Capstone Project Report

Road Users Protection Improvement

Problem statement

In UK, over 1700 people were killed by traffic accidents every year, and hundreds of thousands injured. These accidents have torn families apart, troubled the communities, and damaged properties. There are many causes to traffic accidents such as bad road using habits or unsafe road design. It is everyone's responsibility to keep the road safe.

The UK Department for Transport revealed that the number of casualties remained broadly consistent in the last 10 years, except for a drop in 2020 due to lockdown for the coronavirus pandemic. The Department also discovered that the number of accidents changed along with traffic volumes.

In this project, we will look into different factors of the road, vehicle and casualty in order to find a pattern of the accidents. Then we can use that information to develop strategies to

1. reduce the number of accidents
2. protect road users when an accident happens

Stakeholders

The Department for Transport has the biggest responsibility to road safety. In addition, there are currently 24 agencies and partners to support the UK transport network. They can improve road safety by infrastructure and transport network design.

Business question

The business questions that need to be answered are:

- What are the common factors that lead to road accidents?
- What is the value of saving lives?
- How to reduce the number of accidents?
- How to reduce the accident severity?

Regarding the value of saving lives, economists have already figured out the answer. According to Viscusi (2003) the Value of Statistical Life (VSL) in UK was $4.2 million US dollars in 2000. After taking inflation into account, that VSL is over $11 million US dollars nowadays (1).

$$4.2 \times (inflation\ in\ UK) \times (increase\ of\ US\ dollar\ purchasing\ power) \quad (1)$$
$$4.2 \times 1.72 \times 1.58 \approx 11.41$$

Data question

- Relationship between each factor and: Number of Road Accidents
  Casualty Severity
- What roads prompt to severe accidents?
- What types of road users are more likely to get into an accident on those roads?

To answer these questions, we looked into the data of:
Accident Detail
Casualty Detail
Road Detail
Pedestrian Crossing Detail
Light Condition
Weather Condition
Road Surface Conditions
Special Conditions at Site
Vehicle Detail
Driver Detail

The model should produce a better result in predicting casualty severity than baseline accuracy.

Data

Datasets:
Road Safety Data - Accidents 2019
Road Safety Data - Vehicles 2019
Road Safety Data - Casualties 2019
Source: Department for Transport
Link: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data
Volume: 117536 accidents, 71 Attributes
Reliability: Very reliable
Quality: High
Data Generation: Info gathered by police
Ongoing basis: No. New report published in every September

