Car Accident Severity Analysis

Sophia Chapman
October 2020

Introduction

Every year, more than 1200 Australians lose their lives on the roads. This not only creates enormous pain and suffering for the families who lose loved ones to these accidents, but echoes of these losses account for much grief and suffering in the wider community. Furthermore, these deaths have a significant impact on both the local and national economy, due to not only the lost productivity of the lives cut short by fatal motor vehicle accidents, but also due to injuries, temporary or permanent disabilities and the emotional trauma experienced by both the individuals involved in the accidents and their families and loved ones.

To date, in Australia, there has been a major policy focus on 3 areas:

1. Speeding prevention and speed limit compliance;
2. Drink driving prevention; and
3. Enhanced restrictions on younger drivers (including extended duration of Probationary periods, lower alcohol restrictions, limits on passengers, etc).

However, other factors such as the weather, lighting conditions and road conditions have not received the same amount of attention.

By utilizing data science techniques to assess a robust dataset that includes a range of factors, policy makers can understand what factors correlate with a high accident rate, and a high accident rate resulting in death, and determine where to direct policy focus. This data may also be used to integrate recommendations into map and/or weather software, to direct drivers on routes that avoid areas with known conditions that may lead to accidents if at all possible (e.g. poor lighting, weather, road conditions, etc). It may also allow better utilization of emergency services resources and minimization of emergency services response times by allowing emergency services planners to better predict where their services will be needed in response to car accidents.

Key stakeholders for this project include:

- Road safety policy makers;
- Mapping and direction planning software developers;
- Emergency services response planners; and
- Road users.

Currently, the data gathered in Australia does not consider factors like the weather, the road conditions, or lighting conditions. As such, analogous data from the US will be utilized to help direct Australian data gathering efforts.

Data

The Seattle SDOT Traffic Management Division, Traffic Records Group has collated data from 2004 to present, which will be used as a basis for initial data modelling in the absence of Australian data. This data is stored here.

It should be noted that this data has been pre-treated to amalgamate fatality data with injury data, and has scrubbed data where the outcome was unknown. It therefore categorises each element as either resulting in property damage (1) or physical injury (2) (unlike the statement in the meta data file).

Approximately 5000 elements (out of nearly 2 million) are missing collision type information, accounting for ~2.5% of the data. However, this is not relevant to the question at hand (how weather, lighting conditions and intersection types impact the likelihood of a collision and the likelihood of a collision resulting in an injury). Similarly, the data on driver intoxication is inconsistent, utilizing both a 1/0 and Y/N format; and some location data is also missing; but again, these are not one of the independent or dependent variables in this analysis and will not be utilized in the project.

Feature Selection

The accident severity code (SEVERITYCODE), road conditions (ROADCOND), weather conditions (WEATHER), and lighting conditions (LIGHTCOND) will be the main focus of this analysis. Therefore, elements where these codes are not defined (ie "other"), unknown or blank has been scrubbed for that portion of the analysis. Removing this data resulted in the loss of 24716 out of 194673 elements, or approximately 12.7% of the data. This is viewed as preferable to the distortion of results due to missing data.

Further analysis will be performed on data where crashes with obvious causes of driver error (ie speeding (SPEEDING), intoxication (UNDERINFL) and inattention (INATTENTIONIND)) are eliminated; given that one of the intents of this study is to assess the potential for a predictive algorithm for crashes, these factors will not be known when predicting when a crash is likely, and so crashes where these are a factor should not contribute to the dataset for any predictive algorithm.

It should be noted that this data is not without its concerns. As the SPEEDING data is indicated as a binary (Y/N), rather than providing (where available) the magnitude of the speeding, this may limit the accuracy of the MLA, since the risk profile associated with driving 1-2 km/h over the speed limit is significantly different to driving 20-30km/h over the limit. Furthermore, some of the parameters rely on subjective information. Whilst weather reports can be utilized to establish the WEATHER parameter, the road condition and lighting condition will be subjective to the person who made the report. For example, there is no clear distinction where dusk ends and night begins; what seems like a wet road in winter to one person may seem like an icy road to another; both a road with a small puddle and a flooded road may be characterised as having standing water, etc.

