**Agenda:**

1. Github update:
   1. I have added you as collaborators to the new Github due to errors uploading to the previous.
2. Discussion of finetuned model comparison across different missing data thresholds.
   1. When data is split by country
   2. When data is split by year
3. Discussion of thesis questions
4. Discussion of conference
   1. Acceptable to change abstract?
5. Discussion of next steps

**Finetuned model comparison across different missing data thresholds**

*Process*

1. All data collected between 1985 and 2018 was taken from the merged dataset.
2. The following features were removed
   1. 'Number of maternal deaths’
   2. ‘Lifetime risk of maternal death (1 in: rate varies by country)’
   3. ‘Lifetime risk of maternal death (%)’
   4. Modelled MMR ratio
3. All rows missing a national MMR estimate were removed and all feature columns with no data were removed
   1. At this point, the dataset contained *2,816 samples and 723 columns*
4. This dataset was then split into train/test sets in a 90:10 ratio in two separate ways (generating two distinct train/test sets)
   1. Split by year
      1. All data collected between 1985 and 2014 was placed in the train set and all data collected between 2015 and 2018 was placed in the test set.
         1. I included data from 2015 in the test set, as it was one of the four years containing less than 50% missing data
         2. This produces an 88:12 split
      2. Purpose: to evaluate whether the machine learning models could predict the national MMR estimate for a region within the current study period
   2. Split by country
      1. Process:
         1. The dataset was split into four mini-datasets, one containing all the countries in a specific World Bank defined income level.
         2. Each of these datasets was split into train/test sets with a 90:10 split.
         3. Independence between sets was guaranteed by a specific country only being allowed to belong to either the train or test set.
         4. The 4 train sets were merged & the 4 test sets were merged.
      2. Purpose: to determine whether the trained model could be used to predict future MMR estimates given values for specific socioeconomic and health-related indicators.
5. Each of the train datasets were split into 5 cross-validation folds (80:20 split)
   1. The same country/year could only be in the one of the sets depending on whether the initial train/test split was done by country or year.
   2. Does the validation set need to be independent of the train set?
      1. For years, they are randomly chosen, so not guaranteed to be independent, because they depend on the previous?
6. Iteratively removing missing values
   1. Done individually for each train/validation split
   2. No removal, 95% threshold, 85% threshold
   3. This process was not applied to the validation and testing sets
      1. The models were evaluated on the validation and testing sets with no missing data removed so that all test/validation subsets are comparable.
         1. Is the highlighted sentence a robust enough justification to use in the methods?
7. Xgboost, LightGBM, and scikit learn’s random forest were trained on each train/validation split
   1. Their hyperparameters were fine-tuned over 300 Optuna trials, with the best trial decided by the lowest mean-squared error
   2. This random forest implementation was the only model that could not work with categorical data
      1. A master dictionary associating all country names with a specific number was created. This was used to encode the country names in each of the folds.
      2. This encoding was only done for random forest to ensure the other models could benefit from the categorical data’s full information.
8. The hyperparameters from the best, fine-tuned model for each combination of (train/test split, cross-validation fold, and missing data threshold) was downloaded from Gadi.
   1. The models were re-trained using these hyperparameters on the training data from the model’s associated fold and then evaluated on the associated test set.
   2. Evaluation metrics included: mean average error, mean squared error, root mean squared error, R2 score, and a symmetrical version of the MAPE score ()
   3. To confirm the RMSE process: is it just the square root of the mean square error?

*Split by Country*

*LightGBM*

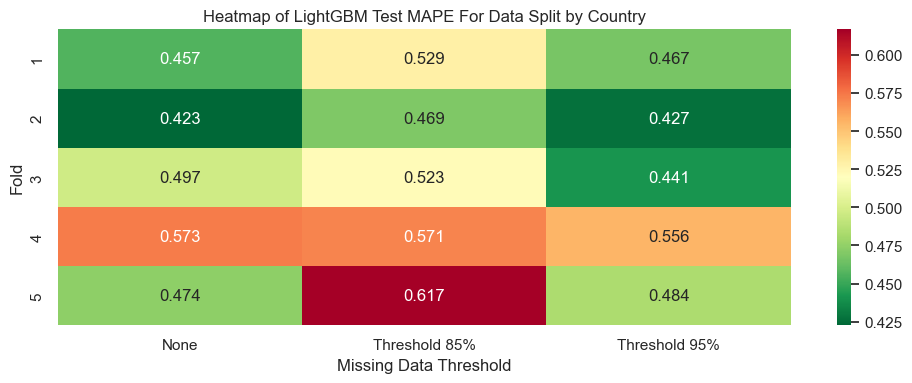
The following table shows the hyper-parameter choices of the fine-tuned models:

A screenshot of a computer

AI-generated content may be incorrect.

