An Investigation Into What Causes Crash Injuries, Fatalities

Using association analysis on crash data from the City of Chicago

Nabil Darwich  
 Department of Computer Science  
 George Mason University  
 Fairfax VA USA  
 ndarwich@gmu.edu

Hamza Mughal  
 Department of Computer Science  
 George Mason University  
 Fairfax VA USA  
 hmughal2@gmu.edu

ABSTRACT

Being able to save human life is one of the most important goals facing various communities. An unfortunately common reality in which lives are lost is due to car accidents, and there are always public service announcements to advise on avoiding them. In this paper, an analysis is conducted regarding the true factors that lead to injuries and fatalities when it comes to car crashes. Data from the City of Chicago, during the summer of 2018, is used to find out what some of these causes are, and a model is built to determine accident severity outcomes given the elements involved. The summer season was picked to minimize external factors such as snow leading to crashes, and the most recent summer was that of the current year, 2018. A goal from this paper is to raise awareness regarding what can be done to avoid such tragedies.

1 Data Analysis

1.1 - Visualization of Car Accidents in Chicago

During the summer of 2018, there were approximately thirty thousand recorded car crashes throughout Chicago, as reported by city officials. Recorded information about these crashes ranged from data about the incident’s occurrence, such as geolocation, to data about the driver, such as their sex.

Naively mapping the latitude and longitudes with matplotlib’s pyplot yielded an image similar to the shape of Chicago as it can be seen here:



Not being able to discern much information from this map, though, we decided to plot each accident against an actual map of Chicago. This was accomplished using Baseplot in python along with associated helper files. Mapping the accidents on an actual map yielded much more informational results:



While this graph was helpful in visualizing our data, it was very intriguing how one location in the Southeast had very few crashes. Thus, an investigation into why that is the case took place. The population map of Chicago revealed that only a few individuals lived in that location, with a vast majority of them choosing to live along the center of Chicago, rather than its outskirts.



From this map, the suspicion that the majority of car accidents would occur throughout the densest areas of Chicago naturally came about, and mapping the accidents against the population densities seemed to corroborate our hypothesis:



As can be seen from the map above, the outskirts of Chicago have fewer accidents, with the majority of accidents being in the center of Chicago, where it is most dense in population.

**1.2 - Analysis of Chicago Accidents**

Prior to data mining work, it was also insightful to find that some obvious patterns in the data were confirmed when the data was graphed. But in other times, the answer did not seem as intuitive as originally thought. Using the summer 2018 data, this is what we found:



Figure 1: Injury Severity vs. Posted Speed Limit

In here, the intuitive thought that roads that have higher speed limits are involved in more severe crashes was confirmed. The x-axis in Figure 1 is the posted speed limit where a crash occurred, and the y-axis is the % occurrence of each accident severity. As can be seen, accident severities gradually get worse as the speed limit is increased. There are a few outliers when it came to higher speed limits, due to lack of data involving the same road conditions. As an example, only one crash happened when the speed limit was 70 during the summer of 2018, and fortunately, no serious injuries occurred. Due to this anomaly, however, a suggestion that roads that have a speed limit of 70 have a 100% survival rate may arise, which is definitely a false conclusion.



Figure 2: Weather Condition vs. Crash Severity

A somewhat intuitive graph that was also generated involved the severity of the crash when it came to weather conditions. As it can be seen, rainy days involved much more injuries than clear/overcast days. But what came in as very surprising is the amount of injuries and fatalities that occurred during foggy/smoky/hazy days, which far exceeds rainy days. However, this graph may also lead to a few false conclusions, as one can draw that injuries never occur during days with snow or hail, but again, this was only due to the infrequency of data elements in that subset, as snow/hail rarely occurred in Chicago during the summer of 2018.



Figure 3: Lighting vs. Crash Severity

The last graph that was generated during pre-data mining analysis was the lighting conditions at the time when the crash occurred. The results here were also consistent with natural biases, where crashes that occurred at night involved much more injuries than daylight. However, it was surprising to find out that in pitch darkness, without any streetlights to light up the path, crashes that involved injuries were relatively infrequent when compared to crashes that occurred during daylight. A possible explanation for this phenomenon may be that streets with pitch darkness are not as populous as streets where streetlights have been installed, and individuals there are much more cautious when driving.

2 Data Mining

2.1 - Data Preprocessing

Preprocessing was the most time consuming and crucial step in our process. Modifying the data set such that it keeps the same meaning and be interpreted by a data mining algorithm is definitely a challenging task faced by many data scientists and statisticians.

**2.1.1 – Selecting a Training Sample**

The original City of Chicago car crash data set was divided into 3 separate datasets that contained more than 200,000 that date as far back as 2014. Those datasets were contained information about the following:

1. The crash itself (date, weather, lighting, primary cause, …) [1]

2. The individuals involved (age, sex, injury classification, …) [2]

3. The vehicles involved in the crash (make, model, year, …) [3]

We have come to notice that vehicular data was largely sparse and infeasible to deal with and discretize without any general vehicle descriptions (i.e.: size of the vehicle). So, we agreed to only deal with information about the crashes and individuals involved.

Next. we have come to realize that different seasons often meant different reasons for crashes. For example, roads will be in much rougher conditions during winter months, and that information is not necessarily captured by the datasets. We did not want these details to bias our results, so, we decided to only deal with summer data for this project, with dates ranging [6/21, 9/21]. Finally, when it came to the year, we decided to select 2018 as our training year to build the most accurate and up to date model that reflected current regulations and road conditions. We decided to use summer 2017 data for testing, and summer data from the other years was used for our training model.

