## 5. Evaluation

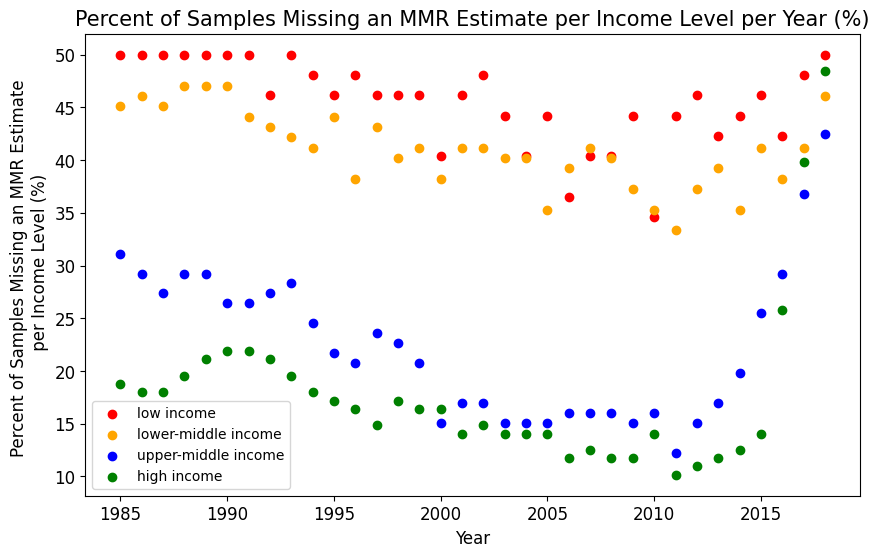
In this section, I first discuss the results of my exploratory data analysis (5.1) and investigation into the distribution of data between the training/validation and test sets (5.2). I then examined how Random Forest, XGBoost, and LightGBM models performed when fit on training datasets from different cross-validation folds and generated with different feature subsets and missing data removal strategies (5.3). Next, I explored the effect of incorporating these base estimators into different stacking and voting ensembles (5.4). Finally, the architecture (5.5) and performance (5.6) of the best-performing stacking ensemble were analysed, before the MMR estimates from this best-performing model were compared to existing MMR estimates from the BMat, CODEm, and GMatH models in the literature (5.7). Models’ MMR estimates for the country-level prediction and forecasting tasks were analysed separately.

### 5.1 Exploratory Data Analysis

First, I present all results from the exploratory data analysis that were not already given in the methods. These findings contextualised model performance.

#### 5.11 Analysis of Trends in Missing MMR Values

The proportion of country, year samples missing an associated MMR estimate out of all samples from the same income level was referred to as “the proportion of missing estimates” in the following analysis. This proportion varied widely across income levels, with the greatest difference observed between lower-middle and upper-middle income countries (Figure 10). The proportion of missing estimates decreased as income level increased. Additionally, this proportion decreased between 1985 and 2010 for each income level. For example, between 1985 and 2010, the proportion of missing estimates decreased from 50 to 35% in the low-income data, 45 to 35% in the lower-middle income data, 31 to 16% in the upper-middle data, and 19 to 14% in the high-income data. The proportion of missing estimates started increasing for all income levels post-2011, with the greatest increases observed in high and upper-middle income countries (38 and 30 percentage points, respectively).



**Figure 10:** The percent of samples in the input dataset missing MMR estimates before cleaning or pre-processing. Results were presented per year between 1985 and 2018 and per income-level (red for low-income, orange for lower-middle, blue for upper-middle, and green for high).

#### 5.12 Key Statistics in the Merged Input Data Before Pre-Processing

To better understand the input data, I calculated key summary statistics about a few of the features indicated by the literature to have a meaningful relationship with MMR (Table 10). Generally, health outcomes improved as income level increased. Standard deviation in the feature decreased as income level decreased. While many of the important variables had low rates of missing data, some of the socio-economic and quality of care features had increasing proportions of missing data for higher income levels. For example, the dataset for the lowest income countries was missing 58% of measurements for ‘women participating in own health care decisions (% of women age 15-49)’ while the highest income dataset was missing 99.9%.

The national, ground truth MMR estimates were subject to large outliers, as the mean values were larger than the median values for all income levels (Table 10). Additionally, standard deviation for the MMR estimates was large. The difference between mean and median, as well as the magnitude of standard deviation, decreased as income level increased.

**Table 10:** Mean, median, standard deviation and proportion of missing data of features with a meaningful relationship to MMR. The key summary statistics were presented per income level.

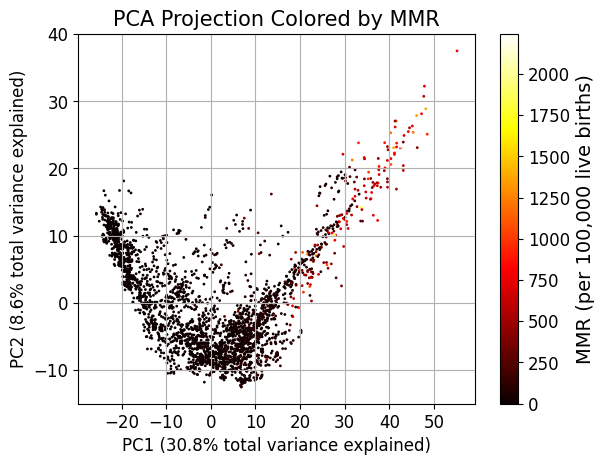
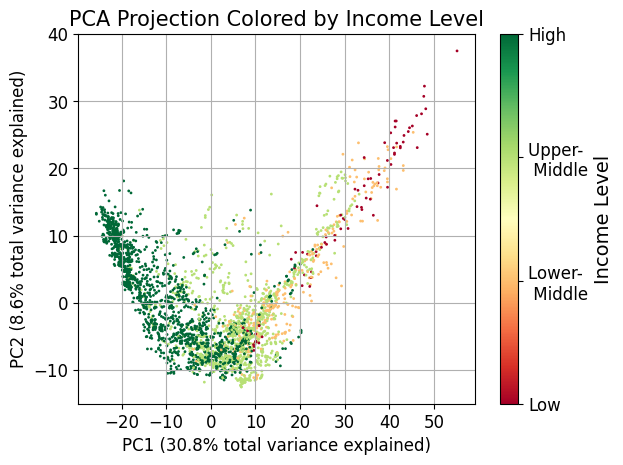
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Feature** | **Income Level** | **Mean** | **Median** | **Standard Deviation** | **Proportion of Missing Data (%)** |
| WHO national MMR estimates (ground truth) | Low | 657 | 617 | 453 | 0 |
| Upper-Middle | 197 | 55 | 260 | 0 |
| Lower-Middle | 51 | 38 | 55 | 0 |
| High | 15 | 8 | 21 | 0 |
| Infant mortality rate (per 1,000 live births) | Low | 63 | 65 | 29 | 0 |
| Upper-Middle | 43 | 39 | 23 | 0 |
| Lower-Middle | 24 | 19 | 15 | 0 |
| High | 9 | 7 | 7 | 2 |
| Pregnant women receiving prenatal care (%) | Low | 74 | 85 | 23 | 28 |
| Upper-Middle | 81 | 86 | 18 | 65 |
| Lower-Middle | 92 | 96 | 10 | 78 |
| High | 93 | 97 | 8 | 95 |
| Women participating in own health care decisions (% of women age 15-49) | Low | 55 | 60 | 22 | 58 |
| Upper-Middle | 65 | 67 | 22 | 86 |
| Lower-Middle | 84 | 84 | 8.7 | 97 |
| High | 91 | 91 | NaN | 99.9 |
| Communicable, maternal, neonatal, & nutritional diseases prevalence in females (age standardized, per 100,000 population) | Low | 79,399 | 84,661 | 14,140 | 77 |
| Upper-Middle | 73,030 | 73,279 | 9,389 | 84 |
| Lower-Middle | 62,248 | 63,092 | 10,658 | 86 |
| High | 38,835 | 36,807 | 11,821 | 87 |
| Survival to age 65, female (% of cohort) | Low | 59 | 58 | 13 | 0 |
| Lower-Middle | 71 | 73 | 12 | 0 |
| Upper-Middle | 79 | 80 | 8 | 0 |
| High | 87 | 88 | 5 | 0 |
| Unmet need for contraception | Low | 27 | 28 | 7 | 33 |
| Upper-Middle | 22 | 23 | 8 | 72 |
| Lower-Middle | 13 | 12 | 6 | 88 |
| High | 13 | 10 | 10 | 97 |

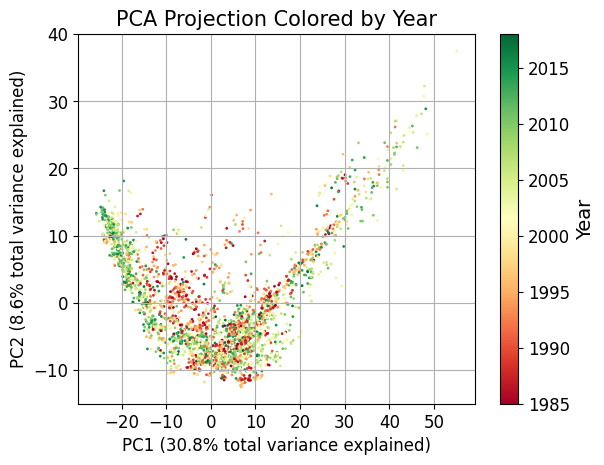
#### 5.13 Principal Component Analysis

The feature dataset was projected onto its first two principal components to better visualise patterns and clusters in the data. The dense cluster in the bottom center-left of Figures 11a and 11b corresponded to higher income countries with low MMRs. This cluster extended upwards to the plot’s centre-left. A country’s income level tended to decrease and its MMR tended to increase travelling up and to the right of Figures 11a and 11b. There were no similarly large clusters when the datapoints were coloured by year (Figure 11c). However, there was slight clustering at the leftmost and rightmost edges of the U-shape, which corresponded to more recent years. The inner-part of the U above the valley represented years further in the past.

a)

b)

****

c)

**Figure 11:** PCA projection of the feature dataset onto its first 2 principal components, which captured 30.8 and 8.6% of the data’s total variation, respectively. The PCA projection was coloured by the samples’ a) MMR, b) income level, and c) year.

### 5.2 Data Distribution Between the Train/Validation and Test Sets

This sub-section compared the ground truth MMR distribution in the train/validation and test sets to provide foundation for later discussions about model performance. This comparison was performed separately for the datasets used to conduct country-level prediction and forecasting.

#### 5.21 Country-Level Prediction

The ground truth MMR distributions in the train/validation and test sets overlapped for lower-middle, upper-middle, and high-income countries (Figure 12). More specifically, these datasets had similar quartile 2 (Q2) MMR values for both their train/validation and test sets. However, the Q3 and maximum MMR values in these countries’ train/validation data were higher than in their test data. For example, the Q3 MMR for lower-middle countries was 283 in the train/validation set versus 60 in the test set. The differences between train/validation and test MMR data were greater for lower-middle income countries than upper-middle and high-income. The train/validation set for these three income levels also contained outliers with higher MMR values than the associated test sets.

There was a greater difference between MMR estimates in the train/validation and test sets filtered for low-income data. For example, the train/validation data had a Q2 MMR value of 610 while the test set had a Q2 of 772. The test set’s Q3 and Q1 similarly exceeded the train/validation set’s Q3 and Q1 by 126 and 103, respectively. As a result, the test set’s MMR distribution was shifted higher than the train set’s distribution. Given the small number of low-income samples, which cover a wide range of MMR values (Table 10), samples with high MMRs may have been included in the test set by chance. However, the maximum MMR value in the low-income train/validation data was greater the maximum in the corresponding test set.

A graph of a graph showing the value of a company

AI-generated content may be incorrect.

**Figure 12:** Minimum, quartile (Q) 1, Q2, Q3, and maximum ground truth MMR for the train/validation and test sets used for country-level prediction. The train/validation and test sets were filtered for each income level. The y-axis was shown with a log-scale. The boxplot bars representing the minimum MMR sometimes appeared cut-off because they extended to zero.

#### 5.22 Forecasting

The income-level specific test distributions were not within the corresponding train/validation distributions (Figure 13). The low-income data’s train/validation set had smaller Q1 and Q3 values than its associated test set. In contrast, the train/validation sets filtered for lower-middle and high-income data had larger Q1 and Q3 values than their associated test sets. For example, the high-income test set’s Q3 exceeded the Q3 in the associated train/validation set by 27. The test set filtered for upper-middle income data had a higher Q3 and lower Q1 than the associated train/validation set, meaning the test set completely encompassed the train/validation set.

The test set’s Q2 MMR value was lower than the train/validation set’s Q2 MMR value for all income levels. The difference was greatest for the low-income data, where the test set’s Q2 was 104 points larger than the train/validation set’s Q2. In contrast, this difference ranged between 3 and 15 for lower-middle, upper-middle, and high-income countries. The outliers and maximum MMR values in the test sets were often larger than those in the train/validation sets.

The offsets between the train/validation and test distributions meant the samples used to assess the model’s out-of-sample performance were not reflected in its training data.

A graph with different colored squares

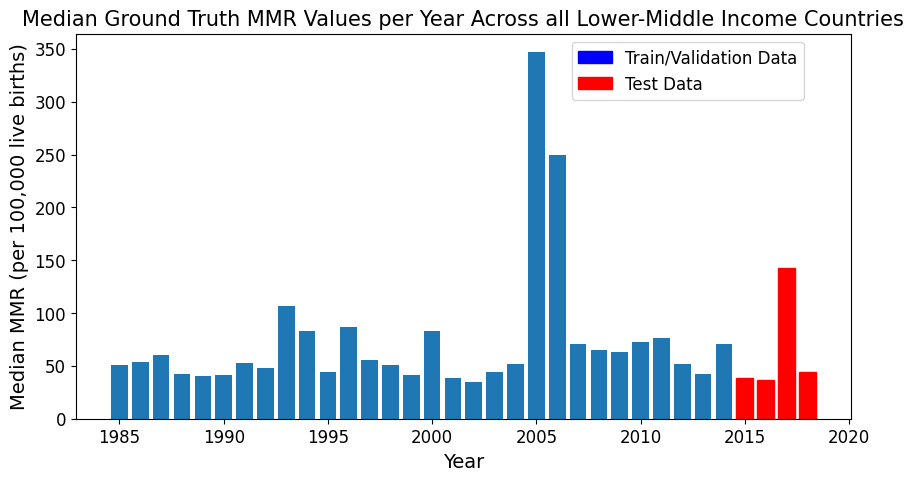
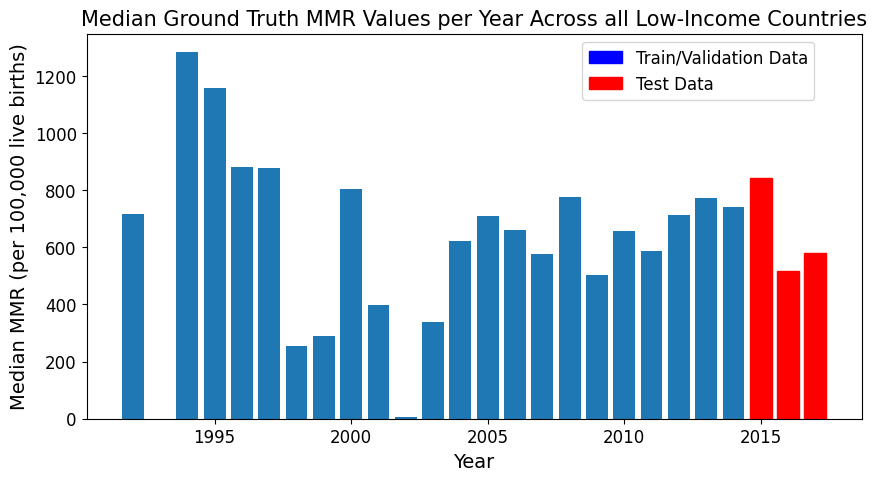
AI-generated content may be incorrect.

**Figure 13:** Minimum, quartile (Q) 1, Q2, Q3, and maximum ground truth MMR for the train/validation and test sets used to perform forecasting. The train/validation and test sets were filtered for each income level. The y-axis was shown with a log-scale. The boxplot bars representing the minimum MMR sometimes appeared cut-off because they extended to zero.

The test set filtered for low-income data had median MMR values within the spread of the associated train/validation set (Figure 14a). However, it did not have examples of the train/validation set’s sporadic decreases in MMR, potentially explaining why the test set’s Q1 was higher than the train/validation set’s Q1. The test set for lower-middle, upper-middle, and high-income countries all contained an outlier year (2017 or 2018) with a much higher MMR value than typically observed in the train/validation set (Figures 14b, 14c, and 14d). While the lower-middle income test set’s outlier was contained within the train/validation distribution, the outlier years from the upper-middle and high-income test sets were not. Nevertheless, the other years in the upper-middle and high-income test sets were generally lower the train/validation set, especially for the high-income test set. This explains why the Q1 and Q2 metrics for the upper-middle and high-income test sets were lower than the Q1 and Q2 values in the associated train/validation sets, despite the presence of the outlier year.

b)

a)



A graph of blue and red bars

AI-generated content may be incorrect.

d)

c)

**Figure 14:** Median ground truth MMR per year for a) low, b) lower-middle, c) upper-middle, and d) high-income countries in the train/validation (blue) and test (red) sets used for forecasting.

I further investigated the outlier year (2018) in the high-income data due to the large discrepancy between the train/validation and test sets. Only two countries (Oman and Uruguay) had non-NAN test MMRs for 2018. These countries’ ground truth MMRs were larger than the norm for high-income countries, consistently ranging between 14 and 25 throughout the test set (Table 11). In the non-outlier test years, the median ground truth MMR did not reflect these high values because it was brought down by low MMRs in other high-income countries. For example, in 2015 Australia and Norway had ground truth MMRs of 3 and 0, respectively. Thus, the large median ground truth MMR observed in 2018 for the high-income countries was not due to a change in circumstances within a certain country. It was solely due to data only being reported from countries with MMR values on the higher end of the spectrum.

**Table 11:** The ground truth MMR values for Oman and Uruguay between 2015 and 2015. These were the only two high-income countries with non-NAN data for 2018.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **Ground Truth MMR Value (per 100,000 live births)** | | | |
| **2015** | **2016** | **2017** | **2018** |
| Oman | 21 | 20 | 14 | 23 |
| Uruguay | 19 | 25 | 19 | 17 |

### 5.3 Performance of Single Random Forest, XGBoost and LightGBM Models

This section discusses the performance of Random Forest, XGBoost, and LightGBM models trained on cross-validation data curated with different feature subsets and missing data removal techniques. As described in Section 4.412, performance was measured using mean relative error (MRE), MSE, RMSE, MAE, and R2. However, to keep this chapter concise, only MRE and MSE were presented, with the other metrics given in Appendix 9.1. While MRE described the model’s error as a proportion of the ground truth and prediction values, MSE provided information about the model’s performance when predicting for outliers. Combining the two metrics provided a comprehensive measure of model performance.

