## 5. Results

### 5.1 Exploratory Data Analysis

First, I present all results of the exploratory data analysis that were not given in the methods. These findings contextualised model performance, as discussed in Section 6.

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

The following analysis used the input dataset before cleaning or pre-processing to better understand the pattern of missing MMR data, and thus which samples were most likely to be dropped during cleaning. 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 between income levels observed between lower-middle and upper-middle income countries (see Figure 10). The proportion of missing estimates decreased as income level increased between 1985 and 2010. Additionally, the proportion of missing estimates for each income level decreased over this time. More explicitly, 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).

A graph of different colored dots

AI-generated content may be incorrect.

**Figure 10:** The proportion of samples missing MMR estimates per income level per year between 1985 and 2018 in the input dataset before cleaning or pre-processing. The proportion of missing data across low-income countries was plotted in red, lower-middle in orange, upper-middle in blue, and high in green.

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

As described in Section 4.22, I calculated key summary statistics about a few of the more meaningful features in the cleaned dataset to increase understanding of the input data (see 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 important 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 level countries was missing 58% of data regarding ‘women participating in own health care decisions (% of women age 15-49)’ while the dataset for the highest income countries was missing 99.9%.

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

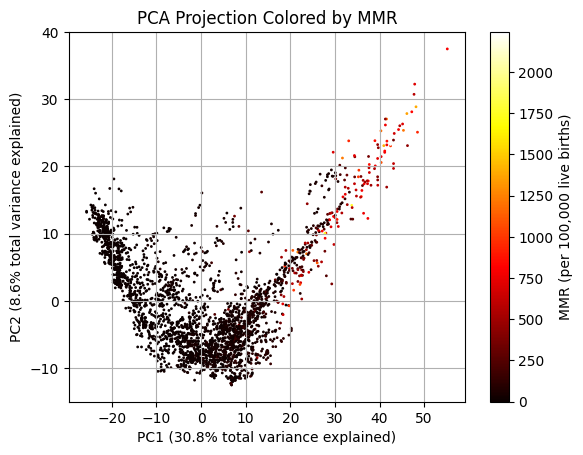
**Table 10:** Mean, median, standard deviation and proportion of missing data of features with a meaningful relationship to MMR. The key summary statistics were presented per income level.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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

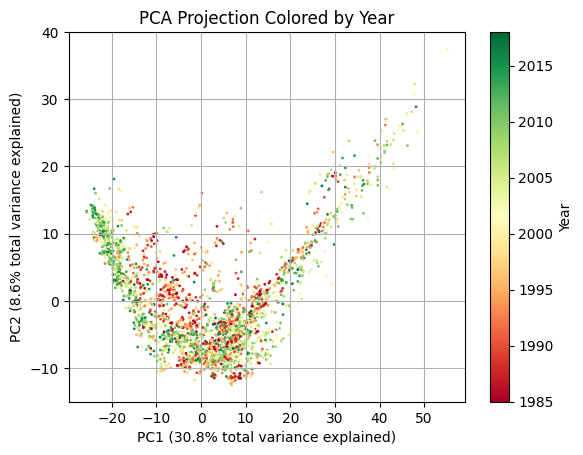
The feature dataset was projected onto its first two principal components for better visualisation of patterns and clusters in the data (See Figure 11). The dense cluster of data in the bottom center-left of Figures 11a and 11b corresponded to the countries with the lowest MMRs and highest income levels. This cluster extended up to left of the plot’s middle. 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, where the edges corresponded to more recent years (light-green/yellow dots) and the inner-part of the U above the valley corresponded to years further in the past (red/orange dots).

b)

 A graph with a chart and a diagram

AI-generated content may be incorrect.

a)

****

c)

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

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

#### 5.21 Country-Level Prediction

The ground truth MMR distribution in the train/validation set overlapped the distribution in the test set for samples from lower-middle, upper-middle, and high-income countries (see Figure 12). More explicitly, the quartile 2 (Q2) values for the lower-middle, upper-middle, and high-income countries were generally very similar between the train/validation and test sets. For instance, there was no difference in Q2 between the datasets for lower-middle income countries. However, the train/validation data for these income levels contained samples with higher MMR estimates than the associated test set, as the Q3 and maximum MMR values were higher for the train/validation data than the test data. For example, the Q3 MMR for lower-middle countries was 283 in the train/validation set and 60 in the test set. The differences between the train/validation and test MMR data decreased for the upper-middle and high-income countries. The the train/validation set for these three income levels also contained outliers with higher MMR values than the associated test sets. These results indicated that the test MMR distribution for lower-middle, upper-middle, and high-income countries generally lay within the train/validation MMR distribution, where the distributions had similar averages, but the train/validation distribution had more MMR estimates with higher values.

There was greater difference between MMR estimates in the train/validation data and test data distribution for low-income countries. For example, the train/validation data had a Q2 MMR value of 610 while the test distribution had a Q2 MMR value of 772. Similarly, the test set’s Q3 and Q1 were greater than the train/validation set’s Q3 and Q1 by 126 and 103, respectively. As a result, the test dataset’s MMR values were shifted higher than the train dataset’s MMR values. This may have been due to the small number of low-income samples, which cover a wide range of MMR estimates (Table 10). Thus, the small subset of samples with MMR values at the higher end of the range may have been included in the test set by chance. However, as a note, the extreme MMR values in the train/validation set for low-income countries were greater than the extremes in the corresponding test dataset.

A graph showing a number of different colored squares

AI-generated content may be incorrect.

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

#### 5.22 Forecasting

The Q1 to Q3 distribution of MMR values in the train/validation versus test sets used for forecasting were often offset, meaning the test distribution was not encapsulated within the train/validation distribution (see Figure 13). For example, the Q3 MMR value for the high-income train/validation set was lower than the Q3 value for the test set by 27. Additionally, the Q1 MMR value for lower-middle, upper-middle, and high-income countries was larger in the train/validation set than the test set. The outliers and maximum MMR values in the train/validation dataset were also often larger than those in the test dataset.

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

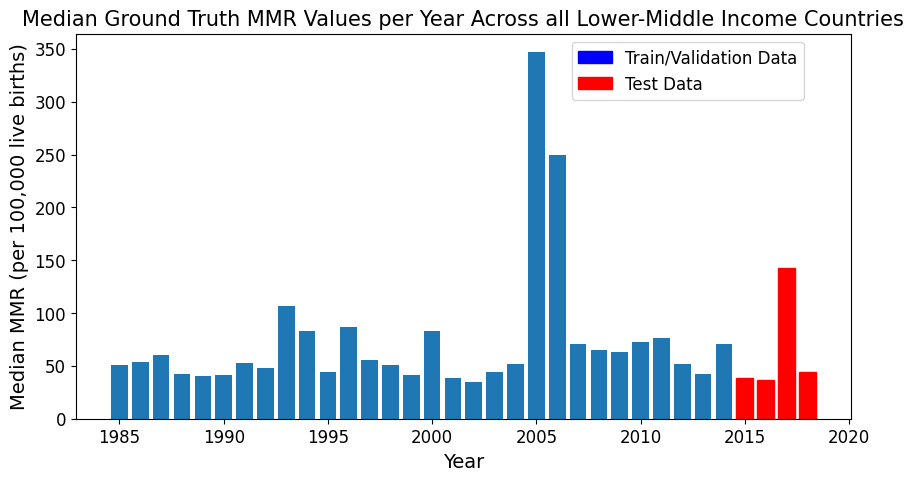
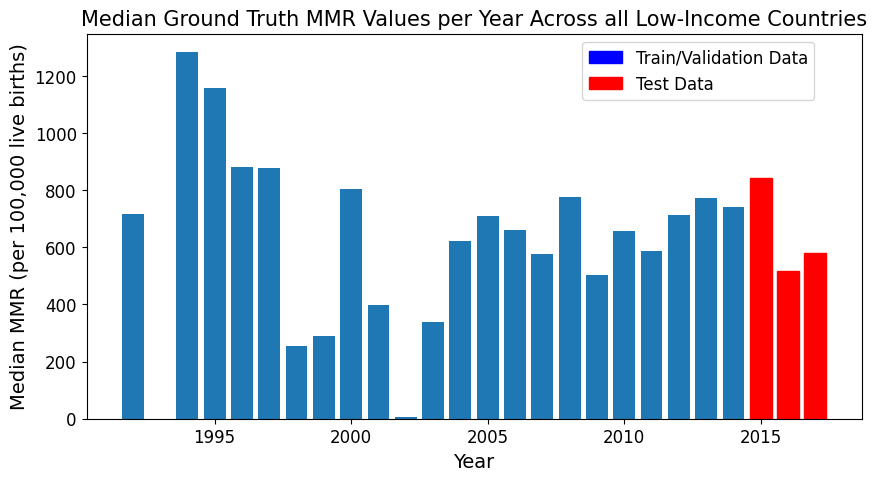


**Figure 13:** Minimum, quartile (Q) 1, Q2, Q3, and maximum ground truth MMR values for each income level across the train/validation and test sets used for forecasting. The y-axis was shown with a log-scale. The boxplot bars giving the minimum values sometimes appeared to be cut-off because they extended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

The differences between the MMR estimates in the train/validation and test sets used for forecasting were further visualised in Figure 14. The low-income test set had median MMR values within the spread of the train/validation set (Figure 14a). However, it did not have 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 values per year across all a) low-income, b) lower-middle income, c) upper-middle income, and d) high-income countries. The bars represented the train/validation data (blue) and test data (red) 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 data. Only two countries (Oman and Uruguay) had non-NAN test MMR values for 2018. As shown in Table 11, these countries’ MMR ground truth values were larger than the norm for high-income countries throughout the test set, consistently ranging between 14 and 25. In the non-outlier test years, the median ground truth MMR did not reflect these high values because it was brought down by low MMR values in other high-income countries. For example, in 2015 Australia and Norway had ground truth MMR values 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 that tended to have 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 Base Model Predictive Performance

This section discusses the predictive performance of the base Random Forest, XGBoost, and LightGBM models trained on cross-validation data curated using different combinations of feature selection and missing data removal techniques. No results from stacking and voting ensembles were presented in this section. As described in Section 4.412, predictive performance was measured using mean relative error, MSE, RMSE, MAE, and R2. However, to keep this chapter concise, only the models’ mean relative error (MRE) and MSE were presented, with the other metrics given in Appendix 9.1. While MRE provided information about the model’s predictive error relative to the size of the ground truth values and predictions, MSE provided information about outliers. Combining the two metrics provided a comprehensive understanding of model performance.

Each combination of model type, feature selection method and missing data removal threshold had 5 associated training folds. The combination’s performance was the average test performance of the models trained on these folds, with the average plotted on the graphs in this section. The error bars used on the following graphs represented the standard deviation of the model’s error across the 5 folds.

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

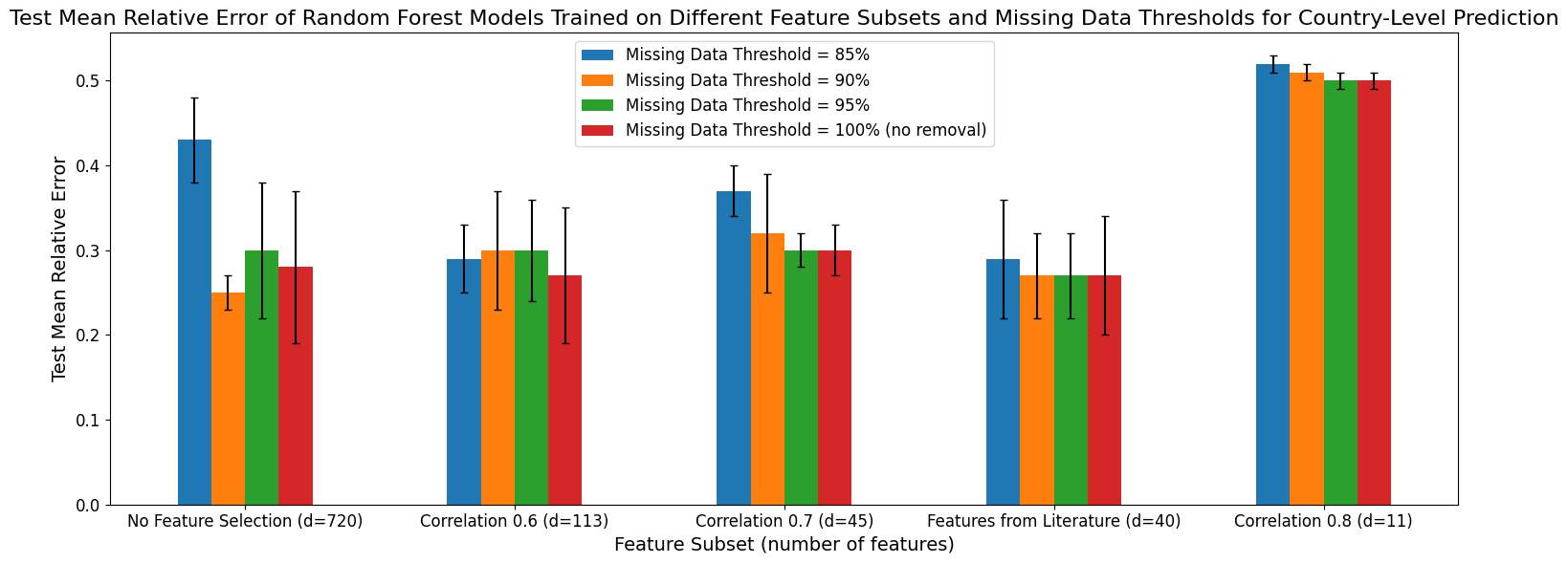
Performance of Random Forest, XGBoost, and LightGBM models were plotted separately to visualise how each model type performed when fit on datasets curated with different pre-processing techniques. The best performing model type was discussed later.

