**Agenda:**

* Thesis timeline update
  + Finished rough drafts of background information and literature review.
  + In the process of writing methods and analysis, with expectation that I will finish by our meeting next week (Week 9 Tuesday)
  + First draft of discussion finished for Week 10 Tuesday
  + Week 10-Week 12 drafting and feedback
* Experiment with different ensemble configurations and data permutations
* Discussion of prediction error for split by year vs split by country
* Discussion of SMOTE
* Discussion of presentation of literature comparison
* Thesis structure questions
* Poster feedback

**Results from different ensemble compositions**

Tested random forest ensemble performance using a variety of different ensemble compositions.

* All base predictors (method discussed up until now)
* Randomly permuted base predictors to determine whether order of base predictors impacts which base models are most ‘valued’ by the ensemble.
  + Whether the XGBoost models were ‘first seen’ by the ensemble, and since no other model ‘added value’ by further decreasing loss (as performance was similar across the base estimators), the XGBoost remained the most valued.
* Just XGBoost base predictors
* Just Random Forest base predictors
* Just LightGBM base predictors

*Split by country:*

A table with numbers and letters

AI-generated content may be incorrect.



* The XGBoost only ensemble performs slightly better than using all the base predictors
  + Makes sense, as these are the models that the random forest ensemble placed highest importance on.
  + Should I redo the final bit of the sensitivity analysis, making this new version of the ensemble the ‘best’ version of the model?

A graph of blue rectangular bars

AI-generated content may be incorrect.

Model importance scores and choice of most important models largely the same between the permuted and original random forest ensemble model. *I do not have a good explanation for why these XGBoost base predictors are chosen above any other model.*

Most important models in original. Most important models in permutation

A table with numbers and a number on it

AI-generated content may be incorrect. A table of numbers with numbers

AI-generated content may be incorrect.

*Split by year:*

A table with numbers and a few black text

AI-generated content may be incorrect.

