Data Scientist technical exercise

Government Digital Service

Billy Liggins

Traffic accidents in the UK

“There were 26,610 people killed or seriously injured (KSI) reported to the police in the year ending June 2018.”

Taken from Department of Transport 2018 report found here

Can a data solution be created to aid responders to such accidents?

Dataset

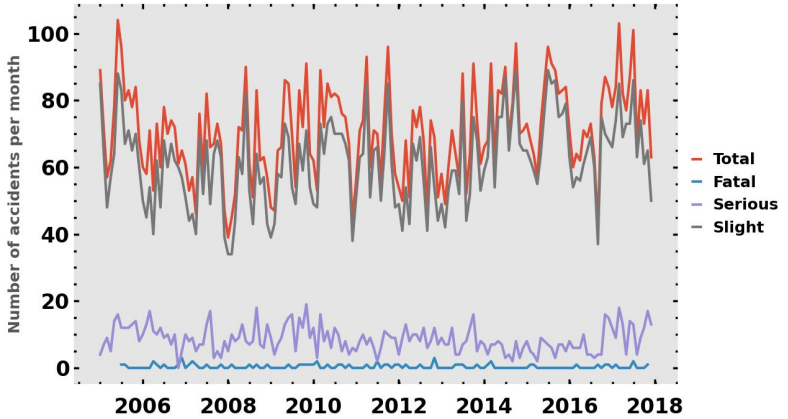
- Throughout this presentation I will be using data collected, via API calls, from https://opendata.camden.gov.uk.
- Namely the Road Collision Casualties In Camden dataset.
  - This dataset contains metadata on road collisions casualties in Camden dating back to 2005.
    - Including timestamps, accident locations and weather conditions at the time of the accident.
- I will also be using:
  - Other datasets from https://opendata.camden.gov.uk,
  - London borough ward boundaries data from data.london.gov.uk
  - Live weather data from https://openweathermap.org/ supplied under a free tire membership API calls.