* The choice of number of trees and depth was not consistent across folds.
* Generally, the chosen boosting type was ‘dart’ (dropout added to the model)
* Number of trees only got close to max of 300 once
* Max tree depth got close to the max of 25 once
* Bagging frequency and fraction varied across their full range

Summary of heatmaps below: There was some observed differences in the evaluation metrics between folds. The ‘most different’ fold was not always the same across different missing data thresholds. How much variation is too much? Generally, the threshold 95% seemed to have the lowest variation between folds.



A chart of data loss

AI-generated content may be incorrect.

A chart with numbers and a number of data

AI-generated content may be incorrect.

A chart of a number of data

AI-generated content may be incorrect.

A chart of data loss

AI-generated content may be incorrect.

As shown below, the evaluation metrics are all highly correlated, as expected.

A screenshot of a graph

AI-generated content may be incorrect.

*XGBoost*

* As before, the number of trees and maximum tree depth varied substantially between folds. However, the maximum values did tend to be lower than those used to train the LightGBM models.
* While the DART boosting algorithm still dominated, gbtree was used more frequently.
* Learning rate varied by orders of magnitude between folds.

A screenshot of a table

AI-generated content may be incorrect.

As shown below, there are noticeable differences in the evaluation metrics across folds. Larger inter-fold differences appeared to occur when rows/columns with a greater than threshold proportion of missing data were removed.

A chart with numbers and a number of data

AI-generated content may be incorrect.

A chart of data loss

AI-generated content may be incorrect.

A green and red chart

AI-generated content may be incorrect.

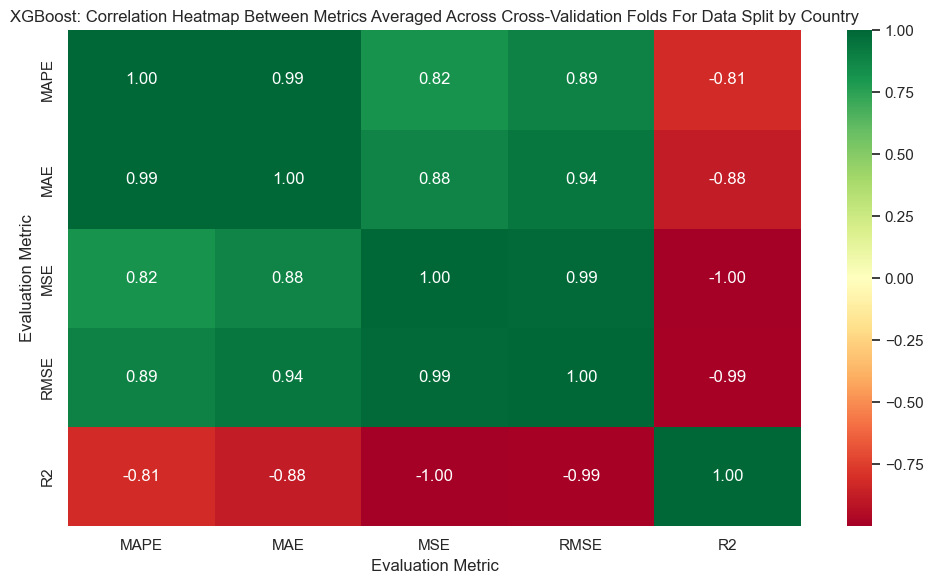
A chart of data loss

AI-generated content may be incorrect.

A chart of data loss

AI-generated content may be incorrect.

While the evaluation metrics were highly correlated, they were less strongly correlated than for LightGBM.



