**Predicting the severity of car accidents**

**in Great Britain**

Yue Wang

September 1, 2020

1. **Introduction**
   1. **Background**

Road safety and the occurrence of car accidents has been one of the biggest concerns across the world. According to the World Health Organization (WHO), there are approximately 1.35 million people die each year as a result of road traffic crashes and they cost most countries 3% of their gross domestic product. On average, five people die every day on the road in Great Britain and countless more are seriously injured. Britain's road safety record has stagnated in recent years, with the number of road deaths remaining broadly constant for several years.

* 1. **Who would be interested**

It would be helpful if the UK local government can gain more understanding from the car accidents that already occurred so they can take immediate and effective actions to further reduce the severity of incidents on the roads. Looking at historical data behind the car accidents would also benefit the local insurance companies for them to work out better model on premium evaluation and claim payment.

* 1. **Objective**

There are various of factors could contribute to determining the severity of car accidents in UK, some are related to people (drivers, passengers etc.) and the others are related to road type or nature. This project aims to predict how these factors would affect the severity of the car accidents.

1. **Data Description**
   1. **Data sources**

The dataset to be analysed in this project can be found from Kaggle [here](https://www.kaggle.com/akshay4/road-accidents-incidence) . It was fetched

from UK government open data sources and primarily captures road accidents in UK between 1979 and 2015 and has 70 features/columns with around 250,000 records.

* 1. **Data Cleaning**

It is clear that the dataset fetched from Kaggle has already transformed the string values under features into numbers. There are a few features related to people involved in the car accidents, while some other features describe the road or weather conditions. However, we don’t need all 70 features, many of them can be redundant due to either not fitting our project purpose or having less impact/missing data on our analysis.

The objective of this project is to analyse the severity of car accidents, but the data include various types of vehicles on road. I found that number 9 represent ‘car’ under ‘Vehicle Type’, so we only keep those records for car accidents and drop the others.

* 1. **Feature Selection**

Some of the features are grouped by another feature, for example, ‘Age band of driver’ is obtained by splitting the feature ‘Age of driver’ into 12 classes. In this case I would only keep one of them as they are providing same information. I decided to use ‘Age band of driver’ feature in my analysis.

Features related to casualties’ details are not very important in this project as most likely they don’t cause the accident (unless they are drivers), and also it is easy to notice that there is a lot of missing values for casualties. Therefore, we remove these features and only focus on drivers’ details.

For the similar reason, we are not looking into too much road details such as road number and locations but will retain the features like speed limit and road surface condition which are obviously having impact on the accidents. Also, we are not interested in the policy force and officer who dealt with the accidents, as the result, the related features can be excluded from the analysis.

Base on the above consideration, I discard the redundant features and retain 11 features (Table 2). There are 3 features contain missing or out of range value. After checking, I have recorded the percentage of missing value in these features in below table:

Table 1: Features with missing or out of range value

|  |  |  |
| --- | --- | --- |
| Feature | Number of missing or out of range value | % occupied in total 205,852 records |
| Sex of driver | 14 | 0.0068%(Not including unknown value) |
| Age band of driver | 23,350 | 11.3% |
| Road surface condition | 390 | 0.189% |
| Weather condition | 3340 | 1.62% |
| Light condition | 2759 | 1.34% |

Based on the above table, the records with missing or out of range value for ‘Sex of driver’, ‘Road surface condition’, ‘Weather condition’ and ‘Light condition’ are small therefore those records were removed. However, the percentage for ‘Age band of driver’ is high so I decided to use the mean value 6 (actual mean 6.0265) for those records.

I noticed that the date value under ‘Time’ feature is all wrong, it seems that the dataset only wants us to look at the time value, so I have cleaned up the column to only contain the time value.

The final features are listed in below table:

Table 2: Features selected and the corresponding key values

|  |  |
| --- | --- |
| **Kept Features** | **Key Value** |
| Accident Severity | 3 values:   1. Fatal 2. Serious 3. Slight |
| Age band of Driver | 11 groups from age 0 to 75, and over 75.  5 years in each band |
| Sex\_of\_driver | 3 values:   1. Male 2. Female 3. Not known |
| Day\_of\_week | 7 numbers that represents each day of the week |
| Speed\_limit | Actual speed limit in the road |
| Light\_conditions | 4 values:   1. Daylight 2. Darkness – lights lit 3. Darkness – lights unlit 4. Darkness – lighting |
| Weather\_conditions | 8 values:   1. Fine no high winds 2. Raining no high winds 3. Snowing no high winds 4. Fine and high winds 5. Raining and high winds 6. Snowing and high winds 7. Fog or mist 8. Other |
| Road\_surface\_conditions | 7 values:   1. Dry 2. wet or damp 3. Snow 4. Frost or ice 5. Flood over 3cm. deep 6. Oil or diesel 7. Mud |