**Agenda:**

* Investigation of random forest stacking ensemble model
* Sensitivity analysis for split by country data
* Possible approaches to uncertainty
* Having a performance baseline
* Questions about literature review and structure of thesis
* Thesis writing timeline
* Conference

*Investigation of random forest stacking ensemble model*

Previously, we had seen that the random forest stacking ensemble model had significantly higher performance than the other ensemble and non-ensemble based models, as shown below.

A graph of different colored bars

AI-generated content may be incorrect.

I investigated when the random forest model’s performance may have been so much higher by exploring whether it was making smart predictions about which base estimator models it should use. To do this, I used random forest’s built-in ‘feature importance’ attribute (calculated under the hood by the scikit learn platform). Feature importance is the mean ‘of accumulation of the impurity decrease within each tree.’

A graph with blue lines

AI-generated content may be incorrect.

As shown in the plot above, the random forest ensemble model only places importance on features from a subset of models. While there were some small importances for random forest models and other models than those listed below, the following models had the highest feature importance values.

0-19: lightgbm all features

20-39: lightgbm literature features

200-219: xgboost all features

220-239: xgboost literature features

240-259: xgboost forest correlation 60

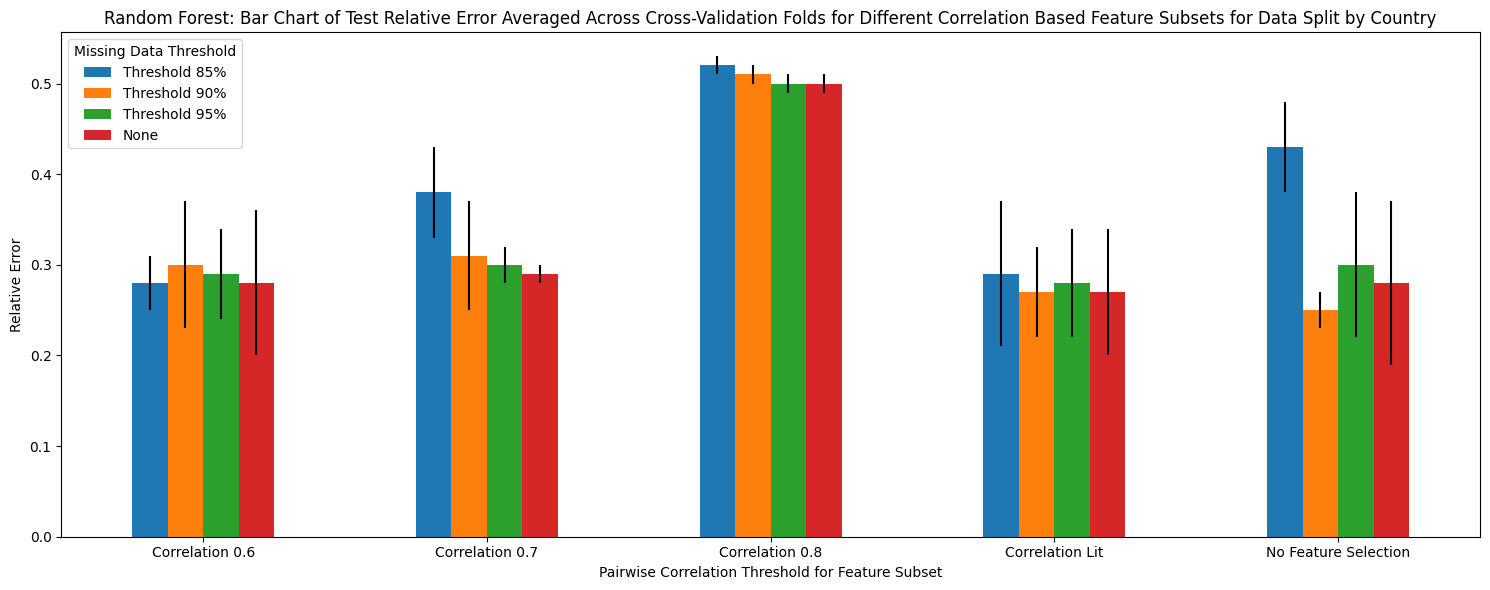
It makes sense for the xgboost models to be chosen, as the error bars show standard deviation across the folds, so the ensemble model could have chosen the models trained on the fold with the smallest standard deviation. For example, the model with the highest feature importance was the xgboost model from fold 2, 95% threshold, no feature selection. I don’t completely understand why none of the random forest base estimators were given high feature importance, as some of them performed just as well as the others. None of the higher correlation models were chosen due to worse performance.