*Random Forest*

* The 95% threshold had the lowest number of estimarors, while the 85% threshold appeared to have the most instances of the highest. All 3 thresholds had fluctuating numbers of trees and maximum depths.
* Minimum number of samples required to split an internal node tended to be high
* Bootstrapping was almost always used.
* The proportion of samples used to train each new tree varied across folds.

A table with numbers and text

AI-generated content may be incorrect.

The models trained on data that had no missing data threshold had the lowest fluctuation in evaluation metrics across folds and the 85% threshold had the highest.

A chart of a forest test

AI-generated content may be incorrect.

A graph of data analysis

AI-generated content may be incorrect.

A green and red chart

AI-generated content may be incorrect.

A green and red chart

AI-generated content may be incorrect.

A graph of data analysis

AI-generated content may be incorrect.

Very strong correlation between evaluation metrics.

A screenshot of a graph

AI-generated content may be incorrect.

*Models Comparison*

Potentially, the 85% threshold had the highest variation because the data removed depends on what data is placed in train set.

I took the average and standard deviation of the evaluation metrics across the five cross-validation folds and presented them in a bar chart to enable model comparison.

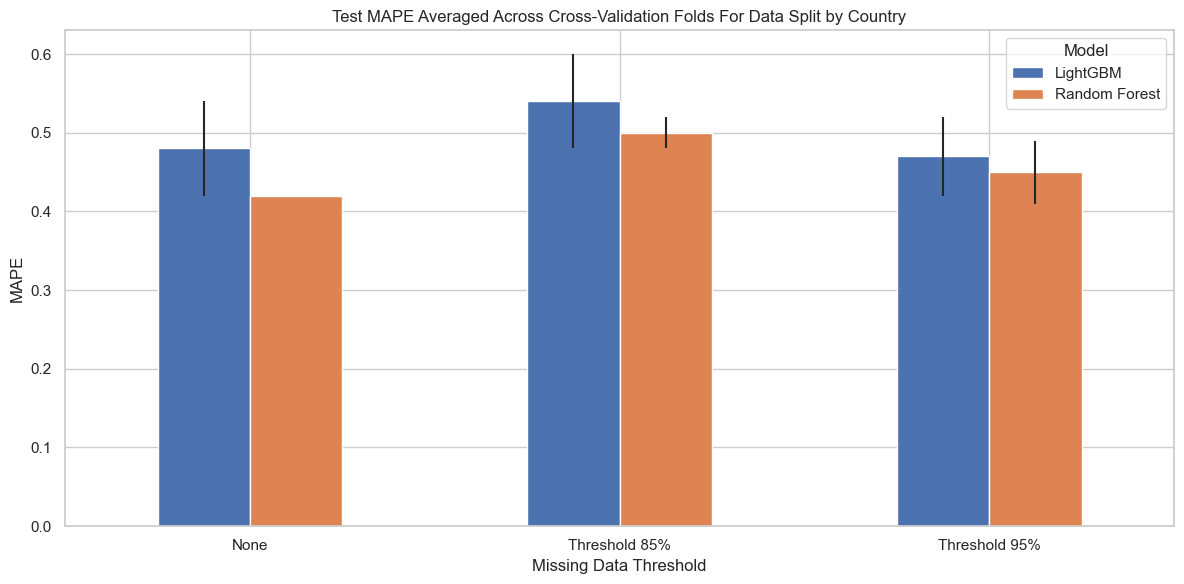
XGBoost performed the worst of the models over all evaluation metrics, as demonstrated in the following bar chart. Only one of the graphs is shown for concision.

A graph of a bar chart

AI-generated content may be incorrect.

To make the difference between LightGBM and Random Forest clearer, I removed XGBoost from the graphs.

* Threshold = 85% always performed the most poorly.
* Random Forest had the lowest MAPE and MAE scores for threshold = None.
* LightGBM had the best MSE, RMSE, and R2 scores for threshold=95%
* Standard deviations were comparable.



A graph of a bar graph

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

A graph with blue and orange bars

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

*Split by Year*

*LightGBM*

* Number and depth of trees varied across the folds, with higher numbers more common. (Given more at higher depths, should I increase depth above 25?)
* Boosting algorithm was often dart, unless threshold=95, when it was often gbdt.
* Bagging frequency tended to be lower and bagging fraction tended to be higher, but again, varied across their ranges.
* Learning rate was orders of magnitude higher than for other train/test split, and generally varied in the same order of magnitude.