##### 5.311: Random Forest

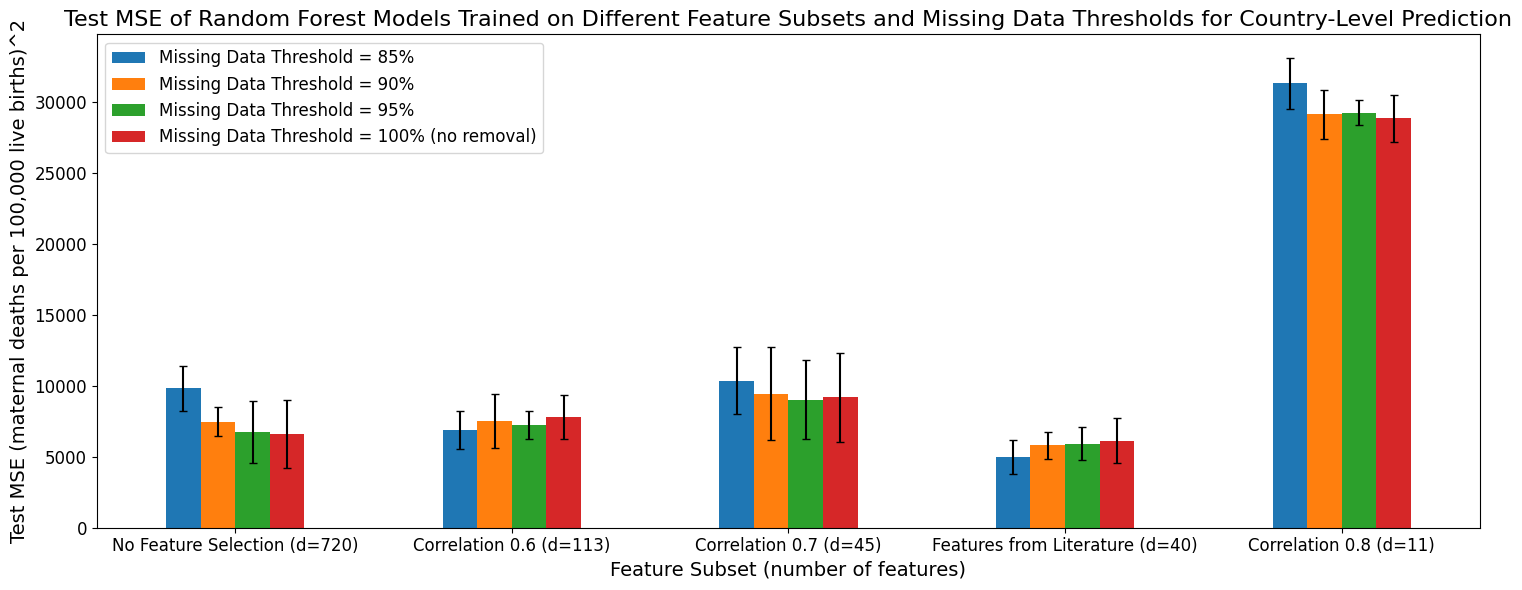
Random Forest models trained on different feature subsets generally had similar performance, especially when considering standard deviation in their performance (see Figure 15). The models’ MRE typically ranged between 0.25 and 0.32 across the different feature subsets, and their MSE generally ranged between 5,000 and 10,000. The exception was models trained on feature subsets that only contained variables that had an absolute pairwise Pearson’s correlation coefficient with MMR greater than 0.8 (referred to as ‘Correlation 0.8). 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 model with the lowest MRE (0.25) was trained on data curated without feature selection and with a missing data threshold of 90% (Figure 15a). However, the model with the lowest MSE (4,986) was trained on the subset of features hand-picked from the literature and with a missing data threshold of 85% (Figure 15b). In general, the models’ MSE scores were most consistently lower when trained on the feature subset hand-picked from the literature, with this trend present but less noticeable for the models’ MRE. Thus, models trained on hand-picked features from the literature may have less outlier-induced error.

The variation between trends in the models’ MSE and MRE scores was also noticeable in the differences between models’ predictive performance on different missing data thresholds. For example, the relative ordering of best to worst performing model trained on the same feature subset but different missing data thresholds changed when considering MSE versus MRE. For example, 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 threshold 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 a missing data threshold of 85%. Therefore, a missing data threshold of 85% may be less affected by outliers while no missing data removal may be more affected. 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.

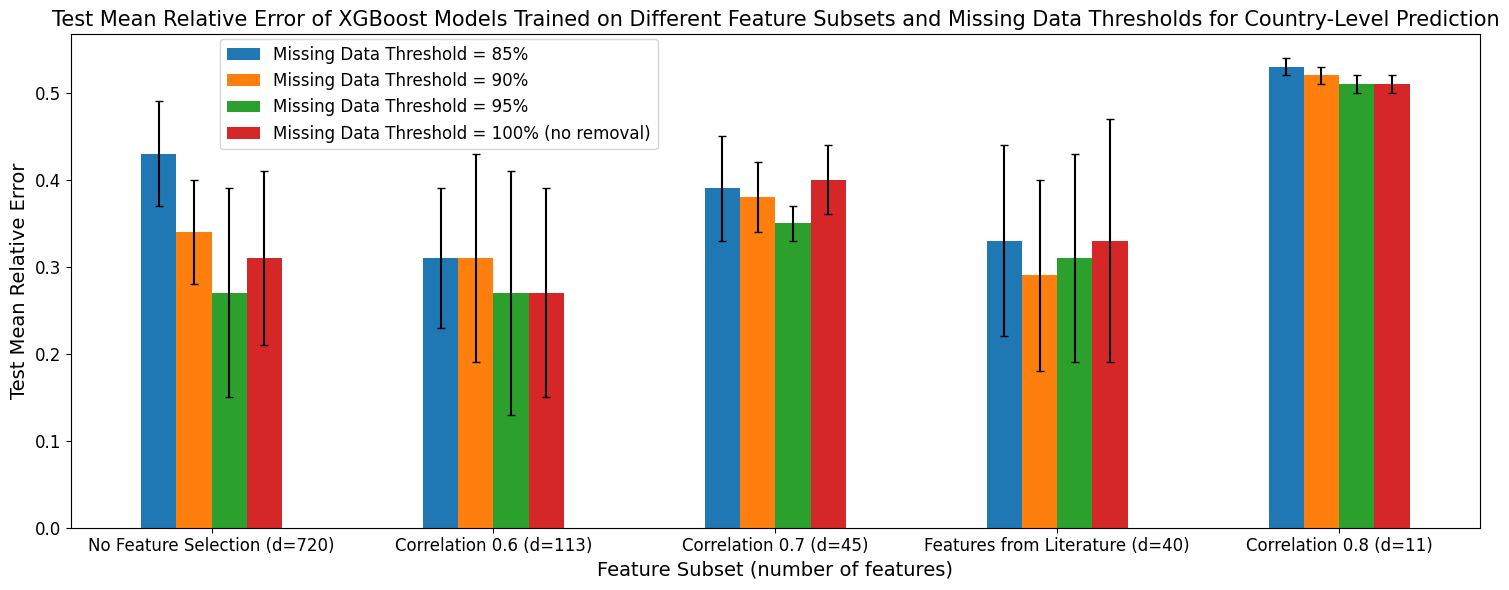
##### 5.312: XGBoost

XGBoost models had similar trends in their predictive performance as the Random Forest models (see Figure 16). For example, the XGBoost models similarly had the lowest performance in both MRE and MSE 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. Similar to the Random Forest models, XGBoost models trained on a missing data threshold of 85% had the highest MRE across most feature subsets, but no consistent trend was observed in their MSE.

The MRE for XGBoost models ranged from 0.27 to 0.43 when excluding datasets with the ‘Correlation 0.8’ feature subset, with both the lower and upper bounds on the error higher than for the Random Forest models. The MSE for XGBoost models ranged from 4,000 to 10,000 for all feature subsets excluding ‘Correlation 0.8’, with a slightly smaller lower bound that the Random Forest models.

One of the major differences between the XGBoost and Random Forest models was the magnitude of standard deviation, with XGBoost having larger standard deviation in its predictive performance on different cross-validation folds than Random Forest models. For example, the standard deviation in MSE for XGBoost models trained with no feature selection ranged from 2,271 to 5,037 versus 1,021 to 2,379 for Random Forest models similarly trained with no feature selection. Additionally, Random Forest models’ higher performance with the hand-picked feature subset was not observed with XGBoost models, especially with the latter’s higher standard deviations.

As observed with the Random Forest models, there was no universally best performing feature subset or missing data threshold, especially given the XGBoost models’ wide standard deviations. For example, there were 3 models with the same lowest MRE (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.



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

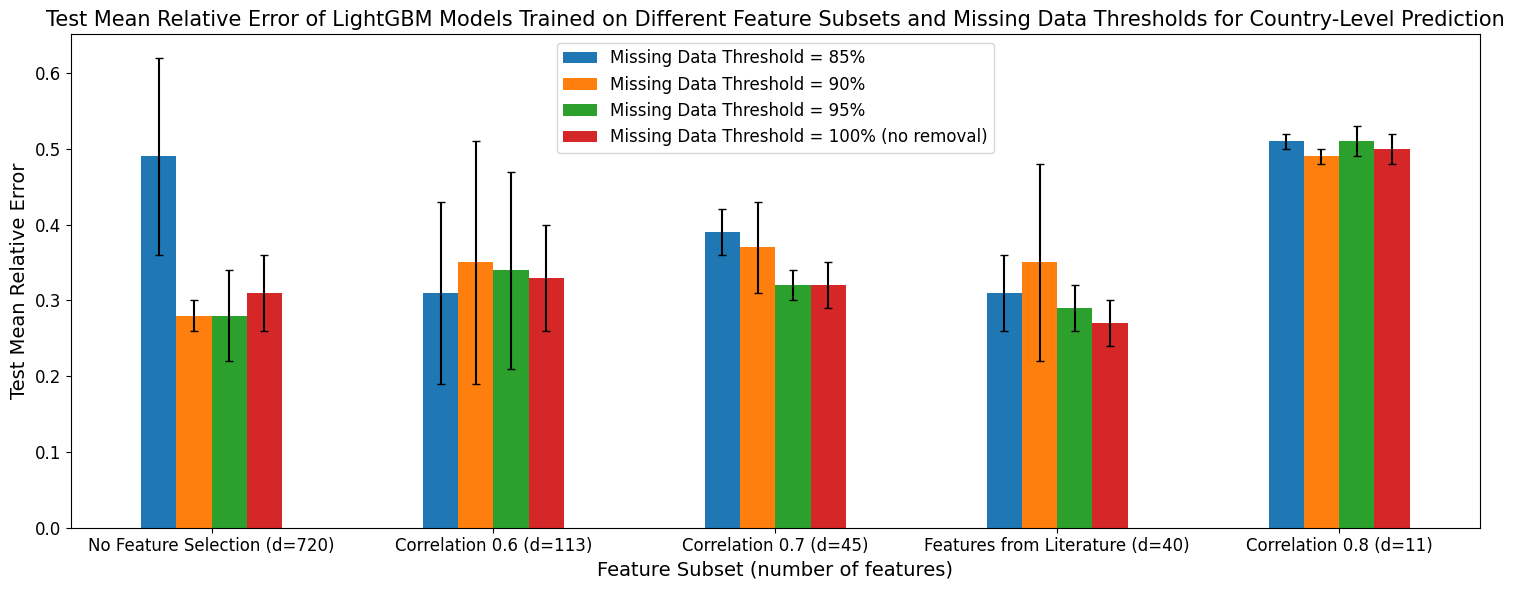
The LightGBM models had similar performance trends as the XGBoost and Random Forest models. For example, the models had the worst performance on the ‘Correlation 0.8’ feature subset and among the worst performance on the ‘Correlation 0.7’ subset. Additionally, models trained on a missing data threshold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

Excluding performance on datasets with ‘Correlation 0.8’ feature selection, the MRE for LightGBM models ranged from 0.27 to 0.49 and the MSE varied 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.

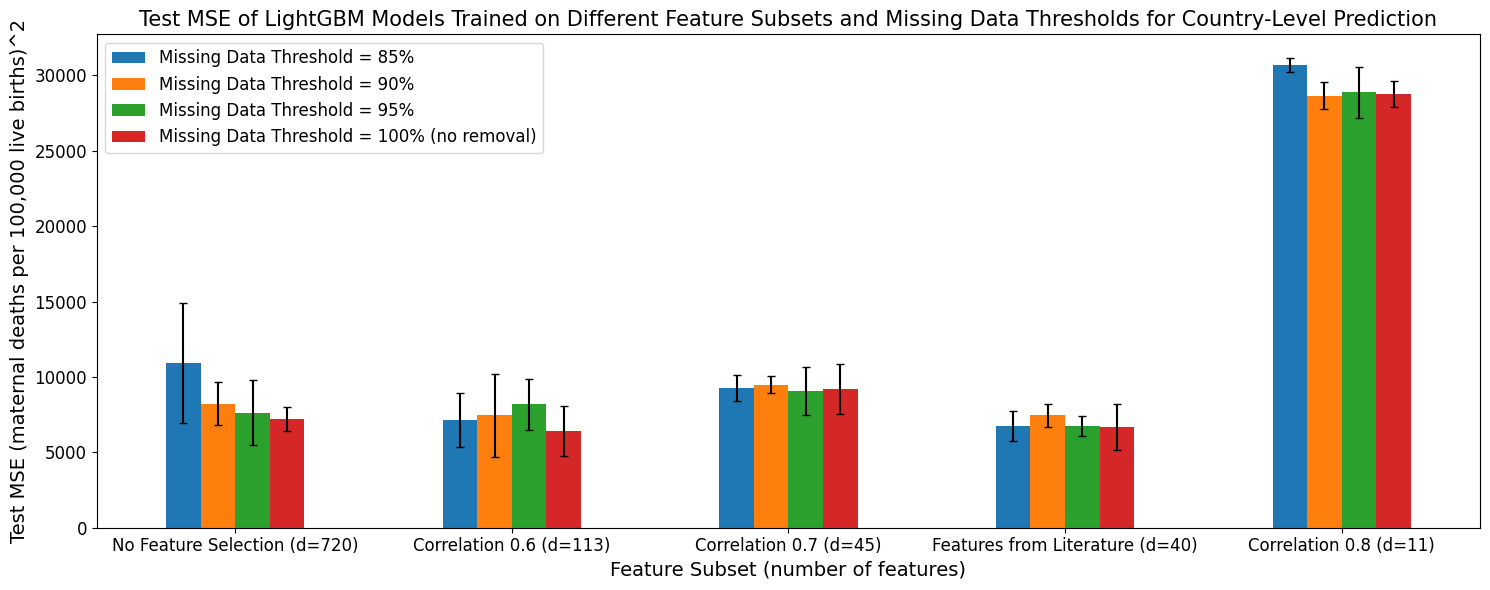
The LightGBM models had more reasonable standard deviations than the XGBoost models but higher standard deviations that Random Forest models. For instance, LightGBM models trained on datasets with no feature selection had standard deviations in their MSE ranging from 777 to 3,989. Similar to the XGBoost models and unlike Random Forest models, LightGBM models did not consistently have higher performance on the hand-picked feature subset.