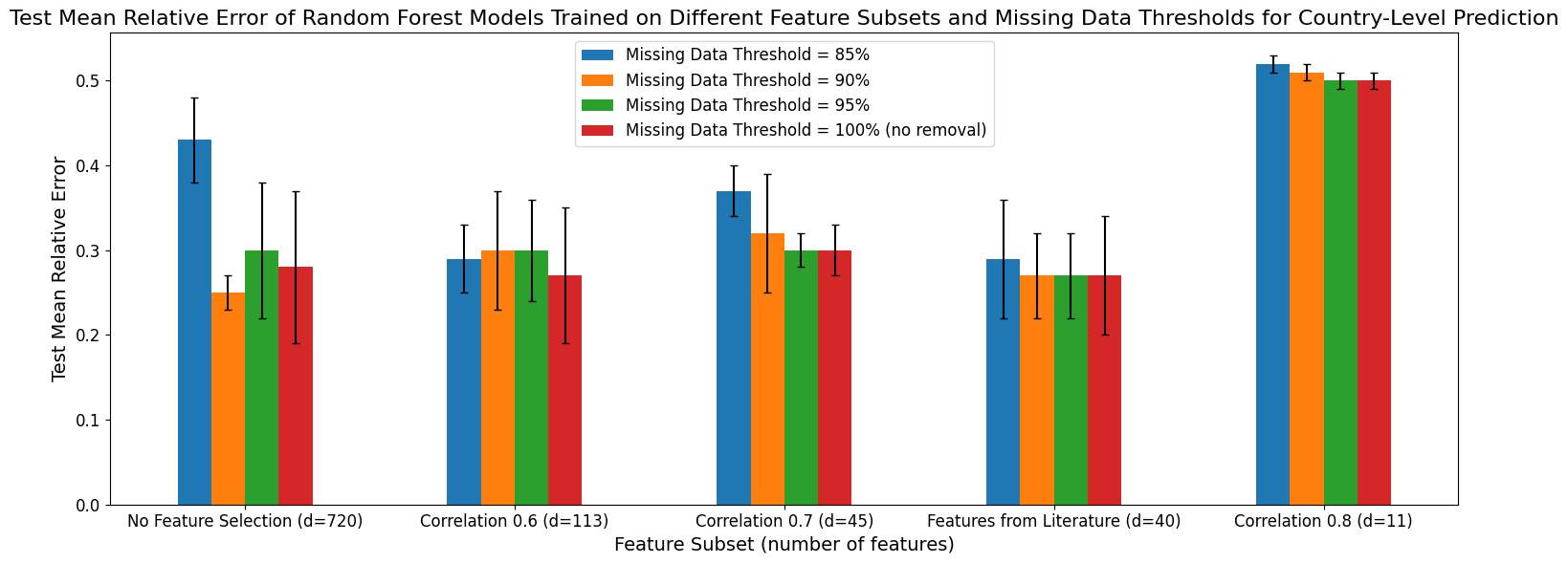
#### 5.31 Base Estimator Performance on Different Feature Subsets and Missing Data Removal Thresholds for Country-Level Prediction

##### 5.311: Random Forest

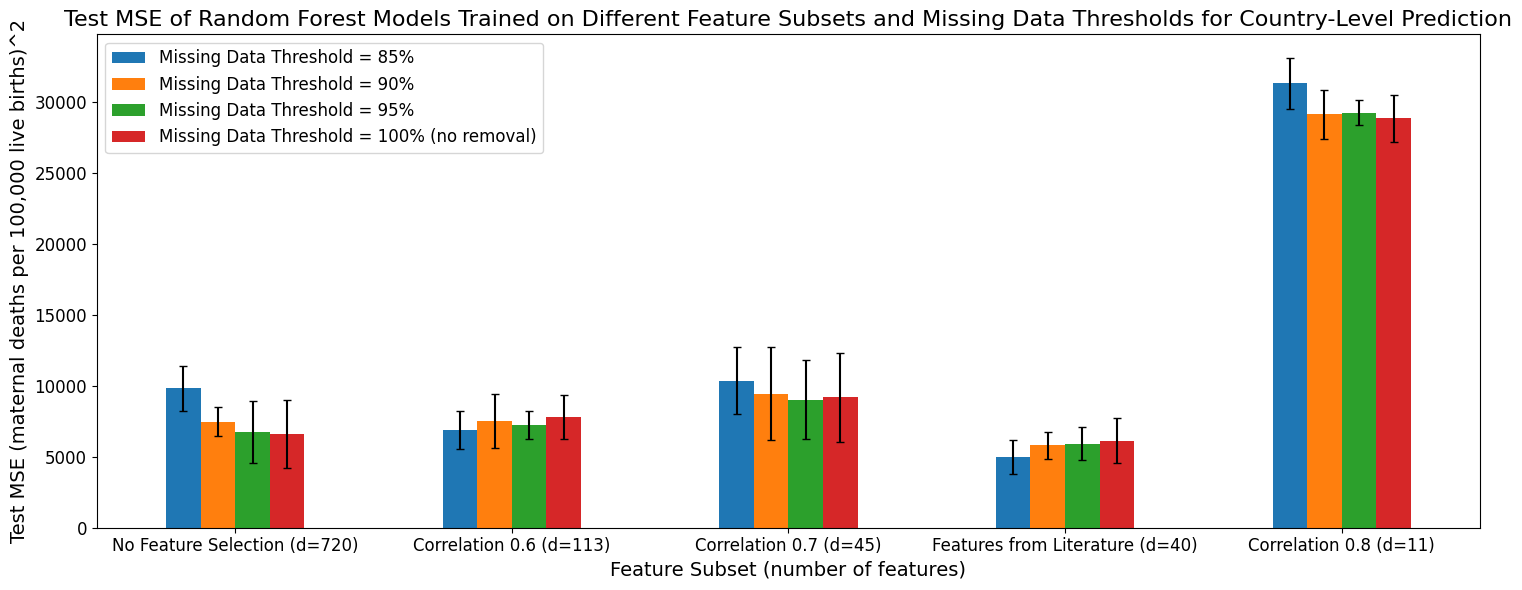
Random Forest models trained on different feature subsets generally had similar performance, especially when accounting for standard deviation in their performance (Figure 15). The models’ MRE and MSE typically ranged from 0.25 to 0.32 and 5,000 to 10,000 across the different feature subsets, respectively. The exception was models trained only on the ‘Correlation 0.8’ feature subset. In this case, models had notably lower predictive performance, with mean relative error at least 0.5 and MSE at least 28,000. Models trained on the ‘Correlation 0.7’ feature subset generally had the second largest errors across both metrics.

The Random Forest models with the lowest MRE (0.25) were trained on datasets formed without feature selection and with a missing data threshold of 90%. However, the models with the lowest MSE (4,986) were trained on the subset of features hand-picked from the literature and with a missing data threshold of 85%. The set of features hand-picked from the literature was most consistently associated with low MSE scores, and to a lesser degree low MRE scores. Thus, models trained on this feature subset may have had less outlier-induced error.

The relative ordering of best to worst performing missing data threshold changed when considering MSE versus MRE. For example, Random Forest models trained on a missing data threshold of 85% typically had higher MRE than models trained on higher missing data thresholds. Additionally, models trained with no missing data removal had the lowest, or tied for the lowest, MRE. In contrast, models trained with a missing data threshold of 85% had both the highest and lowest MSE scores, depending on the feature selection method. The same applied for no missing data removal. Therefore, outliers may have had a greater effect on models trained with a missing data threshold of 85% than models trained with no missing data removal. However, this comparison must be taken with caution, as the standard deviation in both error metrics always overlapped with that of models trained on other thresholds.



a)



b)

**Figure 15:** a) Mean relative error and b) mean-squared error for Random Forest base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

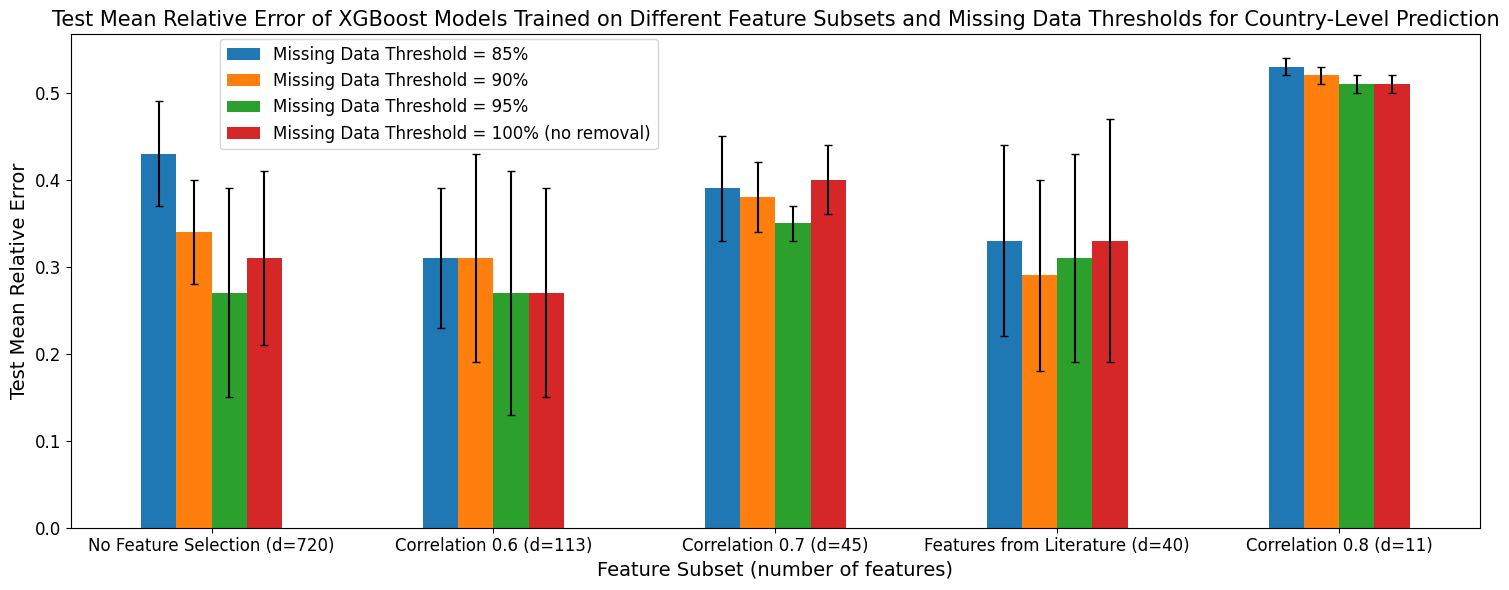
##### 5.312: XGBoost

The XGBoost models had similar trends in their performance as the Random Forest models (Figure 16). For example, the XGBoost models also incurred their highest MRE and MSE scores when trained on the ‘Correlation 0.8’ feature subset. Additionally, they generally had the second highest error in both metrics when trained on the ‘Correlation 0.7’ feature subset and tended to have lower MSE when trained with the hand-picked feature subset. Similar to the Random Forest models, XGBoost models trained on a missing data threshold of 85% had the highest MRE across most feature subsets with no consistent trend observed for MSE.

Excluding the high-error models trained on the ‘Correlation 0.8’ feature subset, the MRE and MSE of XGBoost models ranged from 0.27 to 0.43, and 4,000 to 10,000, respectively. The MRE range’s lower and upper bounds were higher than for the Random Forest models. The MSE range’s lower bound was slightly smaller than for the Random Forest models. Three XGBoost models had the same lowest MRE score (0.27). They were trained on the ‘Correlation 0.6’ feature subset (missing data thresholds 95% and 100%) and no feature selection (missing data threshold 95%). The models with the lowest MSE (4,185) were trained with the hand-picked feature subset and a 90% missing data threshold.

One of the major differences between the XGBoost and Random Forest models was their magnitude of standard deviation, with XGBoost models showing larger differences between their performance on different cross-validation folds. For example, standard deviation in the MSE of XGBoost models trained with no feature selection ranged from 2,271 to 5,037. In contrast, this standard deviation varied between 1,021 and 2,379 for Random Forest models.

As observed for the Random Forest models, there was no universally best performing feature subset or missing data threshold, especially given XGBoost models’ wide standard deviations.



a)

b)

A graph of different colored bars

AI-generated content may be incorrect.

**Figure 16:** a) Mean relative error and b) mean-squared error for XGBoost base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

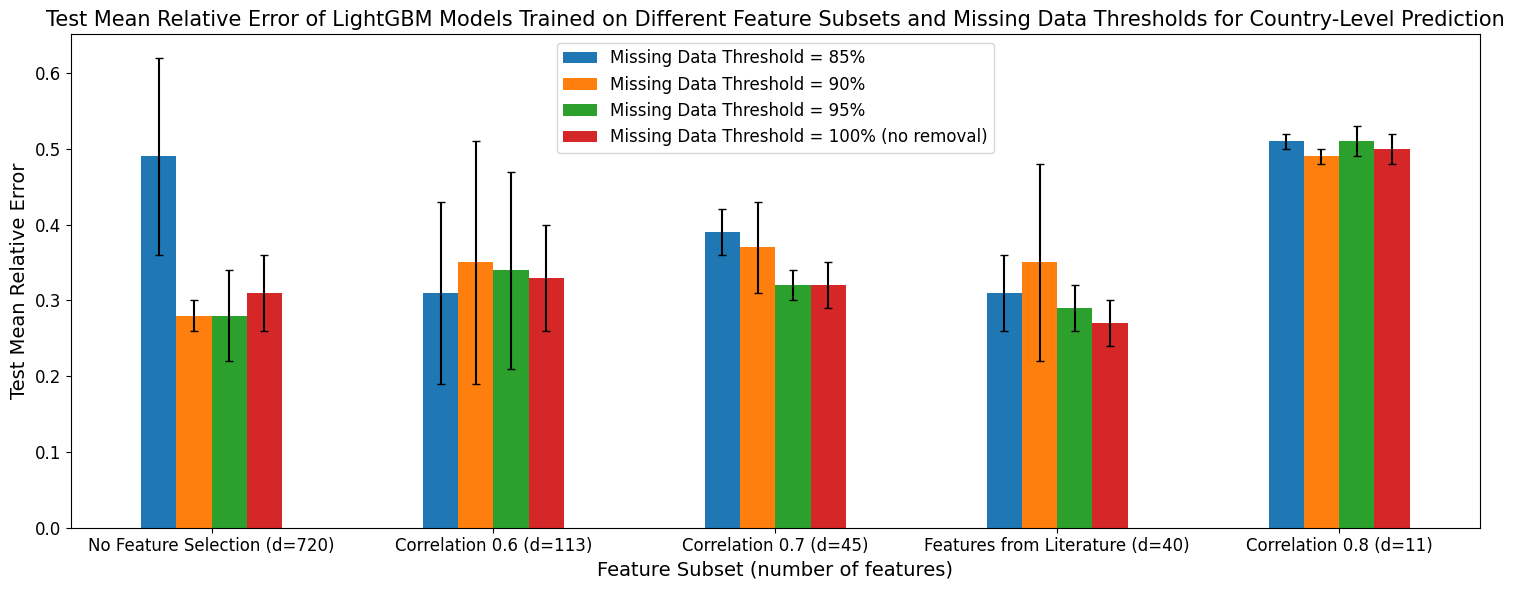
##### 5.313: LightGBM

The LightGBM models had similar performance trends as XGBoost and Random Forest. For example, they had their worst performance on the ‘Correlation 0.8’ feature subset and among their worst performance on the ‘Correlation 0.7’ subset. Additionally, models trained on a missing data threshold of 85% and no feature selection had the worst MRE performance across all three model types (Figures 15, 16, 17). Similar to the XGBoost models and unlike Random Forest models, LightGBM models did not consistently have higher performance on the hand-picked feature subset. As with the Random Forest and XGBoost models, LightGBM models did not have a consistently best performing missing data threshold or feature subset.

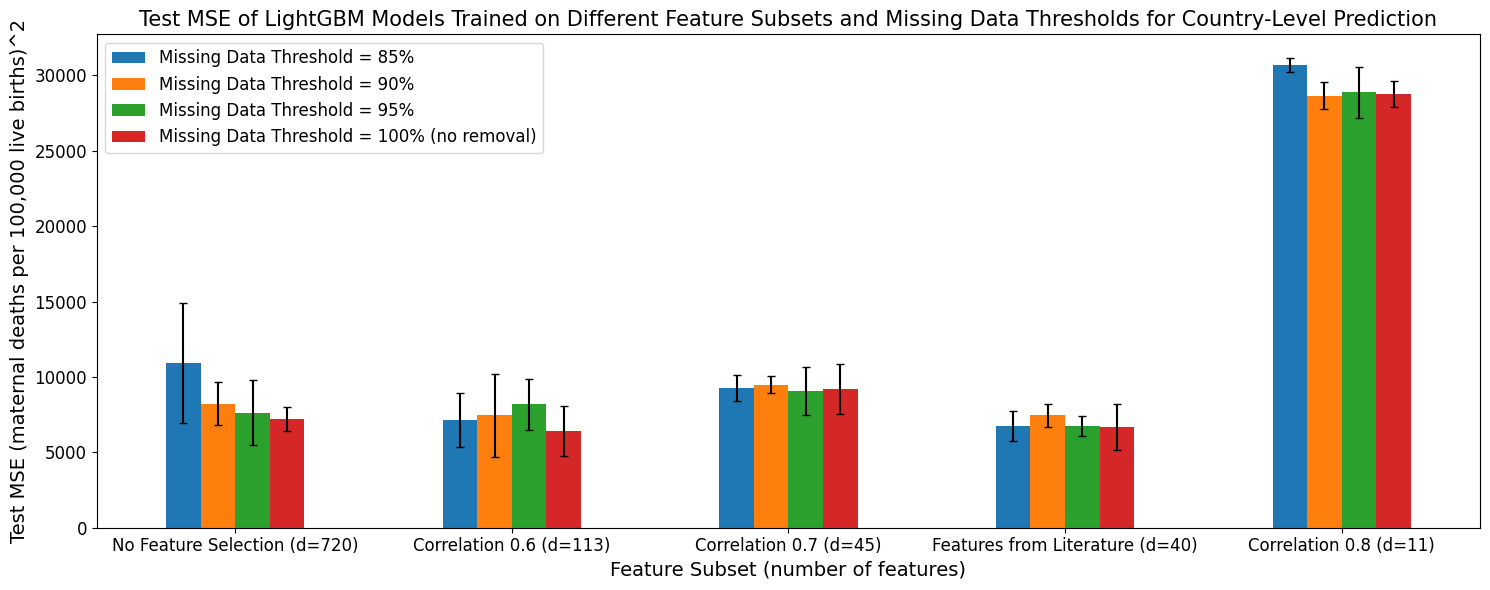
Excluding performance on ‘Correlation 0.8’ feature subsets, LightGBM models had MRE between 0.27 and 0.49 and MSE between 6,000 and 11,000. Both ranges were higher than for the Random Forest and XGBoost models, although the lower bound of the MRE range was the same as for XGBoost. The standard deviation in LightGBM’s performance was smaller than for the XGBoost models but higher than for the Random Forest models. For instance, the standard deviation in MSE for LightGBM models trained with no feature selection ranged from 777 to 3,989, compared to 2,271 to 5,037 for XGBoost and 1,021 to 2,379 for Random Forest.

