From Data to Insights: A Comprehensive Analysis of Traffic Accident Data in the UK (2000 - 2022)

Louise Malm & Nikolay Moshenskiy
GitHub Repository

Task 2 - Business understanding

1. Identifying business goals

1.1. Background

In the UK, road accidents cause several hundreds of deaths and tens of thousands of severe injuries annually. This project revolves around the analysis of UK road accident data from 1979 to 2022, with a special focus on data between 2000 and 2022. The analysis aims to uncover patterns, trends, and factors contributing to road accidents, especially those leading to fatal outcomes. In addition to analysis, a tool classifying an accident as fatal or non-fatal is expected to provide additional insights into the most significant features contributing to fatal accidents.

In the UK, maintenance of roads is handled by the Department for Transport[1] together with National Highways[2] and the Welsh and Scottish Governments[3,4]. As this project is not related to any business, the aforementioned organizations will henceforth be referred to as stakeholders.

1.2. Goals

Ultimately, the goal for the stakeholders is to enhance road safety and reduce the occurrence of fatal accidents. This is achieved by highlighting critical factors such as light, road conditions, speed limits, etc., that potentially can be improved. Together with the gained knowledge of where and when fatal accidents are more likely to occur, these insights are expected to help identify the risk factors leading to serious and potentially fatal accidents. This will help the stakeholders make data-driven decisions by implementing proper road design and maintenance, as well as pinpointing areas where more resources are needed.

1.3. Success criteria

Success would be evident through the development and implementation of measures addressing the identified risk factors, if these implementations lead to a measurable decrease in the number of fatal accidents. This impact however can only be observed by subsequent analysis of road accident data, taking into consideration natural variation and other factors which may influence the occurrence of fatal road accidents (for example, during Covid-19 pandemic the overall occurrence of road accidents drastically decreased).

2. Assessing the situation

2.1. Inventory of resources

The resources for this project include historical road accident data from the UK, collected between 1979 and 2022, python libraries for data processing (pandas, geopandas), visualisation (plotnine) and machine learning (scikit-learn) as well as QGIS software for spatial data processing.

2.2. Requirements, assumptions, and constraints

The data is used with the assumption that it accurately reflects road safety conditions. However, as old data may not fully reflect current road safety dynamics, the special focus was on more recent years (2000 - 2022). Since the data is publicly available, it already adheres to data privacy and ethical considerations.

2.3. Risks and contingencies

There is a risk of the data being outdated and not representative, thus the contingency plan is to use a smaller subset of the more recent data.

2.4. Terminology

Road accident severity scale:

1: fatal accident (at least one fatality)
2: serious accident (there are severely injured casualties, no fatalities)
3: slight accident (only slightly injured casualties, no severely injured, no fatalities)

Casualties: here casualties refer to victims.

2.5. Costs and benefits

Costs: Time and resources allocated for data analysis.

Benefits: Potentially improved road safety and reduced societal costs associated with road accidents.

3. Defining the data-mining goals

3.1. Data-mining goals

The goals of the data-mining is to (1) identify the most significant predictors of accident fatality, and (2) understand the relationship between various factors and accident outcomes. The first is achieved by training a classifier and extracting the feature importance from it, while the second is achieved by exploratory data analysis of the extracted features among others.

3.2. Data-mining success criteria

The success of the first goal, the classifier, is evaluated based on how well it can predict the correct class (fatal or non-fatal accident). Evaluation metrics include accuracy, AUC, and f-score. The success of the second goal is more difficult to access, but will include identification and pinpointing of factors and geographical areas which should be targeted to improve road safety.

Task 3 - Data understanding

1. Gathering data

1.1. Data requirements

For the scope of this project, data about road accidents in the UK is needed. The dataset should be sufficiently large to be useful for machine learning (i.e., it should contain enough data for model training, validation and testing), and information regarding fatality is required for the classification task. Moreover, the data needs to contain features which may be useful for determining accident severity, such as road type, weather conditions, date and time of accident, geographical information, etc.

In general, for machine learning and to determine patterns, the more data available, the better. At the same time, it is important to ensure that the data carries relevance for the task. Here, we want to gain knowledge needed to improve road safety, thus, the data used should not be too old. Gathering data as .csv or .xlsx files is preferred.

1.2. Verify data availability

Road accident data containing required information is publicly available from the UK government database (https://www.data.gov.uk/). Labelled data, gathered from road accidents spanning the years 1979 to 2022, is available for download in .csv format. This dataset offers a comprehensive overview of road accidents in the UK over the last four decades.

1.3. Selection criteria definition

Three linked datasets containing data about collisions, vehicles and casualties from road accidents occurring in the UK between 1979 - 2022 can be downloaded from https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data.

Together, these datasets provide comprehensive information regarding UK road accidents. However, only data between the years 2000 - 2022 will be used, as the older data may not provide relevant information. Since the data is obtained from a governmental source, we expect the data is reliable and authentic. In addition to the three datasets, a table with metadata is also available.

