Applied Data Science Capstone Final Report

Introduction

Arguably the most prominent issue at the intersection of transport and public health is that of poor air quality in municipalities with high levels of motorized vehicle usage. While this is indeed an important component of the larger discussion surrounding climate change, another negative consequence of ‘car dominance’ in the transportation network is the alarming levels of injuries and fatalities due to vehicle-vehicle and vehicle-person collisions.

In response to this growing crisis, cities and countries across the globe have pledged themselves to shifting the prevailing norm around traffic safety from the *inevitability* of deaths to their *preventability*.[[1]](#footnote-1) This mission, which originates from Sweden during the 1990s, is known as Vision Zero. Broadly speaking, it promotes multidisciplinary discussions on a variety of topics, including the re-design of roadways and public policy, to improve overall quality of life.[[2]](#footnote-2)

One of the major roadblocks, however, is limited financial and knowledge capital to execute abstract ideas into concrete actions. Thus, it is important for governments to collect and provide data so that researchers can analyze it using various data science techniques and inform decision-makers about the ‘best’ pathways towards achieving Vision Zero outcomes.

There are two core questions to this analysis. First, what is the optimal set of predictors for a classification model of crash severity and how might these results translate into meaningful policy recommendations? Second, using this final predictor set, is it possible to identify meaningful geographic clusters to prioritize funding for the most at-risk communities? [Note: the specific modeling approaches will be mentioned here in the full report]

This report features a mini case study on New Zealand, which was selected for three reasons: (a) it is where the greatest film trilogy of all-time, *The Lord of the Rings*, was filmed; (b) Jacinda Ardern is an amazing prime minister and has been nominated for the 2020 Nobel Peace Prize; and (c) the country’s fjords are some of the most stunning natural wonders of the world. Although there is reason to doubt the complete generalizability of findings, especially in the absence of uniform data collection procedures, the research approach sets an example for others to follow.

Data Source

The Waka Kotahi New Zealand Transport Agency provided the Crash Analysis System Data as part of the countrywide open source platform for government-funded projects. It consists of 730,280 observations of 70 variables beginning from 2000 and through August 2020. The dependent variable of interest, *crash severity*, has four categories, with counts summarized in Table 1 below. This is determined by the worst injury sustained in the crash at time of entry. The data is clearly imbalanced, with about 71% of cases resulting in no injuries.

**Table 1**. Summary of crash level severity in New Zealand.

|  |  |
| --- | --- |
| **Crash Severity** | **Count** |
| Non-injury | 518,194 |
| Minor | 164,847 |
| Serious | 40,601 |
| Fatal | 6,638 |

Furthermore, the following variables are conjectured to be potentially significant in the final classification model. This is adapted directly from the open data website.[[3]](#footnote-3)

