,416	699	172	32	2,319
Taranaki Region	1,599	756	217	27	2,599
Tasman Region	580	357	82	11	1,030
Waikato Region	7,215	3,693	961	191	12,060
Wellington Region	6,557	2,435	531	46	9,569
West Coast Region	459	268	86	12	825
Grand Total	62,564	28,481	6,749	1,017	98,811

Count %

0.0% 100.0%

(Table 2: Number of Crash Severities Categorised by Regions)

One of the leading causes of car collisions is excessive speeding, which frequently leads to accidents (NZTA). In Table 3, it's evident that approximately 70% of car crashes in New Zealand occurred in areas with speed limits ranging from 50 to 99 km/h, which include urban and open roads. Additionally, more than a quarter of these accidents took place in areas such as motorways and expressways where the speed limit exceeds 100 km/h. While road accidents happening in locations with less than 50 km/h speed limit only, such as school zone, coast areas, or emergency spots, comprises about 3% of the total collision count, it still resulted in fatality and serious injuries of about 1% and 18% respectively.

Regional Crash Count based on Speed Range

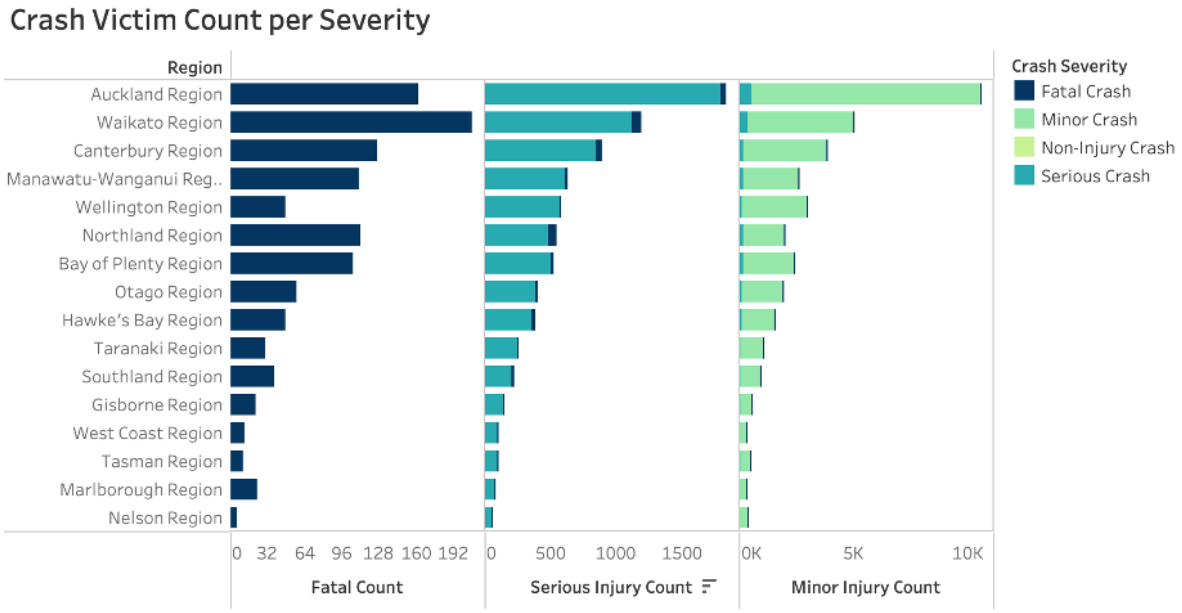
Region	Speed Range			
	Below 50	50-99	100 above	Grand Total
Auckland Region	976	24,090	4,835	29,901
Bay of Plenty Region	63	4,441	2,017	6,521
Canterbury Region	277	6,956	2,318	9,551
Gisborne Region	14	935	505	1,454
Hawke's Bay Region	31	2,943	1,420	4,394
Manawatu-Wanganui Reg..	39	4,397	2,448	6,884
Marlborough Region	56	677	285	1,018
Nelson Region	39	821	93	953
Northland Region	57	2,431	2,230	4,718
Otago Region	218	3,088	1,709	5,015
Southland Region	20	1,331	968	2,319
Taranaki Region	53	1,613	933	2,599
Tasman Region	15	577	438	1,030
Waikato Region	318	6,456	5,286	12,060
Wellington Region	569	7,288	1,712	9,569
West Coast Region	14	290	521	825
Grand Total	2,759	68,334	27,718	98,811

Count %

0.0% 100.0%

(Table3: Regional Crash Count based on Speed Range)

The degree of reported crashes ranges from non-injury incidents to fatal accidents. Auckland, due to its popularity, large population, and high traffic volume, accounts for 30% of the reported crashes, while Waikato and Wellington contribute 12% and 10%, respectively. In regions where the "100 above" speed range category records higher crash counts, such as Auckland, Waikato, and Canterbury, there is a corresponding increase in the occurrence of severe crashes, including fatal and serious accidents. This suggests that high-speed crashes are associated with higher severity.

