SG - UK Road Accidents – Report

Contents

Introduction

In 2018 (the last year for which there are complete annual records), there were 1,784 reported road deaths and 25,511 serious injuries in car accidents in the United Kingdom. This level is similar to that of all 6 years since 2012 and, although there was a decline in the first decade of the 21st century, deaths have stabilised around this mark.

Although the decline that occurred from 2000 to 2012 is overwhelmingly positive, 1,784 road deaths and 25,511 serious injuries is still too many. These incidents cause untold emotional damage to those surrounding the casualties, as well as a significant amount of damage to the economy of the United Kingdom. Indeed, a report from Fonesca Law (a solicitors firm in the UK) in 2014 (when road deaths were roughly the same as they were in 2018) argues that road accidents cost the national economy almost £15 billion, partially due to vehicle and property damages, police costs and insurance costs.

The aim of this project, therefore, is to perform data analysis and visualisation which indicates the features which make a road accident more likely, and to build a set of machine learning models, which can predict accident severity.

Having observed the analysis and used the model, it is hoped that drivers will consequently be able to change their driving habits accordingly, and thus be less likely to have an accident. If there is sufficient uptake of the analysis and the model, it might be possible to reduce the number of road deaths and road injuries per year by a significant margin. However, even if just a few people observe the data analysis and use the model, and change their behaviour, leading to higher levels of safety on the UK's roads, this would be justification enough for this project.

Data

The Nature of the Available Data

Data on the subject of road accidents is readily available through the UK government. Indeed, each year the Department for Transport releases information on road accidents and road casualties, downloadable in CSV format. Each annual accident dataset tends to include around 120,000 observations (rows) and exactly 32 attributes (columns). For each accident that occurs, the Department records location, accident severity, the number of vehicles involved, the number of casualties, the day of the week, the time, the road classes (i.e. motorway, A-road etc.) of each road involved, road type (roundabout, one-way street, dual

carriageway etc.), the speed limit, details regarding the junction, the type of controls at the junction (traffic lights, none etc.), the light conditions, the weather conditions, the road surface conditions, whether there were any hazards on the carriageway, and whether the accident occurred in an urban or rural (or unassigned) area. The 120,000 observations from a single year should be more than enough for effective data analysis and the construction of a good machine learning model. The features which I think will be most telling are speed limit, light conditions, junction details, weather conditions, urban or rural area and road surface conditions.

Though the data is readily available, it is formatted in such a way that is not conducive to immediate analysis. There are many missing values (coded as various things, from -1 to 6 depending on the attribute), variables whose datatype is not conducive to analysis, unnecessary attributes (such as 'Did_Officer_Attend_Scene_of_Accident') and variables which should not be used because they might unfairly affect the model (for instance, using 'Number_of_Casualties' as a predictor for accident severity would be unwise, since it is something which one cannot know before one actually has an accident, and because it would lead to a less useful model). Consequently, a large amount of data pre-processing will have to occur before anything can be done.