A screenshot of a table

AI-generated content may be incorrect.

* Evaluation metrics again varied across the folds, with threshold=95 or none appearing to have the tightest range.

A chart with numbers and a number of data

AI-generated content may be incorrect.

A chart with numbers and a number of data

AI-generated content may be incorrect.

A chart of data loss

AI-generated content may be incorrect.

A chart of data analysis

AI-generated content may be incorrect.

**A chart with numbers and a number of data

AI-generated content may be incorrect.**

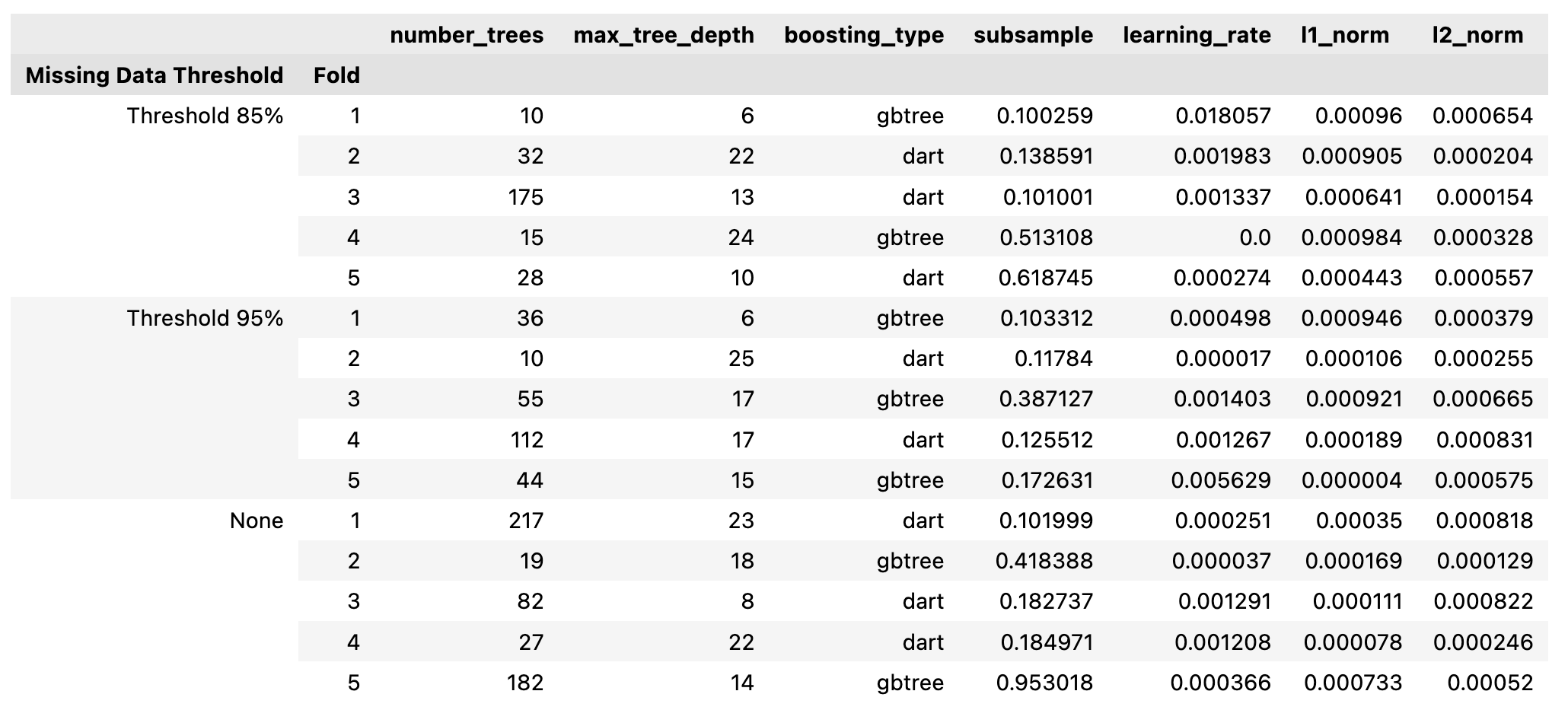
Strong correlation between metrics, although less strong than for split by country. Strong correlations between MAPE/MAE, RMSE/MSE, R2/RMSE, R2/MSE.