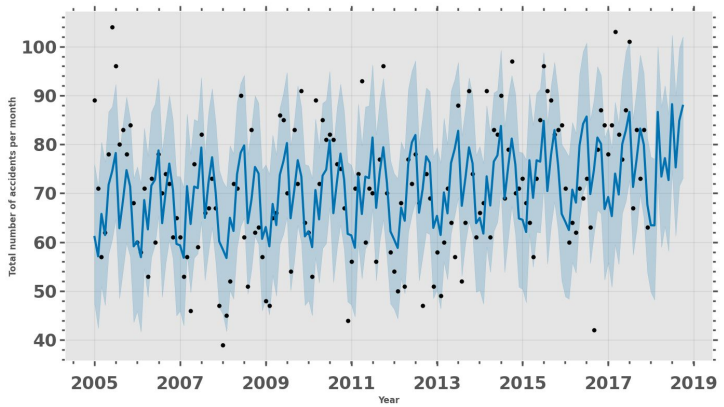
First step is to understand our data

Inspecting the data

Accidents over time

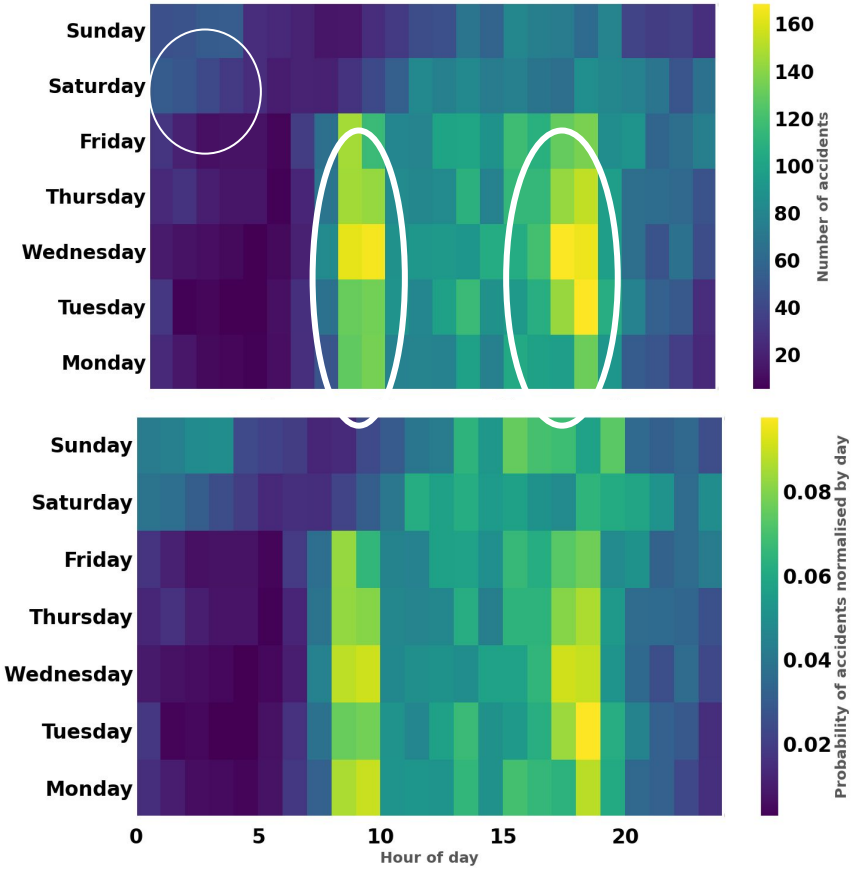
By month





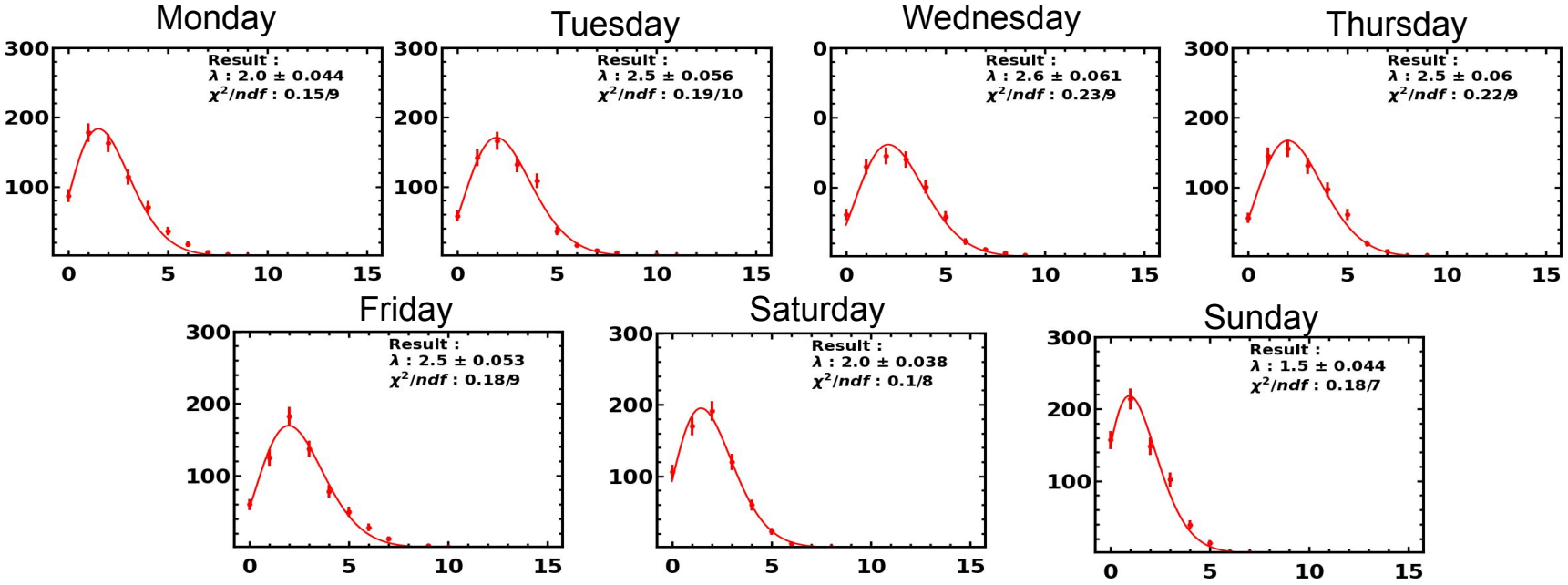
Facebook “fbprophet.Prophet” 10 month prediction

