Data Incubator Project Proposal: Building a predictive model of road traffic accidents in the United Kingdom

Robert Martin-Short

October 2018

Introduction

Road traffic accidents pose one of the most significant risks to health in modern societies. In 2017 alone, 4.57 million people in the United States (US) required medical attention after having been involved in a traffic accident, generating medical bills exceeding $413 billion [6]. About 40,000 people per year are killed on American roads due to motoring accidents [6]. In the United Kingdom (UK), the figure is lower at about 2000 deaths per year. However, taking into account the relative population sizes and mean annual miles driven per person, the mean accident risk is similar in both regions. In the UK especially, there is concern that this figure is not falling fast enough to meet government targets [10].

The problem is arguably even more severe in developing and newly industrialized countries, which have experienced rapid growth in vehicle ownership in recent decades. Indeed, the World Health Organization estimates that as much as 5% of the GDP of some low to middle income counties is lost to road traffic deaths and injuries [9].

Many authorities collect detailed information about road traffic accidents and make this data publicly available. This provides an opportunity for data scientists to develop predictive models of traffic accidents. Such models have the potential to save lives by allowing emergency services to optimally distribute their resources or local authorities to make targeted improvements to road infrastructure in accident hotspots, Many studies in the academic literature use statistical learning techniques to predict accident rates or fatalities (e.g. [11],[12]), and the insurance industry readily uses such models to calculate premium costs.

However, there is a lack of publicly available, easy-to-use online tools for predicting road traffic accidents. People could use such tools to inform their choices about where to live and work, plan their commuting routes and more generally to understand how traffic accident risks vary over time.

Researchers at the geographical information systems company ESRI have recently developed an IOS app that uses accident prediction models to help drivers find the safest route to their destination [4]. Nevertheless, this is an emerging technology and there is a great deal of scope for work on improving the models and understanding their applicability in different regions.

If selected to participate in the Data Incubator program, I propose to build a road traffic accident prediction model trained on an excellent database of accidents in the United Kingdom [3]. I will also use infrastructure data [8], weather feeds [1] and population information [7], which will give my model temporal and spatial variability. This idea was inspired by reading a Medium blog post about predicting car accident risk in Utah [5], which provides a high level overview of the modelling approach.

The ultimate aim would be to use my model to produce frequently-updated heat maps of accident probability given current conditions, which users could browse via an online tool. A possible extension of this project could be to take the model trained in UK data and test its ability to predict accidents in the United States, where similar data is available for many states [2]. This would allow me to comment on the generalizability of such models.

Datasets and computational resources

The United Kingdom Department for Transport maintains a database of all road traffic accidents reported to the police in Great Britain (excluding Northern Ireland) since 1979 [3]. This includes the coordinates, time and severity of the incident, vehicle and road-type information and weather conditions. Also provided are tables of demographic information about the causalities involved in each incident and specifications of the vehicles. This is a rich dataset that could be used to answer a wide range of questions about road traffic accidents. As part of an exploratory data analysis I downloaded the accident reports spanning 2010-2017, which amounts to 1.14 million rows (approx. 300Mb). A visualization of the accident locations is shown in Figure 1

Figure 1: Visualization of the locations of all reported road traffic accidents in Great Britain for the years 2010-2017. This was generated using Datashader. The brightness of each pixel corresponds to the number of accidents that occurred within that area. There is clearly a strong correlation between population density and accident density, with clusters of accidents occurring in major cities.

Many factors could affect the probability of accidents along a particular stretch of road [5]. These include the time of day (which may be a proxy for how busy the road is), the weather and road conditions, proximity to distractions such as signs and billboards, proximity to complicated intersections, road type, curvature and state of repair and even the demographics of the region (young men are known to be more prone to road accidents, for example [10]). In order to take weather conditions into account, I will use the weather reports in the traffic accident database in addition to temperature records, which are available as hourly timeseries from the UK Meteorological Office [1] and can