Data science process

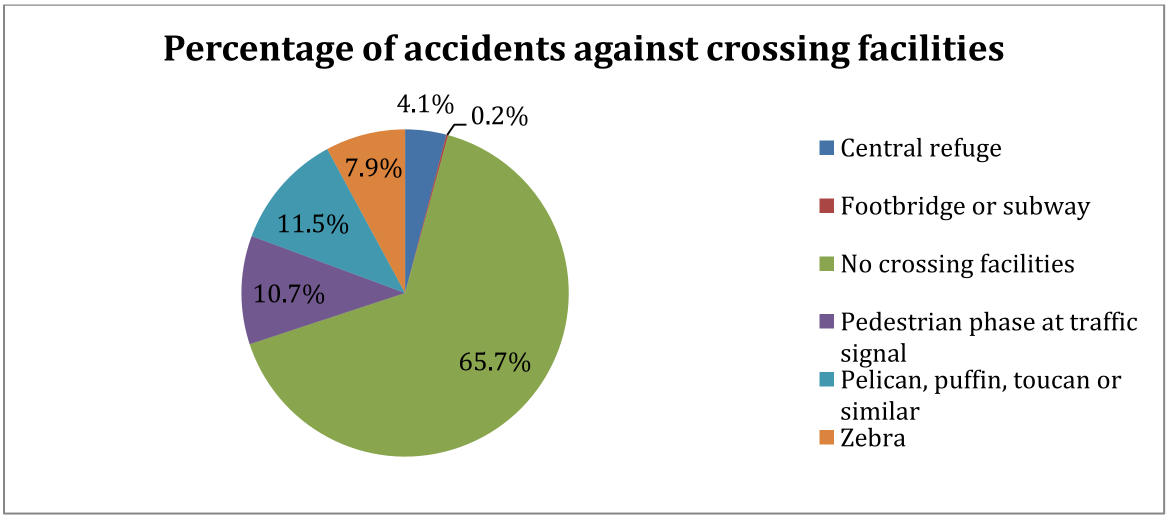
Data analysis

Firstly, we searched for null values in the datasets. No null value was found since the datasets used “-1” to represent missing or out-of-range data. Then we searched for the number “-1” in the datasets, and found that the column “Junction_Control”, represented the safety measure at a junction, had a significant amount of “-1”. To avoid losing too much data, we assumed that all the “-1” under this column were uncontrolled junction “4”. After that, only 7.6% accidents contained the value “-1”, which were acceptable to be discarded.

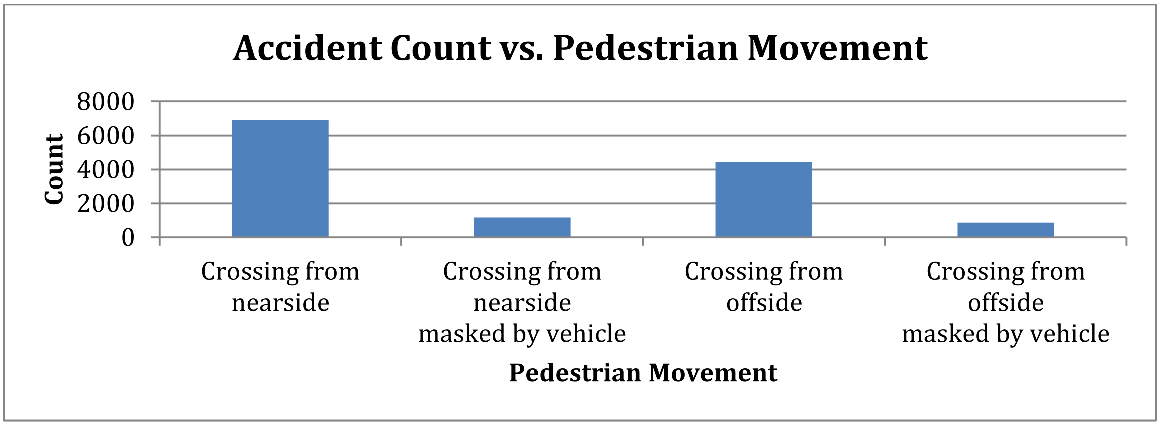
Since the three datasets are relational, we created a SQL database for Exploratory Data Analysis.

The highlights of the Exploratory Data Analysis (EDA) are:

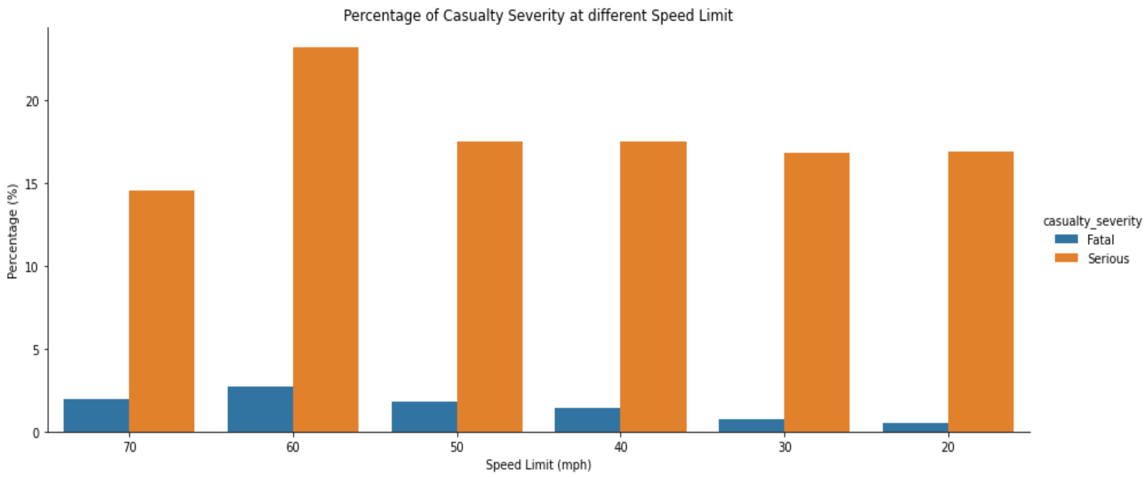
- Most accidents happened on urban major and unclassified roads, corresponding to the finding “number of accidents changed along with traffic volumes”.
- Most accidents happened on roads with speed limit 30 mph (~50 kph). Other roads had similar number of accidents.
- ~43% accidents happened not at a junction, while ~28% happened at T or Staggered junction. Crossroads and Roundabouts combined was ~15%, and the rest were fairly low. Most of the junctions had no control or controlled by give way signs, only ~18.4% were controlled by traffic lights.
- Most accidents happened at / near a junction when the vehicles were approaching / waiting near the junction, or in the mid-junction. Fewer accidents happened when the vehicles were entering or leaving a junction.
- Most casualties were not pedestrian, although pedestrian had higher risk of getting serious or fatal damage.
- Using crossing facilities can greatly reduce the chance of having an accident. Footbridge and Subway provide the best protection, with an accident rate of only ~0.2%. All other crossing facilities had ~10% or less accident rate, although Central Refuge and Zebra Crossing were slightly better than Traffic Lights and Pelican Crossing.



- Pedestrian crossing the road from driver's nearside was more dangerous from offside, but not very significant. However, when pedestrians were masked by vehicles, the accident number was much lower. It could be because the pedestrians and drivers were more alert near parked vehicles.



- When pedestrians were not crossing, even they were standing / playing on the road or walking along the road, the chance of road accident was still quite low (less than 10% combined).
- Weather and road surface condition did not seem to be major factors in road accidents; most accidents happened in good condition (daylight, fine weather, dry road). The weather condition might have influence on traffic volume, which affected the accident rate.
- Speed Limit directly affected the rate of having serious / fatal damage during an accident. When speed limit was higher, fatal rate was also higher. If we reduce the speed limit from 60 mph (~96.6 kph) to 50 mph (~80 kph), the chance of severe casualty will be lowered by ~6.6%.



- Car Occupant, Cyclist and Pedestrian were the top three road users in accidents. About 57.4% accidents involved car occupants, but they were mostly slight injuries (~87.3%). About 14.2% accidents involved pedestrians, and they had the highest rate of serious and fatal injuries (~30%). About 10% accidents involved cyclists, and their rate of serious and fatal injuries was ~24.2%, better than pedestrians.

Road User	Percentage of Accidents	Serious / Fatal Casualty Rate
Car Occupant	57.4%	12.7%
Pedestrian	14.2%	30%
Cyclist	10%	24.2%

The EDA pipeline is reusable. This is achieved by creating a SQL database and renaming the columns to maintain consistency. Some functions for data visualisation were created to reduce the size of the script. If new codes are added in the future dataset, we only need to add the new codes to the lookup variable.

Modelling

Since we would like to study what roads prompted to severe accidents, we used Road Class, Road Type, Speed Limit, Junction Detail, Junction Control, Crossing Facilities, Light Conditions, Weather Conditions and Road Surface Conditions as main features for the model. All those features were independent to one another and had no interaction between them, except Weather and Road Surface condition which did not play an important role.

