Traffic Accidents in the UK - Severity Analysis using Machine Learning

Introduction

The UK collects a detailed and comprehensive set of data on road traffic accidents, casualties and vehicles involved, and releases it periodically through the Open Data portal(https://data.gov.uk/). This report will review the factors that contribute to the most severe accidents involving fatalities, and present causal models that would allow for the prediction of when, where and how these accidents are most likely to occur, with a view to minimising them in the future. For example, links between fatalities and poor weather conditions could inform changes to government policy and legislation to reduce speed limits in poorer weather, as is seen in continental Europe. The interest in this will be that preventing serious accidents clearly has a positive impact on society, health and wellbeing.

Data

We started by sourcing the 2018 accidents data from the UK open data website, and reviewed the quality, balance and completeness of the data set. The UK data benefits from being nicely formatted and structured, although there are a number of fields with missing data (coded as -1 in the dataset, so not immediately obvious when looking for null or blank records). There are a decent number of variables to construct a model with as well, although intuitively some (weather, road conditions) are going to be more predictive than others (latitude, longitude, road number). Geospatial data will be useful to produce some nice visualisations even if not useful for model predictions.

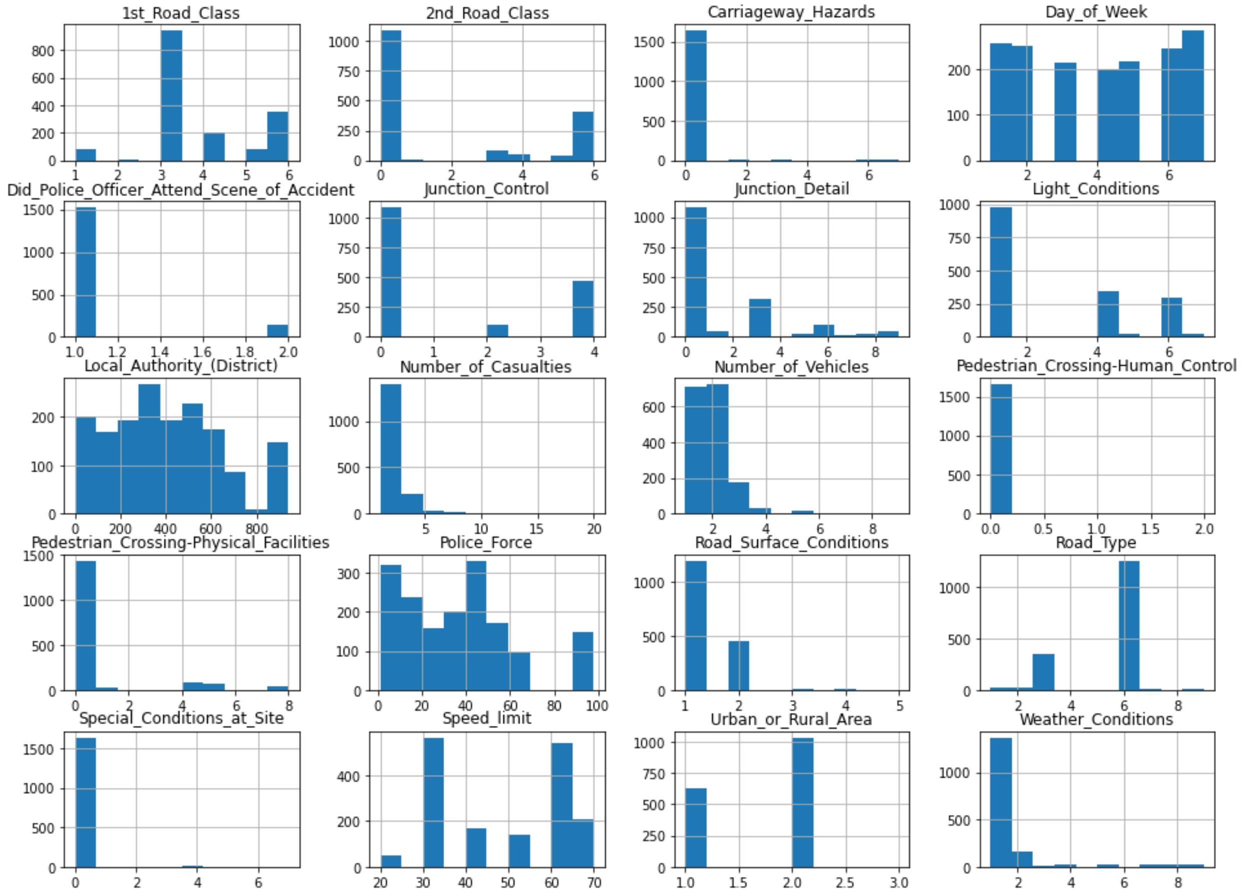
In terms of balance of the dataset, particularly on the severity metric, which is going to be our target variable, there are far fewer category 1 accidents in our dataset, which are the fatalities. Category 2 are the severe accidents, and category 3 are slight injuries. Definitions, are here: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/743853/reported-road-casualties-gb-notes-definitions.pdf. Essentially, severe injuries require hospitalisation, and slight ones do not. On the face of it, this gives us a challenge. Although the dataset is relatively large, the number of fatalities is a small proportion of the overall dataset (about 1.4%).

There are lots of options at this stage. We could rebalance the dataset by oversampling the fatalities, or undersampling the severe and slight accidents. We could train simpler models using an area under the ROC curve metric, or we could use tree based models like gradient boost or random forest. It is also a multi-class classification problem, as we're going to want to predict whether accidents are fatal, severe or slight, which is going to give us some further challenges. We might also look at converting the severity metric into something continuous and treating this as a regression problem to predict that continuous variable.

We create an array of the potential factors likely to be predictive, by removing some of the more descriptive variables like the geospatial data, the road numbers, and the target variables and then look at how many -1 values we have in each potential factor to see how much missing data we have.

Junction_Control and 2nd_Road_Class not surprisingly has a lot of missing data as only relevant for accidents at junctions. We will assume that the -1's in those columns are the same as zero's, meaning not at or within 20m of a junction. We'll see when we build out the models how predictive they are and then make a call as to whether to keep them in the modelling variable candidates or not. For every other variable, there are few enough missing values that we drop the rows with missing values entirely as unlikely to materially change the dataset. Doing this reduces the row count by about 5k rows out of 122k total, so less than 5% of the total data.

With our data set cleansed, missing values removed or replaced with appropriate alternatives, we performed some simple visual inspection of the data to see what we can learn from it before we get into model building.



It is clear from some of these histograms that some of the data is quite skewed. For example, the Pedestrian Crossing variable is almost always zero, suggesting fatal accidents don't happen often at pedestrian crossing sites. It is therefore unlikely to be predictive. It may be interesting to build a model on the subset of accidents that do happen at those sites, but for now we'll drop the variable.

Looking at some of the others, and as we're interested in a causal relationship, it makes sense to drop the variable that says whether a police officer attended the scene or not, as that's after the accident. Also, the local authority and police force variables are just proxies for location, so we'll drop those too.