The LightGBM models with the lowest MRE were trained on the hand-picked feature subset with no missing data threshold (0.27). In contrast, the LightGBM models with the lowest MSE were trained with no feature selection and with a missing data threshold of 95%. These combinations of pre-processing techniques also produced the best performing XGBoost models. However, wide standard deviations in error prevented these techniques from being conclusively designated as the highest performing combination, especially given they did not produce the best performing Random Forest models.

a)



b)



**Figure 17:** a) Mean relative error and b) mean-squared error for LightGBM base estimators fit on different feature subsets and missing data thresholds for country-level prediction.

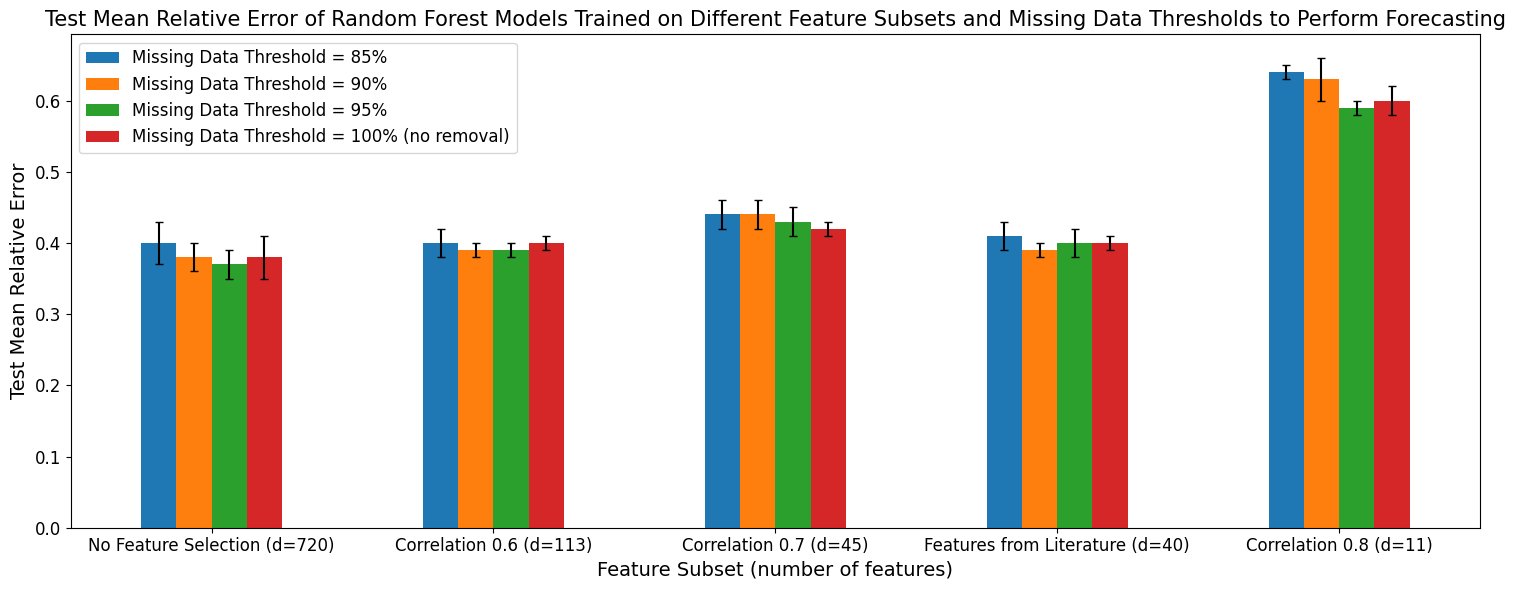
#### 5.32 Base Estimator Performance on Different Feature Subsets and Missing Data Removal Thresholds for Forecasting

##### 5.321: Random Forest

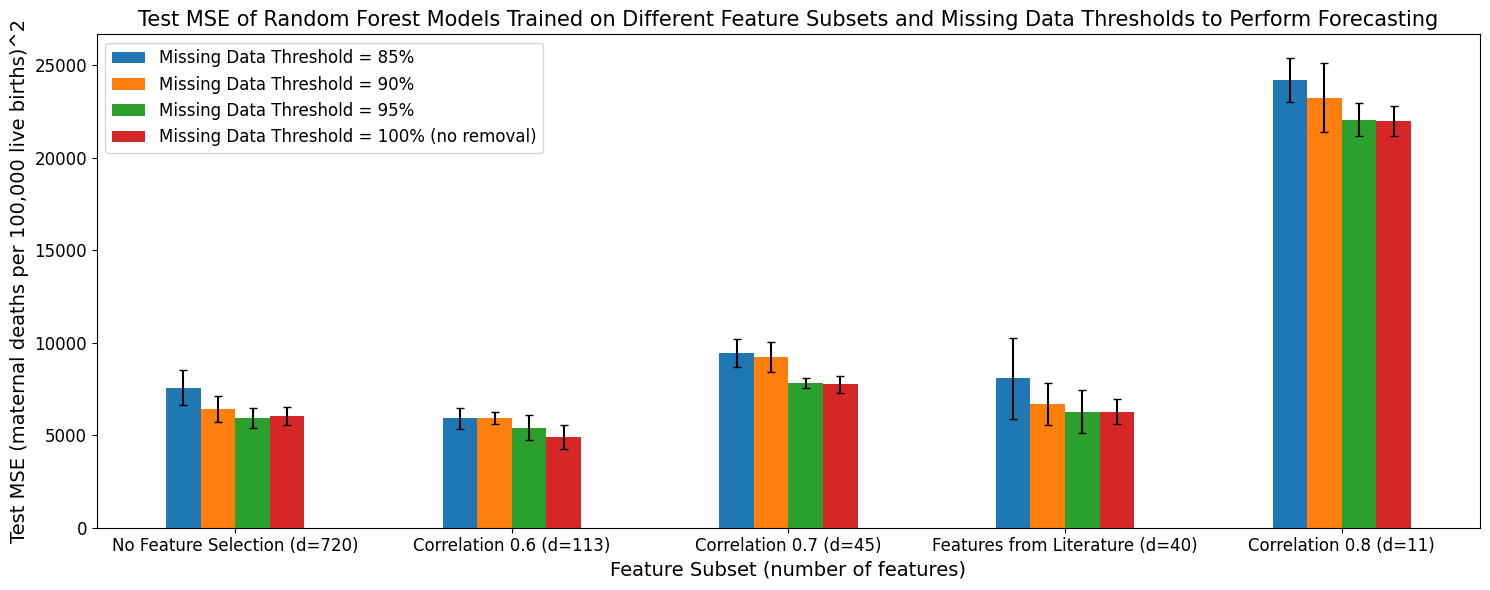
Random Forest models trained to perform forecasting using different feature subsets had very similar MRE scores, with MRE ranging from 0.37 to 0.40 (excluding models trained with the ‘Correlation 0.8’ feature subset) (Figure 18a). The Random Forest models with the lowest MRE (0.37) were trained with no feature selection and a missing data threshold of 95%.

There was more variation in the models’ MSE scores, indicating differences in the effect of outliers on the different pre-processing methods (Figure 18b). The Random Forest models’ MSE scores varied from 4,900 to 9,500, excluding errors from models trained on the ‘Correlation 0.8’ subset. The ‘Correlation 0.6’ feature subset generally produced the lowest MSE scores (all below 6,000), with its stronger performance more consistent when measured with MSE than MRE. Thus, it may more effectively handle outliers.

The Random Forest models trained to perform forecasting had the highest MRE and MSE scores on the ‘Correlation 0.8’ feature subset, like the models used for country-level prediction. As previously observed, the standard deviation in the error metrics prevented one missing data threshold from consistently producing the highest model performance.



a)



b)

**Figure 18:** a) Mean relative error and b) mean-squared error for Random Forest base estimators fit on different feature subsets and missing data thresholds for forecasting.

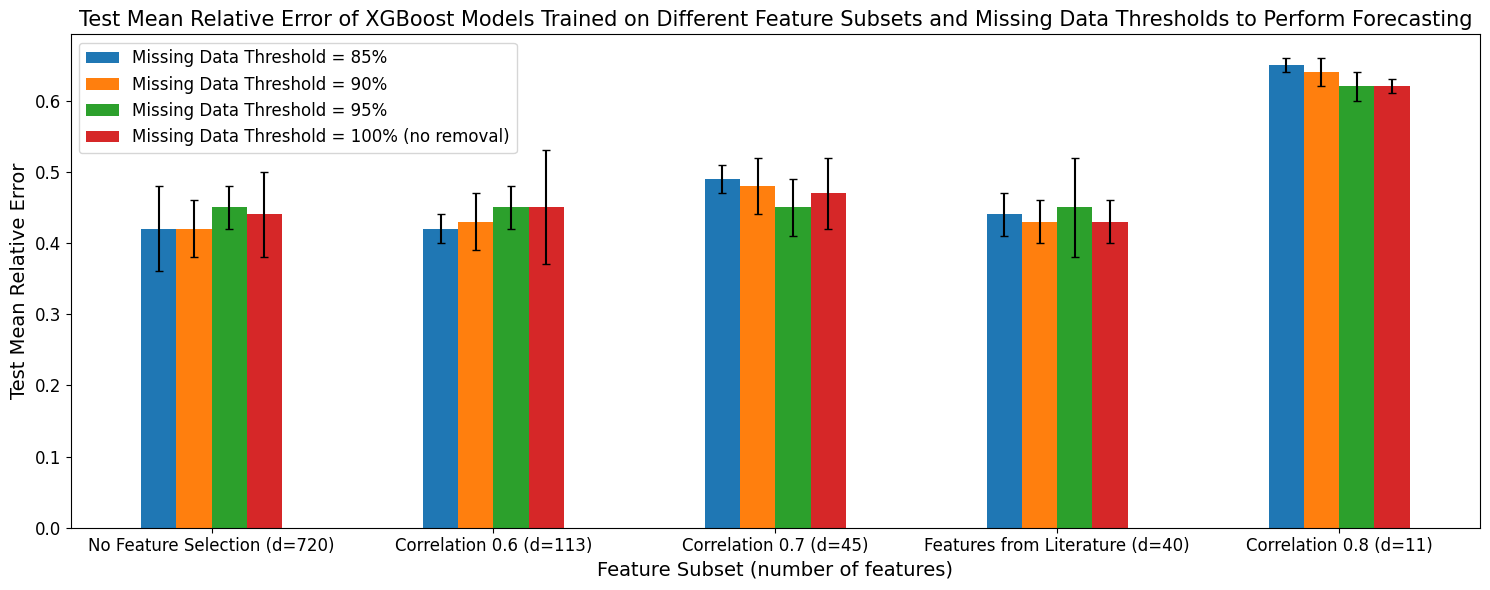
##### 5.322: XGBoost

Similar to the Random Forest models discussed above, the XGBoost models trained to perform forecasting had very similar MRE scores across the different feature subsets (Figure 19a). More specifically, when excluding the low performance ‘Correlation 0.8’ feature subset, their MRE ranged from 0.42 to 0.49, which was higher than that observed for the Random Forest models used for forecasting. This lack of variation made it difficult to identify a feature subset and missing data threshold that consistently had the lowest MRE, especially when taking into account the standard deviation in each error estimate across the cross-validation folds.

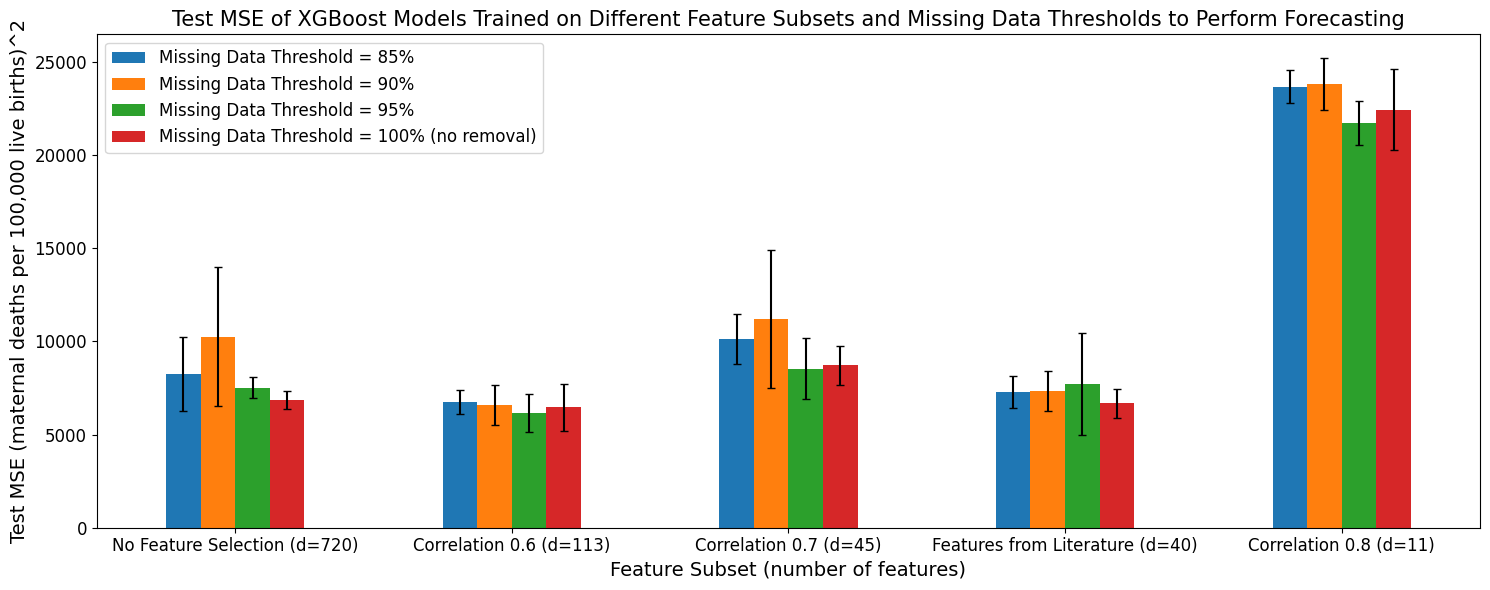
The MSE score for XGBoost models ranged from 6,100 to 11,200 (excluding models trained on the ‘Correlation 0.8’ feature subset) (Figure 19b). This range had larger lower and upper bounds than the MSE range for the Random Forest models. In general, XGBoost models also had higher standard deviation in their MSE scores than the analogously trained Random Forest models (497 to 3,734 versus 270 to 2,188). As observed for these Random Forest models, the XGBoost models trained on the ‘Correlation 0.6’ feature subset generally had lower MSE scores (all less than 7,000).

Three XGBoost models tied for the lowest MRE (0.42). They were trained on datasets with no feature selection (missing data thresholds 85% and 95%) and the ‘Correlation 0.6’ feature subset (missing data threshold 85%). The XGBoost model with the lowest MSE (6,163) was trained on data with the ‘Correlation 0.6’ feature subset and a missing data threshold of 95%.

a)



b)



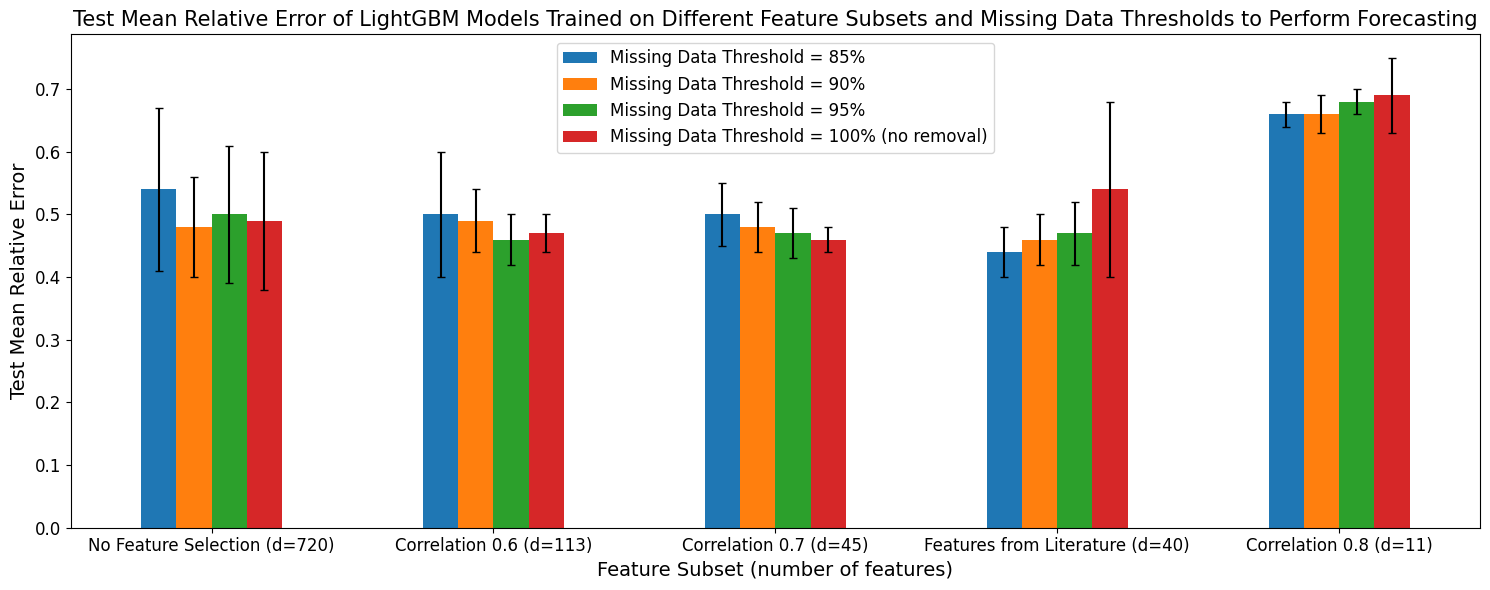
**Figure 19:** a) Mean relative error and b) mean-squared error for XGBoost base estimators fit on different feature subsets and missing data thresholds for forecasting.

##### 5.323: LightGBM

As observed for the Random Forest and XGBoost models, the LightGBM models had relatively uniform MRE scores across different versions of the input dataset (Figure 20a). Excluding models trained on the ‘Correlation 0.8’ feature subset, the LightGBM models had MRE scores between 0.44 and 0.54. This range was higher than for the Random Forest models and had greater upper and lower bounds than the XGBoost models’ MRE. While the LightGBM MRE range was wider than the XGBoost and Random Forest MRE ranges, it was still relatively small. In combination with the large standard deviations in the error metrics, this meant that no single feature subset or missing data threshold consistently had the highest performance.