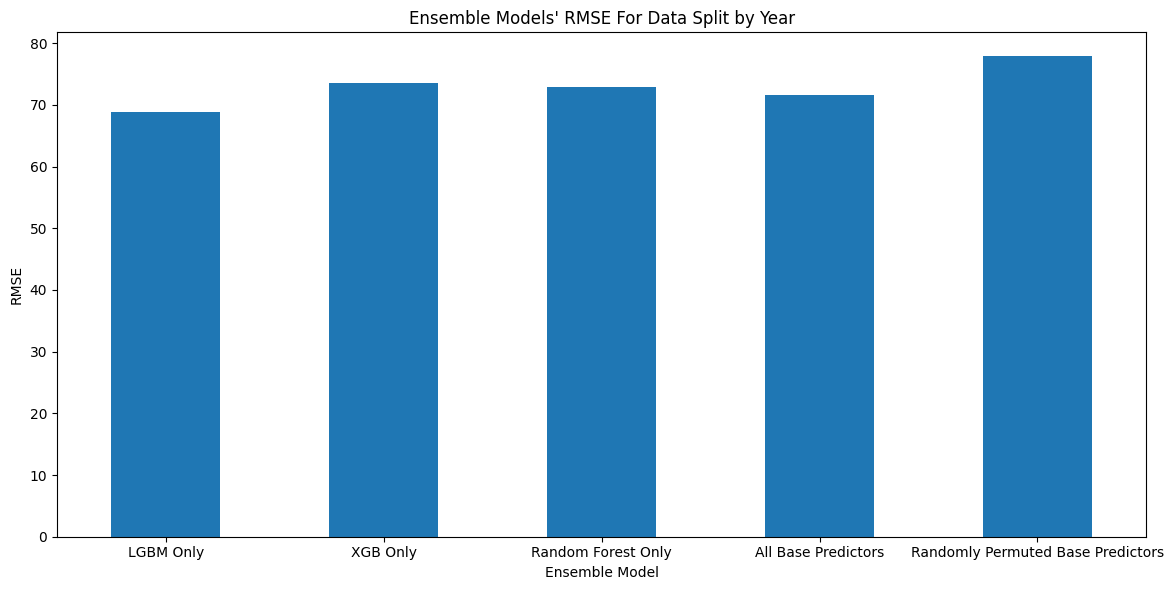


* How would you choose which is the best performing model?
  + Is this enough of a difference to switch from using all the base estimators to just a single type?
* What is strange about this is, according to the models valued most highly by the random forest ensemble when all 3 base model types were used, the Random Forest was barely used and the XGBoost was most used.
  + Although the random forest was still heavily used.
  + I do not have a good explanation for why LightGBM performed above the others and had a lower MSE than XGB only (which was the metric used to finetune the ensemble), but was still chosen less than the XGB model.

A graph with blue lines

AI-generated content may be incorrect.

A graph of blue rectangular shapes

AI-generated content may be incorrect.

As shown in the following table, permuting the order of base predictors affected the importance score magnitudes (more of a difference between most important and second most important models), but similar models were still deemed more important. **I do not have a good justification for this.**

Most important models in original. Most important models in permutation

A table of numbers with numbers on it

AI-generated content may be incorrect. A table of numbers and a few digits

AI-generated content may be incorrect.

* Randomness in training process
* Random subsetting of predictions

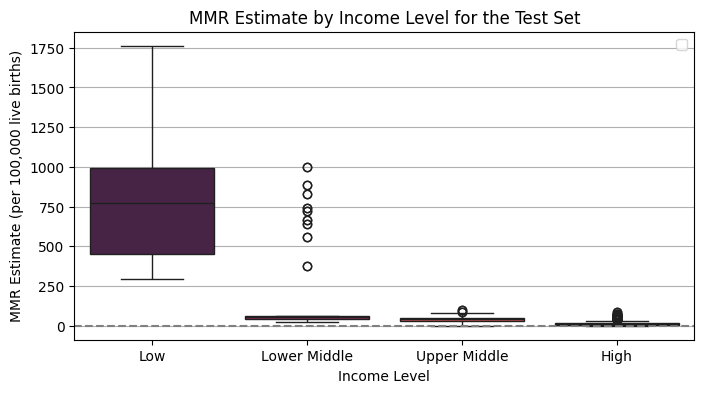
**Discussion of prediction error for split by year versus split by country**

A graph of a graph with different colored squares

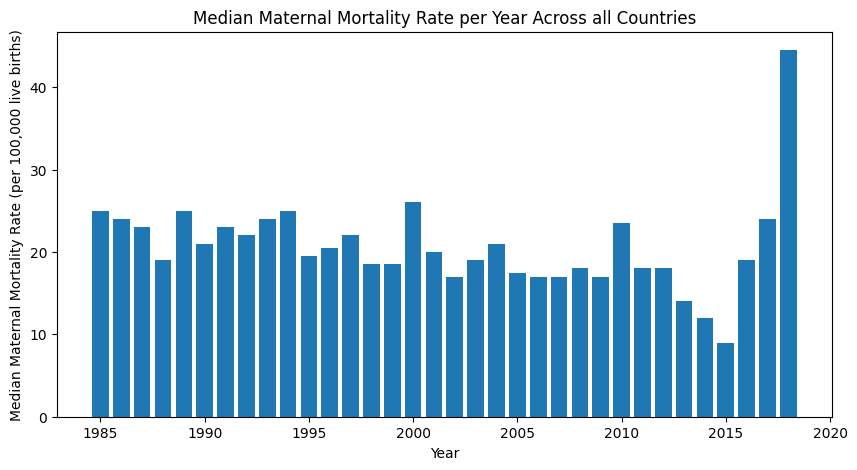
AI-generated content may be incorrect.

* Missing data analysis
  + The lowest-income countries have the highest predictive error, likely due to low-quality, sparse data and a small sample size, causing overfitting to training set.
  + High-income and lower-middle income have lowest error
    - High-income likely due to small MMR range, PCA clustering and relatively large amount of data
    - However, outliers due to ‘noise’ caused by the inclusion of data from other income levels.
    - Lower-middle due to the model picking up some kind of trend (no specific reason, is this necessary in the thesis?)
  + As shown below, the lower-middle test values are entirely within encapsulated in Q0 to Q3 of train/validation set, allowing the model to generalise well.
  + However, the test set has higher mean and Q3 for low-income than train/validation, potentially helping to explain the lower performance.