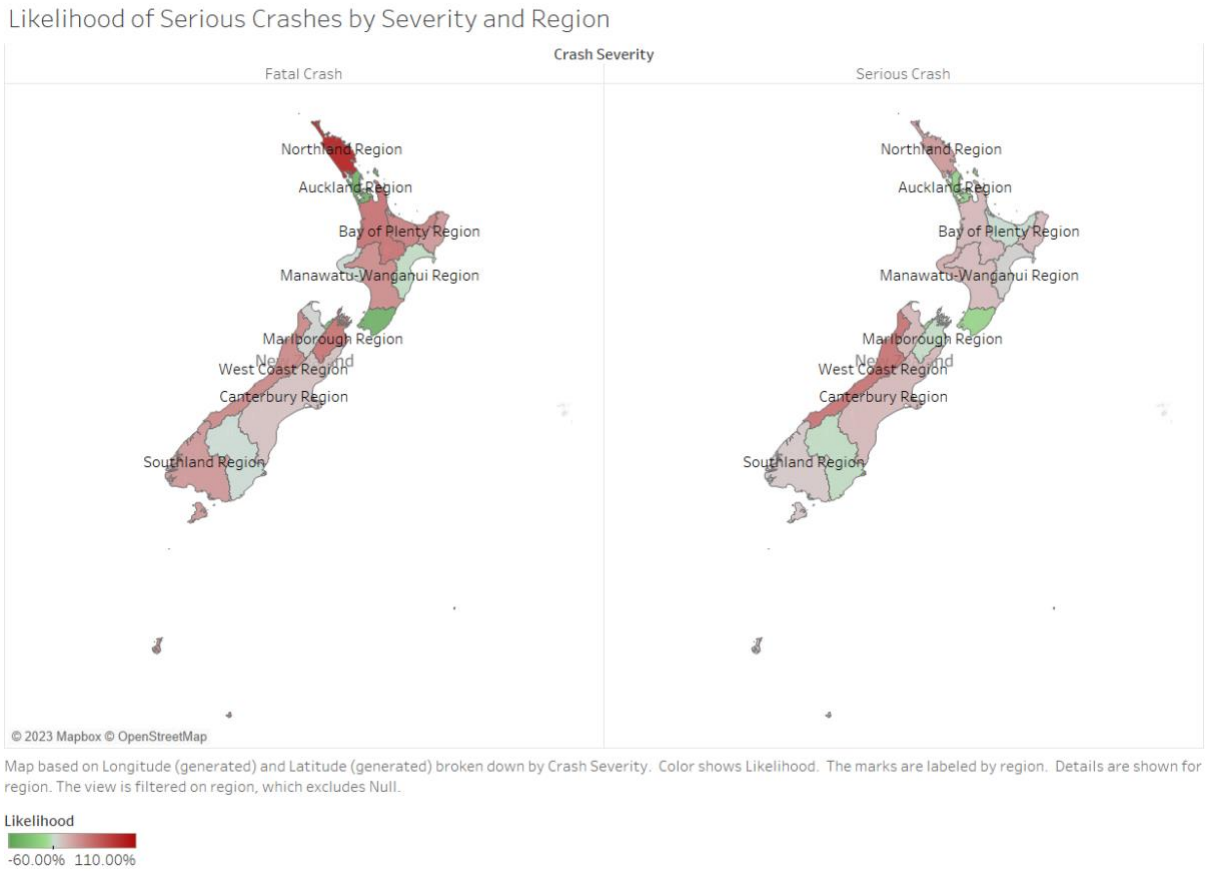


(Figure 1: Crash Victim Count per Severity)

In terms of the number of casualties, fatal crashes caused about 2% death toll from the total crash victim count. Mortalities resulting from road accidents are a significant concern across various regions in New Zealand. Waikato, Auckland, and Canterbury have relatively high counts of fatal crashes. Reducing the number of fatalities should be a priority, and it may involve initiatives to address issues like over speeding, distracted driving, and road infrastructure improvements.

Serious injuries result from both fatal and serious crashes, with the former contributing to 5% and the latter contributing to 95% of these injuries. These injuries can have long-lasting physical, emotional, and economic impacts on individuals and communities. Reducing major injuries may require improvements in vehicle safety features, road design, and medical response times.

Minor crashes account for the majority of minor injuries although there are relatively minimal counts of serious injuries. High counts in Auckland, Waikato, and Canterbury suggest a need for continued efforts to enhance road safety, including awareness campaigns, driver education, and enforcement of traffic laws.



(Figure 2: Likelihood of Serious Crashes by Severity and Region)

While count measures provide immediately interpretable action points, likelihood measures provide a complement that characterises the relationship between the region and crash severity regardless of their absolute counts. This method was adapted from

GlobalWebIndex's index metric (GWI). This was done by taking the proportion of severity for each region, and dividing that value by the proportion of crashes per region.

These values are then plotted on the map, with red values showing regions that are more likely to have serious crashes. In the plot, Northland Region is the one most characterised as more likely for crashes to be Fatal Crashes, which may be worth inspecting why this could be the case. Perhaps there are characteristics of the road, the driver, or other external factors that affect crashes in regions with high likelihood for serious crashes.

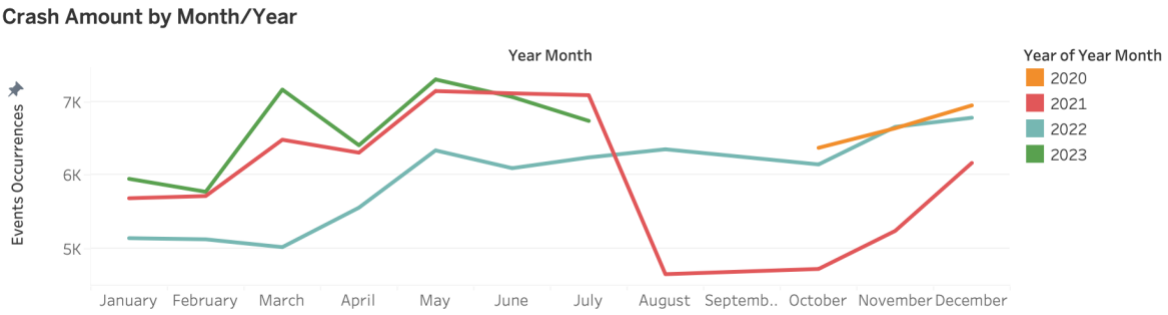
In contrast to the count-based metrics, Auckland is less characterised by the likelihood of Fatal Crashes despite its absolute counts. This could mean the volume and population within Auckland may be contributing more to the count of crashes than the characteristic of the region itself.

(Figure 3: Year-on-Year Crash Count)

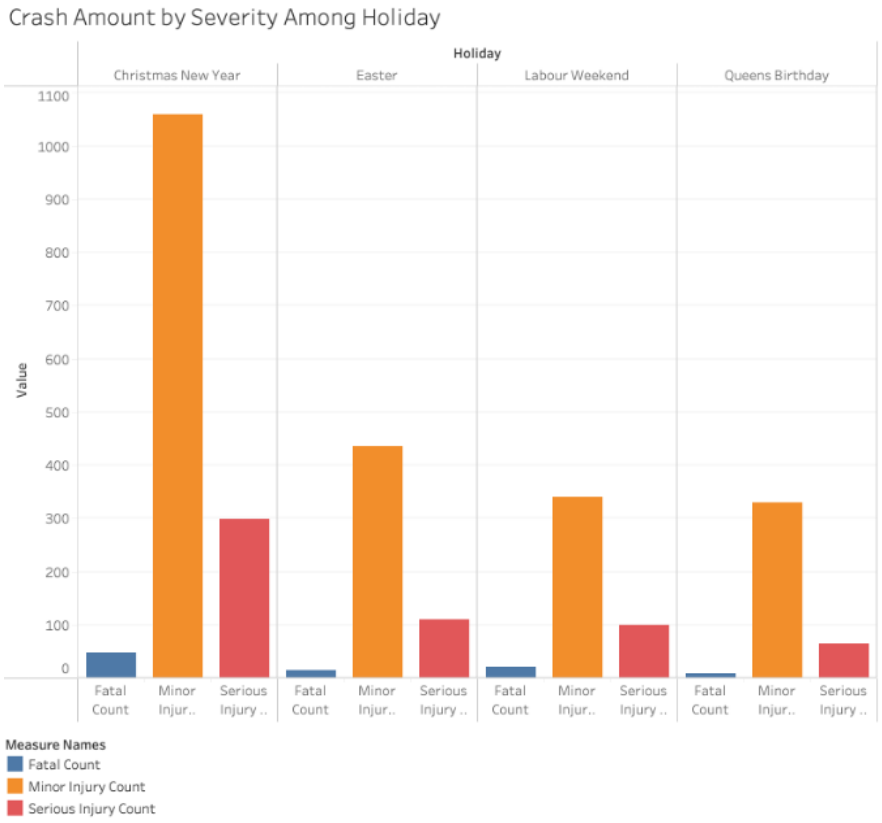
Overall, the New Zealand government's Road to Zero strategy has shown positive progress in reducing overall crash counts nationwide between 2020 and early 2023. However, it is still a considerable distance from achieving its vision of fully eliminating or reducing fatalities and severe injuries from vehicular accidents by 40%. Realising this aspiring goal will require significant effort and unwavering commitment from both the government and the public. Implementing this project on a nationwide scale is imperative to ensure the successful execution of all relevant action plans and to ensure that the budget is efficiently utilised in creating safer roadways for all.