Correlation between day and hour



Are accidents randomly distributed over time?

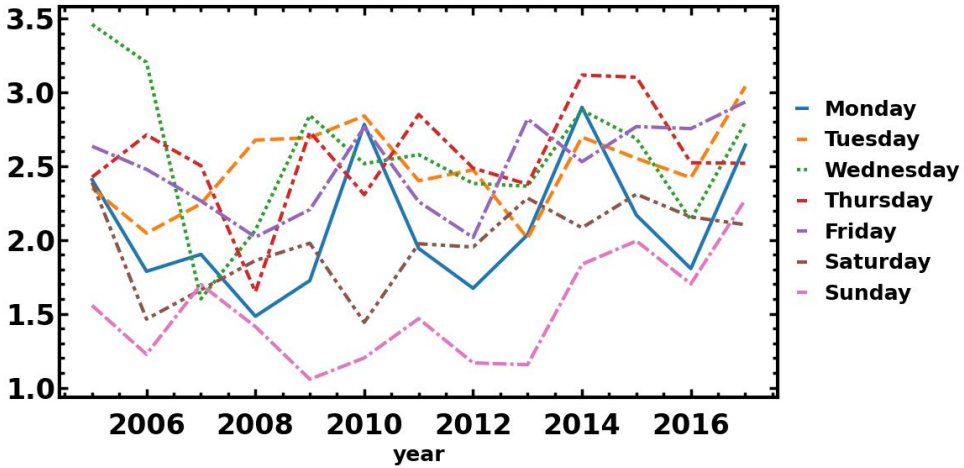
- If accidents are random, the number of accidents per (**comparable**) samples of time, will follow a **Poisson distribution**.
- Need to compare similar samples of time, with the assumption that conditions (e.g. traffic levels, driver behaviour) are constant over the set of samples.
- Individual **days of the week** fit this criteria.

