ASSESSING ACCIDENT SEVERITY   
IN TRAFFIC COLLISIONS

Sue Neary

**1. PROJECT INTRODUCTION**

**THE PROBLEM**

Driving bears a risk of collisions, with accidents sadly resulting in 148 road fatalities in Ireland in 2019 (reference: www.rsa.ie). The technology for self-driving cars is fast advancing but, for the vast majority of drivers, it will be a while yet before self-driving cars are accessible. This project seeks to understand the risk factors which increase the severity of a road collision, in a bid to better understand how to make driving a safer activity.

**WHO WILL BE INTERESTED?**

The results of this project will be of interest to city councils who may be able to educate members of the public with road safety advice based on research. Through developing a better understanding of what increases the severity of car accidents, we can make our roads safer. By understanding what factors contribute to car accidents, city councils can inform the public on what to be aware of, to hopefully prevent a number of collisions or to reduce their severity.

The results will also be of interest to car manufacturers who can use this information to develop new features to make safer cars. This includes self-driving cars – for example, if environmental conditions increase the risk of a more severe collision taking place, this needs to be understood in the development of the self-driving technology.

**2. DATA**

**DESCRIPTION OF THE DATA**

This dataset shares information on vehicle collisions as provided by the Seattle Police Department and recorded by Traffic Records. The dataset has 38 attributes and 194,673 rows. There are a mix of numerical and categorical attributes, and some have missing data.

The attribute ‘SEVERITYCODE’ indicates the assigned severity of the accidents, with the numbers representing the following accident severity levels:

* 3—fatality
* 2b—serious injury
* 2—injury
* 1—prop damage
* 0—unknown

The dataset includes a number of other attributes containing further detail and information about recorded vehicle collisions in Seattle City. This project seeks to determine if any of these other attributes can operate as a predictor for Accident Severity.

Some potential hypotheses could suggest that environmental conditions on the day of the collision (for example weather, road conditions, light conditions and so on) or driver conditions (for example speeding on the part of the driver, or whether the driver was under the influence) might be predictors of the severity of an accident. This analysis will consider whether environmental conditions influence the likelihood of a more serious collision occurring.

The data dictionary can be found here:

<https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>

**HOW THE DATA WILL SOLVE THE PROBLEM**

The dataset in question only includes accidents which have a Severity Code (attribute SEVERITYCODE) of 1 or 2, indicating collisions resulting in ‘Property damage’ or ‘Injury’ respectively. As such, the model will attempt to determine what factors increase the likelihood of a more serious accident taking place, where an injury occurs rather than just property damage.

One avenue that might be explored is whether road conditions impact the severity of a collision. The following are the potential values for the attribute ‘ROADCOND’, indicating ‘Road Condition’:

* Dry
* Ice
* Oil
* Other
* Sand/Mud/Dirt
* Snow/Slush
* Standing Water
* Unknown
* Wet
* (blank)

This project will comprise some initial exploratory data analysis to better understand the dataset. Once this data exploration is complete, a model will be developed and then evaluated against the problem of: can we predict accident severity level?

**3. METHODOLOGY OF THE ANALYSIS**

**EXPLORATORY DATA ANALYSIS: Understanding ‘SEVERITYCODE’**

The head method was used to assess the first five rows of data, to develop an understanding of what fields comprised the dataset. The key attribute in the dataset, and the target attribute for much of the analysis, is ‘SEVERITYCODE’. The possible entries for ‘SEVERITYCODE’ are ‘1’ or ‘2’. ‘1’ indicates an accident involving property damage and ‘2’ indicates and accident involving injury.

After assessing the fields shown in the head of the dataset, the datatypes were printed, to gain further insight into the attributes in the dataset. This indicated that the field ‘SEVERITYCODE’ was type int64, or an integer. While 1 and 2 are integers, in this instance there are representing categories of accident severity (‘property damage’ and ‘injury’ respectively). While the larger number (2) indicates a more serious accident outcome, the number 2 does not hold a numerical meaning here. Indeed, the metadata file shows the additional information regarding ‘SEVERITYCODE’:

* 3—fatality
* 2b—serious injury
* 2—injury
* 1—prop damage
* 0—unknown

Given that serious injury is represented by the entry ‘2b’, it is understood that these numbers are categories, rather than integers. This informs the approaches taken to analyse the data.

