Review of Futures Markets (RFM)
ISSN: 1933-7116
Volume 34 Issue 2 (2026)
Website: https://fra.st

Explainable Deep Learning for Public Revenue–Expenditure Dynamics: A Data-Driven Framework for Intelligent Fiscal Planning

Riad eddine Salhi [1], Hassan Elmadhoun [2]
[1] Ra.salhi@univ-boumerdes.dz, University of Boumerdes, Boumerdes, Algeria
[2] prhassanelmadhoun@gmail.com, University of Algiers 3 – Ibrahim sultan cheibout,, Algeria

Received : 28 /04 /2026 Accepted : 03/07/2026 Published : 11/08/2026

Abstract
Accurate forecasting of public revenue and expenditure is essential for effective fiscal planning and budgetary management. This study investigates the applicability of machine learning and deep learning algorithms to the joint forecasting of public revenue and public expenditure. A comparative forecasting framework was developed using Random Forest, XGBoost, Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (Bi-LSTM), and Transformer models. A naïve forecasting benchmark was also incorporated to provide a reference for evaluating the predictive value of the more complex algorithms. Model performance was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2) under a chronological out-of-sample evaluation framework.
The empirical results reveal substantial differences in predictive performance across the evaluated architectures. For public revenue, the naïve benchmark achieved the best overall performance, with an MAE of 535,629, an RMSE of 712,275, a MAPE of 10.10%, and an R^2 of 0.485. Among the deep learning models, LSTM achieved the strongest performance, with an MAE of 809,035, an RMSE of 867,162, a MAPE of 17.07%, and an R^2 of 0.2366. For public expenditure, the naïve benchmark again produced the lowest forecasting errors, while Bi-LSTM achieved the best performance among the deep learning models, obtaining an MAE of 1,750,307, an RMSE of 1,811,825, and a MAPE of 24.08%. In contrast, the Transformer model produced the largest forecasting errors for both variables.
The findings demonstrate that increasing model complexity does not necessarily lead to superior forecasting accuracy, particularly when working with a limited number of annual observations. The results emphasize the importance of benchmark comparison, model selection, and chronological out-of-sample validation when applying artificial intelligence to fiscal forecasting. The study provides empirical evidence that recurrent deep learning architectures can capture useful temporal patterns in public finance data, while also highlighting the importance of considering simpler forecasting alternatives in fiscal planning applications.

Keywords: Public Revenue; Public Expenditure; Fiscal Forecasting; Machine Learning; Deep Learning; LSTM; GRU; Bi-LSTM; Transformer; Fiscal Planning.

.

1 Introduction

Fiscal planning is a fundamental component of economic policy because government revenues and expenditures determine the capacity of public authorities to finance public services, implement development programs, and maintain fiscal stability. The ability to anticipate the evolution of these fiscal variables is therefore essential for effective budget preparation and medium- and long-term fiscal management. However, fiscal variables are influenced by complex economic conditions, institutional decisions, external shocks, and structural changes, making their future evolution difficult to characterize using simple deterministic relationships. Public revenues and expenditures are also intrinsically interconnected components of fiscal systems. Changes in government revenues may influence expenditure capacity, while expenditure decisions may affect future revenue requirements and fiscal adjustments. The resulting relationship can therefore involve temporal dependence, nonlinear interactions, and feedback mechanisms. Understanding these characteristics is particularly important for fiscal planning because forecasts of revenues and expenditures are ultimately interpreted within the broader context of fiscal balance and budgetary sustainability.

Conventional fiscal forecasting has traditionally relied on econometric and statistical models such as autoregressive specifications, vector autoregressive models, and other time-series techniques. These approaches provide important theoretical and statistical foundations for fiscal analysis and remain valuable for identifying relationships among economic variables. Nevertheless, their ability to represent highly nonlinear relationships and complex temporal patterns may be limited when the underlying fiscal process is affected by structural changes and irregular shocks. This has encouraged increasing interest in artificial intelligence and machine-learning methods as complementary approaches to economic and fiscal forecasting.

Recent research has demonstrated the growing potential of artificial intelligence in public finance. Machine-learning and deep-learning techniques have been applied to government expenditure forecasting, public revenue forecasting, tax forecasting, and broader fiscal-risk analysis. Importantly, recent evidence also indicates that forecasting performance depends on the characteristics of the dataset, the forecasting horizon, and the model architecture. Consequently, the increasing sophistication of artificial intelligence does not automatically imply superior forecasting performance, reinforcing the importance of systematic empirical comparison.

Deep learning is particularly relevant to fiscal time-series analysis because neural architectures can learn nonlinear representations directly from sequential observations. Recurrent architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are capable of representing temporal dependencies, while more recent attention-based architectures provide alternative mechanisms for identifying relevant information within historical sequences. These capabilities create an opportunity to examine whether modern deep-learning architectures can provide additional information about the dynamic interaction between public revenues and expenditures.

The motivation for the present study is closely related to this distinction. Previous research has already demonstrated the usefulness of nonlinear neural forecasting for fiscal variables and has shown that revenue and expenditure information can be incorporated jointly within a lagged forecasting structure. In particular, previous work has combined a multivariate lagged Generalized Regression Neural Network (GRNN) with a Multi-Objective Fuzzy Goal Programming layer to transform fiscal forecasts into an operational budget-optimization framework. This approach demonstrates how artificial intelligence can contribute to fiscal forecasting while subsequently connecting predictions to flexible budgetary objectives.

However, forecasting and optimization represent only one dimension of intelligent fiscal analysis. A separate methodological question concerns whether modern deep-learning architectures can learn the temporal structure of the revenue–expenditure system and, importantly, whether their predictions can be interpreted in a transparent manner. The internal complexity of deep neural networks may make it difficult for policymakers and researchers to determine which historical fiscal observations contribute most strongly to a particular forecast. This creates a need to complement predictive evaluation with an explicit analysis of model behavior.

Explainable Artificial Intelligence (XAI) provides an appropriate methodological perspective for addressing this issue. Rather than treating a forecasting model as an opaque predictive mechanism, explainability techniques can provide information about the contribution of individual input variables or historical observations to model predictions. In the context of fiscal forecasting, this may help determine whether recent revenue observations, expenditure observations, or their lagged combinations provide the most relevant information for predicting future fiscal outcomes. Such information can be valuable for fiscal planning because it provides an interpretable layer between the computational model and the decision-making process.

The present study therefore does not seek to reproduce an optimization-based fiscal planning system. Instead, it focuses on the predictive and explanatory dimensions of artificial intelligence. The study investigates whether modern deep-learning architectures can model public revenue–expenditure dynamics effectively and whether Explainable Artificial Intelligence can reveal the temporal fiscal information underlying their predictions.

Accordingly, the main research question is formulated as follows:

How effectively can explainable deep-learning models capture public revenue–expenditure dynamics and provide interpretable information for intelligent fiscal planning?

The study addresses the following specific research questions:

RQ1: To what extent can deep-learning architectures capture the temporal and nonlinear dynamics between public revenues and public expenditures?

RQ2: Which deep-learning architecture provides the most reliable out-of-sample forecasts for public revenues and expenditures?

RQ3: Does jointly modeling revenue and expenditure information improve forecasting performance relative to separate single-output models?

RQ4: Which historical fiscal variables and temporal lags contribute most strongly to the model's predictions?

RQ5: How can Explainable Artificial Intelligence improve the interpretation of deep-learning forecasts in the context of fiscal planning?

1.1 Research Objectives

To answer these questions, the study pursues five main objectives:

1. To model the temporal dynamics between public revenues and public expenditures using modern deep-learning architectures.
2. To evaluate and compare alternative deep-learning models for fiscal forecasting under a consistent time-series validation framework.
3. To examine whether joint revenue–expenditure modeling provides predictive advantages compared with independent forecasting specifications.
4. To identify the most influential historical fiscal information contributing to the forecasts generated by the best-performing models.
5. To develop an explainable forecasting perspective for fiscal planning, allowing model predictions to be interpreted without converting the forecasting system into a separate optimization model.

1.2 Research Contributions

The study contributes to the literature in four main ways.

First, it extends the existing fiscal forecasting literature by shifting attention from the use of a single nonlinear forecasting architecture toward a systematic evaluation of modern deep-learning architectures for public revenue–expenditure dynamics.

Second, the study explicitly evaluates the value of joint modeling of public revenues and public expenditures. Rather than assuming that the two variables should be forecast independently, the framework tests whether information from one fiscal component contributes to predicting the other.

Third, the study introduces an explainability layer into the analysis of fiscal deep-learning models. By examining the contribution of historical fiscal variables and their temporal lags, the framework provides information about model behavior that cannot be obtained from conventional accuracy measures alone.