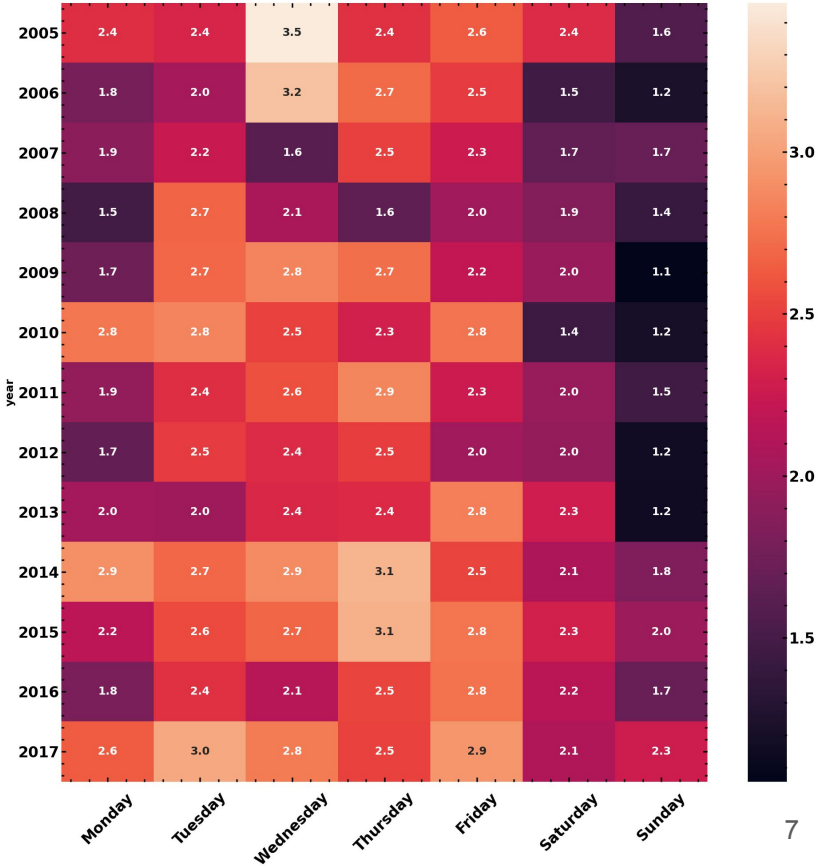


Conclusion : The data supports that accidents on a day of the week are Poisson in time

Do accident rates change over time?

- Do drivers behaviours change?
  - Change in rate
- Can I find evidence of external forcings?
  - Change in shape, i.e. not Poisson
- Slice data by year and day, and perform fits.