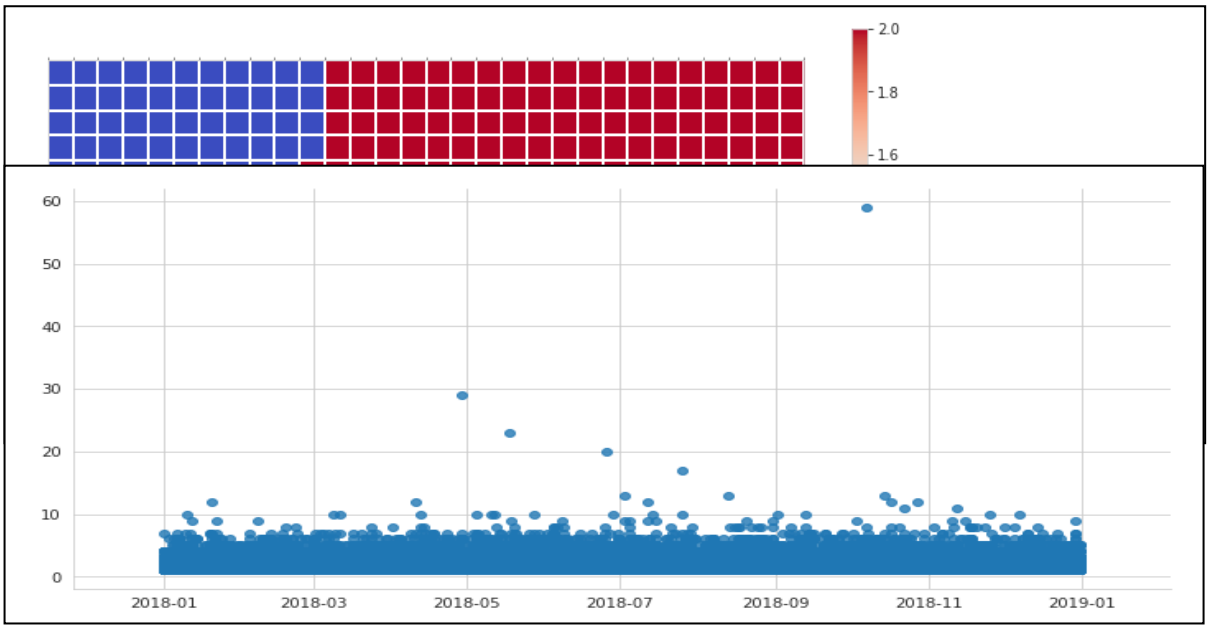
That said, I am optimistic about the results of using this dataset, and I look forward to observing the results of this project.

Methodology

The project used data visualisation and analysis techniques (bar charts, scatter plots, waffle charts etc.) on the Speed Limit, Lighting Conditions, Date, Urban or Rural Area, Number of Vehicles, Day of Week, and Road Conditions features. Full details of the methods of this project can be found in the Github repository for the project, at github.com/shayg1019/Coursera_Capstone.

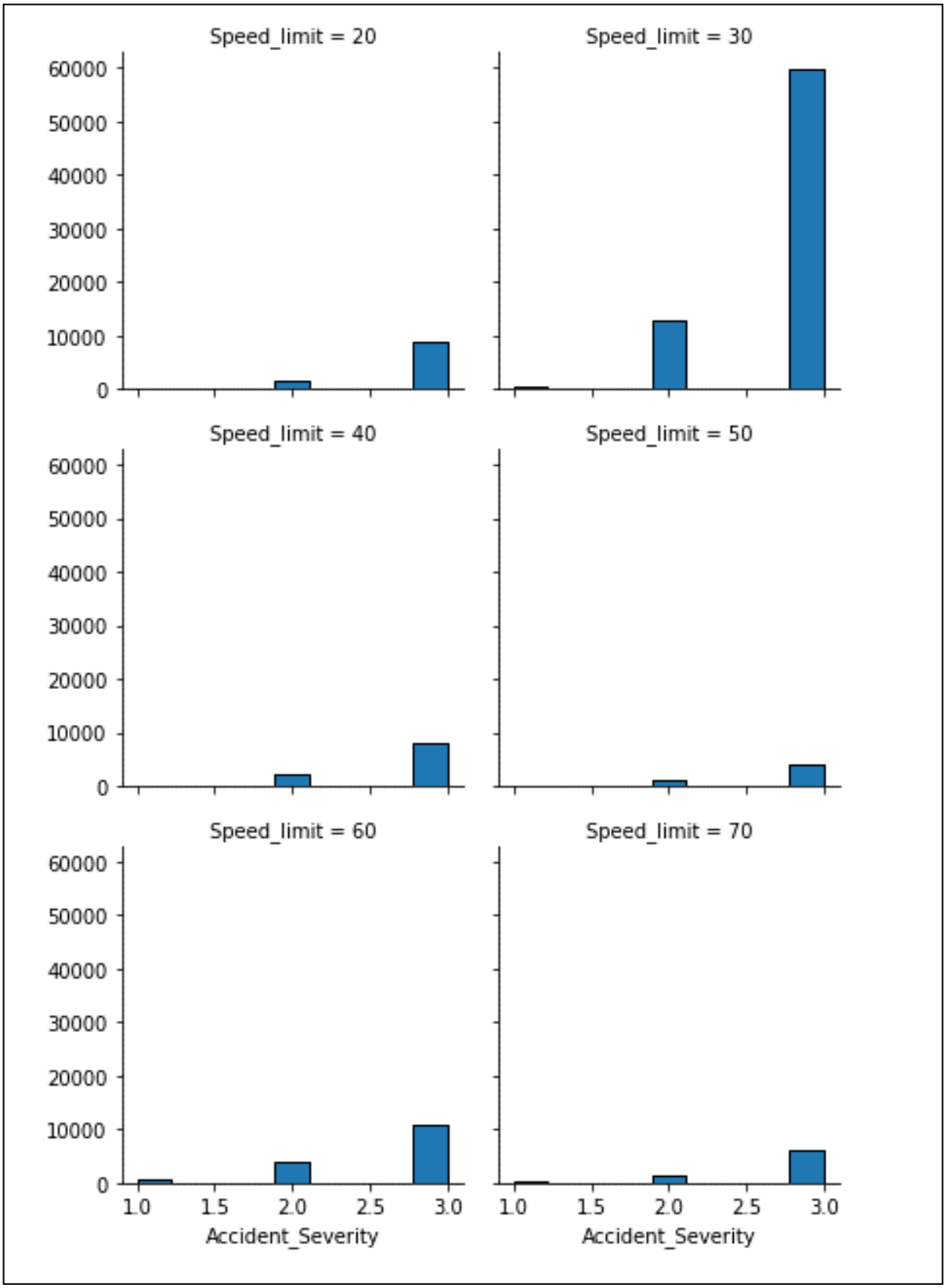
The charts themselves can also be found at the aforementioned Github repository, but just to give the reader an impression of what they can expect, a few charts can be found below.