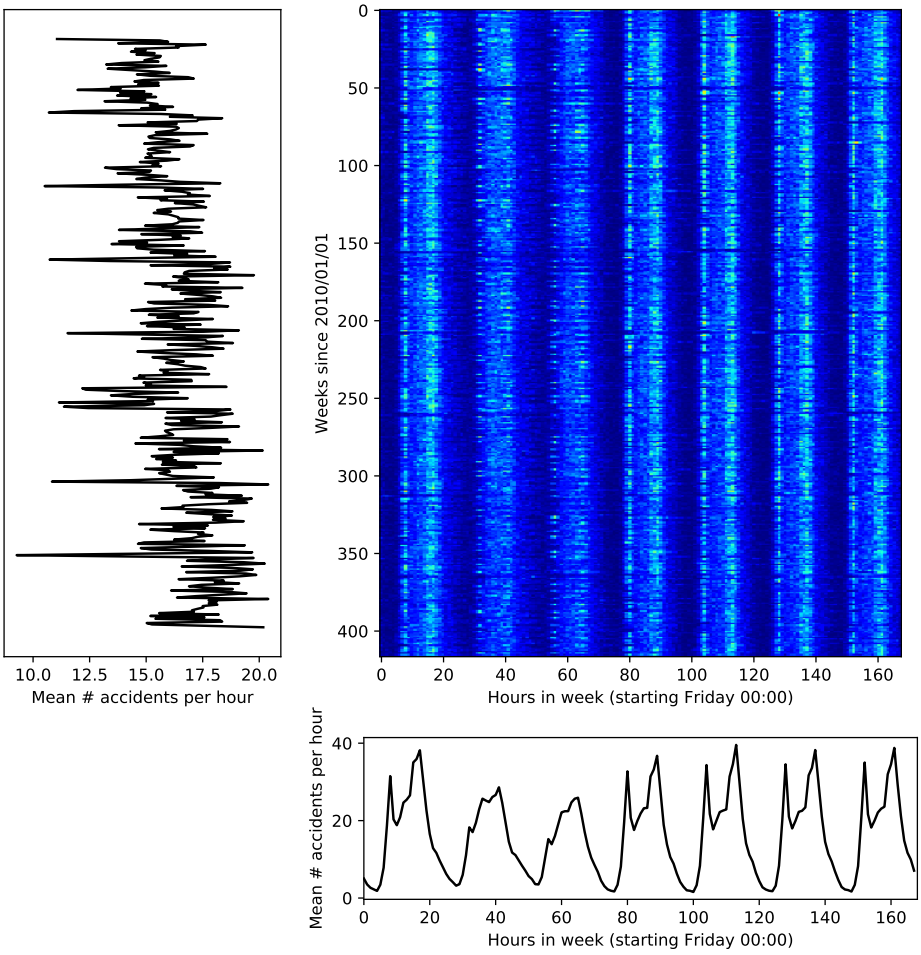

Figure 2: All accidents in Great Britain between 2010 and 2017, grouped by hour. The central, colored plot shows accidents as a function of hour within a week (x-axis) and week within the 7 year period (y-axis). Brighter colors indicate a larger number of accidents within that one hour window. The lower graph shows the hourly mean over all weeks, while the the right hand graph shows the weekly mean for all hours. The pattern of accidents within each week is clearly periodic, with lows in the early hours of the morning and peaks during the morning and evening rush hour on weekdays. Weekends do not feature such peaks. It is also clear that some weeks are distinctly more 'accident prone' than others, and that there is a long term decrease in accident numbers over time.

be interpolated between weather stations across the county. In order to extract features of the road segments, I will use a freely available road network map, which is provided by the Ordinance Survey as an ESRI shapefile [8]. Finally, I will be able to obtain demographic information about the region surrounding the road segments from the Office of National Statistics, which provides census data [7].

The total size of all datasets, including the training data for the prediction model (see the next section for how I plan to construct this) will not exceed 5Gb. Thus most analysis should be achievable

on a modern laptop computer. However, depending on the statistical learning approach taken I may need to utilize a cloud computing service to train the prediction model.

To repeat, the goal of this project is to build a reliable accident prediction model using a small number of important features that are readily available in other parts of the world. This will allow me to test its applicability to other road networks, and more generally to test the idea that accident prediction models built using data from one country can still be useful elsewhere. Thus, subject to time constraints I may attempt to test my model's ability to predict traffic accidents in the US, using datasets available from DATA.GOV [2].

Modelling approach

This project would aim to take a similar modelling approach to that described by [5]. I will aim to predict the probability of an accident occurring along some small segment of road (length on the order of 500m) during a one-hour period. Thus, this is a problem of supervised binary classification, although it could potentially be extended to multi-class classification if also consider the severity of the accident. For each accident in the dataset, I will assign a road segment ID, a series of static features of that road segment and a series of dynamic features such as weather conditions, time of day, day of week etc. It will also be essential to generate negative examples (e.g. datapoints where accidents did not occur) for the training set. There are many possible ways to do this, one of which is to randomly choose examples from the accident dataset and then randomly alter their location, time and weather condition to create a new datapoint. This should be done repeatedly until the training set contains several times (perhaps 4-5 times) the number of negative examples as positive examples. This approach avoids the problem of including every hour-segment combination in the network, which would produce a large class imbalance. It could also be extended by preferentially weighting our negative examples towards time/segment combinations that are more common in the accident database. These presumably correspond to times/segments that are busier.