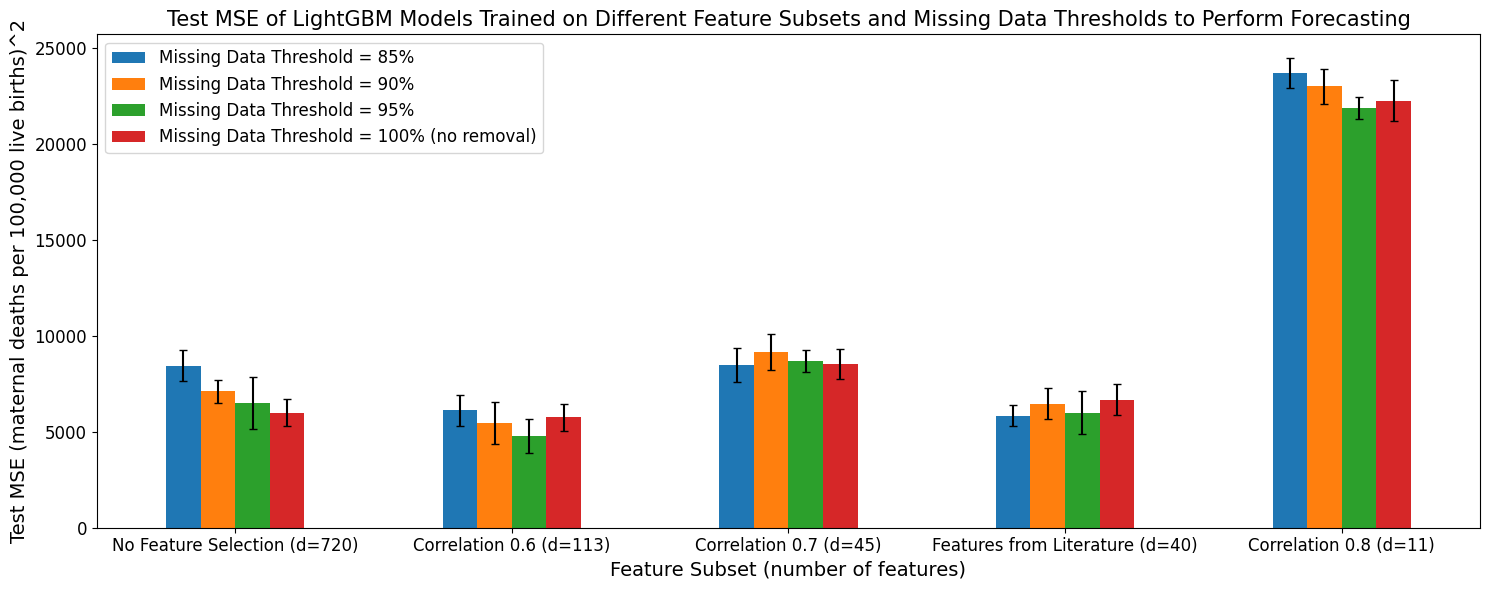
The LightGBM models’ MSE ranged from 4,773 to 9,156, excluding models trained on the ‘Correlation 0.8’ feature subset (Figure 20b). This was similar to the Random Forest models and lower than the XGBoost models. The standard deviation in the LightGBM models’ MSE ranged from 546 to 1,336. This range was fully contained within the analogous ranges for the XGBoost and Random Forest models. The ‘Correlation 0.6’ feature subset produced low MSE scores more consistently than any other feature subset, as observed for the XGBoost and Random Forest models. No single missing data threshold consistently produced the lowest MSE scores for the LightGBM models.

The LightGBM models with the lowest MRE were trained on the ‘Correlation 0.6’ feature subset (missing data threshold 95%) and ‘Correlation 0.7’ feature subset (no missing data threshold). The lowest MSE scores were also observed in LightGBM models trained on the ‘Correlation 0.6’ feature subset and 95% missing data threshold.



b)

a)



**Figure 20:** a) Mean relative error and b) mean-squared error for LightGBM base estimators fit on different feature subsets and missing data thresholds for forecasting.

#### 5.33 Comparisons Between Random Forest, XGBoost, and LightGBM Performance on Different Feature Subsets and Missing Data Removal Thresholds

The previous two sections showed that no single feature selection method or missing data threshold produced the highest performance. The analysis also described how all models fit on the ‘Correlation 0.8’ feature subset had higher error. The hand-picked feature subset more consistently produced low error for country-level prediction, and the ‘Correlation 0.6’ feature subset more consistently produced low error for forecasting.

In this section, I compared the Random Forest, XGBoost, and LightGBM models directly. While the following plots contained a lot of detail, the most salient information was the difference between the various models (plotted in different colours). See Appendix 9.2 for comparisons between the models’ MAE, RMSE, and R2 scores.

##### 5.331: Country-Level Prediction

The LightGBM and XGBoost models had the highest, or tied for the highest, MRE in almost every scenario. The Random Forest models thus often had the lowest MRE across the 5 cross-validation folds. However, the standard deviation in the XGBoost models’ MRE indicated they achieved lower MRE scores on specific folds when trained with no feature selection or on the ‘Correlation 0.6’ and hand-picked feature subsets.

The XGBoost models had the lowest MSE when trained with no feature selection or on the ‘Correlation 0.6’ and hand-picked feature subsets. While the standard deviations for the different model types overlapped, the XGBoost models’ MSE standard deviation indicated higher performance on specific cross-validation folds. Generally, when XGBoost did not have the highest performance, the LightGBM and Random Forest models performed similarly. While the LightGBM models had the highest MSE on when trained with no feature selection or on the hand-picked feature subset, they rotated with the Random Forest models for the worst MSE performance on the other feature subsets.

Despite XGBoost’s high fold-specific performance, none of the model types had consistently superior performance across all data pre-processing technique combinations, especially when considering the overlapping standard deviation in the models’ error metrics (Figure 21).

A graph of different colored lines

AI-generated content may be incorrect.

a)

A screenshot of a graph

AI-generated content may be incorrect.

b)

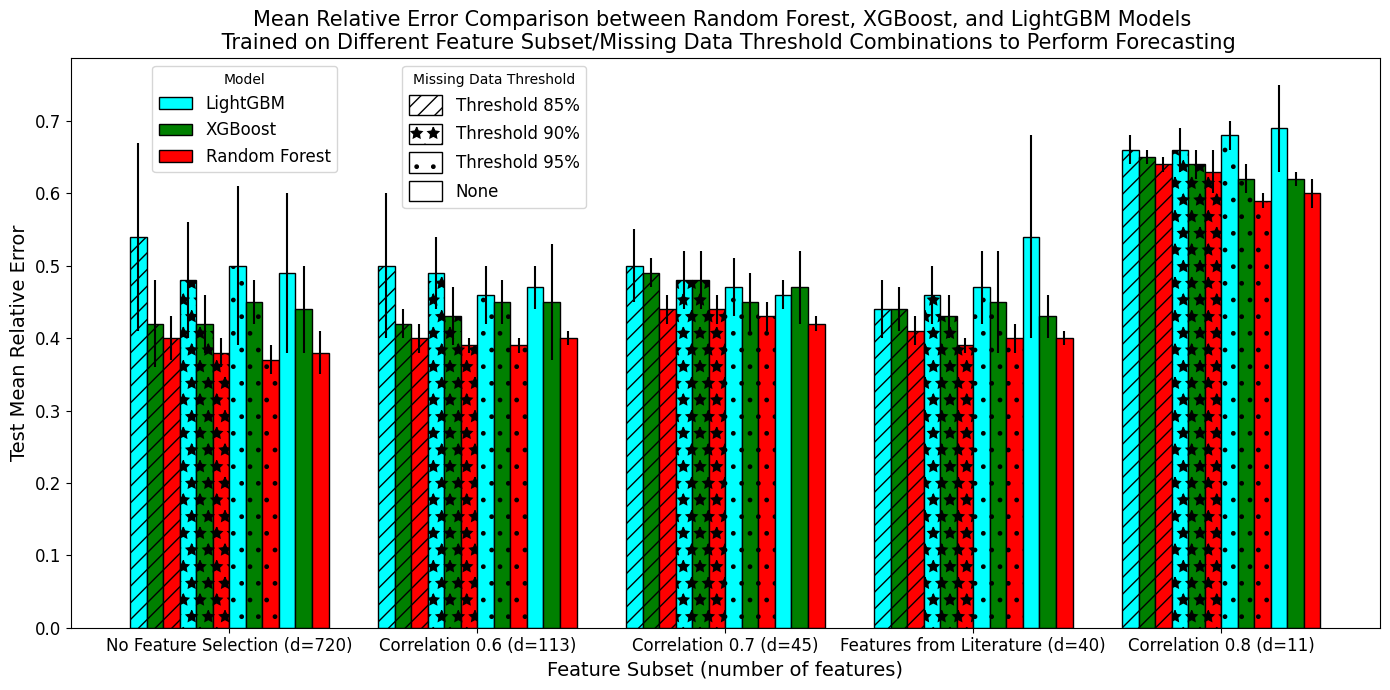
**Figure 21:** a) Mean relative error and b) mean-squared error for Random Forest (red), XGBoost (green), and LightGBM (blue) models fit on different feature subsets and missing data thresholds to perform country-level prediction.

##### 5.332: Forecasting

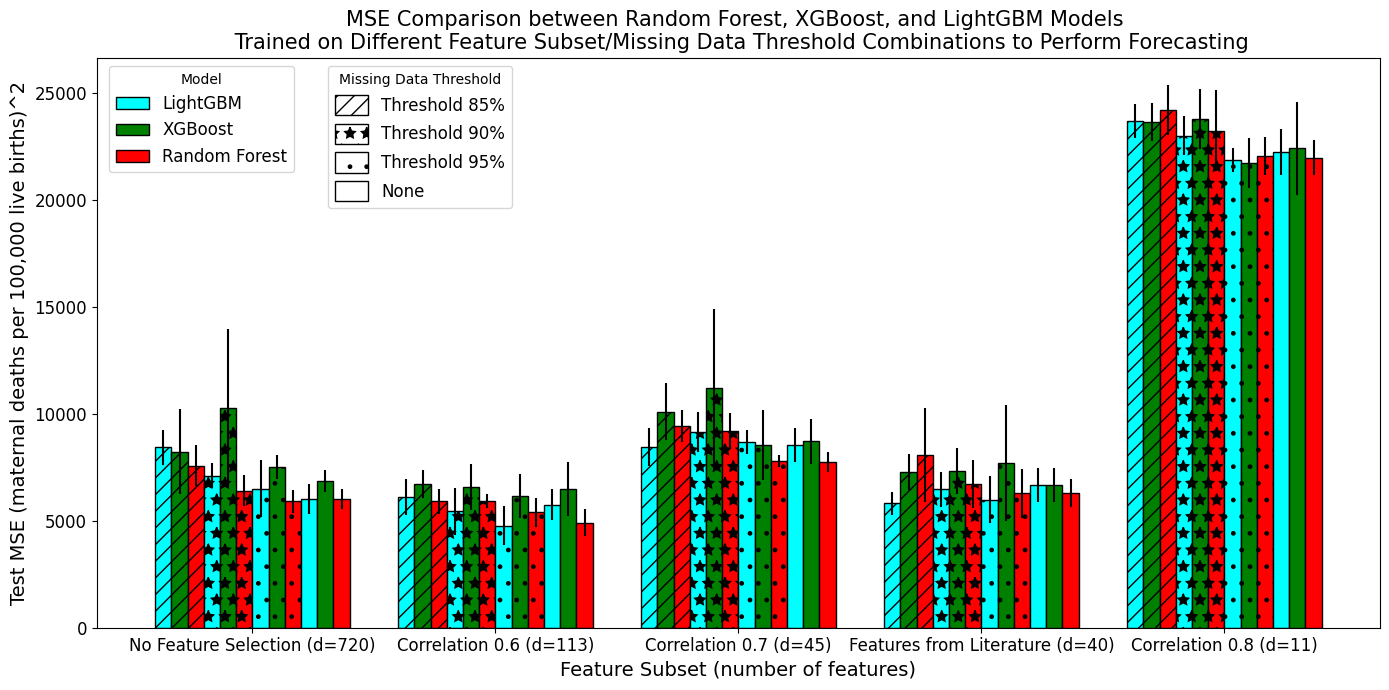
LightGBM models often had the highest MRE, and Random Forest models often had the lowest. Unlike the scenario described above, the XGBoost models’ MRE standard deviation did not indicate their consistently having high performance on specific folds.

XGBoost models trained for forecasting had either the highest or second-highest MSE, with the former occurring more consistently. The LightGBM and Random Forest models performed similarly, with strong overlap in their standard deviations.

Thus, no single model type consistently had the lowest error across both MRE and MSE (Figure 22). While Random Forest models more consistently had lower MRE scores, they had similar MSE performance to the LightGBM models, indicating their potential susceptibility to outliers.



a)



b)

**Figure 22:** a) Mean relative error and b) mean-squared error for Random Forest (red), XGBoost (green), and LightGBM (blue) models fit on different feature subsets and missing data thresholds to perform forecasting.

### 5.4 Performance of Stacking and Voting Ensembles that Combined Predictions from the Base Estimators Trained on Various Input Datasets

The observation that no single model type consistently had the highest performance motivated experimentation into use of a stacking or voting ensemble to combine predictions from Random Forest, XGBoost, and LightGBM models trained on the various input datasets. These models were referred to as “base estimators” from this point forward. As a note, the Random Forest Stacking Ensemble was fit on the predictions of 300 base estimators. In contrast, the Random Forest models detailed above were base estimators fit on feature data.

#### 5.41 Stacking and Voting Ensemble Performance When Trained on All Base Estimators

Voting and stacking ensemble performance was measured according to Section 4.422, with their RMSE, MAE, and R2 scores reported in Appendices 9.311 and 9.312.

##### 5.411: Country-Level Prediction

The stacking and voting ensemble models trained for country-level prediction achieved MRE scores between 0.07 and 0.33 (Figure 23a). The Random Forest Stacking Ensemble achieved the lowest MRE score while the SVM Stacking Ensemble had the highest. The stacking and voting ensembles had MSE scores between 2,161 and 7,100 (Figure 23b), where the Elastic Net Stacking Ensemble had the lowest MSE while the Voting Ensemble had the highest.

The Random Forest Stacking Ensemble’s MSE was approximately 1.3 times greater than the Elastic Net Stacking Ensemble’s MSE (1,689 versus 2,161). In contrast, the Random Forest Stacking Ensemble’s MRE was roughly 2.8 times smaller than the Elastic Net Stacking Ensemble’s MRE (0.07 versus 0.19). Thus, the benefit of using the Random Forest Stacking Ensemble to reduce MRE was greater than the benefit of using the Elastic Net Stacking Ensemble to reduce MSE. Additionally, MRE provides a better holistic understanding model of model performance, while MSE tends to exaggerate outliers, indicating the Random Forest Stacking Ensemble had better performance on the dataset as a whole. Thus, **the Random Forest Stacking Ensemble was chosen as the best-performing ensemble**.

A graph with blue squares

AI-generated content may be incorrect.

a)

A graph with blue squares

AI-generated content may be incorrect.

b)

**Figure 23:** a) Mean relative error and b) mean-squared error for voting and stacking ensembles trained on all base models to perform country-level prediction.

##### 5.412: Forecasting

The stacking and voting ensembles trained to perform forecasting achieved MRE scores ranging from 0.37 to 0.56 and MSE scores between 5,100 and 8,000 (Figure 24). **The Random Forest Stacking Ensemble was the best-performing model** and the SVM Stacking Ensemble was the worst-performing model in terms of both MRE and MSE.

**A graph of blue rectangular objects

AI-generated content may be incorrect.**

b)

a)

**A graph of blue rectangular objects

AI-generated content may be incorrect.**

**Figure 24:** a) Mean relative error and b) mean-squared error for voting and stacking ensembles trained on all base models to perform forecasting.

#### 5.42 Weighting Given to Each Base Estimator in the Stacking and Voting Ensembles

To better understand the performance differences between the various ensembles, I explored which base estimators were weighted most heavily by each ensemble. I did not further investigate the SVM Stacking Ensemble because the Scikit Learn implementation lacked a ‘feature importance’ method. Each of the 300 base estimators were referenced using a number between 0 and 299. LightGBM base estimators were numbered 0-99, Random Forest base estimators numbered 100-199, and XGBoost base estimators numbered 200-299.

The Random Forest Stacking Ensemble (RFSE) only placed importance on a subset of base estimators regardless of whether it was trained for country-level prediction or forecasting (Figure 25). It primarily drew strength from the XGBoost base estimators, with a smaller subset of LightGBM models used. It placed very little emphasis on Random Forest base estimators. A greater number of base estimators were used by the RFSE to perform forecasting than country-level prediction.

Unlike the RFSE, the Elastic Net Stacking Ensemble derived support from most base estimators, with importance placed on all model types (Figure 26). This difference was shown clearly by how the Elastic Net Stacking Ensemble placed high importance on some Random Forest base estimators. However, like the RFSE, the Elastic Net Stacking Ensemble placed only a small amount of importance on a subset of base estimators. In contrast, the Voting Ensemble placed a very small, but relatively equal, amount of importance on all base estimators, with only a few base estimators contributing little to the final prediction (Figure 27).

b)

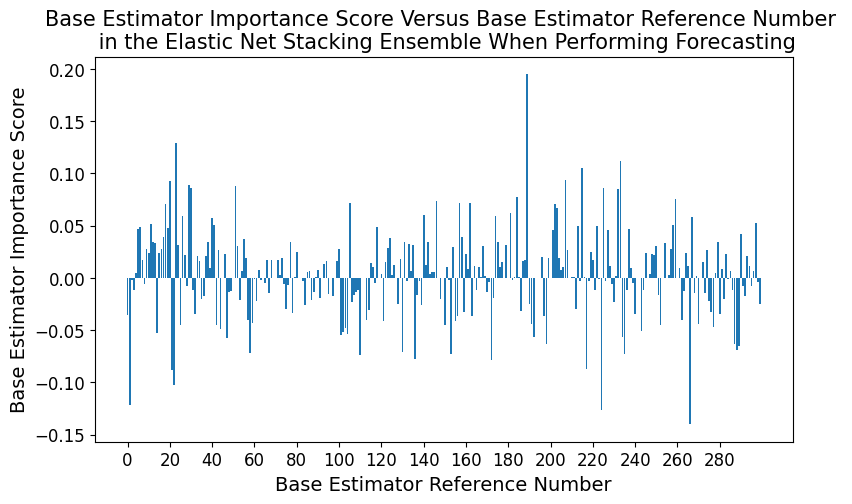
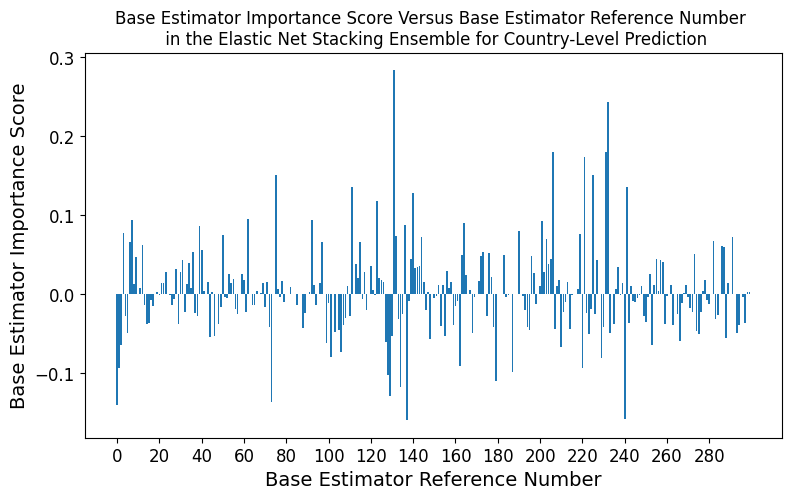
a)

A graph with numbers and a number

AI-generated content may be incorrect.A graph with blue lines

AI-generated content may be incorrect.

