

IBM Data Science Specialization
Capstone Project

Predicting Road Accident Severity Using Machine Learning Models

A Case Study on Road Accident Statistics in the
United Kingdom

Zoltan Hradszky

2020

Contents

Contents	i
Figures	i
I. Business Understanding	1
Global Context and Relevance	1
Case Selection	1
Application and Target Audience	1
II. Data Understanding	2
About the Dataset	2
Data Characteristics	3
Casualty Severity	3
Accidents in Rural and Urban Areas	4
Accidents by Road Type	5
Casualty Type	5

Figures

1. Figure: List of Variables	3
2. Figure: Distribution of Casualty Severity	4
3. Figure: Accident severity in rural and urban areas	4
4. Figure: Accident severity by road type	5
5. Figure: Severe or fatal accidents among different road users.	6
6. Figure: Distribution of accident severity by casualty types	6

I. Business Understanding

Global Context and Relevance

Road accidents have been in the spotlight due to the increasing number of global traffic accidents recorded each year. According to the World Health Organisation (WHO), there were 1.35 million road traffic deaths in 2016 and road traffic injuries became the leading cause of death for children and young adults (aged 5-29) in recent decades. This shows that it is crucially important to study the factors leading to severe or fatal road accidents and having a better understanding of the circumstances may allow to further improve preventive measures and develop new safety technologies.

Case Selection

The United Kingdom (UK) ranks among the top countries for road safety with the second lowest road accident fatality per million inhabitants in the European Union. (European Commission, 2017) There were more than 122 thousand road accidents recorded in 2018 with 25,511 reported severe injuries and 1,784 deaths in the UK. Until 2010, fatality rates had been constantly declining and were stabilised on its current rate in the past decade. According to the Department of Transport, extreme weather events have slightly influenced the annual figures, but the overall trends have not changed significantly in this period.

The Department of Transport claims that the highest fatality rate is observed among vulnerable road user groups - namely pedestrians, pedal- and motorcyclist, as they are more exposed in an event of colliding with other vehicles or objects. Although, there is no single underlying factor that drives road casualties, instead there are several influences can be observed. These are mostly distance travelled per driver, the mix of transport modes used, driver's behaviour, the mix of groups using the road (young inexperienced and older drivers) and external effects, such as weather and road conditions.

Application and Target Audience

With each accident being unique and several factors are hard to measure (e.g. fatigue) it is challenging to develop a generalised machine learning (GML) model that can predict accident severity with a high precision. However, this project may demonstrate a GML model that can predict road accident severity based on a number of features, including weather, road

conditions, driver characteristics' and the type of road users involved in the accident. The model can be used to develop an early warning system for road users, which could be of high interest for government officials, local authorities (hospitals and law enforcement), as well as for car manufacturing companies wishing to further develop the drivers' assistance features of their models.

II.Data Understanding

About the Dataset

For this project the latest road accident and safety statistics is used from 2018, covering the territory of Great Britain. The dataset is published annually by the Department of Transport, obtaining the data through the reporting system used by local authorities and also self-reported data by drivers. In the case of severe and fatal accidents, most of the data are recorded by authorities, while slight accidents are usually reported by the drivers.

Considering that the database includes various features regarding the environment, drivers' and casualties' characteristics¹, it is suitable to develop a machine learning model. It provides a comprehensive description of the circumstances, making it fit to examine the relationship between the selected predictors and the target variable.

The data is published in three datasets annually:

1. The first dataset describes the accidents, the number of casualties and in terms of our target variable, which is accident severity, it records, whether at least one party was severely injured or died. In cases, when no such outcome was observed, it is indicated as slight injury. The dataset also doesn't include accidents with property-damage only, only ones, where at least one party was injured.
2. The second dataset describes the driver's behaviour, age, gender and the vehicle's manoeuvre and condition. Unfortunately, there is no record of substance abuse or exceeding the speed limit for 2018, which could significantly improve the accuracy of the model.
3. The third dataset describes the casualties, their mode of transport, as well as their behaviour during the accident (e.g. crossing road, loitering on central carriageway refuge, etc.).

¹ The variable lookup is available through the following [link](#).

The final dataset used for the analysis is merged from these three datasets and the selected variables are transformed into binaries.