Below is a waffle chart for accidents in urban and rural areas, in the UK area Warwickshire.



Above is a scatter plot of number of accidents against date.

Below is a series of bar charts of speed limit against accident severity.



Analysis of this, and other data analyses and visualisations, can be found in the results section of this report.

The project also involved the creation of various machine learning models, which predicted ccident severity by analysing Number of Vehicles, Day of Week, Road Type, Speed Limit, Junction Details, Light Conditions, Weather Conditions, Road Surface Conditions and Urban or Rural Area. Once again details of the methods of this project can be found in the Github repository for the project, at github.com/shayg1019/Coursera_Capstone.

The machine learning techniques used were K-Nearest-Neighbour, Decision Trees, and Support Vector Machine. The results were positive, with the models attaining high accuracy scores with each evaluation. Again, for full details, go to github.com/shayg1019/Coursera_Capstone.

Results and Discussion

The results of the data analysis and visualisaton, and of the modelling, are promising; they will be summarized and discussed here.

These results reveal several changes which drivers can make in order to reduce their risk of having an accident, or a severe accident, on the road in the UK.

1. Results of the Speed Limit Analysis

This analysis showed that one is most likely to have an accident with slight injury in a 30 mph area, and most likely to have a fatal accident in a 60 mph area.

It interestingly showed that you're more likely to have a fatal accident in a 30 zone than a 40 zone; in analysing this, we must remember that this dataset includes pedestrian deaths, and that most 30 areas are in residential zones, where pedestrians may be present, while most 40 zones are on A Roads and B Roads, where they are less likely to be.

Drivers should be careful in 30 zones for obstacles, and in 60 zones for serious hazards. Interestingly, the fact that there were fewer accidents in 70 zones than 60 zones would imply that fast a-roads (with worse maintenance and more bends) are more dangerous than motorways.

2. Results of the Lighting Conditions Analysis

This analysis showed that most accidents occur in daylight (1). This is unsurprising, as it seems heuristically sensible to assume that most car journeys take place in daylight. More interestingly, it seems that very few accidents take place in darkness, with lights unlit (5). This would suggeset that the British public tends to abide by this particular road safety guideline very well. There are plenty of accidents which occur in darkness, with lights lit (4) (this correlates with the notion that many accidents occur in 30 mph zones, where street lights are in place), and fewer in darkness with no lighting (6). Perhaps this reflects the notion that people are happier to drive at night in urban areas (which will be lit) than rural areas (which will not be lit).