Car Crash records monthly Trend

As observed in Figure 4, there is a slight increase in accidents during the months of April to May and November to December annually. These specific time frames correspond to Easter and Christmas holidays, suggesting that holidays may play a significant role in contributing to the increased incidence of car accidents. Additionally, from Figure 5, we can observe that when considering different car crash severity, Christmas is the holiday with the highest car crash amount, followed by Easter.



（Figure 4: Crash Amount by Month/Year)



(Figure 5: Crash Amount by Severity Among Holiday)

Accidents tend to occur more frequently during holiday periods for a variety of reasons. Firstly, there is a notable increase in traffic volume during holidays, as more vehicles occupy the roads, thereby elevating the overall risk of accidents. Secondly, many individuals find themselves driving on unfamiliar roads during vacations, leading to navigation challenges and increased likelihood of wrong decisions. Furthermore, long hours of driving during the holidays, often during early mornings or late evenings, can result in driver fatigue, which impairs judgement and reaction times. Additionally, holiday driving can be characterised by heightened stress and frustration due to factors like heat, traffic jams, noisy children, family tensions, and general end-of-year fatigue. Moreover, individuals on holiday may tend to let down their guard when it comes to road safety, engaging in behaviours such as speeding,

driving while fatigued, or neglecting seatbelt use. Lastly, there is an increased incidence of drink-driving during holiday periods, further exacerbating the risk of accidents.

Several measures have already been implemented to address the challenges posed by increased holiday traffic and mitigate potential accidents. The NZ Transport Agency Waka Kotahi has taken proactive steps by publishing a journey planner, providing valuable insights into the country's popular routes and forecasting peak traffic times. This includes an interactive traffic prediction map for holiday periods, which utilises past travel patterns to anticipate heavy traffic periods. Furthermore, real-time updates on route changes, delays, closures, and incidents are readily available to travellers. Given the dynamic nature of predicted peak times influenced by incidents, weather conditions, and driver behaviour, it is strongly recommended that individuals consult the real-time journey planner before embarking on their journeys to access the latest information regarding road works, traffic conditions, road closures, and potential delays, ensuring safer and more efficient travel during holidays.

Car Crash records hourly Trend

We analysed the police data on vehicle collisions and observed an increase in the number of incidents per hour, particularly between 15PM and 18PM. This suggests that these hours might be a significant indicator.

Exploring plots depicting the distribution of incidents by units, weeks combinations would also be a valuable endeavour.

From Table 4, it's evident that in the Auckland region, there is a significantly high number of collision records between 15PM and 18PM. Not only in Auckland, but also in other regions. This finding isn't surprising, given Auckland's dense population. Following Auckland are the Canterbury and Wellington regions.

Crash Record by Hour

Recorded Occ Type Group / Hour Band
Vehicle Collision

% of Total Number of R..
2.63% 26.69%

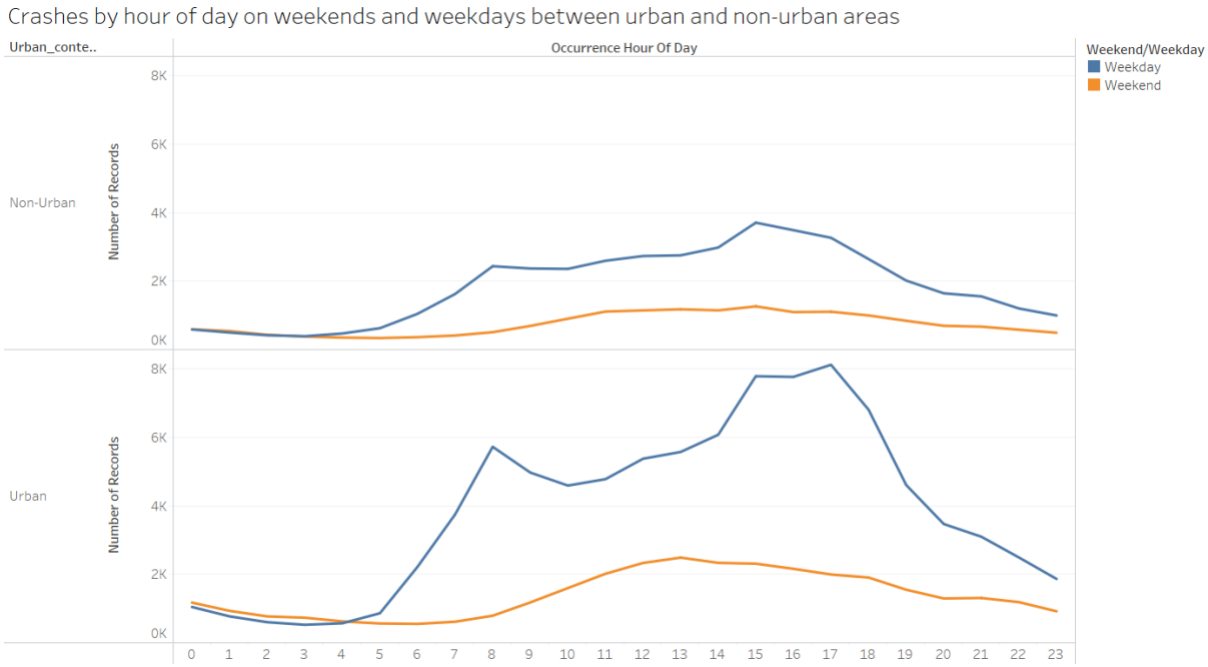
Regions	00:00-02:59	03:00-05:59	06:00-08:59	09:00-11:59	12:00-14:59	15:00-17:59	18:00-20:59	21:00-23:59
Auckland Region	2,914	2,220	7,918	10,429	12,824	16,042	11,305	5,985
Canterbury Region	855	687	1,882	3,016	4,056	4,981	3,012	1,915
Wellington Region	744	527	2,037	2,971	3,720	4,477	2,722	1,424
Waikato Region	843	701	1,946	2,829	3,617	4,179	2,704	1,559
Bay of Plenty Region	654	458	1,476	2,254	2,577	3,081	1,985	1,242
Manawatu-Wanganui Reg..	547	471	1,155	1,800	2,258	2,670	1,583	999
Northland Region	383	292	793	1,254	1,498	1,676	1,056	713
Hawke's Bay Region	367	276	733	1,104	1,364	1,751	997	589
Otago Region	337	240	698	1,093	1,298	1,651	967	618
Taranaki Region	229	156	401	675	798	874	550	366
Southland Region	167	135	304	444	594	714	461	312
Gisborne Region	122	88	194	356	371	477	333	188
Nelson Region	50	57	121	253	317	414	209	130
Tasman Region	52	40	135	251	275	401	233	132
Marlborough Region	52	40	130	256	304	380	185	108
West Coast Region	43	39	88	170	225	268	144	83
Chatham Islands Territory.		3	3	1	4	5	6	2

（Table 4: Crash Record by Hour)

Based on Table 4, we can observe that the crash records are highest in the time interval from 15:00 to 18:00 for almost all regions. When we carefully consider this, we can identify several reasons for this phenomenon.