Data Characteristics

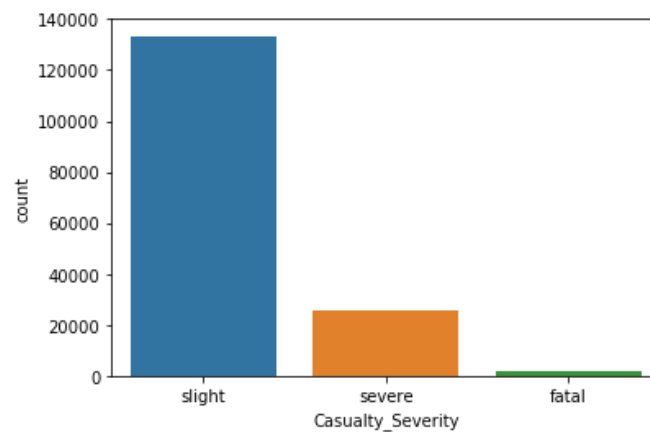
Variable name	Type	No. of Categories
Accident Index	Unique Identifier	–
Vehicle Type	Nominal	20
Vehicle Leaving Carriageway	Nominal	9
Age Band of Driver	Nominal	11
Sex of Driver	Nominal	3
Pedestrian Location	Nominal	11
Accident Severity	Nominal	3
Number of Casualties	Numeric	–
Number of Vehicles	Numeric	–
1 st Road Class	Nominal	6
Road Type	Nominal	7
Speed limit	Numeric	–
Junction Detail	Nominal	9
Light Conditions	Nominal	5
Weather Conditions	Nominal	9
Road Surface Conditions	Nominal	7
Urban or Rural Area	Nominal	3
Casualty Type	Nominal	21

1. Figure: List of Variables

Figure 1 lists the variables, which are used for the model development. Most of the variables are on a nominal scale, where each number corresponds to a class. The difference between each value cannot be interpreted on a nominal scale, therefore, they need to be transformed into binary variables. In order to determine the classes, the one hot encoder was used, and several categories were merged, when they shared certain characteristics. For instance, in the case of vehicle classes, motorcycles are distinguished in 5 categories, so they were merged into one. In other cases, only some classes were kept as binary variable to describe the accident and the others were simply dropped (e.g. vehicle leaving carriageway).

Casualty Severity

The target variable is casualty severity, which is split into three categories – slight, severe and fatal. Figure 2 below shows the distribution of these categories from the casualty dataset.

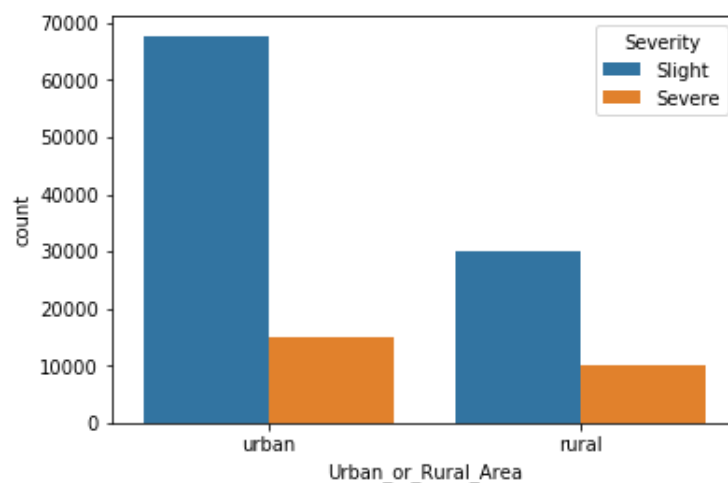


2. Figure: Distribution of Casualty Severity

Having a peek at the above figure, it becomes visible that the distribution of accident severity is extremely skewed. This is also referred to as Poisson distribution, when one category has a high number of observation and the frequency of observation rapidly drops in further categories.

Since there are relatively few fatal accidents observed in the dataset, it is reasonable to merge severe and fatal categories into one category. This will allow better visualisation and a more balanced sample for further statistical analysis. Also, having a binary target variable is more suitable to be used for the selected GML models.

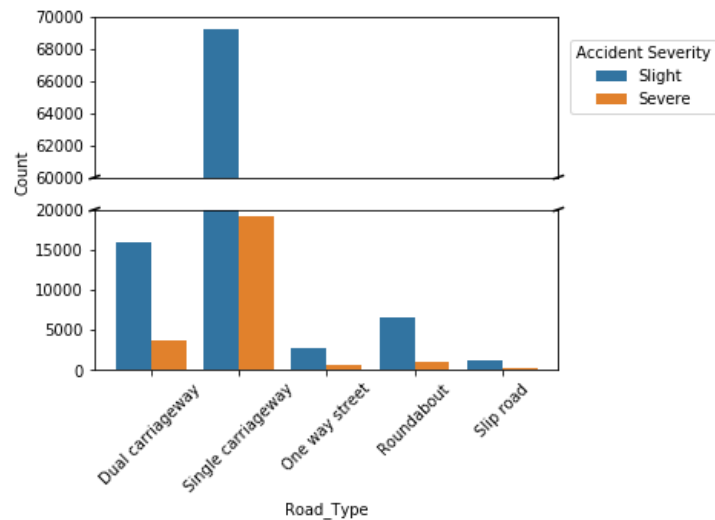
Accidents in Rural and Urban Areas



3. Figure: Accident severity in rural and urban areas

The figure 3 shows that while there are more accidents in urban areas, they less likely have a severe or fatal outcome than the ones in rural areas. This can be explained by higher traffic, but lower speed limits. In rural areas due to lower traffic, cars may less likely collide, but there will be a higher chance of sever or fatal outcome due to higher speed limits.