**EXPLORATORY DATA ANALYSIS: Basic Dataset Understanding**

The correlation method was run to provide a table of correlation figures. However, this did not provide meaningful information in relation to accident severity, represented by SEVERITYCODE. As mentioned, although SEVERITYCODE has a datatype of integer, it is actually categorical. Were this a truly numerical field, it could have been helpful to assess the correlation via the ‘corr’ method.

The describe method was used to assess both numerical and categorical data.

The value counts were checked for the following attributes:

* SEVERITYCODE - A code that corresponds to the severity of the collision
* SEVERITYDESC - A detailed description of the severity of the collision
* WEATHER - A description of the weather conditions during the time of the collision
* ROADCOND - The condition of the road during the collision
* LIGHTCOND - The light conditions during the collision

From checking the value counts, it is clear that there are more instances of property-only damage than injury:

* Severity Code 1: 136485
* Severity Code 2: 58188

The ‘SEVERITYCODE’ attribute was converted to a string datatype. While possible entries are 1 and 2, this is a categorical variable representing accidents with either ‘property damage’ or ‘injury’ taking place. Converting the datatype facilitates the appropriate analysis of the data.

**EXPLORATORY DATA ANALYSIS: Looking at Weather, Road Condition & Light Condition**

The attributes ‘ROADCOND’, ‘LIGHTCOND’ and ‘WEATHER’ were then assessed in relation to ‘SEVERITYCODE’. As these are all categorical variables, seaborn’s countplot function was used to group the possible values and show the counts for each of these.

**Count: Weather grouped by Severity Code**

For ‘WEATHER’, we can see that for accidents resulting in both SEVERITYCODE 1 and 2, the most common weather condition recorded was ‘Clear’. For both codes 1 and 2, the second most common weather condition was ‘Raining’ and the third most common weather condition recorded was ‘Overcast’.

Chart

Description automatically generated

**Count: Severity Code grouped by Weather**

**Chart

Description automatically generated**

**Count: Road Condition grouped by Severity Code**

For ‘ROADCOND’. the most common road condition recorded for both SEVERITYCODE 1 and 2 was ‘Dry’, followed by ‘Wet’ as the second most common road condition recorded.

Chart, bar chart

Description automatically generated

**Count: Severity Code grouped by Road Condition**

Chart, histogram

Description automatically generated

**Count: Light Condition grouped by Severity Code**

For ‘LIGHTCOND’, the most common light condition recorded for both SEVERITYCODE 1 and 2 was ‘Daylight’, followed by ‘Dark - Street Lights On’.

Chart, bar chart

Description automatically generated

**Count: Severity Code grouped by Road Condition**

Chart

Description automatically generated

Having assessed the value counts for ‘WEATHER’, ‘ROADCOND’ and ‘LIGHTCOND’, it’s notable that these instances do not occur in equal proportions. As a result, some additional barplots were generated which calculated the percentage of occurrences of SEVERITYCODE 1 and 2 for the various categories in ‘WEATHER’, ‘ROADCOND’ and ‘LIGHTCOND’.

**Percentage: Weather**

The Weather plot shows that for most weather conditions the majority of collisions recorded result in a level 1 Severity Code. However, the exception to this is ‘Partly Cloudy’, the only weather condition where there are more level 2 Severity Code collisions, accounting for 60% of collisions recorded in Partly Cloudy weather.

Chart, bar chart

Description automatically generated

**Percentage: Road Condition**

The Road Condition plot shows that the road condition with the highest proportion of Level 2 Severity Code collisions (37.5%) is ‘Oil’, followed in second place by ‘Wet’ (33.19%). This provides additional insight on the earlier bar plot which showed the most common road condition for Level 2 Severity Code collisions was ‘Dry’. However, the most common road condition for Level 1 Severity Code collisions was also ‘Dry’, likely due to the fact that the majority of recorded collisions took place with dry road conditions. There were 124,510 collisions recorded with dry road conditions, accounting for 64% of the records in the dataset.

Chart, bar chart

Description automatically generated

**Percentage: Light Condition**

The Light Condition plot shows that the condition with the highest proportion of Level 2 Severity Code collisions (36.36%) is ‘Dark – Unknown Lighting’.

