Forecasting the South African Unemployment Rate: A Comparative Analysis of ARIMAX and LSTM Models

Israel Maingo * and Leonard Marevhula

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0965.v1

Keywords: ARIMAX; LSTM; machine learning; time series forecasting; unemployment rate

 doi:10.20944/preprints202604.0965.v1

Article

Forecasting the South African Unemployment Rate: A Comparative Analysis of ARIMAX and LSTM Models

Israel Maingo * and Leonard Marevhula

Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou 0950, Limpopo, South Africa; leonard.marevhula@univen.ac.za

* Correspondence: 21012713@mvula.univen.ac.za; Tel.: +27-72-454-1021

Abstract

This study looks into the predictive performance of linear econometric and deep learning methodologies for the South African unemployment rate quarterly data. In this paper, the Autoregressive Integrated Moving Average with exogenous variables (ARIMAX) model was compared to the Long Short-Term Memory (LSTM) network using unemployment rate quarterly data. Exploratory Data Analysis (EDA) suggested that the unemployment rate series is non-stationary, with structural breaks around 2020 and time-varying volatility. Stationarity tests established the need for differencing, whereas diagnostic tests revealed the presence of autocorrelation and ARCH effects in the raw data. The ARIMAX model added labour market covariates, and the differenced Not Economically Active (*NEA*) variable was statistically significant, whereas *Discouraged* workers were not. Although the ARIMAX model provided a good in-sample fit, residual diagnostics showed deviations from normality. Out-of-sample forecast study revealed moderate predictive accuracy, with relatively substantial forecast errors and increasing prediction intervals over time. In contrast, the LSTM model showed significant learning capacity, with early convergence and well-behaved residuals that meet both independence and homoskedasticity criteria. The model achieved significantly lower forecast errors, with RMSE, MAE, and MAPE values much lower than those of the ARIMAX model. Comparative forecast analysis using Diebold-Mariano (DM) test and model confidence Set (MCS) method and bootstrap confidence intervals consistently demonstrated the statistical superiority of the LSTM model. The findings give strong evidence that the LSTM model outperformed the ARIMAX model for projecting South African unemployment rate. The findings emphasise the importance of nonlinear modelling approaches in capturing complex labour market dynamics while also demonstrating the limitations of classic linear models. These findings also emphasise the importance of using nonlinear machine learning algorithms in macroeconomic forecasting.

Keywords: ARIMAX; LSTM; machine learning; time series forecasting; unemployment rate

1. Introduction

1.1. Overview

Unemployment rate is still an important macroeconomic issue, more especially in emerging nations such as South Africa where labour market imbalances prevent development in the economy and stability in the society. Precise unemployment rates forecasting is essential for both policy development, labour market planning, and macroeconomic management. Adjusting unemployment dynamics is challenging because of structural changes, nonlinear interactions, and external shocks like financial crises and the COVID-19 pandemic. Economic time series can be unstable and uncertain, making it harder to make accurate predictions. Traditional econometric models, especially the Autoregressive Integrated Moving Average (ARIMA) model established by [1], are commonly employed for time

 doi:10.20944/preprints202604.0965.v1

series prediction because of their strong theoretical background and interpretability. The ARIMAX model and other extensions allow for the integration of macroeconomic information in forecasting.

Recent advances in machine learning have introduced powerful alternatives for forecasting complex time series, with Long Short-Term Memory (LSTM) networks demonstrating strong capability in capturing nonlinear dynamics and long-term dependencies [2]. Empirical evidence from financial and economic forecasting shows that LSTM and related models often outperform traditional econometric approaches, particularly in complex and data-rich environments [3–5]. In unemployment forecasting, studies incorporating macroeconomic and real-time indicators have improved predictive accuracy [6,7], while recent machine learning applications provide further evidence of superior performance. For example, ref. [8–11] showed that LSTM and hybrid models consistently outperform traditional linear models, particularly in capturing nonlinear patterns and structural changes such as the COVID-19 shock. These findings highlight the limitations of linear models and the advantages of machine learning approaches in modelling unemployment dynamics. While some research has looked at unemployment forecasting in South Africa using machine learning techniques or conventional econometric models [12,13], these approaches are frequently used separately and have a narrow focus. The increasing interest in comparative modelling frameworks across various economic contexts is demonstrated by the numerous international research that have also looked at unemployment predictions using both econometric and machine learning methodologies. Specifically, current research mostly uses univariate frameworks and does not offer thorough comparisons between linear and nonlinear models in a single empirical context. Furthermore, less is known about the function of labour market-specific factors in the South African setting, such as the Not Economically Active (*NEA*) population.

In light of this, there is a significant gap in the research about how well linear and nonlinear models predict unemployment, especially in the context of South Africa. Predictive accuracy is lower during times of economic upheaval because traditional models frequently fail to account for nonlinear dynamics, structural discontinuities, and time-varying volatility. In order to close this gap, this study uses unemployment data from South Africa to compare the predicting ability of the Long Short-Term Memory (LSTM) network and the ARIMAX model.

1.2. Literature Review

A rising corpus of research contrasting conventional econometric models with cutting-edge machine learning techniques has resulted from the increased complexity of macroeconomic systems. The limitations of traditional models like ARIMA and ARIMAX in capturing nonlinear interactions, structural fractures, and time-varying volatility, especially in labour market dynamics, even though they offer theoretical interpretability. On the other hand, because of their capacity to represent intricate nonlinear patterns and long-term dependencies in time series data, machine learning techniques like Long Short-Term Memory (LSTM) networks and other approaches have become increasingly popular.

Ref. [14] used ARIMAX models based on data from 1991 to 2021 to evaluate unemployment predicting performance during the COVID-19 timeframe.. The study found that while the ARIMAX model provides reasonably accurate forecasts in the pre-COVID-19 period, its predictive accuracy deteriorates significantly during and after the pandemic. These results demonstrate how linear models are limited in their ability to capture structural breaks and sudden economic shocks.

By combining Google Trends data with conventional macroeconomic variables, ref. [15] examined unemployment nowcasting in Ghana.. Using ARIMA, ARIMAX, and VAR models, the study found that the inclusion of search-based information significantly improved forecasting performance, with VAR achieving the lowest error across multiple evaluation metrics. The results highlight the importance of incorporating exogenous data sources in enhancing the accuracy of unemployment predictions.

Ref. [12] investigated unemployment forecasting in South Africa using ARIMA models in the presence of outliers and unexpected events. Using quarterly data from 2010 to 2020, the study applied robust estimators to account for data contamination and determined that the best model was an ARIMA(1,1,1) model. The results showed that, although traditional estimation methods are sensitive

 doi:10.20944/preprints202604.0965.v1

to outliers, the use of robust estimators improved forecasting accuracy, as evidenced by lower RMSE, MAE, and MAPE values.

Ref. [5] investigated inflation forecasting using a large U.S. macroeconomic dataset and applied machine learning techniques including random forests, boosting, and LASSO regression. Using high-dimensional data, the study found that machine learning models significantly outperform traditional linear econometric models in out-of-sample forecasting accuracy.

Ref. [8] investigated unemployment forecasting using monthly data from the United States, United Kingdom, France, and Italy spanning 1983-2022. The study applied LSTM, GRU, and a hybrid LSTM–GRU model and found that the hybrid deep learning model outperformed standalone models in terms of RMSE, MAE, and MAPE, demonstrating the advantage of combining nonlinear architectures.

Ref. [3] analysed S&P 500 stock market data from 1992 to 2015 using LSTM networks. The study compared LSTM with logistic regression and found that LSTM significantly improved predictive performance due to its ability to capture nonlinear dependencies.

Ref. [11] analysed unemployment forecasting in the United States using data from 1978-2023 and compared ARIMA, LSTM, and intervention models. The study found that while ARIMA captured general trends, LSTM provided superior predictive accuracy, particularly during periods of structural change such as the COVID-19 shock.

Ref. [10] examined unemployment forecasting in Indonesia using annual data from 1970-2023. The study applied multiple machine learning models, including gradient boosting and hybrid ARIMA-ML approaches, and found that machine learning models significantly outperformed traditional econometric models, with gradient boosting achieving the highest predictive accuracy.

Ref. [16] examined inflation forecasting using LSTM and traditional econometric models. Using multi-feature datasets, the study found that while LSTM models capture nonlinear patterns, hybrid models combining machine learning and traditional approaches often yield better performance depending on the dataset characteristics.

Ref. [17] examined hybrid ARIMA and neural network models and found that combining linear and nonlinear methods leads to improved predictive performance.

Ref. [7]used Google search data to forecast unemployment in the United States and demonstrated that real-time indicators significantly enhance predictive performance.