Accidents by Road Type



4. Figure: Accident severity by road type

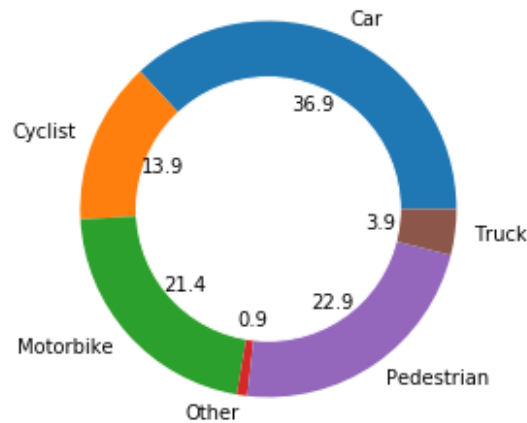
The above figure shows that single carriageway roads have the most accidents observed. However, road type is not necessarily a good predictor per se, but it might be fit as joint predictor when examining its relations with other variables. For instance, if a car encounters an unexpected event (e.g. the car slip on the slippery road) it might more easily hits an object or another car in the opposing traffic, as there are no physical barriers to stop it.

Also, it should be noted that most roads in the urban setting are single carriageway roads, therefor the sample may be biased. But if the relationship between the road type, weather conditions are measured in a rural area, it might provide a more accurate description of a severe or fatal accident.

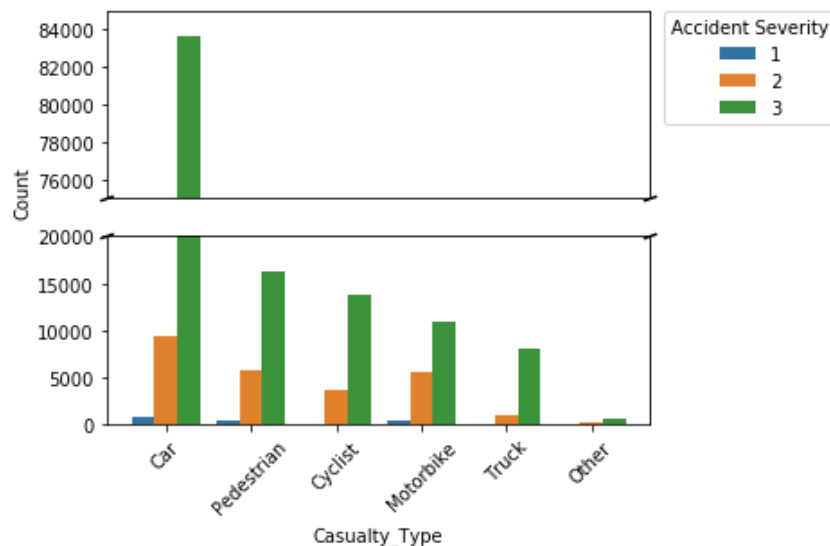
Casualty Type

As other sources claim, vulnerable road users are the ones, who will most likely have serious injuries, when they collide. A pedestrian hit by a car may more likely to have severe injuries than the car driver and a motorcyclist may also have a severe accident, if they hit an object or another car with high speed.

Severe casualties in different road user groups



5. Figure: Severe or fatal accidents among different road users.



6. Figure: Distribution of accident severity by casualty types

Figure 5 shows that the overall distribution of severe and fatal accidents among different road users. It is visible that among fatal and severe accidents, car passengers were exposed the most to such outcomes. However, figure 6 shows that there is a significantly higher number of observations among car drivers and overall, car passengers are less likely exposed to severe or fatal accidents than vulnerable road users.

Conclusion

While evidence shows that road accidents cannot be defined by a simple set of characteristics, some variables tend to show a greater impact on the target variable than others. The most significant variables so far are area type – urban or rural area – and the

participation of vulnerable road users in an accident. It is important to note that none of these conditions can determine the outcome of an accident per se, but correlating them with other factors may provide more precision and helps us to develop an accurate model.