Firstly, during this time interval, people are often at their most fatigued. After a day of work or study, energy levels tend to decline, increasing the risk of fatigue-related driving accidents.Secondly, it's a time when many students are getting out of school, and parents are picking them up. The presence of a large number of students and parents on the roads can lead to increased congestion and potential risks, including the possibility of children crossing roads unpredictably.Additionally, there are seasonal factors to consider. During this time range, the angle of the sun can result in direct glare that impairs drivers' vision, potentially leading to crashes. Lastly, this time interval coincides with rush hour for some companies, leading to heavy traffic and congestion, which also contributes to an increased risk of crashes. These factors combined make the period between 15:00 and 18:00 a high-risk time for road crashes in various regions.

Figure 6 displays the trends in car crashes on both weekdays and weekends, 24 hours a day, in Urban and Non-Urban areas. It is not surprising to note that, overall, car crash rates are lower in Non-Urban areas compared to Urban areas. This can be attributed to the higher population density and increased human activity in Urban areas. Furthermore, it is evident that both Urban and Non-Urban areas experience a significant increase in car crashes during the weekday hours of 6-9 AM. This can be logically attributed to the morning commute when people are driving to work, resulting in a higher number of car crashes. This trend is particularly pronounced in Urban areas, as a majority of companies and workplaces are located there. Additionally, there is another notable increase in car crash rates during the hours of 3-5 PM on weekdays. This can be attributed to factors discussed earlier, including driver fatigue, school dismissal times, and the impact of sunlight. During the weekend, the overall trend shows lower car crash rates compared to weekdays, reflecting the fact that many individuals are at home and not commuting to work.



(Figure 6 : Crashes by hour of day on weekends and weekdays between urban and non-urban areas)

B. Using random forest to classify crashes as serious or non-serious

We trained several models of the classifier. We started with a multi-class classifier to directly label crashes with the original classes (Fatal, Serious, Minor, and Non-Injury); however, this model had poor accuracy and recall, which meant it could not predict much of the actual classes, and allows for a lot of serious crashes to be overlooked.

Model	**Class Weight**	**Downsampling**	**Test Accuracy**	**Recall**
4 classes	Reciprocal	None	0.2969	0.25 (Fatal), 0.41003 (Serious)
2 classes, de-duplicated	Reciprocal	None	0.51	0.59038
2 classes, de-duplicated	Reciprocal	1.5x	0.3171	0.86090
2 classes, de-duplicated	Reciprocal of log	1.5x	0.5604	0.62179
2 classes, de-duplicated	None	1.5x	0.7558	0.38590
2 classes, de-duplicated	Reciprocal	3x	0.4144	0.71795
2 classes, de-duplicated	Reciprocal of log	3x	0.633	0.52500

(Table 5: Classification models and effects of class balancing)

We then went from multi-class to a binary classification problem after grouping Fatal with Serious as Serious and Non-Injury with Minor crashes as Non-Serious. This made sense as the first two groups are of stronger interest as they have higher social cost. The majority class (Non-Serious) was also de-duplicated to reduce the noise in an effort to balance the

classes. The accuracy and recall improved; however, there are still 40% of serious crashes being misclassified as non-serious.

The next model adds downsampling of the majority class up to 1.5x the minority class count. This greatly improved recall (i.e. reducing uncaught serious crashes), but suffered in overall accuracy, which means we are also predicting a lot of non-serious crashes as serious. This means the actual features of a serious crash are yet to be characterised.

One other method to improve the model was adjusting the class weight, as using the reciprocal of class counts may have had too large of an effect on weighting given the large gap between the two class counts. We tried using the reciprocal of the log of counts, which makes the gap smaller, but uses the reciprocal to favour the Serious crashes. This yielded the best model so far, as it had higher accuracy and a fair amount of recall.

To demonstrate the effects of class weights, we tried a version without it. The accuracy shot up; however, it was mostly from predicting the majority class correctly, and contains plenty of uncaught Serious crash cases.

The next iteration eases on the downsampling from 1.5x to 3x the minority class count. Compared to its 1.5x counterpart, it had a higher accuracy, but lower recall. However, this model still has less than 50% of predictions being correct. Another attempt is to use the reciprocal of log class weight. This improved the accuracy, exceeding the 50% mark; however, has a much lower recall than the 1.5x counterpart. At that point, nearly 50% of actual Serious cases are uncaught.

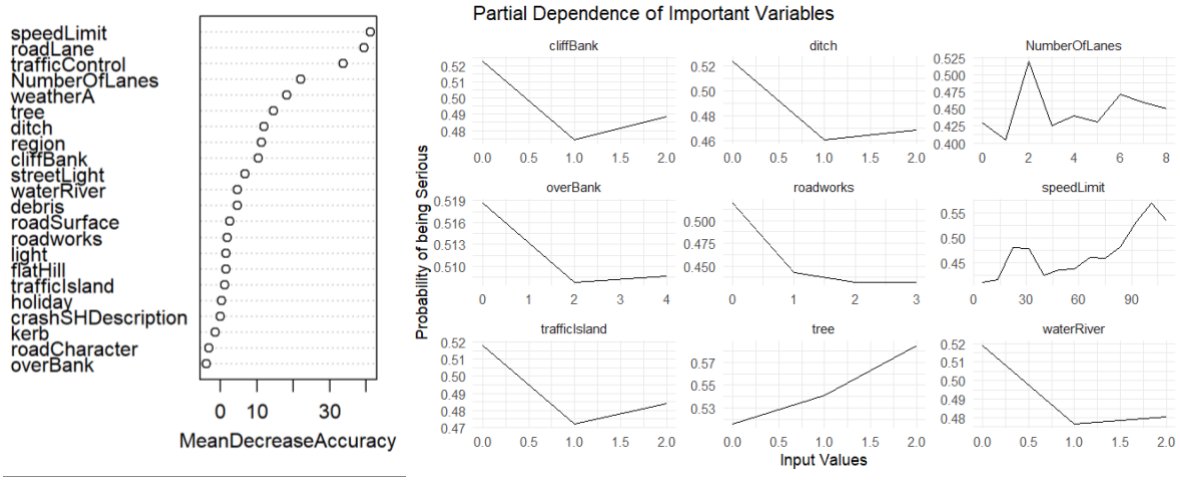
We settled on the binary model with a de-duplicated and downsampled majority to 1.5x the minority class, weighted by the reciprocal of the log of class counts.

	Reference / Actual	
Prediction	Non-Serious	Serious
Non-Serious	10,158	590
Serious	8,138	970

(Table 6: Confusion matrix of the best model)

The most important features to predicting Serious crashes consist of the road conditions and possible obstructions (both artificial and natural) such as speed limit, road configuration (roadLane), traffic control, number of lanes, weather, trees, ditch, region. The scale on the graph represents how much error is introduced when we do not consider these variables. To visualise how each variable affects the probability of a crash being serious, we used a partial dependence plot. While this oversimplifies the capabilities of the Random Forest, it gives a more interpretable view of the effects of each predictor in isolation. We can see some signs

that serious crashes involve less obstructions, 2-lane roads, and higher speed limits. It would then be good to provide extra precaution to those areas.



