Predicting Pedestrian Counts using Machine Learning

Molly Asher[1], Yannick Oswald[2], Nick Malleson[2]

[1] School of Earth and Environment, University of Leeds
[2] School of Geography, University of Leeds

These slides: https://urban-analytics.github.io/dust/presentations.html

UNIVERSITY OF LEEDS

Context

Accurately predicting number of pedestrians is both *important* and *challenging*

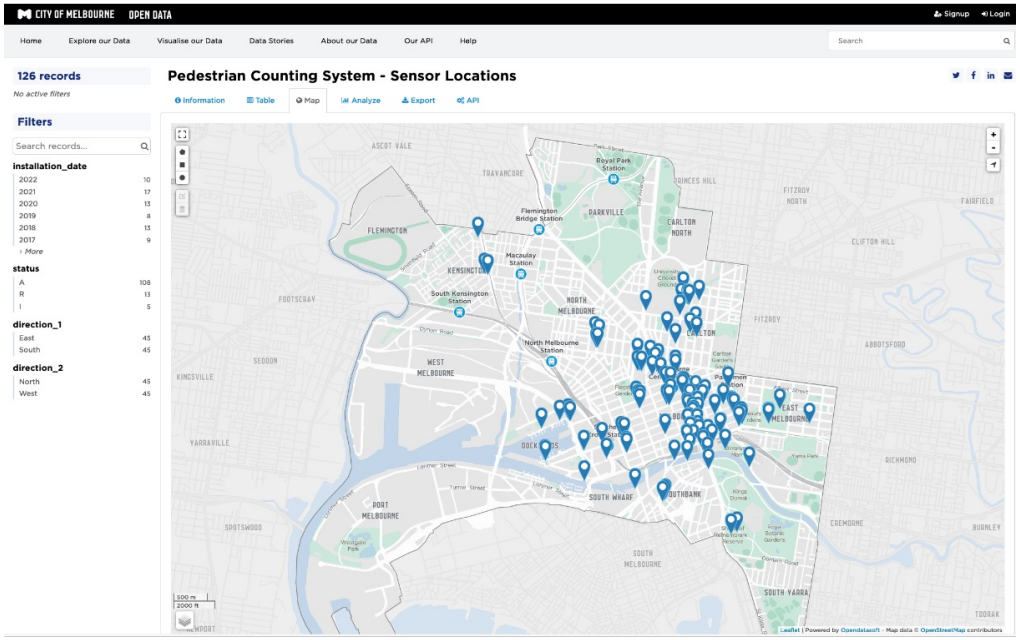
Aims and objectives:
This work is using machine learning to:

- Better understand the impact of the built environment and other contextual factors on pedestrian counts
- Predict the number of pedestrians at un-sampled locations under different conditions
- Evaluate the success of past events

UNIVERSITY OF LEEDS

Melbourne Open Data

- Melbourne Open Data Portal for open data:[1]
  - Land-use, litter, built environment, roads, bike sharing, air quality, etc.
- Network of pedestrian sensors:
  - 18 sensors in 2009
  - 82 sensors in 2022
- Record bi-directional pedestrian movements 24h/day every hour

