Spatio-Temporal Machine Learning for Economic Forecasting: A Cross-Indicator and Cross-Country Analysis Using World Bank Data

Jiadong Zhang*

May 2025

Abstract

This paper develops a spatio-temporal framework for forecasting key macroeconomic indicators using machine learning techniques. Leveraging World Bank panel data from 1960 to 2020, we investigate three core tasks: (1) cross-indicator prediction—forecasting one major indicator based on others; (2) national-level time series forecasting; and (3) cross-national forecasting for major economies. Through comparative analysis of models including XGBoost, Random Forest, and linear baselines, we find that most structural indicators are highly predictable, while GDP growth remains volatile and difficult to model. We introduce a novel feasibility rate metric to assess reliability across multiple performance dimensions. Our results highlight both the promise and the limitations of data-driven methods in economic forecasting and underscore the importance of model selection, data context, and indicator stability.

*Email: dereksodo@gmail.com

Contents

1 Introduction

Macroeconomic indicators provide a comprehensive lens for assessing the economic and social development of countries. This study leverages World Bank panel data from 1960 to 2020 to identify key national indicators and evaluate their predictability using machine learning. The focus is on cross-indicator prediction: can one major indicator be reliably predicted from the others?

2 Literature Review

Recent advancements in machine learning (ML) have significantly influenced macroeconomic forecasting. This section reviews key studies that have explored the integration of ML techniques into economic prediction models.

2.1 Nonlinearity and Regularization in ML Forecasting

Goulet Coulombe et al. (2019) investigate the efficacy of ML in macroeconomic forecasting, emphasizing the importance of capturing nonlinear relationships in economic data. Their study concludes that nonlinearity is a crucial factor in improving forecast accuracy. They also highlight that traditional factor models serve as effective regularization tools within ML frameworks, aiding in managing model complexity and preventing overfitting. The authors advocate for the use of K-fold cross-validation as a best practice for model evaluation and selection. Their findings suggest that ML models, when properly regularized and validated, can outperform traditional econometric models, especially during periods of economic uncertainty and financial stress [1].

2.2 Automating Forecasting with ML Techniques

Hall (2018) explores the application of ML methods to macroeconomic forecasting, focusing on the automation of model selection and parameter tuning. The study demonstrates that ML algorithms can process vast and complex datasets, identifying patterns that traditional models might overlook. Hall's analysis reveals that ML models can outperform both simple time-series models and consensus forecasts from professional economists, particularly in predicting short-term economic indicators like the unemployment rate. The research underscores the potential of ML to enhance forecasting accuracy by reducing reliance on manual model specification and expert judgment [2].

2.3 ML Applications in China's GDP Forecasting

Yang et al. (2024) apply various ML models to forecast China's quarterly real GDP growth, assessing their performance against traditional econometric models and expert forecasts.

Their study finds that ML models generally achieve lower forecast errors, particularly during stable economic periods. However, during economic inflection points, expert forecasts may exhibit greater accuracy due to a more nuanced understanding of the macroeconomic environment. Additionally, the authors employ interpretable ML techniques to identify key variables influencing GDP fluctuations, providing insights into the underlying drivers of economic change [3].

2.4 Synthesis and Implications for Current Research

The reviewed studies collectively highlight the transformative impact of ML on macroeconomic forecasting. They demonstrate that ML models, with their ability to capture complex nonlinear relationships and process large datasets, can enhance forecast accuracy beyond traditional methods. These findings inform the current research by underscoring the importance of incorporating ML techniques into economic prediction models, particularly for analyzing cross-indicator relationships, time series data, and cross-country economic dynamics. While the empirical studies reviewed above emphasize the technical advances of ML-based forecasting, it is equally important to align these findings with economic theory. For instance, the persistent unpredictability of GDP growth observed in both past studies and our own results echoes theoretical insights from the Solow growth model [4], which attributes long-term economic growth primarily to exogenous technological progress—a factor that is inherently difficult to observe or predict. Likewise, indicators such as life expectancy and energy consumption—which consistently achieve low RMSE/STD and MASE and high R^2 and DA—reflect long-term structural trends that are more stable and easier to model. The declining predictive value of agricultural output aligns with structural transformation theory, which explains the shift of economic activity from agriculture to industry and services as economies develop. By framing ML findings within established theoretical paradigms, this study highlights not only algorithmic performance but also its macroeconomic interpretability. [4, 5]

3 Data and Methods

3.1 Data Source

- World Bank Open Data, 1960–2020, including G20 expect African Union.
- Main dataset: [World Bank Data by Indicators](https://github.com/light-and-salt/World-Bank-Data-by-Indicators) (GitHub repository)
- We selected features with more than 60% of relevant data present to minimize interpolation errors, resulting in a subset of 13 indicators. From these, we identified 10 relatively independent features for modeling, denoted $\{F_1, F_2, \ldots, F_{10}\}$.[1]

[1] Check Table 1 for details.

Indicator Code	Indicator Name
SP.DYN.LE00.IN	Life expectancy at birth, total (years)
SP.URB.TOTL.IN.ZS	Urban population (% of total population)
NV.AGR.TOTL.ZS	Agriculture, forestry, and fishing, value added (% of GDP)
EG.USE.PCAP.KG.OE	Energy use (kg of oil equivalent per capita)
FS.AST.PRVT.GD.ZS	Assets of private sector banks to GDP (%)
NE.IMP.GNFS.ZS	Imports of goods and services (% of GDP)
NY.GDP.MKTP.CD	GDP (current US$)
NE.EXP.GNFS.ZS	Exports of goods and services (% of GDP)
NY.GDP.MKTP.KD.ZG	GDP growth (annual %)
EN.ATM.GHGT.KT.CE	Total greenhouse gas emissions (kt of CO_2 equivalent)

Table 1: Indicator Table

3.2 Data Preprocessing

- Interpolated missing values for convenience.
- Constructed a country-year-feature panel: each row is a unique (country, year) pair.

3.3 Machine Learning Models

The following models are compared[2]:

- Linear Regression (LR)
- Ridge Regression
- Lasso Regression
- Elastic Net
- Support Vector Regression (SVR)
- Random Forest (RF)
- K-Nearest Neighbors (KNN)
- XGBoost
- Locally Weighted Regression (LWR)

[2]See /src/DataProcessing/models.py for parameters

3.4 Experimental Setup

- **Year ranges:**
  - Full period: 1960–2020
  - Recent period: 2010–2020
- **Cross-Validation:** 5-fold cross-validation is used for each prediction, averaging metrics across folds.
- **Evaluation Metrics:**
  - Standardized error (RMSE/STD) [6], threshold = 1
  - Coefficient of Determination (R^2) [7], threshold = 0.6
  - Mean Absolute Scaled Error (MASE) [8], threshold = 1
  - Directional Accuracy (DA) [8], threshold = 0.7
  - **Guiding Metric (Heuristic Indicator):** For each model and indicator, we define a guiding metric score as the weighted score consisting of 4 metircs. Specifically, we compute:

$$\alpha_1 = -\frac{\frac{\text{RMSE}_i}{\text{STD}_i} - 1}{1}$$

$$\alpha_2 = \frac{R_i^2 - 0.6}{0.6}$$

$$\alpha_3 = -\frac{\text{MASE}_i - 1}{1}$$

$$\alpha_4 = \frac{\text{DA}_i - 0.7}{0.7}$$

$$\text{Guiding Score} = \sum_{i=1}^{4} \frac{\beta_i}{1 + \exp(-\alpha_i)}$$

    β_i is the weight assigned to each standard metric, and in this paper we choose $\beta = \{2.0, 2.0, 1.0, 5.0\}$ because when making policies the direction seems more important than the actual value. When all metrics are at their thresholds, Guiding Score = 2.0, also the threshold for Guiding Metric.