It should also be noted that this dataset is not a balanced dataset. There is no indication on the average distribution of weather (per driver hour) which would allow the impact of weather on the overall likelihood of a crash to be ascertained, and the sample size for each weather condition, lighting condition and road condition varies between tens of thousands (for clear weather) and less than 5 (for smoke/smog).

Despite these concerns with the variable sample sizes, and the limits around the subjectivity and quality of the data, it is still worthwhile pursuing an investigation to determine the potential accuracy of an MLA, as even with the limitations around the data quality, the results will indicate whether it is worth pursuing additional data utilizing less subjective measures for the same parameters.

Methodology

Jupyter Notebooks using Python have been utilized for this data analysis.

Data Cleaning and Processing

After importing the data from the Seattle SDOT, the size of the matrix was checked using the SHAPE command. The elements where weather, road conditions or lighting were 'Unknown', 'Other', or missing were expunged, as were the columns not relevant to the central question used in the analysis to speed up data processing time (specifically geographical coordinates X and Y, the report numbers, and information relating to parked cars and crosswalks). The matrix shape was checked as a confirmation that the required elements and columns had been dropped.

The data was then converted from text values to numerical values to allow for algorithmic analysis using the following key:

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

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

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

In the columns for speeding, inattention and intoxication, Y and N inputs were replaced with 1 and 0 respectively to allow for numerical processing.

A smaller dataframe with only relevant variables and unique keys was produced from this dataframe.

This dataframe was further broken down into separate dataframes for the less severe crashes (severity 1) and more severe crashes (severity 2), with an additional subset dataframe made up of only crashes that did not involve speeding, intoxication or inattention.

Data Visualisation and Analysis

In order to facilitate analysis of the impact of the parameters in question, the cleaned data was visualised to get a better understanding of the orders of magnitude. An initial review of the main cleaned dataframe shows the dataframe was undertaken using histogram plots to visualise the size of each part of the dataset and the proportions of the data, for both the cleaned dataframe (including data from accidents involving driver inattention, intoxication and speeding) and the reduced dataframe (excluding data from accidents involving driver inattention, intoxication and speeding).

Grouping by severity code for the lighting conditions, the weather conditions and the road conditions and trending the relative proportions of the more severe and less severe accidents (both with and without data from accidents involving driver inattention, intoxication and speeding) to provide insight into whether the data supports the idea that certain weather, road or lighting conditions affect the likelihood of a severe crash.

Machine Learning Algorithm

The datasets were divided randomly into training and testing data, with a train/test split of 20% selected.

Multiple Machine Learning Algorithms (MLAs) were developed and tested for the datasets. K Nearest Neighbour analysis (KNN analysis), Decision Tree analysis, Support Vector Machine analysis (SVM) and Logistical regression analysis were created for the dataset using a training subset. Their accuracy was assessed using a Jaccard score, F1 score and (for logistical regression) log loss score. The false positive (severity 2 predicted and severity 1 actual) and false negative (severity 1 predicted and severity 2 actual) rate was also checked, as for some purposes, a less accurate model with a lower false negative rate may be more useful.

The MLA models utilized the logic developed in SKLearn, Pylab and SciPy.

Results

Initial Data Analysis

An initial check of the data showed all elements were unique, and that there were more than twice the number of lower severity (severity 1) crashes as more severe crashes (severity 2) as shown below in Figure 1.

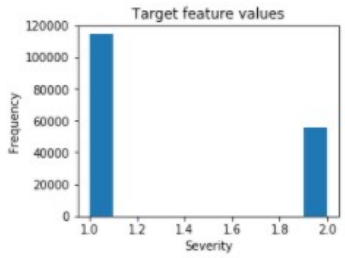


Figure 1 - Total Severity Data

This data shows that the sample size of the severity 1 (over 114,000) and severity 2 (over 55,000) crashes is sufficient to warrant further analysis.

Plotting a histogram showing the breakdown of the weather conditions, lighting conditions and road conditions show that the sample size of some of the identified conditions may not be large enough to draw conclusions, as shown in Figures 2, 3 and 4.