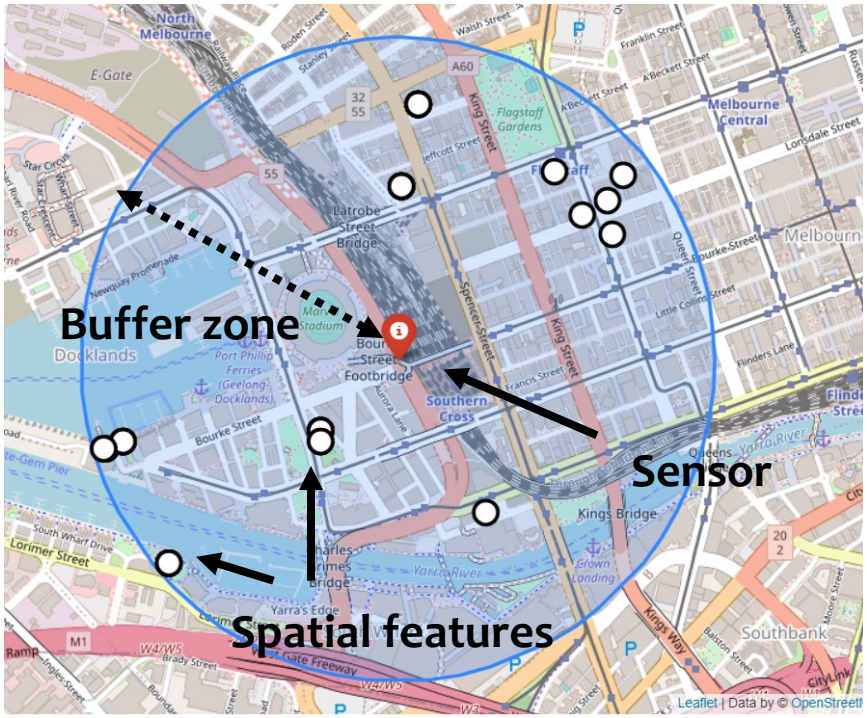


Locations of sensors in the City of Melbourne[1]

1 https://www.data.vic.gov.au/discover-city-melbourne-open-data

UNIVERSITY OF LEEDS

Melbourne Open Data



Example buffer zone within which spatial features are linked to sensors

- Numerous additional open data sets, including:
  - Weather
  - Street furniture (benches, bins etc)
  - Buildings
  - Landmarks
- Buffer zones drawn around sensors to link them to features describing urban environment in vicinity

UNIVERSITY OF LEEDS

Modelling Overview

- Dependent variable:
  - number of pedestrians per sensor per hour

- Explanatory variables:
  - time of day (hour, day, month, year)
  - weather conditions (temperature, humidity, wind speed)
  - road betweenness (a measure of how well integrated the nearest road is to the rest of the network)
  - local built environment variables (number of trees, benches, buildings, public transport, etc., etc.)

- Trained on available sensor data (4 million rows)

- Later used to predict at locations without sensors

UNIVERSITY OF LEEDS

Model selection

- Candidate models evaluated using 10-fold cross-validation
- Error metrics (RMSE and MAE) calculated on the predicted counts-per-hour of pedestrians from 10-fold cross-validation of each model against actual values from the sensor data

Error metrics

Model	MAE	RMSE
Linear regression	268.40	370.54
Random Forest regression	89.88	179.62
XGBoost	121.35	207.40

Model selection

- Candidate models evaluated using 10-fold cross-validation
- Error metrics (RMSE and MAE) calculated on the predicted counts-per-hour of pedestrians from 10-fold cross-validation of each model against actual values from the sensor data

Error metrics

Model	MAE	RMSE
Linear regression	268.40	370.54
Random Forest regression	**89.88**	**179.62**
XGBoost	121.35	207.40

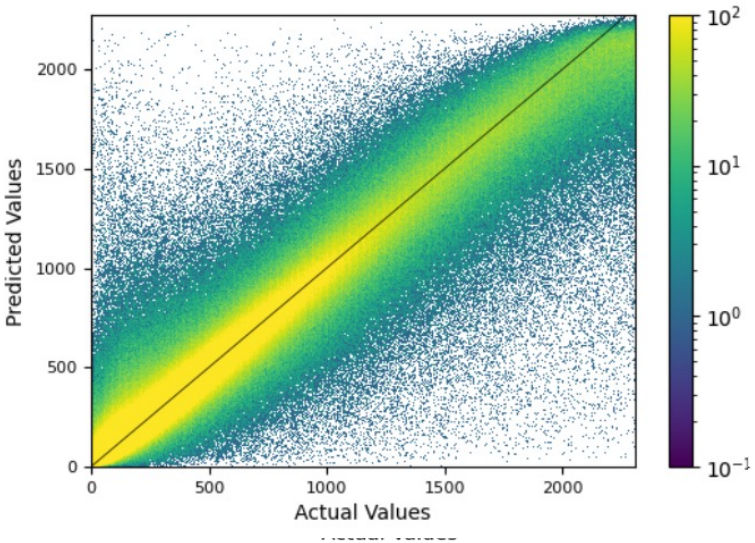
Random forest regressor selected as best performing model