- **Visualization:** For each model and year range, bar plots of the metrics are generated, with feasible region thresholds indicated.

4 Cross-Indicator Forecasting

4.1 Prediction Task

For each indicator F_k, we predict its value for each country-year using the remaining 9 indicators as input features. The process is repeated for all $k = 1, \ldots, 10$.

4.2 Prediction Results

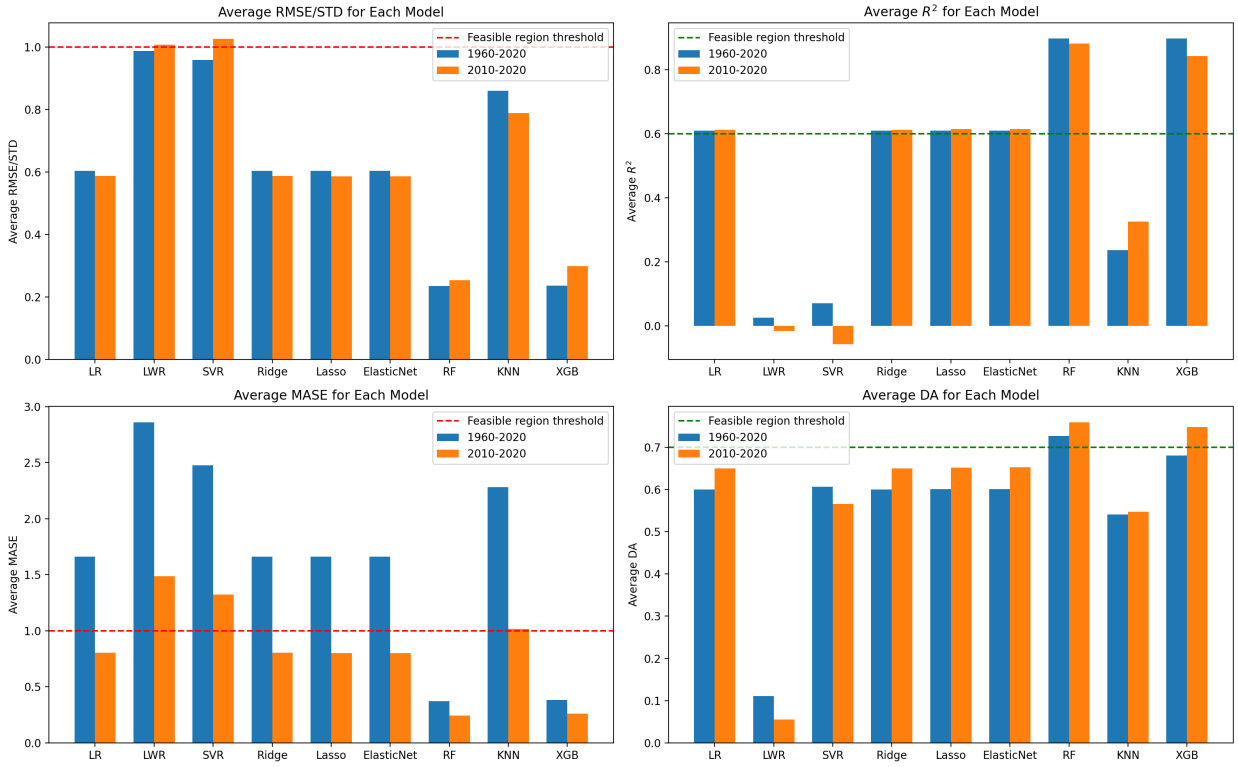


Figure 1: Comparison of Model Performance: RMSE/STD, R^2, MASE and DA (1960–2020, 2010–2020)

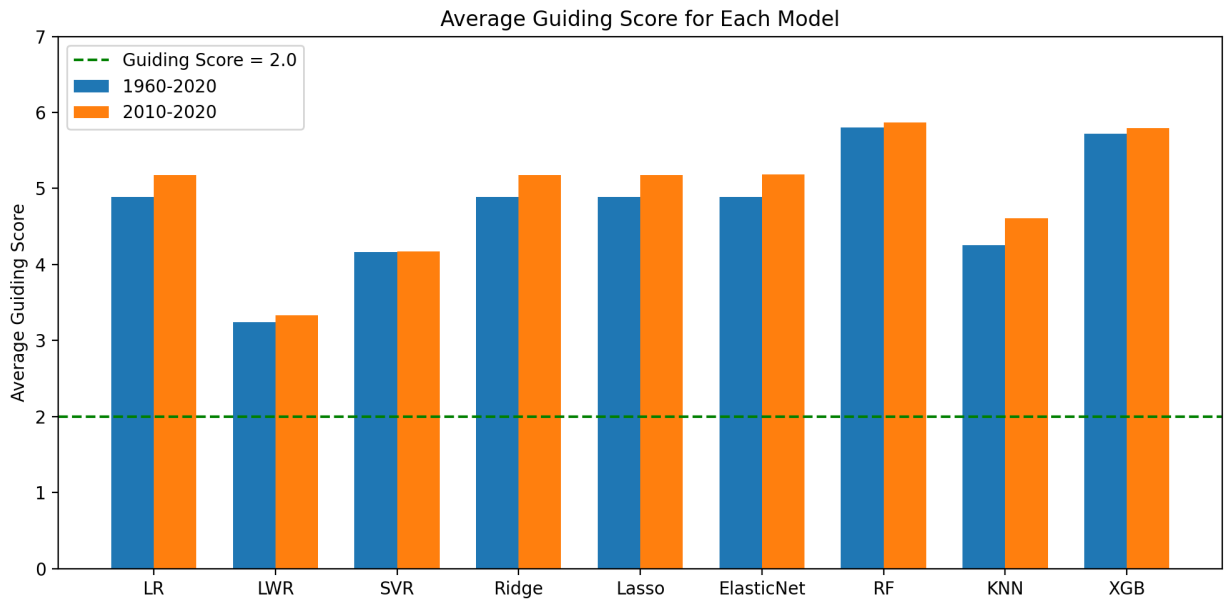


Figure 2: Comparison of Model Performance: Guiding Score (1960–2020, 2010–2020)