A graph of different colored bars

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

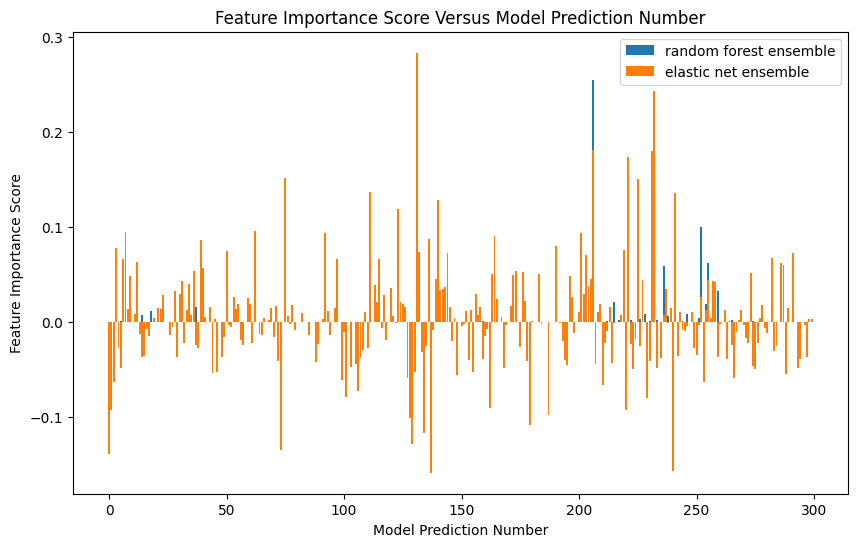


In contrast to the clear preference for certain base estimators shown above, the voting ensemble model used the predictions of each model relatively equally (feature importance determined by the weighting of each base estimator).

A graph with blue and orange lines

AI-generated content may be incorrect.

Similarly, the linear regression-based ensemble model derived support from most of the models, having either a positive or negative coefficient attached to each model. Interestingly, the base models with the highest associated coefficients in the linear regression ensemble were not always the same as those with the highest coefficients in the random forest ensemble model, as, unlike the latter, the former valued some of the random forest base estimators.



I conducted the same analysis for the ensemble-based models produced for the data split by year. Depending on the metric used, the random forest ensemble model either had a very slight edge over the base estimators or it was clearly better, as shown below. Given that it was always either as well performing as the other best model, or clearly better than, depending on the metric, I have chosen it as the best performing model.

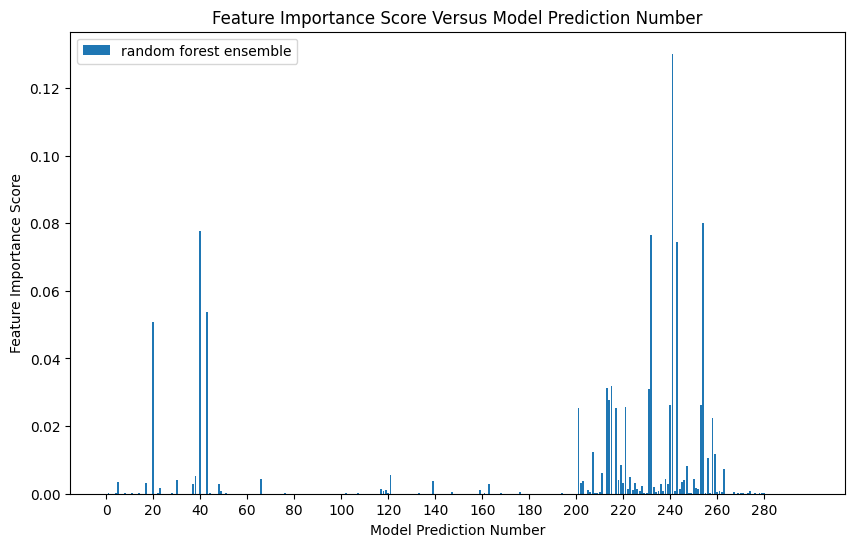
A graph of different colored bars

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

Again, there were few random forest base estimators with high feature importance, with most of the high feature importance scores belonging to the xgboost models. However, in contrast to the ‘split by country’ data, here more of the xgboost models appear to be given relevance.

**

This is explained by more similar performance across the base xgboost models, as shown below. Potentially, this may be due to the feature subsets containing the features most important for future predictions.

