**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 |

1. **Data Analysis.**
   1. **Relationship between severity of car accident and sex of drivers.**

From the pie charts (figure 1) we can see that the accidents caused majorly by drivers’ ages are between 21 to 45, however this is also because that most drivers’ ages are with this range. The percentage of this group of people in the car accidents reduced when the accident severity level increased from Slight to Fatal, while drivers with age above 75 increased from 3.1% to 8.2%. This indicates that people over 75 are more likely to deliver serious accidents.

**A close up of a logo

Description automatically generated A close up of text on a white background

Description automatically generated**

**A picture containing umbrella

Description automatically generated**

Figure 1: Relationship between accident severity and age band of driver

* 1. **Relationship between severity of car accident and accident day of the week**

There is no significant difference on the accident severity level between the days of the week, but we can still observe that weekend is relatively having more slight accidents comparing to Monday which has higher chance on fatal accidents. (Figure 2)

**A picture containing drawing

Description automatically generated**

Figure 2: Relationship between accident severity and accident day of the week

* 1. **Relationship between severity of car accident and sex of driver**

In the occurrence of car accidents, most of the drivers are male, this is also due to that male are the main drivers on the road. However, from the chart (Figure 3) we can see that with the increase of the accident severity level, the proportion of female driver is getting lesser. It could be that female drivers are more conservative and have more concern on road safety comparing to male drivers.

**A screenshot of a cell phone

Description automatically generated**

Figure 3: Relationship between accident severity and sex of driver

* 1. **Relationship between severity of car accident and speed limit**

Most of the car accidents happened when the speed limit of the road is 30km/h (Figure 4), it seems that the low speed limit has caused the lack of the road vigilance, with the majority of the accidents with limit 30km/h are slight. However, when the speed limit is above 60km/h, the occurrence of fatal accidents has significantly increased. This is reasonable considering the risk of high speed.

A screenshot of a cell phone

Description automatically generated

Figure 4: Relationship between accident severity and speed limit

* 1. **Relationship between severity of car accident and light conditions, weather conditions and road surface conditions**

The 3 bubble charts each represent the records of light, weather and road surface conditions in the car accidents, the size of bubble demonstrate the amount of accidents in each severity level, 1 is slight and 3 is fatal.

From what we can see on the charts, most of the car accidents happened with a day light, fine weather with no wind and dry road condition, but we cannot draw any conclusion from this as these 3 conditions seems most common under each category.

Although no significant difference, accident severity tends to be more serious under darkness conditions (1st sub chart), especially when it is darkness with lighting. When accident happened, if the weather is high winds, fog or mist, the severity is more likely to be fatal than good weather. With regard to the road surface condition, comparing to other types of road surface condition, dry surface seems to have a little bit higher proportion of slight accidents.

A close up of a map

Description automatically generated A close up of a map

Description automatically generated

A close up of text on a white background

Description automatically generated

Figure 5: Relationship between accident severity and light conditions, weather conditions and road surface conditions

1. **Modeling**

Due to the predicting value ‘accident severity’ being grouped in 3 classes – Fatal, Serious and slight, I used classification model in this project. I randomly split the datasets into train and test sets with 4 to 1 ratio and trained the data in the models including K Nearest Neighbour, Decision Tree, Support Vector Machine and Logistic Regression. Looping through the according parameters, I find the best value with highest accuracy in each model.

Among these models, Decision Tree and Logistic Regression model performed the best, both have accuracy at around 86.7357% (Table 3)

Table 3: Performance of classification models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K Nearest Neighbour | Decision Tree | Support Vector Machine | Logistic Regression |
| Accuracy | 85.8329% | **86.7357%** | 80.6238% | **86.7357%** |

1. **Result and Discussion**

We have achieved high accuracy from Decision Tree and Logistic Regression model, however the dataset that we analysed on was not perfect and there are many things we need to be careful with. For example, when we cleaned up the data, we replaced the missing or out of range value with mean for feature ‘age band of driver’, this assumption could possibly make the dataset very biased. It also comes to my attention that among the features I chose, there are some value of features are not identified, such as ‘unknown’ value for sex of driver, the insufficiency of important information is a concern.

In addition to the above, we cannot ignore the truth that in the dataset, one accident doesn’t necessarily contribute only one row of records. In other words, those records are not completely mutually exclusive, a deeper analysis on this would be required to receive more accurate information.

**6 Conclusions**

To conclude, I have analysed the relationship between the car accident severity with features in relation to drivers and environmental conditions. The classification models were built to predict the accident severity given the related features.

This would be helpful for the local government to come up with some ideas regarding how to reduce the chance of car accidents and their severity. Based on the result, the pricing teams in the general insurance companies are able to calculate the reasonable premium rates which is also competitive to the market.