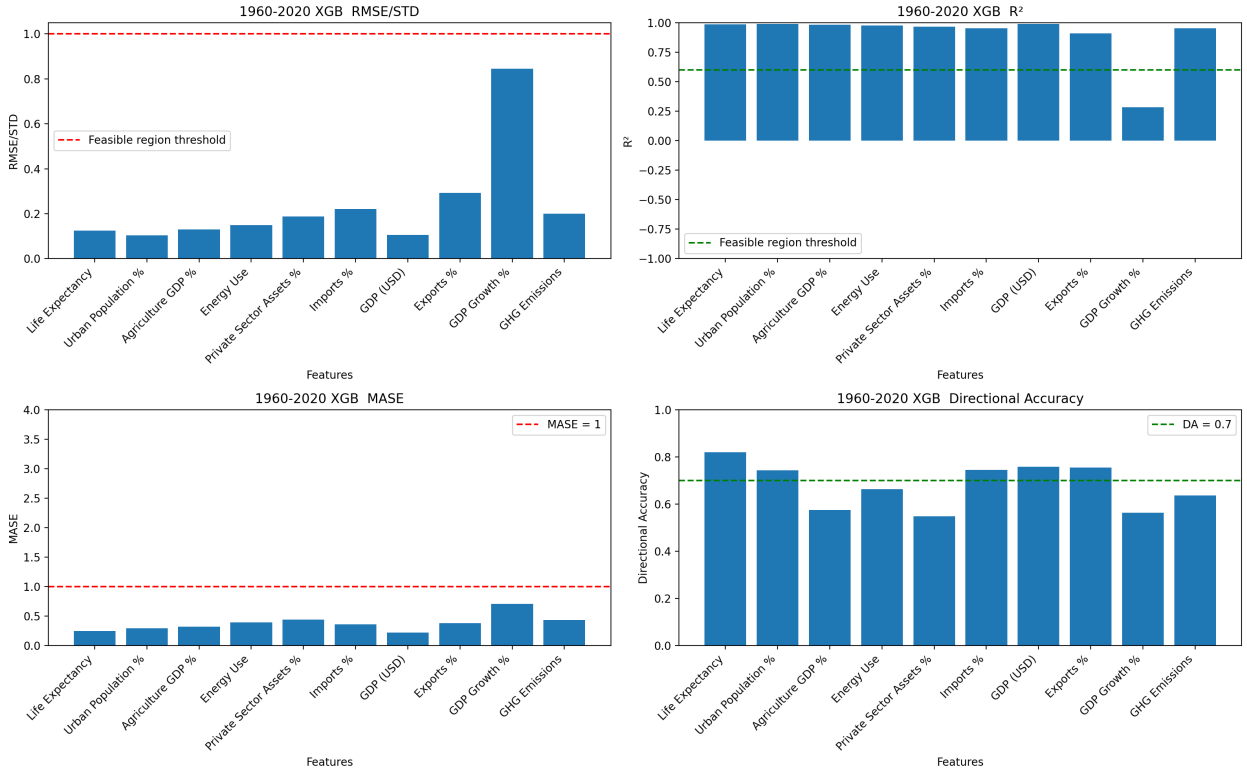


Figure 3: Prediction performance of XGBoost for each feature (1960–2020), including RMSE/STD, R^2, MASE, and Directional Accuracy.

From Figure 1 and Figure 2, We can divide the chosen Machine Learning algorithms into 3 different categories:

- RF and XGB have best performances, both of which has a low standardized error, MASE and high R^2 and DA, with a guiding score over 2.5
- LR, Ridge, Lasso, ElasticNet have very similar performances, with a guiding score about 2.0 (2010-2020)
- LWR, SVR, KNN have comparatively low performances. This is in part because the sample size (M = 1220 or 220) is rather small compared to input features (N = 10)

Figure 3 summarizes the prediction performance of XGBoost for each of the 10 selected indicators over the full period (1960–2020). The results indicate that XGBoost achieves high accuracy for most structural indicators, with standardized errors (RMSE/STD) well below 1 and R^2 values typically above 0.6. Notably, indicators such as life expectancy, urban population share, and energy use are predicted with particularly high precision. In contrast, GDP growth (annual %) stands out as the only indicator with consistently poor predictive performance, exhibiting both high error and low explanatory power. The poor predictability of the GDP growth rate compared to other major indicators is primarily due to its intrinsic volatility, exposure to a broad set of unobserved influences, and its weak contemporaneous linkages with slow-moving structural features. This is a well-documented phenomenon in economic modeling [9, 10], where forecasting economic growth remains an exceptionally challenging task.

4.3 Hyperparameter Tuning

To ensure robust performance from the ensemble models, we conducted hyperparameter tuning for both XGBoost (XGB) and Random Forest (RF) using grid search with 5-fold cross-validation.

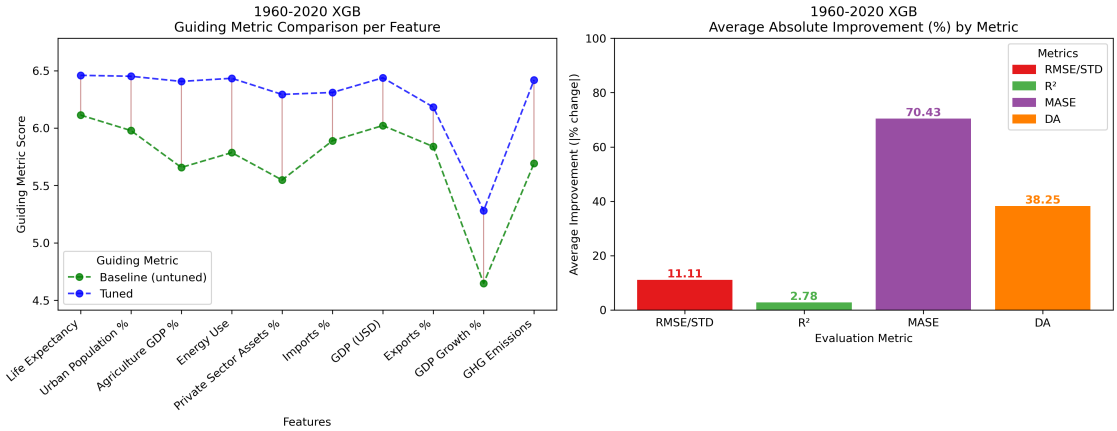
For XGBoost, the primary hyperparameters adjusted include:

- `n_estimators`: Number of boosting rounds.
- `max_depth`: Maximum depth of each tree.
- `learning_rate`: Step size shrinkage used in updates.
- `subsample`: Fraction of observations to be randomly sampled for each tree.
- `colsample_bytree`: Fraction of columns to be randomly sampled for each tree.

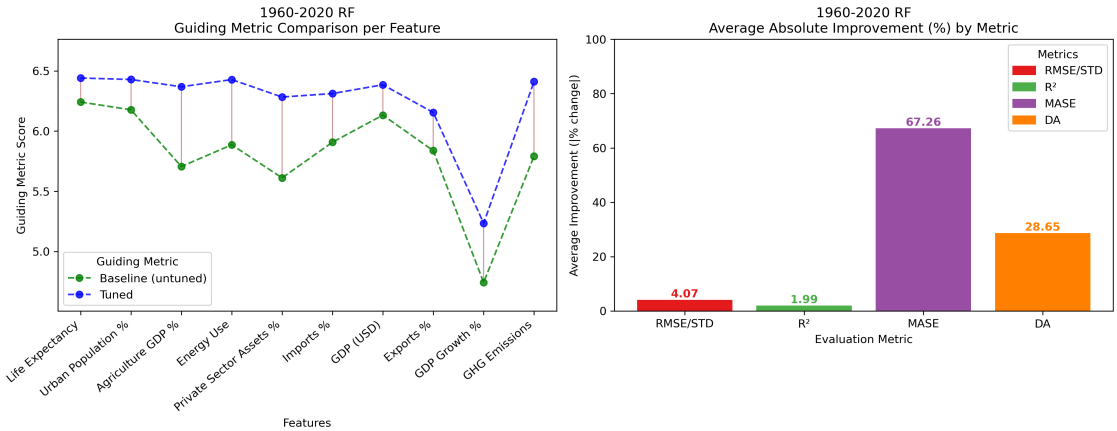
For Random Forest, the tuning focused on:

- `n_estimators`: Number of trees in the forest.
- `max_depth`: Maximum depth of the tree.

- `min_samples_split`: Minimum number of samples required to split an internal node.
- `max_features`: Number of features to consider when looking for the best split.



(a) XGBoost: Tuning Performance (1960–2020)



(b) Random Forest: Tuning Performance (1960–2020)

Figure 4: Comparison of Hyperparameter Tuning Effects for XGBoost and Random Forest (1960–2020)

Figure 4 illustrates how hyperparameter tuning improves model performance. Across all indicators, XGBoost consistently outperforms Random Forest, and both models benefit significantly from tuning.

Metric-wise, the most notable improvements appear in Directional Accuracy (DA), which rises by 30–40%. This suggests that tuning not only improves numerical accuracy but also strengthens the models' ability to correctly predict the direction of change—a crucial feature for real-world policy and decision-making. Improvements in RMSE/STD and MASE are also evident, especially for structural indicators. In contrast, R^2 gains are modest, indicating that tuning has limited effect on variance explanation but substantial effect on practical usability.

