Can we predict injuries in car crashes in Cincinnati ?

Eoin Mc Cormack
Masters Student
Dublin City University
Dublin, Ireland
eoin.mccormack35@mail.dcu.ie

***Abstract*— Using data analysis techniques on a dataset of historic road traffic accidents in Cincinnati we look to predict the likelihood of suffering an injury in a car crash. The proposal to predict injuries centers around removing columns that duplicate other columns, removing columns that do no offer salient information. With the irrelevant information removed we then explore the remaining columns to look for gaps in data or data that appears to be wrong and see if we can put a meaningful value in the cell. Once left with relevant and cleaned data we will model the data and use linear regression to determine the likelihood of someone being injured or killed in a car crash in Cincinnati.**

I. Introduction

Since the invention of mechanically propelled vehicles there have been hundred of millions of Road Traffic Accidents on the roads all around the globe and due to the weight and speed of motor vehicles tragically a lot of lives have been lost and a lot of people have suffered injuries, some life threatening. In 2020 in the United States [1] "*38,680 people died in motor vehicle traffic crashes—the largest projected number of fatalities since 2007*". This is a worrying statistic given the traffic on roads around the World was a fraction of what it normally was due to the Covid-19 Pandemic and Insurers were refunding their customers due to low volume of usage of their vehicles [2] "*Those relief measures generally shaved 15 to 25 percent off customers' premium payments for one or more months during the spring and returned $14 billion to policyholders*".

Apart from the human cost there has been a massive financial cost to the insurance industry due to treatment of injuries, palliative care and legal costs. The impact of any increase in claims payments will inevitably result in premium increases which can lead to loss of customers who are looking to keep their premiums at a similar or lower level to the previous year.

Over a period of ten years data pertaining to road traffic accidents has been gathered and saved to the dataset Cincinnati Car Crash EDA w/ Folium | Kaggle, this information contains a wide array of details about the accidents. The dataset contains a lot of columns information about these accidents which needs to be whittled down to only leave information we feel will help with the task of prediction. As with all datasets we faced missing values and incomplete values in various formats which need to be either analysed cleaned or removed.

Once we have scrubbed, and salient data we can start the machine learning task of endeavouring to predict under what circumstances a person in a vehicle is likely to suffer an injury or fatality.

II. RELATED WORK

This involved reviewing some similar publications focusing on Road Traffic Accidents and ways to avoid them or try to mitigate them.

The first study I reviewed was seeking to use weather and crash reports to determine if wet weather makes crashes more likely [3] *"Of two studies that focus specifically on fatal traffic crashes, one finds an increase in the crash rate of over 100% during rainy conditions [6], and the other finds an increase in one country (Denmark) and no significant change in two other countries (Norway and Sweden) [10]."* This paragraph contrasts the effect of rainy conditions on the roads in three different countries, the way the roads are laid in Denmark could result in a finish less well able to deal with water. Our dataset shows that using the mean there is only a slightly higher chance of injury / death between wet and dry road, however water on the road drastically increases the likelihood of an injury or death, assumably due to tyres not being able to grip as well in standing water.

ROADCONDITIONS	INJURIES
DRY	0.165613
WATER (STANDING, MOVING)	0.328571
WET	0.177942

Fig1: Mean of an injury occurring on dry / wet conditions

It is widely held that women are safer drivers and less fatalities on roads are women, this is backed up by [4]

Total motor vehicle deaths*	
Male	Female
26,040	10,766
25,634	10,420

Fig 2: Total fatalities by gender on US roads in 2019

My observation of injuries / death by gender show some interesting results:

GENDER	INJURIES
FEMALE	0.197827
MALE	0.142902

Fig 3: Mean injuries by gender

Data Mining Methodology

Approached this work using KDD Methodology.

In my previous career as an insurance / reinsurance underwriter I gained a wealth of knowledge in the inner workings of the insurance industry and motor insurance is an area I had a lot of exposure to. It was abundantly clear that men's insurance was more expensive and even bizarrely if a man was the only person insured on his car the premium would come down if he added a woman on to his insurance. When I was looking for a dataset I searched Kaggle for insurance and I came across the dataset I have used.

Data cleaning:

Imported the data and took a look at head to ensure it had imported correctly, then I used info to see the structure of the data and noticed some columns were objects, some int and some float.

Used drop to remove columns that did not offer any meaningful data.

Injuries needed some work as it contained numbers and descriptions in a String, however 5 related to 'no apparent injury' and 'fatal injury' so I updated all rows to either injury(including fatal injuries) or none.

Road Surface contained '6 – other' and '9 – other'

Other columns contained a dash, a substring was used to get the String after the dash

Fillna was used to replace na with unknown

Data selection:

I used a Logistic Regression approach using Sklearn, I felt this was the best fit for my model as any number of classes could result in an injury.

I created train and test datasets and then x and y for train and test.

I tried Logistic Regression fit but this did not work as the age was in a range a String format and it was looking for a float. In an effort to resolve this I wrote a method to use the mean of the age range and create a new age column which was a float type, however, I faced the same issue with another column.

Then I attempted to convert my dataset to one hot, however it was complaining about the indexes of the data, I was unable to resolve this.

My next attempt was to use Label Encoder to convert the data set but again fit complain about the data.

Finally, I created a new numerical column for each remaining column. I created a list for each column using unique to capture all the values. From the list I used the list index of the value to populate the new numerical columns. Using this approach the model worked.

EVALUATION / RESULTS

Tweaking the train and test split, the best results were achieved with a 30% test / train split, they are as follows:

- Accuracy score 0.85
- Precision score 0.60
- Recall score 0.24
- F1 score 0.35

The following github repository contains all information: https://github.com/steevo51/eoinmccormack-15112128-CA683.git

FUTURE WORK

Once the exams are out of the way I will come back to this and try to get some results and an evaluation.

REFERENCES

[1] NHTSA (2021) 2020 Fatality Data Show Increased Traffic Fatalities During Pandemic
Available at:
2020 Fatality Data Show Increased Traffic Fatalities During Pandemic | NHTSA

[2] AARP – P. Kiger and A. Markowitz (2020) Big Auto Insurers Phase Out Refunds for Policyholders. Available at: https://www.aarp.org/auto/car-maintenance-safety/info-2020/coronavirus-car-insurance-premium-refund.html

[3] M. A. Abdel-Aty and R. Pemmanaboina, "Calibrating a real-time traffic crash-prediction model using archived weather and ITS traffic data," in IEEE Transactions on Intelligent Transportation Systems, vol. 7, no. 2, pp. 167-174, June 2006, doi: 10.1109/TITS.2006.874710.

[4] IIHS (2021) Fatality Facts 2019 Males and females. Available at:
Males and females (iihs.org)