Ref. [13] examined unemployment forecasting using both traditional econometric and machine learning approaches with multivariate data obtained from the South African Reserve Bank. The study highlights the limitations of univariate and vector autoregression (VAR) models, noting that they rely on restrictive assumptions such as stationarity and equal interaction among variables. Using machine learning techniques, including LSTM, GRU, LASSO, and elastic net, the results showed that deep learning models, particularly LSTM and GRU, achieve significantly lower forecast errors compared to VAR models. Furthermore, feature selection methods identified domestic output and government expenditure as key determinants of unemployment, demonstrating the advantage of machine learning approaches in capturing complex relationships and identifying important predictors.

Despite these advancements, limited studies provide a direct and comprehensive comparison between econometric and deep learning models in unemployment forecasting, particularly in the South African context. Furthermore, the role of labour market variables such as the Not Economically Active (**NEA**) population remains insufficiently explored. This study addresses these gaps by providing a robust comparative analysis of ARIMAX and LSTM models within a unified framework.

1.3. Research Contributions and Highlights

1.3.1. Research Contributions

This study makes several key contributions to the literature on macroeconomic forecasting and labour market analysis. First, it addresses a critical gap in the literature by providing a unified empirical comparison between a traditional econometric model (ARIMAX) and a deep learning model (LSTM) in forecasting the South African unemployment rate. While existing studies typically examine these approaches in isolation, this study offers a direct and comprehensive comparison within a

 doi:10.20944/preprints202604.0965.v1

single framework, thereby providing clearer evidence on their relative performance. Second, the study extends existing unemployment forecasting models by incorporating labour market covariates, particularly the Not Economically Active (**NEA**) population, and demonstrates its statistically significant role in explaining unemployment dynamics. This contributes additional economic insight beyond purely statistical modelling. Third, the study employs a rigorous and multifaceted evaluation framework, including out-of-sample forecasting, Diebold-Mariano tests, Model Confidence Set (MCS) procedures, bootstrap confidence intervals, Murphy diagrams, and non-parametric tests. This ensures that the comparison between models is statistically robust and not driven by random variation or model-specific bias. Fourth, the empirical findings provide strong evidence that while ARIMAX models capture linear dependencies and offer interpretability, they are unable to adequately model nonlinear dynamics and structural breaks, particularly during periods of economic disruption such as the COVID-19 shock. In contrast, the LSTM model demonstrates superior performance by effectively capturing complex temporal dependencies, resulting in significantly improved forecasting accuracy. Finally, this study contributes to the growing body of literature advocating for the integration of machine learning techniques into macroeconomic forecasting. By providing both methodological and empirical evidence, the study highlights the practical relevance of nonlinear models for improving forecasting accuracy and supporting more informed labour market policy and decision-making.

1.3.2. Research Highlights

- A comparative analysis of ARIMAX and LSTM models is conducted for forecasting the South African unemployment rate.
- The unemployment rate quarterly data is found to be non-stationary, with structural breaks and time-varying volatility.
- The Not Economically Active (*NEA*) variable is determined as a significant predictor of unemployment rate dynamics.
- The ARIMAX model captures linear relationships but exhibits limited out-of-sample predictive accuracy.
- The LSTM model outperforms traditional forecasting model (ARIMAX) in terms of RMSE, MAE, and MAPE.
- Statistical tests (DM and MCS tests) confirm the superiority of the LSTM model.
- The results highlight the importance of nonlinear modelling in capturing complex labour market dynamics.
- Machine learning approaches (i.e., LSTM) provide more accurate and reliable forecasts for macroeconomic variables.

The rest of the paper is divided into three sections. Section 2 presents the data and methodology. The empirical results and analysis are presented in Section 3. then the Discussion and Conclusion are presented in Section 4.

2. Data and Methodology

2.1. Data Source

The time series data used for this study was quarterly data on the unemployment rate in South Africa from 2008 Q1 to 2025 Q3, resulting in 71 data points. The unemployment rate is the percentage of labour force that is unemployed, as per the labour market statistics of the country. Besides the dependent variable, additional labour market variables were included in order to make models more comprehensive. They consist of the number of discouraged work-seekers and Not Economically Active (NEA) populations. All the data was sourced from the official database of Statistics South Africa (Stats SA) and can be accessed freely using this website: https://www.statssa.gov.za. The dataset was split into an 80% training set, comprising 56 observations from 2008 Q1 to 2021 Q4, and a 20% testing set, consisting of 15 observations from 2022 Q1 to 2025 Q3.

 doi:10.20944/preprints202604.0965.v1

2.2. Compact Modelling Framework for ARIMAX and LSTM Forecasting

The complete modelling framework used in this study is shown in Figure 1, which shows the step-by-step procedure from data preparation to final model evaluation. The system incorporates preprocessing, diagnostic testing, and out-of-sample forecasting techniques while integrating both linear (ARIMAX) and nonlinear (LSTM) methodologies. To determine which model is the most dependable, the generated predictions are assessed and contrasted using accuracy metrics, statistical tests, and visual analysis.

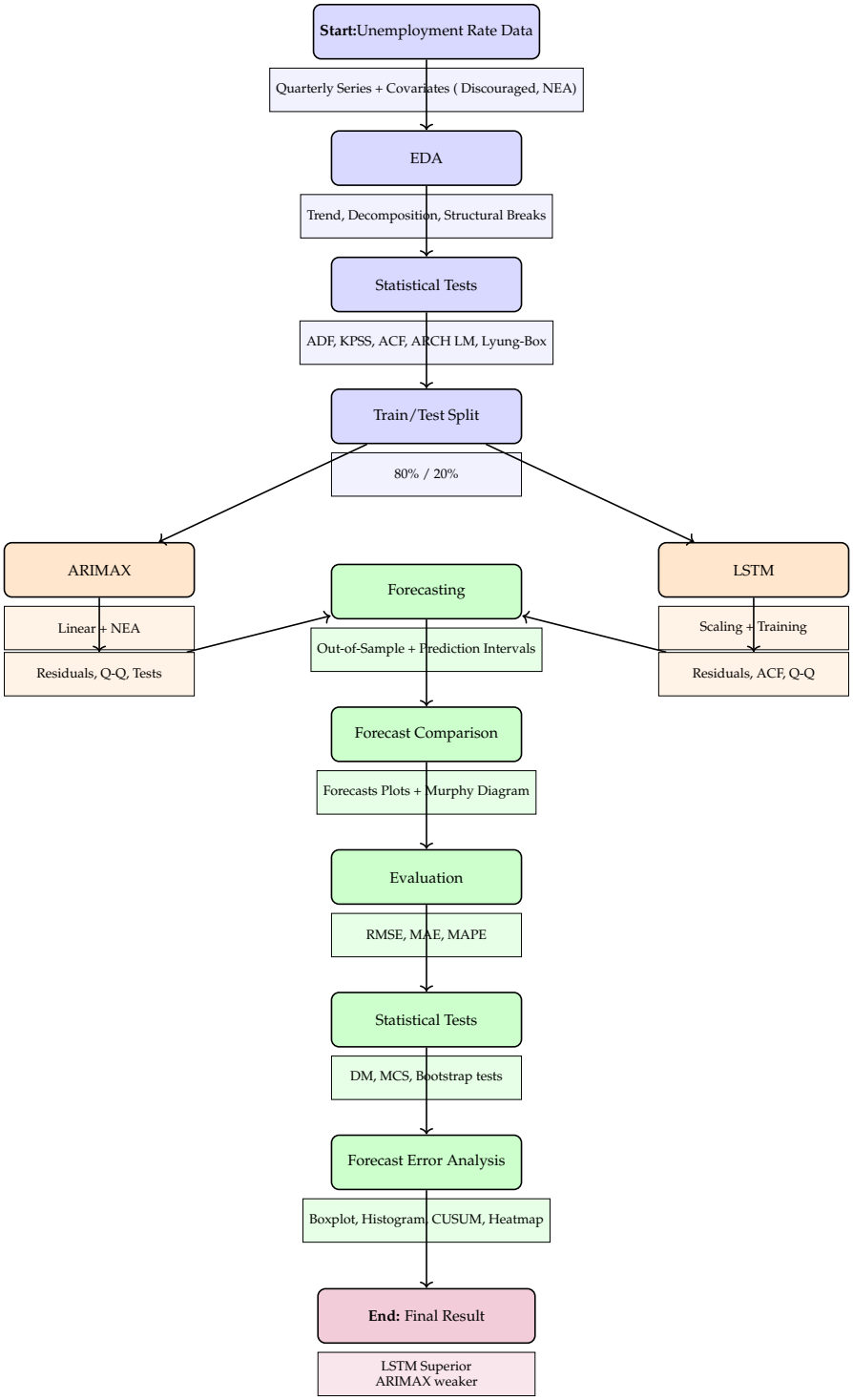

Figure 1. Compact Modelling Framework for ARIMAX and LSTM Forecasting.

 doi:10.20944/preprints202604.0965.v1

2.3. Models

2.3.1. ARIMA Model

A popular statistical technique for time series forecasting is the ARIMA model, which was first created by [1]. It is predicated on the idea that future observations can be predicted using historical values and prediction errors. The expression for the general ARIMA(p, d, q) model is:

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t, \tag{1}$$

where ϵ_t is a white noise process, d is the order of differencing needed to reach stationarity, and B stands for the backshift operator.