Overall, tuning contributes most to directional consistency and error reduction, making ensemble models more robust and interpretable across diverse indicators.

Figure 4 provides a detailed visualization of the performance improvements after tuning. The left panel illustrates that while the magnitude of improvement of guiding score varies across indicators, all ten features experience a substantial average enhancement. This confirms that tuning has universal benefit, though its effect size depends on feature characteristics.

The right panel decomposes improvements across the four evaluation metrics. Notably, tuning yields minimal changes in R^2 (near 2%), modest gains in RMSE/STD (ranging from 5% to 10%), and substantial improvements in Directional Accuracy (DA), which increase by approximately 30%–40%. The most dramatic effect is observed in Mean Absolute Scaled Error (MASE), where XGBoost and Random Forest achieves a nearly 70% improvement. These results highlight how hyperparameter tuning differentially impacts specific model objectives and offer insights into which dimensions of forecast accuracy are most tunable.

4.4 Summary

Chapter 4 evaluated the feasibility of cross-indicator forecasting using machine learning models applied to national-level economic and social data. By predicting each indicator from the remaining features, we tested nine models across two time periods (1960–2020 and 2010–2020), and evaluated their performance using a composite guiding score.

XGBoost and Random Forest consistently outperformed other models, supporting the broader literature on the strengths of ensemble methods in capturing nonlinear macroeconomic patterns [11, 12]. Structural indicators—such as life expectancy, urban population, and energy use—were highly predictable, showing low error and strong directional accuracy. In contrast, GDP growth remained inherently difficult to forecast due to volatility and external shocks [9, 10].

Model performance also improved in the 2010–2020 window, likely reflecting better data quality and macroeconomic convergence across countries [13,14]. These results emphasize the importance of model type, indicator stability, and temporal context in determining forecast reliability.

In summary, ensemble-based machine learning models, when properly tuned and evaluated with a balanced metric like the guiding score, offer a robust framework for forecasting slow-moving national indicators. This lays a solid foundation for the time-series and cross-national forecasting explored in subsequent chapters.

5 Country-Level Time Series Forecasting

5.1 Dataset and Feature Construction

This chapter uses the same interpolated dataset from 1960 to 2020 as in Chapter 4, including all available countries. For each of the 10 selected key indicators, we constructed lag-based

time series features to facilitate temporal prediction modeling. Specifically, we created lag-2 and lag-3 versions of all other features (excluding the target) to predict each target feature value year by year.

The resulting time series dataset preserves the year and country code metadata, and ensures that the predictive modeling process incorporates temporal dynamics. All features were standardized prior to modeling to avoid scale issues.

5.2 Models and Design Logic

We evaluated six models for each country and each target indicator:

- **Naive Forecast**: Uses the value of the previous year as the prediction for the next year.
- **ARIMA**: A univariate autoregressive model applied independently to each target series.
- **Rolling XGB-lag2**: Same as XGB-lag2, but predictions are generated year-by-year using only prior data up to that year (rolling forecast).
- **Rolling XGB-lag3**: Extends the rolling forecast logic to lag-3 input features.

The rolling forecast strategy better simulates real-world forecasting where future data is unavailable during training. The comparison across static and rolling variants allows us to evaluate model generalizability and robustness over time.

5.3 Model Comparison

Country-Level Model Performances Analysis

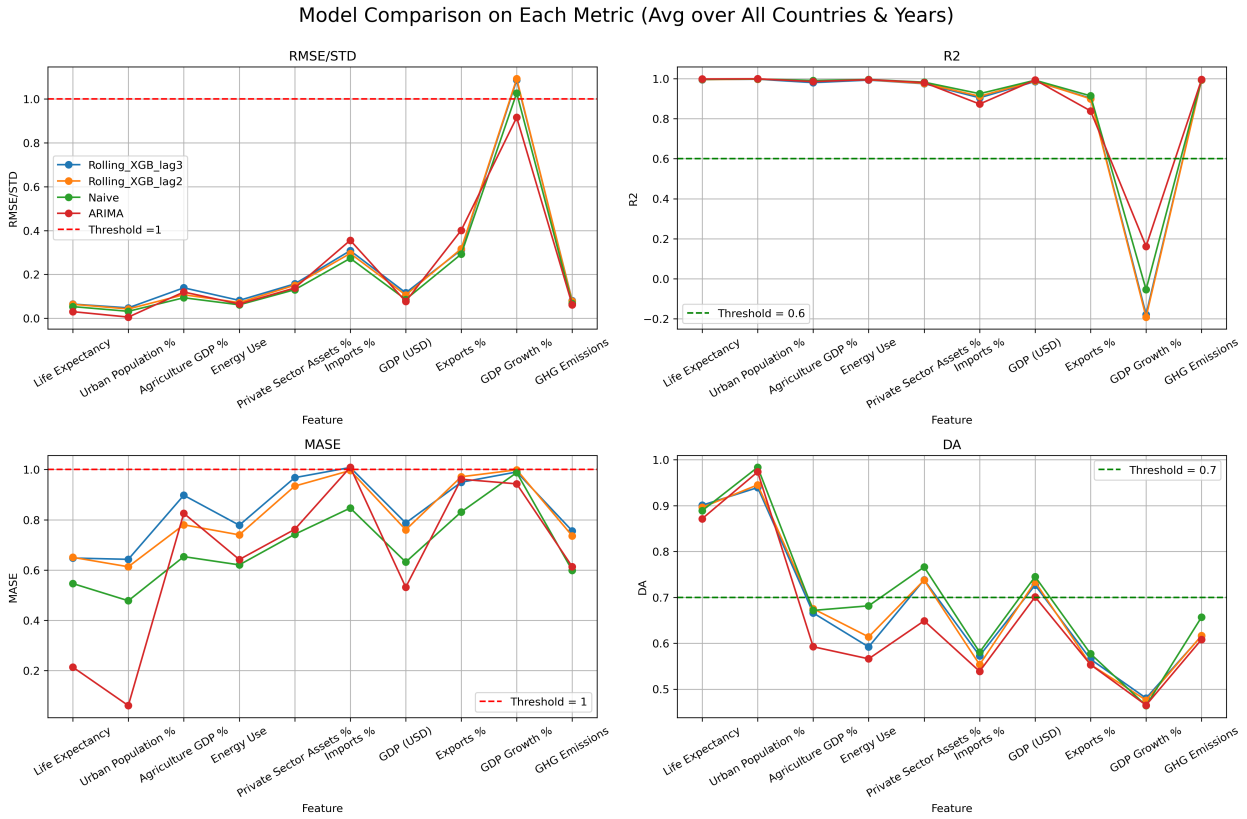


Figure 5: Comparison of model performance across indicators using four metrics: RMSE/STD, R^2, MASE, and Directional Accuracy (DA).

We identify three key findings:

First, the Naive Forecast ranks first in Directional Accuracy (DA) for nearly all features, which directly elevates its overall guiding score. This can be attributed to the fact that most national economic indicators—except GDP growth—tend to evolve steadily over short time horizons. In other words, $F_i^{(j)}$ for a feature i in year j is approximately a linear function of j, resulting in similar signs between the predicted change $\Delta F_i = F_i^{(j)} - F_i^{(j-1)}$ and the actual change $\Delta F_i' = F_i^{(j+1)} - F_i^{(j)}$. Because our guiding metric assigns the highest weight (β=5) to DA, the Naive method, despite its simplicity, achieves the top overall score. This inertia-driven advantage is also supported by economic theory. According to structural transformation theory [5], indicators such as urbanization, agricultural share, and energy use follow stable, long-term trajectories during development, making them easier to predict directionally using simple heuristics.

