UK Sport Presentation

# In which events is there a risk of missing a medal in Paris?

* Built model with assumption:
  + Probability of winning a medal is a function of the results at the World championships from the previous 2 years
* Can only predict where a boat has been entered for the previous 2 years
  + Womens 8 had no entry in 2022
* Chose last 2 years as first year in Olympic cycle tends to have hang overs from previous cycle
  + Also allows for consistency across cycles with taking COVID into account

# How was the model built?

Web scraping

* Scraped all WC and Olympic results from World Rowing website
* Only looked at last 12 races (semi-finals or heats if no semis)
* Obtained rank order from finishing positions
  + Rankings do not necessarily reflect closeness of competition
* Tried to create identifier to stipulate how close boats were
  + Times could be used for A final
  + How do you compare boat 6 with boat 7
    - Times of B final may be at a quicker time of day
  + Percentage of winning time was average of fastest time across semi-finals and A final

Produced rank order

* For each World championships and Olympic games, this produced:
  + Rank order by event
  + Relative speed by event

Data transformation

* Transformed data so that WC years appeared as properties of the Olympic year

Plotted results to find binding box

* Plotted relative times of previous 2 years with outcome at Olympic games
  + Hypothesis was that there was some bounding box in the top right-hand corner, where if you were within that box, it meant a medal was more likely.
* Did same exercise with finishing positions.
  + Hypothesis was that there was some bounding box in the bottom left corner, where if you were within that box, it meant a medal was more likely.

Created logistic regression model

* Built binary classifier Logistic regression model
  + Binary classifier for 1 = win medal
  + Tried Naïve Bayes but logistic regression gave better results
* Built on entire model
  + Did not have time to do train/test split
  + Probably overfitted
* Recall is not good – not picking up a lot of medals won

# How could model be improved?

* Are boats in previous years a fair reflection on current boats
  + What happens when personnel change
* If model was built on relative speed, it could return how much quicker a boat needs to go.
  + How do you measure relative speed?
    - Accommodate changes in conditions
  + Should relative speed by on a logarithmic scale?
    - Gaps get smaller in a tail wind
* If the model included intra-season races, would it be possible to see how quick boats are going during the season