**Table 2**. Variables that will be utilized in the machine learning model development process.

|  |  |
| --- | --- |
| Attribute Name | Description |
| bicycle | Derived variable to indicate how many bicycles were involved in the crash. |
| bridge | Derived variable to indicate how many times a bridge, tunnel, the abutments, handrails were struck in the crash. |
| bus | Derived variable to indicate how many buses were involved in the crash. |
| carStationWagon | Derived variable to indicate how many cars or station wagons were involved in the crash. |
| cliffBank | Derived variable to indicate how many times a 'cliff' or 'bank' was struck in the crash. |
| crashYear | The year in which a crash occurred, if known. |
| debris | Derived variable to indicate how many times debris, boulders or items dropped or thrown from a vehicle(s) were struck in the crash |
| ditch | Derived variable to indicate how many times a 'ditch' or 'drainage channel' was struck in a crash. |
| fence | Derived variable to indicate how many times a 'fence' was struck in the crash. This includes letterbox(es), hoardings, private roadside furniture, hedges, sight rails, etc. |
| flatHill | Whether the road is flat or sloped. Possible values include 'Flat or 'Hill'. |
| guardRail | Derived variable to indicate how many times a guard or guard rail was struck in the crash. |
| holiday | Indicates where a crash occurred during a 'Christmas/New Year', 'Easter', 'Queens Birthday' or 'Labour Weekend' holiday period, otherwise 'None'. |
| houseOrBuilding | Derived variable to indicate how many times a houses, garages, sheds or other buildings were struck in the crash. |
| kerb | Derived variable to indicate how many times a kerb was struck in the crash, that contributed directly to the crash. |
| light | The light at the time and place of the crash. Possible values: 'Bright Sun', 'Overcast', 'Twilight, 'Dark' or ' Unknown'. |
| moped | Derived variable to indicate how many mopeds were involved in the crash. |
| motorcycle | Derived variable to indicate how many motorcycles were involved in the crash. |
| NumberOfLanes | The number of lanes on the crash road. |
| overBank | Derived variable to indicate how many times an embankment was struck or driven over during a crash. This variable includes other vertical drops driven over during a crash. |
| parkedVehicle | Derived variable to indicate how many times a parked or unattended vehicle was struck in the crash. This variable can include trailers. |
| phoneBoxEtc | Derived variable to indicate how many times a telephone kiosk traffic signal controllers, bus shelters or other public furniture was struck in the crash. |
| pedestrian | Derived variable to indicate how many pedestrians were involved in the crash. This includes pedestrians on skateboards, scooters, and wheelchairs. |
| postOrPole | Derived variable to indicate how many times a post or pole was struck in the crash. |
| region | Identifies the local government (LG) region. |
| roadLane | The lane configuration of the road. Possible values: '1' (one way), '2' (two way), 'M' (for where a median exists), 'O' (for off-road lane configurations), ' ' (for unknown configurations). |
| roadMarkings | The road markings at the crash site. Possible values: 'Ped Crossing', 'Raised Island', 'Painted Island', 'No Passing Lanes', 'Centre Line', 'No Marks' or ' Unknown'. |
| roadSurface | The road surface description applying at the crash site. Possible values: 'Sealed' or 'Unsealed'. |
| roadworks | Derived variable to indicate how many times an object associated with 'roadworks' (including signs, cones, drums, barriers, but not roadwork vehicles) was struck during the crash |
| schoolBus | Derived variable to indicate how many school buses were involved in the crash. |
| slipOrFlood | Derived variable to indicate how many times landslips, washouts, or floods (excluding rivers) were objects struck in the crash |
| speedLimit | The speed limit in force at the crash site at the time of the crash. May be a number, or 'LSZ' for a limited speed zone. |
| strayAnimal | Derived variable to indicate how many times a stray animal(s) was struck in the crash. |
| streetLight | The street lighting at the time of the crash. Possible values 'On', 'Off', 'None' or ' Unknown'. |
| suv | Derived variable to indicate how many SUVs were involved in the crash. |
| trafficControl | The traffic control (ctrl) signals at the crash site. Possible values are 'Traffic Signals', 'Stop Sign', 'Give Way Sign', 'Pointsman', 'School Patrol', 'Nil' or ' N/A'. |
| trafficIsland | Derived variable to indicate how many times a traffic island or medians (excluding barriers) was struck in the crash. |
| trafficSign | Derived variable to indicate how many times 'traffic signage' (including traffic signals, their poles, bollards, or roadside delineators) was struck in the crash. |
| train | Derived variable to indicate how many times a train, rolling stock or jiggers was struck in the crash, whether stationary or moving. |
| tree | Derived variable to indicate how many times trees or other flora were struck during the crash. |
| truck | Derived variable to indicate how many trucks were involved in the crash. |
| urban | A derived variable using speed limit. Possible values are 'Urban' (urban, speedLimit < 80) or 'Open Road' (open road, speedLimit >=80 or 'LSZ'). |
| vanOrUtility | Derived variable to indicate how many vans or utility vehicles were involved in the crash. |
| waterRiver | Derived variable to indicate how many times a body of water (including rivers, streams, lakes, the sea, tidal flats, canals, or swamps) was struck in the crash. |
| weatherA | Indicates weather at the crash time/place. Values that are possible are 'Fine', 'Mist', 'Light Rain', 'Heavy Rain', 'Snow', 'Unknown'. |
| weatherB | The weather at the crash time/place. Values 'Frost', 'Strong Wind' or 'Unknown'. |

Of course, future exploratory data analysis tasks could indicate that one or more of the above variables should either be (a) transformed/recorded for or (b) excluded from the classification task. Meanwhile, the *region* variable will be central to the secondary clustering task that will help pinpoint ‘priority areas’ for traffic safety interventions.

1. https://visionzeronetwork.org/about/what-is-vision-zero/ [↑](#footnote-ref-1)
2. https://toolkits.ite.org/visionzero/ [↑](#footnote-ref-2)
3. https://opendata-nzta.opendata.arcgis.com/pages/cas-data-field-descriptions [↑](#footnote-ref-3)