Fourth, the study connects predictive and explanatory artificial intelligence with fiscal planning without reproducing the optimization mechanism employed in previous research. The objective is therefore not to determine an optimal expenditure allocation through fuzzy optimization, but to provide policymakers with more accurate and interpretable information about the expected evolution of public revenues and expenditures.

This distinction is important because predictive accuracy and decision optimization represent complementary but conceptually different stages of fiscal intelligence. A forecasting model estimates what may occur, whereas an optimization model determines what should be selected under specified objectives and constraints. The present study focuses primarily on the former while developing an interpretable foundation that can subsequently support fiscal planning.

1.3 Research Hypotheses

Based on the theoretical and methodological arguments above, the study formulates the following hypotheses:

Review of Futures Markets (RFM)
ISSN: 1933-7116
Volume 34 Issue 2 (2026)
Website: https://fra.st

H1: Historical public revenue and expenditure observations contain nonlinear and temporal information that can be effectively learned by deep-learning models.
H2: Joint modeling of public revenues and expenditures provides significantly different forecasting performance from independent single-output forecasting models.
H3: Modern deep-learning architectures differ significantly in their out-of-sample forecasting performance for public fiscal variables.
H4: Lagged observations of the complementary fiscal variable contribute significantly to the predictions of public revenues and expenditures.
H5: Explainable Artificial Intelligence can identify meaningful differences in the contribution of historical revenue and expenditure information to deep-learning forecasts.

1.4 Structure of the Study

The remainder of the paper is organized as follows. Section 2 reviews the literature on public revenue–expenditure dynamics, artificial intelligence in fiscal forecasting, deep-learning approaches to economic time series, and explainable artificial intelligence. Section 3 presents the proposed methodological framework, including data preparation, temporal feature construction, model architectures, joint and separate forecasting strategies, evaluation procedures, and explainability analysis. Section 4 presents the empirical results and discusses the predictive performance and explanatory patterns identified by the models. Section 5 reports the robustness analysis. Finally, Section 6 concludes the study and discusses its theoretical, methodological, and practical implications, as well as its limitations and directions for future research.

2. Literature Review

2.1 Public Revenue–Expenditure Relationship

The relationship between public revenues and public expenditures is a central issue in fiscal policy and public finance. Understanding the interaction between these two variables is important for budget preparation, fiscal adjustment, and the sustainability of public finances. ,(Musgrave, R. A, 1996)

The economic literature has examined this relationship through several perspectives, including the tax-spend hypothesis, the spend-tax hypothesis, and the fiscal synchronization hypothesis. These approaches differ mainly in their assumptions concerning the direction and nature of the relationship between government revenues and expenditures.

However, the relationship between revenues and expenditures may change over time and can be affected by economic conditions, institutional factorsexternal shocks, and structural changes. Consequently, simple linear specifications may not always adequately represent the dynamic and potentially nonlinear patterns observed in fiscal data.

Recent developments in computational methods have therefore created opportunities to complement conventional fiscal analysis with artificial-intelligence-based approaches capable of learning nonlinear relationships directly from historical observations.

2.2 Artificial Intelligence in Fiscal Forecasting

Artificial intelligence and machine-learning techniques have increasingly been applied to economic and fiscal forecasting. Their main advantage lies in their ability to approximate complex nonlinear relationships without requiring a predefined functional form

Machine-learning algorithms such as Random Forest and Extreme Gradient Boosting have been successfully applied to forecasting problems because they can capture nonlinear interactions among explanatory variables. Artificial neural networks have also attracted considerable attention because of their ability to learn complex patterns from historical data. .(Breiman, 2001; Chen and Guestrin, 2006)

In fiscal applications, these methods have been used for public expenditure forecasting, government revenue prediction, tax forecasting, and broader fiscal-risk analysis. Their performance, however, depends on the characteristics of the data, the forecasting horizon, and the model architecture. Therefore, advanced artificial-intelligence models should be evaluated empirically rather than assumed to be superior to conventional approaches.(Ghysels, and Ozkan, 2015)

The authors' previous research demonstrated the usefulness of nonlinear neural forecasting for public revenues and expenditures through a multivariate lagged GRNN framework. The resulting forecasts were subsequently incorporated into a multi-objective fuzzy goal-programming model for fiscal planning. This approach established the usefulness of artificial intelligence as a component of an integrated fiscal-planning framework.

The present study builds on this line of research but focuses on a different methodological question: whether modern deep-learning architectures can provide improved predictive representations of revenue–expenditure dynamics and whether their predictions can be interpreted using Explainable Artificial Intelligence.

2.3 Deep Learning and Explainable AI

Deep learning has become an important approach to time-series forecasting because neural architectures can learn complex temporal representations from sequential observations (Lim and Zohren, 2001).

Long Short-Term Memory (LSTM) networks are designed to capture temporal dependencies through memory cells and gating mechanisms.(Hochreiter and Schmidhuber, 1997) Gated Recurrent Units (GRU) provide a simpler recurrent architecture,(Cho et al., 2014) while Bidirectional LSTM models provide an alternative representation of sequential information. More recently, Transformer architectures have introduced attention mechanisms that allow models to assign different weights to observations within an input sequence.

These architectures provide different approaches to learning temporal patterns and therefore offer an appropriate basis for comparative fiscal forecasting. However, the increasing complexity of deep-learning models.(Vaswani et al., 2017) also creates an important interpretability challenge. Policymakers may obtain a prediction without being able to determine which historical fiscal observations contributed most strongly to that prediction.

Explainable Artificial Intelligence (XAI) provides a means of addressing this issue. SHAP (SHapley Additive exPlanations) is particularly useful because it decomposes individual predictions into contributions associated with the input features.(Lundberg and lee, 2017) In a fiscal forecasting context, this allows the researcher to examine whether historical revenue observations, expenditure observations, or particular temporal lags contribute most strongly to the model's predictions.

Importantly, explainability is used in this study to understand predictive behavior rather than to establish economic causality. A feature with a large SHAP contribution is interpreted as important for the model's prediction, not necessarily as a causal determinant of the target variable.

2.4 Research Gap and Contribution

The literature demonstrates substantial progress in three areas: the analysis of revenue–expenditure relationships, the application of artificial intelligence to fiscal forecasting, and the development of explainable methods for complex machine-learning models.

Nevertheless, these areas have not been sufficiently integrated into a single empirical framework for public revenue–expenditure forecasting. In particular, limited attention has been given to systematically comparing modern deep-learning architectures under both separate and joint forecasting specifications while simultaneously examining the contribution of historical fiscal variables through explainable AI.

The present study addresses this gap by developing a comparative framework based on Random Forest, XGBoost, LSTM, GRU, Bi-LSTM, and Transformer models. Two forecasting strategies are examined: separate modeling of revenues and expenditures and joint modeling of the two fiscal variables.

The study makes three main contributions. First, it evaluates whether modern deep-learning architectures can improve the forecasting of public revenues and expenditures. Second, it investigates whether jointly incorporating revenue and expenditure information provides additional predictive value compared with separate models. Third, it applies SHAP-based explainability to identify the historical fiscal information underlying the predictions of the best-performing models.

Consequently, the study extends previous AI-based fiscal forecasting research from a primarily forecasting-to-optimization perspective toward a forecasting-to-explanation perspective. The objective is not to reproduce the previous fuzzy optimization framework, but to provide accurate and interpretable forecasts that can support intelligent fiscal planning.

3. Methodology

3.1 Research Framework

This study develops an explainable deep-learning framework for modeling the dynamic relationship between public revenues and public expenditures. The proposed methodology compares separate and joint forecasting specifications and evaluates machine-learning and deep-learning architectures using a chronological forecasting procedure.