**A screenshot of a graph

AI-generated content may be incorrect.**

*XGBoost*

* N\_trees tended to be low/middle of range while depth varied across range.
* Subsample tended to be low.
* Only slight preference of dart over gbtree (often 3:2 for each type of threshold)
* Learning rate was lower than for LightGBM and varied over orders of magnitude more often.



* Only small fluctuations of metrics across folds, with smallest for threshold=none.
* Very low R2 values across the board.

A chart with numbers and a number of data

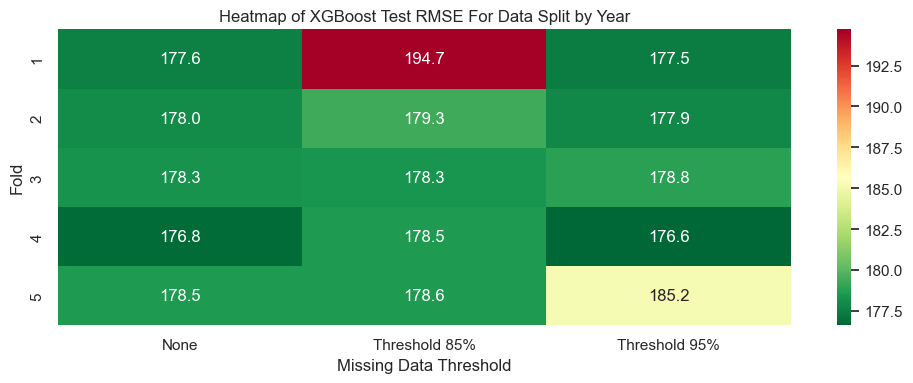
AI-generated content may be incorrect.

A chart with numbers and a red and green box

AI-generated content may be incorrect.

A green and red chart

AI-generated content may be incorrect.



A chart of data analysis

AI-generated content may be incorrect.

The evaluation metrics are almost perfectly correlated (more strongly than for LightGBM).

A screenshot of a graph

AI-generated content may be incorrect.

*Random Forest*

* Number of trees and depth ranged across folds, with depth tending to be smaller for threshold=85% and larger for threshold=none.
* Bootstrapping was always true, with threshold=85% having a smaller max\_samples proportion while the other two tended to have a larger proportion.
* Min-sample split tended to be low.

A table with numbers and letters

AI-generated content may be incorrect.

* Fluctuations across folds, with largest changes for threshold=85%.

A chart with numbers and a number of data

AI-generated content may be incorrect.

A graph of data with numbers

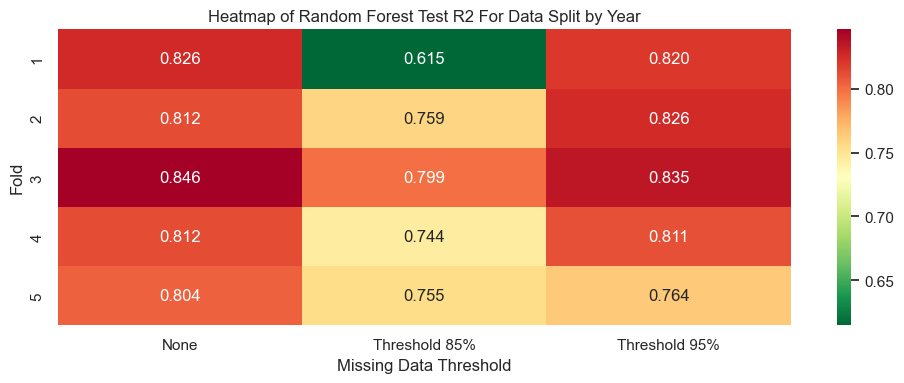
AI-generated content may be incorrect.

A chart of data loss

AI-generated content may be incorrect.

A chart with numbers and a red and green box

AI-generated content may be incorrect.



Almost perfect correlation between metrics (again larger than for LightGBM).

A screenshot of a graph

AI-generated content may be incorrect.

*Models Comparison*

Potentially, the 85% threshold had the highest variation because the data removed depends on what data is placed in train set.

As with the split by country train set, the XGBoost model performed substantially worse than the other models, as shown by the MAPE scores.