UNIVERSITY OF LEEDS

Model evaluation

- Predicted counts-per-hour of pedestrians plotted against actual values from the sensor data
- Most predictions fall around the diagonal (x=y), giving confidence that model is not biased towards smaller or larger counts

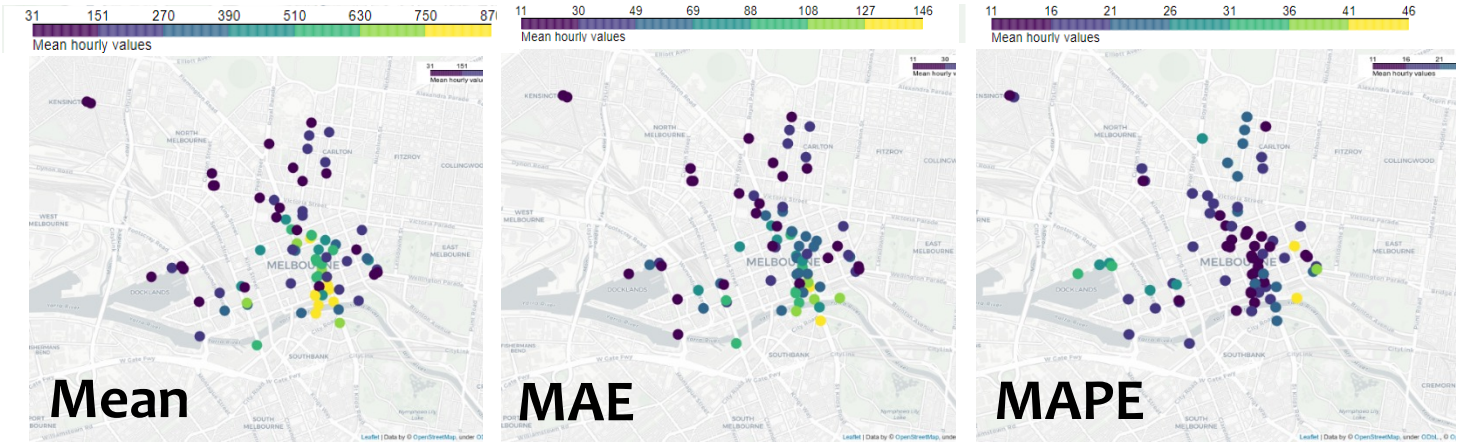
Random forest regressor



MAE	RMSE
89.88	179.62

UNIVERSITY OF LEEDS

Spatial variation



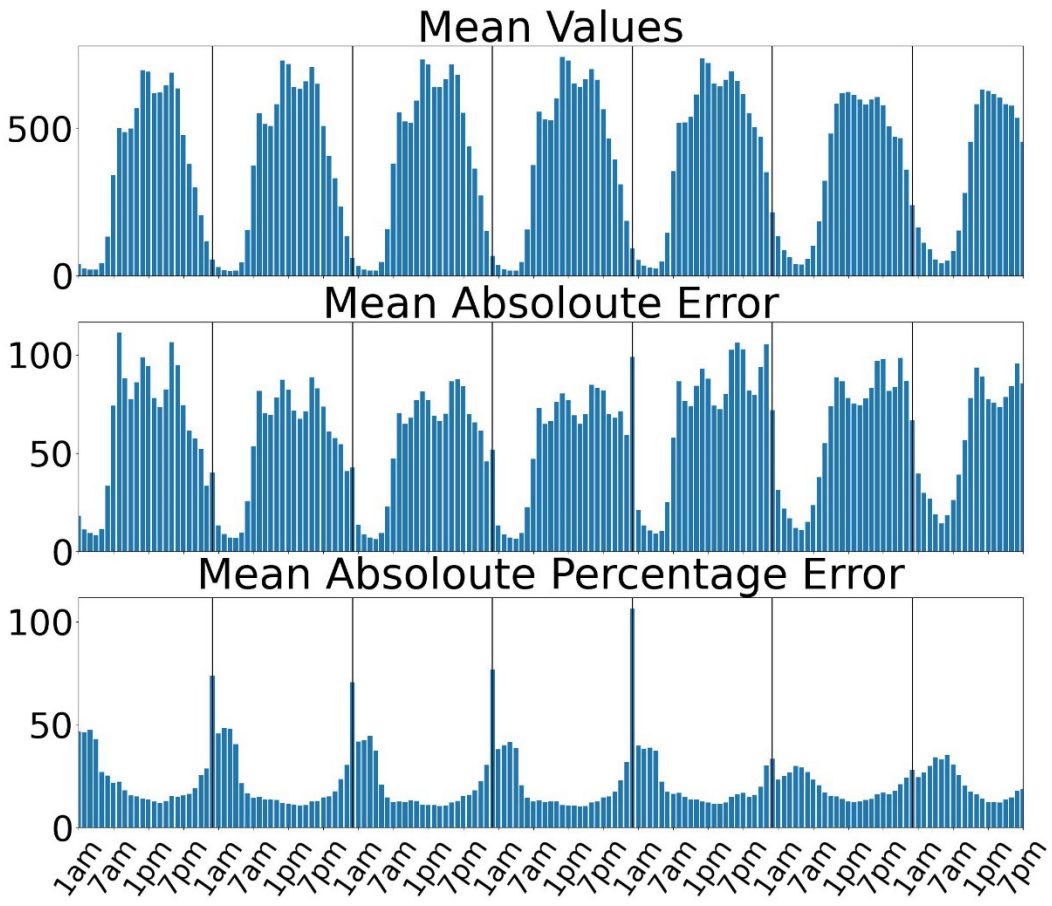
Central and southern sensors capture highest footfall