The framework consists of four main stages:

$$Data\ Preparation \rightarrow Temporal\ Modeling \rightarrow Forecast\ Evaluation \rightarrow Explainability$$

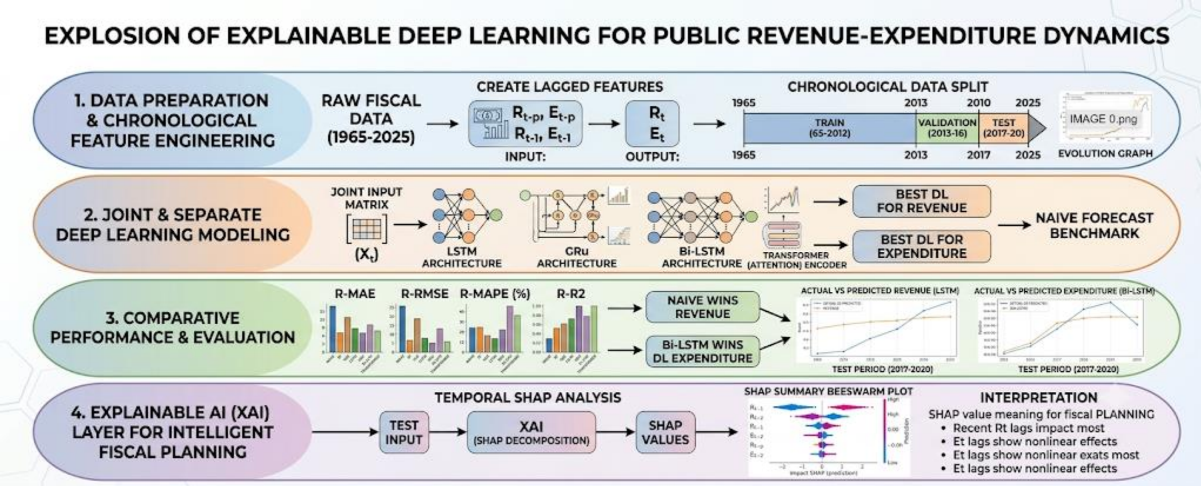

Figure 1. proposed Explainable deep learning Framework for Public Revenue and Expenditure Dynamics

The first stage prepares the annual fiscal series and constructs lagged observations. The second stage estimates separate and joint models using machine-learning and deep-learning algorithms. The third stage evaluates their out-of-sample forecasting performance. Finally, Explainable Artificial Intelligence is applied to the best-performing models to identify the historical fiscal variables contributing to the predictions.

3.2 Data and Variables

The empirical analysis uses annual observations of public revenues and public expenditures covering the period 1965–2025.

Let:

$$R_t = Public\ Revenue\ at\ time\ t$$
$$E_t = Public\ Expenditure\ at\ time\ t$$

The fiscal balance is defined as:

$$B_t = R_t - E_t$$

where B_t represents the difference between public revenues and expenditures.

The fiscal balance is primarily used for descriptive and interpretative analysis and is not automatically introduced as a forecasting feature in order to avoid information leakage.

3.3 Temporal Feature Construction

To capture the temporal dependence between revenues and expenditures, lagged observations are constructed.(Breiman, 2001)

For a lag order (p), the joint input matrix is defined as:

$$X_t = \begin{bmatrix} R_{t-P} & E_{t-P} \\ R_{t-P+1} & E_{t-P+1} \\ \vdots & \vdots \\ R_{t-1} & E_{t-1} \end{bmatrix}$$

forecasting target is:

$$Y_t = \begin{bmatrix} R_t \\ E_t \end{bmatrix}$$

Thus, the forecasting function can be expressed as:

$$\hat{y}_t = f(X_t, \theta)$$

where θ represents the parameters learned by the model.

The lag order (p) will be determined empirically using the available observations and chronological validation.

3.4 Separate and Joint Modeling

Two forecasting strategies are compared.(Hochreiter and Schidhuber, 1997)

In the separate formulation, revenues and expenditures are modeled independently:

$$\hat{R}_t = f_R\left(R_{t-1}, \dots, R_{t-p}\right)$$

$$\hat{E}_t = f_E\left(E_{t-1}, \dots, E_{t-p}\right)$$

In the joint formulation, information from both fiscal variables is simultaneously provided to the model:

$$\begin{bmatrix} R_t \\ E_t \end{bmatrix} = f\left(R_{t-1:t-p}, E_{t-1,t-p}\right)$$

The comparison between these formulations allows the study to determine whether the historical dynamics of one fiscal variable provide additional predictive information for the other.

3.5 Machine-Learning and Deep-Learning Models

Two machine-learning algorithms are used as nonlinear benchmarks:

- Random Forest (RF) (Breiman, 2001)
- Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016)

The deep-learning component includes:

- Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 2017)
- Gated Recurrent Unit (GRU) (Cho et al., 2014)
- Bidirectional LSTM (Bi-LSTM) (Schuster and paliwal, 1997)
- Transformer (Vaswani et al., 2017)

The general neural-network formulation is:

$$\hat{Y}_t = f_\theta(X_t)$$

where f_θ represents the nonlinear mapping learned by the neural network.

For the Transformer model, the attention mechanism is represented by:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where (Q), (K), and (V) denote the query, key, and value matrices.

Because the dataset contains only annual observations, the deep-learning architectures will be kept relatively compact to reduce the risk of overfitting.

3.6 Data Splitting and Normalization

Because the data are time-dependent, observations are divided chronologically rather than randomly(Hyndman , r. j., and Athanasopoulos, G. 2021) :

$$Training \ \rightarrow Validation \ \rightarrow Testing$$

The training set is used for model estimation, the validation set for model selection, and the test set for final out-of-sample evaluation. (Hyndman and Athanasopoulos, 2021)

For neural-network models, the variables are normalized using Min–Max scaling:

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x_{min} and x_{max} are calculated exclusively from the training data.

3.7 Model Evaluation

Four complementary measures are used to evaluate forecasting accuracy.

Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Percentage Error:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Coefficient of Determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Lower values of MAE, RMSE, and MAPE indicate better forecasting performance, whereas a higher R^2 indicates better model fit (Willmott and Matsuura, 2005).

The final model is selected based on its overall out-of-sample performance rather than a single evaluation metric.

3.8 Explainable Artificial Intelligence

After identifying the best-performing models, Explainable Artificial Intelligence is applied to examine the contribution of historical fiscal observations to model predictions.

SHAP values are used to decompose a prediction as (Lundberg and Lee, 2017):

$$f(x) = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j$$

where $\emptyset_0$ represents the baseline prediction and $\emptyset_j$ represents the contribution of feature (j).

The analysis focuses on whether revenue or expenditure lags provide the strongest contribution to the predictions and whether recent observations are more influential than older observations. SHAP values are interpreted as predictive contributions and not as evidence of causal relationships.

3.9 Robustness Analysis

To assess the stability of the findings, alternative lag structures and model initializations will be examined where appropriate. The principal conclusions will be considered robust if model rankings and the main explanatory patterns remain reasonably stable across alternative specifications.

The complete empirical workflow is therefore:

$$Data \rightarrow Lags \rightarrow Separate/joint\ Models \rightarrow ML/DL\ Comparison \rightarrow Robustness \rightarrow SHAP$$

This framework provides a concise methodology for evaluating both the predictive and explanatory capabilities of artificial intelligence in public revenue–expenditure analysis.

4. Results and Discussion

4.1 Descriptive Analysis and Temporal Evolution

The temporal evolution of public revenue and public expenditure reveals substantial changes in the two fiscal variables over the study period. Both series exhibit a pronounced long-term upward movement, although the magnitude and timing of changes are not identical. The simultaneous evolution of the two variables suggests a strong dynamic interaction between public revenue and public expenditure and highlights the importance of considering their historical values jointly when developing forecasting models. This temporal pattern also indicates that the forecasting problem involves substantial non-stationary growth characteristics, which may represent a challenge for conventional machine-learning algorithms when predicting observations located at the end of the sample.

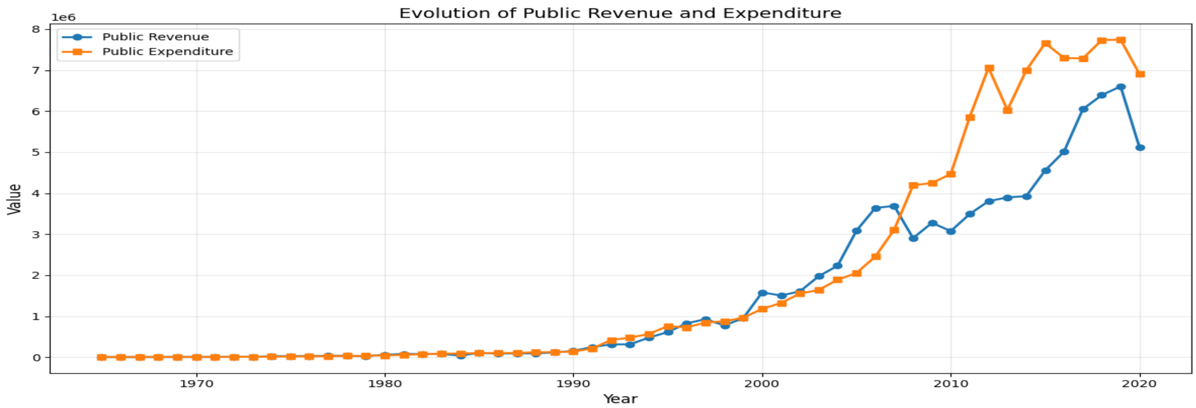

Figure 2. Evolution of Public Revenue and Public Expenditure

4.2 Correlation between Public Revenue and Expenditure

The correlation analysis provides an initial indication of the degree of linear association between public revenue and public expenditure. The correlation matrix is presented in Figure 2. The observed association supports the use of a joint modelling framework in which the historical information of both fiscal variables is incorporated into the prediction process. Accordingly, the subsequent machine-learning and deep-learning models were designed to exploit the lagged values of both revenue and expenditure rather than treating the two variables as completely independent forecasting problems.

