# **What insight does the UK accidents data give us?**

By Shivani Sanjay Mahaddalkar

**Introduction**

Did you know that the United Kingdom boasts some of the safest roads in the world, with only a few countries ranking higher in terms of road user safety than the United Kingdom? Despite this, accidents continue to be quite common, with 122,635 reported in 2018 - or almost 336 every day. Between 1951 and 2006 a total of 309,144 people were killed and 17.6 million were injured in accidents on British roads. With increased emphasis on road safety and awareness, these fatalities could be avoidable. Therefore, it is necessary to determine what improvements can be made.

This could be due to the fact that there are more cars on the road as people enjoy the sunlight. Or it could be due to things like sun dazzle, when light bouncing off dirty wind screens causing dangerous visibility issues. Or do we blame bad weather and small wonder winters? Or perhaps the constant showers resulting in slippery roads? Well, to solve this conundrum, I analysed a dataset of all traffic accidents reported from 2012 to 2014. On understanding the factors contributing to the accidents and their severity, more regulations could be introduced to safeguard against road hazards. This is my report to understand the causes behind the reasons for road accidents.

**Objective**

This analysis aims to create a model that understands how factors like the road conditions, weather conditions, highways, play a role behind the severity of the accidents. It aims to use these factors to predict the severity of an accident, into three classes, minor, medium, and major. The severity is determined by the number of people injured and how fatally they have been injured.

Intuitively, bad weather conditions, high speed limits, bad road conditions should lead to more severe accidents, this analysis aims to confirm that intuition.

Using R software, we aim to build decision trees and Naïve Bayes model to predict the severity of an accident when a few ambient conditions are known.

**Data Description**

The data set is sourced by user Dave-Fisher Hickey on Kaggle. The data is originally sourced from the UK Department of Transport website. The data originally contained over 450,000 data entries with 33 columns. The data set has 33 columns, including columns like ‘Road Type’, ‘Speed Limit’, ‘Light Conditions’, ‘Weather Conditions’, etc. which on preliminary glance seem like factors that could have a major impact on an accident occurring and its severity. The data had entries from the year 2012-2014.

On some exploratory analysis of how well balanced the columns were and how useful a certain attribute would be in determining the accident severity, certain columns were chosen further. The columns used for modeling were Road Conditions, Weather Conditions, Speed Limit, Rural/Urban and Day of the week.

To sample a balanced data set, 500 data points for each level of accident severity were chosen.

**Data Preparation**

1. Sampling the data set

To check if the sample represents the overall population of the data, the histograms of the population and the sample were compared. One such example is the weather conditions which is shown below.

Chart, bar chart, treemap chart

Description automatically generated

Figure . Distribution of Weather Conditions in Population

Chart, bar chart

Description automatically generated

Figure . Distribution of Weather Conditions in Sample

Like the comparison of the graphs above, all the sample and population distribution graphs were compared and were found to have a similar distribution. This meant that the sample represented the population as much as possible and there was no loss in data.

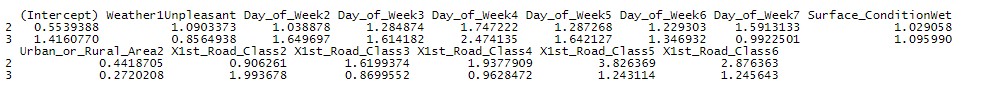
1. Creating data partition

The data was then divided into training and testing data, with 75% of the data sampled into training data and the rest 25% was sampled into testing data.

**Models**

1. Multinomial Logistic Regression

The exponent of coefficients for the multinomial regression model are as follows:



The accuracy for the training model was 46.04%. The confusion matrix is given below:

Text

Description automatically generated with medium confidence

The accuracy for the testing model was 44.53%. The accuracy is not very different from the training model, which means that the data has neither been over fitted nor under fitted. The confusion matrix is given below:

Table

Description automatically generated

1. Decision Tree:

A plot of a pruned decision tree is given below. The structure implies that urban or rural setting is the root node, while the others are internal nodes. Some interesting insights here are, during a weekend if the weather is unpleasant, the accident severity is either medium or major, whereas on the weekdays, unpleasant weather causes minor accidents.

Text, letter

Description automatically generated

Diagram

Description automatically generated

After conducting three-fold cross validation for different confidence parameters, the best model has an accuracy of 44.53% on the testing data. The top ten models are given below:

Table

Description automatically generated

The three-fold confusion matrix for the best model is:

A picture containing text, receipt, screenshot

Description automatically generated

1. Naïve Bayes:

Using three-fold cross validation, we find the best model for Naïve Bayes which loops through various parameters.

A picture containing text, receipt, screenshot

Description automatically generated

The model shows an accuracy of 45.6%, which is only slightly greater than our other two models.

**Model Evaluation**

The accuracy of the models are given below:

|  |  |  |
| --- | --- | --- |
| Model | Training Accuracy | Prediction Accuracy |
| Logistic Regression | 46.04% | 44.53% |
| Decision Tree | 45.33% | 44.36% |
| Naive Bayes | 45.6% | 42.22% |

**Discussion**

Some interesting insights from the models are:

1. Thursdays and Fridays increase the chance of a medium to major accident as compared to a minor one.
2. The likelihood of a minor accident in a rural area is more, which could be because of lesser traffic and more open spaces in rural areas to swerve and avoid major collisions with either other vehicles or road-side objects.
3. In urban areas, during unpleasant weather, the likelihood of a medium to major accident is very high.
4. Road Class 1,2,3, have higher likelihood of minor accidents as compared to major ones.
5. Road Class 4,5 and 6 show increased likelihood of severe accidents.
6. It seems that the road surface conditions do not play a large role in determining the severity of the accident. It is equally likely that the accident could be any of the levels of severity irrespective of the road surface conditions.

**Conclusion**

The accuracy seems to not improve even on tuning the parameters for all the models, which could mean that there are possibly some other factors that play a role in determining the severity of a vehicular accident.

Having determined which classes of roads demonstrate higher chances of accidents, more traffic rules could be enforced specific to the road classes. On particular days of the week, higher traffic patrolling could be done to avoid rash driving on the people’s part. Urban areas could use a lot more traffic law enforcement if a lot of severe accidents occur in those areas.

Even though accidents should be avoidable, it seems perplexing that there is still a lot of noise, which is making it difficult to assess what are the factors that determine accidents exactly. Further information about whether a person of a particular age group is more likely to drive carelessly, if inebriation is one of the reasons for accidents, or the roads genuinely had blind spots could further better the models’ performances.

Even one life lost in a vehicular accident is too many and therefore we need more data to determine the cause leading to these unfortunate incidents. Future work for such a project includes more data collection and building models to incorporate that data to determine if they prove to be helpful in the predictions or not.

**References:**

1. <https://www.gov.uk/government/statistics/reported-road-casualties-in-great-britain-annual-report-2018>
2. https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales?select=accidents\_2005\_to\_2007.csv
3. <https://en.wikipedia.org/wiki/List_of_accidents_and_disasters_by_death_toll>