(Figure 7: Variable importance and partial dependency for best classification model)

C. Using random forest to predict the number of casualties

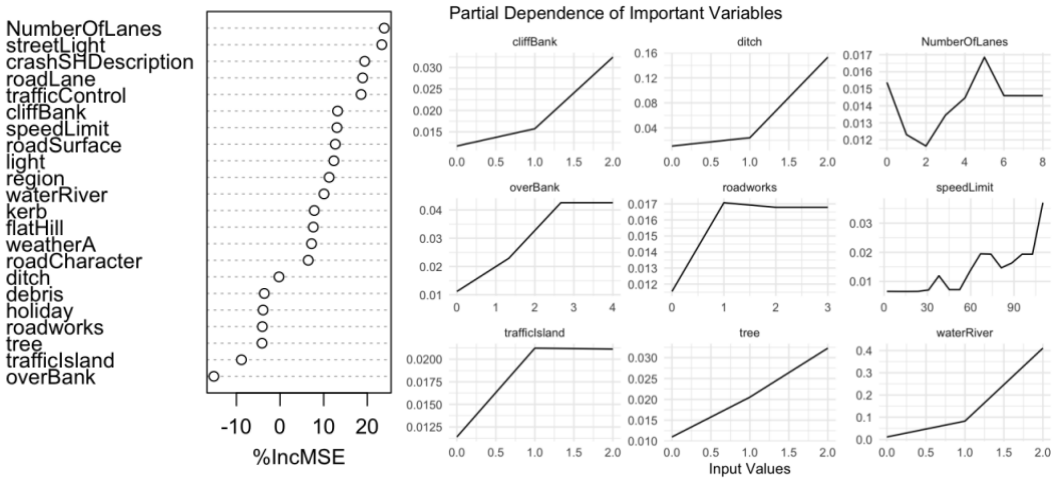
We were able to run a random forest model on the dataset to predict the number of severe injuries. This allows us to run the algorithm to predict different response variables such as the other injuries as a regression task, or predicting the crash severity as a classification task.

Similar inputs were used in the regression models, but we are now predicting the number of casualties and damage as outcomes of crashes. Our regression models got fairly low test MSEs, which means our predictions were not far from the actual values, and means our models have some predictive power to them. We then take the important features for each model to create a final summary.

Model	Test MSE	Model	Test MSE
Fatal Count	0.014	Object Damage	0.013
Serious Injury	0.102	Vehicle Damage	0.414
Minor Injury	0.417	Property Damage	0.257
Pedestrian Casualties	0.046	Stray Animal	0.008

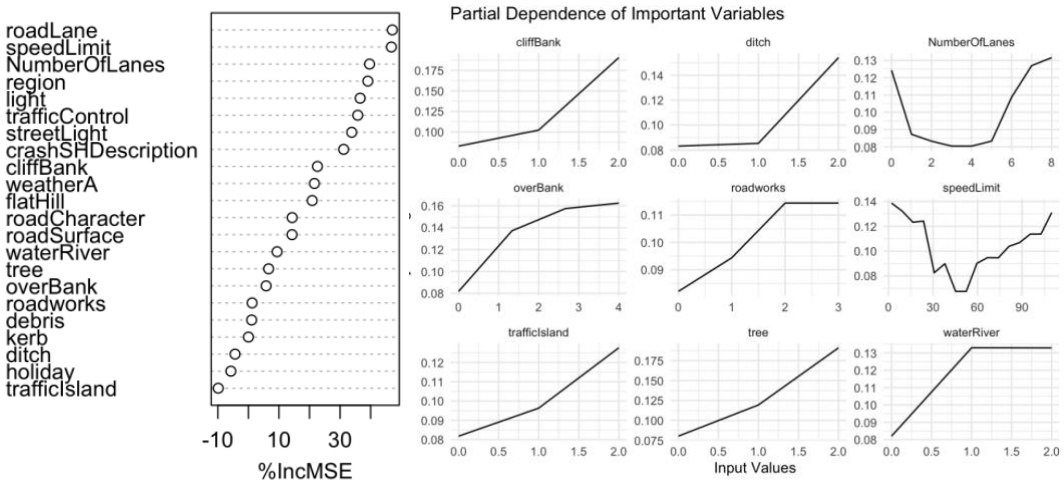
(Table 7: Summary of MSEs for each regression model)

Fatal injuries are associated with the road conditions such as the lane configuration, number of lanes, lighting, and whether the street is a highway or not. In general, more obstructions, lanes, and higher speeds are associated with higher fatal injury counts.



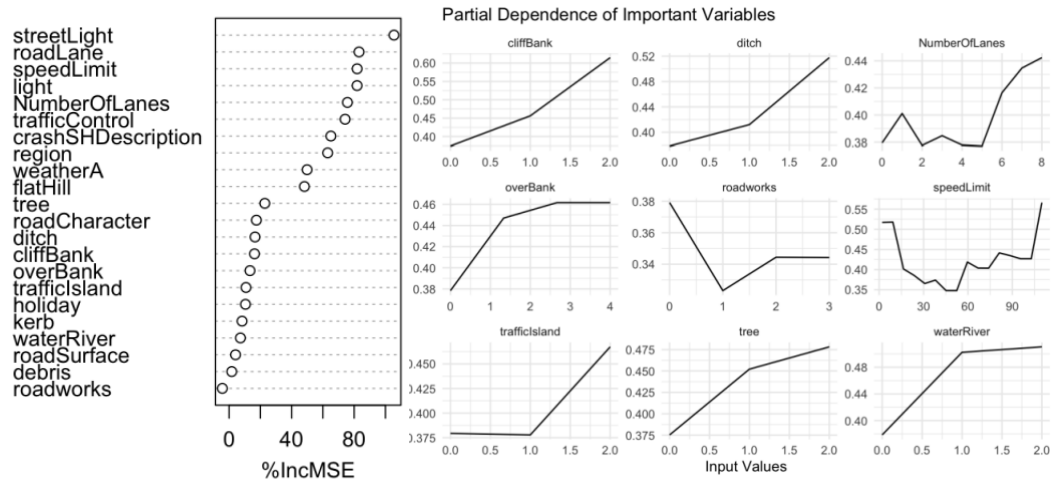
(Figure 8: Variable importance and partial dependency for Fatal Injury Count)

The features for serious injury count are associated with the road conditions such as the lane configuration, speed limit, and number of lanes. In general, more obstructions, lanes, and higher speeds are associated with higher serious injury counts.