Nevertheless, the LightGBM models with the lowest MRE were trained on datasets with the hand-picked feature subset and no missing data threshold (0.27). In contrast, the LightGBM models with the lowest MSE were trained on datasets with no feature selection and a missing data threshold of 95%. These combinations of pre-processing techniques also produced the best performing XGBoost models. However, the wide standard deviations prevent a conclusion of these techniques being the highest performing combination among all the models, especially given these techniques 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

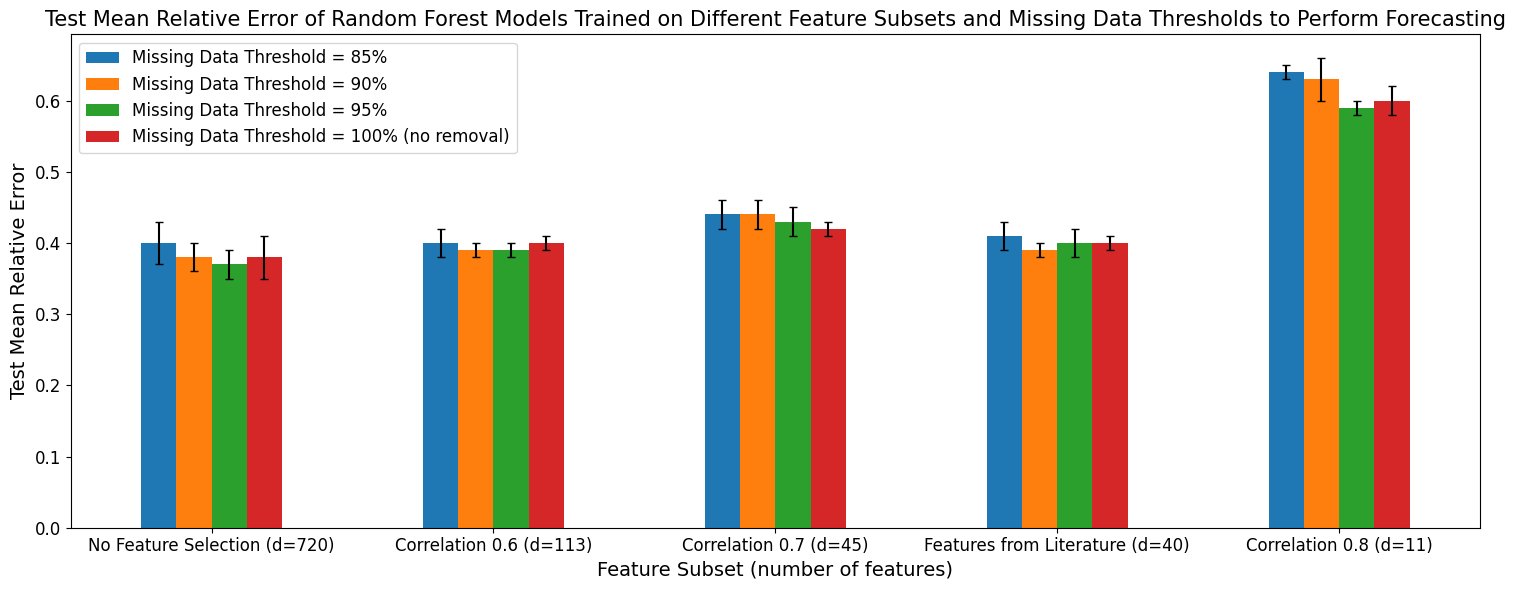
A similar analysis was conducted for models used for forecasting.

##### 5.321: Random Forest

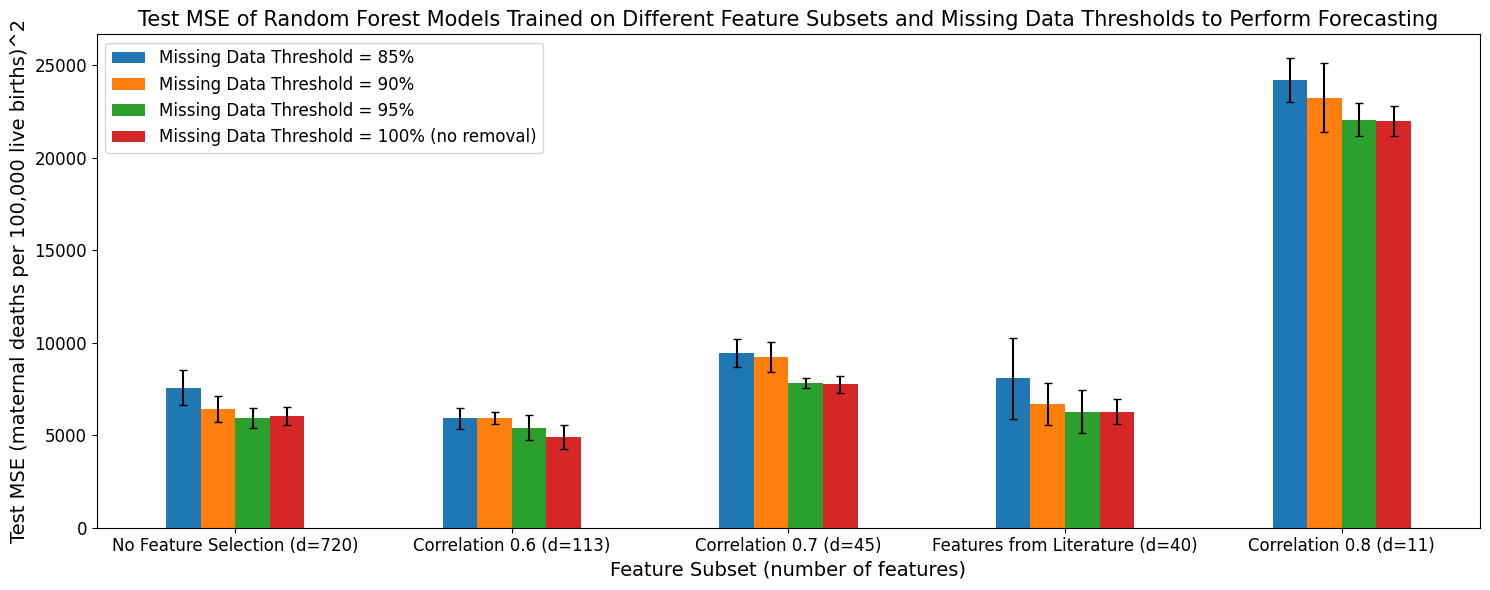
As discussed with respect to the models trained for country-level prediction, Random Forest models trained to complete forecasting models had the highest MRE and MSE scores on datasets curated with a ‘Correlation 0.8’ feature subset. However, in general, the predictive MRE of the Random Forest models trained for forecasting were more uniform and higher, ranging from 0.37 to 0.40, when excluding models trained with the ‘Correlation 0.8’ feature subset (Figure 18a). Thus, there was very little difference in performance between Random Forest models trained on the different feature subsets and missing data thresholds, again excluding the ‘Correlation 0.8’ subset, especially when considering the overlapping standard deviations. 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 techniques (Figure 18b). The Random Forest models’ MSE scores varied from 4,900 to 9,500. 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 mean relative error. Thus, it may more effectively handle outliers.

As previously observed, the standard deviation in the prediction 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.

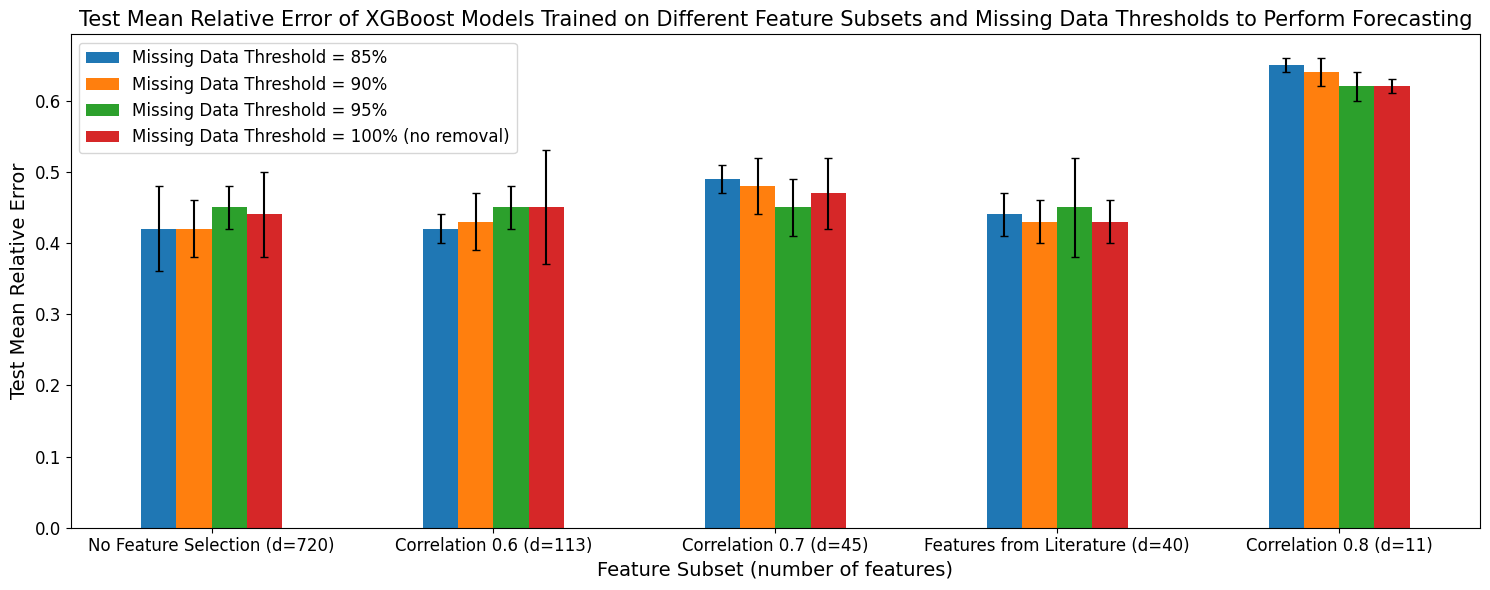
##### 5.322: XGBoost

Similar to the Random Forest models discussed previously, the XGBoost models produced for forecasting had a more uniform and higher MRE distribution that the XGBoost models produced for country-level prediction (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 with consistently lower MRE, especially when taking into account the standard error in each error estimate across the cross-validation folds.

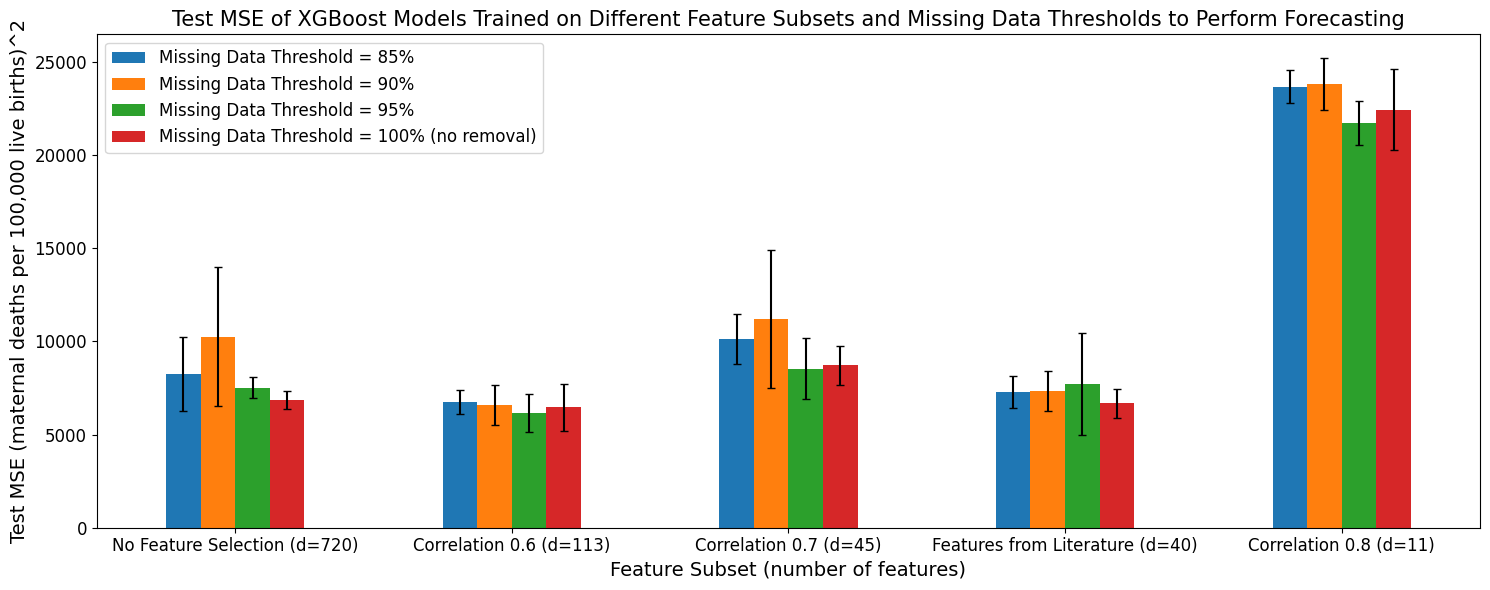
The MSE score for XGBoost models ranged from 6,100 to 11,200 when excluding models trained on the ‘Correlation 0.8’ feature subset (Figure 19b). This range had larger lower and upper bounds than the MSE range for the Random Forest models and for the XGBoost models trained for country-level prediction. However, as noted in Section 5.321 above, the XGBoost models trained on data with the ‘Correlation 0.6’ feature subset had MSE scores that were more consistently low, all less than 7,000. As observed previously, the XGBoost models had higher standard deviation in their MSE scores (497 to 3,734) across cross-validation folds than the Random Forest models (270 to 2,188).

3 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)



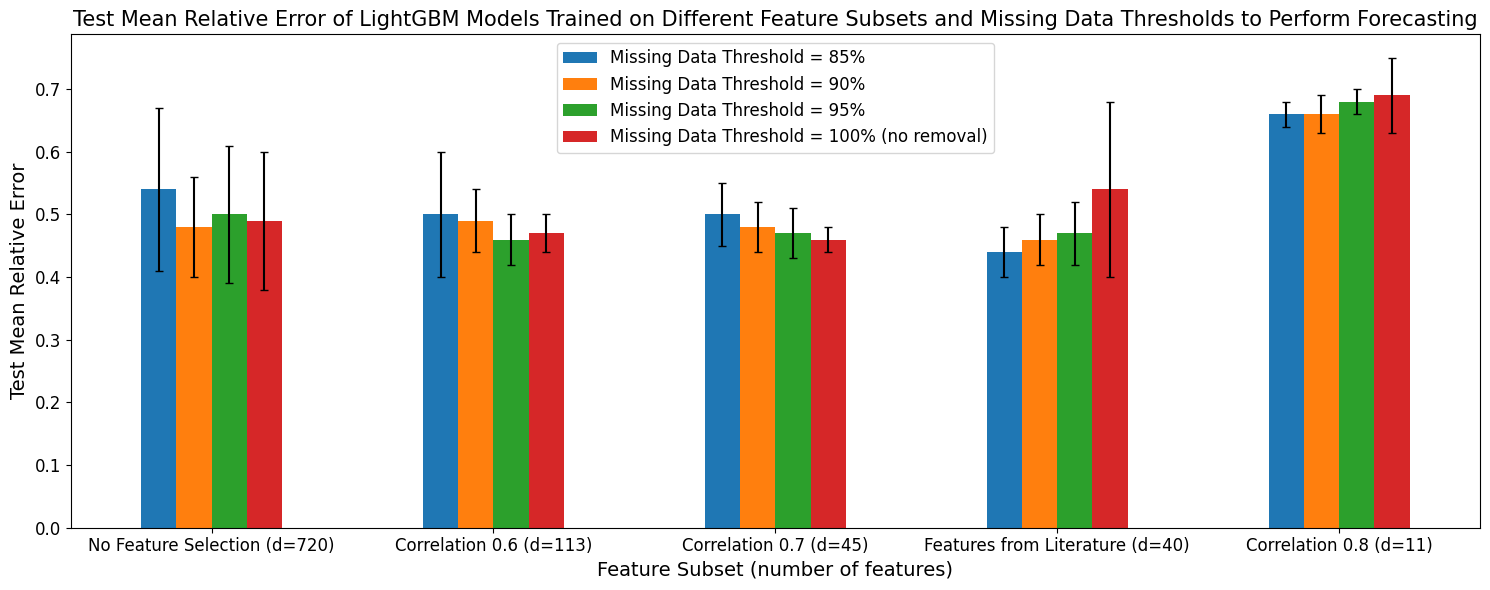
**Figure 19:** a) Mean relative error and b) mean-squared error for XGBoost base estimators fit on different feature subsets and missing data thresholds for forecasting.