Drivers should not assume that driving in the day precludes you from having an accident, and should be aware of the relative hazards of driving in the dark.

3. Results of the Date Analysis

There doesn't seem to be a noticeable correlation between accident severity and date at this level.

This would suggest that drivers needn't worry too much about the time of year when they set out (though they should worry about other factors which relate to time of year, such as weather conditions).

There are individual days when there are few accidents (e.g. Christmas Day), but these are outliers.

4. Results of the Urban or Rural Area Analysis

This analysis would suggest that more accidents occur in urban areas than rural areas by a significant margin. The gap in slight accidents is large, but the gap in serious accidents and fatalities is much lesser. This is likely to be reflecting higher speeds in rural areas.

Drivers ought to be careful in urban areas, and not get complacent about the low speed limits and street lighting.

As one can see, in Warwickshire, a county with some urban and some rural areas, the skew is not towards urban areas, but rural areas.

This result, in comination with the bar chart earlier, would suggest that people cannot be complacent in rural areas, and should be looking out for different risks depending on where they live.

For Warwickshire, and most likely other midlands counties with similar structures, like Gloucestershire and Worcestershire, you're more likely to have an accident in a rural area, but can't take your eye off the ball in urban areas either.

5. Results of the Number of Vehicles Analysis

This analysis shows that a lot of accidents occur involving 2 vehicles, and that the next most accidents occur with just a single vehicle.

This is an interesting discovery. It implies two major things. Firstly, the fact that a very large number of accidents are slight and between two vehicles suggests that a lot of these indicents might involve one car rear-ending another (i.e. a fender bender). This should be noted when one reads accident data, and accounted for, when one drives. Secondly, the large number of injuries involving one vehicle implies that drivers should be very careful even when the road is empty - lapses in concentration, lost control etc. can lead to a single vehicle having an accident a lot of the time.

6. Results of the Day of Week Analysis

The most interesting thing which this shows is that there are fewer accidents on days 1 and 7 (Saturday and Sunday) i.e. the days when people are not going to work. This would suggest that driving in a rush (derived from needing to get to work, or trying to get home asap) and driving on days when there are more people on the roads, can be more dangerous.

Drivers should be particularly cautious on weekdays.

7. Results of Road Conditions Analysis

The most important takeaway from this visualisation is the fact that dry conditions (1) do not protect you from being in an accident!

The results of the model are also very promising. All three are very effective at predicting accident severity, as the strong jaccard and F1 scores for the KNN model, the strong F1 and jaccard scores for the Decision Tree model, and the strong F1 score for the SVM model all show. The specific scores can be seen at the aforementioned GitHub link.

The fact that these scores are all strong shows that data pre-processing was effective, and that the large number of features used allowed for a good set of models.

It also shows that if you know these features (which you can before you set out on a drive), you can use the models to see how severe an accident you might have, if you were to have an accident, which could feasibly be very useful.

Conclusion

The aim of this project was to accurately predict the severity of an accident given a number of features, and to discern a number of ways, through data analysis, in which drivers can change their behaviour to be safer on the roads. The project has fulfilled both of these aims.

In closing, it seems most apt to provide a concise set of findings, which can be easily read by drivers, and used to change their behaviour:

1. Be cautious of obstacles and hazards in 30mph zones
2. Drive carefully on A and B roads with 60mph speed limits - they're more dangerous than motorways!
3. Driving in daylight does not preclude you from having an accident, but you should be careful at night.
4. You should be careful in both urban and rural areas, though your likelihood of having an accident in either depends on where you live.
5. A large number of slight accidents occur with 1 or 2 vehicles - be aware of other cars on the road and be careful even when the road is empty - lapses in concentration, lost control etc. can lead to single vehicle accidents frequently.
6. Drivers should be particularly cautious on weekdays
7. Dry conditions do not preclude you from having an accident.