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%).

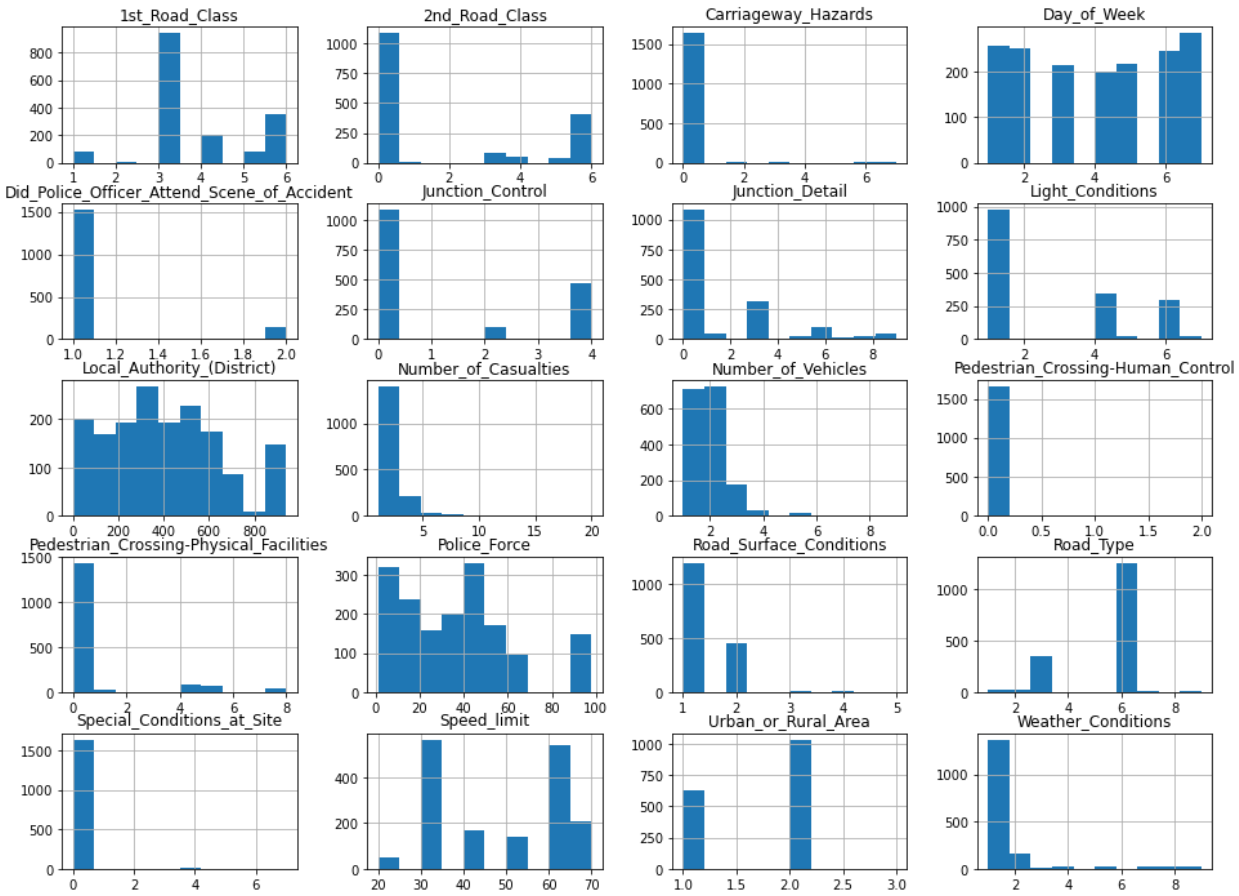
There are lots of options at this stage. We could rebalance the dataset by oversampling the fatalities, or undersampling the severe and slight accidents. We could train simpler models using an area under the ROC curve metric, or we could use tree based models like gradient boost or random forest. It 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.