The ARIMA model is constructed from three components: the autoregressive (AR) part, the moving average (MA) part, and the integration (I) component. The current observation's reliance on its historical values is captured by the autoregressive component:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t, \tag{2}$$

Although the MA component simulates reliance on historical forecast errors:

$$y_t = \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}. \tag{3}$$

To achieve stationarity, the series is differentiated d times:

$$y'_t = (1-B)^d y_t. \tag{4}$$

Putting these components together gives us the full ARIMA representation:

$$y'_t = \sum_{i=1}^{p} \phi_i y'_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}. \tag{5}$$

In practice, the Box–Jenkins methodology is followed, which involves model identification using ACF and PACF plots, parameter estimation typically via maximum likelihood estimation, and diagnostic checking using statistical tests such as the Ljung–Box test. Although ARIMA models are highly interpretable and effective for linear structures, they are limited in capturing nonlinear relationships and require the data to be stationary.

To account for the influence of labor market variables, the linear component is extended to an ARIMAX model:

$$y_t = \phi(B)(1-B)^d y_t + \sum_{k=1}^{K} \beta_k x_{k,t} + \epsilon_t, \tag{6}$$

where $x_{k,t}$ represents the k-th covariate (Discouraged Work-Seekers and Not Economically Active (NEA) population) and β_k its corresponding coefficient. This extension allows the model to estimate the impact of each covariate and determine their statistical significance in explaining unemployment dynamics.

2.3.2. LSTM Model

In order to address the vanishing gradient issue, ref. [2] introduced the LSTM network, an advanced variant of the recurrent neural network. LSTM neural networks are ideal for time series forecasting because, unlike conventional neural networks, they include a memory cell that enables information to remain over long sequences. The gates of an LSTM cell control the data flow, which allows it to function. There are a number of gates involved in this process, such as the forget gate,

 doi:10.20944/preprints202604.0965.v1

which controls the discarding of certain pieces of information, the input gate, which chooses what new data should be saved, and the output gate, which governs the passing on of data to the next time step.

The LSTM algorithm is defined mathematically as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \tag{7}$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \tag{8}$$
$$\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c), \tag{9}$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t, \tag{10}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \tag{11}$$
$$h_t = o_t \odot \tanh(C_t). \tag{12}$$

The LSTM can capture intricate nonlinear relationships and long-term dependencies in the data by using this gating mechanism to selectively preserve pertinent information while eliminating irrelevant signals.

2.3.2.1 LSTM Architecture Diagram

The LSTM cell's internal structure is shown in Figure 2. The forget gate f_t controls the prior cell state C_{t-1}, which establishes the percentage of historical data to be retained. Concurrently, the input gate i_t regulates the quantity of new data that is integrated into the cell state, represented by the candidate memory $\tilde{C}_t$. The current cell state $\tilde{C}_t$ is updated by the interaction of these elements. The data to be sent to the hidden state h_t, which functions as both the input for the subsequent time step and the output of the present one, is then decided by the output gate o_t. Long-term dependencies and nonlinear dynamics in time series data can be efficiently captured by the LSTM because of this gating mechanism.

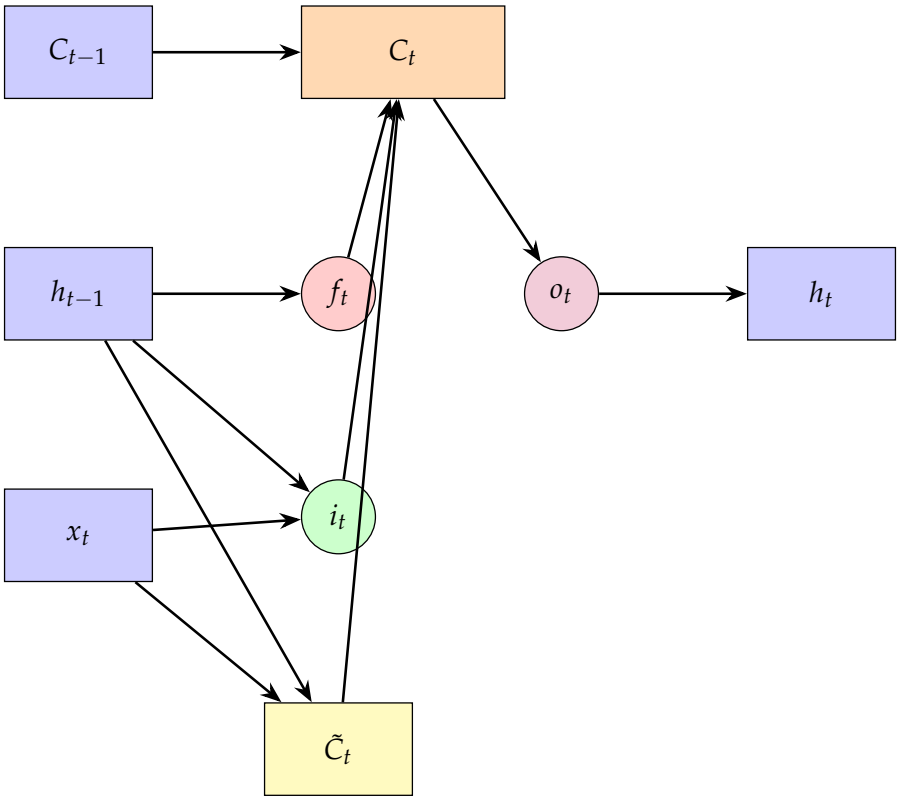

Figure 2. Architecture Diagram of the LSTM Cell.

 doi:10.20944/preprints202604.0965.v1

2.3.2.2 LSTM Pseudocode

Algorithm 1 summarises the processes used to train the LSTM model. The algorithm describes essential procedures, such as model development, sequence generation, data preparation, and hyperparameter optimisation.

Algorithm 1 LSTM Model for Unemployment Rate Forecasting

Input: Time series data y_t
Output: Trained LSTM model
Convert y_t into a time series object
Split data into training (80%) and testing (20%)
Normalise data using Min-Max scaling:

$$y^* = \frac{y - y_{\min}}{y_{\max} - y_{\min}}$$

Define lag length $L = 8$
Transform data into supervised sequences:
for $i = 1$ to $N - L$ **do**
    $X_i = (y_i, y_{i+1}, \ldots, y_{i+L-1})$
    $Y_i = y_{i+L}$
end for
Reshape input into 3D array: $(samples, timesteps, features)$
Define hyperparameter grid:
Units $\in \{50, 100\}$
Dropout $\in \{0.2, 0.3\}$
Batch size $\in \{8, 16\}$
for each combination of hyperparameters **do**
    Initialise LSTM model:
        LSTM layer (return sequences)
        Dropout applied
        Second LSTM layer
        Dense output layer
    Compile model with Huber loss and Adam optimizer
    Train model using:
        Early stopping (patience = 10)
        Validation split = 20%
    Record validation loss
end for
Select model with lowest validation loss
Return: Best trained LSTM model

2.3.2.3 LSTM Hyperparameter Tuning Table

The hyperparameter search space used for tuning the LSTM model is presented in Table 1. A grid search approach was employed to identify the optimal combination of parameters based on validation loss.

Table 1. LSTM Hyperparameter Grid for Model Tuning.

Hyperparameter	Values Tested
Number of Units	50, 100
Dropout Rate	0.2, 0.3
Batch Size	8, 16
Epochs	100
Loss Function	Huber Loss
Optimizer	Adam (learning rate = 0.001)
Lag Length	8
Early Stopping	Patience = 10
Validation Split	20%

 doi:10.20944/preprints202604.0965.v1

2.4. Forecast Accuracy Measures

To evaluate the predictive performance of the competing models, three widely used error metrics are employed, namely the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These measures quantify the deviation between the actual observed values and the corresponding forecasts.

This evaluation forecast's metrics errors are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \tag{13}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t| \tag{14}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \tag{15}$$

where y_t represents the actual value, $\hat{y}_t$ denotes the forecast value, and n is the number of observations.

2.5. Prediction Interval Coverage

The accuracy of prediction intervals is evaluated using the Prediction Interval Coverage Probability (PICP), which measures the proportion of observed values contained within the interval bounds.