Chart, bar chart

Description automatically generated

**INFERENTIAL STATISTICS**

A number of Chi Square tests were run to see if Weather, Road Condition and Light Condition have an impact on Accident Severity.

**Chi Square Hypotheses**

Weather Chi Square

* Null hypothesis 1: Weather does not affect Severity Code
* Alternate hypothesis 1: Weather does affect Severity Code

Road Condition Chi Square

* Null hypothesis 2: Road condition does not affect Severity Code
* Alternate hypothesis 2: Road condition does affect Severity Code

Light Condition Chi Square

* Null hypothesis 3: Light condition does not affect Severity Code
* Alternate hypothesis 3: Light Condition does affect Severity Code

**Chi Square Results**

A screen shot of a social media post

Description automatically generated

The Chi Square test results indicate that Weather, Road Condition and Light Condition each (separately) impact on the Severity Code of a traffic collision.

**MACHINE LEARNING: Logistic Regression**

A subset of the original dataframe was created to contain only the fields relevant to the analysis (Severity Code, Weather, Road Condition, Light Condition) and null values were dropped from this new dataframe. The data was then separated into the dependent variable (the variable to be predicted: Severity Code) and the independent variables (Weather, Road Condition and Light Condition).

All of the data being analysed is categorical:

* Severity Code
* Weather
* Road Condition
* Light Condition.

As such, the data was then one hot encoded to facilitate implementing machine learning approaches.

A logistic regression was performed, resulting in a score of 0.9733333333333334. This indicates a high level of accuracy in prediction.

**Why choose Logistic Regression for Machine Learning?**

Logistic regression was performed due to all of the relevant attributes being categorical in nature, and because the attribute to be predicted (SEVERITYCODE) is binary. By one hot encoding the attributes SEVERITYCODE\_1 and SEVERITYCODE\_2, it’s possible to look just at SEVERITYCODE\_1. Therefore, rows with a value of 0 for SEVERITYCODE\_1 are collisions with a Severity Code of 2, and rows with a value of 1 for SEVERITYCODE\_1 are collisions with a Severity Code of 1. There are no rows with a missing value for Severity Code, so every row is either Severity Code 1 or Severity Code 2.

**Logistic Regression – Just Weather**

Weather has a large number of categories, and some of these categories do not have that many records. As such, Weather was then recategorised into just two categories: Clear and Not Clear. The value counts for these new categories were as follows:

Clear 111008

Not Clear 78329

A logistic regression was then performed on this new data with Weather categorised as Clear/ Not Clear, resulting in a score of 0.9733333333333334.

Chart, bar chart

Description automatically generated

This graph indicates that the majority of recorded collisions are Severity Code 1, and that the more common Weather condition is Clear rather than Not Clear. However, collisions with a Severity Code of 2 have a higher proportion of Clear weather conditions.

**MACHINE LEARNING: Decision Trees**

A decision tree for the Weather Clear/ Not Clear data was created, with a test size of 30%.

**Weather Decision Tree**

A picture containing object, clock

Description automatically generated

**Why choose Decision Trees for Machine Learning?**

A decision tree was chosen again due to the categorical nature of the attributes in the dataset. Initially a decision tree was produced including all original categories of Weather. The tree proved too large, and so a new tree was created with the Clear/Not Clear Weather data.

**RESULTS**

The results of this analysis show that Weather, Road Condition and Light Condition all impact the likelihood of an accident having a more severe outcome, as indicated by a SEVERITYCODE of 2 (injury) recorded instead of 1 (property damage).

Further analysis could be done to assess these categories in further detail, and to explore how these categories interact with each other.

**DISCUSSION**

Based on the results, a recommendation to the city would be to consider educating the population on road safety, highlighting the fact that Weather, Road Condition and Light Condition all play a part in increasing the likelihood of a more severe collision taking place.

Clear conditions often result in a higher level of accident severity – it could be extrapolated that in clear conditions, drivers feel more confident in road safety and are likely to drive faster or pay less attention as a result.

Measures were taken to handle the categorical nature of the data in this dataset. Another recommendation would be to consider other data sources regarding road safety, which may provide additional opportunities for other machine learning approaches.

**CONCLUSION**

In conclusion, this topic will be of interest to city governments with a view to educating the population on road safety and on how Weather, Road Condition and Light Condition impact the Severity Code of traffic collisions.