Second, across the other three metrics—RMSE/STD, R^2, and MASE—all four models (Naive, ARIMA, Rolling XGB lag2, and lag3) exhibit comparable performance. As shown

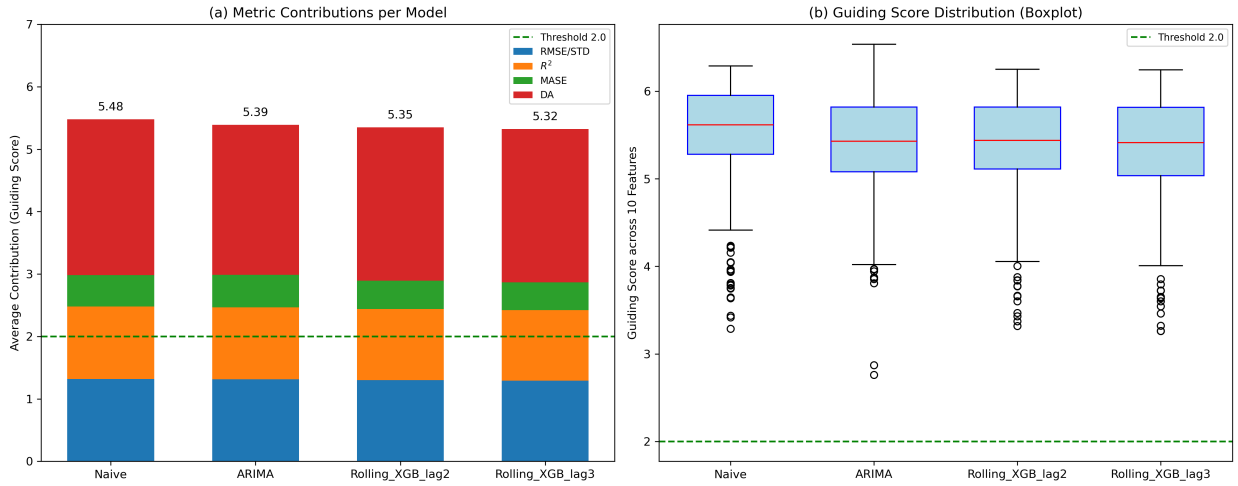


Figure 6: Left figure shows comparison of model performance across indicators using guiding metric. Right figure shows the distribution of guiding scores for each model.

in Figure 6 (left), their average scores in these dimensions differ only marginally. Furthermore, Figure 5 supports this result: most features surpass the threshold for RMSE/STD and R^2, indicating strong performance across the board, but fall short in MASE when predicting volatile indicators such as GDP growth. This aligns with prior literature underscoring the unpredictability of economic growth due to policy shifts, external shocks, and cyclical disturbances [9].

Finally, the boxplot in Figure 6 (right) reveals substantial differences in distribution. Although the Naive Forecast has the highest median score and a compact interquartile range, it also produces more low-score outliers. ARIMA yields the highest maximum scores and performs exceptionally well on trend-dominated features like urbanization, benefiting from its built-in trend and seasonality components. However, it also exhibits the widest variance, indicating less robustness. Rolling XGB models show tighter distributions and fewer outliers, suggesting better generalization in high-dimensional, nonlinear contexts. Their advantage lies in modeling complex interactions among indicators—an area where tree-based ML models excel [11, 12].

These findings highlight the critical role of DA in short-term forecasting and suggest that simple methods can remain competitive when variables evolve monotonically. Conversely, complex models offer better scalability for non-stationary or feature-rich environments. Future improvements may include volatility-aware or Bayesian structural break models [15, 16] to better capture unpredictable features like GDP growth.

Indicator-Specific Feasibility by Country

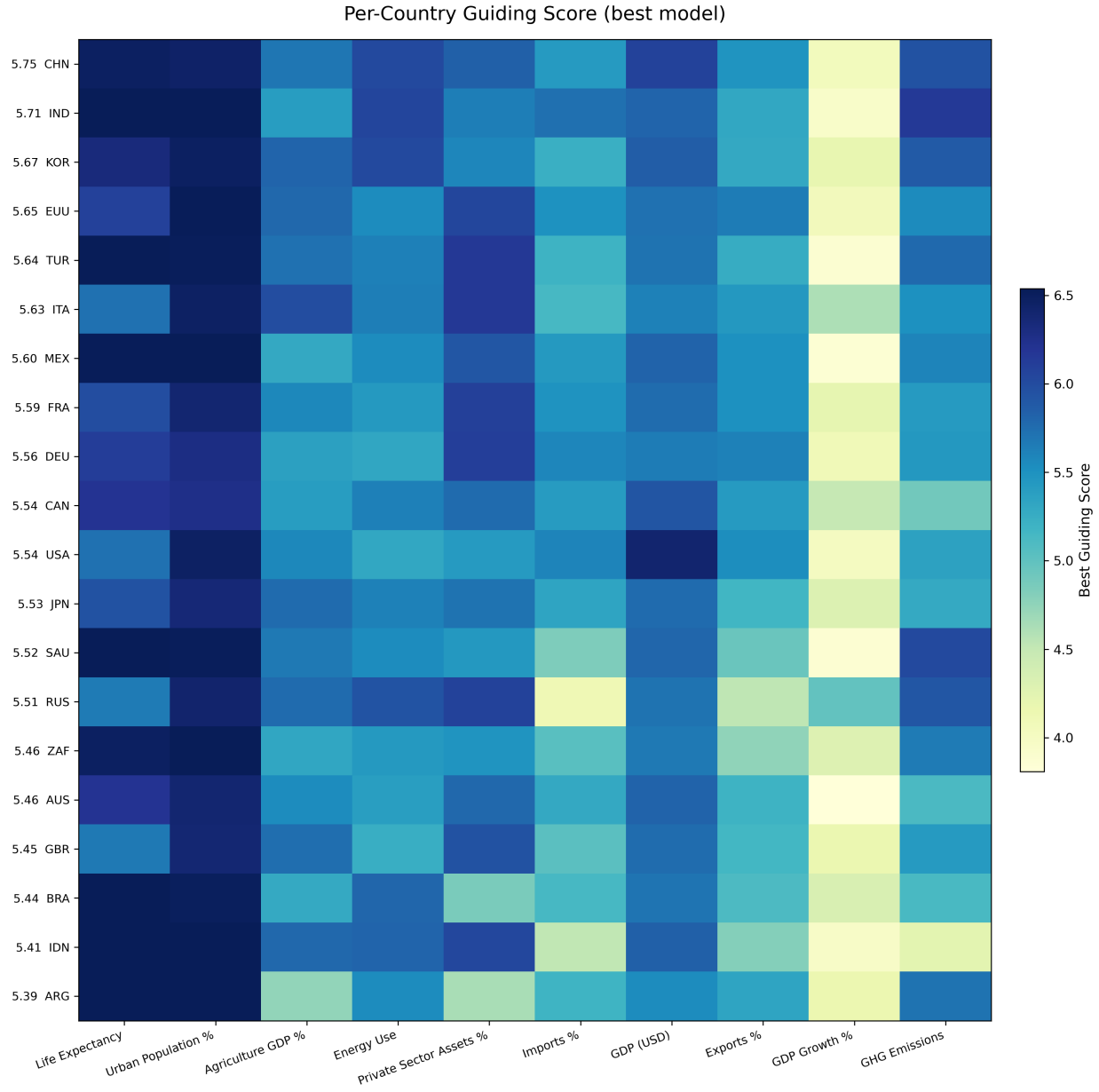


Figure 7: Heatmap of Per-Indicator Feasibility Scores by Country. For each (country, indicator) pair, we select the best-performing model and report the corresponding guiding score. Darker colors indicate more reliable prediction.

Figure 7 provides a granular view of how forecastability varies both across indicators and across economies.

Three patterns stand out.