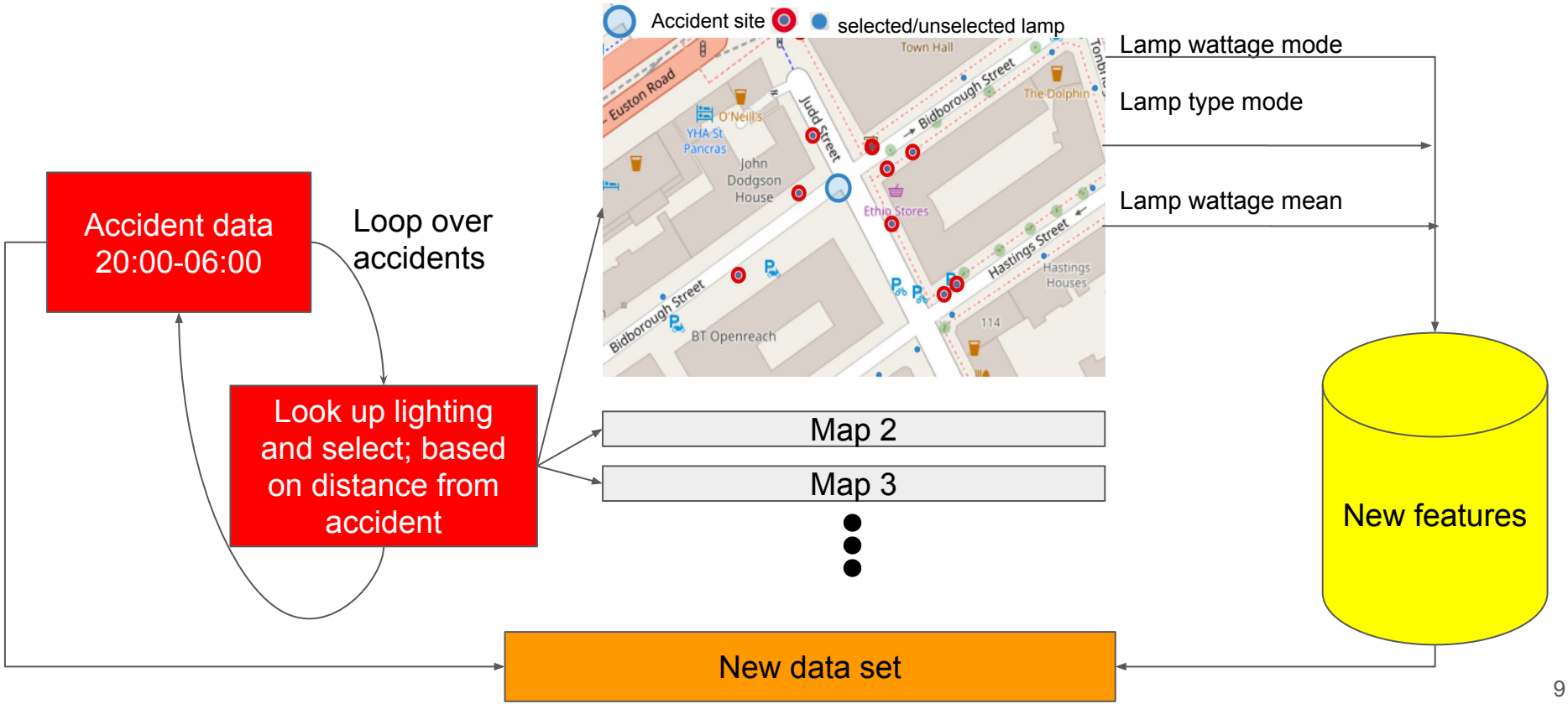


Conclusion : Rates vary over time however well within the fit errors (Needed more time to check this)

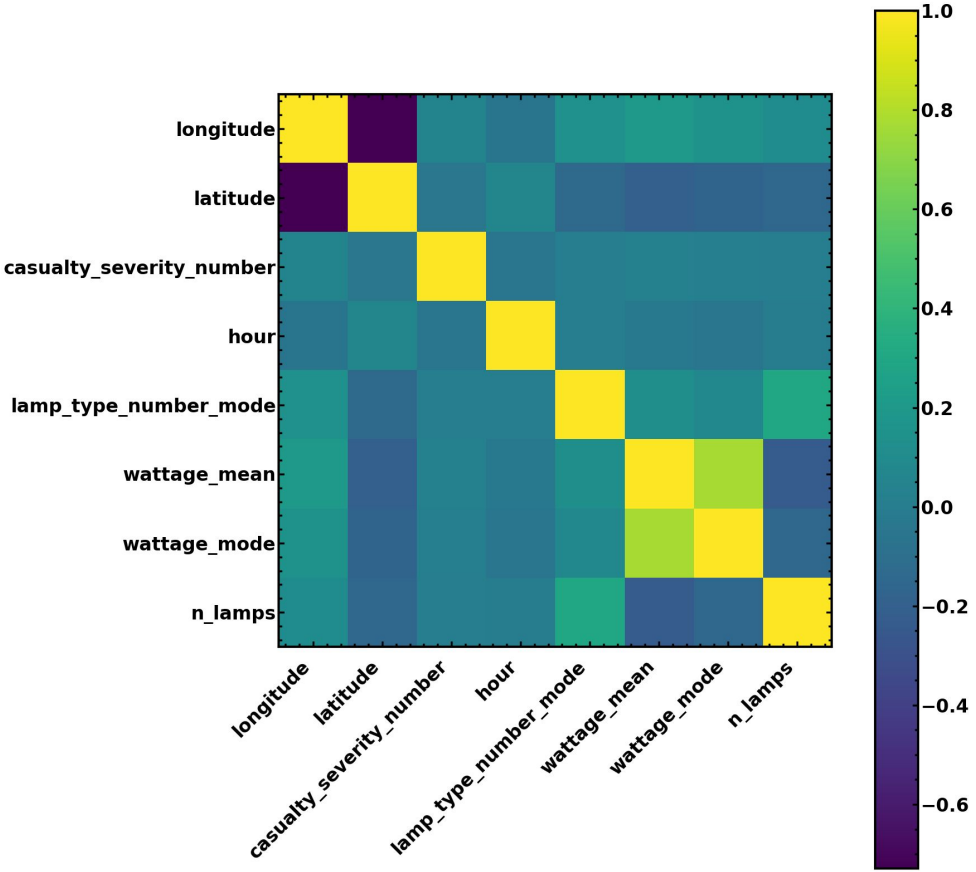
Feature engineering

Do lighting conditions correlate with casualty severity?