*A graph of different colored bars

AI-generated content may be incorrect.*

*Sensitivity Analysis*

To perform a sensitivity analysis, I completed the following steps:

1. Keep the fold divisions for datasets with no missing data removed.
2. Filter for subset of data aligning to sensitivity analysis
   1. Only one income level or only post-2000 or pre-2000
3. Perform missing data thresholding on each fold of the filtered data
   1. This generates 4 missing data versions, with 5 folds each, for each of the sensitivity analysis subsets.
4. Train each of the 300 base models on these datasets
5. Get the models to predict on the entire training/validation set filtered for their specific sensitivity subset.
6. Split these predictions (80:20) into a training and validation subset, then finetune the random forest ensemble on the training predictions 1000 times, comparing trial models’ validation scores to pick the best ensemble model.
7. Get the base models to predict on the test data filtered for their sensitivity analysis, then have the random forest ensemble predict the final MMR outcome using these test predictions.

A graph of blue bars

AI-generated content may be incorrect.

Strangely, this shows that the model trained on all the data performed better than on the sensitivity analysis models trained on just low, lower middle, or upper middle-income data. This may have been due to overfitting, as the entire dataset only includes 78 samples from low-income countries, which likely would have had more missing data. When this missing data is used in the context of countries with much less missing data, the fact of there being missing data can be used by the model as a datapoint, allowing it to perform better. Additionally, the test set restricted only to low income data has 10 samples, preventing a test of the model’s true performance. Potentially, the lowest income model performed better than the lower middle income model because the former has trends unique to the lowest income countries, while the latter encapsulates a greater range of countries with only 310 datapoints in the entire database.

Unfortunately, due to needing to finetune each of these models 1000 times, they are still running. I will share the results as soon as possible. As a note, the test set for the ‘split by year’ data is only from 2015 to 2018, and so the ‘post’ data would have no testing data. Should I omit this sensitivity analysis?

*Possible approaches to uncertainty*

Prediction uncertainty:

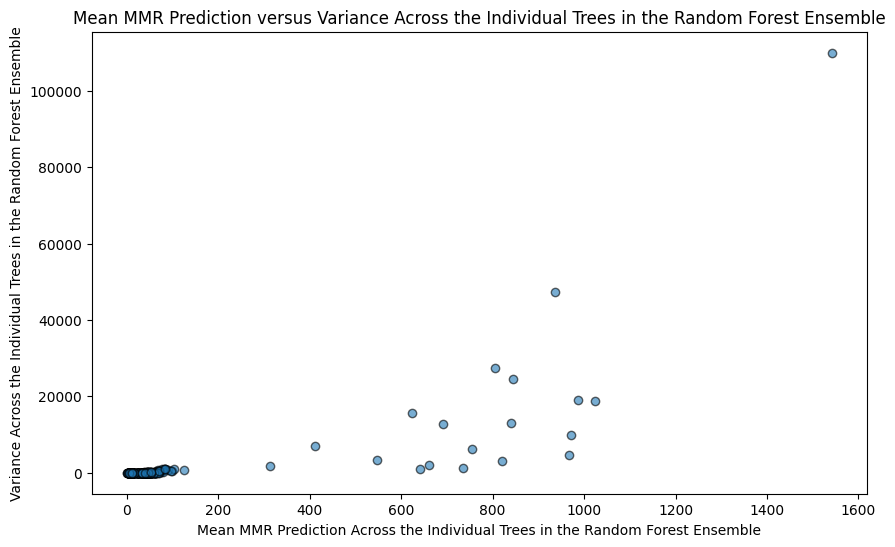
* Chua et al., 2023, Nature Biomedical Engineering
* Potentially relevant methods for measuring prediction uncertainty:
  + ‘ensemble-based prediction discrepancy’A black text on a white background

    AI-generated content may be incorrect.
  + ‘patch based methods’

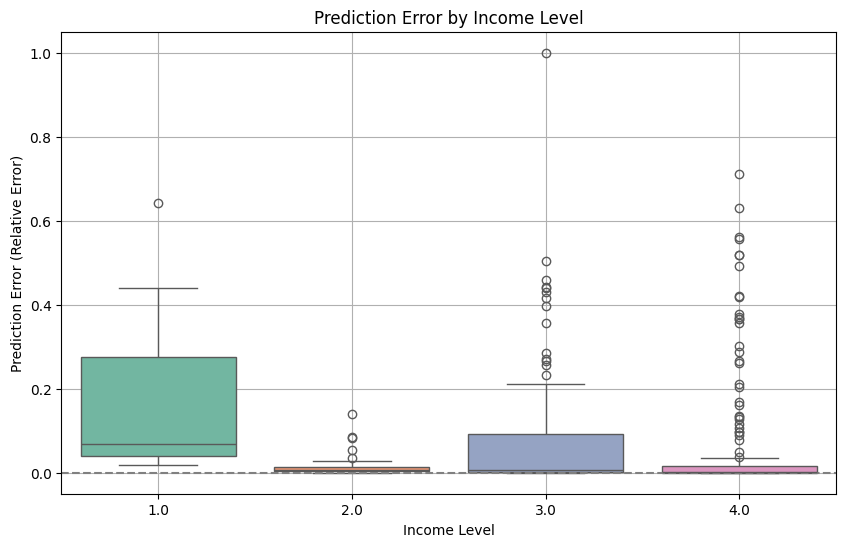
A close-up of a text

AI-generated content may be incorrect.