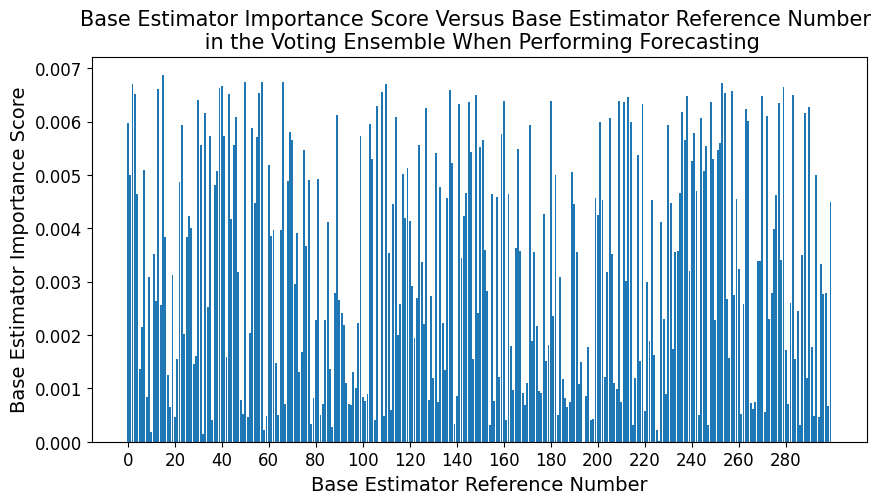
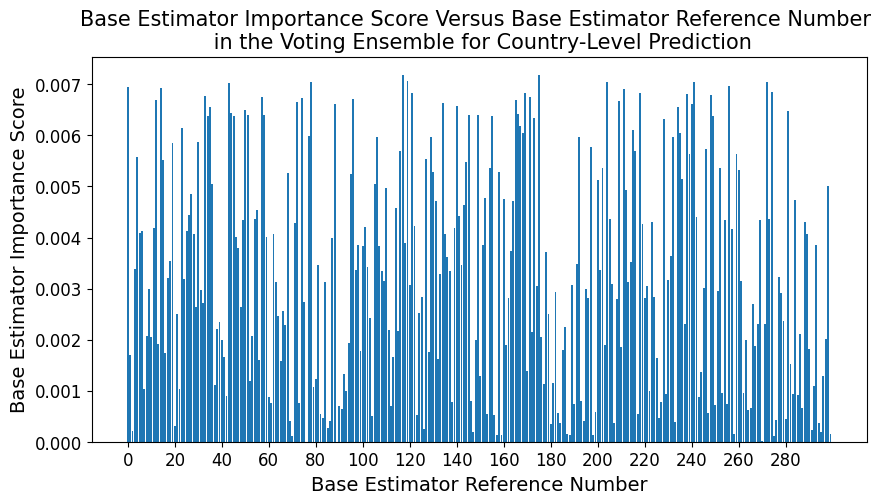
**Figure 25:** Importance score for each of the 300 base estimators used in the Random Forest Stacking Ensemble trained for a) country-level prediction and b) forecasting.



a)

b)

**Figure 26:** Importance score for each of the 300 base estimators used in the Elastic Net Stacking Ensemble trained for a) country-level prediction and b) forecasting.



a)

b)

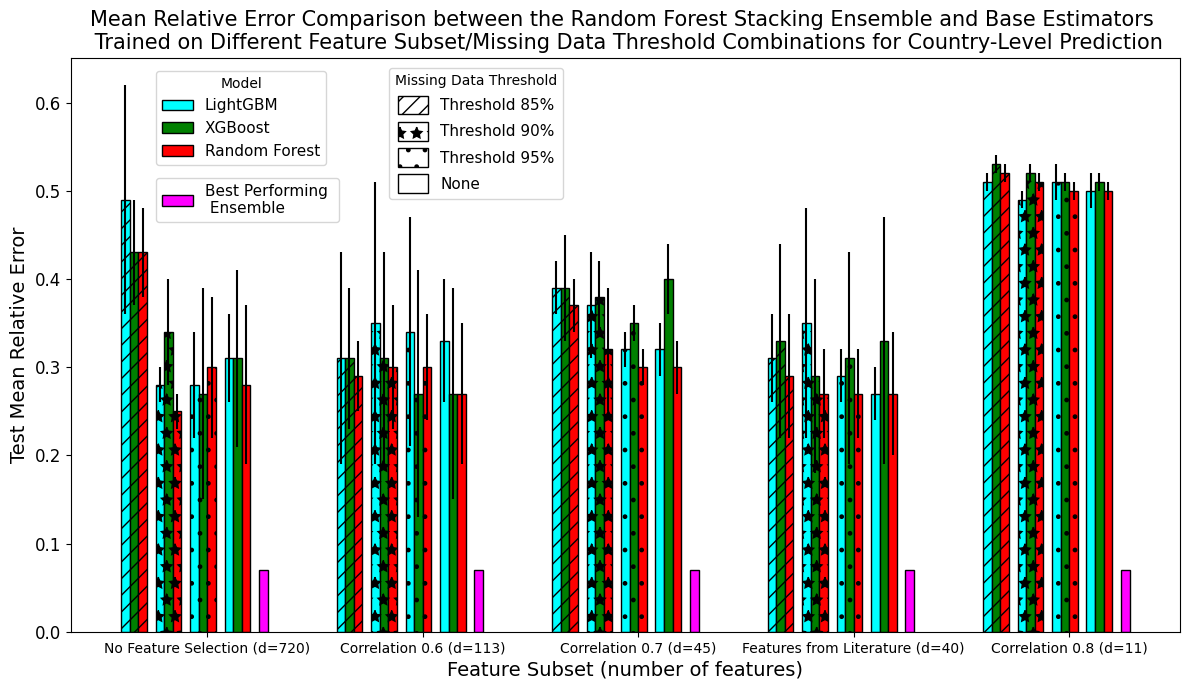
**Figure 27:** Importance score for each of the 300 base estimators used in the Voting Ensemble trained for a) country-level prediction and b) forecasting.

#### 5.43 Performance Comparison of the Best Performing Stacking/Voting Ensemble and the Single Base Estimators

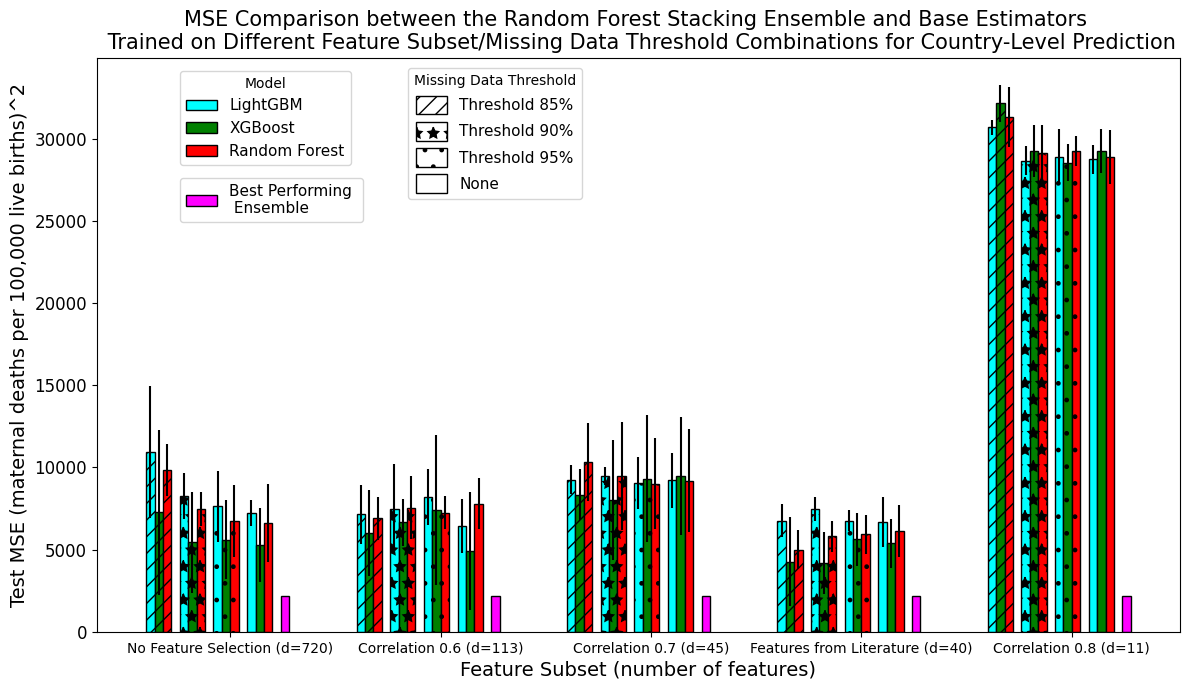
As described in Section 5.41, the best performing stacking and voting ensemble was the Random Forest Stacking Ensemble (RFSE). Its predictive error was compared with that of its base estimators to establish whether stacking reduced error. While the following plots contained a lot of detail, the most important information conveyed was the difference between RFSE and its base estimators (light purple versus red, green, and light blue). See Appendix 9.321 and 9.322 for comparisons using MAE, RMSE, and R2.

##### 5.431: Country-Level Prediction

The Random Forest Stacking Ensemble greatly reduced both MSE and MRE for models trained for country-level prediction (Figure 28). More explicitly, the RFSE achieved an MRE of 0.07 compared to the best MRE achieved by a base estimator of 0.25. Similarly, the RFSE had an MSE of 2,161 while the lowest MSE produced by a base estimator was 4,185. Thus, **the Random Forest Stacking Ensemble was superior to the base estimators for country-level prediction.**



a)



b)

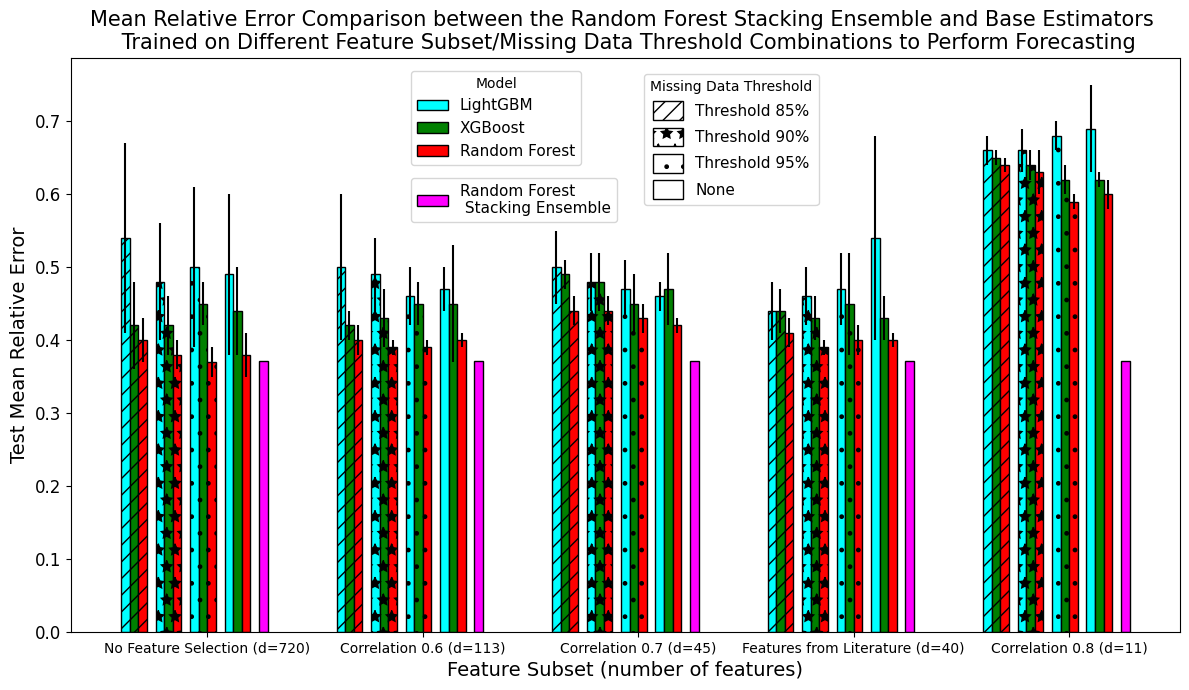
**Figure 28:** a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble (purple) and the Random Forest (red), XGBoost (green), and LightGBM (blue) base estimators trained for country-level prediction.

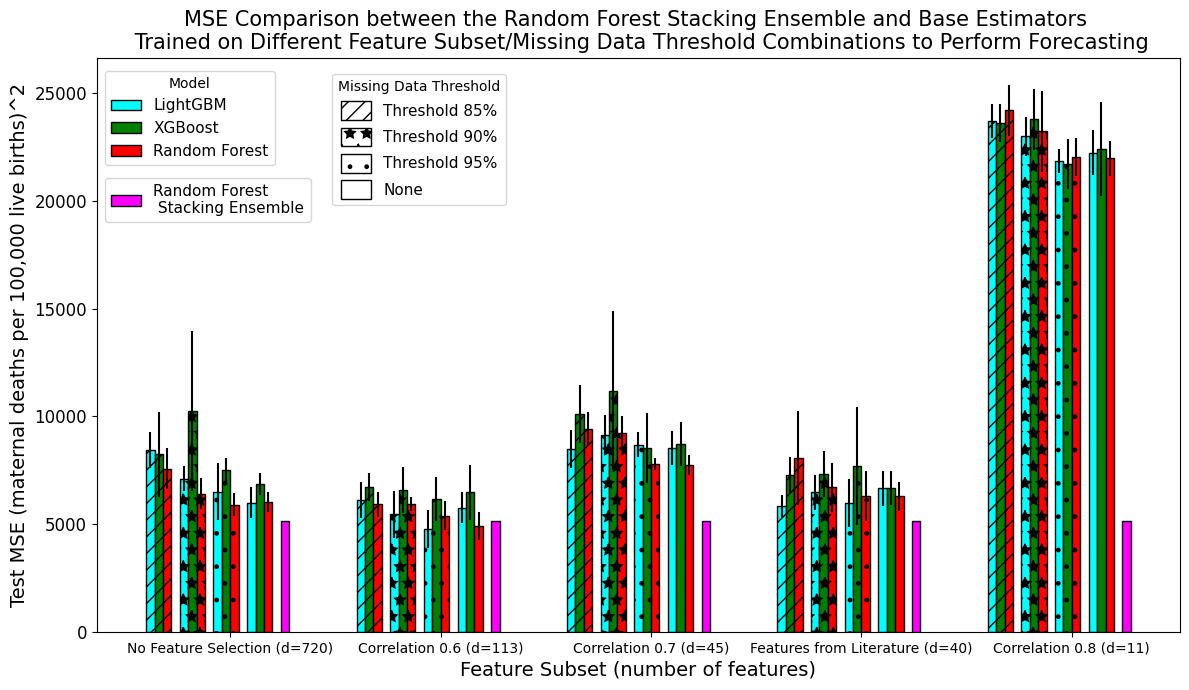
##### 5.432: Forecasting

The Random Forest Stacking Ensemble trained to perform forecasting did not always produce greatly lower error than the base estimators (Figure 29). The RFSE’s best MRE of 0.37 was the same as the MRE produced by the Random Forest base estimator trained with no feature selection and a missing data threshold of 95%. The RFSE’s lowest MSE of 5,134 was larger than the MSE produced by the LightGBM base estimator (4,773) trained on the ‘Correlation 0.6’ feature subset with a missing data threshold of 95%.

Despite these less promising re4sults, **the Random Forest Stacking Ensemble was still considered the ‘best performing model’** because of the substantial improvement it produced for country-level prediction. The RFSE was used for forecasting as well **for consistency and because it did not increase MRE or raise MSE by a notable amount**, as its MSE was only 1.08 times greater than the best base estimator’s MSE.

a)





b)

**Figure 29:** a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble (purple) and the Random Forest (red), XGBoost (green), and LightGBM (blue) base estimators trained to perform forecasting.

### 5.5 Investigation into the Random Forest Stacking Ensemble’s Architecture

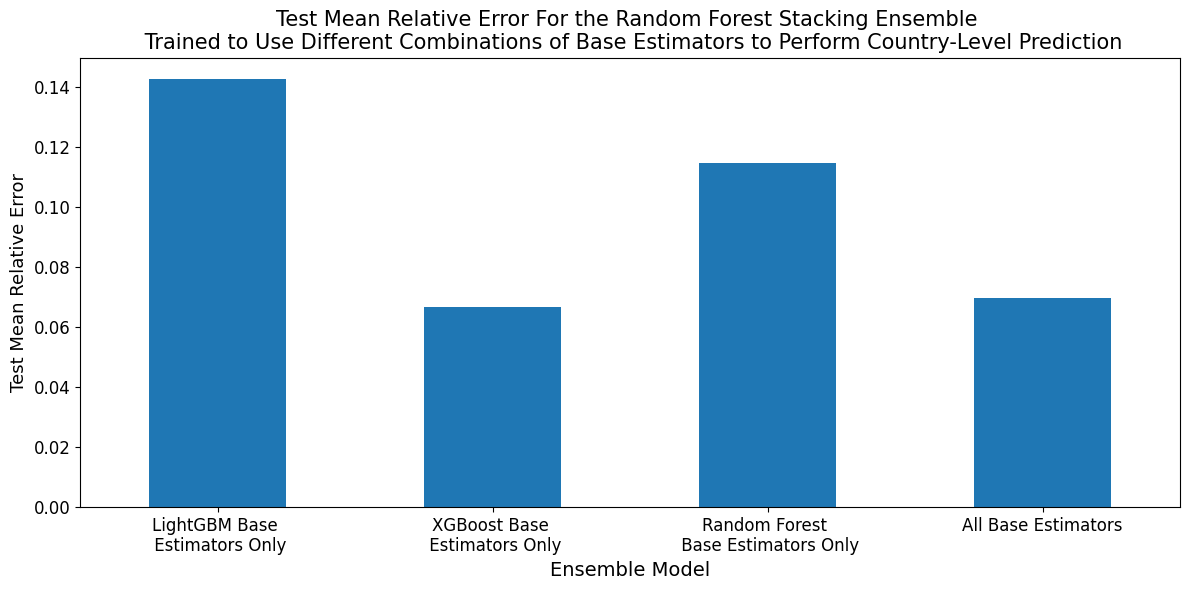
The Random Forest Stacking Ensemble’s architecture was investigated further to better understand its performance and determine if it could be improved.

