Car Accident Severity Analysis

Sophia Chapman

Coursera Capstone Project

Why Predict Car Accident Severity?

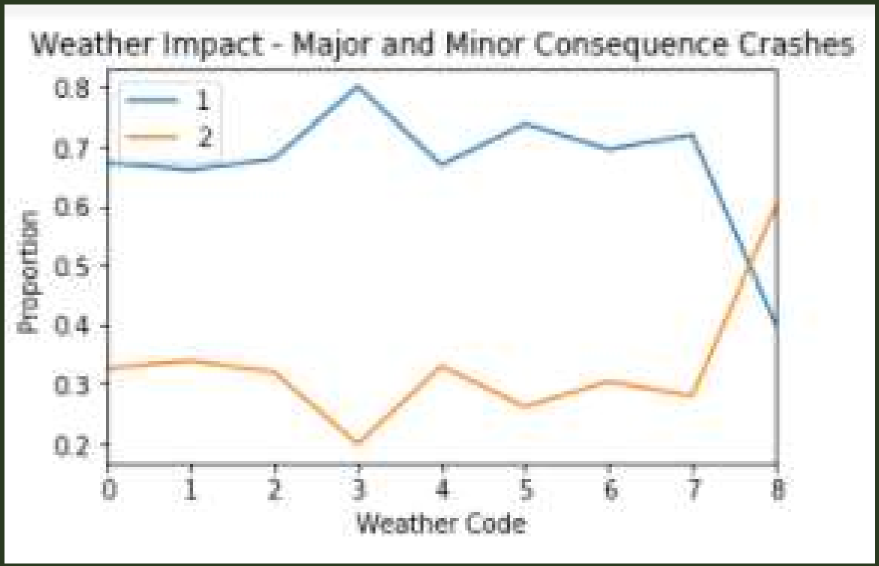
- Australia loses over 1200 citizens every year to road deaths
- Weather, road conditions and lighting conditions all contribute to accident severity
- If an MLA can be produced for weather, road conditions and lighting conditions to predict car accident severity, it can be used by:
  - Mapping software to divert traffic away from locations where severe accidents are more likely
  - Emergency services responders to ensure emergency responders are situated closer to locations where severe accidents are more likely
  - Policy makers to direct policy focus
  - Road users to identify when it may be advisable to avoid non-essential journeys

Data Acquisition and Cleaning

- Data has been obtained from Seattle Department of Transport from 2004 - 2018
  - Australian data is not of the same detail or quality – but successful MLA may be used as impotence to obtain Australian data
- Elements with missing weather, lighting conditions or road conditions were removed
  - 167,427 elements total with no duplicate elements
- Lighting conditions, weather and road conditions were numerically encoded
  - Potential confounding factors like speeding, driver distraction and intoxication were encoded as 1 (for Yes) or 0 (for No) to allow sensitivity analysis for MLA
  - Some data sample sizes are limited (e.g. partly cloudy weather has 5 elements)
- Data is normalised to visualise proportion of severe (Severity 2) vs less severe (Severity 1) crashes

Weather Impact

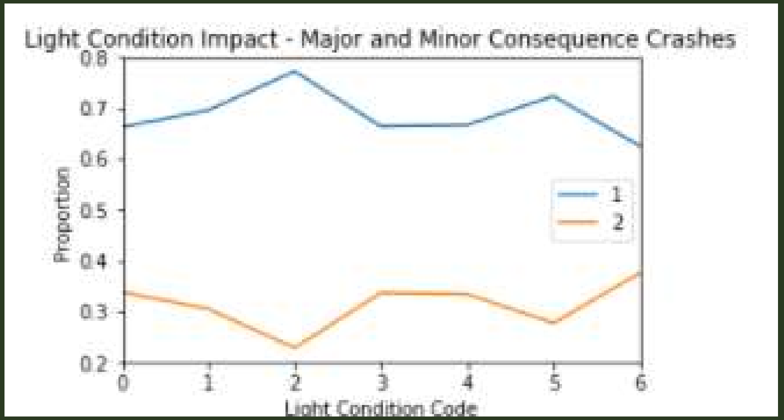
- Clear, Raining and Overcast conditions have similar proportions of severe accidents
  - Raining and Overcast have slightly higher likelihood of severe accident, but difference is small
- Snowing conditions have lowest proportion of severe accident
  - Sample size is statistically significant
  - This may be counterintuitive to expectations for many stakeholders
- "Partly Cloudy" data (5 elements – group 8) is not statistically significant



