An Investigation Into What Causes Crash Injuries, Fatalities

Using crash data from the City of Chicago (Summer 2018)

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.



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 part in our implementation. 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 scientists and engineers.

**2.1.1 – Selecting a sample**

The original City of Chicago car crash data set was divided into 3 separate files that contained more than 200,000 entries for each file that date as far back as 2014. The crashes were split into a file that had information about the incident itself (e.g: time it took place), another file that had information about the individuals involved (e.g: age), and another file that had information about vehicles involved (e.g: Make and Model). As vehicular data was very sparse and often uninterpretable, we opted not to analyze vehicles, but focus on the subjects involved and the crash details. When it came to that, a few inconsistencies were noticed with some older data, such as not being as regular nor as frequently updated, so the decision to choose a subset of the original data, dating from 6/21/2018 to 9/21/2018 was made. Having recent data meant the data that was dealt with met the most up-to-date standards and was largely error free, which helped enormously with our methods. So, the data that was selected was Summer 2018 data for Car Crashes and Individuals Involved, the two subsets respectively being available in the files *data/crashes.csv* and *data/people.csv*.

**2.1.2 – Numbering Labels**

Words are very meaningful to humans, but computers

**2.1.3 – Combining datasets**

After all numbering was done

**2.1.5 – Selecting Features**

The data we now had was almost ready to go through mining.

**2.1.6 - 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.

2.2 Transformation

Below are steps to place alt-txt value in **MS Word 2013/2016**. To add alternative text to

2.3 Data Mining

a picture in Word 2013/2016, follow

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.

ACKNOWLEDGMENTS

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.

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 and F.N.M Surname, 2018 Article Title, The title of book two (2nd. ed.). Publisher Name, City, State, Country.

[4] Ian Editor (Ed.). 2018. *The title of book two* (2nd. ed.). University of XXX Press, City, Chapter 100. DOI: <http://dx.doi.org/10.1000/0-000-00000-0>.

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK’18). ACM, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

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