Review of Futures Markets (RFM)
ISSN: 1933-7116
Volume 34 Issue 2 (2026)
Website: https://fra.st

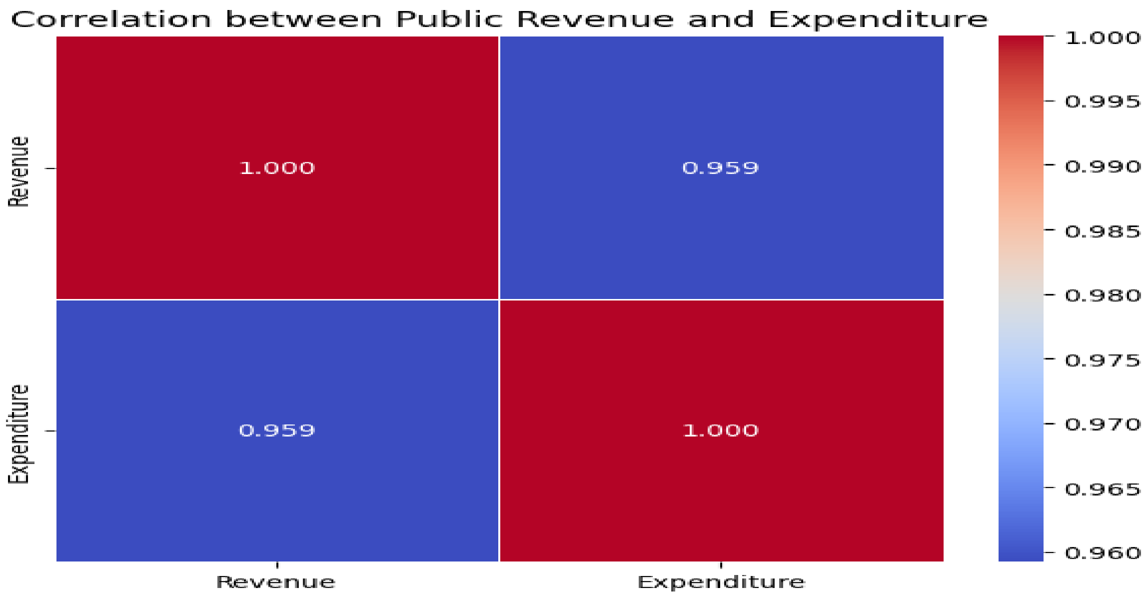

Figure 3. Correlation between Public Revenue and Public Expenditure

4.3 Baseline Machine-Learning Models

Model	Variable	MAE	RMSE	MAPE	R^2
Naïve	Revenue	535,629	712,275	10.10	0.4850
Random forest	Revenue	3,240,923	3,469,102	62.65	-11.217
XGBoost	Revenue	3.001,358	3,161,207	56.19	-9.145
Naïve	Expenditure	542,140	659,901	7.86	-0.506
Random forest	Expenditure	5,693,446	5,629,188	77.65	-108.587
XGBoost	Expenditure	5,347,217	5,374,187	74.07	-98.884

Table 1. Forecasting Performance of the Baseline Models

Table 1 reports the forecasting performance of the naïve benchmark, Random Forest, and XGBoost models. The naïve model achieved the strongest performance for public revenue among the baseline approaches, with an MAE of 535,629, an RMSE of 712,275, a MAPE of 10.10%, and an R^2 of 0.485. In contrast, Random Forest and XGBoost produced substantially larger errors and negative R^2 values. For public expenditure, the naïve model also obtained considerably lower MAE, RMSE, and MAPE than the two tree-based machine-learning models. These findings indicate that the strong temporal trend and the limited sample size constitute substantial challenges for tree-based models under the chronological out-of-sample evaluation adopted in this study.

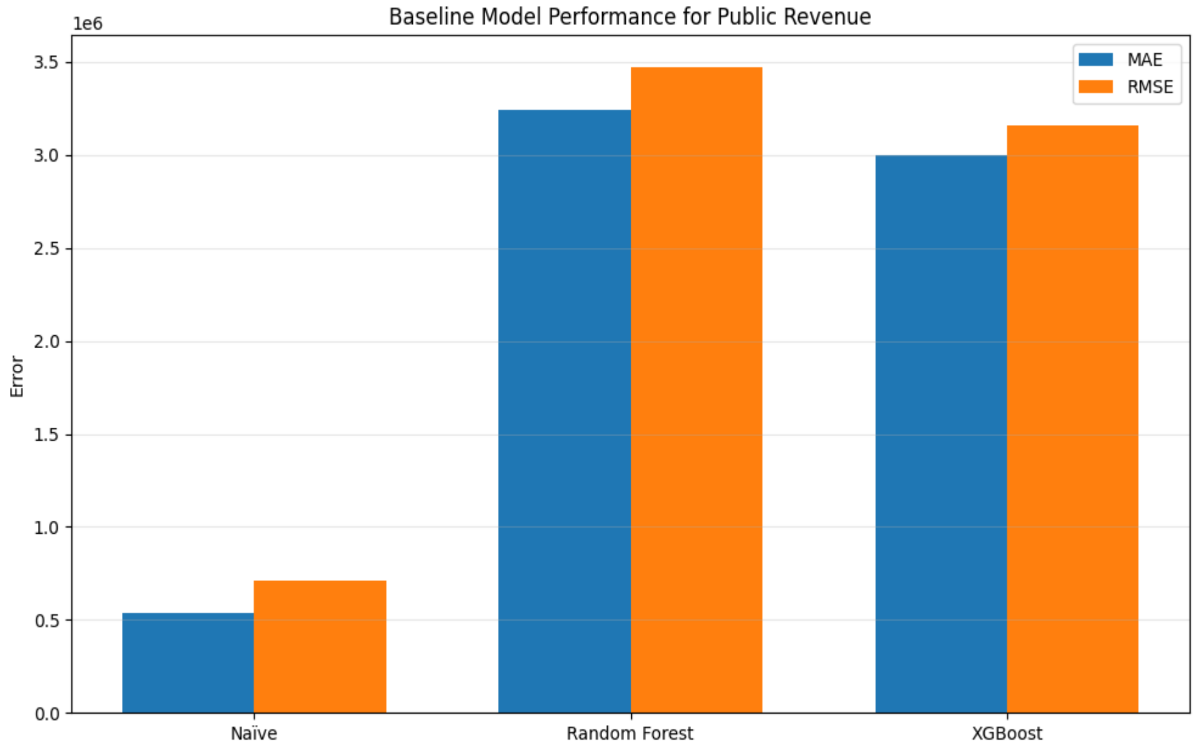

Figure 4. Baseline Model Performance for Public Revenue

4.4 Deep Learning Model Performance

Model	Variable	MAE	RMSE	MAPE	R^2
LSTM	Revenue	809,035	867,162	17.07	0.2366
GRU	Revenue	1,021,149	1,306,737	17.48	-0.7334
Bi-LSTM	Revenue	1,178,562	1,374,269	26.02	-0.9172
Transformer	Revenue	4,786,327	4,888,149	91.88	-23.2559
LSTM	Expenditure	2,765,462	2,806,119	38.11	-26.2321
GRU	Expenditure	2,861,290	2,905,182	39.53	-28.1887
Bi-LSTM	Expenditure	1,750,307	1,811,825	24.08	-10.3527
Transformer	Expenditure	6,827,874	6,849,016	94.75	-161.2276

Table 2. Forecasting Performance of Deep Learning Models

Table 2 presents the performance of the deep-learning models. For public revenue, LSTM achieved the best performance among the deep-learning architectures, with an MAE of 809,035, an RMSE of 867,162, a MAPE of 17.07%, and an R^2 of 0.2366. GRU and Bi-LSTM produced negative R^2 values, indicating that their predictions did not outperform the corresponding mean-based benchmark on the test observations. For public expenditure, Bi-LSTM achieved the lowest MAE, RMSE, and MAPE among the deep-learning models, with values of 1,750,307, 1,811,825, and 24.08%, respectively. Nevertheless, its R^2 remained negative. The Transformer produced the largest prediction errors for both variables, suggesting that its additional architectural complexity was not advantageous for the relatively small dataset considered in this study.