Weather Condition	Encoded Value
Clear	0
Raining	1
Overcast	2
Snowing	3
Fog/Smog/Smoke	4
Sleet/Hail/Freezing Rain	5
Blowing Sand/Dirt	6
Severe Crosswind	7
Partly Cloudy	8

Lighting Condition Impact

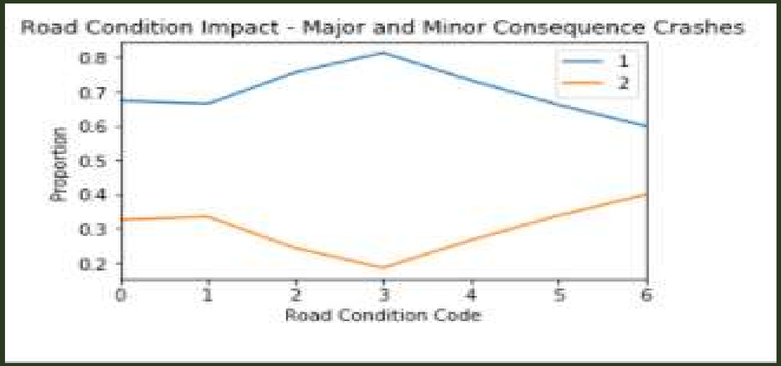
- Other than “Dark – Unknown Lighting”, “Daylight” had the highest proportion of severe crashes
  - “Dark – Unknown lighting” is a relatively small sample (<50 elements out of >167,000) and is not statistically significant
- “Dark – No street lights” had the lowest proportion of severe crashes, followed by “Dark – Street Lights off”
  - Again, this may be counterintuitive for many stakeholders – further investigation is required as to why this is



Light Condition	Encoded Value
Daylight	0
Dark - Street Lights On	1
Dark - No Street Lights	2
Dusk	3
Dawn	4
Dark - Street Lights Off	5
Dark - Unknown Lighting	6

Road Condition Impact

- Wet conditions had the highest proportion of severe crashes, followed closely by Dry conditions
- Ice and Snow/Slush had the lowest proportion of severe crashes
- Oil, Standing water, and Sand/Mud/Dirt did not have enough data points to be considered statistically significant



Road Condition	Encoded Value
Dry	0
Wet	1
Ice	2
Snow/Slush	3
Standing Water	4
Sand/Mud/Dirt	5
Oil	6

Machine Learning Algorithms (MLAs)

- 4 MLAs utilized for this assessment
  - K Nearest Neighbour (KNN)
  - Support Vector Analysis (SVM)
  - Decision Tree
  - Logistic Regression
- Data split into training subset (20%) and testing subset (80%)
- MLAs developed for complete cleaned dataset and subset only of data without speeding, indicators of inattention and driver intoxication
  - Speeding, driver inattention and driver intoxication increase the likelihood of a crash being severe in any conditions – analysis was needed to establish whether this impacted MLA accuracy
- Accuracy assessed using Jaccard Similarity Score, F1 Score and (for Logistic Regression) Log Loss
- False Positive (predicts Severity 2 and actual was severity 1) and False Negative (predicts Severity 1 and actual was severity 2) rates assessed

Machine Learning Algorithms - Results

- Minimal difference in accuracy between MLAs developed for data sets with and without speeding, intoxication and driver distraction elements
- Decision Tree consistently had the best accuracy – but SVM had the lowest false negative rate
  - Different stakeholders may have different needs
  - Some stakeholders may need to prioritise lower false negative rate over model accuracy

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.668687	0.552274	NA
Decision Tree	0.672599	0.544307	NA
SVM	0.608261	0.577430	NA
Logistic Regression	0.672894	0.541344	0.631161

MLA Accuracy Analysis – Including Speeding, Intoxication and Distraction Elements

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.680530	0.561346	NA
Decision Tree	0.682814	0.554932	NA
SVM	0.598358	0.571265	NA
Logistic Regression	0.682505	0.553713	0.62404

MLA Accuracy Analysis – Excluding Speeding, Intoxication and Distraction Elements

Conclusions and Next Steps

- Although MLA has been developed to determine whether a crash is likely to be severe, its accuracy is limited (Jaccard Similarity Score of 0.67) and is based on the Seattle context
- Further data is required to be gathered in the Australian context – MLA may be refined and applied by Australian stakeholders
  - Data needs to be gathered until sample sizes for each condition are statistically significant
- Qualitative assessment required to understand why some (typically viewed as more dangerous) conditions result in a lower severe crash proportion to ensure policy application does not create complacency leading to a severe crash increase