- Metadata, including type and location, of all council owned street lighting can be found on opendata.



Do lighting conditions correlate with casualty severity?cont.



	Casualty severity	Lamp type mode
Casualty severity	1	0.008465
Lamp type mode	0.008465	1

Conclusion : The data shows no evidence of correlation to the new features.

Possible data solution

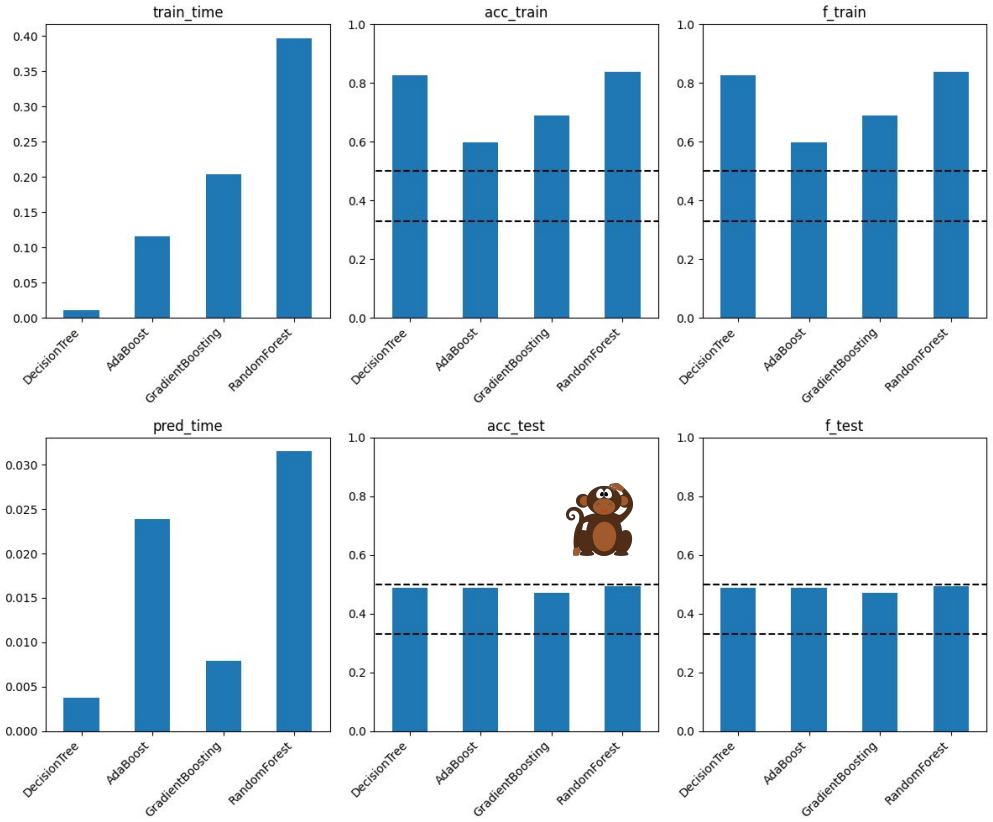
Can I assess if an accident will be fatal in real time ?

- It has been shown that supervised machine learning models can achieve high success rates when predicting casualty severity using similar data (see here & here).
- However, these models require complete information on an accident.
- I wanted to see if it was possible to predict an accidents severity from real world, real time data that I have to hand.
- Started with features: day, hour, location (borough ward), weather conditions.
  - All given in the dataset.
  - All findable in real time:
    - Current weather conditions can be accessed through free tier APIs provided by openweathermap.

Can I assess if an accident will be fatal in real time? cont.