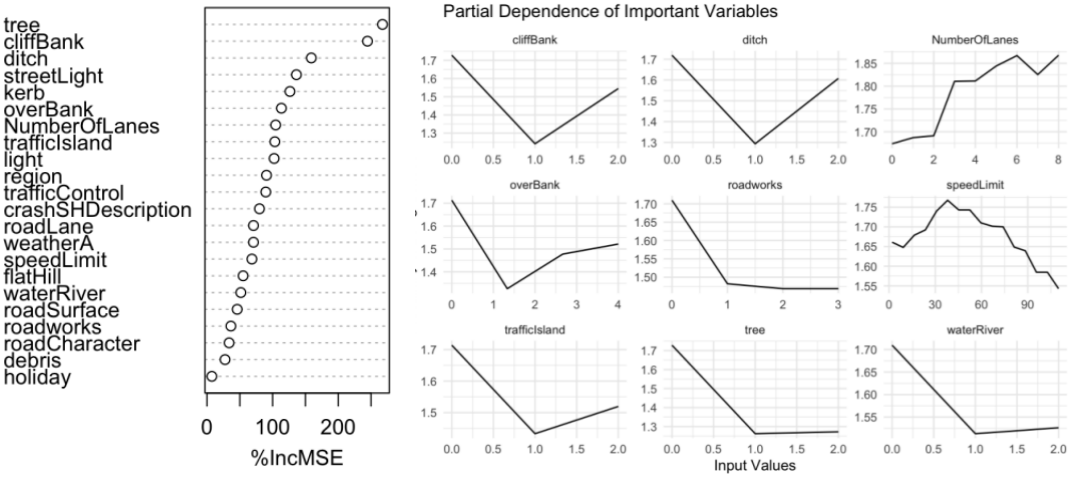
(Figure 9: Variable importance and partial dependency for Serious Injury Count)

Minor injury counts are also associated with road conditions as important variables. Compared to serious injury counts, road works are less characteristic of being able to predict minor crashes.



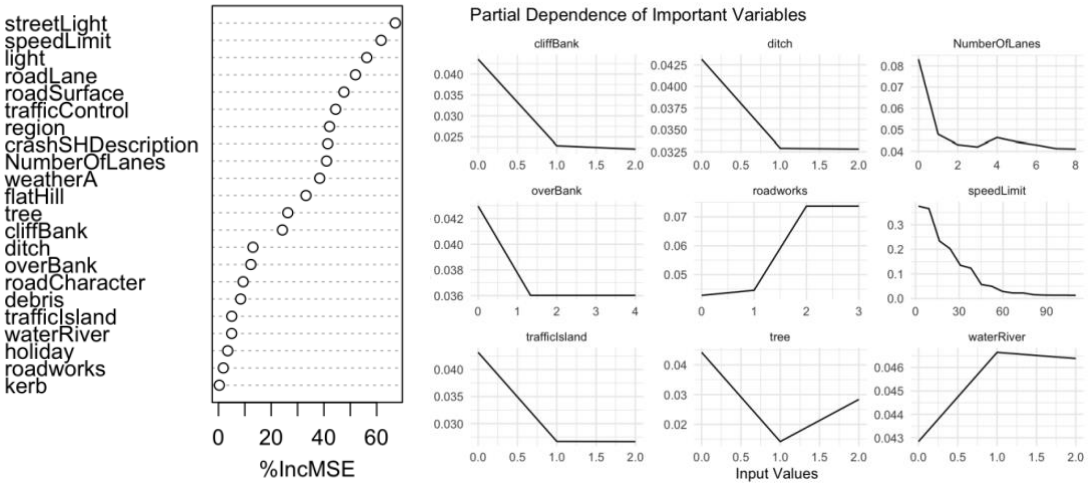
(Figure 10: Variable importance and partial dependency for Minor Injury Count)

Vehicle Damage had mostly environmental factors as important features. The absence of certain objects like cliffs, ditches, roadworks, and trees increased the number of vehicle damage. However, the number of lanes and speed limit exhibited a different trend, as the casualties were higher at lower speed limits, and higher number of lanes.



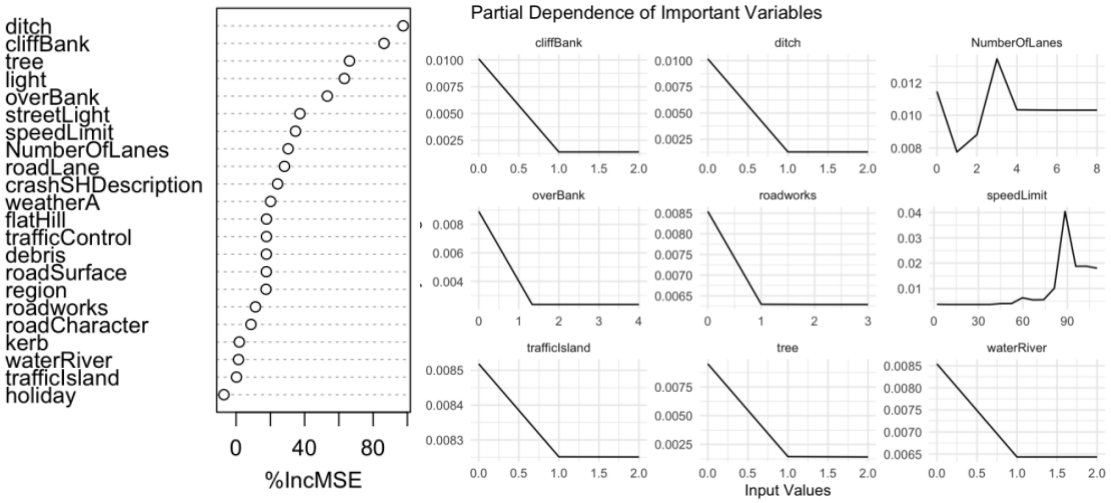
(Figure 11: Variable importance and partial dependency for Vehicle Damage)

For pedestrian casualties, road conditions were more prominent, such as street lighting, speed limits, light, lane configuration, and road surface. A higher number of pedestrian casualties are associated with more trees and bodies of water present in the crash. More casualties are associated with less lanes and lower speed limits. These may point to urban and suburban areas being important areas for pedestrian casualties.



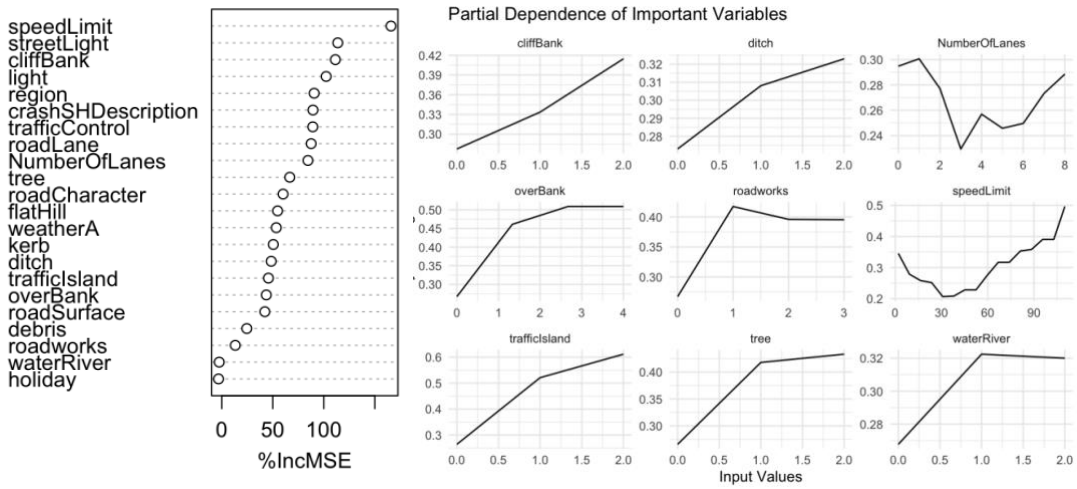
(Figure 12: Variable importance and partial dependency for Pedestrian Casualties)

The model for stray animals associates environmental factors like ditches, cliffs, trees, lighting, and banks as important features. Given the plot, it looks like areas with more lanes and higher speed limits are associated with more stray animals being involved in crashes.