Patterns of absolute error follow (roughly) the mean, with some deviations

Several sensors with much larger percentage error

UNIVERSITY OF LEEDS

Temporal variation



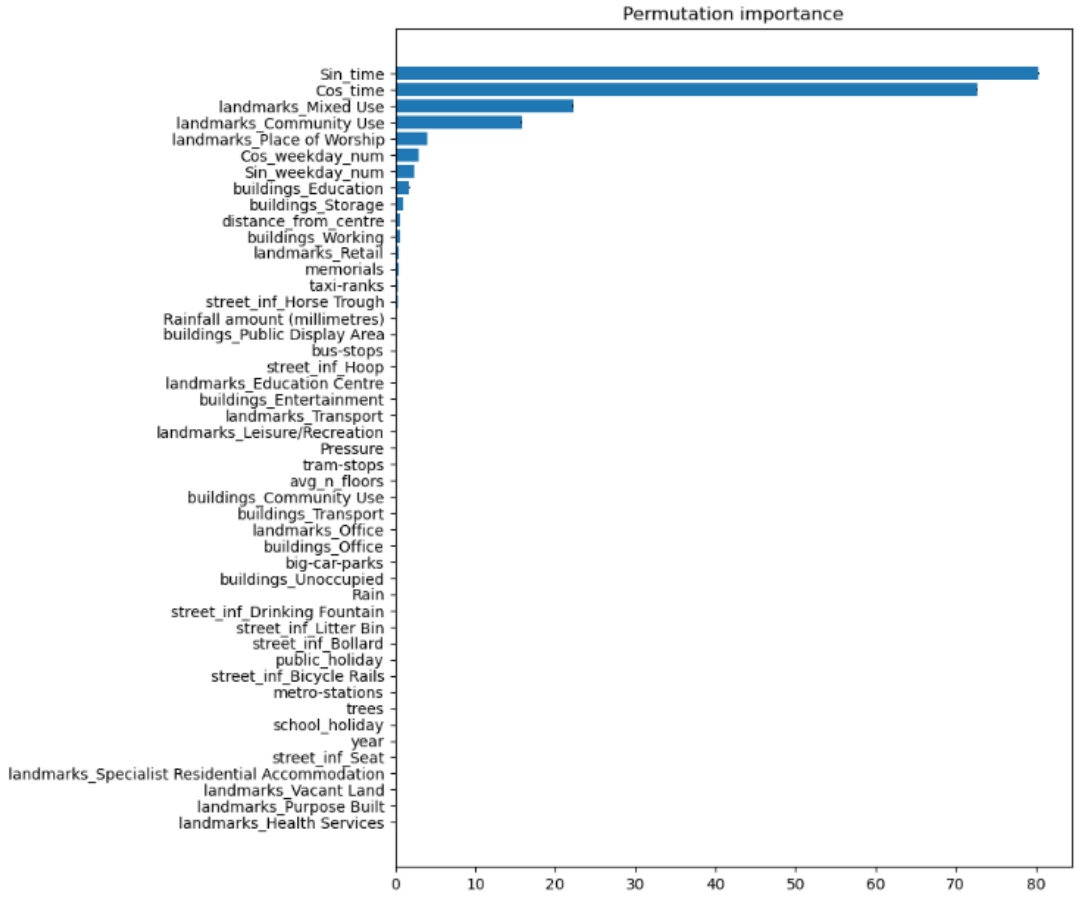
Reflect typical city centre patterns

Similar patterns to mean (larger counts = larger errors)

Largest errors at night

UNIVERSITY OF LEEDS

Feature importance

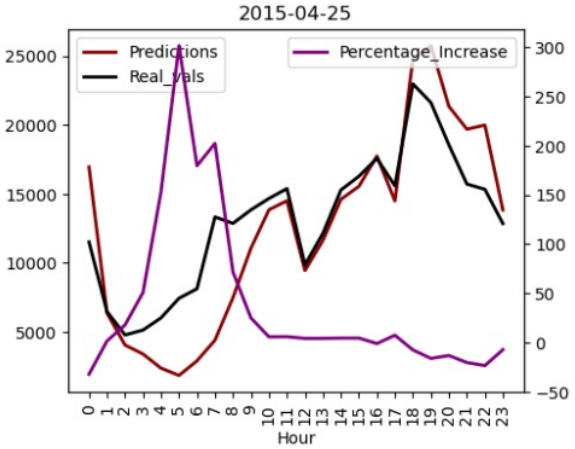


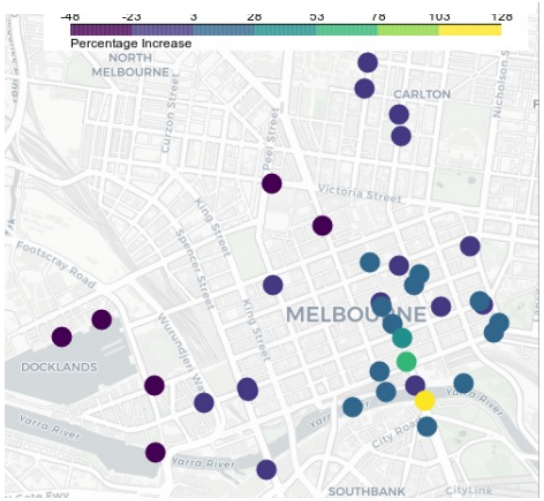
- Ranks features contribution to the model's predictions

- Most important features:
  - Hour of day
  - Landmarks (mixed, community use, places of worship)
  - Weekday
  - Educational buildings
- (surprisingly) lower importance features:
  - Betweenness
  - School/public holiday

UNIVERSITY OF LEEDS

Evaluating events





- Model can be used as a tool to evaluate success of events

- E.g. Anzac Day Parade:
  - 5% more footfall in whole city over 24h
  - 72% more footfall from 3-10am
  - 128% more footfall at a sensor in south-east near parade location

UNIVERSITY OF LEEDS

Conclusions

- Ongoing work aiming to:
  - accurately predict the number of pedestrians in time and space at un-sampled locations under different conditions
  - better understand the impact of the built environment and other contextual factors on pedestrian counts
  - Evaluate the success of past events
- Model performs reasonably well overall
- Some spatial and temporal variations in prediction error
- Beginning to make inferences about impact of urban environment

Predicting Pedestrian Counts using Machine Learning

Molly Asher[1], Yannick Oswald[2], Nick Malleson[2]

[1] School of Earth and Environment, University of Leeds
[2] School of Geography, University of Leeds

These slides: https://urban-analytics.github.io/dust/presentations.html

UNIVERSITY OF LEEDS