In order to determine the importance of each feature, we converted the data back to numerical code. Then we used Forward Feature Selection to pick the suitable features for the model. Within all the features, Speed Limit played the most important role in determining Casualty Severity and Road User Type.

This problem was a supervised classification (multi-class) problem, therefore we used Decision Tree, Random Forest, Extra Trees, Ada Boost, LightGBM and K-Nearest Neighbors to learn the data. Their performance is shown below.

Baseline accuracy is 79.56%

	Accuracy (%)	Process Time (s)
decision tree	79.80	0.009
random forest	79.80	1.138
extra tree	79.80	0.426
ada boost	79.80	2.084
light gbm	79.80	0.573
k-nearest neighbors	79.80	3.929

Unfortunately, the models could only predict slight casualty due to the overwhelmingly large amount of it (~80%). No model could do better than baseline accuracy.

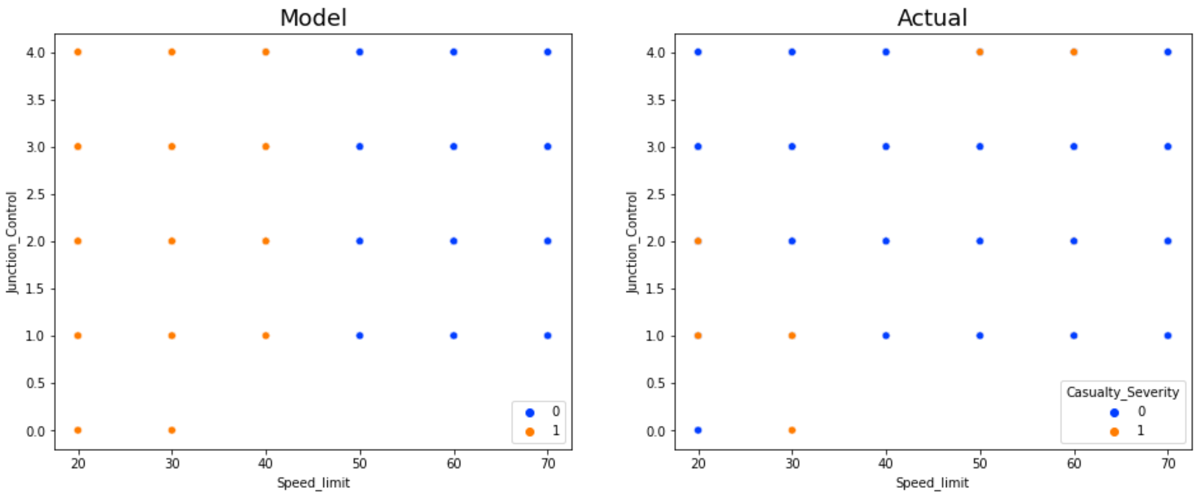
We also predicted what types of road users are more likely to get into an accident according to the road.

Baseline accuracy is 49.84%

	Accuracy (%)	Process Time (s)
decision tree	51.31	0.083
random forest	51.46	2.651
extra tree	51.38	0.793
ada boost	33.43	7.188
light gbm	51.47	5.577
k-nearest neighbors	51.08	2.924

The models could only get half of the prediction right, around 1.5% better than baseline accuracy.

Since the supervised methods could not provide a good result, we tried an unsupervised method, K-means Clustering. Casualty Severity groups were reduced to two, where serious and fatal were combined. The model split the data into two groups merely by Speed Limit, which was not accurate. Although it could successfully reflect that Speed Limit had the greatest influence on Casualty Severity.