(i) Structural indicators are universally easy to predict. Life expectancy and urban-population share (first two columns) display uniformly dark shades across nearly every country, reflecting their smooth, monotonic trajectories and the ability of even simple models to capture them. This observation is consistent with the "conditional convergence" hypothesis for demographic and infrastructure variables [17].

(ii) Macro-financial variables exhibit strong cross-country heterogeneity. Columns for private-sector assets, exports, and imports appear markedly lighter for commodity-oriented economies such as RUS [18,19] and IDN [20] —an outcome that mirrors the high terms-of-trade volatility typical of resource-dependent growth [18]. In contrast, advanced economies with diversified trade baskets—such as Germany (DEU), Japan (JPN), and the United States (USA)—achieve guiding scores above 5.5, consistent with IMF Article IV evidence that their external sectors are broadly diversified and less sensitive to commodity-price swings [21–23]. However, note that countries with the highest guiding scores generally submit more complete data series to the World Bank; the reduced need for interpolation in turn lowers measurement error and boosts forecast reliability.

(iii) GDP growth remains the Achilles' heel of economic forecasting. The penultimate column is the lightest for almost every country, reaffirming our earlier conclusion that growth is intrinsically noisy and difficult to pin down. Even for historically steady performers such as CHN or IND, the best guiding score rarely exceeds 4.2. This pattern echoes the well-documented "growth-forecast puzzle" discussed by [9].

Taken together, Figure 7 suggests that cross-country heterogeneity in forecastability is driven less by modelling technique and more by structural economic characteristics—specifically, the degree of exposure to external shocks and the maturity of demographic transitions. Policymakers can thus place greater confidence in forecasts for slow-moving structural variables, while treating projections for volatile indicators with appropriate caution.

5.4 Case Study: G7 and EU

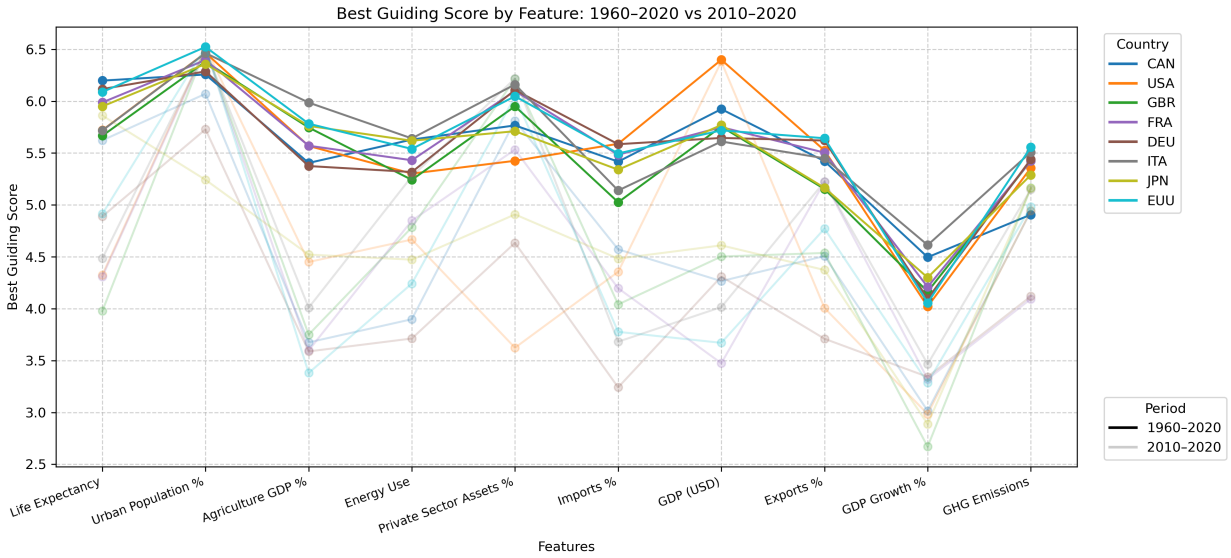


Figure 8: Guiding Score comparison across key indicators for selected G7 and EU countries. Each polyline represents a country, with higher scores indicating more reliable prediction. Solid lines correspond to the full period (1960–2020), while faded lines represent the recent decade (2010–2020).

Figure 8 reinforces earlier findings: life expectancy and urban population share remain the most predictable indicators, with scores exceeding 5.5 and 6.0 across all countries and time periods. In contrast, GDP growth consistently exhibits the lowest scores, rarely surpassing 4.5, confirming its volatility [9]. Most G7 economies show minimal performance drop between the two time periods, indicating data stability and model robustness over time.

5.5 Case Study: BRICS-5

From Figure 9, structural indicators again stand out: urbanization and life expectancy exceed the guiding threshold of 5.5 and 6.0 across all five countries. However, BRICS economies display greater cross-feature and cross-period variation, especially for financial indicators like private sector assets and exports, reflecting their more dynamic and transitional economic structures [24]. India and South Africa exhibit sharper declines in the recent decade (2010–2020), likely due to external shocks and data limitations. The persistently low scores for GDP growth echo global patterns and further underscore the limits of model-based forecasting for cyclical variables [10].

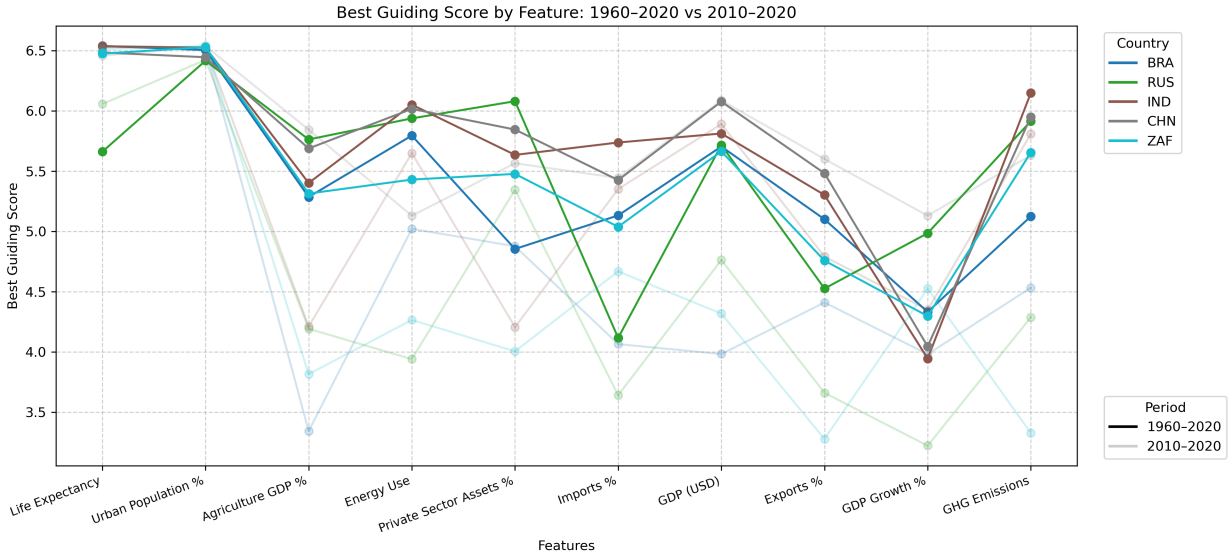


Figure 9: Guiding Score comparison across key indicators for BRICS-5 countries (Brazil, Russia, India, China, South Africa).