A graph with a bar chart

AI-generated content may be incorrect. 

* Predictive analysis
  + Error increases with income level, which was surprising.
    - As shown in the following figure, this was likely due to the median MMR being higher for the test years.
    - Could also be due to difficulty of determining a per-year trend (no year-specific clustering on the PCA)
    - Is this enough of an explanation?
  + Higher predictive error will accumulate and likely result in a higher relative error for higher-income countries than lower-income.
    - More specifically, for high-income countries, the predictive error will be divided by the max(true, predicted) MMR estimate. Given the datapoint is a high-income country, the MMR is likely to be low. Thus, the error is being divided by a smaller number, causing relative error to increase.



* Large outlier in last year (show per income plots)
  + Check if they changed the definition?
  + Explore which specific countries were involved in the spike

A graph with a bar chart

AI-generated content may be incorrect. A graph with a bar and a number of bars

AI-generated content may be incorrect.

**SMOTE Discussion**

|  |  |  |  |
| --- | --- | --- | --- |
| **Income Level** | **Number of Samples** | | **Proportion of Samples Remaining (%)** |
| **Before Removing Samples with MMR Missing** | **After Removing Samples with MMR Missing** |
| **Low** | 884 | 78 | 8.8 |
| **Lower-Middle** | 1734 | 310 | 17.88 |
| **Upper-Middle** | 1802 | 996 | 55 |
| **High** | 2176 | 1405 | 65 |

*Problems caused by small proportion of low-income samples*

* Lack of generalisation to low-income countries, who would benefit most from this technology
* Performance measures may not reflect poor performance on low-income samples due to their relative scarcity
  + Although this may be somewhat balanced by the larger magnitude of the errors, as the low-income samples tend to have much larger MMR estimates.

*Oversampling by randomly replicating low occurrence datapoints*

* Can cause overfitting given the small number of cases covered by the low occurrence datapoints, with replication potentially amplifying noise

*Synthetic minority oversampling technique (SMOTE)*

* Synthetic samples are generated using interpolation on the K nearest neighbours of low occurrence instances.
* Limitations:
  + Can generate overlapping and noisy samples.
    - This is particularly important in my dataset, as the PCA shows that the MMR estimates for the upper-middle, lower-middle, and low-income classes overlap. This means that the neighbours of low-income samples could actually be from another income level, meaning that the generated sample would not be representative of the low-income level samples.
  + Prior work has found that applying SMOTE on high-dimensional data produces no real benefit in classification performance.
    - Lower performance of SMOTE for high-dimensional data is partially attributed to hubness, where the same small subset of datapoints are more frequently chosen as the neighbours, and thus are used to generate a higher proportion of the synthetic samples. This can bias the synthetic samples, especially if the datapoints in the hub are unrepresentative.
    - As a result, many methods propose use of feature selection or dimensionality reduction techniques before applying SMOTE.
    - Particularly limitation for my data, given that I have 721 features.
  + A review discusses how a variety of techniques have been developed that perform data cleaning and filtering before applying SMOTE to increase data quality and thus have more representative points.
  + In a review of SMOTE that covered 15 years of progress, there was no explicit mention of modifications to SMOTE using missing data.
* My thoughts (not from the paper):
  + Imputation could add additional bias, causing the generated samples to be unreflective of the true data distribution.
  + These limitations could result in biased and/or inaccurate generated samples. Thus, SMOTE was not used in this project, but would be an interesting avenue to explore in the future, as potential modifications could be made to the primary SMOTE method to allow it to work with sparse and overlapping data.