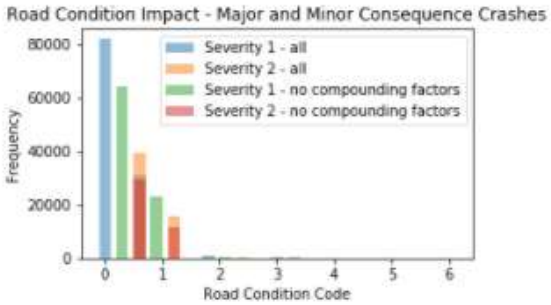


Figure 2 - Road Condition Breakdown

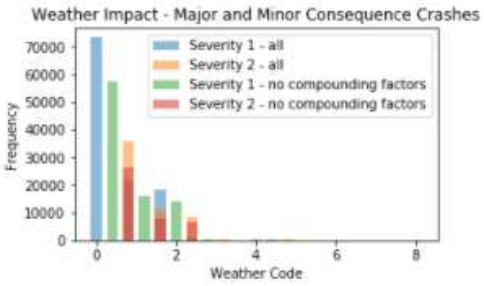


Figure 3 - Weather Condition Breakdown

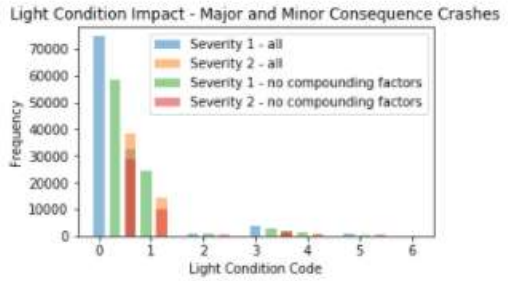


Figure 4 - Light Condition Breakdown

The overwhelming majority of light condition data classifies elements as either "Daylight", "Dark – street lights", "Dusk" or "Dawn", representing 55,049 out of 55,683 severity 2 crashes, and 112,378 out of 114,274 severity 1 crashes. The overwhelming majority of weather condition data classifies elements as "Clear", "Raining" or "Overcast", representing 55,284 out of 55,683 severity 2 crashes, and 113,112 out of 114,274 severity 1 crashes. Similarly, the vast majority of road conditions are classified as "Dry" or "Wet", representing 55,191 out of 55,683 severity 2 crashes, and 112,623 out of 114,274 severity 1 crashes.

The analysis of the elements not classified in the above groups should be taken with care, as they represent a relatively small subset of the data and the elements in other categories of lighting, road conditions and weather may not be statistically significant.

Of particular concern are the lighting condition of "Dark – Unknown Lighting" (lighting condition 6), and the weather conditions of "Severe Crosswind" and "Partly Cloudy" (weather conditions 7 and 8), as the sample size of the severity 1 and severity 2 crashes of each of these condition subsets are less than 10 elements. Therefore, it is not possible to reliably draw any conclusions around the impact of these conditions. However, this analysis may provide insight into where additional data gathering may be useful, and so these categories will not be excluded from the data analysis.

If historical weather data is available, future analysis may be performed to ascertain the degree of misclassification of weather data (for example, where partly cloudy weather was classified as clear).

Impact of Weather, Lighting Conditions and Road Conditions

The normalized value count of each weather, road and lighting condition were trended to allow visual comparison of the proportion of more severe crashes to the severity 1 crashes. These are shown below in Figures 5, 6 and 7.

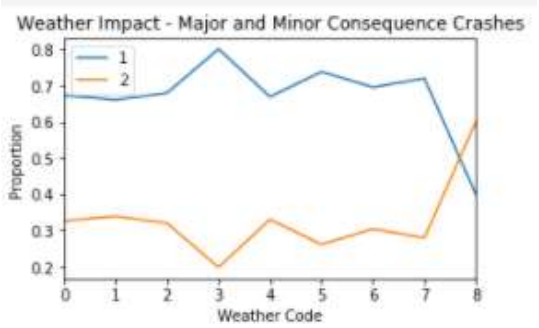


Figure 5 – Normalised Weather Crash Split

Figure 5 shows that the highest proportion of severe crashes occurs during partly cloudy conditions (weather condition 8), and that raining and overcast conditions (conditions 1 and 2 respectively) and sleet/hail conditions (condition 4) have a higher proportion of severe crashes than clear conditions (condition 0). It also shows that snowing conditions (weather condition 3) have a lower severe crash rate.