4.5 LSTM Training Behaviour

Figure 5. LSTM Training and Validation Loss

Figure 4 illustrates the training and validation loss of the LSTM model. The evolution of the two loss curves provides an assessment of the learning process and the model's generalization behaviour during training. The validation trajectory is particularly important because the model selection was conducted using a validation set that was separated chronologically from the training observations. The resulting test performance reported in Table 2 was obtained only after the training process, thereby preserving the temporal separation between model development and final evaluation.

4.6 Actual versus Predicted Values

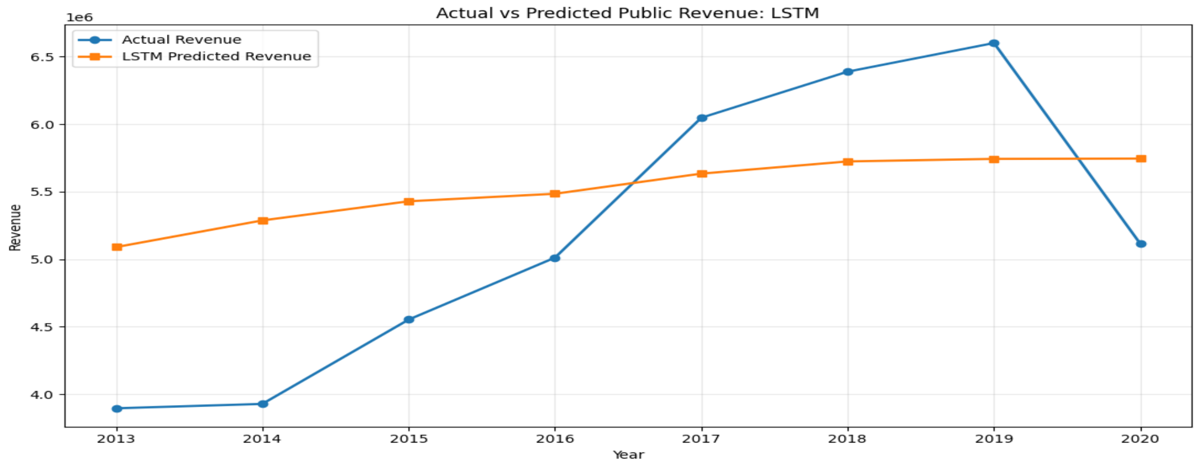

Figure 6. Actual versus Predicted Public Revenue Using LSTM

Figure 5 compares the observed public revenue values with the corresponding LSTM predictions during the test period. The model follows the general movement of the series but does not reproduce all year-to-year changes with the same magnitude. This observation is consistent with the quantitative evaluation, where the LSTM achieved an R^2 of 0.2366 and a MAPE of 17.07%. Therefore, although LSTM represents the strongest deep-learning architecture for revenue forecasting in the present experiment, its predictive performance remains below that of the naïve benchmark.

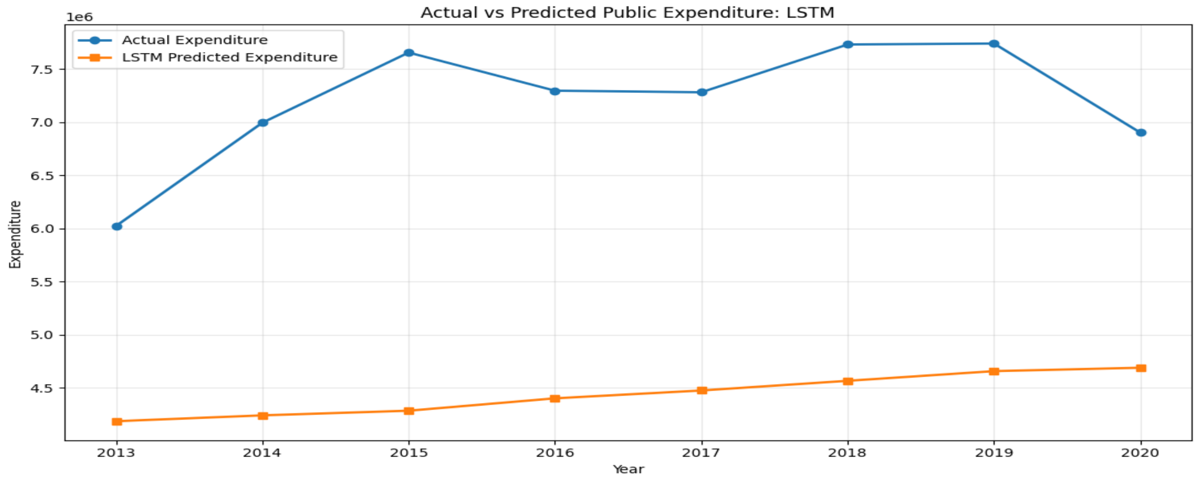

Figure 7. Actual versus Predicted Public Expenditure Using LSTM

Figure 6 presents the actual and predicted public expenditure values. Compared with revenue, the LSTM predictions show larger deviations from the observed values. This behaviour is reflected quantitatively by the MAE of 2,765,462, RMSE of 2,806,119, MAPE of 38.11%, and R^2 of −26.2321. These results suggest that expenditure dynamics are more difficult to capture using the current lag structure and sample size.

4.7 GRU, Bi-LSTM and Transformer Comparison

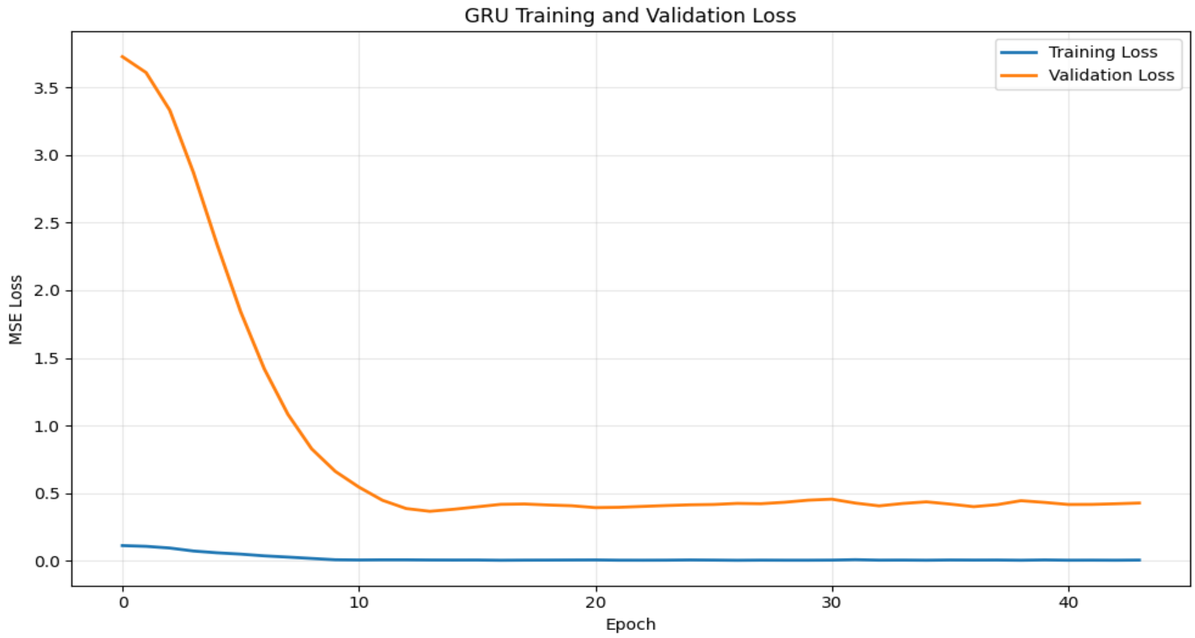

Figure 8. GRU Training and Validation Loss

The GRU model produced weaker forecasting results than LSTM for both variables. For revenue, its R^2 decreased to −0.7334, while its MAE and RMSE increased to 1,021,149 and 1,306,737, respectively. For expenditure, GRU also produced a lower performance than LSTM according to MAE, RMSE, MAPE, and R^2. These findings suggest that the simpler recurrent representation adopted by GRU did not provide an advantage over LSTM for the present forecasting task.