*Undersampling*

* Removing samples from the more common income levels to reduce imbalance in the data.
* However, this can reduce generalisation if important samples are removed
  + Important given the dataset is a little bit small
* <https://jair.org/index.php/jair/article/view/11192/26406>
* [SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary](https://jair.org/index.php/jair/article/view/11192)

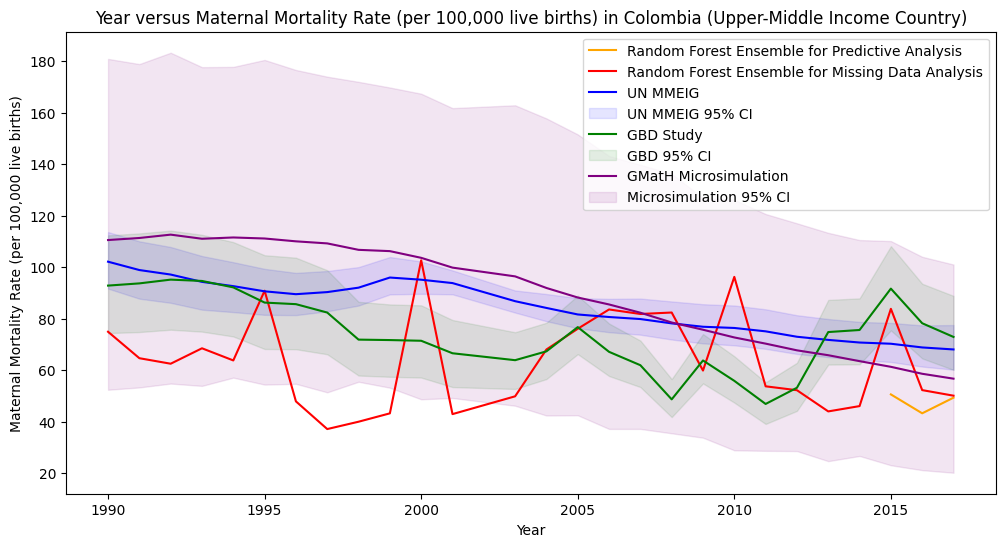
**Discussion of presentation of literature comparison**

* Choosing specific countries
* High income I underestimate and am outside the confidence intervals (except for GMatH, which takes up the entire plot – exclude?)
  + Potentially due to underestimation in supplied data, as documented by literature

A graph with numbers and lines

AI-generated content may be incorrect.

* Upper-middle:
  + What kind of analysis could I make here?
  + My results are generally less smooth, with peaks around when the proportion of missing data is smaller, potentially indicating that the model increases its estimates in response to missing data.



* Low and lower-middle income
  + Tend to overestimate the literature
  + Potentially due to low-quality or sparse data

A graph showing the number of birth rate

AI-generated content may be incorrect. A graph showing the number of birth rate

AI-generated content may be incorrect.

However, there does not appear to be any significant relationship between proportion of missing data and MMR ratio

A graph with colored dots

AI-generated content may be incorrect.

* Questions:
  + How to represent confidence intervals given that the GMatH intervals are so wide
  + I have plotted the ‘split by year’ and ‘split by country’ here on the same graph, but you can barely see the year because the test data is 2015-2018. Should I plot this separately?
    - What about for the poster?

**Questions about thesis structure**

* Should I give explanation of why I did not use SMOTE in the methods or **discussion**?
* Do I need to explain why I removed the country and year columns from the input dataset?
  + Do not need to explain why
* Should I put the results of the exploratory data analysis (e.g. PCA, trends in missing data, number of samples lost due to data cleaning (e.g. removing samples with a missing MMR estimate)) in the methods, results or discussion?
* Generally, would you recommend trying to combine figures for conciseness or spread them out for clarity and to highlight nuances?
* Should I give the features selected in each feature subset method in the methods or appendix? (Is this a uniform decision, or should I put the subsets with <45 features in the methods and the rest in the appendix?
  + Link github to text for data availability
  + Add section to methods giving code availability
* In the background information, I give an explanation about MMR and why estimation methods for MMR are important. Will I move this entire section to the Introduction, or do I keep as is and just put a summary in the introduction?