##### 5.323: LightGBM

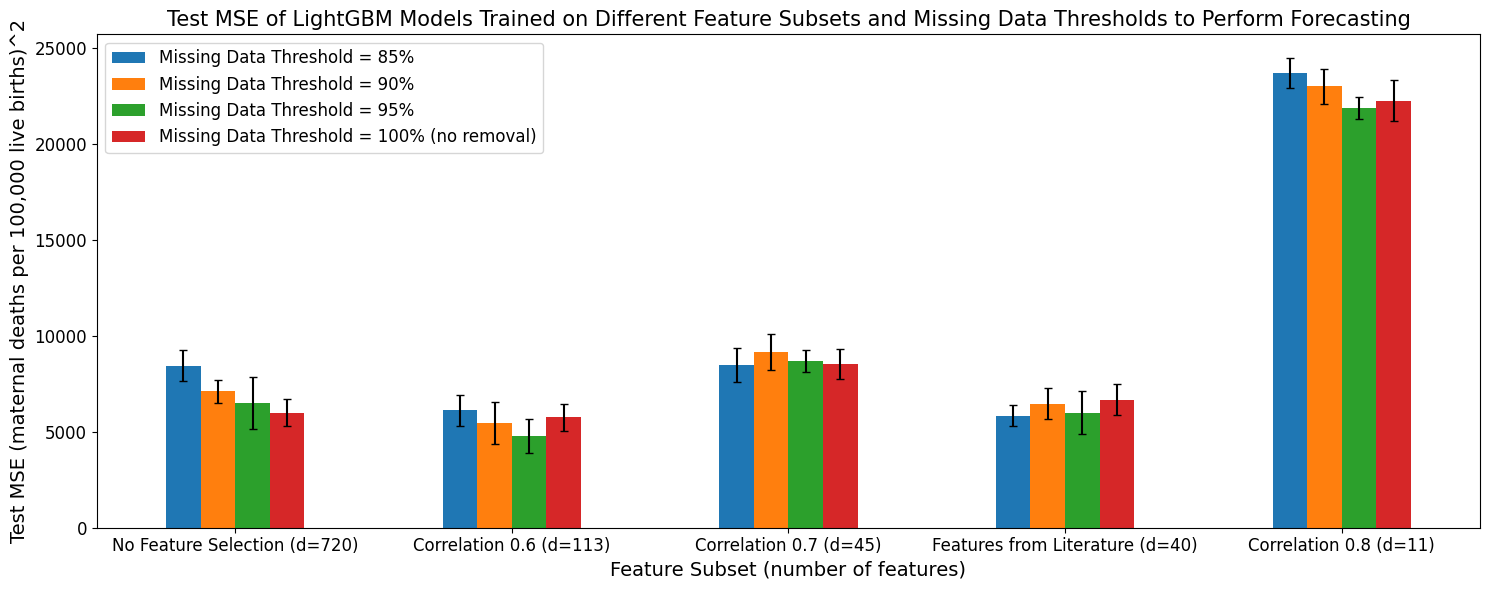
As observed previously, the LightGBM models trained for forecasting purposes had more uniform, higher MRE scores when compared to LightGBM models trained for country-level prediction (Figure 20a). When excluding the low performing ‘Correlation 0.8’ feature subset, the LightGBM models trained for forecasting had a MRE between 0.44 and 0.54. This was higher than the range for the Random Forest models trained for forecasting. Additionally, the upper and lower bounds of this range were greater than the bounds for MSE of the XGBoost models trained for forecasting. The MRE range for these LightGBM models was wider than for the XGBoost and Random Forest models. However, the range was still relatively small and given the large standard deviations in the error metrics, no single feature subset or missing data threshold consistently had the highest performance.

The MSE for LightGBM models trained for forecasting purposes ranged from 4,773 to 9,156 when excluding models trained on the ‘Correlation 0.8’ feature subset (Figure 20b). This was similar to the MSE range for Random Forest models trained for forecasting, and lower than the range XGBoost models trained for the same purpose. The bounds were also smaller than the bounds on the MSE observed for LightGBM models trained for country-level prediction. The standard deviation of the MSE scores for LightGBM models trained for forecasting ranged from 546 to 1,336. This lower bound was higher than that for the XGBoost and Random Forest models trained for the same purpose, but lower than these models’ upper bounds. Despite this smaller range, LightGBM models trained on data with the ‘Correlation 0.6’ feature subset had MSE scores that were more consistently low than the models trained on other feature subsets, as observed for XGBoost and Random Forest models trained for the same purpose. There was no single missing data threshold that consistently produced the lowest MSE scores for these 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.



a)



b)

**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 Comparative Base Estimator 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, with the models trained for forecasting (PA) generally having higher predictive errors than models trained for country-level prediction (MDA). However, the analysis showed that all models trained on the ‘Correlation 0.8’ feature subset had higher predictive errors. It also showed that the hand-picked feature subset more consistently produced low error for MDA, and the ‘Correlation 0.6’ feature subset more consistently produced low error for PA.

In this section, I compared the Random Forest, XGBoost, and LightGBM models directly. While this section’s plots contained a lot of detail, the most salient information was the difference between the various models (plotted in different colours). As in Section 5.2, only MRE and MSE were shown. See Appendix 9.2 for the other metrics.

##### 5.331: Country-Level Prediction

One model type did not have consistently superior performance across all data pre-processing technique combinations, especially when considering the models’ overlapping standard deviations in their error metrics (see Figure 21, below).

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. However, the standard deviation range of the Random Forest models’ MRE often had a lower bound that was higher than that of the XGBoost models for datasets curated with no feature selection, the ‘Correlation 0.6’ feature subset, and the hand-picked feature subset. Therefore, while the Random Forest models had the lowest MRE when error was averaged across its cross-validation folds, some XGBoost models trained on specific folds had higher performance.

In contrast, the XGBoost models had the lowest MSE on datasets curated using no feature selection, the ‘Correlation 0.6’ feature subset, and the feature subset hand-picked from the literature. While the standard deviation on these lowest errors overlapped with the standard deviations of the other models on the same feature subsets, the lower bound on the XGBoost models’ MSE standard deviation was lower, indicating very high performance on specific cross-validation folds. Thus, XGBoost may have higher performance on outlier data. Generally, when XGBoost did not have the highest performance, the LightGBM and Random Forest models had performed similarly. While the LightGBM models had the highest MSE on datasets curated with no feature selection and the hand-picked feature subset, they rotated with the Random Forest models for the worst MSE performance on the other feature subsets.

A graph of different colored lines

AI-generated content may be incorrect.

a)

b)

A screenshot of a graph

AI-generated content may be incorrect.

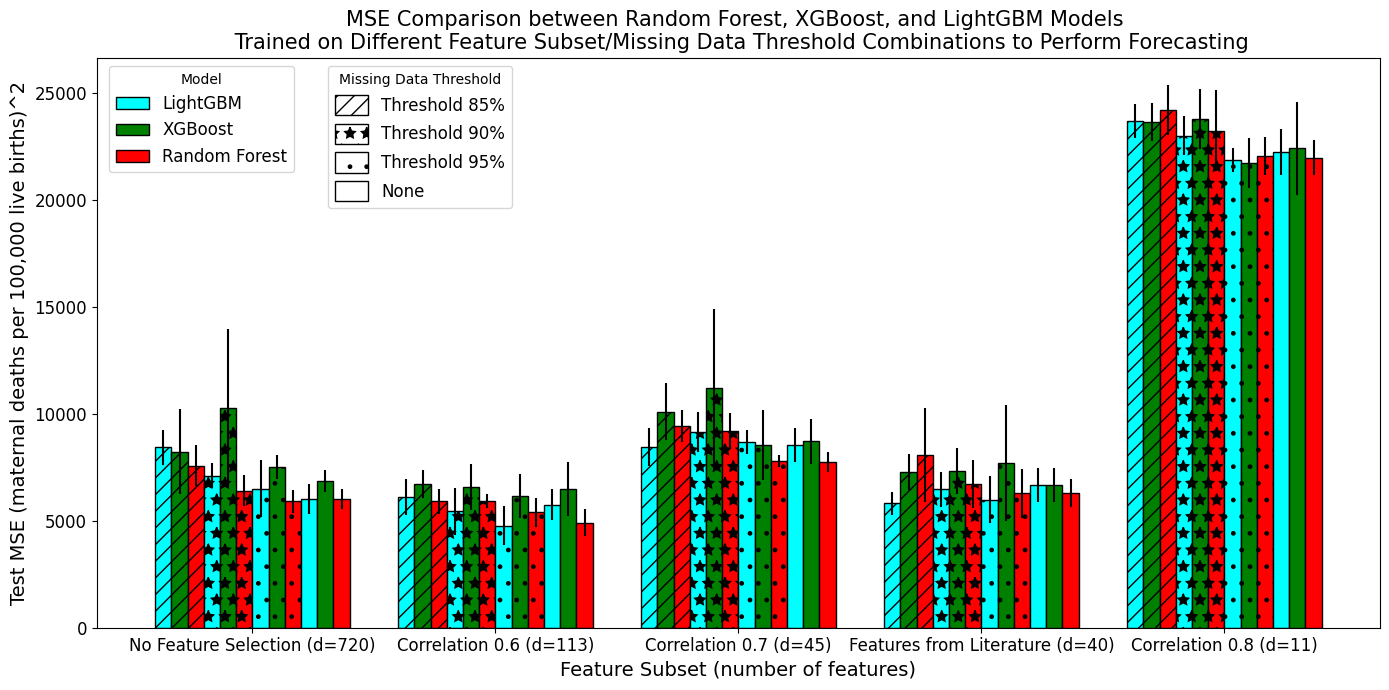
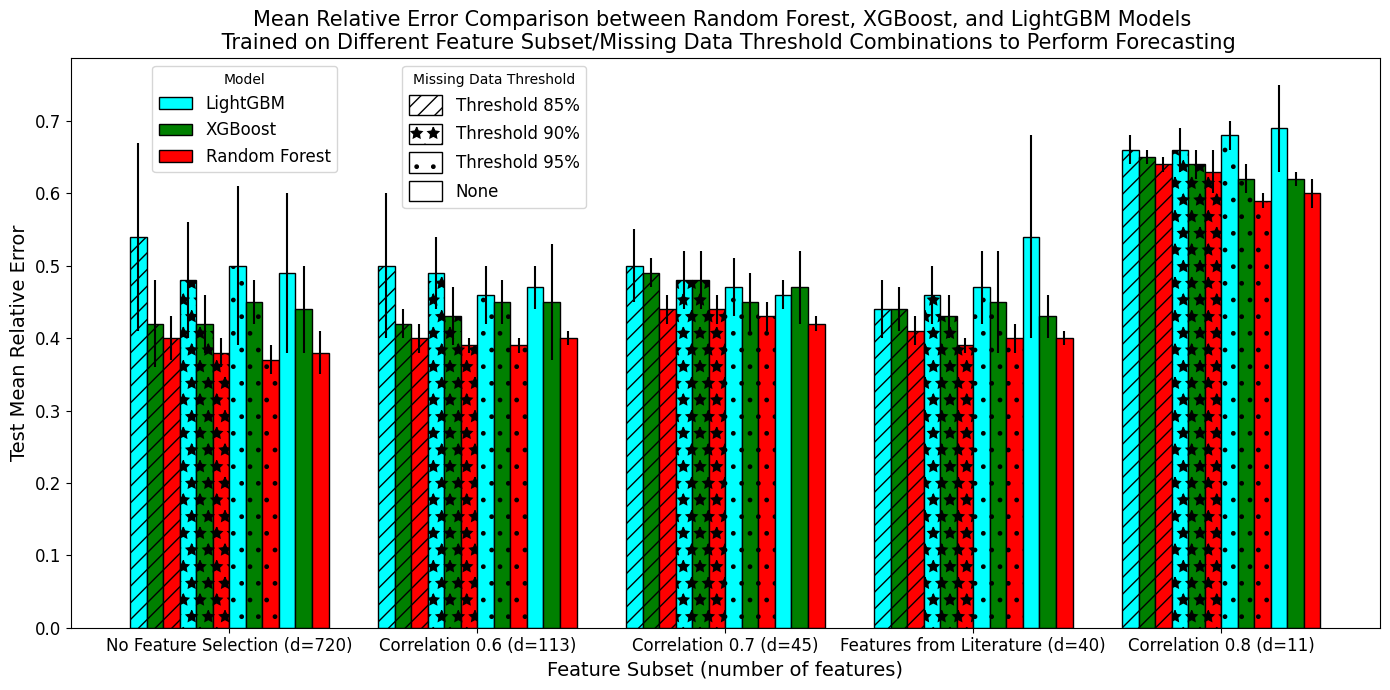
**Figure 21:** a) Mean relative error and b) mean-squared error for all base estimators fit on different feature subsets and missing data thresholds for country-level prediction. Random Forest models were represented with red, XGBoost with green, and LightGBM with light blue.

##### 5.332: Forecasting

Similar trends were observed for Random Forest, XGBoost, and LightGBM models trained for forecasting, with no model type consistently having the lowest error (see Figure 22).

LightGBM models often had the highest MRE, and Random Forest models often had the lowest. XGBoost models had the highest MRE values more rarely when trained for PA than MDA. The standard deviation for XGBoost models often covered a higher range of MRE values than the Random Forest models’ standard deviation, unlike the MDA models.

In contrast to when they were trained for country-level prediction, XGBoost models trained for forecasting had either the highest or second-highest MSE, with the former case occurring more consistently. This indicates that XGBoost models were more susceptible to outliers when trained for forecasting. The LightGBM and Random Forest models performed similarly, with strong overlap in their standard deviations.



a)

b)

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

### 5.4 Stacking and Voting Ensemble Results

Use of a stacking or voting ensemble model to combine the Random Forest, XGBoost, and LightGBM models presented in the previous section was motivated by the observation that no single model type consistently had the highest performance.