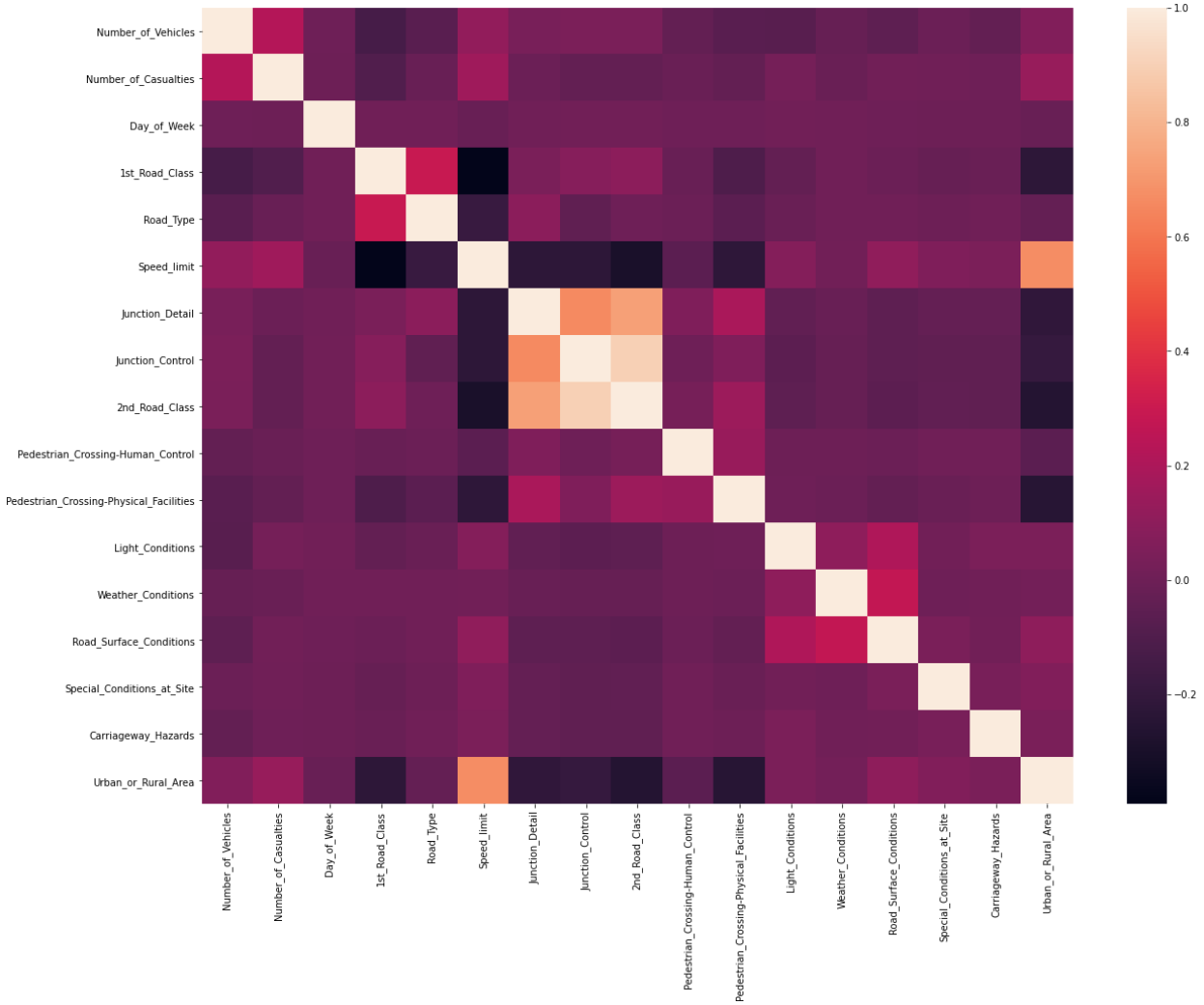


It is clear from some of these histograms that some of the data is quite skewed. For example, the Pedestrian Crossing variable is almost always zero, suggesting fatal accidents don't happen often at pedestrian crossing sites. It is therefore unlikely to be predictive. It may be interesting to build a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