Let y_t be the observed value at time t, with lower and upper bounds provided by L_t and U_t. Define the indicator:

$$I_t = \begin{cases} 1, & \text{if } L_t \leq y_t \leq U_t, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

The coverage probability is computed as:

$$\text{PICP} = \frac{1}{T}\sum_{t=1}^{T} I_t, \tag{17}$$

where T is the sample size.

For 95% prediction intervals, a value of PICP close to 0.95 indicates good interval performance.

2.6. Predictive Accuracy Statistical Test

The Diebold-Mariano (DM) test, proposed by [18], is a statistical test that is used to determine whether the predictive accuracy of two competing models' forecasts is significantly different or not.

Let u_t and v_t represent the forecast errors obtained from Model 1 and Model 2, respectively, at time t.

The loss differential is defined as:

$$\Delta_t = L(u_t) - L(v_t), \tag{18}$$

where $L(\cdot)$ represents a loss function. In this study, the squared error loss function $L(x) = x^2$, corresponding to the Mean Squared Error (MSE), is employed in constructing the loss differential for the DM test.. The sequence $\{\Delta_t\}$ is referred to as the loss differential series.

The average loss differential is given by:

$$\bar{\Delta} = \frac{1}{T}\sum_{t=1}^{T}\Delta_t, \tag{19}$$

where T stands for the sample size.

 doi:10.20944/preprints202604.0965.v1

To account for possible serial correlation in the loss differential series, the variance is estimated using autocovariances. Specifically, the autocovariance at lag $j \geq 1$ is defined as:

$$\Gamma_j = \text{Cov}(\Delta_t, \Delta_{t-j}). \tag{20}$$

The DM test statistic is then constructed as:

$$DM = \frac{\bar{\Delta}}{\sqrt{\hat{\sigma}^2/T}}, \tag{21}$$

where $\hat{\sigma}^2$ is a consistent estimate of the variance of Δ_t, typically incorporating autocovariances up to lag $h-1$. In practice, the forecast horizon is set to $h = 1$, corresponding to a one-step-ahead forecast..

The null hypothesis of the DM test is that both models have equal predictive accuracy:

$$H_0 : \mathbb{E}(\Delta_t) = 0, \tag{22}$$

against the alternative:

$$H_1 : \mathbb{E}(\Delta_t) \neq 0. \tag{23}$$

Under the null hypothesis, the DM statistic asymptotically follows a standard normal distribution:

$$DM \sim N(0,1). \tag{24}$$

The null hypothesis is rejected if the absolute value of the DM statistic exceeds the critical value from the standard normal distribution, i.e., if $|DM| > z_{\alpha/2}$, where α is the chosen level of significance.

A key assumption underlying the validity of the DM test is that the loss differential series $\{\Delta_t\}$ is stationary.

2.7. Model Confidence Set

The Model Confidence Set (MCS) procedure introduced by [19] provides a systematic framework for identifying a subset of models that are statistically indistinguishable in terms of predictive performance. This approach relies on sequential hypothesis testing to construct a set of superior models for which the null hypothesis of Equal Predictive Ability (EPA) cannot be rejected at a chosen significance level. The EPA test can be implemented using any suitable loss function, including squared or absolute loss functions.

Let $\mathcal{M}^0$ denote the initial collection of candidate models, and let $m = |\mathcal{M}^0|$ represent the total number of models under consideration. Define the loss associated with model i at time t as $\ell_{i,t}$. In this study, the loss function is specified as the squared forecast error, i.e., $\ell_{i,t} = e_{i,t}^2$.

The following represents the pairwise loss differential between models i and j:

$$\delta_{ij,t} = \ell_{i,t} - \ell_{j,t}, \quad i,j = 1,\ldots,m,\ t = 1,\ldots,T. \tag{25}$$

The average loss of model i relative to all competing models is defined as:

$$\delta_{i,t} = \frac{1}{m}\sum_{j\in\mathcal{M}} \delta_{ij,t}, \quad i = 1,\ldots,m. \tag{26}$$

Assume that $\bar{\delta}_{ij} = \mathbb{E}(\delta_{ij,t})$ and $\bar{\delta}_i = \mathbb{E}(\delta_{i,t})$ represent the expected loss differentials and are finite and time-invariant. The EPA hypothesis for the model set $\mathcal{M}$ can be formulated in two equivalent ways:

$$H_{0,\mathcal{M}} : \bar{\delta}_{ij} = 0 \quad \text{for all } i,j = 1,\ldots,m, \tag{27}$$

$$H_{A,\mathcal{M}} : \bar{\delta}_{ij} \neq 0 \quad \text{for some } i,j = 1,\ldots,m, \tag{28}$$

 doi:10.20944/preprints202604.0965.v1

or alternatively,

$$H_{0,\mathcal{M}} : \bar{\delta}_i = 0 \quad \text{for all } i = 1, \ldots, m, \tag{29}$$

$$H_{A,\mathcal{M}} : \bar{\delta}_i \neq 0 \quad \text{for some } i = 1, \ldots, m. \tag{30}$$

Based on these formulations, two test statistics are constructed. The first is based on pairwise comparisons:

$$\tau_{ij} = \frac{\bar{\delta}_{ij}}{\sqrt{\widehat{\mathrm{Var}}(\bar{\delta}_{ij})}}, \tag{31}$$

and the second is based on the average loss across models:

$$\tau_i = \frac{\bar{\delta}_i}{\sqrt{\widehat{\mathrm{Var}}(\bar{\delta}_i)}}. \tag{32}$$

Here, $\bar{\delta}_i = \frac{1}{m}\sum_{j\in\mathcal{M}} \bar{\delta}_{ij}$ represents the sample average loss of model i relative to other models, while $\bar{\delta}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \delta_{ij,t}$ captures the mean loss differential between models i and j. The quantities $\widehat{\mathrm{Var}}(\bar{\delta}_i)$ and $\widehat{\mathrm{Var}}(\bar{\delta}_{ij})$ denote bootstrap-based variance estimates.

Following [20], a block bootstrap procedure with 2,000 replications is employed, where the block length is determined by the maximum number of significant parameters obtained from fitting an autoregressive model to the series $\delta_{ij,t}$. Two overall test statistics are then defined as:

$$T_{\mathcal{R},\mathcal{M}} = \max_{i,j\in\mathcal{M}} |\tau_{ij}|, \tag{33}$$

$$T_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} \tau_i. \tag{34}$$

In this study, the $T_{\max,\mathcal{M}}$ statistic is employed to identify inferior models within the candidate set.

The statistic $T_{\mathcal{R},\mathcal{M}}$ is based on pairwise loss comparisons, whereas $T_{\max,\mathcal{M}}$ relies on the aggregated loss measure. These statistics are used to iteratively eliminate inferior models from the set $\mathcal{M}$.

The MCS procedure proceeds sequentially by removing the model that exhibits the poorest performance at each iteration until the null hypothesis of equal predictive ability can no longer be rejected. The elimination rules corresponding to the two statistics are given by:

$$e_{\mathcal{R},\mathcal{M}} = \arg\max_{i\in\mathcal{M}} \left\{ \sup_{j\in\mathcal{M}} \frac{\bar{\delta}_{ij}}{\sqrt{\widehat{\mathrm{Var}}(\bar{\delta}_{ij})}} \right\}, \tag{35}$$

$$e_{\max,\mathcal{M}} = \arg\max_{i\in\mathcal{M}} \frac{\bar{\delta}_i}{\sqrt{\widehat{\mathrm{Var}}(\bar{\delta}_i)}}. \tag{36}$$

This iterative elimination continues until a subset of models remains for which the EPA hypothesis cannot be rejected, forming the final MCS.

3. Empirical Results and Discussion

3.1. Exploratory Data Analysis (EDA)

The series of unemployment rates shows a mean of 27.74 with a standard deviation of 3.80. This evidence indicates that the series is relatively dispersed (refer to Table 2). A positive skewness value (0.49) reflects the slight right skewness of the data, while a negative kurtosis (-1.13) means that the series has a platykurtic distribution, which means the series will show light tails and be less peaked than the normal distribution. Overall, this series shows moderate dispersion and light tail distribution similar to other unemployment time series found empirically.

 doi:10.20944/preprints202604.0965.v1

Table 2. Summary Statistics of Unemployment Rate (%).

Statistic	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD	Skew	Kurt
Value	21.50	24.90	26.70	27.74	31.90	35.30	3.80	0.49	-1.13

Based on Figure 3, the time series graph shows an obvious increasing trend in the unemployment rate; hence, the series is non-stationary in terms of the mean value. It is evident from the graph that there is a sharp drop in the years leading to 2020 and, thereafter, a sharp increase. It can be interpreted that there is a structural shock, such as the economic impact of the COVID-19 pandemic, that leads to the stability of the series at the new higher level. The above characteristics coincide with the descriptive statistics provided in Table 2, where the variance (SD = 3.80) is high and the skewness (+0.49) implies some periods of a high unemployment rate. The negative value of kurtosis (-1.13) means the low level of observations' concentration at the mean value. Finally, there is no evident seasonal pattern within the quarters.