As a potential method for estimating uncertainty in my random forest ensemble, I took the variation of the prediction across all trees in the ensemble for each MMR prediction and plotted below. The graph shows the variance of the individual trees versus mean MMR prediction across the trees. It demonstrates that the ensemble had higher variation across its component trees for higher MMR predictions, which correspond to (generally) lower income countries with less data and thus more uncertainty.



The effect of this is shown below, as lower income countries had greater relative test prediction error in the final random forest ensemble result. Interesting that the second-lowest income level also had low error.



Is this a possible way of representing uncertainty in the model?

*Literature review and having a ‘baseline’ for comparison of model’s results*

In our previous meeting, we discussed the importance of having a ‘baseline’ to use for comparison against my model’s performance. From a literature review, I have the following options. Would any of these be acceptable to serve as a ‘baseline’?

* The UN and the Global Burden of Disease Study have produced modelled MMR estimates using a mixture of Bayesian hierarchical regression and gaussian regression models. They do not include ‘error’, but they have 95% uncertainty intervals for their estimates.
  + I could compare the percentage difference between their mean estimate and my predictions as well as the percentage of my estimates that are within their uncertainty bounds.
  + As a note, I am predicting the ‘national’ estimates, which many articles in the literature describe as under-reporting maternal mortality. Thus, both of the above models include mechanisms to adjust for under-reporting and data quality issues, meaning that there are likely to be differences between their results and mine. Is this solely a discussion point?
* The GMatH model is a microsimulation of women’s reproductive lifecycles used to generate alternate maternal mortality predictions. They provide the mean absolute error of their model’s predictions on a test set of civil data for the number of maternal deaths.
  + While they do not have a specific mean absolute error for MMR (as shown below), they have one that looks like it is maybe the average across a couple of similar metrics.
  + A black and white text

    AI-generated content may be incorrect.
* The one paper I found on using machine learning to predict maternal mortality in specific areas of Kenya (published 2024) had a R2 score and mean squared error in the MMR. I could compare these against my results.

*Questions about literature review and structure of thesis*

I am currently deciding the best way to structure my background information and literature review. Do you think that the following breakdown of information make sense?

*Background:*

1. What is maternal mortality and the maternal mortality rate (MMR)
   1. Why is it important to reduce the MMR?
   2. What data sources are typically used to estimate the MMR?
   3. Why is lack of data making this difficult?
2. Machine learning to be able to predict information in data-sparse fields
   1. Say that machine learning has been increasingly used in public health and to predict missing data (then say more on this later in the literature review)
   2. Define supervised learning
      1. Include information about how data is split into train/val/test sets and evaluated
   3. Define base models used
      1. Linear regression based
      2. Tree based
         1. Xgboost
         2. Random forest
         3. Lightgbm
      3. Support vector machines
      4. Do I need to discuss other types of models (i.e. models not used in this study)
   4. Define ensemble learning
      1. Stacked
      2. Voting

*Literature Review of Related Work (Machine Learning for Public Health)*

1. Existing methods for estimating the MMR
   1. Description of each of the 3 statistical models
   2. Small comparison of their results
   3. Limitations of the models
      1. Should the accuracy (if given) of the models be given here or in the discussion?
      2. Should this be throughout the description of the models, or at the end here?
2. Why machine learning may perform well for this problem
   1. How would you recommend justifying the use of machine learning for this problem in light of these sophisticated methods?
3. How machine learning has been used in the maternal health domain

*Thesis writing timeline*

Based on your past experiences supervising students, when would you expect for me to have a full draft? Shen would you feel comfortable to start providing feedback?

My general plan:

* By the end of the mid-semester break, I will have finished writing the literature review, background information
* Methods done by the end of week 7
* Results section by the end of week 8. I will also have a first draft of my poster.
* By the end of week 9, I will have a first draft of my complete thesis (with the caveat that I may still be adding small sections to the discussion)

*Conference final decision*