In order to build a robust predictive model I would aim to compare the performance of a range of supervised classification algorithms on a holdout dataset. Tree-based classifiers such Random Forest and Gradient Boosting would be a useful starting point here, since they are relatively intuitive and can provide lists of feature importance. I would aim to use tree-based selection and/or a generic algorithm to select an optimal combination of a small number of features that are most relevant to the prediction problem. Ideally, this will make my model testable on datasets from other regions. Hyperparameter tuning would be carried out using scikit-learn `Pipeline` objects and the `GridSearchCV` tool. This classification problem could also be attempted using an Artificial Neural Network (ANN), which I would build using TensorFlow. In order to assess the performance of each model I would use standard metrics such as precision, recall and F1 score. At this stage of the project I would aim to consult with peers and supervisors at the Data Incubator program in order to select appropriate parameters and improve the model as far as possible.

An exploratory analysis of the data reveals that time of day and day of week will be important features in determining accident probability, which is confirmed by previous work [5]. Figure 2 shows that the hourly accident rate peaks during the weekday afternoon commute, and is lowest during the early hours of weekdays. Some weeks are also significantly more accident-prone than than others. These are likely times at which the roads are particularly busy, such as during public holidays.

Deliverables

The outcome of this project will be a model to predict the likelihood of a road traffic accident at each United Kingdom road segment as a function of static variables concerning the road segment and dynamic variables such as time of day, weather conditions and day of week. I would aim to display this model via a web-application developed with Flask and Heroku. I envisage the application to consist of a map of the road network colored by accident probability. A user could investigate a region of interest by inputting variables such as date, time and weather conditions, following which the map

would display the model's prediction of accident probability. It may also be possible to have the web-application access weather and road-condition forecasting data, in which case it could be used to predict road traffic accident probabilities a short time into the future. Finally, if the model works well for the UK and requires only a small number of input features, it should be straightforward to test on other datasets. The outcome of this part of the project will consist of a series of testing scores corresponding to various traffic accident datasets from the United States [2], but maps of these networks could eventually also be incorporated into the web application.

Concluding thoughts

This is a relatively ambitious but exciting project in traffic accident prediction that would benefit greatly from the guidance of mentors at the Data Incubator. The outcome would be a web-application relevant to road users in the Unite Kingdom and potentially some interesting insights into the generalizability of accident prediction models. I have already begun to collect, clean and explore the required data and if selected for the program will be able to start on the analysis and building of the training set ahead of time. In preparation for this project I would also aim to take online tutorials in Flask and Heroku, with which I only have limited experience.

I appreciate the time taken by Data Incubator staff to read this proposal and thank them for considering my application.

References

[1] Ceda archive, *MIDAS: UK Hourly Weather Observation Data*, October 2018. URL: http://catalogue.ceda.ac.uk/uuid/916ac4bbc46f7685ae9a5e10451bae7c?search$_u rl =$

[2] Data.gov, *DATA.gov public catalogs*, October 2018. URL: https://catalog.data.gov/dataset?tags=traffic.

[3] Data.gov.uk, *UK Reoad Safety Data*, October 2018. URL: https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents.

[4] Espi blog article, *Routing around future accidents*, October 2018. URL: https://community.esri.com/groups/applications-prototype-lab/blog/2018/04/02/routing-around-future-accidents.

[5] Medium blog article, *Using Machine Learning to Predcit Car Accident Risk*, October 2018. URL: https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57.

[6] National safety council article, *2017 Estimates Show Vehicle Fatalities Topped 40,000 for Second Straight Year*, October 2018. URL: https://www.nsc.org/road-safety/safety-topics/fatality-estimates.

[7] Office for national statistics, *2011 Census data*, October 2018. URL: https://www.ons.gov.uk/census/2011census/2011censusdata.

[8] Ordinance survey open roads dataset, October 2018. URL: https://www.ordnancesurvey.co.uk/opendatadownload/products.htmlOPROAD.

[9] South china morning post article, *Traffic's toll: road accidents kill 700 people a day in China*, October 2018. URL: https://www.scmp.com/news/china/society/article/1952218/traffics-toll-road-accidents-kill-700-people-day-china.

[10] David D Clarke, Patrick Ward, Craig Bartle, and Wendy Truman. Killer crashes: fatal road traffic accidents in the uk. *Accident Analysis & Prevention*, 42(2):764–770, 2010.

[11] So Young Sohn and Hyungwon Shin. Pattern recognition for road traffic accident severity in korea. *Ergonomics*, 44(1):107–117, 2001.

[12] Robert W Thomas and José M Vidal. Toward detecting accidents with already available passive traffic information. In *Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual*, pages 1–4. IEEE, 2017.