- Data prep:
  - Made all features binary categorical data.
  - Merged “Fatal” and “Serious” severity category into one.
    - Improving training data size.
  - Made weather categories more course:
    - Reducing the dimensionality
    - Improving compatibility with weather API
- Produced models with both balanced and unbalanced datasets.
- Trained 4 sklearn models:
  - DecisionTreeClassifier
  - AdaBoostClassifier
  - GradientBoostingClassifier
  - RandomForestClassifier

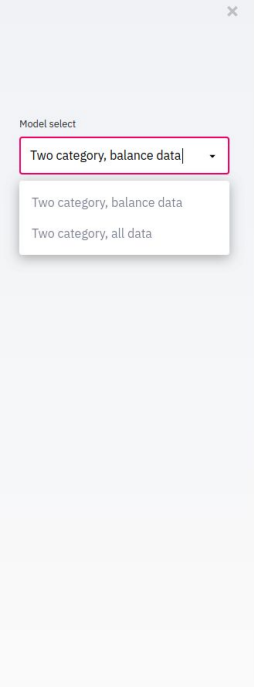
Balanced training data



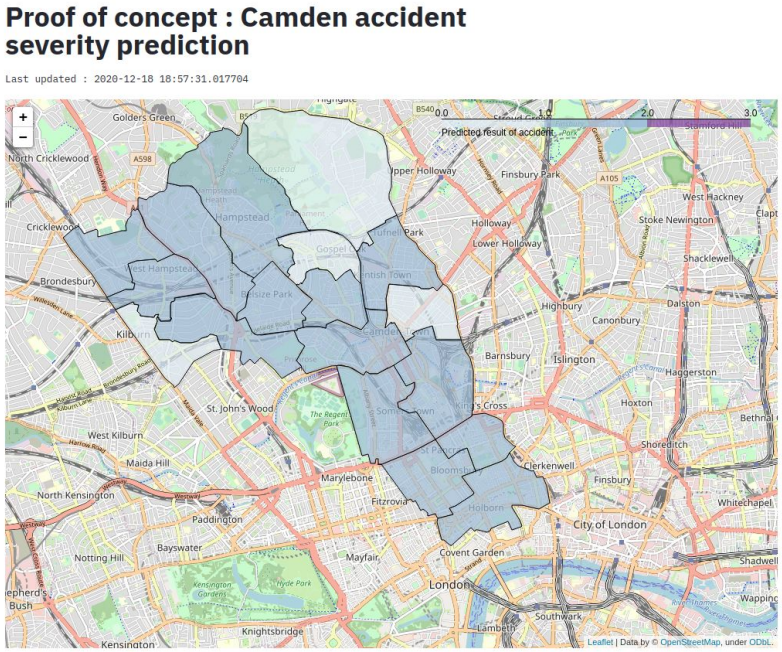
Proof of concept: Live web application

Although our model has no power, I created a web app to show how I would see it going to production.

- Web app queries the weather API and evaluates the selected model on each borough ward
- The app shows the outcome of the model overlaid over a map.
- The darker color indicates the model predicts a serious or fatal accident would occur.



Live at
www.gds-demo.herokuapp.com



Future work and potential use cases

- It is **clear** that the development of such a **model** will require a large amount of **feature engineering** work (such as the lighting correlation analysis) and an **improved dataset**.
- Possible feature engineering work may include:
    - Traffic flow data modelling
    - Location modelling
    - Geographic modelling
- More data is needed; two approaches:
    - <u>More data:</u>
        - Requires multiple boroughs to have an asset like “opendata.camden”
    - <u>Monte carlo modelling :</u>
        - In which work such as the Poisson timing analysis becomes valuable

Use cases (possible stakeholders) :

Conclusion

- I have shown that accidents are time dependent, with peeks which intuitively follow high traffic rates.
- However, they are random events and follow a Poisson distribution, with rates stable over 10s of years.
- Feature engineering: there is no correlation between street lighting and accident severity.
- The dataset is not sufficient to build a simple supervised machine learning model, with current attempts yielding random choice behaviour.
- Proof of concept: Web app in “production-like” environment.

Thank you for your time

Questions ?

Code can be found on github at https://github.com/BillyLiggins/gds_demo

Back up slides