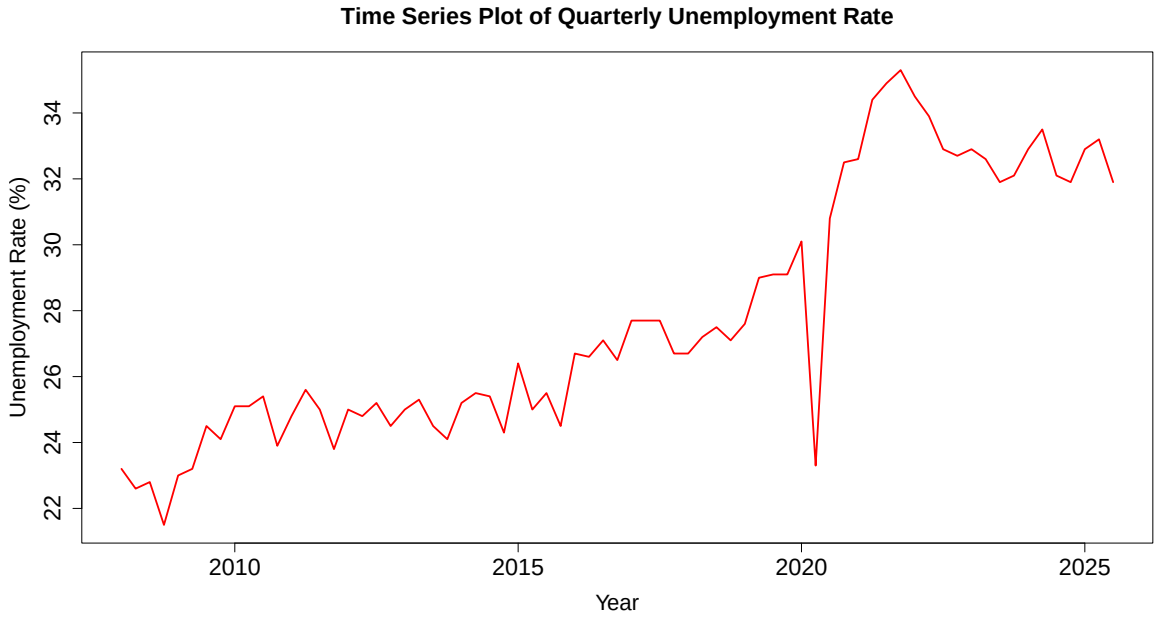

Figure 3. Time Series Plot of Quarterly Unemployment Rate.

The decomposition graph from Figure 4 illustrates the strong upward trend component as the main driving force behind the changes in the unemployment rate, proving the existence of non-stationarity in the time series data. A pronounced structural break takes place in the year 2020, shown through the significant departure in both the trend and remainder components, implying the existence of an exogenous shock. As for the seasonal component, it demonstrates steady and constant values at each period, implying the existence of consistent quarterly seasonality, even though the amplitude of its fluctuations is rather low in comparison with the trend. The remainder component has a tendency to be close to zero, which implies that most of the variability is accounted for in the trend and seasonal components.

As per Table 3, the unemployment rate data series is non-stationary. This conclusion is made based on the findings from the Augmented Dickey-Fuller (ADF) test in which we are unable to reject the null hypothesis of unit root since the p-value (0.4325) is higher than the critical value of 0.01. Likewise, the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test also rejects the null hypothesis of stationarity.

 doi:10.20944/preprints202604.0965.v1

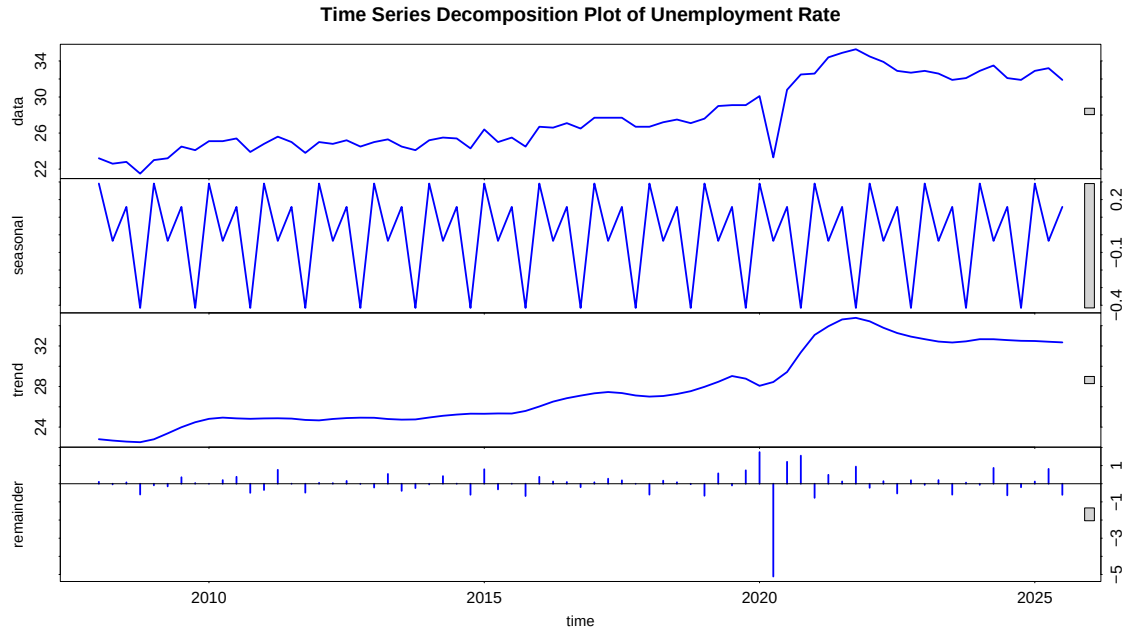

Figure 4. Time Series Decomposition Plot of Unemployment Rate.

Table 3. Stationarity Test Results for Unemployment Rate Time Series.

Test	Test Statistic	Lag	p-value	Conclusion
ADF	-2.3508	4	0.4325	Non-stationary
KPSS	1.653	3	$<$0.01	Non-stationary

The ACF and PACF plots of the unemployment rate in Figure 5 reveal that the series has very strong persistence, meaning that it is not stationary. The ACF plot has a very slow decay and features many significant lags, implying that there is a unit root in the series and that any shock to it will have a long-lasting effect on it. However, the PACF has only one significant spike at lag 1, while the other lags do not show significance.

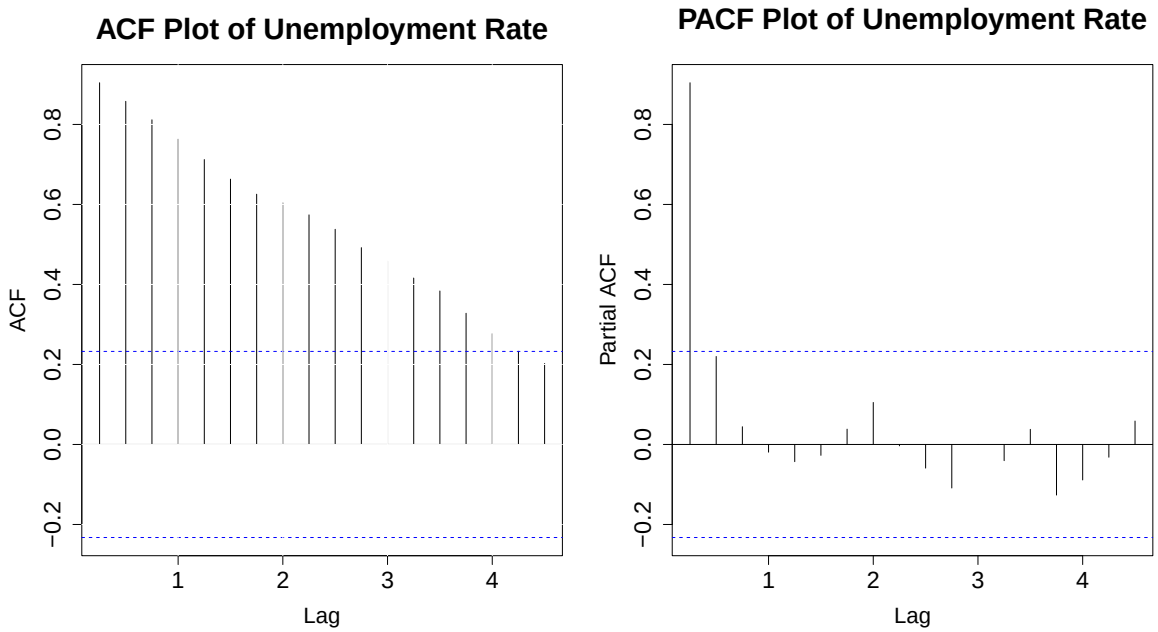

Figure 5. ACF and PACF Plots of the Unemployment Time Series Data.

 doi:10.20944/preprints202604.0965.v1

As per Table 4, the null hypothesis of no autocorrelation, using the Ljung-Box test, is rejected ($p < 0.001$), which indicates that the unemployment rate time series data has autocorrelation. Moreover, the null hypothesis of no ARCH effect is rejected ($p < 0.001$), which means there is conditional heteroskedasticity present in the unemployment time series data. It can be concluded from these results that the series data is not a white noise process but rather exhibits volatility; hence, there is justification for using models which capture autocorrelation and volatility.