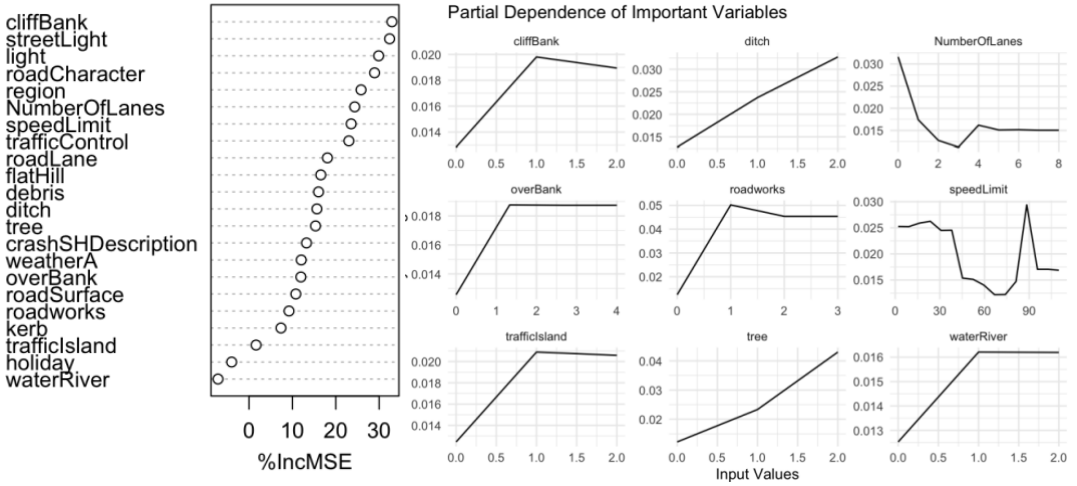
(Figure 13: Variable importance and partial dependency for Stray Animal Casualties)

When it comes to property damage, road conditions and environmental factors are the most important features, such as speed limit, street lighting, cliffs, lighting, region, and whether the road was a street or highway. There is an association for road obstructions to have higher property damage. Higher speed limits, and a general uptrend in the number of lanes are also associated with more property damage.



(Figure 14: Variable importance and partial dependency for Property Damage)

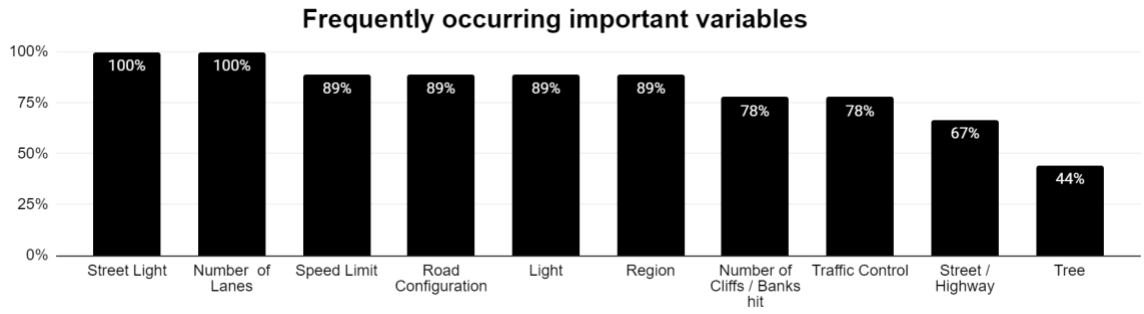
Object damage is associated with environmental factors and road conditions, with an association to more road obstructions contributing to more damage.



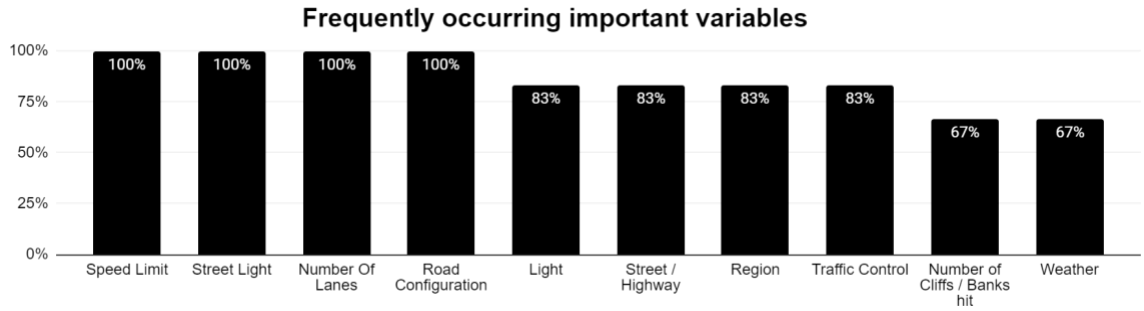
(Figure 15: Variable importance and partial dependency for Object Damage)

IX. Summary of findings

From our random forest models, we found recurring variables of high importance across different crash outcomes. Street conditions, such as Street Light, Speed Limit, Number of Lanes, and Road Configuration are among the most frequently highly ranked variables in terms of importance. They are followed by environmental factors such as Region, Light, Cliffs/Banks.



If we were to look at the most frequently highly ranked variables for severe crashes, injuries, and sentient casualties, we see most of the same variables.



A more nuanced view would be to look at each variable and see what they can predict. Authorities can prioritise the inspection of features that address more target variables, such as Street Light, Region, Lanes, and Speed Limit.

Street Lighting

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
5. Stray Animals
6. Vehicle Damage
7. Property Damage
8. Object Damage

 # Trees

5. Stray Animals
6. Vehicle Damage
7. Property Damage
8. Object Damage
9. Crash Severity

Region

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
6. Vehicle Damage
7. Property Damage
8. Object Damage

 # Cliffs / Banks

1. Fatal Count
5. Stray Animals
6. Vehicle Damage
7. Property Damage
8. Object Damage
9. Crash Severity

 # Lanes

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
5. Stray Animals
6. Vehicle Damage
7. Property Damage

Road Configuration

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
5. Stray Animals
7. Property Damage

Speed Limit

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
5. Stray Animals
7. Property Damage
8. Object Damage

Traffic Control

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
7. Property Damage
8. Object Damage

Street/Highway

1. Fatal Count
2. Serious Injury Count
3. Minor Injury Count
4. Pedestrian Casualties
7. Property Damage

X. Recommendations

Addressing Declining Car Crash Trends

While New Zealand has witnessed a progressive decline in car accidents from 2021 to early 2023, this trend should be actively sustained and enhanced. To achieve this, it is strongly recommended that the action plans outlined in the Road to Zero program are expedited and rigorously executed. This program has already laid a solid foundation for reducing road accidents and improving safety. By fast-tracking its implementation, New Zealand can expect to see even better outcomes in terms of minimising road accidents and their associated impacts.