**2.1.2 – Combining Crashes and People**

With two distinct datasets, crashes and people, our next challenge was combining them into one. The two tables both featured a column called RD\_NO, which stood for the report number of the crash. For each individual, the report number was the crash they were involved in, so multiple individuals often showed up with the same report number. With that in mind, a left outer join was done from people onto crashes based on the report number column being equivalent. This step created a table that had information about every individual merged with the crash they were involved in.

**2.1.3 – Discretization (binning)**

With crashes and people now merged into one, allowing machines to operate on the data was next. For that to happen we had to transform each cell from a label {DARKNESS, DAWN, DAYLIGHT, DUSK} to a number {0, 1, 2, 3}. We decided to enumerate the values for each column and doing so initially required a manual inspection of the data and recording of those unique values. Once they were recorded, a mapping was created to allow for a translation from raw strings to numerical data.

Later on in our process, we felt the compulsion to also discretize age data, which was already numerical, by grouping it {CHILD, YOUNG ADULT, SENIOR, etc..} then enumerating it as well.

**2.1.5 – Feature Reduction**

Numerous columns were found to have no use as they were either redundant, irrelevant, or represented a consequent that we’re not aiming to determine. Examples of such columns included duplicated columns, report numbers, and their the individual was transported to an Emergency Room. What we did was simply delete those columns.

2.2 - Transformation

**2.2.1 - Oversampling**

When briefly looking over the data, we noticed a large discrepancy between the number of non-incapacitating and fatal injuries. It is rather difficult to successfully determine causes of fatal injuries due to the massive bias towards one class. If we were to train any classifier as is, there most certainly will be overfitting where we will then see horrible prediction results.

To try to lessen the bias towards the non-incapacitating, we are choosing to oversample the data so that there is a much more even distribution between all classes by using a random sampler from sklearn. This will let the classifier not be overfitted/underfitted so that we can more accurately predict when fatalities may occur during car crashes.

**2.1.6 – Refining Features**

Running data mining algorithms made us realize that more features needed to be evicted. The reason for the eviction often was their sheer commonness or the lack of information they helped bring.

2.3 - Data Mining

**2.3.1 – Weka**

Data mining was done using the Weka software released by the University of Waikato in New Zealand under the GNU public license. Under the explorer tool, the primary applications we will be utilizing is classification and association.

**2.3.2 – Weka Data Format (ARFF)**

Weka primarily uses a file format known as Attribute-Relation File format (ARFF). Because our test and training files were in the CSV file format, we had to convert them into ARFF format.

**2.3.2 – Association Analysis**

Before dealing with any type of classification, we were interested in what antecedents lead to fatalities. To accomplish this, class association rules (CAR) was set to true with the class index being set to 18 to indicate Injury Classification as the consequent (as a minor note, setting a specific consequent will only output Confidence so we will not be displaying any other metrics such as Lift). The minimum confidence is set to 0.3, and the lower bound minimum support is set to .0001 (this means that for an item to be considered frequent it must be appear in about .0001 % of the data).

After running association analysis with these parameters on the training data, we selected a few interesting rules that were found:

1. TRAFFIC\_CONTROL\_DEVICE=NO\_CONTROLS LIGHTING\_CONDITION=DARKNESS LIGHTED ROAD ROAD\_DEFECT=NO DEFECTS SEX=M ==> INJURY\_CLASSIFICATION=FATAL conf:(0.87)
2. FIRST\_CRASH\_TYPE=FIXED OBJECT SEX=M ==> INJURY\_CLASSIFICATION=FATAL conf:(0.85)
3. LIGHTING\_CONDITION=DAYLIGHT FIRST\_CRASH\_TYPE=ANIMAL==> INJURY\_CLASSIFICATION=FATAL conf:(0.33)

The first rule states that crashing in a lighted non-defective road at night and being male leads to fatal injury. The second rule states that crashing into a fixed object (e.g: tree) and being male leads to fatal injury. The third rule states that crashing into an animal in daylight leads to fatal injury. Even though the third rule has an extremely low confidence compared to the two other rules, it is still interesting to note despite the low likelihood of it occurring.

**2.3.3 – Classification**

Additionally, Weka was used to do classification. Weka contains a plethora of classification algorithms at our use, however we only decided to use a handful. Given our dataset (knowing that we had the ground truths and a decent number of features), we decided on primarily using Naïve Bayes, Random Tree, and Random Forest.

**2.3.4 – Naïve Bayes**

Naïve Bayes is a basic probability-based classifier with assumes the independent nature of features. This independent assumption makes the classification very fast. The formula by which Naïve Bayes operates is:



The advantage of using Naïve Bayes with our dataset is that it is very fast, linearly scalable with the number of features, and can easily be used for multi class classification problem prediction.

Given that we have a multi class problem, it was quite easy to determine that Naïve Bayes would be one of the algorithms to use during classification.

2.4 - Evaluation

these steps:

1. In a Word 2013/2016 document, insert a picture.
2. Right click on the inserted picture and select the **Format Picture** option.
3. In the settings at the right side of the window, click on the "Layout & Properties" icon (3rd option).
4. Expand **Alt Txt** option.
5. In the "Title:" and "Description:" text boxes, type the text you want to represent the picture, and then click "Close".

*1.1.1 Heading Level 3.* Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

*1.1.1.1 Heading Level 4.*Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here. Insert paragraph text here.

3 Testing Against Summer 2017

ACKNOWLEDGMENTS

Large thanks to the City of Chicago for making the three data sets publicly available for research. Unlike previous datasets we have researched, the City of Chicago data was very structured, uniform, and consistent. The amount information that was recorded in the datasets was immensely helpful in deriving these conclusions

REFERENCES

[1] FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.

[2] F.N.M Surname, Article Title, https://www.acm.org/publications/proceedings-template.

[3] F.N.M Surname, Article Title, https://www.acm.org/publications/proceedings-template.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00