6 Bayesian Latent Structure Model for Multi-Country Economic Forecasting

6.1 Model Motivation and Setup

Notation and Dimensions. Let m be the number of features and n the number of countries. For each country c at time t, let $\mathbf{Y}_{t,c} \in \mathbb{R}^m$ be the observed feature vector. Define a fixed feature mask $\mathcal{I} \subseteq \{1, \ldots, m\}$ of size d (e.g., import/export-related indices) used for cross-country influence. For any country j, define $\mathbf{Z}_{t,j} \in \mathbb{R}^m$ by setting:

$$(\mathbf{Z}_{t,j})_i = \begin{cases} (\mathbf{Y}_{t,j})_i & \text{if } i \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

For any country c, define the stacked cross-feature matrix $X_{t,c} \in \mathbb{R}^{(n-1)\times m}$ as:

$$X_{t,c} = \begin{bmatrix} \mathbf{Z}_{t-1,1}^\top \\ \vdots \\ \mathbf{Z}_{t-1,c-1}^\top \\ \mathbf{Z}_{t-1,c+1}^\top \\ \vdots \\ \mathbf{Z}_{t-1,n}^\top \end{bmatrix}$$

In this chapter, we propose a Bayesian latent structure framework designed to forecast key economic indicators across multiple countries by integrating time-lagged dependencies and

cross-national interactions. Our approach aims to strike a balance between interpretability, forecasting accuracy, and theoretical grounding.

Let $\mathbf{Y}_{t,c}$ denote the standardized vector of economic indicators for country c in year t. We assume the observed data can be explained by latent shared influences through a generative model of the form:

$$\mathbf{Y}_{t,c} = \sum_{i=1}^{\tau} A_{t,c}^{i} \mathbf{Y}_{t-i,c} + \sum_{j \neq c} M_{j,c} \mathbf{Z}_{t-1,j} + \boldsymbol{\epsilon}_{t,c}, \quad \boldsymbol{\epsilon}_{t,c} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$$

Here, the dimensions of the variables are as follows: $\mathbf{Y}_{t,c} \in \mathbb{R}^m$ represents the observed indicators for country c at time t, $A_{t,c} \in \mathbb{R}^{m \times m}$ is the transition matrix for self-history, $M_{j,c} \in \mathbb{R}^{m \times m}$ maps the external features $\mathbf{Z}_{t-1,j} \in \mathbb{R}^m$ from country j to the output space, and $\boldsymbol{\epsilon}_{t,c} \in \mathbb{R}^m$ is the noise term.

The model incorporates three main components:

- **Self-History Influence:** We include the past τ years of the country's own indicators $\mathbf{Y}_{t-i,c}$, aggregated with exponentially weighted powers of matrix A. To ensure diminishing influence over time and model stability, we assume $\det A < 1$.

- **Cross-Country Influence:** For external effects, we use the sparse cross-feature vectors $\mathbf{Z}_{t-1,j} \in \mathbb{R}^m$ as defined above, with only a predefined subset of d components (corresponding to selected cross features) nonzero. This sparse representation ensures that $\mathbf{Z}$ has the same dimension as $\mathbf{Y}$, but only encodes relevant cross-national influences. We define a set of matrices $\{M_{j,c} \in \mathbb{R}^{m \times m} \mid j \neq c\}$ that model how the retained features from each country j influence the target country c. The total cross-country effect is computed as:

  $$\sum_{j \neq c} M_{j,c} \mathbf{Z}_{t-1,j}$$

  This setup allows each $M_{j,c}$ to focus on relevant dimensions, while preserving compatibility with $\mathbf{Y}_{t,c}$.

- **Noise Term:** The additive noise $\boldsymbol{\epsilon}_{t,c}$ accounts for unexplained variability, assumed to be i.i.d. Gaussian with zero mean and isotropic covariance.

In this paper, we select $\tau = 2$ based on empirical evidence that XGB_lag2 and XGB_lag3 models show minimal performance difference in earlier experiments, indicating that higher powers of A (i.e., A^3 and beyond) have negligible contribution. We set $d = 2$ to focus on two key cross-country indicators—import and export proportions—which capture essential aspects of international economic interactions.

In this formulation, we assume a uniform cross-country influence mechanism across all countries, which allows simplification in later estimation steps.

6.2 Simplified Estimation with Shared Cross-Country Influence

6.2.1 Assumptions and Notation

We assume that all countries share the same cross-country influence weights $M \in \mathbb{R}^{m\times m}$ applied to a masked cross-feature sum. This allows us to define:

$$\mathbf{Q}_{t,c} = \sum_{j\neq c} \mathbf{Z}_{t-1,j}$$

where each $\mathbf{Z}_{t-1,j}$ is a masked version of $\mathbf{Y}_{t-1,j}$ retaining only d predefined features. The vector $\mathbf{Q}_{t,c} \in \mathbb{R}^m$ is used to represent the aggregated cross-country input for predicting $\mathbf{Y}_{t,c}$.

6.2.2 Ridge Regression Estimation

The prediction for each country c at time t is:

$$\hat{\mathbf{Y}}_{t,c} = A^1\mathbf{Y}_{t-1,c} + A^2\mathbf{Y}_{t-2,c} + \sum_{j\neq c} M_j\mathbf{Z}_{t-1,j}$$

We denote the prediction residual as:

$$\mathbf{r}_{t,c} = \mathbf{Y}_{t,c} - \hat{\mathbf{Y}}_{t,c}$$

The objective is to minimize the total loss over all years and countries using a ridge penalty:

$$\mathcal{L} = \sum_{t,c} \|\mathbf{r}_{t,c}\|_2^2 + \lambda_A \left(\|A^1\|_F^2 + \|A^2\|_F^2\right) + \lambda_M \sum_j \|M_j\|_F^2$$

where λ_A and λ_M are regularization parameters.

To estimate the parameters, we organize the inputs into a linear model.

For each (t, c), define the concatenated input vector:

$$\mathbf{h}_{t,c} = \begin{bmatrix} \mathbf{Y}_{t-1,c}^T \\ \mathbf{Y}_{t-2,c}^T \\ \{\mathbf{Z}_{t-1,j}\}_{j\neq c}^T \end{bmatrix} \in \mathbb{R}^{m(\tau+n-1)}$$

and corresponding parameter block:

$$\mathbf{W}_c = \begin{bmatrix} (A^1)^T \\ (A^2)^T \\ \{M_j\}_{j\neq c}^T \end{bmatrix} \in \mathbb{R}^{m(\tau+n-1)\times m}$$

The prediction becomes:

$$\hat{\mathbf{Y}}_{t,c} = \mathbf{W}_c^\top \mathbf{h}_{t,c}$$

Letting $\mathbf{H}_c$ be the stacked input matrix and $\mathbf{Y}_c$ the corresponding outputs over time:

$$\mathbf{H}_c = \begin{bmatrix} \mathbf{h}_{\tau+1,c}^\top \\ \vdots \\ \mathbf{h}_{T,c}^\top \end{bmatrix} \in \mathbb{R}^{(T-\tau)\times m(\tau+n-1)}, \quad \mathbf{Y}_c = \begin{bmatrix} \mathbf{Y}_{\tau+1,c}^\top \\ \vdots \\ \mathbf{Y}_{T,c}^\top \end{bmatrix} \in \mathbb{R}^{(T-\tau)\times m}$$

Initially, we set the regularization strength for both the self-influence terms A and cross-influence terms M to be the same, i.e., $\lambda_A = \lambda_M$. Moreover, we let $\lambda_A = \lambda_M = 1.0$. Then the objective is:

$$\mathcal{L}_c = \|\mathbf{Y}_c - \mathbf{H}_c\mathbf{W}_c\|_F^2 + \lambda\|\mathbf{W}_c\|_F^2$$

and the closed-form ridge solution is:

$$\mathbf{W}_c = (\mathbf{H}_c^\top\mathbf{H}_c + \lambda I)^{-1}\mathbf{H}_c^\top\mathbf{Y}_c$$

This simplification facilitates direct estimation using standard ridge regression. The guiding scores of this baseline are shown in Figure 10.