Table 4. Diagnostic Test Results for Unemployment Rate Time Series

Test	Test Statistic	Lag	p-value	Conclusion
Ljung–Box	497.79	20	<0.001	Autocorrelation present
ARCH LM	51.153	12	<0.001	ARCH effects present

The structural break plot presented in Figure 6 shows a clear change in regime in the level of unemployment at around the year 2020. Before the structural break, the plot shows a stable pattern of the series with an upward trend that moves around a low average of 25.5. After the structural break, however, there is a steep fall followed by a sharp rise in the series, after which it becomes stable at a high average of 33. This plot shows a clear structural break, likely due to an exogenous shock that permanently changed the unemployment rate.

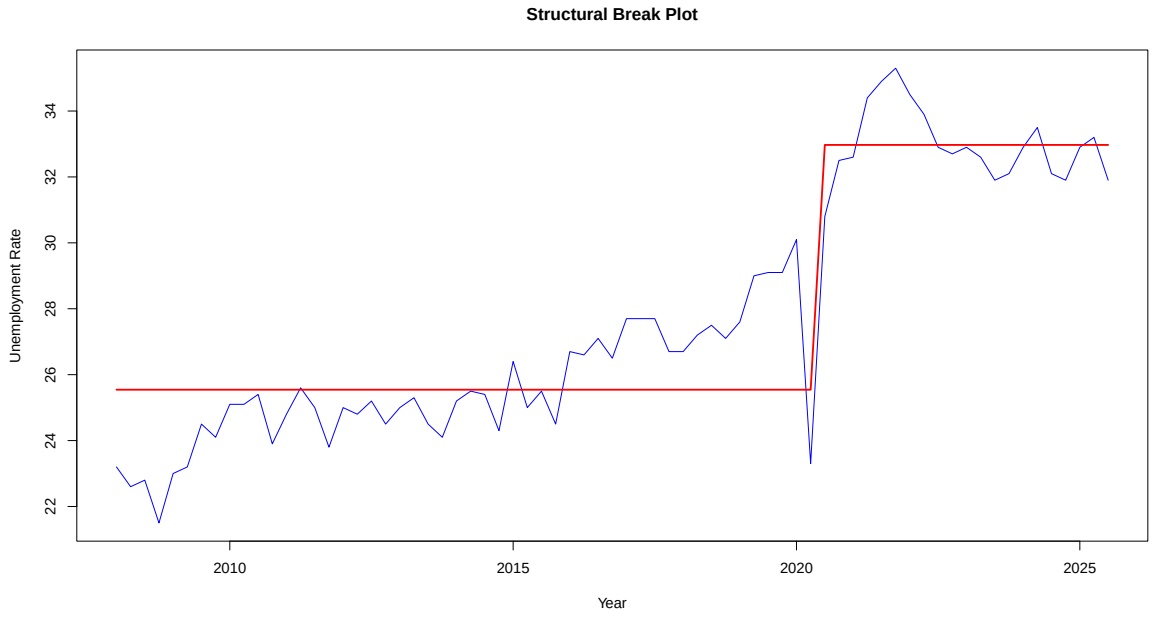

Figure 6. Structural Break Analysis Plot.

Figure 7 illustrates the time-varying volatility, displaying the rolling four-quarter standard deviation of the unemployment rate. For most of the period within the sample, volatility has been fairly constant at low levels. However, there is one point in time when the level of volatility peaks, specifically in the period between 2020 and 2021, when the standard deviation rises to the highest level within the sample period. This represents the point of increased volatility. It would suggest that there is an economic shock at play during this period. Afterwards, volatility falls and goes back to normal levels.

 doi:10.20944/preprints202604.0965.v1

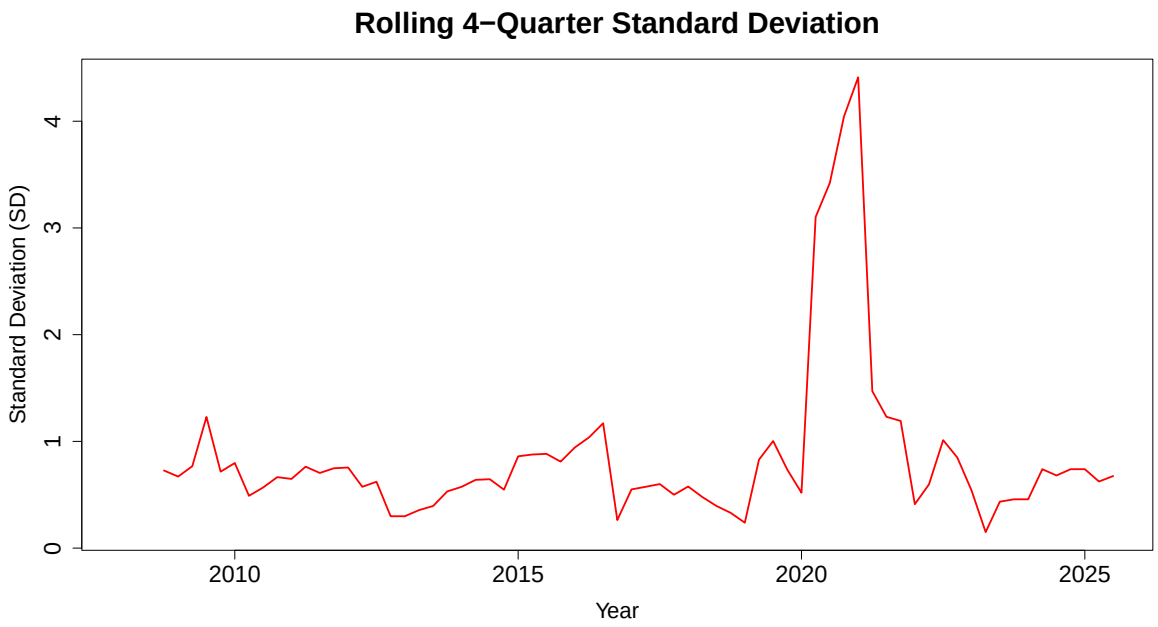

Figure 7. Rolling Plot of Standard Deviation.

As shown in Figure 8, the unemployment rate is displayed based on its trend and including the time period during the outbreak of COVID-19 pandemic. The trend line until 2020 shows the increasing movement of the unemployment rate with slight variations that suggest a steady labour market environment. After the emergence of COVID-19, there is an initial decrease in the unemployment rate before a steep rise occurs in it, which then peaks soon after. In the latter time periods, the unemployment rate starts to become slightly stabilised yet higher than the previous levels.

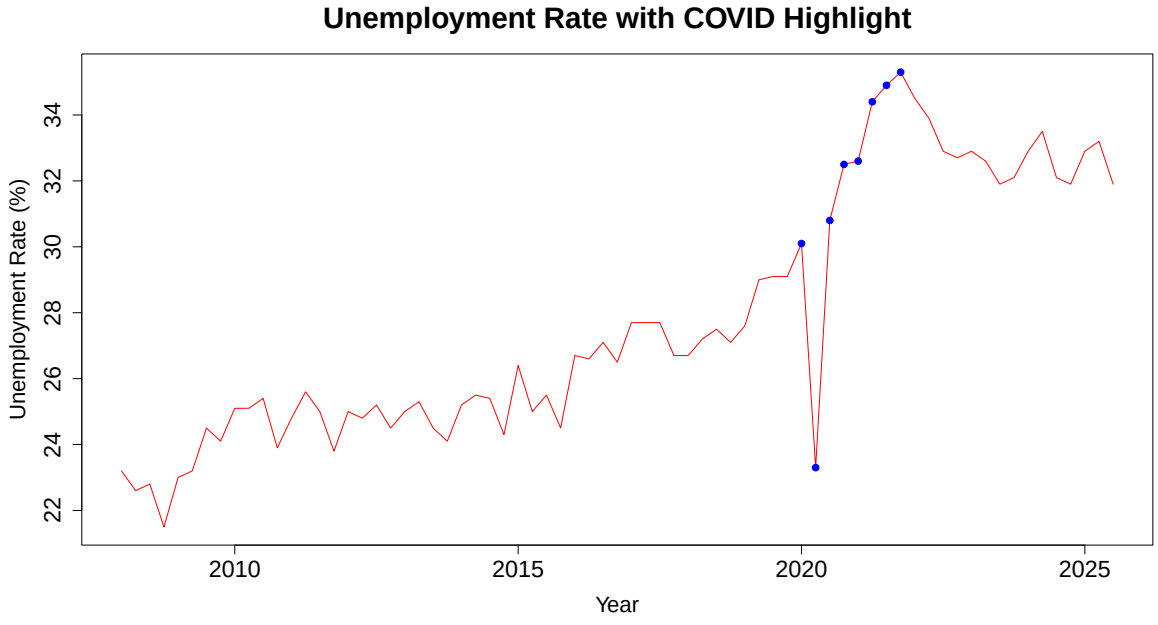

Figure 8. Time Series Plot of Unemployment Rate with COVID Highlight.