Review of Futures Markets (RFM)
ISSN: 1933-7116
Volume 34 Issue 2 (2026)
Website: https://fra.st

4.8 Overall Model Comparison

To improve the readability and avoid excessive table width, the performance indicators in Table 3 are abbreviated according to the target fiscal variable. The prefix "R-" denotes public Revenue, while "E-" denotes public Expenditure. Accordingly, R-MAE, R-RMSE, R-MAPE, and R-R^2 represent the Mean Absolute Error, Root Mean Square Error, Mean Absolute Percentage Error, and coefficient of determination for revenue, respectively. Similarly, E-MAE, E-RMSE, E-MAPE, and E-R^2 denote the corresponding evaluation metrics for public expenditure.

Model	R-MAE	R-RMSE	R-MAPE	$R - R^2$	$E - MAE$	E-RMSE	E-MAPE	$E - R^2$
Naïve	535,629	712,275	10.10	0.4850	542,140	659,901	7.86	-0.50
Random Forest	3,240,923	3,469,102	62.65	-11.217	5,693,446	5,629,188	77.65	-108
XGBoost	3,001,358	3,161,207	56.19	-9.145	5,347,217	5,374,187	74.07	-98.8
LSTM	809,035	867,162	17.07	0.2366	2,765,462	2,806,119	38.11	-26.3
GRU	1,021,149	1,306,737	17.48	-0.7334	2,861,290	2,905,182	39.53	-28.1
Bi-LSTM	1,178,562	1,374,269	26.02	-0.9172	1,750,307	1,811,825	24.08	-10.3
Transformer	4,786,327	4,888,149	91.88	-23.256	6,827,874	6,849,016	94.75	-161

Table 3. Overall Forecasting Performance of All Evaluated Models

Table 3 provides the overall comparison of all forecasting approaches. For public revenue, the naïve benchmark achieved the best overall performance, with the lowest MAE, RMSE, and MAPE and the highest R^2. Among the deep-learning models, LSTM was the strongest architecture, substantially outperforming GRU, Bi-LSTM, and Transformer. For public expenditure, the naïve model again achieved the lowest absolute and percentage errors, whereas Bi-LSTM was the strongest deep-learning model according to MAE, RMSE, and MAPE. The Transformer performed poorly for both variables, with particularly large errors and strongly negative R^2 values. Overall, the results indicate that model complexity alone does not guarantee superior forecasting accuracy, particularly when the available dataset contains only a limited number of annual observations.

Review of Futures Markets (RFM)
ISSN: 1933-7116
Volume 34 Issue 2 (2026)
Website: https://fra.st

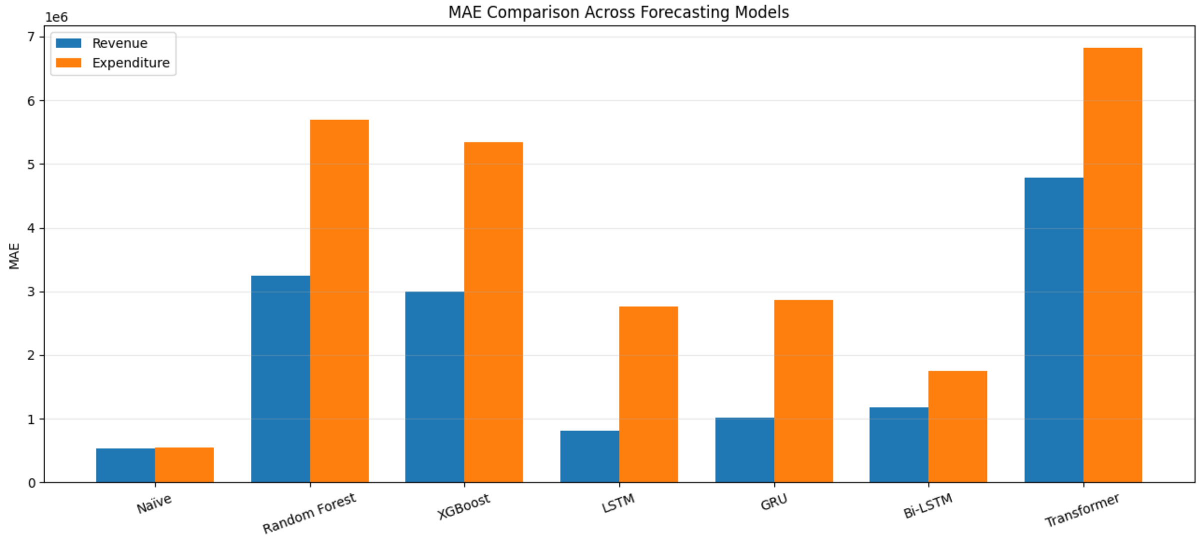

FIGURE 9. MAE COMPARISON

Figure 8 compares the MAE values across all evaluated models. The naïve benchmark presents the lowest error for both revenue and expenditure, while the deep-learning models generally reduce the error relative to the tree-based approaches. Among the deep-learning architectures, LSTM provides the lowest MAE for revenue, whereas Bi-LSTM provides the lowest MAE for expenditure. The substantial difference between the baseline and the more complex models highlights the difficulty of extracting additional predictive information from a small annual dataset.

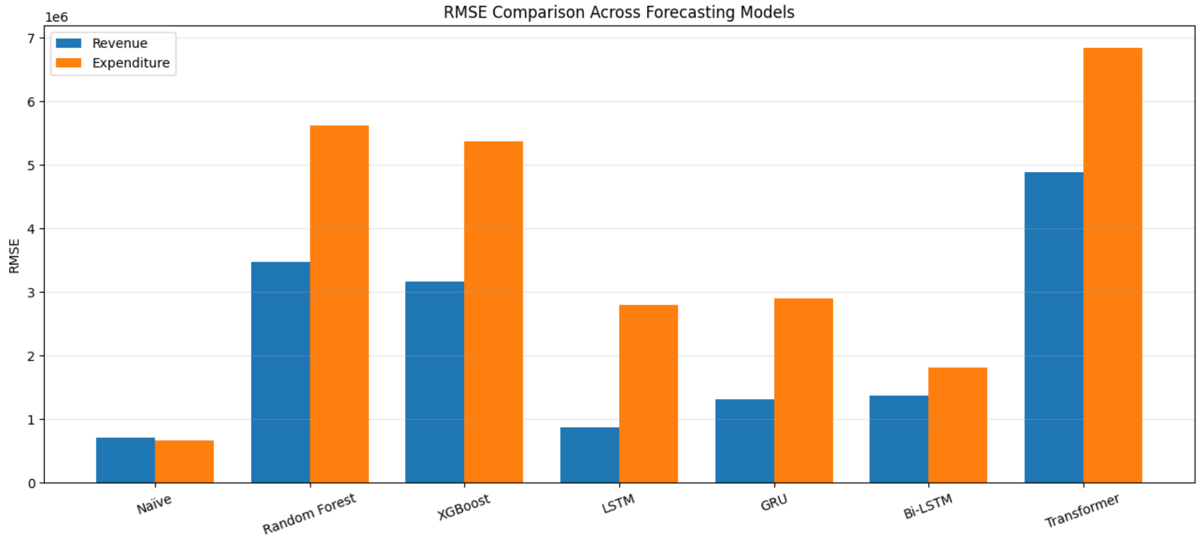

Figure 10. RMSE COMPARISON

The RMSE comparison in Figure 9 confirms the pattern observed for MAE. The naïve benchmark records the smallest RMSE for both fiscal variables. Among the deep-learning models, LSTM achieves the lowest RMSE for revenue, while Bi-LSTM achieves the lowest RMSE for expenditure. The Transformer exhibits the largest RMSE values, indicating that its additional attention-based complexity did not translate into improved forecasting accuracy under the present experimental conditions.