#### 5.51 Random Forest Stacking Ensemble with Different Combinations of Base Estimators

Training the Random Forest Stacking Ensemble (RFSE) on a subset of available base estimators rather than all base estimators, as above, did not greatly improve performance of either country-level prediction or forecasting (Figures 30 and 31).

When training the RFSE for country-level prediction, only using predictions from XGBoost models reduced MRE by roughly 0.3% (from 0.0695 to 0.0667) and decreased MSE by 87 (from 2,161 to 2,074). When training the RFSE for forecasting, only using predictions from Random Forest base estimators reduced MRE by about 0.19% (from 0.3708 to 0.3689) and only using predictions from LightGBM base estimators decreased MSE by 405 (from 5,134 to 4,729). All other combinations of base estimator inputs reduced performance. For example, using a different base estimator subset as input to the RFSE trained for country-level prediction increased its MRE by at least 5% and its MSE by at least 1,200.

The improvements due to using a different subset of base estimators were extremely small, especially when the improvement was put in terms of MRE. Additionally, the best subset of base estimators to use changed for each metric and type of analysis. Choosing to remain with the original RFSE model trained on all base estimators prevented the need to conduct all future experiments on three different stacking ensemble formulations. Given the lack of compute resources at the tail-end of this project, the decision was made to use all base estimators. Additionally, using all available base estimators in an ensemble model more closely follows convention. **Thus, the Random Forest Stacking Ensemble trained on predictions from all base estimators was considered the ‘best-performing’ model from this point forward.**

A graph of blue rectangular bars

AI-generated content may be incorrect.

b)

a)

**Figure 30:** a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble trained to perform country-level prediction using different combinations of base estimators.

A graph of blue bars

AI-generated content may be incorrect.

a)

A graph of blue bars

AI-generated content may be incorrect.

b)

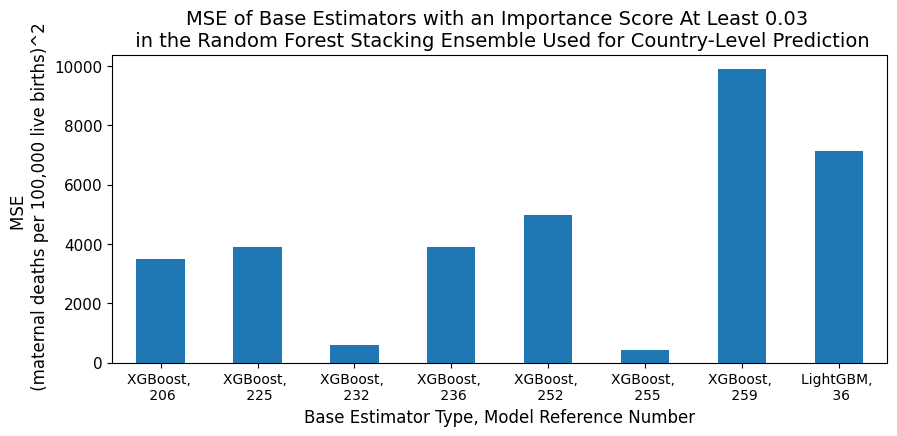
**Figure 31:** a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble trained to perform forecasting using different combinations of base estimators.

#### 5.52 Importance Analysis of the Base Estimators in the Best-Performing Random Forest Stacking Ensemble

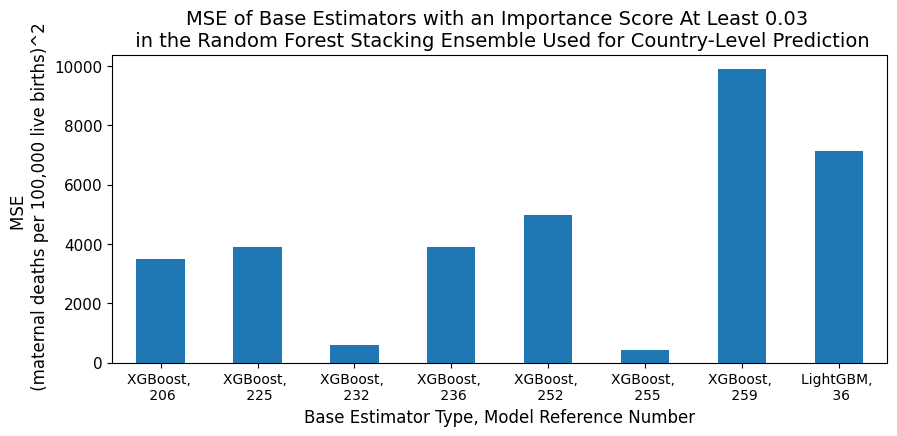
In Section 5.42, I discussed how the RFSE only placed high importance on a small subset of mostly XGBoost and LightGBM models. Given the decision to continue using all base estimators as input to the ensemble, I tested different hypotheses for why the specific subset of base estimators was chosen by the RFSE using the procedure outlined in Section 4.425.

##### 5.521 Differences in Predictive Performance of Most Important Base Estimators

I first determined whether mean predictive error over the test set was the sole predictor of the RFSE’s choice of base estimator. I compared MSE because this was the metric used to train and fine-tune the stacking ensemble. Of the 8 base estimators trained for country-level prediction that were given an importance score of at least 0.03, 7 were XGBoost models and 1 was LightGBM. While many of the MSE scores of the chosen base estimators were low, two had MSE scores of greater than 6,000 and one had an MSE of almost 10,000, which was at the higher end of the range of observed MSE scores (Figure 32a). Of the 10 base estimators trained to perform forecasting that were given an importance score of at least 0.03, 7 were XGBoost models and 3 were LightGBM. Again, while most of these estimators produced MSE at the bottom of the observed range, there was one XGBoost base estimator with an MSE score towards the top end of the range (close to 10,000) (Figure 32b). Thus, **MSE did not completely explain** **how the RFSE gave importance to its base estimators.**



a)



b)

**Figure 32:** MSE for the base estimators given an importance score of at least 0.03 by the Random Forest Stacking Ensemble trained to perform a) country-level prediction and b) forecasting. Each base estimator was identified with its model type and the reference number used in Section 5.43, which specified its ordering in the RFSE’s input data.

##### 5.522 Effect of Permutating the Order of Base Estimators in the RFSE’s Input Data

I next tested whether the Random Forest Stacking Ensemble was biased in its choice of estimator. For example, by default, the first ‘features’ it used to create splits in its decision trees may have had specific positions in its input dataset. Practically, this would mean it first tried to create splits using predictions from base estimators located at default positions in its input data. If none of the base estimators it subsequently trialled produced a split with a lower predictive error, it would remain with the default, biased base estimator selection. This was somewhat likely given the relatively similar performance between base estimators. To test this hypothesis, I randomly permuted the positions of base estimators in the RFSE’s input dataset.

###### 5.5221: Country-Level Prediction

Nine of the ten base estimators given importance scores of at least 0.03 in the original RFSE were also given importance scores of at least 0.03 when base estimator order was permuted (Table 12). After permutation, their importance score magnitudes generally did not change by a large amount. The largest change was in the model given the highest importance, which lost 0.06 importance points after permutation. The RFSE’s predictive accuracy did not change greatly after permutation, indicating that the subset of base estimators used was not a random decision (MRE=0.07, MSE=2,188 versus the original MRE=0.07, MSE=2,161). However, the weighting given to each base estimator may be unstable and subject to change via retraining.

###### 5.5222: Forecasting

Randomising base estimator order had a greater effect of the RFSE trained for forecasting (Table 12).While 10 base estimators in the original RFSE had importance scores at least 0.03, only 4 base estimators had a sufficiently high importance score in the permuted RFSE. Only 2 of these 4 base estimators were also in the list of 10 given high importance under the original ordering, with both of these base estimators’ importance scores increasing by 0.26 to 0.30 points. The RFSE’s predictive accuracy decreased after permutation, with its MSE increasing by roughly 860 points (MRE=0.39, MSE=6,063 versus the original MSE=0.37, MSE=5,134). These changes showed that the RFSE’s choice of base estimators was more affected by ordering and/or was instable and subject to change through retraining.

**Table 12:** The base estimators given an importance score of at least 0.03 by the Random Forest Stacking Ensemble when present in the RFSE’s input data in the original and permuted orders. The model reference numbers were given in terms of the original ordering to allow comparison. For example, if the base estimator originally in the 206th position in the input data was moved to the 2nd position in the permuted order, it was still presented below as the 206th model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Country-Level Prediction** | | | |
| **Random Forest Stacking Ensemble with Original Base Estimator Order** | | **Random Forest Stacking Ensemble with Permuted Order of Base Estimators** | |
| **Model Reference** | **Importance Score** | **Model Reference** | **Importance Score** |
| 206 | 0.26 | 206 | 0.20 |
| 232 | 0.14 | 232 | 0.11 |
| 252 | 0.10 | 236 | 0.11 |
| 255 | 0.06 | 252 | 0.08 |
| 236 | 0.06 | 255 | 0.06 |
| 36 | 0.05 | 225 | 0.04 |
| 259 | 0.03 | 36 | 0.04 |
| 225 | 0.03 | 258 | 0.04 |
| **Forecasting** | | | |
| **Random Forest Stacking Ensemble with Original Base Estimator Order** | | **Random Forest Stacking Ensemble with Permuted Order of Base Estimators** | |
| **Model Reference** | **Importance Score** | **Model Reference** | **Importance Score** |
| 241 | 0.13 | 241 | 0.43 |
| 254 | 0.08 | 243 | 0.33 |
| 40 | 0.08 | 258 | 0.09 |
| 232 | 0.08 | 207 | 0.05 |
| 243 | 0.07 | All remaining models had importance scores < 0.03 | |
| 43 | 0.05 |
| 20 | 0.05 |
| 215 | 0.03 |
| 213 | 0.03 |
| 231 | 0.03 |

##### 5.524 Conclusion of Base Estimator Importance Experiments

I did not find a robust explanation for why the Random Forest Stacking Ensemble placed high importance on a specific subset of base estimators. While some of the subset had low predictive error, others had very high MSE scores. Potentially, even the base estimators with high predictive error performed well for a specific subset of estimates, making them useful for the stacking meta-estimator, which can learn when and how to use them. While the permutation analysis showed the base estimators trained for country-level prediction were not chosen at random, it did reveal instability in their importance scores over different training instances. This was shown more explicitly for the RFSE trained for forecasting. This instability may be due to many of the base estimators having similar performance, allowing them to substitute for each other and/or take some of each other’s importance weighting.

#### 5.53 Feature Importance Analysis for Chosen Base Estimators

One of the primary aims of this research was to determine the socio-economic and health-related variables with the highest predictive power for MMR. The following section presents the features given the highest importance by the two base estimators with the highest weightings in the RFSE. For comparison, I also presented the features given the highest importance by 2 base estimators with low importance in the RFSE and relatively higher MSEs.

##### 5.531 Country-Level Prediction

The two base estimators given the highest importance scores in the Random Forest Stacking Ensemble placed the greatest importance on features detailing women’s level and type of employment, women’s knowledge of contraceptive options, the percentage of women who were teenage mothers, and the country’s World Bank defined income level (Table 13). Health-related variables monitoring the presence of skilled health staff at births, fertility rates, nutritional status, and life expectancy were also highly valued.

While the two base estimators with higher errors and lower importance scores in the RFSE also placed value on variables that monitor contraception prevalence and literacy rates, they focused more on features that monitored the prevalence of different diseases. For example, they placed high importance on features monitoring the rate of still-births and specific nutritional deficiencies as well as infectious disease and maternal disorders prevalence.

Overall, these results indicate that the base estimators given more importance in the RFSE placed more value on socio-economic related variables and aggregate health-trends while base estimators given less importance monitored more specific trends in health and disease.

**Table 13:** The 5 features given the highest importance scores in: (blue) the two base estimators given the highest importance scores in the Random Forest Stacking Ensemble, (orange) a medium-low accuracy base estimator from the ‘Correlation 0.7’ feature subset, and (purple) a low-accuracy base estimator from the ‘Correlation 0.8’ feature subset. All models were used for country-level prediction.

|  |  |
| --- | --- |
| **Base Estimator Model & Importance Score in the RFSE** | **Feature Name** |
| XGBoost, fold 2, missing data threshold 95%, no feature selection  Importance score: 0.26 | Vulnerable employment (% of total employment), female |
| Knowledge of any modern method of contraception (% of all women ages 15-49) |
| Wage and salaried workers (% of total population) female |
| Knowledge of any modern method of contraception (% of all married women ages 15-49) |
| Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant) |
| XGBoost, fold 4, missing data threshold 85%, features hand-picked from literature  Importance score: 0.14 | Country income level |
| Births attended by skilled health staff (% of total) |
| Fertility rate, total (births per woman) |
| Survival to age 65, female (% of cohort) |
| Prevalence of overweight (% of adults) |
| XGBoost, fold 4, missing data threshold 90%, ‘Correlation 0.7’ feature subset  Importance score: 1.44\*10-4 | Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of male population) |
| Contraceptive prevalence, any method (% of married women ages 15-49) |
| Stillbirth rate (per 1,000 total births) |
| Tetanus prevalence (age standardised) (per 100,000 population) male |
| Other infectious diseases prevalence (age standardised) (per 100,000 population), male |
| XGBoost, fold 4, missing data threshold 100%, ‘Correlation 0.8’ feature subset  Importance score: 1.56\*10-6 | Literacy rate, youth total (% of people ages 14-24), female |
| Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) |
| Maternal disorders prevalence (age standardised) (per 100,000 population) |
| Vitamin A deficiency prevalence (age standardised) (per 100,000 population), male |
| Probability of survival to age 5, male |

##### 5.532 Forecasting

As above, the two base estimators with the highest importance scores in the Random Forest Stacking Ensemble placed the greatest importance on features that measured the amount and type of female employment as well as knowledge about contraceptive options and nutritional status (Table 14). There was also slightly more emphasis on long-term conditions, such as measuring mortality due to non-communicable diseases. While the base estimators that added little value to the RFSE placed highest importance on similar features, these estimators placed slightly more emphasis on mortality measures and contained more information about trends in health outcomes for the whole population and for men rather than focusing on women.

**Table 14:** The 5 features given the highest importance scores in: (blue) the two base estimators given the highest importance scores in the Random Forest Stacking Ensemble, (orange) a medium-low accuracy base estimator from the ‘Correlation 0.7’ feature subset, and (purple) a low-accuracy base estimator from the ‘Correlation 0.8’ feature subset. All models were used to perform forecasting

|  |  |
| --- | --- |
| **Base Estimator Model, Importance Score in RFSE** | **Feature Name** |
| XGBoost, fold 1, missing data threshold 90%, ‘Correlation 0.6’ feature subset.  Importance score: 0.13 | Wage and salaried workers (% of total employment), female |
| Vulnerable employment (% of total employment), female |
| Prevalence of stunting, height for age, male (% of children under 5) |
| Contraceptive prevalence, any method (% of married women ages 15-49) |
| Self-employed, total (% of total employment), female |
| XGBoost, fold 3, missing data threshold 95%, ‘Correlation 0.6’ feature subset.  Importance score: 0.08 | Vulnerable employment (% of total employment), female |
| Children in employment (% of children ages 7-14), female |
| Cause of death, by non-communicable diseases, female (% of female population) |
| Yellow fever prevalence (age standardised) (per 100,000 population), female |
| Contraceptive prevalence, any modern method (% of married women ages 15-49) |
| Random Forest base estimator, fold 5, missing data threshold 95%, ‘Correlation 0.7’ feature subset  Importance score: 1.55\*10-7 | Births attended by skilled health staff (% of total) |
| Contraceptive prevalence, any method (% of married women ages 15-49) |
| Stillbirth rate (per 1,000 total births) |
| Mortality rate, under-5, male (per 1,000) |
| Demand for family planning satisfied by any methods (% of married women with demand for family planning) |
| XGBoost, fold 4, missing data threshold 100%, ‘Correlation 0.8’ feature subset  Importance score: 5.28\*10-7 | Maternal disorders prevalence (age standardised) (per 100,000 population), female |
| Literacy rate, youth total (% of people ages 15-24), female |
| Vitamin A deficiency prevalence (age standardised (per 100,000 population), male |
| Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of male population) |
| Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) |

### 5.6 Performance Analysis of the Random Forest Stacking Ensemble

Building on the previous results, section explores the RFSE’s performance.

#### 5.61 Random Forest Stacking Ensemble’s Predictive Error per Income Level

To gain a deeper understanding of how the RFSE performs in different settings, I analysed how its prediction errors changed when estimating MMR for countries from different income levels.

##### 5.611: Country-Level Prediction

Generally, the RFSE’s MRE on the test set decreased as income level increased (Figure 33a). For example, its test MRE was 0.18 for low-income countries but 0.07 for high-income countries. However, the RFSE achieved its lowest test error for lower-middle income countries (0.02). Generally, train and validation MREs were similar and smaller than the test MRE for all income levels, with the exception again being the lower-middle income subgroup. The difference between the train/validation and test MREs was greatest for low-income countries (~0.14).

MSE uniformly decreased as income level increased, with the differences between income levels spanning orders of magnitude (Figure 33b). More specifically, the RFSE incurred an MSE of 62,133 for low-income countries versus an MSE of 6 for high-income countries. The highest standard deviation in MSE was observed for low-income countries. The largest difference in MSE between consecutive income levels occurred between low-income and lower-middle income countries (62,133 to 356).

A graph of different colored bars

AI-generated content may be incorrect.

a)

A graph of a bar graph

AI-generated content may be incorrect.