Figure 9 shows the development over time of the *Discouraged* workers and the Not-Economically Active (*NEA*) population. In the case of the discouraged series, we can observe a rising trend along with moderate oscillations in its development. Hence, there has been an increase in the number of those who abandoned their search for employment because of being unable to find a job for some period. In particular, there was an upsurge around the period of 2020 and afterward, which suggests

 doi:10.20944/preprints202604.0965.v1

increased labour market problems. On the other hand, before 2020, the *NEA* series was rather constant. However, after that point, there was a pronounced upturn due to the pandemic situation.

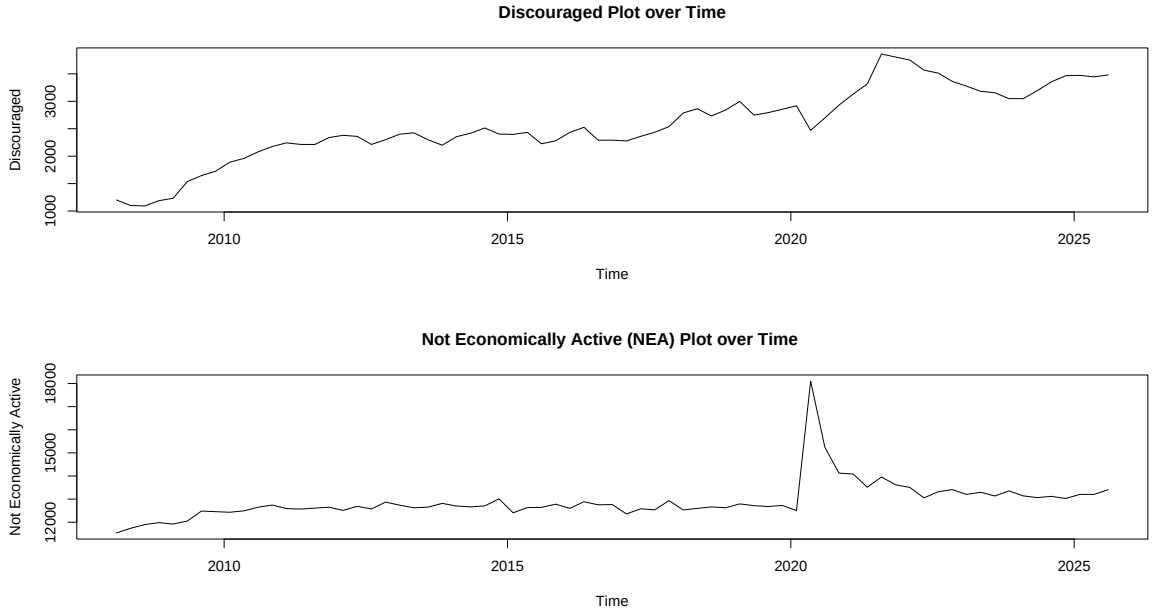

Figure 9. Unemployment Covariates Plot.

In Figure 10, we see the correlation between the *Discouraged* workers and the *NEA* variables. The outcome shows there exists a significant positive correlation between the two variables at a correlation coefficient of about 0.48. This means an increase in the number of the discouraged individuals goes hand in hand with an increase in the number of individuals who are not economically active. The level of correlation is not high enough to conclude that a multicollinearity problem exists between the two variables.

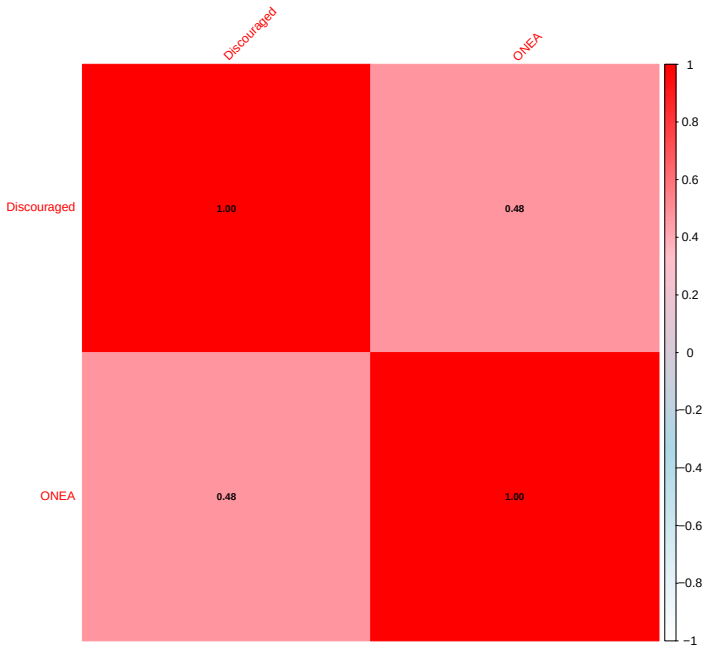

Figure 10. Correlation Matrix of Unemployment Covariates.

As indicated by the multicollinearity results in Table 5, all the covariates have a VIF less than 5. It means there is no multicollinearity among the covariates. In other words, it shows that the covariates

 doi:10.20944/preprints202604.0965.v1

are linearly independent and hence can be included in the model. On the other hand, according to the correlation significance shown in Table 6, the correlations between the covariates are significant at $p < 0.05$. It means that there is a linear relationship between the covariates. Despite their statistical significance, the lack of multicollinearity among them is confirmed by the low VIF values. It means that the relationship between them is not strong enough to result in multicollinearity.

Table 5. Variance Inflation Factor (VIF) Results for Covariates.

Variable	VIF
Discouraged	1.296
NEA	1.296

Table 6. Correlation Significance (p-values) Among Covariates.

	Discouraged	ONEA
Discouraged	0.0000	0.0000
NEA	0.0000	0.0000

From the output of the Granger causality test in Table 7, it is observed that the variable '*Discouraged*' does not have any effect on Granger-causing the unemployment rate since it yields a non-significant p-value ($p = 0.6622$). So, the variable Discouraged's past values do not predict the unemployment rate. However, the variable *NEA* shows statistical significance ($p < 0.001$) in terms of having Granger causation. It means that the previous values of the variable *NEA* do have some prediction power over the unemployment rate.

Table 7. Granger Causality Test Results (Lag Order = 2).

Covariate	F-Statistic	p-value	Conclusion
Discouraged	0.415	0.6622	No Granger causality
NEA	22.857	<0.001	Granger causality present

Test results on stationarity of the covariates in Table 8 show mixed results. In the case of *Discouraged*, the ADF test does not reject the hypothesis of the unit root ($p < 0.01$). This result implies stationarity. However, the KPSS test does not support stationarity of the series ($p < 0.01$). These mixed results imply that there could be structural instabilities, or the data may need transformations. The case with *NEA* is that, using the ADF test, non-stationarity ($p = 0.2664$) is observed. Using the KPSS test as well, non-stationarity ($p < 0.01$) is observed. It means that the covariate *NEA* is non-stationary and needs differencing before including it in the model.

Table 8. Stationarity Test Results for Covariates.

Variable	Test	Test Statistic	Lag	p-value
Discouraged	ADF	-4.6073	4	<0.01
Discouraged	KPSS	1.6066	3	<0.01
NEA	ADF	-2.7590	4	0.2664
NEA	KPSS	0.8977	3	<0.01

The results shown in Table 9 gives robust evidence about the stationary nature of the differenced *NEA* data series. From the ADF test result, the rejection of the null hypothesis of unit root proves that

 doi:10.20944/preprints202604.0965.v1

the series is stationary, as this occurs at a 1% level of significance. The same is supported by the KPSS test, where we fail to reject the null hypothesis of stationarity, given that the p-value is above 0.10.

Table 9. Stationarity Tests for Differenced NEA Series.