2. Describing data

The data used here consists of three three tables: collisions, vehicles, and casualties. Here are the specifications for each full dataset, encompassing records from 1979 to 2022:

1. The collision dataset has 8 809 915 rows and 36 columns, with a size of 1.23 GB as .csv file and 2.4 GB as a pandas DataFrame.

2. The vehicles dataset has 15 725 817 rows and 28 columns, and a size of 1.46 GB as .csv file and 3.3 GB as a pandas DataFrame.

3. The casualties dataset has in total 11 713 001 rows and 19 columns, and the size is 0.8 GB as .csv and 1.7 GB as a pandas DataFrame.

The datasets encompass various features, each specific to its corresponding table. For instance, the collisions table comprises features detailing each unique accident, the vehicle table includes features describing each vehicle involved in the accidents, and the casualties table contains features characterizing all the victims in the accidents. The 'accident_index' feature serves as a unique identifier for each accident and can be utilized for table integration if necessary. The majority of the datasets consist of categorical features like 'road_class,' 'vehicle_type,' 'light_conditions,' etc., which are numerically coded and can be translated using the metadata table. Additionally, the datasets include continuous numerical values such as 'age_of_driver' along with time/date values.

Key features that potentially might be the most useful for this project include longitude, latitude, accident_severity, day_of_week, time, road_class, road_type, speed_limit, junction_detail, junction_control, light_conditions, weather_conditions, road_surface_conditions, urban_or_rural_area, vehicle_type, sex_of_driver, age_of_driver, casualty_class, and casualty_severity.

3. Exploring data

Data exploration focused on the subset spanning 2000 to 2022. According to the metadata, missing or out-of-range data is mostly indicated as -1. Consequently, the collision dataset was analysed to determine the percentage of -1 or NaN values. Some NaN values (0.08%) were detected for coordinates (latitude and longitude attributes) and have to be discarded.

Additionally, other attributes showed a higher share of missing values, such as second_road_number (41.18%), junction_control (38.45%), trunk_road_flag (10.40%). Other attributes have a negligibly small number of missing values. Notably, features like the age of the driver and speed limit include non-realistic values. These values were handled by binning. When preparing the data for machine learning, categorical feature values were renamed before removing all unknowns, which reduced the number of records with approximately 83%. However, since the dataset originally was so large, we believe it is still reasonable to remove the records containing unknown or missing values, and that the remaining data will be sufficient for machine learning. Due to the unbalanced data (few fatal accidents compared to non-fatal), pre-processing the data for machine learning will include over- and undersampling strategies.

4. Verifying data quality

Overall, no severe data issues were identified. The initial exploration revealed a large number of unknown values for certain attributes in the collisions dataset. However, since most of the attributes with a large share of missing values are not going to be used in the project, they do not pose any issues.

In addition, for machine learning even after excluding these unknown or missing values, a sufficient number of records (over 600 000) are left.

Task 4 - Project plan

1. Data preprocessing and overview - (4 hours Nikolay, 10 hours Louise):

1. Load data and filter out the years before 2000, analyse the number of null values (unknown) and leave attributes to be used.
2. Preliminary exploration of data.
3. Perform joins, analyse results, remove unmatching rows if needed.
4. Data preparation for task 3: filter out unnecessary attributes, perform joins if needed, do one hot-encoding.
5. Save subsets of data.

2. Temporal analysis - 15 hours (Nikolay):

Analyse data and do the following visualisations:

1. Total number of accidents by:
    1. Year (bar plot)
    2. Month (bar plot with monthly average value)
    3. Day of the week (bar plot with daily average)
    4. Time of the day (bar plot)
    5. Month vs day of the week (heatmap with number of accidents accidents)

3. Feature importance - 14 hours (Louise)

Identifying the most significant predictors of accident severity, by training a classifier and extracting the feature importance from successful model(s).

4. Accidents severity analysis based on feature importance (7 hours each) (Louise, Nikolay):

Visualise accident severity based on the revealed features.

5. Statistical analysis - 6 hours (Nikolay)

F-test and then T-test on rural/urban accidents (comparing accident severity between urban and rural areas.)

6. Spatial analysis - 6 hours (Nikolay):

Finding the most dangerous locations (fatal accidents clusters): Use K-mean clustering to find locations where most of the fatal accidents happened.

7. Writing readme for GitHub - 2 hours (Louise)

8. Making a poster - (12 hours Louise, 6 Nikolay)

1. Selecting the key findings and deciding the content to feature on the poster.
2. Making the poster.

The main data processing and visualisation to be done using the following Python libraries: numpy, pandas, sklearn, imblearn, plotnine, geopandas, shapely, folium. For additional visualisation and data processing QGIS 3.30.2 will be used.

References

[1] Department for Transport https://www.gov.uk/government/organisations/department-for-transport

[2] National Highways https://nationalhighways.co.uk/

[3] Transport Wales https://www.gov.wales/transport

[4] Transport Scotland https://www.transport.gov.scot/