**Agenda:**

* Methods Outline
* Showing singular and ensemble models retrained without ‘year’ and ‘country’ columns and with 1000 Optuna finetuning trials rather than 300 to decide which models we should use going forward
* Showing of PCA results

**Methods Outline**

**Primary aim:**

* To predict the maternal mortality ratio (MMR) using a mixture of socio-economic and health-related data sourced from the World Bank and WHO.
* Motivation:
  + Provide a method for estimating the MMR of countries without an official estimate to inform policy.
  + To forecast future changes in the MMR to predict future trends and the effects of policies, again to inform public health policy.

**Secondary aim:**

* To identify the variables with the highest predictive power for the MMR, thus highlighting potential important areas that public health policies can target to reduce MMR.

**Data Exploration:**

* *Trends in the missing data* 
  + Missing data per year
  + Missing data per feature histogram
  + Maternal mortality rate missing data
  + Proportion of data missing data remaining, as well as the number of rows, and columns remaining after all rows/columns with greater than *threshold* proportion of missing data are removed.
* *PCA analysis* 
  + With colouring by income level, MMR, and year
  + Give variation explained by each principal component (e.g. bar graph of first 10 principal components)
* *Correlation analysis*
  + Correlation strength versus MMR
    - Name highest correlation
    - Give histogram to show full distribution of correlation strengths
* *Key statistics based on literature review* 
  + Gave mean, median, standard deviation and proportion of missing data in dataset
    - Can compare mean and median for discussion of outliers
  + 20-25 features
  + Emphasise that GBD is very high quality data (include in references)

**Data Cleaning and Pre-processing:**

* Data merging process
* *Removal of missing data for cleaning purposes:*
  + MMR ratio was only measured between 1985 and 2018, so all data outside this window was removed.
  + Removed all rows where the MMR estimate was missing
  + Removed all rows and columns whose data was completely missing.
  + Removal of 3 feature columns too similar to the MMR measure, preventing the model from ‘cheating’.
* Iterative removal of columns and rows with a greater than threshold proportion of missing data

**Feature Selection**

* No feature selection employed, and all features used
* Subset of features based on their importance in the literature
* Subset of features whose absolute pairwise Pearson’s correlation coefficient with the MMR was greater than [0.6, 0.7, 0.8]
  + Producing 3 subsets of features

**Train/Validation/Test Set Curation:**

* Split the 5 versions of the dataset (each of the feature subset versions and no feature subsetting) into train test sets in a 90:10 split by either splitting by year or by country.
* Split these train sets into train/validation sets using 5-fold cross validation in an 80:20 ratio, again either by country or year (where the same year or country was either all in the train set or all in the validation set).
* Apply each level of missing data removal to each of these folds.

**Model Building and Fine-Tuning:**

* Trained a Random Forest, XGBoost, and LightGBM model on each of the available training sets.
  + Hyperparameter finetuning performed using Optuna for 1000 trials using MSE.
  + Evaluation on test set using MSE, RMSE, MAE, Relative Error, and R2.
  + Averaged results across cross-validation folds and took standard deviation for comparison purposes.
* Trained ensemble models based on the predictions for all the best performing models’ predictions on each of the cross-validation folds (with separate models trained for the different missing data thresholds).
  + Voting ensemble and stacking ensemble models (ElasticNet Linear Regression, Random Forest, SVM,).
  + Similarly, finetuned the ensemble models’ hyperparameters using 300 Optuna trials

**Feature Comparison:**

* Looking at which features the models found to be most useful and comparing against those selected for their high correlation and/or high importance in the literature.
  + Either via in-built methods or permutation importance

**Sensitivity Analysis:** need to redo this for the best performing models

* Only trained the best-performing models on data from:
  + The lowest income level
  + The highest income level
  + 2000-2014
* These re-training procedures were done as separate experiments.
* The retrained models’ performance was compared to their performance on the full dataset to determine the models’ sensitivity.

Uncertainty analysis? -> potential to-do

* Resample from the original data and use the best-performing model to predict the MMR based on the resampled data.
  + Do this 1000 times
  + Look at output distribution to extract the 95% confidence interval

**Comparison to literature values:**

* Look at predictions of the MMR made in the literature
  + E.g. on IHME site I can find MMR estimate for different countries in different years
* Compare the predictions of my best-performing to their predictions using absolute difference
  + Can also look to see if my estimate with its uncertainty bounds fall within the uncertainty bounds of the literature’s estimate.

**Questions:**

* Many of my early reports discuss different imputation methods. Should I mention anything about trying these methods, or leave this out completely?

None to threshold 100%

* 0.6 should be closer to ‘no feature selection’
* Rename the correlation lit bar
* Add number of features for each feature selection method into the figure

Ensure that the scaling is the same for the same metric across different models

* Figure a-c for the same metric and model

One figure/metric in main text, others in the supplement

* Explain for training because of training time so could only train on one metric

See feature importance for ensemble random forest

* See if they are all contributing to the reduced error

Sensitivity analysis on rf ensemble

Research uncertainty analysis in literature

Look for anomalous behaviour in figures (and try to explain it)

Note: from last week’s discussion of variance per principal component

A graph with blue squares

AI-generated content may be incorrect.