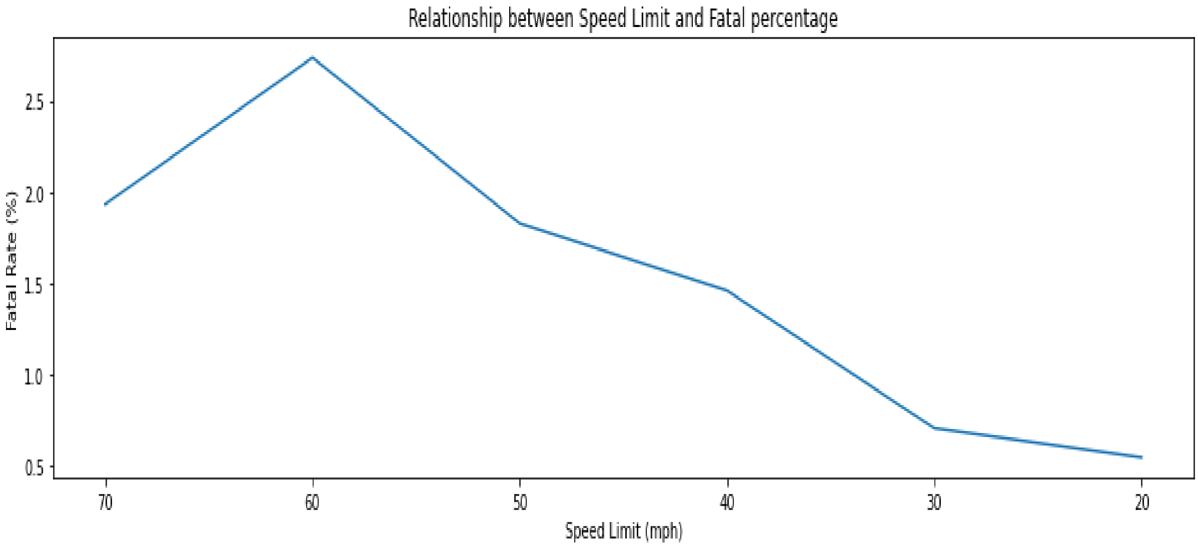


Outcomes

The most important finding of this study is that road accidents mostly happen under normal condition, which means that it is mainly affected by road users' behaviour. When the road users are alert, the number of accidents is lower.

Since pedestrians are the most vulnerable road users, they should always use crossing facilities if possible. Footbridge and subway are the best crossing facilities, while others can also greatly reduce the chance of accident.

Speed Limit has the highest influence on the casualty severity. Reducing speed limit can significantly reduce the rate of serious and fatal damage in an accident.



Data answer

We successfully analysed how different factors affected the number of accidents and casualty severity. However, we were unable to identify what type of roads would lead to more accidents and severe casualties based on this dataset.

Business answer

Now we have evidence that Speed Limit and Crossing Facilities are important to road safety, we can develop improvement plans based on these factors to reduce number of accidents and casualty severity. The improvement will be significant according to the data.

Response to stakeholders

We should make sure that there are sufficient crossing facilities for pedestrians in order to reduce the number of accidents. Therefore our next step will be investigating the crossing facilities condition, distribution and arrangement to find out what we can improve.

Variable Speed Limit (VSL) can be a good way to improve road safety. A VFL is a flexible speed restriction on a road according to the current environmental and road conditions and is displayed on an electronic traffic sign. An example is the South Eastern Freeway in Adelaide, where speed limit is reduced when there is high traffic. By reducing speed limit from 60 mph (~96.6 kph) to 50 mph (~80 kph), the chance of severe casualty will be lowered by ~6.6%.

Lastly, education is also important to improve road users' habits on the road. Posters and public transport advertisement can be a good reminder of appropriate manner to road users.

References

Journal:
Viscusi, W & Aldy, J 2003, ‘The Value of a Statistical Life: A Critical Review of Market Estimates Throughout the World’, *Journal of Risk and Uncertainty*, vol. 27, no. 1, pp. 5-76.

Tools:
Jupyter Notebook
Python 3
SQL
Microsoft Excel

Python Libraries:
numpy
pandas
sqlite3
sqlalchemy
matplotlib
seaborn
time
sklearn
lightgbm