Four primary factors influencing crash severities are street lighting, the number of lanes, speed limits, and road configuration. Addressing these factors is crucial to enhance road safety. To reduce the severity of accidents, it is recommended to invest in improved street lighting, particularly on roads with inadequate lighting. Moreover, attention should be given to road configuration, ensuring it accommodates safe and efficient traffic flow. Roads with two or more lanes should have extra precautions and safety measures in place to ensure safe lane changes and effective congestion management. Areas with high-speed limits should undergo speed reduction considerations and heightened speed enforcement to minimise the impact of accidents. Implementing these recommendations will not only make the roads safer but also reduce the severity of accidents, making New Zealand's roads more secure and less prone to fatal crashes.

Tailoring Safety Interventions to Regional Specifics

Recognizing that the number of crashes and their severities vary significantly from one region to another, it is imperative to tailor safety interventions to the specific characteristics of each area. This includes accounting for local factors such as road conditions, speed limits, and population density. For instance, regions like Waikato and Auckland, with high population density, should have focused safety initiatives due to their higher fatality rates. Additionally, regions with speed limits of 100 km/h and above, associated with severe crash rates, should undergo targeted safety improvements. In contrast, areas like Northland, which exhibit the highest likelihood of severe crashes, should receive specialised attention, with improved

emergency response strategies to minimise the impacts of accidents. Such localised approaches are essential for making road safety more effective and efficient.

Addressing Holiday Period Accidents

To address the issue of increased accidents during holiday periods, several key recommendations are proposed. Firstly, there should be a collaborative effort with law enforcement agencies to enhance police presence on the roads during holiday seasons. This can be achieved through visible police patrols and the implementation of sobriety checkpoints, which can act as deterrents to reckless and impaired driving.

Secondly, launching comprehensive public awareness campaigns is crucial. These campaigns should emphasise safe driving practices during holidays, with a specific focus on the dangers of drunk driving, distracted driving, and speeding. Utilising various media channels, including social media, billboards, and radio, will enable these messages to reach a wide audience effectively.

Additionally, traffic flow management strategies, such as the use of variable message signs and traffic signal coordination, should be implemented to optimise traffic movement and reduce congestion during peak holiday travel times. Ensuring the availability of safe and well-lit rest areas along major travel routes is another important measure to combat driver fatigue, a significant factor in accidents.

Lastly, enforcing strict penalties for driving under the influence (DUI) offences during the holidays is imperative. Conducting sobriety checkpoints and increasing penalties for DUI convictions will act as strong deterrents against impaired driving.

Addressing Hourly Period Accidents

Addressing the elevated number of accidents occurring between 15:00 PM and 18:00 PM requires a multifaceted approach involving various stakeholders, including government authorities, transportation agencies, schools, and the public.

This comprehensive strategy encompasses initiatives such as launching public awareness campaigns to educate drivers about the unique risks during this time frame, encouraging employers to provide workplace flexibility to reduce rush-hour congestion, investing in infrastructure improvements like widening roads and enhancing signage, implementing advanced traffic management systems for real-time optimization, working with schools to improve student transportation safety, increasing law enforcement presence to enforce traffic laws, exploring technological solutions like intelligent traffic signals, promoting seasonal safety measures, conducting thorough data analysis to pinpoint trends and areas of concern, and engaging the community in reporting and addressing traffic-related issues.

The ultimate aim is to establish a collaborative effort that effectively reduces accidents during this critical time period.

XI. Limitations

The project can benefit from more information on the actual social cost of the casualties and damage. This will inform which models to prioritise and optimise. The study currently prioritises the factors for crashes involving sentient beings, while the prioritisation of non-sentient beings were based on the degree of separation from probable involvement of life (ex. Vehicles may be manned or unmanned, so they were placed higher).

Another potential improvement may stem from being able to calculate actual cost of uncaught serious crashes compared to false alarms. This will better inform the process of balancing precision and recall. Future studies may also opt to replace precision with f1 score as the secondary metric to observe.

The study can also benefit from hyperparameter tuning to improve model performance. While the project made adjustments through resampling and class weight applications, experimenting with hyperparameters can make predictions more accurate.

Finally, future researchers can make use of the important variables in conjunction with fields such as geospatial to hone in on areas that fit the bill when it comes to features that are associated with serious crashes, high casualties, and high damage.

Reference

Drivers Licence Holder. NZ Transport Agency Open Data. (n.d.). *https://opendata-nzta.opendata.arcgis.com/documents/driver-licence-holders/about*

2018 Census place summaries. Stat NZ.(n.d.). *https://www.stats.govt.nz/tools/2018-census-place-summaries/ . Retrieved 7 October, 2023*

Road-to-Zero-strategy_final. NZ Transport Agency. December 2019.(n.d.). *https://www.transport.govt.nz/assets/Uploads/Report/Road-to-Zero-strategy_final.pdf*

Speed Limit. NZ Transport Agency Road Code.(n.d.). *https://www.nzta.govt.nz/roadcode/heavy-vehicle-road-code/road-code/about-limits/speed-limits/*

Demand and Activity. . NZ police. (n.d.). *https://www.police.govt.nz/about-us/statistics-and-publications/data-and-statistics/demand-and-activity*

Cas Data field descriptions. Waka Kotahi open data. (n.d.). https://opendata-nzta.opendata.arcgis.com/pages/cas-data-field-descriptions

Guide to treatment of crash locations - definitions - Waka Kotahi NZ (n.d.). https://www.nzta.govt.nz/assets/resources/guide-to-treatment-of-crash-location/docs/definitions.pdf

GWI. (2023). *Understanding index scores – GWI*. GWI.

https://gwihelpcenter.zendesk.com/hc/en-us/articles/4404468207762-Understanding-index-scores

Haynes, R., Lake, I. R., Kingham, S., Sabel, C. E., Pearce, J., & Barnett, R. (2008). The influence of road curvature on fatal crashes in New Zealand. *Accident Analysis & Prevention*, *40*(3), 843–850. https://doi.org/10.1016/j.aap.2007.09.013

Kerner, B. S. (2018). *Breakdown in traffic networks fundamentals of transportation science*. Springer Berlin.

Lewis-Evans, B. (2010). Crash involvement during the different phases of the New Zealand graduated Driver Licensing System (GDLS). *Journal of Safety Research*, *41*(4), 359–365. https://doi.org/10.1016/j.jsr.2010.03.006

Poulsen, H., Moar, R., & Troncoso, C. (2012). The incidence of alcohol and other drugs in drivers killed in New Zealand Road crashes 2004–2009. *Forensic Science International*, *223*(1–3), 364–370. https://doi.org/10.1016/j.forsciint.2012.10.026

Tay, R. (2001). Fatal crashes involving young male drivers: A continuous time Poisson change-point analysis. *Australian and New Zealand Journal of Public Health*, *25*(1), 21–23. https://doi.org/10.1111/j.1467-842x.2001.tb00544.x

Walton, D., Jenkins, D., Thoreau, R., Kingham, S., & Keall, M. (2020). Why is the rate of annual road fatalities increasing? A unit record analysis of New Zealand data (2010–2017). Journal of Safety Research, 72, 67–74. https://doi.org/10.1016/j.jsr.2019.11.003

Varghese, D. (2018). Comparative Study on Classic Machine learning Algorithms. *Towards Data Science.* https://towardsdatascience.com/comparative-study-on-classic-machine-learning algorithms-24f9ff6ab222