A graph of a bar graph

AI-generated content may be incorrect.

XGBoost was excluded from the following plots to better show differences between Random Forest and LightGBM.

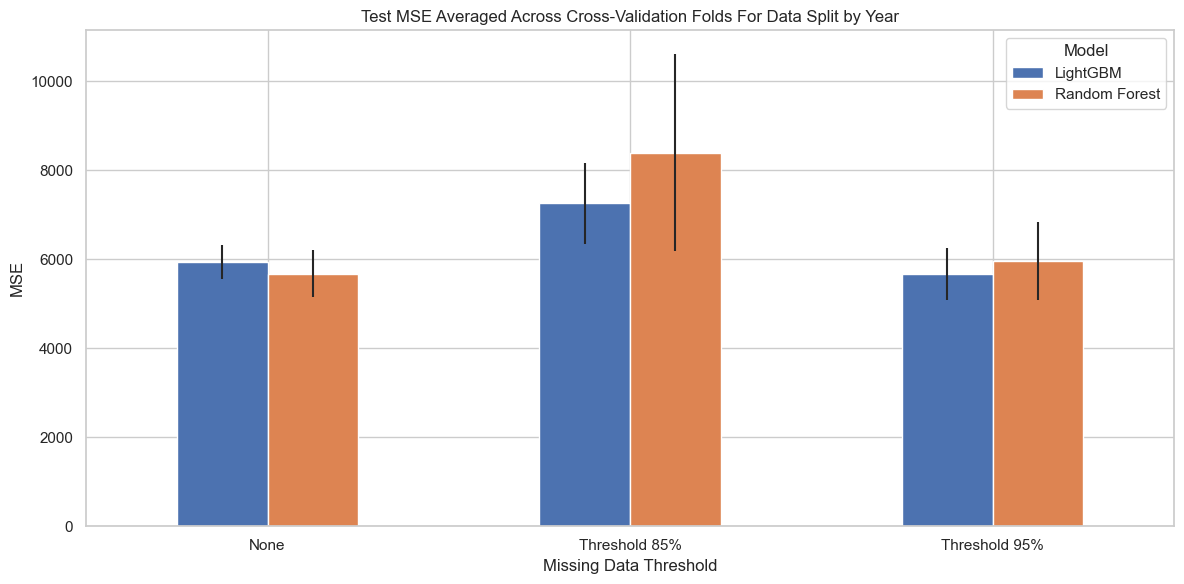
* While the difference is not as stark as for split by country, both models perform the most poorly in threshold=85%
* Random Forest performs better in terms of the MAPE score.
* LightGBM performs better in terms of the MAE score (threshold=95%)
* They perform very similarly for MSE, RMSE, R2
  + LightGBM performs better in threshold=95% while random forest performs better for threshold=None