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

MRE and MSE were used to compare the voting and stacking ensembles’ performance. See Appendix 9.311 and 9.312 for their RMSE, MAE, and R2. Performance was measured according to Section 4.422. 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.

The voting and stacking ensembles’ MRE and MSE were usually lower when trained for country-level prediction (MDA) than for forecasting (PA) (see Figures 23 and 24). For example, the best-performing MDA and PA models achieved MRE of 0.07 and 0.37, respectively. The exception was the voting ensemble, whose MSE was roughly 650 points lower for PA than MDA. This may imply it was more affected by outliers when used for PA than MDA. Additionally, ensembles trained for PA had a smaller range of MRE and MSE values (0.37-0.56 and 5,100-8,000) than ensembles trained for MDA (0.07-0.33 and 2,150-7,100).

The SVM Stacking Ensemble always had the highest MRE and MSE scores. In contrast, the Random Forest Stacking Ensemble (SE) had the lowest mean relative error in both the MDA and PA (see Figure 23a and 24a). It also had the lowest MSE in PA (see Figure 23b and 24b). However, the Elastic Net SE had the lowest MSE for MDA. Given that the difference between the PA Random Forest and Elastic Net SE MSE scores was only approximately 260, a percentage difference of about 5%, **the Random Forest Stacking Ensemble was chosen as the best-performing ensemble** for consistency.

A graph with blue squares

AI-generated content may be incorrect.

a)

b)

A graph with blue squares

AI-generated content may be incorrect.

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

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

#### 5.42 Weighting Given to Each Base Estimator in the Stacking and Voting Ensembles Trained with All Base Estimators

To better understand the reasons behind the performance differences between the various ensembles, I explored which base estimators were weighted most heavily by each ensemble (see Section 4.424 for a description of the method). 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, as indicated on the plots below. 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 only placed importance on a subset of base estimators in both the MDA and PA (Figure 25). Primarily, it drew strength from the XGBoost base estimators, with a meaningfully smaller subset of LightGBM models used. The ensemble placed very little emphasis on Random Forest base estimators, with slightly more emphasis placed in the PA ensemble. A greater number of total base estimators were used when the stacking ensemble was used for PA than MDA (Figure 25b versus 25a).

Unlike the Random Forest Stacking Ensemble, 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 SE placed high importance on some Random Forest base estimators. However, like the Random Forest SE, the Elastic Net SE placed little 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 from all model types contributing little to the final prediction (Figure 27). The importance score distribution was similar for the voting ensembles trained for MDA and PA.

b)

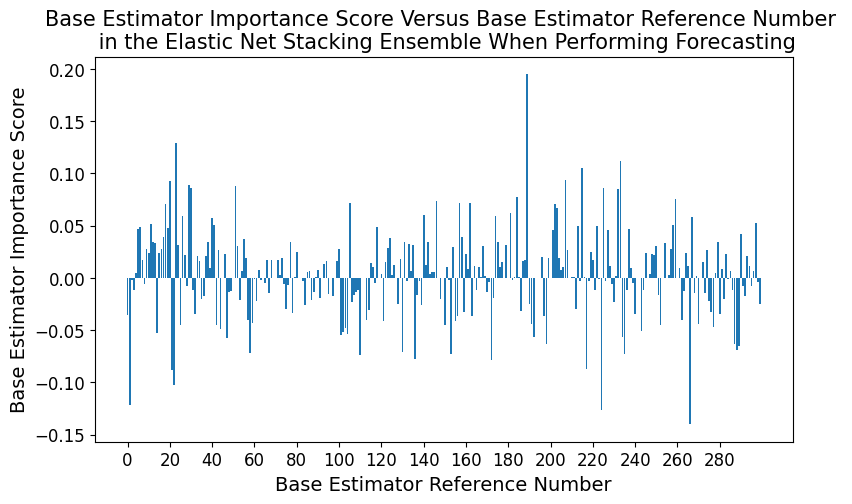
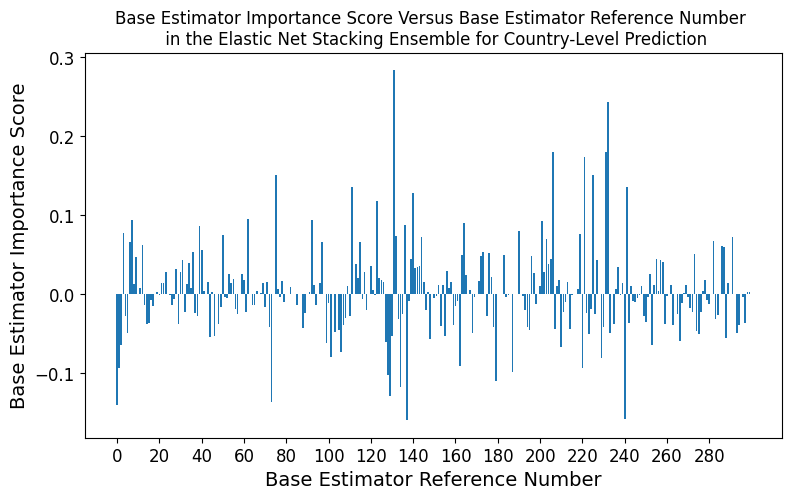
a)

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.



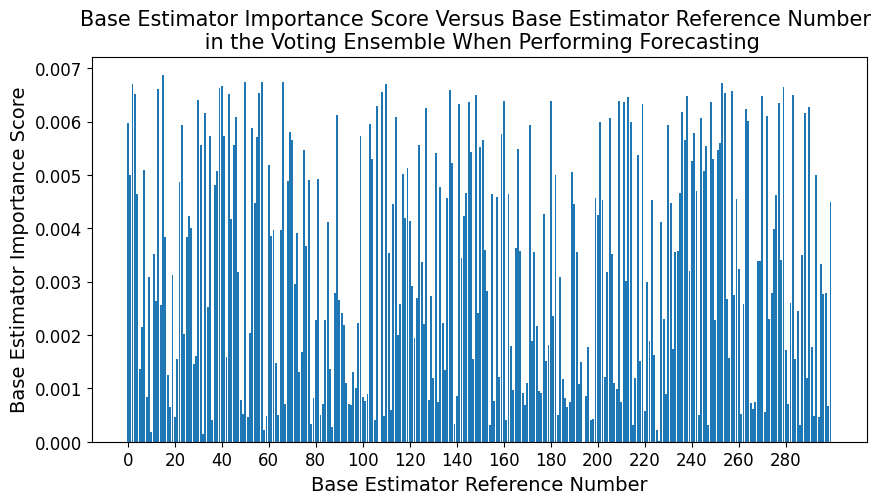
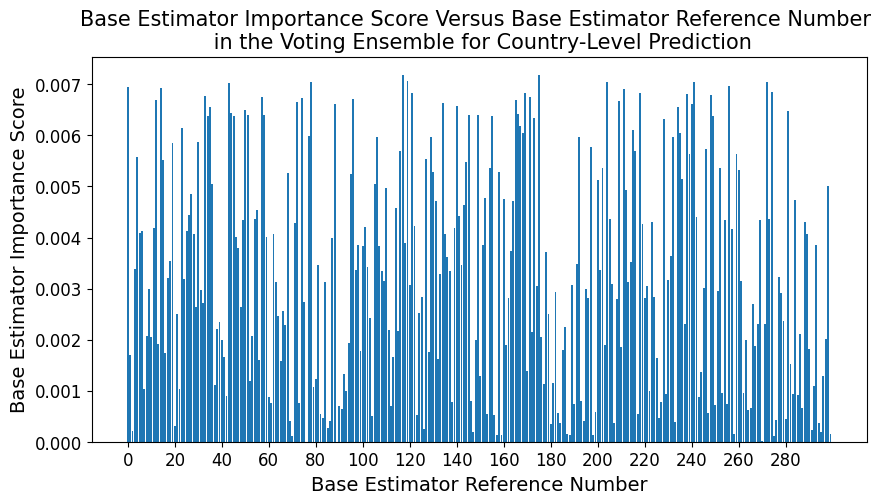
b)

a)

**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)



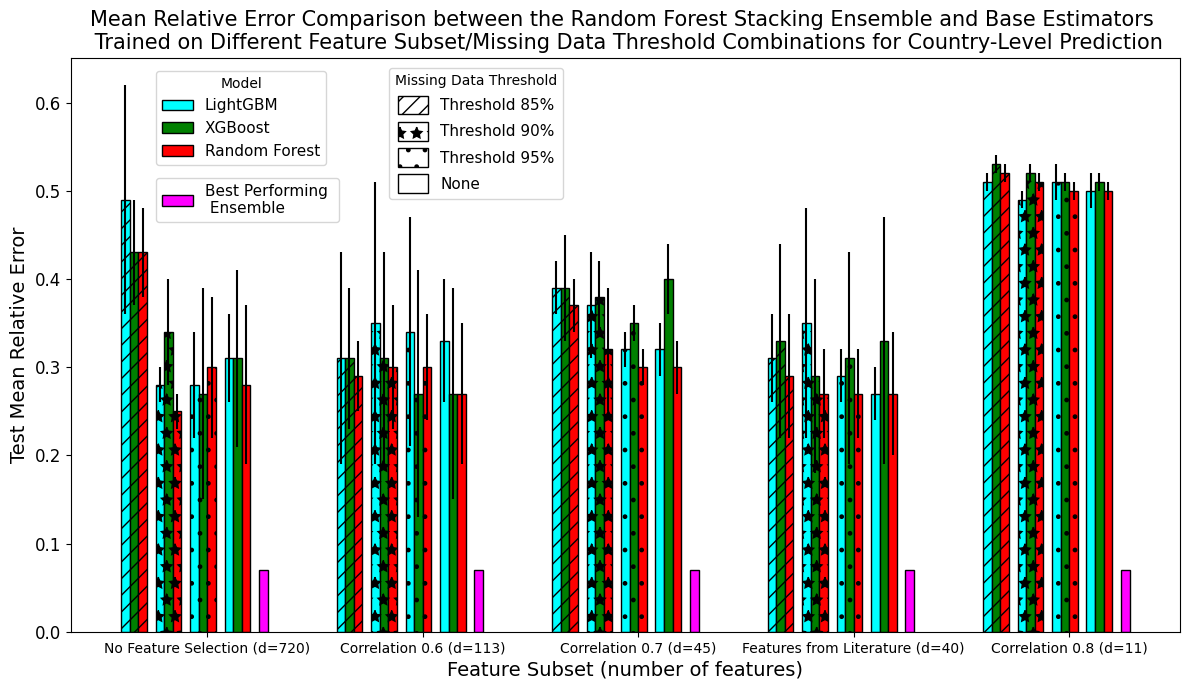
**Figure 27:** Importance score for each of the 300 base estimators used in the Voting Ensemble trained for a) country-level prediction and b) forecasting.

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

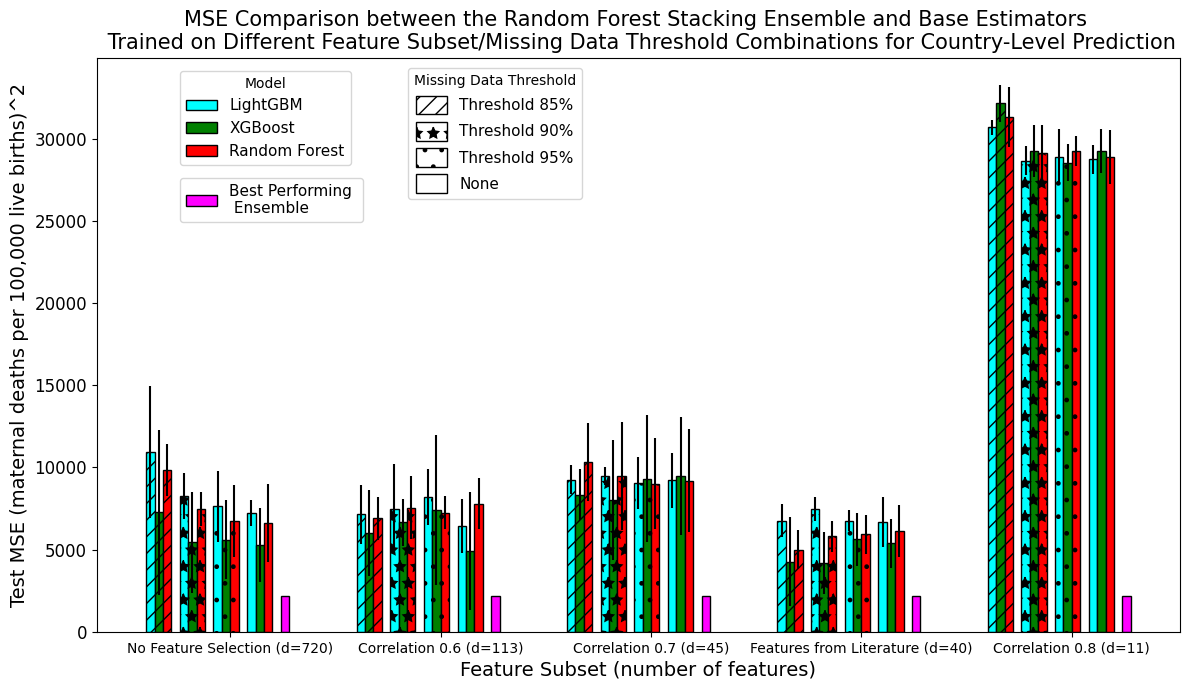
As described in section 5.41, the best performing ensemble was the Random Forest Stacking Ensemble (RFSE). Its predictive error was compared with that of single base estimators to establish whether stacking reduced error. While the following plots contained a lot of detail, the most important information 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

Figure 28 highlights how the Random Forest Stacking Ensemble greatly reduced both MSE and MRE for models trained for country-level prediction. 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)