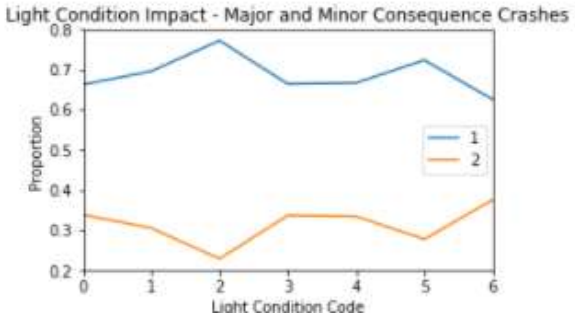


Figure 6 - Normalised Light Condition Crash Split

Figure 6 shows that daylight conditions (condition 0), have a higher severe crash likelihood than almost all other conditions other than condition 6 (dark – unknown streetlights), including dark conditions with streetlights (condition 1), dark conditions without streetlights (condition 2), dusk (condition 3).

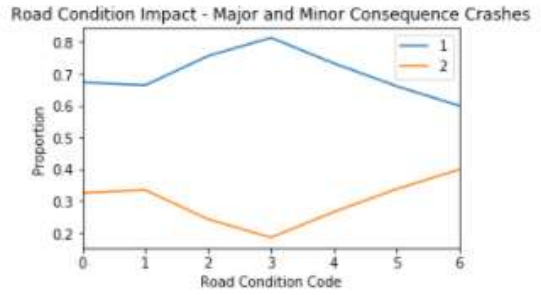


Figure 7 - Normalised Road Condition Crash Split

Figure 7 shows that wet conditions (condition 1), oil on the roads (condition 6) and sand/mud (condition 5) have a higher severe crash proportion than dry conditions (condition 0); however, icy conditions (condition 2), snow/slush (condition 3), and standing water (condition 4) have a lower severe crash proportion.

Similar figures have also been generated for the reduced dataset that excludes accidents where speeding, intoxication or distraction may have been a factor, in figures 8, 9 and 10 below.

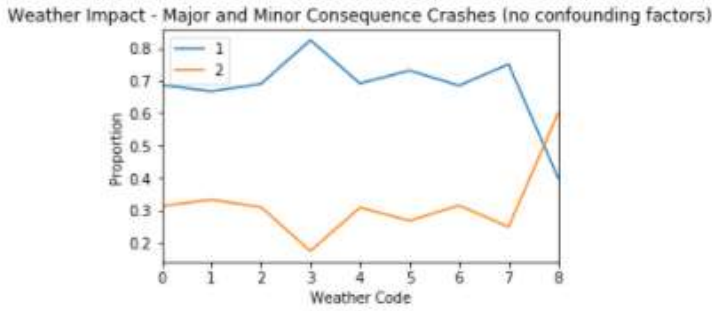


Figure 8 - Normalized Weather Split - No Speeding, Intoxication or Distraction

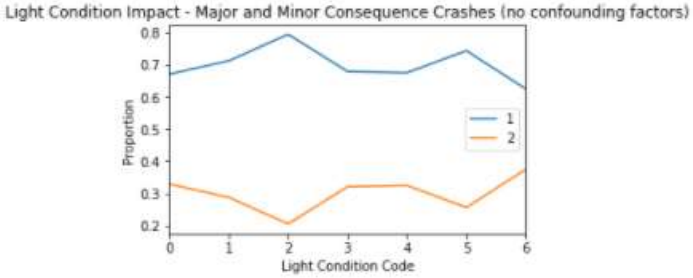


Figure 9 - Normalized Light Condition Split - No Speeding, Intoxication or Distraction