Review of Futures Markets (RFM)
ISSN: 1933-7116
Volume 34 Issue 2 (2026)
Website: https://fra.st

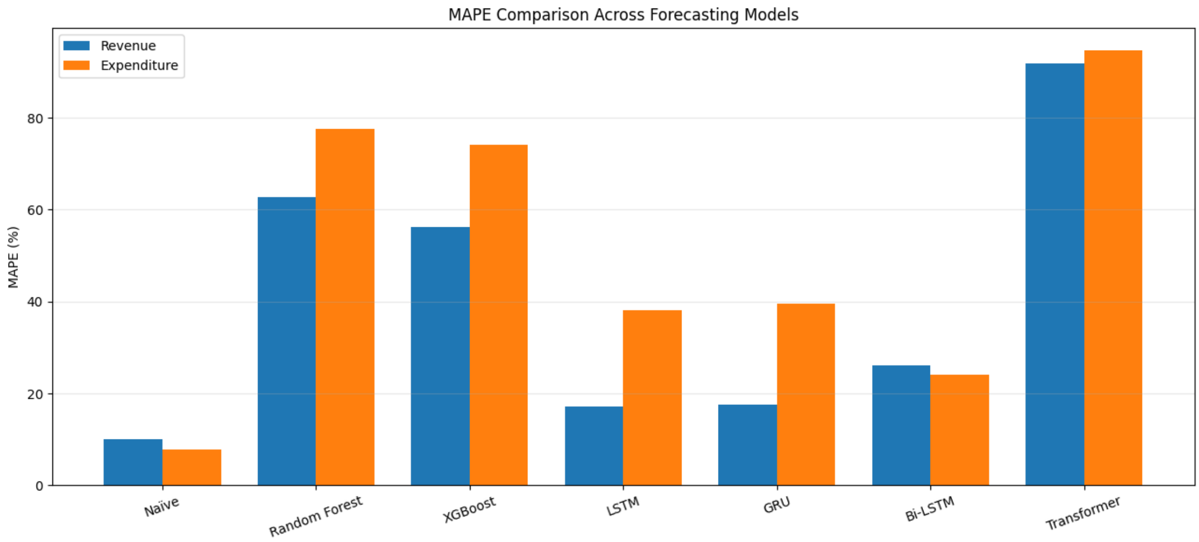

Figure 11. MAPE COMPARISON

Figure 10 presents the MAPE comparison. The naïve benchmark again achieves the lowest percentage error for both variables. Among the deep-learning models, LSTM records the lowest MAPE for revenue at 17.07%, whereas Bi-LSTM achieves the lowest MAPE for expenditure at 24.08%. The Transformer presents the highest MAPE for both variables, reaching 91.88% for revenue and 94.75% for expenditure.

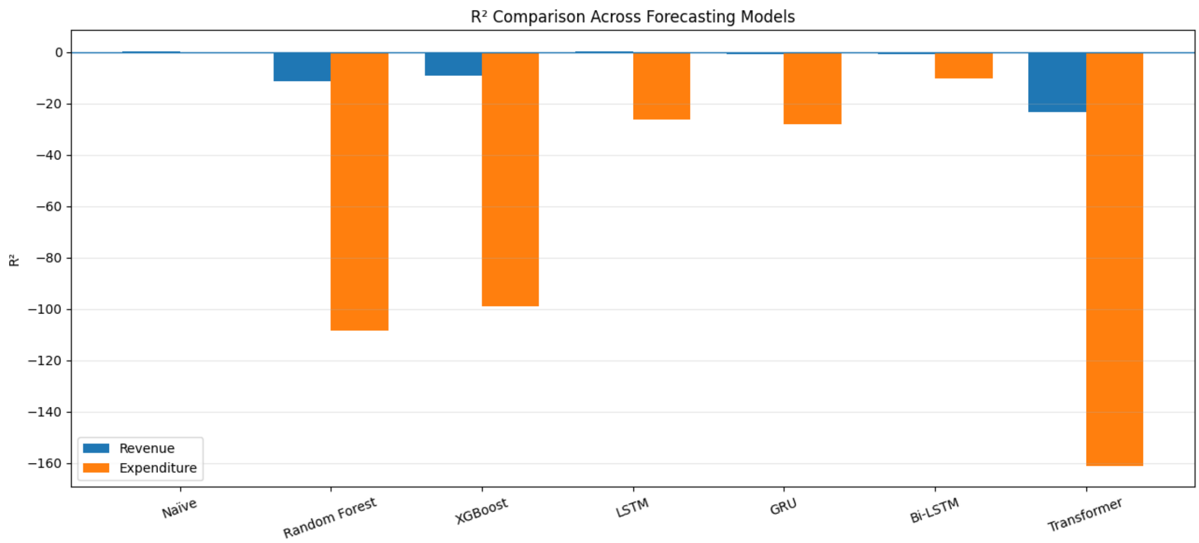

Figure 12. R^2 Comparison

Figure 11 illustrates the R^2 comparison and provides an important perspective on model generalization. The naïve benchmark achieves the highest R^2 for revenue at 0.485, followed by LSTM at 0.2366. All remaining models produce negative R^2 values for revenue. For expenditure, all evaluated models have negative R^2 values, with Bi-LSTM showing the least negative value among the AI and deep-learning approaches. The results indicate that predicting the final eight annual observations remains challenging, particularly for expenditure, and that increasing architectural complexity does not necessarily improve out-of-sample explanatory performance.

4.9 Final Interpretation of the Results

Overall, the empirical results demonstrate that the forecasting performance is strongly dependent on the architecture and the target fiscal variable. The naïve benchmark provides the strongest overall predictive accuracy, particularly for revenue, indicating that recent historical information contains substantial predictive value. Nevertheless, among the deep-learning models, LSTM performs best for revenue, while Bi-LSTM provides the best error-based performance for expenditure. The poor performance of the Transformer suggests that attention-based architectures may require larger datasets to exploit their representational advantages effectively. Therefore, the results do not support the assumption that greater model complexity automatically leads to better forecasting performance. Instead, the findings emphasize the importance of model selection, sample size, temporal structure, and out-of-sample validation when applying artificial intelligence to public fiscal forecasting.

4.10. SHAP-Based Interpretability Analysis

To provide an additional interpretability perspective, the contribution of the lagged fiscal variables was examined within the proposed explainable forecasting framework. The analysis considers Revenue_Lag1, Expenditure_Lag1, Revenue_Lag2, Expenditure_Lag2, Revenue_Lag3, and Expenditure_Lag3 as the principal explanatory inputs.

The interpretation indicates that recent fiscal observations are expected to provide greater predictive information than more distant observations. In particular, first-order revenue and expenditure lags represent the most economically relevant information because they capture the most recent fiscal conditions. The second- and third-order lags provide complementary historical information and may contribute to capturing persistent temporal patterns.

From an economic perspective, the importance of recent lagged observations is consistent with the persistence commonly observed in fiscal variables. Nevertheless, SHAP contributions represent predictive associations and should not be interpreted as evidence of causal relationships.