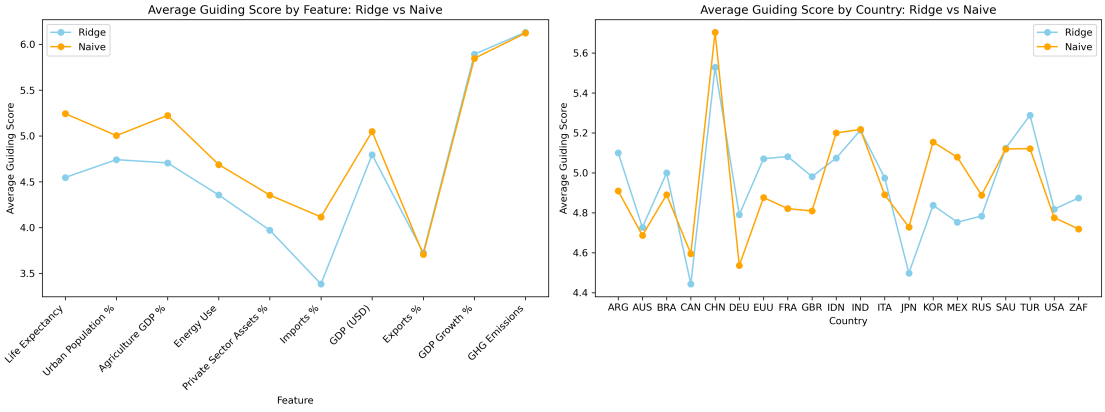


Figure 10: Guiding scores for Ridge vs Naive using uniform regularization

To better capture the different roles of A and M, we then optimize λ_A and λ_M separately. This allows for more flexible control over self and cross influence regularization strengths. The updated guiding score comparison after separate tuning is presented in Figure 11.

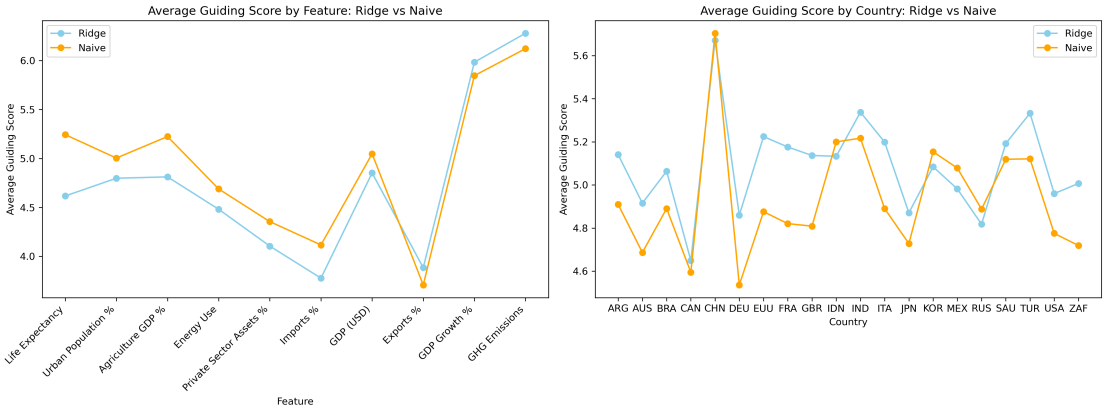


Figure 11: Guiding scores after parameter tuning for each (country, feature) pair.

Figure 11 reveals that tuning (λ_A, λ_M) for each individual (country, feature) pair yields consistent improvements in guiding score relative to the uniform-regularization baseline in Figure 10. This gain is most apparent for structural indicators and stable countries, where tailored regularization allows the model to better separate temporal inertia from cross-national influences.

6.2.3 Model Evaluation and Visualization

7 Generalized Cross-Country Modeling

All cross-country influence is encoded by the sum $\mathbf{Q}_{t,c} = \sum_{j \neq c} \mathbf{Z}_{t-1,j}$, preserving structural sparsity while retaining matrix compatibility.

8 Conclusion

This work demonstrates that most major structural indicators of national development are highly predictable from a small set of other key indicators, especially when using ensemble tree-based machine learning models. The exception is GDP growth rate, which remains notoriously difficult to forecast—consistent with macroeconomic theory and previous empirical research.

Our results suggest that, for long-run cross-country comparative analysis, reliable prediction of most economic and demographic indicators is feasible using standard machine learning approaches and open-access datasets. However, caution should be exercised when interpreting models for inherently volatile outcomes such as economic growth. Overall, this study highlights the promise and limitations of data-driven prediction in international development research and points to several avenues for further methodological and substantive refinement.

Project Repository

The full code, data preprocessing scripts, and results can be found at: [GitHub link will be inserted here].

References

[1] Philippe Goulet Coulombe, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964, 2022.

[2] Aaron Smalter Hall. Machine learning approaches to macroeconomic forecasting. *Economic Review*, 103(4):63–81, 2018.

[3] Yanqing Yang, Xingcheng Xu, Jinfeng Ge, and Yan Xu. Machine learning for economic forecasting: An application to china's gdp growth, 2024.

[4] Robert M. Solow. A contribution to the theory of economic growth. *The Quarterly Journal of Economics*, 70(1):65–94, 1956.

[5] Simon Kuznets. *Economic Growth of Nations: Total Output and Production Structure*. Harvard University Press, 1971.

[6] Tianfeng Chai and Roland R. Draxler. Root mean square error (rmse) or mean absolute error (mae)? arguments against avoiding rmse in the literature. *Geoscientific Model Development*, 7(3):1247–1250, 2014.

[7] Carol Alexander. *Market models: A guide to financial data analysis*. John Wiley & Sons, 2001.

[8] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[9] Prakash Loungani. How accurate are private sector forecasts? cross-country evidence from consensus forecasts of output growth. *International Journal of Forecasting*, 17(3):419–432, 2001.

[10] Michael P. Clements and David F. Hendry. How far can we forecast? *Journal of Forecasting*, 21(1):1–27, 2002.

[11] Sendhil Mullainathan and Jann Spiess. Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

[12] Susan Athey. The impact of machine learning on economics. *The economics of artificial intelligence: An agenda*, 2019.

[13] Richard Baldwin. *The Great Convergence: Information Technology and the New Globalization*. Harvard University Press, 2016.

[14] Morten Jerven. *Poor Numbers: How We Are Misled by African Development Statistics and What to Do About It*. Cornell University Press, 2013.

[15] James H Stock and Mark W Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002.

[16] Gary Koop and Dimitris Korobilis. Forecasting economic time series using flexible bayesian vars. *International Journal of Forecasting*, 29(4):685–700, 2013.

[17] Robert J. Barro and Xavier Sala i Martin. Convergence. *Journal of Political Economy*, 100(2):223–251, 1992.

[18] Raúl Prebisch. *The Economic Development of Latin America and Its Principal Problems*. United Nations Economic Commission for Latin America and the Caribbean, New York, 1950.

[19] Jeffrey D. Sachs and Andrew M. Warner. The curse of natural resources. *European Economic Review*, 45(4–6):827–838, 2001.

[20] Frederick van der Ploeg and Steven Poelhekke. Volatility and the natural resource curse. *Oxford Economic Papers*, 61(4):727–760, 2009.

[21] International Monetary Fund. Germany: 2023 article iv consultation—press release; staff report, 2023. IMF Country Report No. 23/200.

[22] International Monetary Fund. Japan: 2023 article iv consultation—press release; staff report, 2023. IMF Country Report No. 23/270.

[23] International Monetary Fund. United states: 2023 article iv consultation—press release; staff report, 2023. IMF Country Report No. 23/220.

[24] Justin Yifu Lin. New structural economics: A framework for rethinking development. *World Bank Research Observer*, 27(2):193–221, 2012.