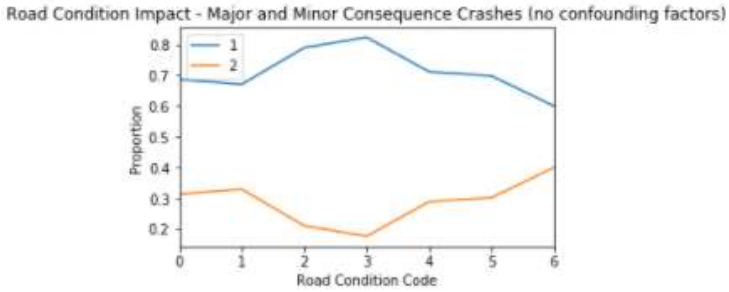
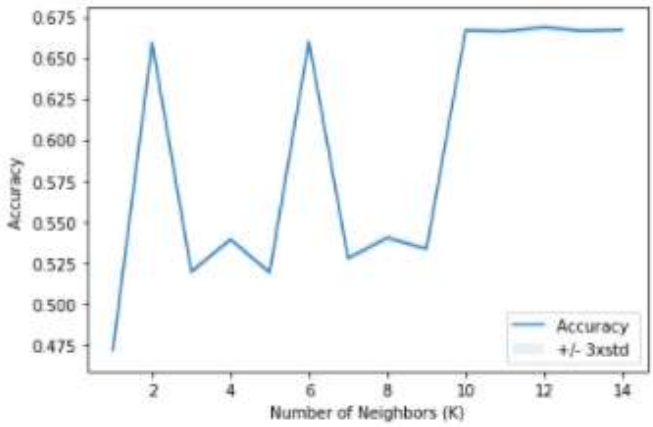


Figure 10 - Normalized Road Condition Split - No Speeding, Intoxication or Distraction

The proportion of severity 2 and severity 1 crashes when excluding data from crashes where speeding, intoxication or were very similar to (but not identical to) when the data was included.

Machine Learning Algorithm Assessment – Including Driver Distraction/Intoxication/Speeding

KNN analysis showed a KNN of 12 gave the highest accuracy, as shown below in Figure 11.




```
The best accuracy was with 0.6686867498234879 with k= 12
```

Figure 11 - KNN Model Accuracy

Using a KNN with a value of 12 gave the following accuracy when tested on the test data and train data:

```
Train set Accuracy:  0.6681425366822344
Test set Accuracy:  0.6686867498234879
Jaccard Similarity:  0.6686867498234879
F1 Score:  0.552273798552947
```

The test data was used to assess false positives and false negatives:

```
Count where predicted matches actual: 22730
Count where severe predicted, minor actual: 399
Count where minor predicted, severe actual: 10863
```

Decision tree analysis was also completed, which produced the following accuracy results utilizing the test dataset:

```
DecisionTree Accuracy:  0.6725994351612145
f1 score :  0.5443074212271202
jaccard similarity score :  0.6725994351612145
```

The test data was used to assess false positives and false negatives:

```
Count where predicted matches actual: 22863
Count where severe predicted, minor actual: 69
Count where minor predicted, severe actual: 11060
```

SVM analysis was also completed, which produced the following accuracy results utilizing the test dataset:

```
F1 Score 0.5774300165482842
Jaccard Similarity Score 0.6082607672393504
```

The test data was used to assess false positives and false negatives:

```
Count where predicted matches actual: 20676
Count where severe predicted, minor actual: 4473
Count where minor predicted, severe actual: 8843
```

Logistic Regression analysis was also completed, which produced the following accuracy results utilizing the test dataset:

```
Jaccard Similarity:  0.6728936220287126
F1 Score:  0.5413441909886492
Log Loss:  0.6311613116629152
```

The test data was used to assess false positives and false negatives:

```
Count where predicted matches actual: 22873
Count where severe predicted, minor actual: 1
Count where minor predicted, severe actual: 11118
```

In summary, for the full cleaned dataframe:

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

A similar analysis was performed for the subset dataframe that excluded accidents that involved speeding, intoxication or driver distraction. The results of the MLA accuracy analysis are summarized below:

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

Discussion and evaluation

Impact of Weather, Lighting Conditions and Road Conditions

Snowy weather (weather code 3) and icy and roads (road code 3 and 4) had the smallest proportion of severe crashes. Given the obstructed visibility and slippery roads that occur during snowy weather, this is what counterintuitive. With over 700 crashes represented in this dataset, it is not appropriate to write this off as statistically insignificant, and warrants further investigation. This investigation does not look into why this would be; but it would be worthwhile pursuing this line of questioning via qualative research to determine whether snowy weather prompts reduced trip frequency, increased driver caution, lower traffic density or some other factor that reduces the severity of accidents.

With only 5 results in the total dataset, there are insufficient elements for the partly cloudy result to be statistically significant; however, further investigation may be warranted, as it showed overwhelmingly the highest proportion of severe crashes.

Similarly, the lower severe crash proportion during dark/low light conditions than during daylight (light condition code 0) was surprising. The sample sizes for the dusk, dawn, daylight and dark – streetlights on conditions are sufficient to suggest statistical significance. Further qualative investigation will be required to ascertain the reason for this.

There was no major difference in proportion of severe crashes for a given road condition, weather condition or lighting condition when excluding data from crashes where speed, driver intoxication or driver inattention.

Machine Learning Algorithms

The highest Jaccard score obtained from the full dataset MLA was 0.67, using the Decision Tree analysis. This score, while less than optimal, shows that MLAs may have some limited use to support the stakeholder use of MLAs. However, the data is not at the point where the MLAs can be directly applied in the Australian context without additional information, since this data was obtained from the US.

Note that, as additional data will need to be gathered to tune the MLAs to the Australian context, lessons can be learned to focus the data gathering onto areas that enhance the MLA operation. Ensuring adequate sample sizes of all datasets, and ensuring that all stakeholders contributing to that data collection (e.g. police) have a common classification methodology will enhance the accuracy of the resultant future MLA.

A review of the F1, Jaccard and Log Loss scores show negligible difference (<1%) between the MLAs generated from the full cleaned data set and the MLAs generated from the dataset that had elements expunged where speeding, intoxication or inattention were factors in the crash. It is therefore more useful to utilize the full dataset, to maximise the available sample size.

Although the best machine learning Jaccard and F1 score has been obtained by using Decision Tree analysis, the false negative rate (predicting a severity 1 crash when, in fact, a severity 2 crash occurred) is higher than in the SVM analysis. Given that the different stakeholders have different needs for this data, some users may find the SVM more appropriate, despite the lower accuracy (for example, mapping and direction software to divert users away from areas with high severe crash potential has a different limit of acceptability for false negatives than software that directs emergency services to stand by near areas of high potential car accident severity).

Conclusion

The analysis of the Seattle Car Accident data suggests that weather, lighting conditions and road conditions have a significant impact on the likelihood of a car accident being severe.

Whilst a Machine Learning Algorithm may not be able to perfectly predict which conditions are more likely to result in severe car accidents, with the highest Jaccard score obtained of 0.67 obtained by the Decision Tree MLA, they provide valuable directional insights into the likelihood of severe accidents under a range of weather, lighting and road conditions. These insights may be utilized by policy makers to direct policy to discourage drivers from driving under the conditions where a severe accident is more likely, and mapping and driver directional software developers to direct drivers away from areas where a severe accident is more likely.

The Decision Tree MLA provided the highest accuracy, and may be of more use to policy makers, mapping software developers and information provided to road users. However, the high false negative rate means that the Decision Tree MLA may not be appropriate for use in emergency services determining when to increase emergency services availability and/or divert emergency services to areas where the MLA predicts a higher severe crash rate. Rather, the SVM model would be better utilized for this purpose.

Valuable insights may be further gleaned from the furthering of data collection in the Australian context. However, this data is not without its problems, due to the subjectivity of "road conditions" and "lighting conditions", and the fact that there is no clear indication of the magnitude of the speeding. Increasing the context available from the data may enhance the accuracy of the MLA.