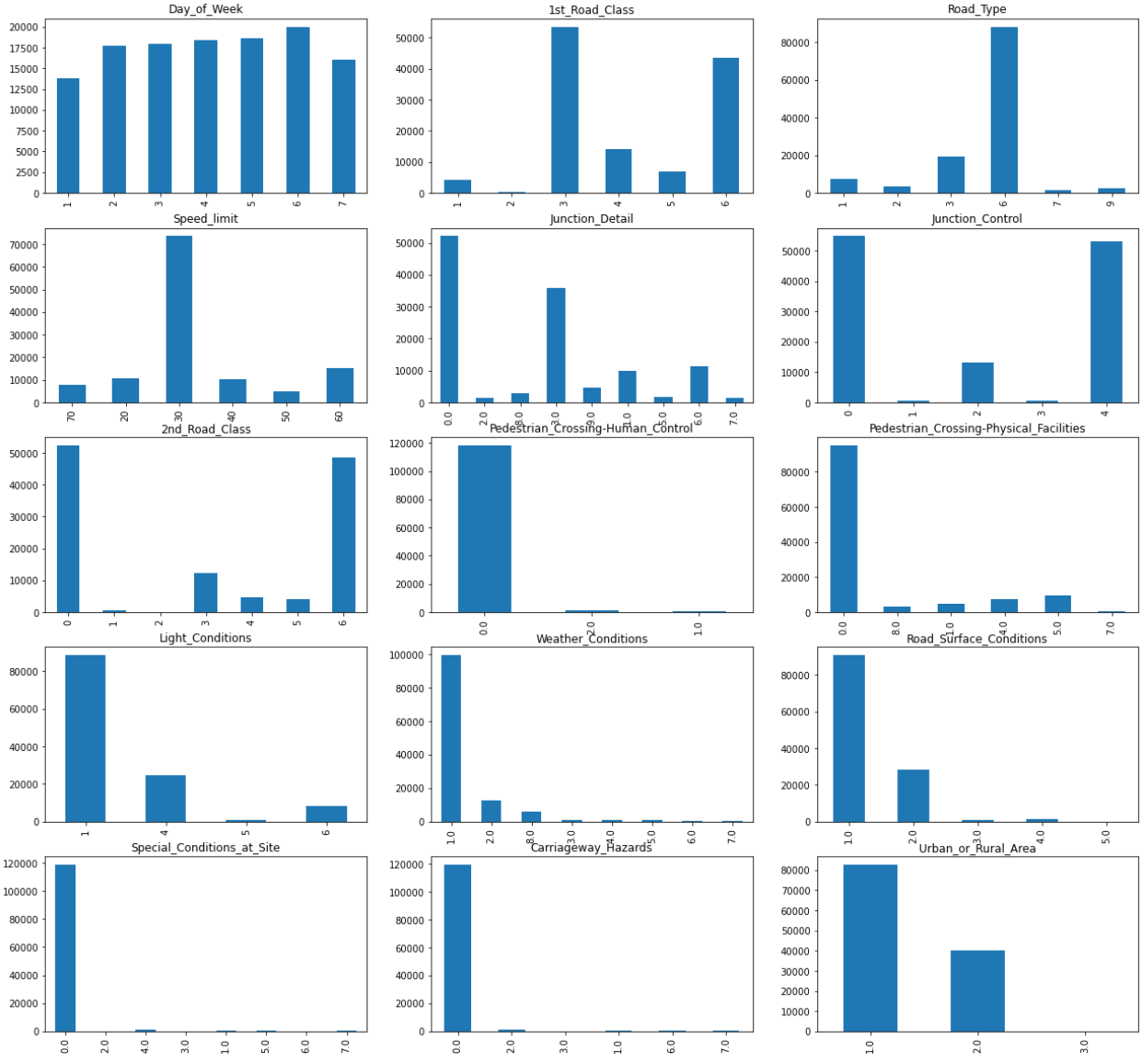
Methodology and Results

Once the dataset was cleansed (see project workbook for details), we did some exploratory data analysis, starting with a correlation analysis between the variables we were planning to use to predict accident severity:



Most variables are not strongly correlated with each other, other than the Junction and 2nd road class ones, which are correlated for obvious reasons (they relate to accidents that do or don't occur at a junction, which happen infrequently, so the large number of 'not at a junction' accidents are driving the high correlation). Speed limit also correlates with the urban or rural flag, again unsurprisingly as urban roads have lower speed limits by design. We saw no good reason at this stage to remove any variables as a result of the correlation analysis.

We then reviewed the frequency of occurrence of each variable in the dataset:



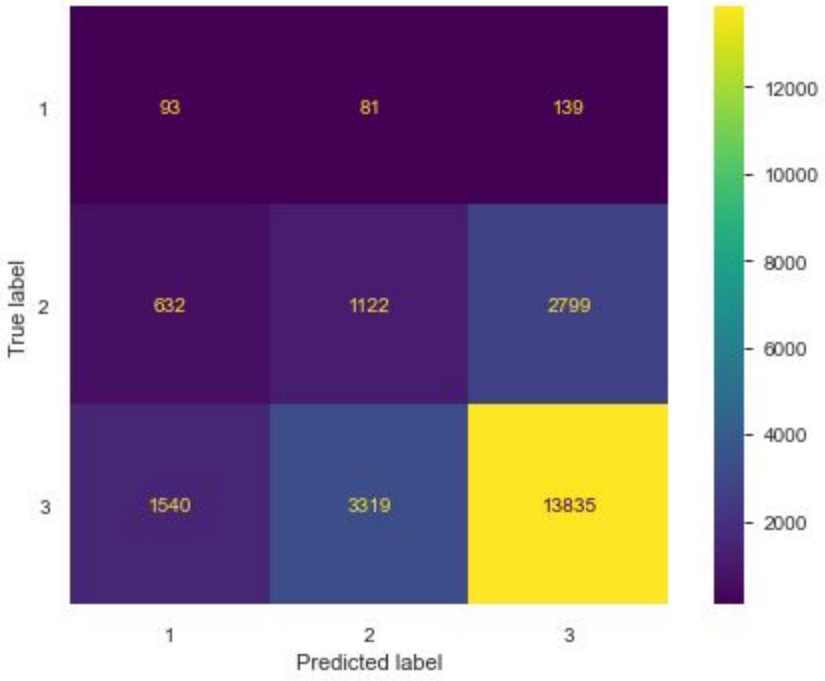
Which shows some interesting findings. Most accidents happen on a Friday, for example, and the fewest on a Sunday. Most accidents happen on A roads or unclassified roads. Most accidents happen at low speed limits. These findings, and similar ones across the variables, likely reflect the most commonly

occuring scenarios (i.e. drivers travelling at low speeds, on the most frequently occuring road types, during the day, in good weather and on good road surfaces). It seems that accidents happen most often in conditions that occur most often, which perhaps isn't surprising. As a consequence, we expect it will be challenging to build a predictive machine learning model, as it is likely there is limited signal in this data, and lots of noise.

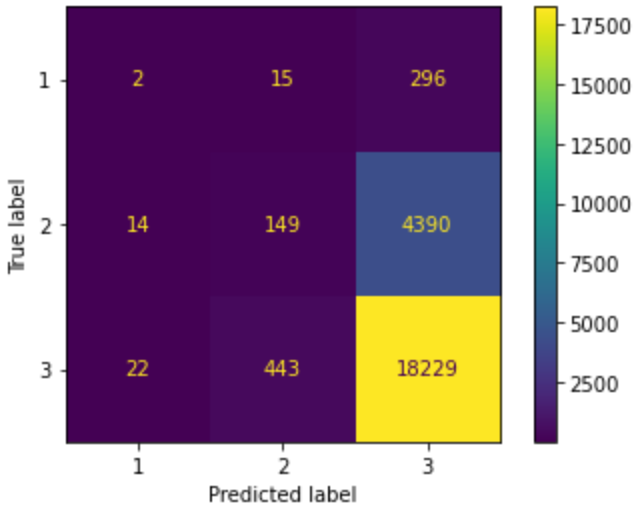
We then one-hot encoded the categorical variables and then built a series of different machine learning models to predict the accident severity, with varying degrees of success. The accuracy scores of each approach are shown below.

Model	Accuracy
Decision tree	0.714219
Linear SVM	0.758319
K Nearest Neighbours	0.758659
Naive Bayes	0.238413
Linear SVM(rbf)	0.793463
Multi Layer Perceptron	0.780136
Random Forest	0.627037
Balanced Undersampling RF	0.404117
Randomised hyperparameter search RF	0.638795

Note that accuracy is a relatively poor measure here, as the class imbalance means that a model can score well solely by predicting which accidents belong to the largest class. As such, we also calculated a confusion matrix for each approach to assess which is performing the best in differentiating the classes. These are shown in the workbook, but the best performing was the hyperparameter tuned Random Forest, which successfully picked out the most severe accidents, albeit with a very high false positive rate:



The model with the lowest false positive rate was the multi-layer perceptron:



Although again, largely attained by over-predicting the least severe accidents.

Conclusions

Although we have had some success in building a model, it is unlikely that any of these would be usable in practice, and further work needs to be done on feature engineering and setting up the problem to build a more robust model. Unfortunately, time doesn't allow us to go into that level of detail. It is likely though that better results could be achieved by recognising that accidents tend to happen where people/vehicles are most frequently present, so setting up the problem instead as an anomaly detection problem rather than a classification problem, to look at deviations from this pattern, may lead to better results.