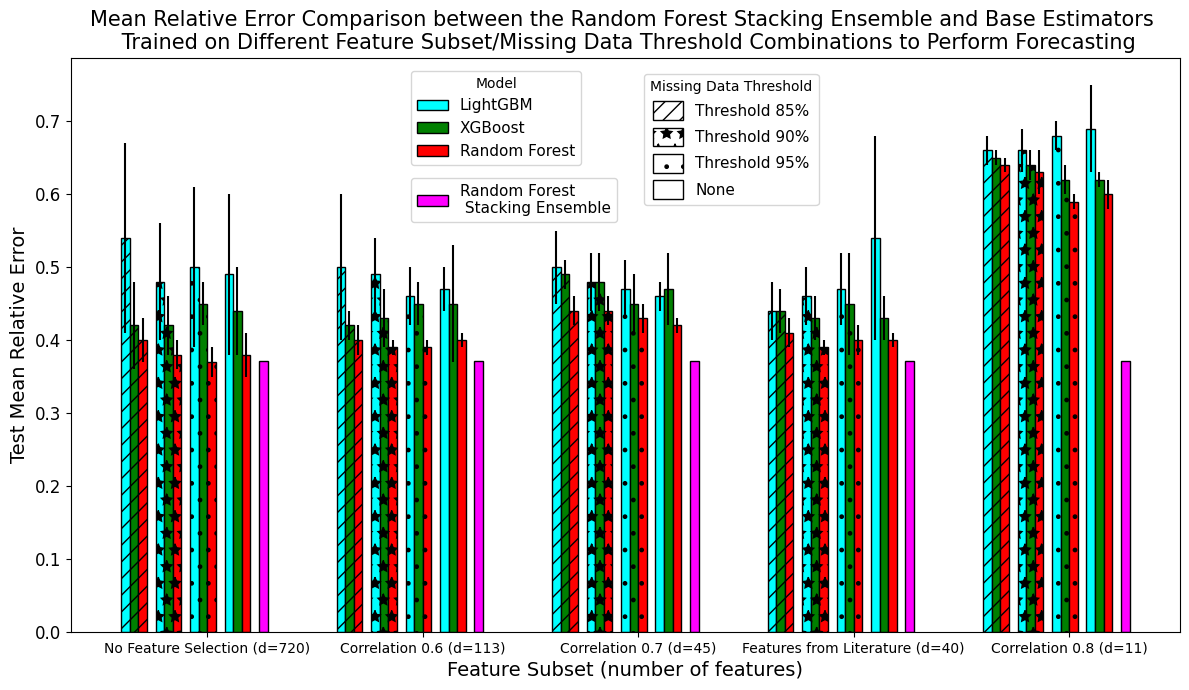


b)

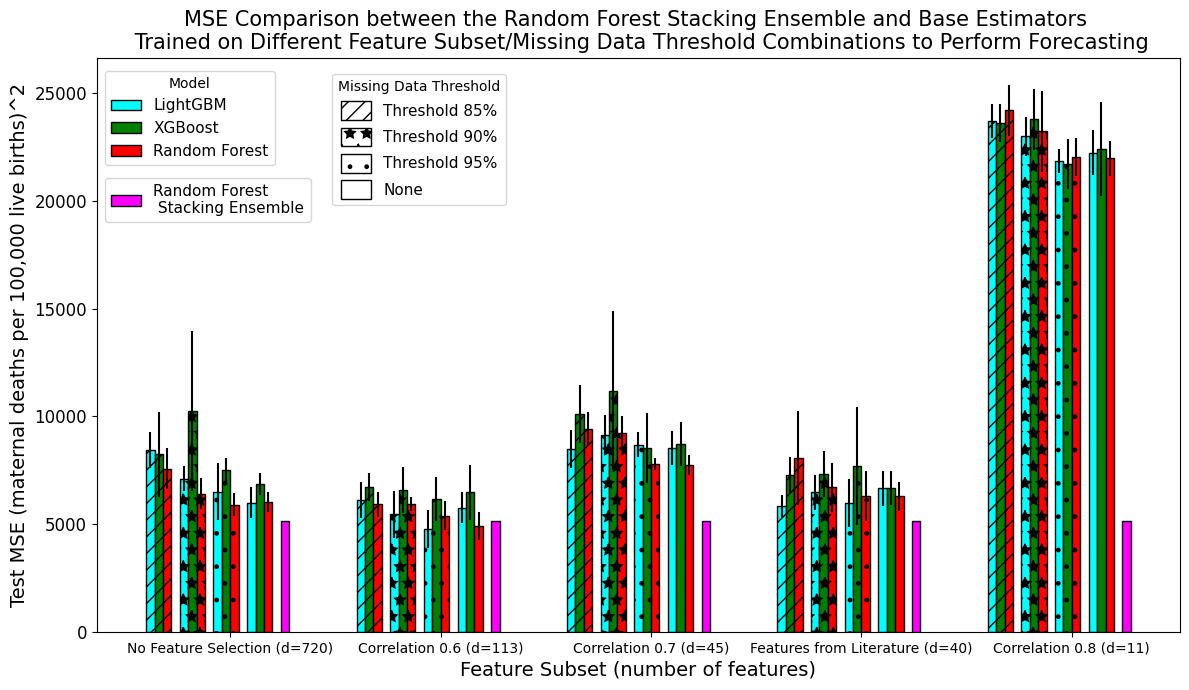
**Figure 28:** a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble and all base estimators trained for country-level prediction. Random Forest base models were represented with red, XGBoost with green, LightGBM with light blue. The Random Forest Stacking ensemble was shown with bright purple.

##### 5.432: Forecasting

In contrast to the Random Forest Stacking Ensemble (RFSE) trained for country-level prediction, the RFSE trained for forecasting does not always produce greatly lower error than the base estimators (see 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%. Its 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%.



a)



b)

**Figure 29:** a) Mean relative error and b) mean-squared error for the Random Forest Stacking Ensemble and all base estimators trained for forecasting. Random Forest base models were represented with red, XGBoost with green, LightGBM with light blue. The Random Forest Stacking ensemble was shown with bright purple.

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

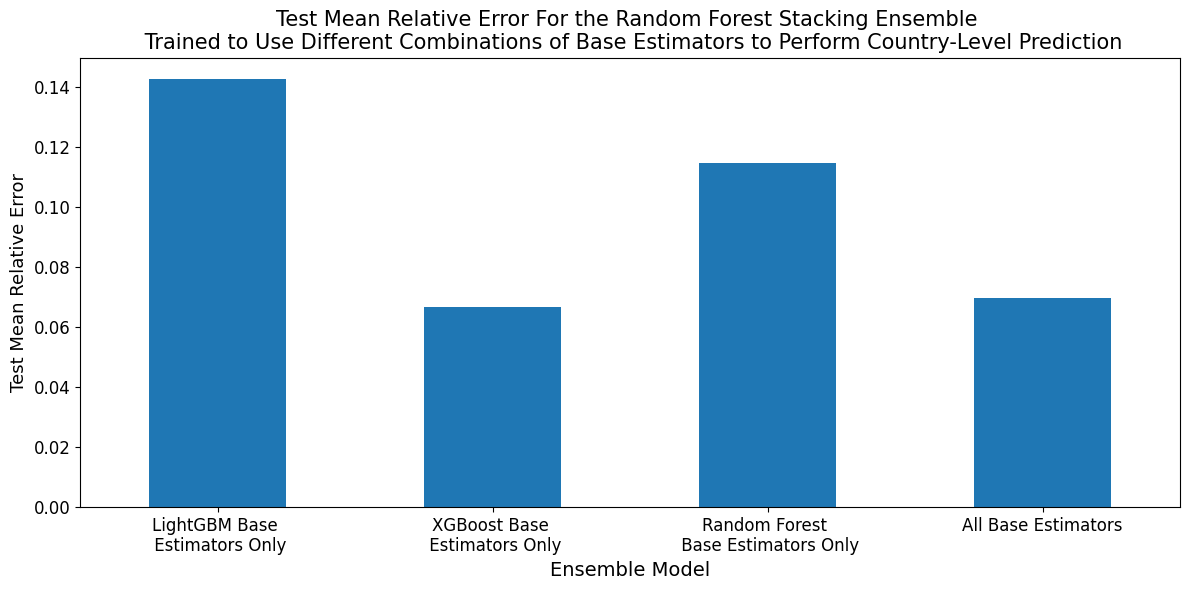
Compared to the best base estimators’ performance, the Random Forest Stacking Ensemble (RFSE) halved the MSE and reduced the MRE by over a factor of 4 when trained for country-level prediction. In contrast, compared to the existing base estimators, the RFSE produced no change in the MRE and increased the MSE by approximately 350 when trained for forecasting. Despite this, the Random Forest Stacking Ensemble was considered the ‘best performing model’ produced through this project. This determination was made due to the substantial improvement generated through using the RFSE for country-level prediction. It was used for forecasting as well for consistency and because the RFSE did not increase MRE or raise MSE by a notable amount.

In this section, 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

Using a different combination of base estimators as the input to the RFSE did not greatly improve performance of either country-level prediction or forecasting (Figures 30 & 31). When training the RFSE for country-level prediction, using only predictions from XGBoost models as input reduced MRE by roughly 0.3% (from 6.95% to 6.67%) and decreased MSE by 87 (from 2,161 to 2,074). When training the RFSE for forecasting, using only Random Forest base estimators as input into the RFSE reduced MRE by about 0.19% (from 37.08% to 36.89%). Using only LightGBM base estimators as input into the RFSE 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 combination as input to the RFSE used 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 the 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.



a)

A graph of blue rectangular bars

AI-generated content may be incorrect.

b)

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

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 on different combinations of base estimators to perform forecasting.

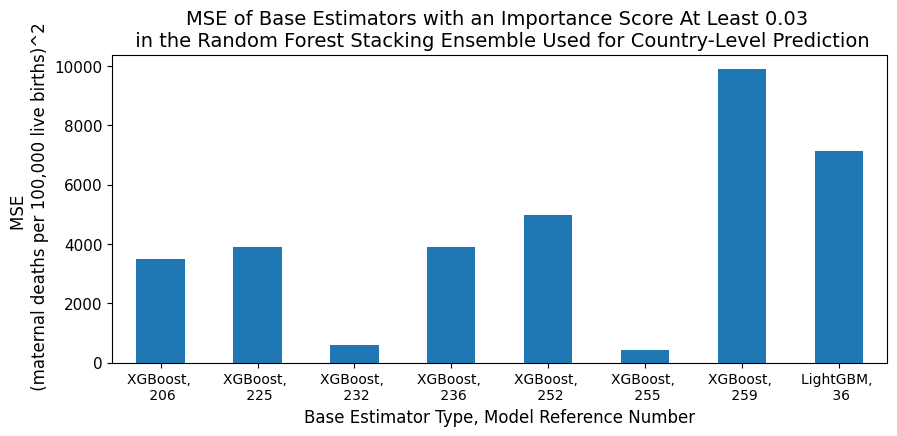
#### 5.52 Base Estimator Importance Analysis in the Random Forest Stacking Ensemble

In Section 5.42, I discuss 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 were chosen by the RFSE using the procedure outlined in Section 4.425.

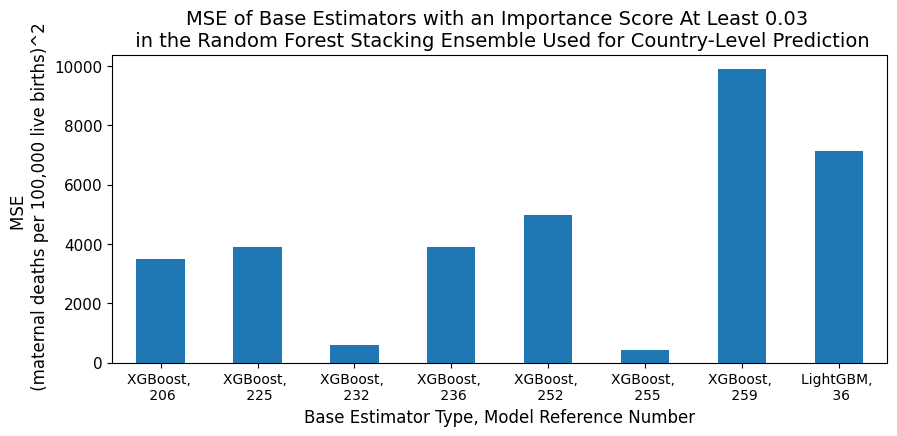
##### 5.521 Differences in Predictive Performance of Most Important Base Estimators

I first determined whether mean predictive error over the test set was the sole predictor of the 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 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 qualifying base estimators for forecasting, 7 were XGBoost 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 used for a) country-level prediction and b) forecasting. Each base estimator was identified with its model type and reference number used in Section 5.43, which specifies its ordering in the RFSE’s input data.

##### 5.522 Effect of Permutating the Order of Base Estimators Input into the Random Forest Stacking Ensemble

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 determined how useful it was to create the splits using predictions from base estimators located at default positions in its input data. If none of the base estimators it subsequently trialled could produce a split with a lower predictive error, it would remain with the default 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.

Nine of the ten MDA base estimators with importance scores at least 0.03 in the original order 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 due to the permutation. The largest change was in the model given the highest importance, which lost 0.06 importance points after permutation. The MDA 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). However, the weighting given to each base estimator may be unstable and subject to change through retraining.

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, 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 models’ importance scores increasing by 26 to 30 points. The PA RFSE’s predictive accuracy decreased after permutation, with its MSE increasing by roughly 860 points. These changes showed that the PA 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 their predictions were 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, it was 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 chose to place 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 than can learn when and how to use them. While the permutation analysis showed the MDA base estimators were not chosen at random, it did show the instability of the importance scores and the effect of retraining on how different base estimators were valued. This was shown more explicitly for the PA models. 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 it also presented the features with the highest importance for 2 base estimators with low weightings in the RFSE. One of these estimators had a high MSE score.

##### 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 analogues were also highly valued.

While the two base estimators with higher errors and lower importance scores in the RFSE also placed variables monitoring 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 still-birth rates, 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 health trends.

**Table 13:** The 5 most important features in the two base estimators with the highest importance scores in the Random Forest Stacking Ensemble (light blue), a medium-low accuracy base estimator from the ‘Correlation 0.7’ feature subset (light orange), and a low-accuracy base estimator from the ‘Correlation 0.8’ feature subset (light purple). These models were all 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 most important features in the two base estimators with the highest importance scores in the Random Forest Stacking Ensemble (light blue), a medium-low accuracy base estimator from the ‘Correlation 0.7’ feature subset (light orange), and a low-accuracy base estimator from the ‘Correlation 0.8’ feature subset (light purple). These models were all 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 fine-grained understanding of the best-performing Random Forest Stacking Ensemble and its architecture, this section explores its performance.

#### 5.61 Random Forest Stacking Ensemble’s Predictive Error on Data from Each 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 RSE’s mean relative error (MRE) on the test set decreased as income level increased (Figure 33a). For example, the test MRE for low-income countries was 0.18, compared to a test MRE of 0.07 for high-income countries. However, the lowest test error was observed when estimating MMRs for lower-middle income countries (0.02). Generally, validation MRE was similar to the train MRE and smaller than the test MRE for all income levels, with the exception again being the lower-middle subgroup. The difference between the train/validation MREs and the test MRE was greatest for low-income countries.