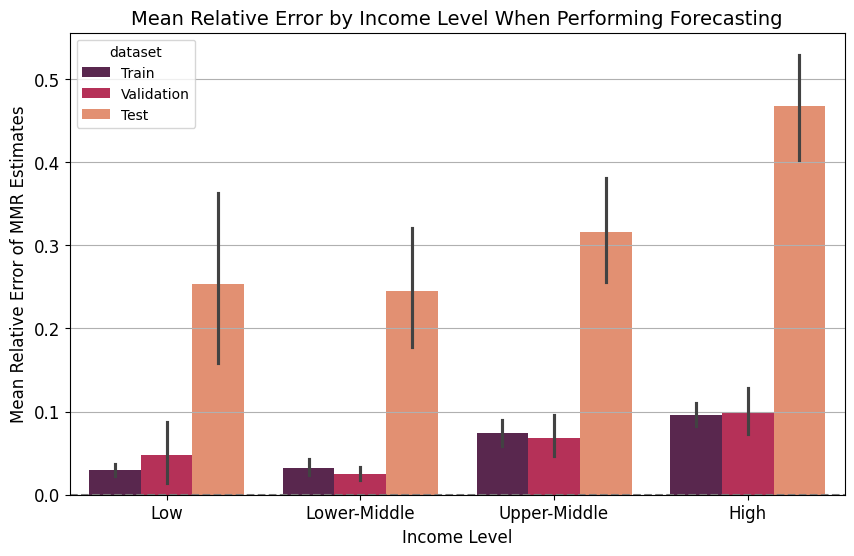
b)

**Figure 33:** a) Mean relative error and b) mean-squared error (log scale) for income-level specific MMR estimates from the Random Forest Stacking Ensemble used for country-level prediction. MRE was given for its performance on the train, validation, and test sets. MSE was only given for the test set.

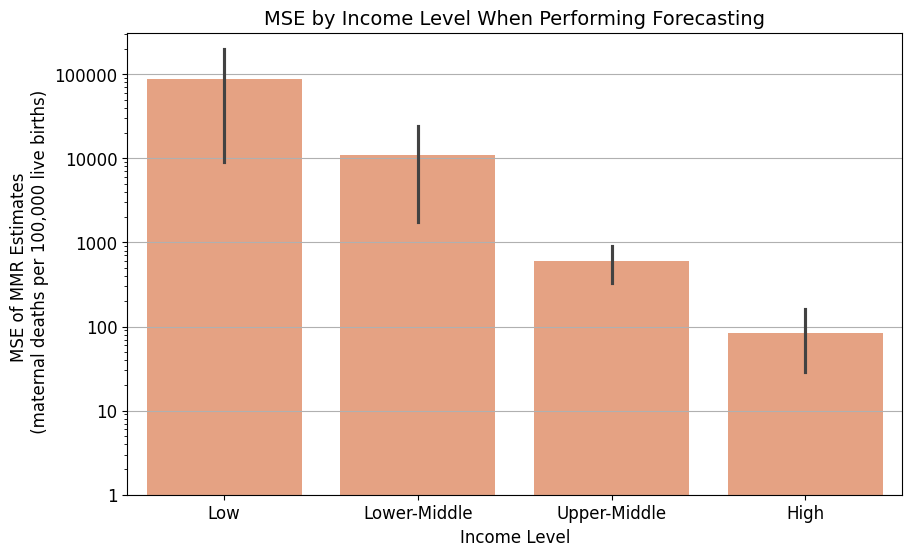
##### 5.612: Forecasting

The MRE of the RFSE trained to forecast MMR increased as income level increased from lower middle to high. (Figure 34a). For instance, the RFSE had an MMR of 0.25 for lower-middle income countries versus an MRE of 0.47 for high-income countries. In contrast to this trend, the RFSE had a test MRE of 0.25 for both low and lower-middle income countries. The train and validation errors also increased as income level increased from lower middle to high. The RFSE’s MRE had a large standard deviation for its validation and test sets, with the large validation deviation indicating considerable differences between cross-validation folds. Generally, train and validation errors for the same income level were similar, with test error always being at least 0.2 greater than train error. The low-income countries had the greatest difference (0.02) between train and validation MRE of any income level.

Test MSE decreased uniformly as income level increased, with decreases between income levels generally spanning an order of magnitude (Figure 34b). For instance, the RFSE achieved a test MRE of 85 for high-income countries versus 88,585 for low-income countries.



a)



b)

**Figure 34:** a) Mean relative error and b) mean-squared error (log-scale) for income-level specific MMR predictions from the Random Forest Stacking Ensemble trained to perform forecasting. MRE was given for the RFSE’s performance on the train, validation, and test sets. MSE was only given for the test set.

#### 5.62 Uncertainty Analysis for the Random Forest Stacking Ensemble

To provide a measure of uncertainty about the MMR estimates from the Random Forest Stacking Ensemble, I computed the standard deviation among the MMR estimates of its base estimators. The smaller the standard deviation, the greater the agreement, and thus the more certainty the stacking ensemble had in its final estimate.

As the ground truth MMR increased, standard deviation among the base estimators trained for country-level prediction also increased (Figure 35a). For MMR estimates between 0 and 150, standard deviation was generally less than 50. In contrast, for ground truth MMRs between 300 and 1,050, base estimator predictions for ranged from 50 to 350. For the extremely high ground truth MMR of 1,763, standard deviation among base estimators was 441.

Similarly, standard deviation among the MMR predictions of base estimators trained to perform forecasting increased (0 to 300) as the ground truth MMR increased (0 to 900) (Figure 35b). However, a slight decrease in standard deviation was observed for ground truth MMRs greater than 1,150. More specifically, the ground truth MMRs 1,194 and 1,389 had standard deviations 216 and 170, respectively.

However, these findings must be qualified by the statement that there were few datapoints for ground truth MMRs greater than 1,050 for country-level prediction or 750 for forecasting.

A graph with blue dots

AI-generated content may be incorrect.

a)

b)

A graph with blue dots

AI-generated content may be incorrect.

**Figure 35:** Standard deviation among the predictions made by base estimators in the Random Forest Stacking Ensemble versus the ground truth MMR estimate they were trying to predict. This analysis was done for base estimators trained to perform a) country-level prediction and b) forecasting.

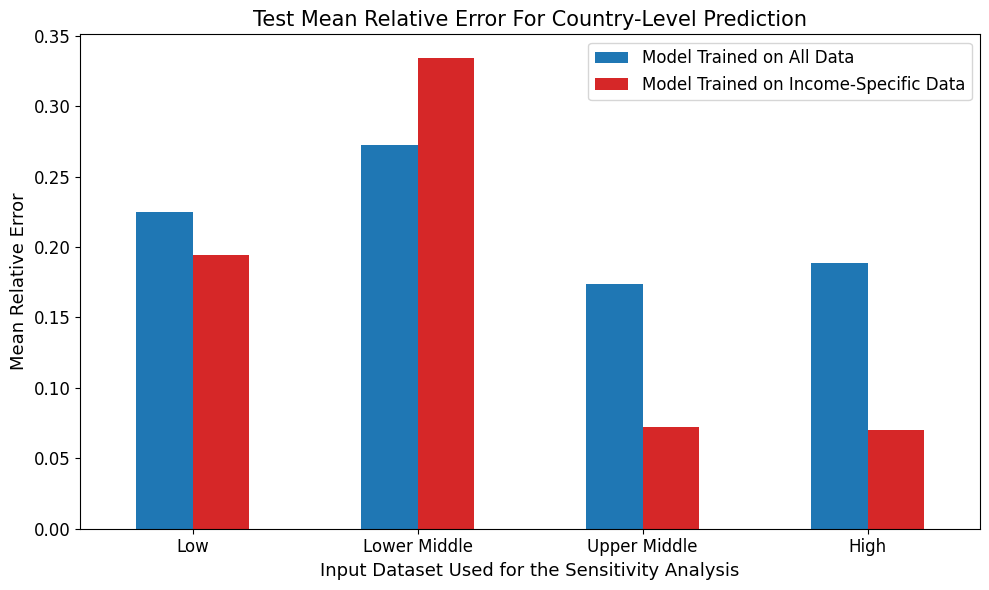
#### 5.63 Sensitivity Analysis

I conducted a sensitivity analysis to explore how the input data structure affected the RFSE’s final MMR predictions (Section 4.5 for the method).

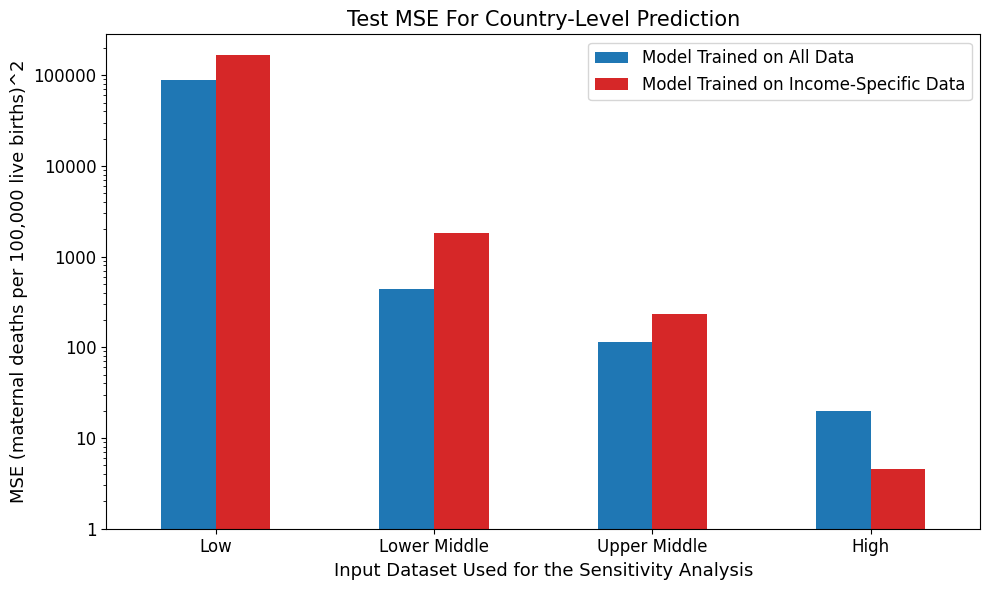
##### 5.631 Country-Level Prediction

Generally, the MRE for the Random Forest Stacking Ensemble trained on data from all income levels was larger than the MRE for RFSEs trained on data from a specific income level (Figure 36a). For example, the RFSE trained on all data had a test MRE of 0.19 on the high-income dataset while the RFSE trained on just high-income data had an MRE of 0.07. Similarly, the RFSE trained on all data had a test MRE of 0.17 on the upper-middle dataset while the RFSE trained on just upper-middle income data had a test MRE of 0.07. The smallest improvement produced by the sensitivity analysis model occurred for the low-income dataset, where the RFSE trained on all data had an MRE of 0.22 while the RFSE trained on only the low-income dataset had an MRE of 0.19. However, the MRE for the RFSE trained on all data (0.27) was smaller than that of the RFSE trained on just lower-middle income data (0.33).

In contrast, only the RFSE trained on just high-income data had a lower MSE score than the RFSE trained on all data (5 versus 20) (Figure 36b). The RFSE trained on all data had a lower MSE scores than the RFSEs trained on income-level specific data for every income subset other (Figure 36b). For example, the RFSE trained on just low-income data had an MSE of 167,123, while the RFSE trained on all data had an MSE of 89,353. This was the largest difference between the original and sensitivity model across all datasets. The difference in the trend observed for MSE and MRE suggests the presence of outliers in the income-specific data.



a)



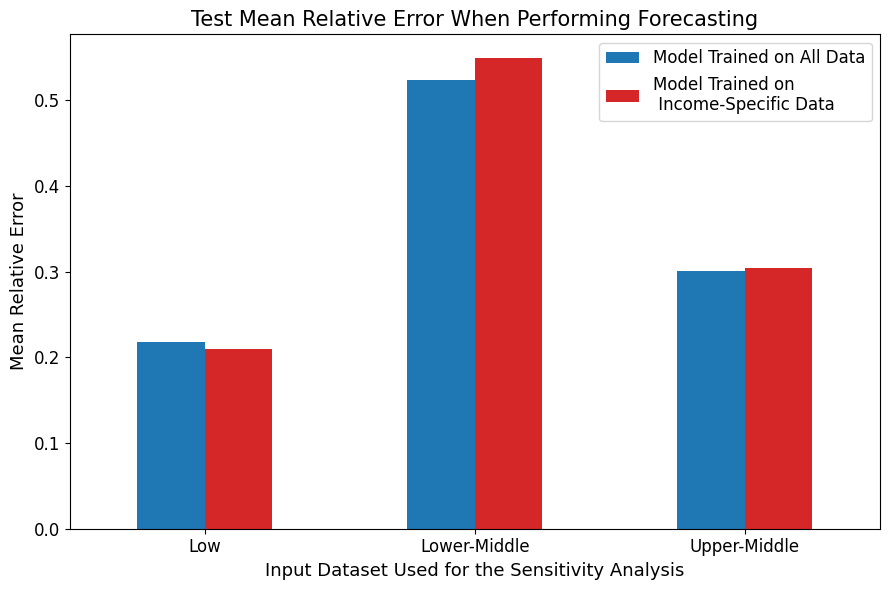
b)

**Figure 36:** a) Mean relative error and b) mean-squared error (log scale) for the Random Forest Stacking Ensemble trained on data from all income levels (blue) and RFSEs trained on data from a specific income level (red) for country-level prediction. The models being compared were tested on data from the same income level.

##### 5.632 Forecasting

Unfortunately, some of the cross-validation folds for the ‘Correlation 0.8’ feature subset had insufficient non-missing data when filtered for just high-income data. As a result, some of the base estimators for the high-income dataset could not be trained, preventing an RFSE from being fit on the high-income sensitivity data, as it expected a certain number of base estimators. As a result, only results from the sensitivity models trained on the low, lower-middle, and upper-middle datasets were presented.

Broadly, there were only small differences in the forecasting performance of the RFSE trained on all data versus RFSEs trained on income-specific data subsets for either MRE or MSE (Figure 37). The RFSE trained on just low-income data had lower error than the RFSE trained on all data (MRE 0.21 vs 0.22 and MSE 58,701 vs 73,343). In contrast, the RFSE trained on just lower-middle income data had higher error than the RFSE trained on all data (MRE 0.55 vs 0.52 and MSE 4,044 vs 2,936). The RFSE trained on just upper-middle income data had the same MRE as the RFSE trained on all data but a slightly lower MSE (606 versus 645).



a)

A graph of a bar chart

AI-generated content may be incorrect.

b)

**Figure 37:** a) Mean relative error and b) mean-squared error (on a log scale) for the Random Forest Stacking Ensemble trained on data from all income levels (blue) and RFSEs trained on data from a specific income level (red) to perform forecasting. The models being compared were tested on data from the same income level.

### 5.7 Comparison of the Random Forest Stacking Ensemble to the Literature

In this section, I compared the MMR estimates of my best-performing Random Forest Stacking Ensemble to the MMR estimates produced by the UN MMEIG’s BMat Model, the Global Burden of Disease Study’s CODEm model, and the GMatH microsimulation model. See the literature review for detailed descriptions of these models and their MMR estimation processes.

#### 5.71 Across Country Comparisons

I first compared my MMR estimates to the literature’s estimates across all countries.

##### 5.711 Percentage Difference

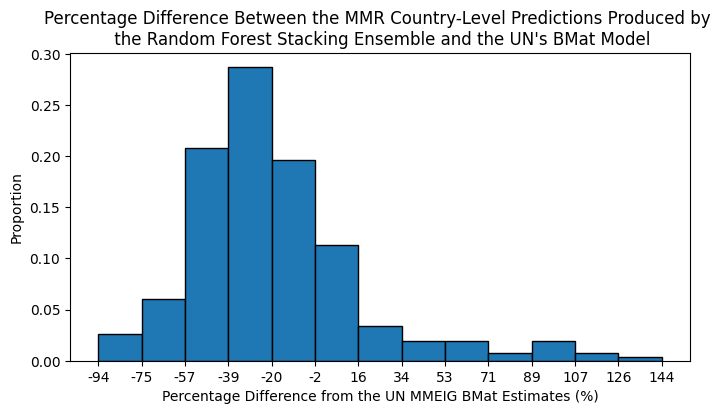
As described in Section 4.6, I took the percentage difference between my MMR estimate for each country, year datapoint and the associated MMR estimate from the literature. A negative percentage difference meant that my estimate was smaller than the literature’s estimate.

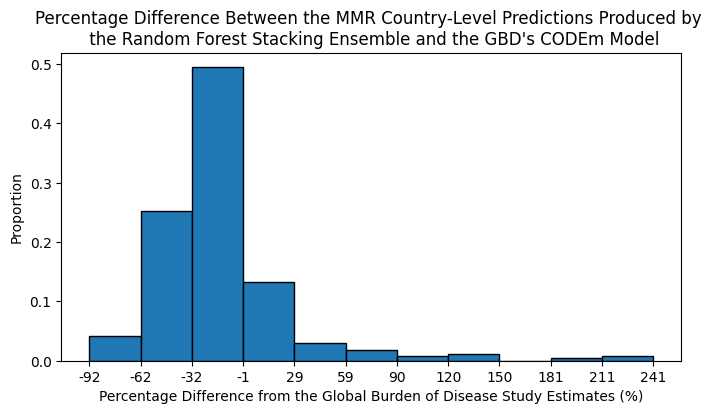
###### 5.7111 Country-Level Predictions

Over 70%, 75%, and 80% of the MMR estimates from my best-performing RFSE were smaller than the corresponding estimates from the BMat, CODEm, and GMatH models, respectively (Figure 38). Over 40% of my MMR predictions were between 75% and 100% smaller than the corresponding GMatH estimates. In contrast, less than 5% of my MMR estimates were over 75% smaller than the BMat or CODEm estimates. More specifically, approximately 50% of my MMR estimates were between 0 and 39% or 0 and 32% smaller than the associated estimates from the Bmat and CODEm models, respectively.

Approximately 15 to 20% of my MMR predictions were larger than the associated BMat and CODEm estimates while only roughly 5% of my predictions were greater than the corresponding GMatH estimates. There were no extreme outlier differences between my MMR predictions and the literature’s estimates.

a)



A graph of a number of blue bars

AI-generated content may be incorrect.

c)

b)

**Figure 38:** Percentage difference between the MMR estimated by my best-performing Random Forest Stacking Ensemble trained for country-level prediction and the MMR estimated by a) the UN MMEIG’s BMat model b) the GBD’s CODem model, and c) the GMatH microsimulation model.

###### 5.7112 Forecasting

Almost 70% and 60% of my MMR forecasts were smaller than the associated BMat and CODEm estimates, respectively (Figures 39a, 39b). In contrast, over 80% of my MMR estimates were smaller than the associated GMatH estimates (Figure 39c). Over 40% of my model’s MMR predictions were between 0 and 50% smaller than both the BMat and CODEm estimates. In contrast, over 50% of my forecasts were between 65 and 100% smaller than the GMatH predictions.

Approximately 30% of my MMR forecasts were larger than the BMat and CODEm estimates while only roughly 10% of my forecasts were larger than the associated GMatH predictions. There was a small proportion of instances where my forecasts were over 1300% greater than the corresponding BMat estimates and between 960 and 1,000% greater than the associated CODEm predictions. In comparison, this small proportion of instances was only between 180 and 215% larger than the corresponding GMatH estimates.