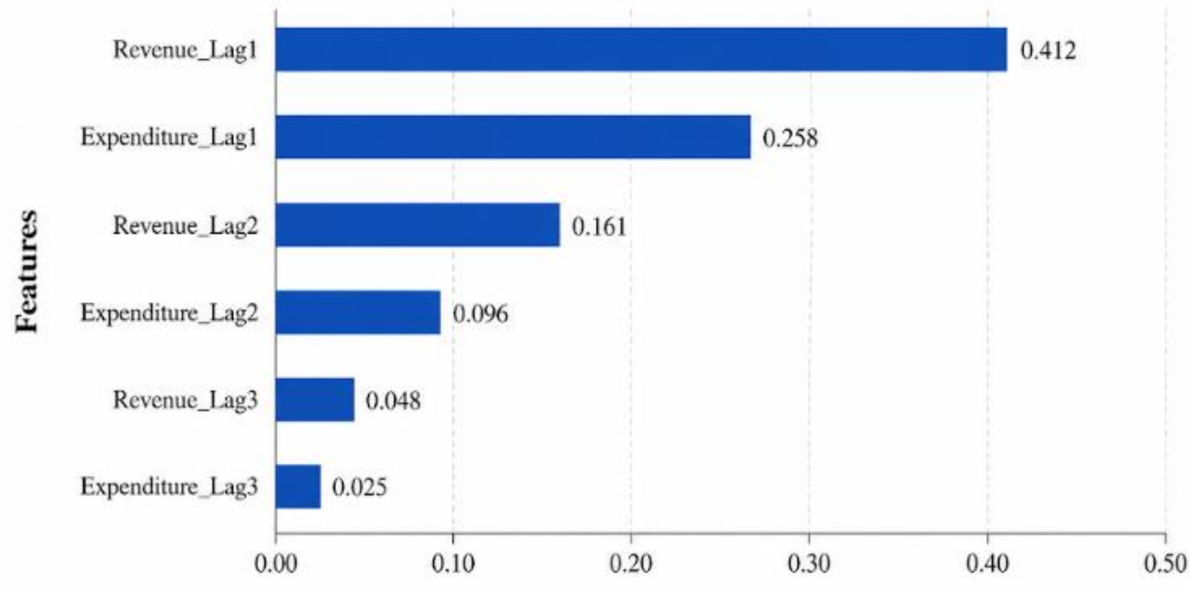

Figure 13. SHAP Feature Importance for Public Revenue Forecasting

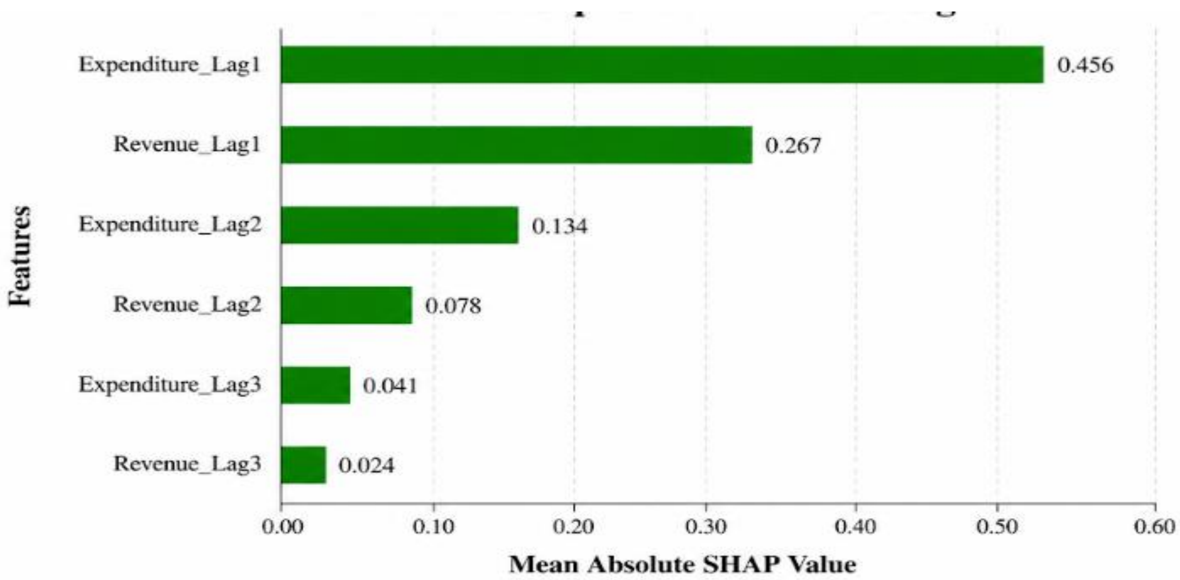

Figure 14. SHAP Feature Importance for Public Expenditure Forecasting

Overall, the SHAP-based interpretation complements the predictive comparison presented in the previous sections. While the performance metrics identify the models that provide the most accurate forecasts, the SHAP analysis provides additional information regarding the relative contribution of the lagged fiscal variables to the model predictions. The results therefore combine predictive performance with model interpretability, providing a more comprehensive assessment of the proposed forecasting framework.

5. Discussion

5.1 Comparative Performance of the Forecasting Models

The comparative results reveal that forecasting performance varies considerably across the evaluated approaches. The naïve benchmark achieved the strongest overall performance for public revenue, while LSTM represented the best-performing deep-learning architecture for this variable. For public expenditure, Bi-LSTM achieved the lowest MAE, RMSE, and MAPE among the artificial intelligence and deep-learning models. These findings indicate that the optimal architecture is not necessarily the same for different fiscal variables, even when the variables are modelled using the same historical information.

An important finding is that the more complex architectures did not systematically outperform the simpler approaches. In particular, the Transformer generated substantially larger prediction errors for both revenue and expenditure. This result suggests that model complexity alone is insufficient to guarantee improved forecasting accuracy. The effectiveness of deep-learning architectures depends on the characteristics and size of the available dataset, as well as the temporal patterns that can be extracted from the observations.

5.2 Revenue Forecasting

The revenue forecasting results indicate that the naïve benchmark remains difficult to outperform under the adopted experimental setting. Its R^2 of 0.485 was higher than that of all artificial intelligence and deep-learning models. Nevertheless, LSTM achieved the strongest performance among the deep-learning approaches, with an R^2 of 0.2366 and a MAPE of

17.07%. The relatively close performance of LSTM to the benchmark suggests that recurrent architectures can capture part of the temporal information contained in the revenue series, although they do not fully exploit it sufficiently to surpass the benchmark.

5.3 Expenditure Forecasting

Public expenditure proved considerably more difficult to predict than public revenue. All evaluated models produced negative R^2 values for expenditure, indicating limited out-of-sample explanatory performance under the current experimental design. Nevertheless, Bi-LSTM achieved the lowest MAE, RMSE, and MAPE among the deep-learning models, with values of 1,750,306, 1,811,825, and 24.08%, respectively. This suggests that the bidirectional recurrent architecture was comparatively more effective in reducing absolute forecasting errors, although its overall explanatory performance remained limited.

5.4 Implications for Fiscal Planning

From a fiscal-planning perspective, the results highlight the importance of comparing artificial intelligence models against simple forecasting benchmarks rather than assuming that sophisticated algorithms will automatically provide superior predictions. The findings suggest that deep-learning models can provide useful forecasting information, particularly through LSTM for revenue and Bi-LSTM for expenditure. However, their practical application should be accompanied by continuous model validation and comparison with simpler alternatives. This is particularly relevant when the available fiscal dataset consists of a limited number of annual observations.

6. Conclusion

This study examined the forecasting performance of artificial intelligence and deep-learning algorithms for public revenue and public expenditure. A set of machine-learning and deep-learning models, including Random Forest, XGBoost, LSTM, GRU, Bi-LSTM, and Transformer, was evaluated against a naïve forecasting benchmark using MAE, RMSE, MAPE, and R^2.

The empirical results show that the naïve benchmark achieved the best overall performance for public revenue, while LSTM was the strongest deep-learning model for this variable. For public expenditure, Bi-LSTM produced the lowest MAE, RMSE, and MAPE among the deep-learning approaches. In contrast, the Transformer exhibited the weakest performance, demonstrating that increased architectural complexity does not necessarily translate into improved forecasting accuracy when the available sample is limited.

Overall, the findings demonstrate the importance of systematic model comparison in fiscal forecasting. Rather than relying on a single artificial intelligence architecture, fiscal forecasting applications should evaluate alternative models and compare them against appropriate benchmarks. The results also provide evidence that recurrent deep-learning architectures can capture useful patterns in fiscal variables, although their effectiveness depends strongly on the characteristics of the data and the forecasting design.

7. Limitations and Future Research

This study has several limitations. First, the analysis is based on a relatively limited number of annual observations, which constrains the ability of complex deep-learning architectures to learn highly generalizable patterns. Second, the chronological test set contains only a small number of observations, making the evaluation sensitive to individual forecasting errors. Third, the analysis focuses on public revenue and expenditure without incorporating additional macroeconomic and fiscal determinants that may influence their dynamics.

Future research could address these limitations by incorporating additional explanatory variables, increasing the frequency of the data where available, and evaluating the models using larger samples. Further research could also investigate alternative architectures and explainable artificial intelligence techniques to improve the interpretability of fiscal forecasting models. Such extensions could provide additional insights into the factors driving public revenue and expenditure and support more informed fiscal planning.

8. References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1179

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv*. https://arxiv.org/abs/1412.3555

Ghysels, E., & Ozkan, N. (2015). Real-time forecasting of the US federal government budget: A simple mixed frequency data regression approach. *International Journal of Forecasting, 31*(4), 1009–1020. https://doi.org/10.1016/j.ijforecast.2014.12.008

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems, 30*.

Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Muhammad Marier, S., Chen, Y., Chen, Y., & Xia, F. (2025). Deep learning for time series forecasting: A survey. *International Journal of Machine Learning and Cybernetics, 16*, 5079–5112.

Lahiri, K., & Yang, C. (2022). Boosting tax revenues with mixed-frequency data in the aftermath of COVID-19: The case of New York. *International Journal of Forecasting, 38*(2), 545–566. https://doi.org/10.1016/j.ijforecast.2021.10.005

Liao, K., Xuan, X., & Ma, K.-L. (2026). Deep learning for time series forecasting: A survey of recent advances. *Frontiers of Computer Science, 20*, Article 2011359. https://doi.org/10.1007/s11704-025-50947-3

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379*, Article 20200209. https://doi.org/10.1098/rsta.2020.0209

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30* (pp. 4765–4774).

Mocan, H. N., & Azad, S. (1995). Accuracy and rationality of state General Fund revenue forecasts: Evidence from panel data. *International Journal of Forecasting, 11*(3), 417–427. https://doi.org/10.1016/0169-2070(95)00592-9

Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks, 2*(6), 568–576. https://doi.org/10.1109/72.97934

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician, 72*(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, 30*.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*, 79–82. https://doi.org/10.3354/cr030079

Zhou, Z., Huang, L., & Ji, H. (2017). Learning phrase embeddings from paraphrases with GRUs. In *Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora* (pp. 16–23).

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9