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

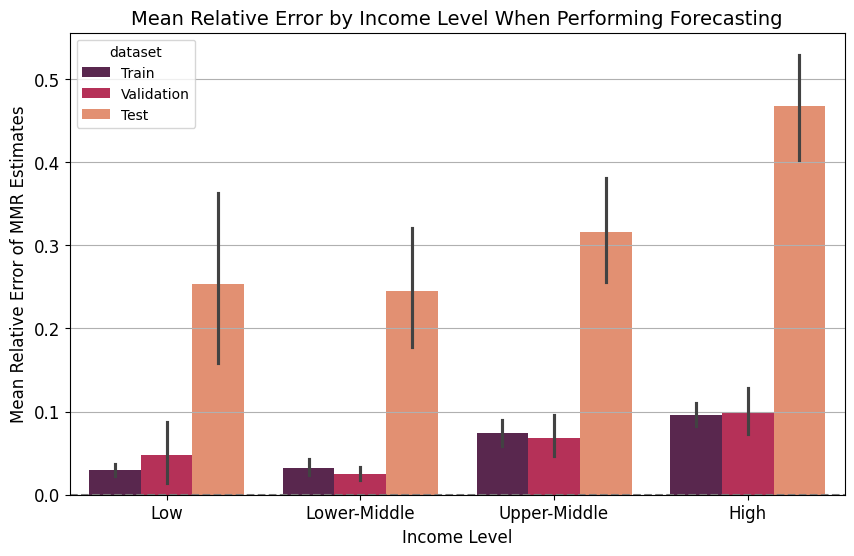
b)

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

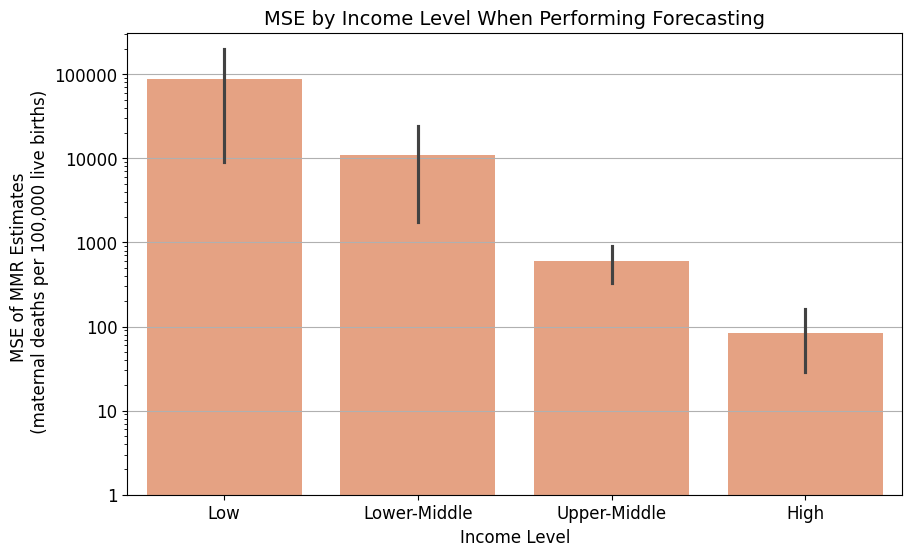
##### 5.612: Forecasting

The mean relative error of the Random Forest Stacking Ensemble trained to forecast MMR increased as income level increased from lower middle to high. (Figure 34a). While the RFSE had an MMR of 0.25 for lower-middle income countries, it had an MRE of 0.47 for high-income countries. Contrary 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 large differences between the cross-validation folds. Generally, train and validation error 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 scores 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, compared to a test MRE of 88,585 for low-income countries.



a)



b)

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

#### 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 the ensemble’s base estimators. The smaller the standard deviation, the greater the agreement, and thus the more certainty the stacking ensemble may have 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, base estimator predictions for ground truth MMR estimates between 300 and 1,050 ranged from 50 to 350. For the extremely high ground MMR of 1,763, standard deviation among base estimators was 441.

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

However, these findings must be qualified by the statement that there were few datapoints for ground truth MMR values greater than 1,050 for country-level prediction and greater than 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.

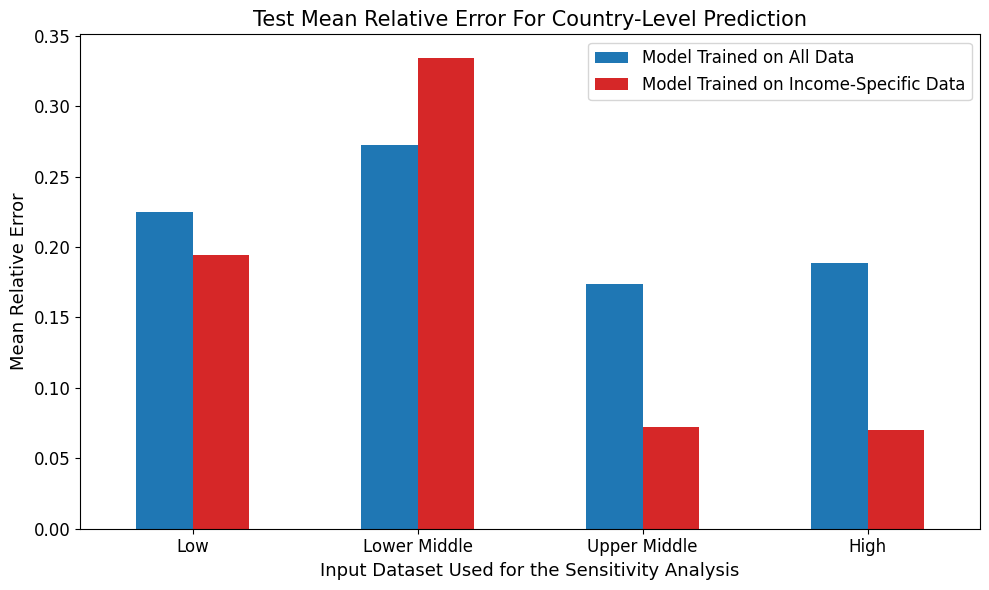
#### 5.63 Sensitivity Analysis

To how the input data structure affects final MMR predictions, I conducted a sensitivity analysis (see Section 4.5 for the method). RFSEs trained on data from a single income-level were referred to as ‘sensitivity models’, while the RFSE trained on data from all income levels was referred to as the ‘original model’.

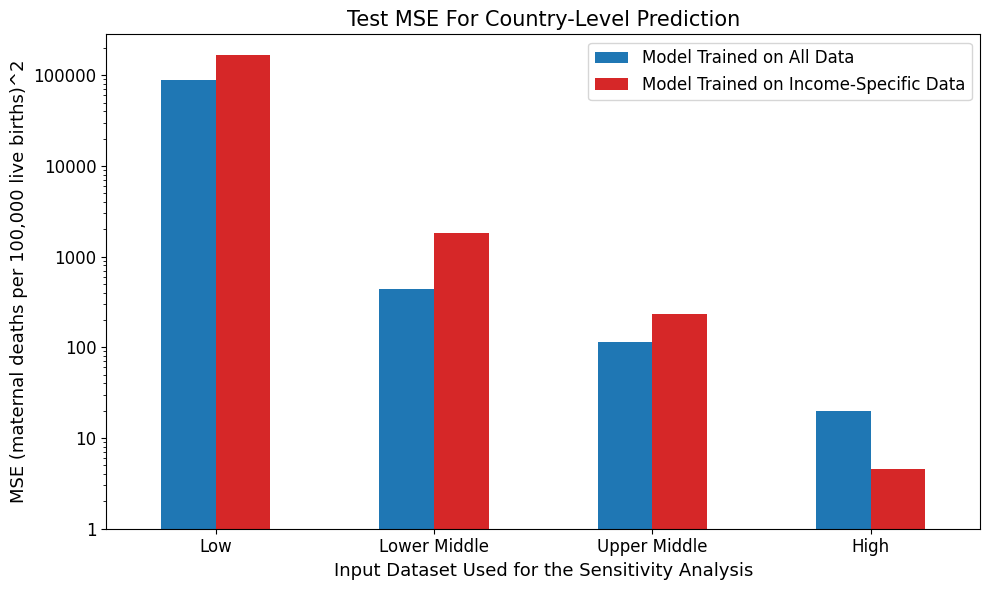
##### 5.631 Country-Level Prediction

Generally, the mean relative error (MRE) for the Random Forest Stacking Ensemble trained and tested on data from all income levels was larger than the MRE for RFSEs trained and tested 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 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 when it was applied to the low-income dataset, where the RFSE trained on all data had an MRE of 0.22 compared to an MRE of 0.19 for the RFSE trained on only the low-income dataset. In contrast to this trend, the MRE for the RFSE trained on all data (0.27) was smaller than the MRE of the RFSE trained on just lower-middle income data (0.33).

In contrast, only the RFSE trained on with solely high-income data had a lower MSE score than the model trained on all data (5 versus 20) (Figure 36b). The RFSE trained on all data had a lower MSE score than the RFSE trained on income-level specific data for all other income levels. For example, the model trained on just low-income data had an MSE of 167,123 compared to the model trained on all data, which 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 versus MRE indicates the presence of outliers in the income-specific datasets.



a)



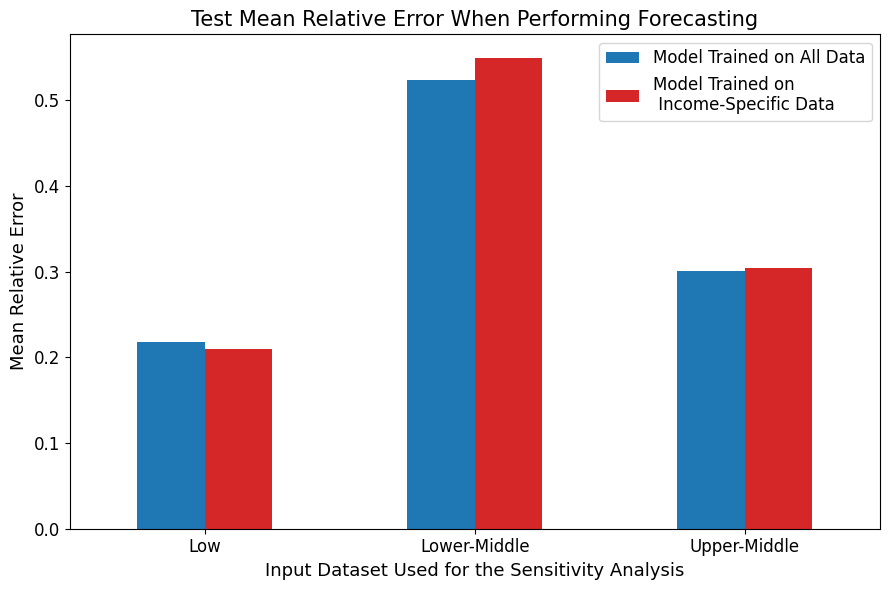
b)

**Figure 36:** 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 RFSE’s trained on data from a specific income level (red). Both models were data from a single income level. These models were trained to perform country-level prediction of MMR estimates.

##### 5.632 Forecasting

Unfortunately, some of the 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). Generally, the low-income sensitivity model had lower error than the original model (MRE 0.21 versus 0.22 and MSE 58,701 versus 73,343) while the lower-middle income sensitivity model had higher error than the original (MRE 0.55 versus 0.52 and MSE 4,044 versus 2,936). The sensitivity model for upper-middle income countries had the same MRE as the original model 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 RFSE’s trained on data from a specific income level (red). Both models were data from a single income level. These models were trained to forecast MMR.

### 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 the aggregate difference between my MMR estimates and the literature’s estimates across all countries.

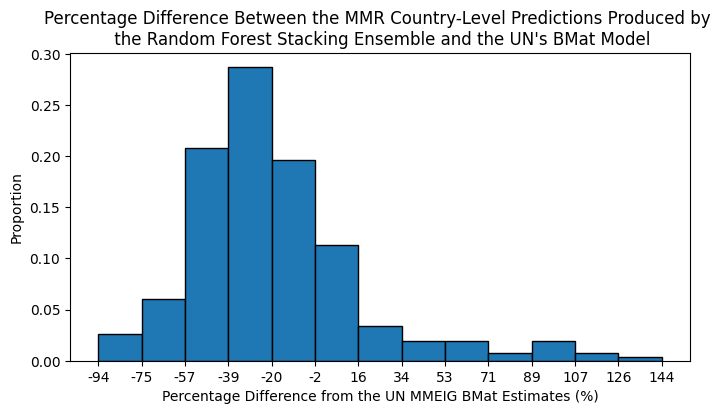
##### 5.711 Percentage Difference

As described in Section 4.6, I took the percentage difference between my MMR estimate for each country, year datapoint 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

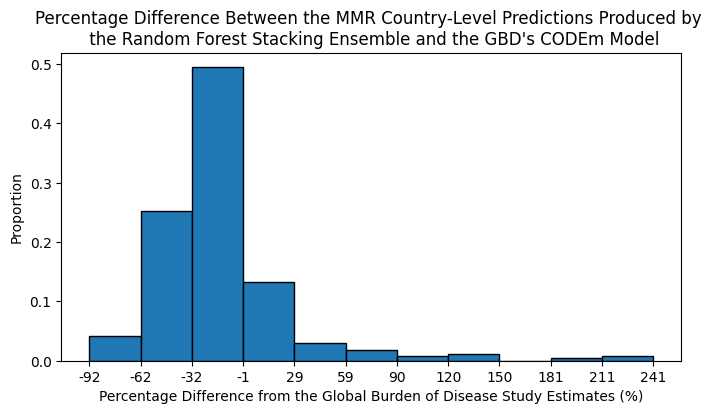
Over 70% of the MMR estimates from my best-performing RFSE were smaller than the corresponding estimates from the UN’s BMat model (Figure 38a) compared to approximately 75% smaller than the associated CODEm estimates (Figure 38b) and roughly 80% smaller than the corresponding GMatH microsimulation estimates (Figure 38c). While over 40% of my MMR predictions were between 75% and 100% smaller than the corresponding GMatH estimates, while less than 5% of my MMR estimates were over 75% smaller than the UN’s BMat estimates or GBD CODEm predictions. In contrast to the GMatH model, approximately 50% of my MMR estimates were between 0 and 39% or 0 and 32% smaller than the associated estimates from the UN Bmat and GBD CODEm models. Therefore, the magnitude difference between my MMR estimates and the GMatH predictions were generally larger than the magnitude differences to the other two models.

Approximately 15 to 20% of my MMR predictions were larger than the associated BMat and CODEm predictions 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)

b)

A graph of a number of blue bars

AI-generated content may be incorrect.

c)

**Figure 38:** Percentage difference between the MMR estimates produced by my best-performing Random Forest Stacking Ensemble trained for country-level prediction and the MMR estimates produced 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% of my MMR forecasts were smaller than the associated BMat estimates (Figure 39a) and almost 60% of my MMR forecasts were smaller than the corresponding GBD CODEm predictions (Figure 39b). In contrast, over 80% of my MMR estimates were smaller than the associated GMatH microsimulation estimates (Figure 39c). Over 40% were between 0 and 50% smaller than both the BMat and CODEm model 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 greater 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 the MMR estimates produced by my best-performing Random Forest Stacking Ensemble trained to perform forecasting and the MMR estimates produced by a) the UN MMEIG’s BMat model b) the GBD’s CODem model, and c) the GMatH microsimulation model.