A graph with numbers and a number of text

AI-generated content may be incorrect.

a)

A graph with numbers and a number of people

AI-generated content may be incorrect.

b)

A graph of a number of blue squares

AI-generated content may be incorrect.

c)

**Figure 39:** Percentage difference between MMR estimated by my best-performing Random Forest Stacking Ensemble trained to perform forecasting and the MMR estimated by a) the UN MMEIG’s BMat model b) the GBD’s CODem model, and c) the GMatH microsimulation model.

###### 5.7113 Summary of Differences

The magnitude difference between the GMatH predictions and both my MMR country-level predictions and forecasts was generally larger than the differences to BMat and CODEm.

##### 5.712 Coverage

67.1% of my RFSE’s country-level MMR predictions were within the 95% confidence intervals (CI) of GMatH’s MMR predictions (Table 15). In contrast, only 22.3% and 29.4% were within the 95% CI of the BMat or CODEm’s MMR estimates, respectively. Similarly, only 20.4%, 33.2%, and 67.1% of the ground truth MMR estimates used to train my RFSE were within these models’ 95% confidence intervals.

A higher proportion of my RFSE’s MMR forecasts were within BMat and GMatH’s 95% CI than the proportion of my model’s country-level predictions. In contrast, a smaller proportion of my model’s MMR forecasts were within the 95% CI of the CODEm model’s estimates. More specifically, 30.9, 23.5, and 81.6% of my models’ MMR forecasts were within the 95% CI of the BMat, CODEm, and GMatH models, respectively. A similar, but slightly higher, proportion of the ground truth MMR estimates used to train my model were within these 95% confidence intervals (32.6%, 25.1%, and 84.9%).

**Table 15:** The percentage of MMR country-level predictions and forecasts from my best-performing Random Forest Stacking Ensemble that fell within the 95% confidence intervals (CI) for the BMat (blue), CODEm (green), and GMatH (purple) models’ predictions. The proportion of ground truth MMR estimates used to train my model that fell within these CI was also presented.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Type of Analysis** | **Percent of RFSE MMR estimates within BMat’s 95% CI** | **Percent of ground truth MMR estimates within BMat’s 95% CI** | **Percent of RFSE MMR estimates within CODEm’s 95% CI** | **Percent of ground truth MMR estimates within CODEm’s 95% CI** | **Percent of RFSE MMR estimates within GMatH’s 95% CI** | **Percent of ground truth MMR estimates within GMatH’s 95% CI** |
| **Country-Level Prediction** | 22.3% | 20.4% | 29.4% | 33.2% | 67.1% | 67.1% |
| **Forecasting** | 30.9% | 32.6% | 23.5% | 25.1% | 81.6% | 84.9% |

#### 5.72 Per-Country Comparison

To better understand the differences between my best-performing model’s predictions and those found in the literature, I compared my model’s MMR estimates for a specific country from each income level to the associated BMat, CODEm, and GMatH estimates.

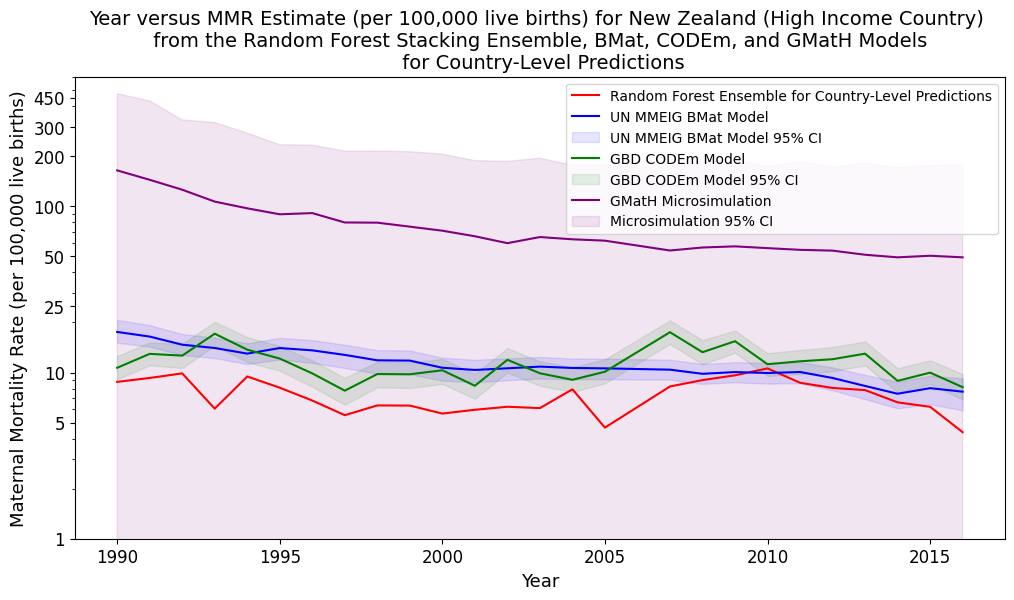
##### 5.721 Comparisons with My RFSE Trained to Perform Country-Level Prediction

Compared to the other models, my Random Forest Stacking Ensemble underpredicted New Zealand’s MMR between 1990 and 2015, with greater underprediction pre-2010 (Figure 40a). While my model’s estimates were outside the 95% confidence intervals (CI) of the BMat and CODEm models, the actual difference in MMR between the estimates was between 5 and 20. The GMatH model strongly overestimated New Zealand’s MMR, predicting MMR to be close to 200 in 1990 and fall to roughly 50 by 2015. In contrast, the BMat and CODEm models did not predict an MMR of higher than 25 for New Zealand in this time interval. These larger GMatH estimates came with a wide 95% CI that enveloped my model’s MMR estimates.

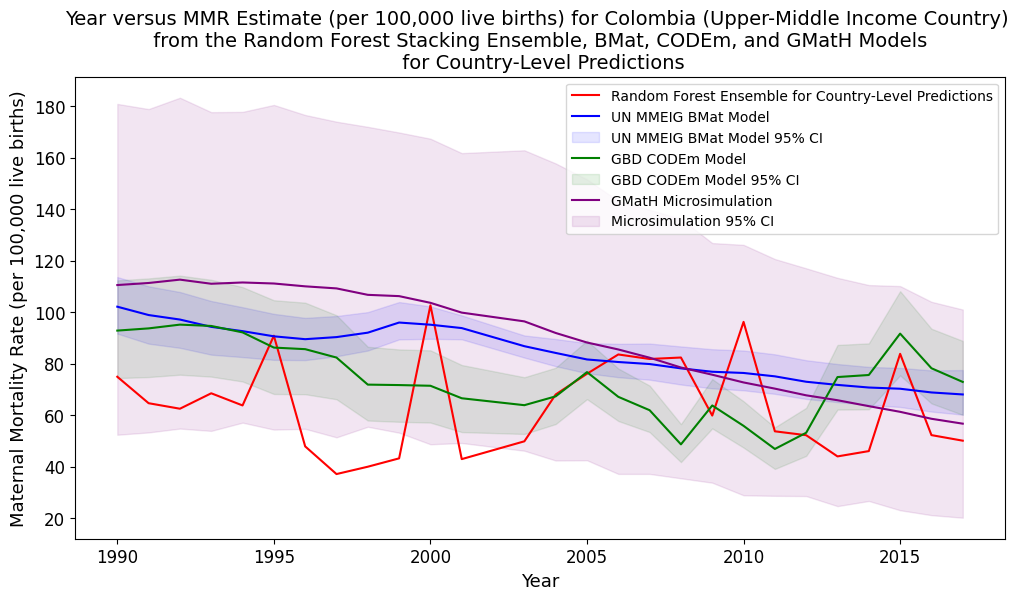
My RFSE’s estimates had greater intersection with the literature’s estimates for Colombia, an upper-middle income country (Figure 40b). My model’s estimates were generally 0 to 40 points off the closest literature estimate. They were generally within the GMatH model’s wide 95% confidence interval, with the GMatH predictions again higher than the other literature estimates. At times, my model’s estimates were within the 95% CI of either the BMat or CODEm models’ predictions. However, my model’s estimates fluctuated more strongly between consecutive years, compared to the smoother literature estimates. The majority of the sharp peaks predicted by my model corresponded to the years with the greatest amount of non-missing data (see Figure 3 and Table 4 in Section 4.21).

My models’ estimates were completely within the 95% CIs of at least one other model when predicting for Kenya and Rwanda, which were designated as lower-middle and low-income countries, respectively (Figures 40c, 40d). However, the magnitude difference between my estimates and the literature’s estimates was in the hundreds, with the greatest difference observed between my estimates and the CODEm predictions. My RFSE’s MMR estimates for these countries was generally greater than the BMat and CODEm estimates. They were more often greater than the GMatH estimates for Rwanda than for Kenya. However, these comparisons must be taken with a grain of salt, as there were only 4 datapoints for each of Kenya and Rwanda over this period of time.

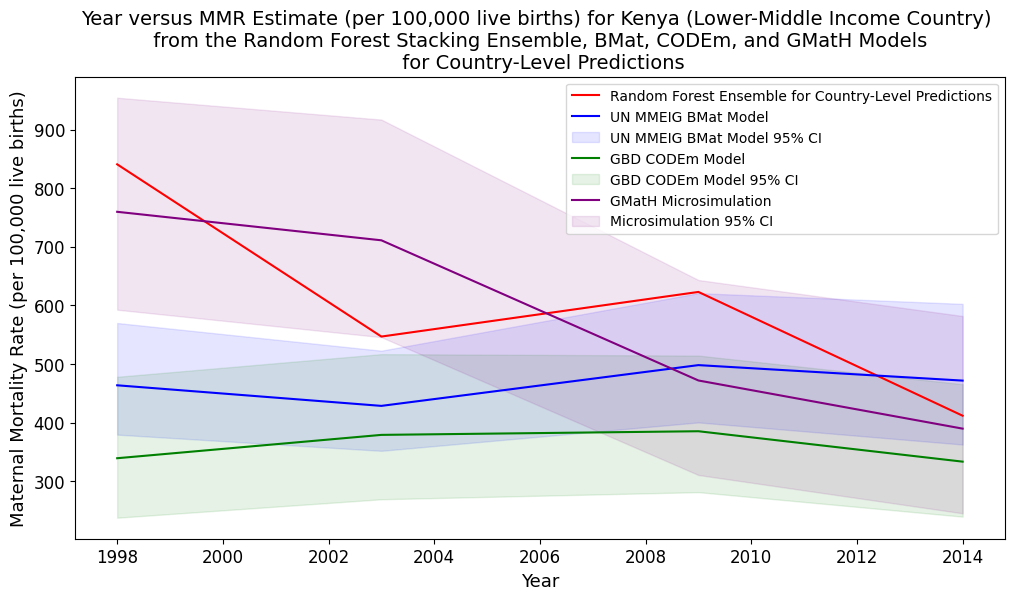
Therefore, it appeared that my model’s country-level predictions were underestimates for higher-income countries but over-estimates for lower-income countries.



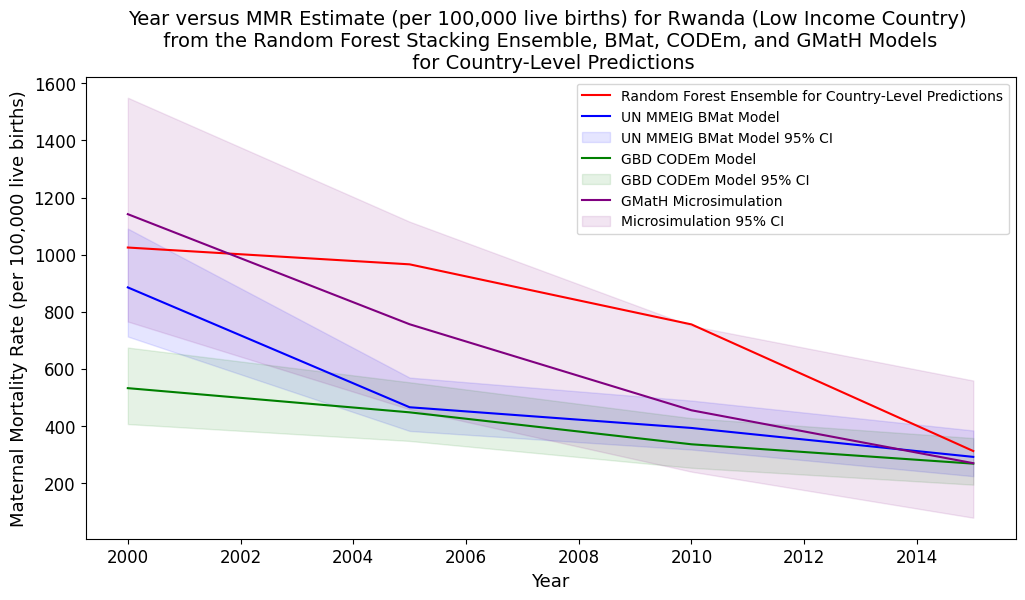
a)

****

b)



c)



d)

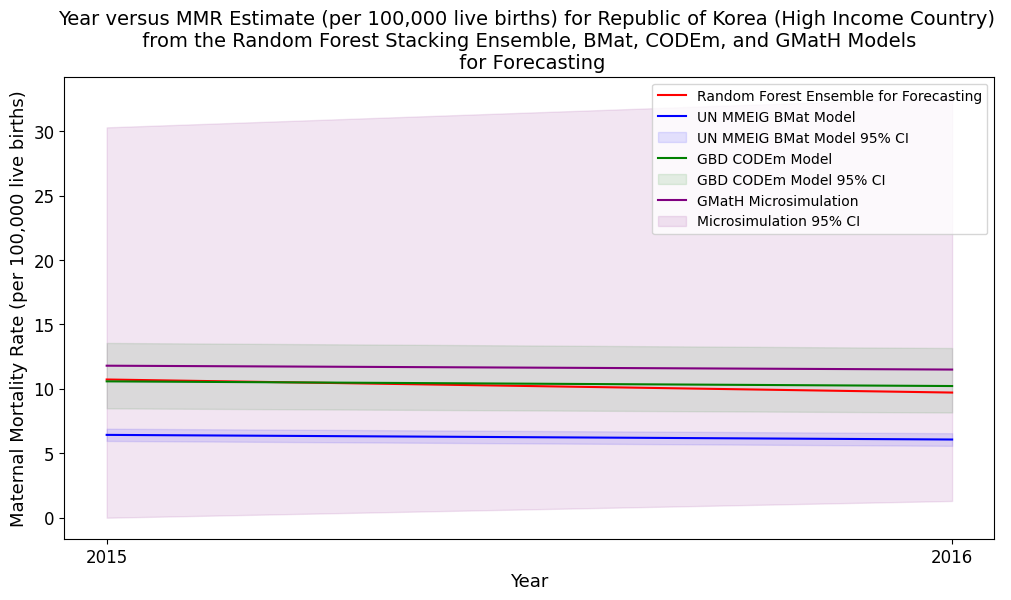
**Figure 40:** Comparison of my best-performing Random Forest Stacking Ensemble’s MMR country-level predictions to the associated estimates from the BMat, CODEm, and GMatH models for one country from each World Bank defined income level. a) New Zealand (high-income) (log scale), b) Colombia (upper-middle income), c) Kenya (lower-middle income), and d) Rwanda (low-income).

##### 5.722 Comparisons with My RFSE Trained to Perform Forecasting

There was less room for comparison between my model’s MMR forecasts and the literature’s estimates because all comparisons were performed on my model’s test values, which were confined between 2015 and 2018. Additionally, not all samples in my test set had non-missing MMR values, meaning some of the countries presented in this section did not have an associated MMR prediction for every year in the test set.

My best-performing Random Forest Stacking Ensemble’s MMR forecasts were always in the 95% confidence intervals (CI) of the literature’s corresponding estimates (Figure 41). For the high and upper-middle income countries (Republic of Korea and Armenia), my model’s MMR forecasts were the second lowest, and either within the CODEm or BMat 95% CIs. The actual difference between my estimates and the CODEm estimates for the Republic of Korea’s MMRs was in the single digits (Figure 41a). My model’s MMR forecasts for Chad were also the second-lowest available, and within GMatH’s 95% CI (Figure 41d). Unfortunately, there was only one test datapoint for Chad and every other low-income country contained in the test set.

In contrast, my model’s MMR forecasts were the highest available for the first half of the testing period for Sri Lanka, a lower-middle income country (Figure 41c). Its estimates in the second half of this training period were very similar to the GMatH and BMat predictions. All of its estimates in the test period were within the 95% CI of the literature models.



a)

A graph with lines and text

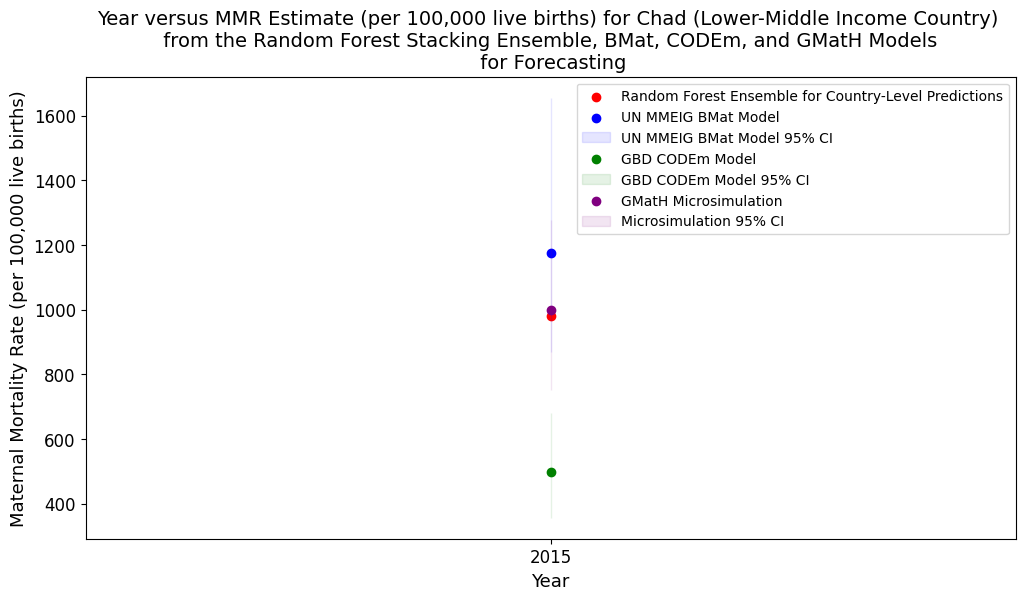
AI-generated content may be incorrect.

b)

A graph with different colored lines

AI-generated content may be incorrect.

c)



d)

**Figure 41:** Comparison of my best-performing Random Forest Stacking Ensemble’s MMR forecasts to the associated estimates from the BMat, CODEm, and GMatH models for one country from each World Bank defined income level. a) Republic of Korea (high-income), b) Armenia (upper-middle income), c) Sri Lanka (lower-middle income), and d) Chad (low-income).