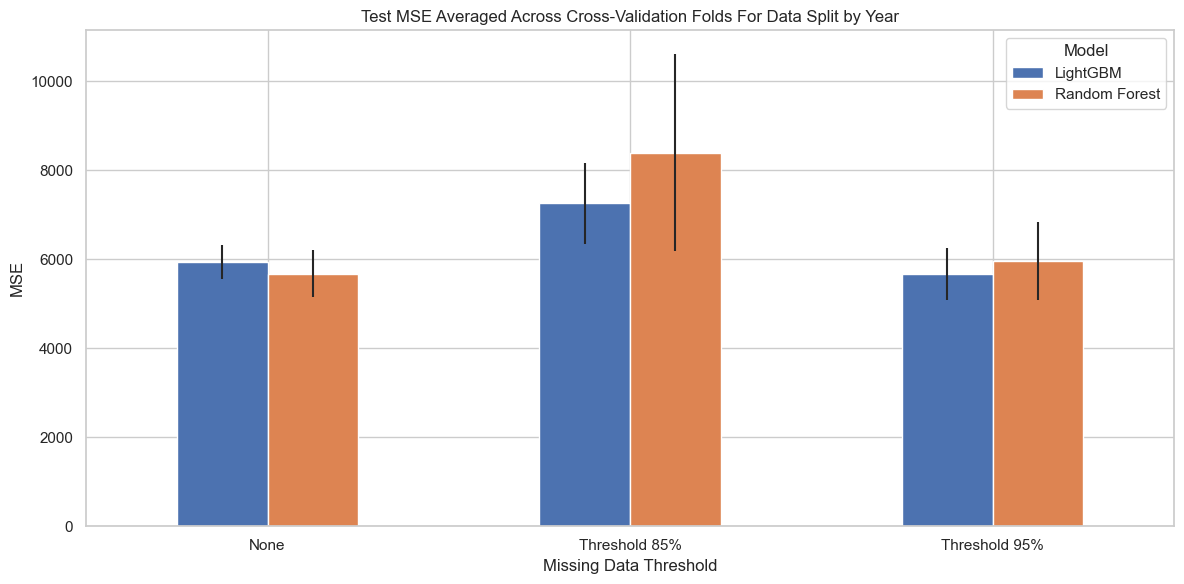
A graph of a bar chart

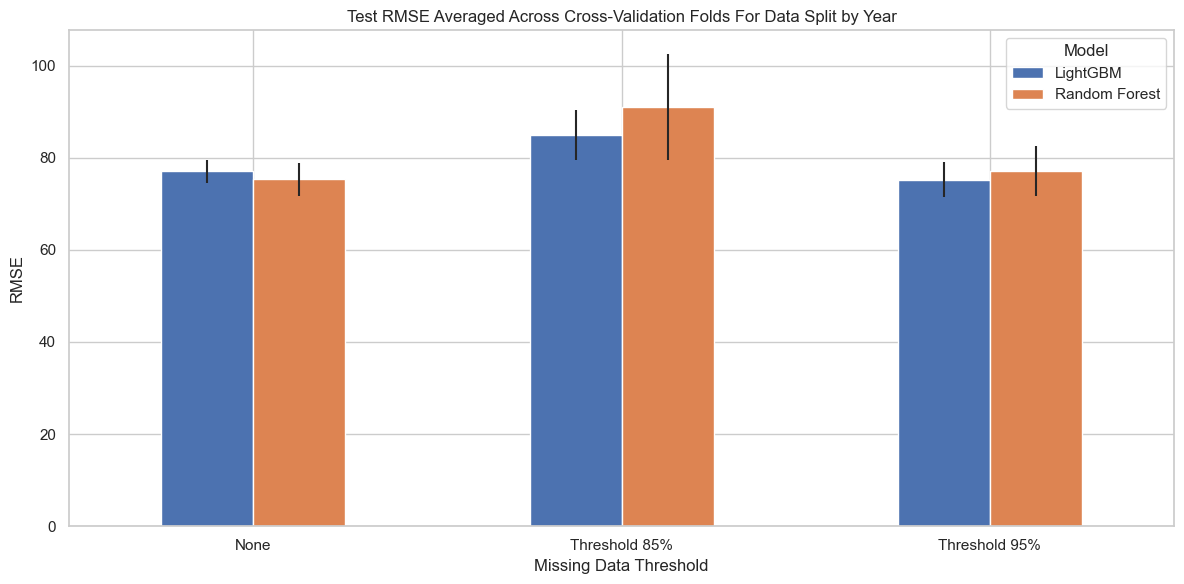
AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.







A graph of a bar graph

AI-generated content may be incorrect.

**Discussion of Thesis Questions**

* How would you recommend best presenting my results from above?
* Do you think I need to still do imputation?
* As an update on how the World Bank/WHO/UN modelled MMR
  + They used a multilevel hierarchical regression model.
    - No machine learning, solely statistics.
    - Small subset of variables I used.
  + (I will be discussing this in the related literature section)
  + Where should I be testing the national estimates versus the modelled estimates?
* Do I need to explicitly mention that I am not using imputation (given that so many other studies in public health do)?
  + Do I put this in methods or related work?
* In my methods, should I show an example plot for a specific fold of what thresholding does to rows and columns?
* How to refer to the mape score when its formula has been made symmetrical?

**Discussion of Conference**

* Ticketing
* Changes to data pipeline and project since submission of the abstract
  + Particularly with respect to size of dataset and breadth of models/imputation techniques used

**Discussion of Next Steps**

* Ensemble models
* SHAP values
* The following week: take best model and conduct sensitivity assessments
  + E.g. train and test on most recent 20 years
  + E.g. train and test on countries from only one income level