##### 5.712 Coverage

A greater proportion of my best-performing RFSE’s country-level MMR predictions were within the 95% confidence intervals (CI) of the GMatH model’s MMR predictions than were within the 95% CI of the BMat and CODEm models (Table 15). More specifically, approximately 22.3%, 29.4%, and 67.1% of my model’s country-level MMR predictions were within the 95% confidence intervals of the BMat, CODEm, and GMatH models, respectively. Similarly, only 20.4%, 33.2%, and 67.1% of the ground truth MMR estimates that my RFSE was trained on were within these models’ 95% confidence intervals.

A higher proportion of my best-performing RFSE’s MMR forecasts were within the 95% CI of the BMat and GMatH model predictions than observed for 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 (light blue), CODEm (light green), and GMatH (light purple) models’ MMR predictions. The proportion of ground truth MMR estimates used to train my model that fell within these CI were 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 gain a more fine-grained understanding of 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 corresponding BMat, CODEm, and GMatH estimates.

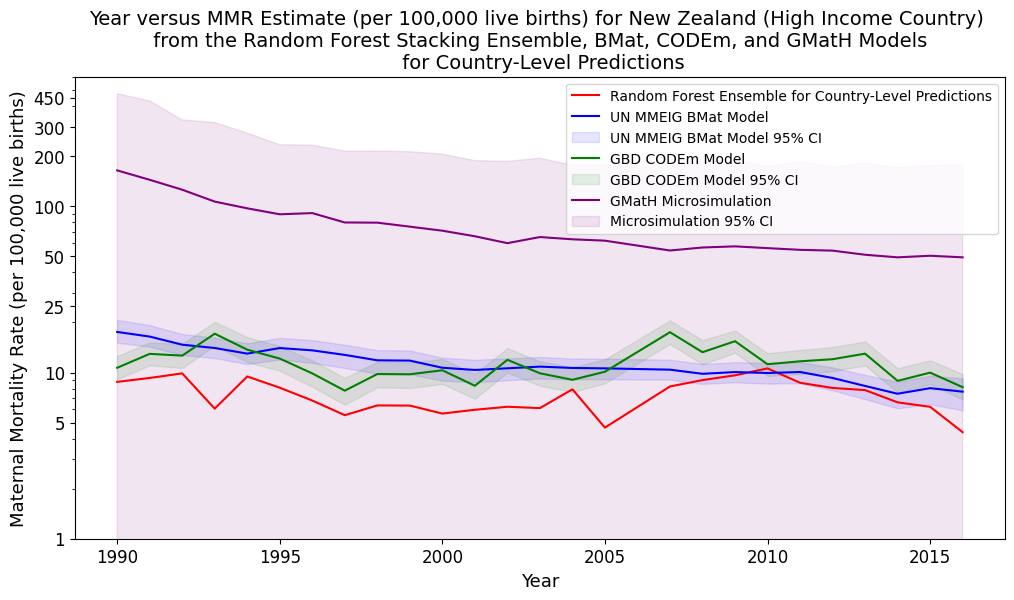
##### 5.721 Country-Level Prediction

Compared to the other models, my Random Forest Stacking Ensemble underpredicted MMR for New Zealand between 1990 and 2015, with greater underprediction pre-2010 (Figure 40a). While my model’s estimates were outside the 95% confidence intervals of the BMat and CODEm models, the actual difference in MMR between the estimates was between 5 and 20. The GMatH microsimulation model strongly overestimated New Zealand’s MMR, predicting MMR to be close to 200 in 1990 and fall to close to 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% confidence interval that enveloped my model’s MMR estimates.

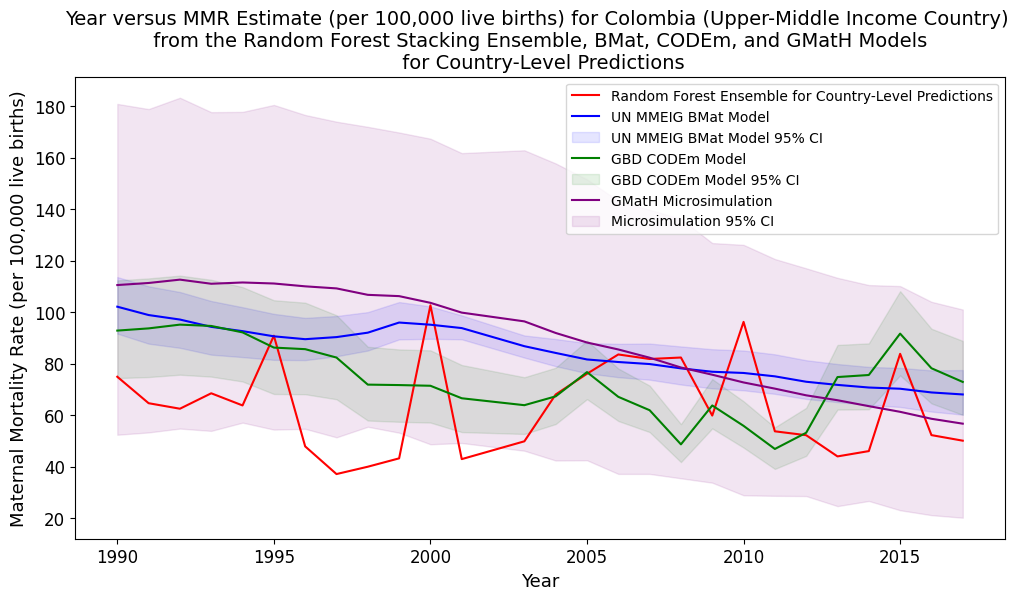
In contrast, my model’s estimates had greater intersection with the literature’s estimates for Colombia, an upper-middle income country (Figure 40b). My model’s estimates were generally 0 to 40 points off the closest literature estimate. They were generally within the GMatH model’s wide 95% confidence interval (CI), 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 model predictions. However, my model’s estimates fluctuated more strongly between consecutive years, compared to the smoother literature estimates. With most of the peaks corresponding 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 4 datapoints of each of Kenya and Rwanda over this period of time.

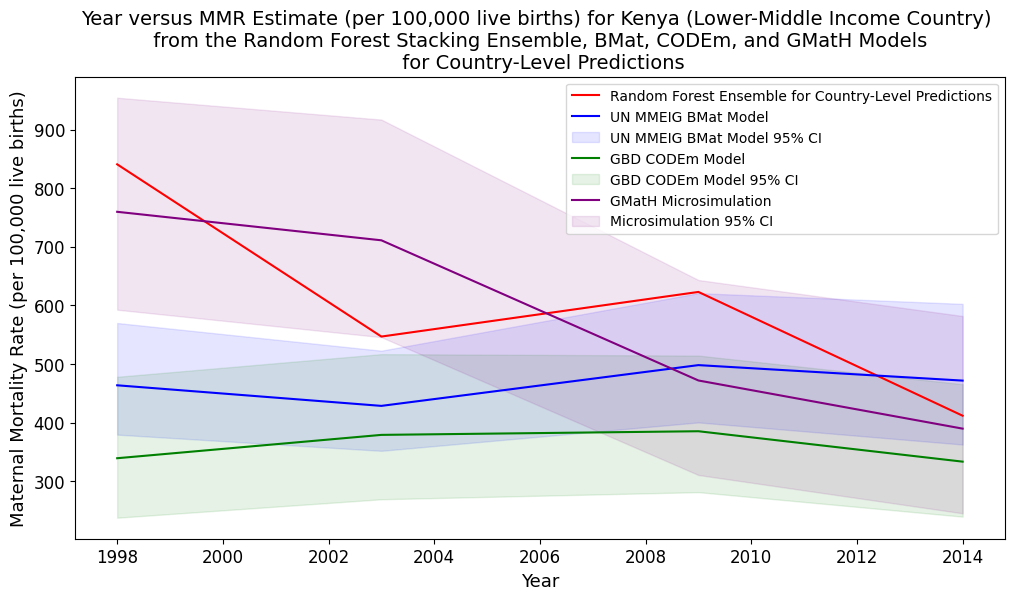
Therefore, it appeared that my model’s country-level predictions were underestimates for higher-income countries but over-estimates for lower-income countries.



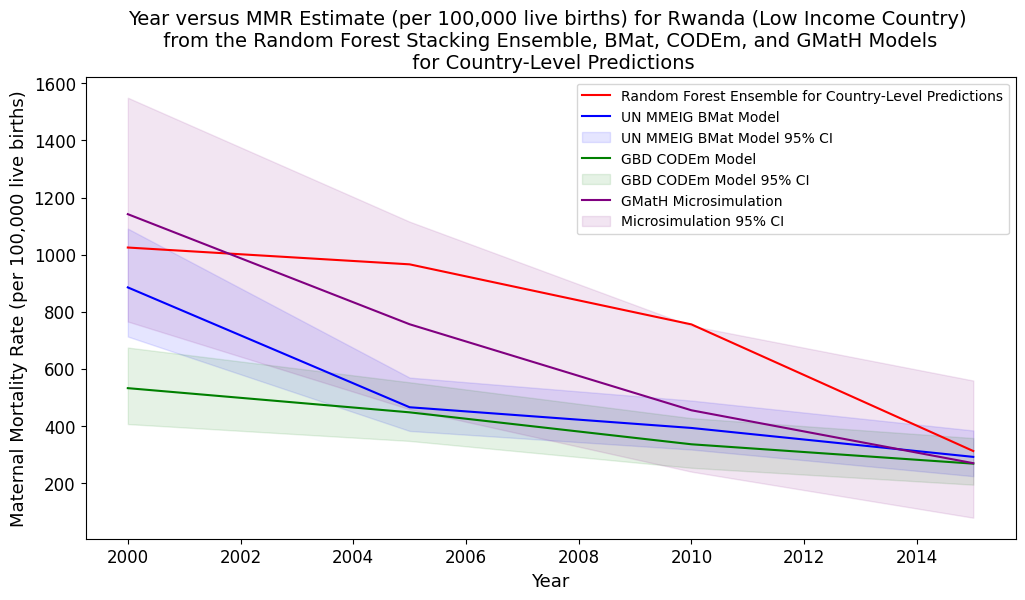
a)

****

b)



c)



d)

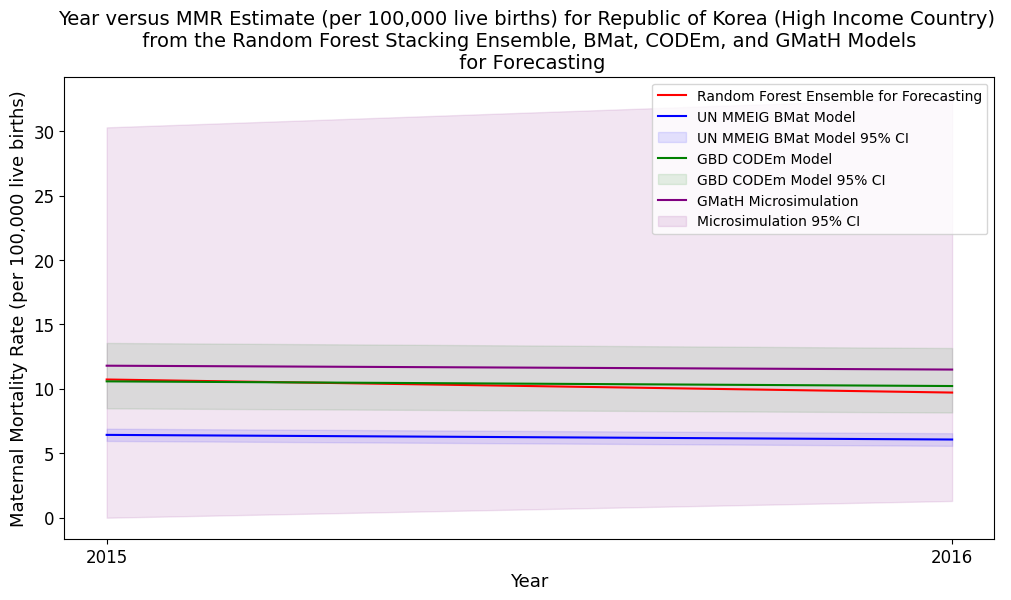
**Figure 40:** Comparison of my best-performing Random Forest Stacking Ensemble’s MMR country-level predictions to the associated estimates from the BMat, CODEm, and GMatH models for one country from each World Bank defined income level. Estimates were presented for a) New Zealand (high-income country)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

##### 5.722 Forecasting

As a whole, there was much 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. The lack of comparison points was underlined by the fact that not all samples in my test dataset had non-missing MMR values, meaning that 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 the 95% CI of the GMatH model (Figure 41d). Unfortunately, there was only one test datapoint available for Chad, and every other low-income country represented in this 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.



b)

a)

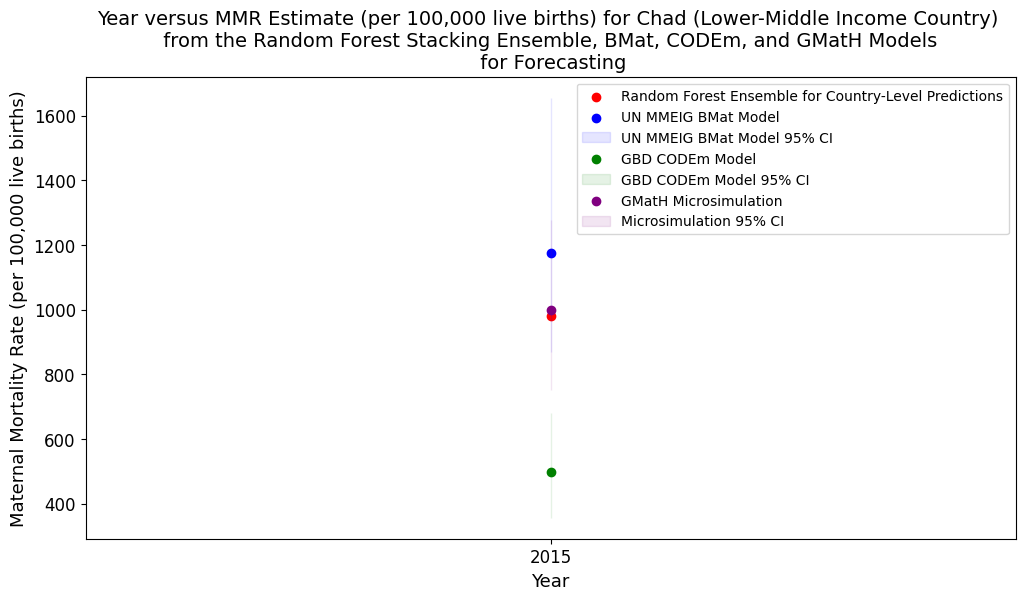
A graph with lines and text

AI-generated content may be incorrect.

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. Estimates were presented for a) the Republic of Korea (high-income country), b) Armenia (upper-middle income country), c) Sri Lanka (lower-middle income country), and d) Chad (low-income country).