Test	Test Statistic	Lag	p-value
ADF	-4.9209	4	< 0.01
KPSS	0.0430	3	> 0.10

3.2. Fitting ARIMAX Model

3.2.1. Parameter Estimates

Coefficients of the ARIMAX model are shown in Table 10. The coefficient of the second moving average with seasonality is statistically significant at 1%, which confirms that there is some seasonal dependence in the unemployment rate data. Regarding explanatory variables, the differenced *NEA* is highly statistically significant, which indicates that the *NEA* rate is strongly positively influencing the unemployment process. On the contrary, the variable of *Discouraged* Workers is insignificant. It means that it has not much impact on the unemployment dynamics when accounting for other effects.

Table 10. Estimated ARIMAX Model Coefficients.

Variable	Estimate	Std. Error	t-value	p-value
SMA(1)	0.1744	0.1319	1.3215	0.1863
SMA(2)	0.5118	0.1875	2.7293	0.0063***
Δ Discouraged	0.0010	0.0008	1.1849	0.2361
Δ NEA	0.0008	0.0002	4.8105	0.0000***

Note: *** p<0.01, ** p<0.05, * p<0.1

The model fit and diagnostic results of the ARIMAX model are presented in Table 11. The ARIMAX model was estimated using the *auto.arima()* function. This procedure selects the best-fitting model using information criteria. The best model found is ARIMA(0,1,0)(0,0,2)[4]. The model fits perfectly since it has quite low values for the information criteria despite having a moderate residual variance. Moreover, it also has excellent in-sample forecasting performance since it has low values for the forecast error metrics. The near-zero value of the autocorrelation of the residuals at lag one also indicates an adequate handling of serial correlation.

3.2.2. ARIMAX Model Diagnostics

Based on the residual diagnostics presented in Figure 11, it can be noted that the ARIMAX model is able to capture the serial correlation in the dataset since most of the autocorrelation coefficients are contained within the significant region, implying no more autocorrelation exists. Moreover, the residual time plot shows that the residuals oscillate around the mean of zero, although some outliers can be seen, especially during the time of the COVID-19 outbreak. Likewise, from the histogram and kernel density overlay, it is clear that the errors are not normally distributed, showing evidence of skewness and heavy-tailed distribution, which is supported by the findings of the normality test.

 doi:10.20944/preprints202604.0965.v1

Table 11. ARIMAX Model Fit and Diagnostic Statistics.

Statistic	Value
Model	ARIMA(0,1,0)(0,0,2)[4]
σ^2	1.309
Log-likelihood	-84.62
AIC	179.23
AICc	180.46
BIC	189.27
ME	0.1565
RMSE	1.0920
MAE	0.6467
MAPE	2.4002
MASE	0.4867
ACF1	-0.0137

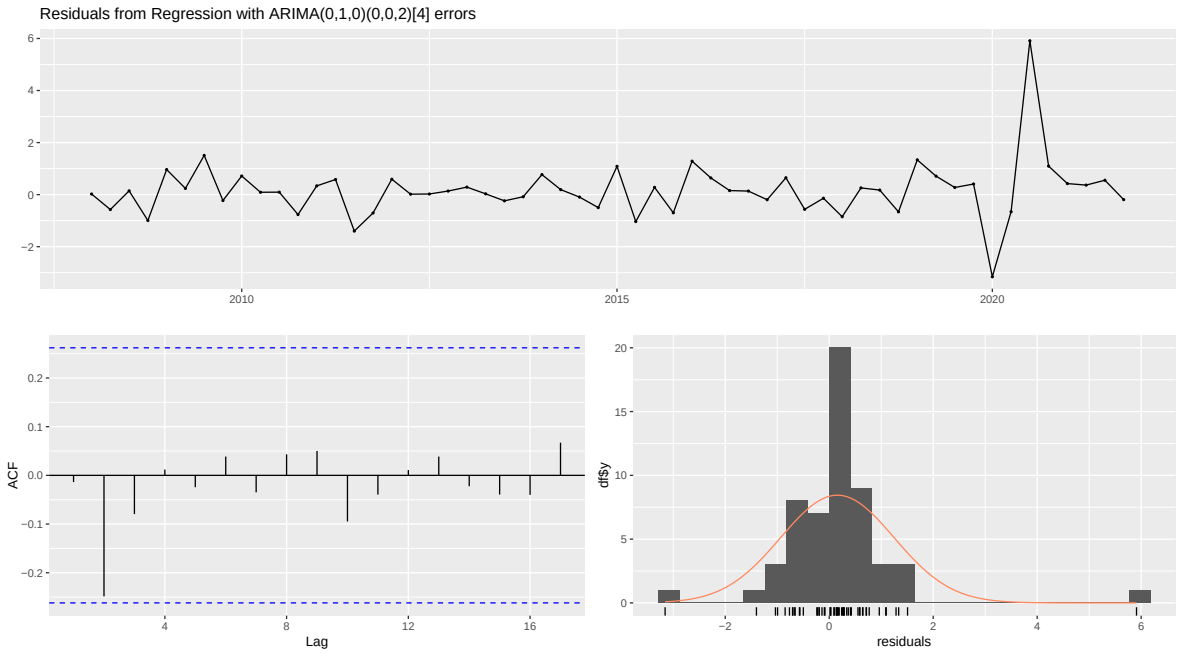

Figure 11. Time Series, ACF, and Histogram with Density Curve Plots of the ARIMAX Residuals.

The normal Q-Q plot presented in Figure 12 shows some clear violation of the normality assumption of the residuals. Even though the data points located in the middle part follow the reference line quite well and show signs of normality near the mean value, the tail points clearly deviate from this theoretical line. More precisely, there is a significant deviation from this line in the upper tail, which means that positive outliers occur, while in the lower tail, there is some degree of deviation as well. It means that the tails of the residuals' distribution can be considered heavy and possibly skewed as well.

 doi:10.20944/preprints202604.0965.v1

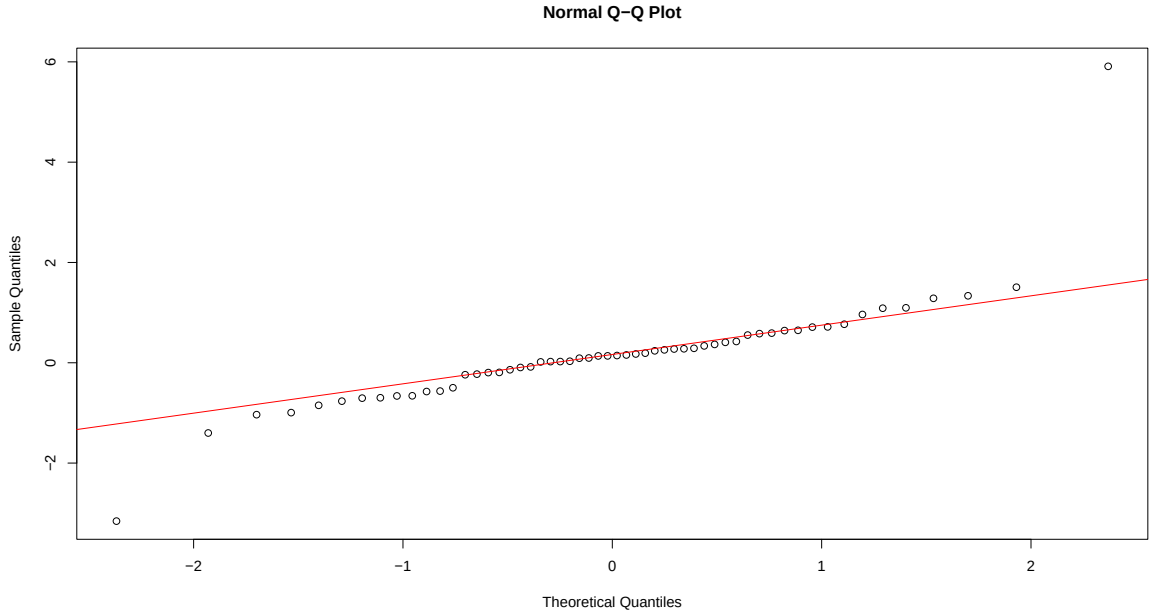

Figure 12. Normal Q–Q Plot of the ARIMAX Residuals.

The results of the Shapiro-Wilk test of normality conducted on the residuals from the ARIMAX model are presented in Table 12. It is evident that the test produces a p-value that is extremely low, which means that the null hypothesis of normality can be rejected. This suggests that the errors are non-normally distributed and could be skewed or have fat tails. This is not an indication that the model used is incorrect; however, it might mean that the assumption regarding the distribution of the errors is not correct.

Table 12. Shapiro-Wilk Normality Test for ARIMAX Residuals.

Test	Statistic (W)	p-value
Shapiro-Wilk	0.7648	4.345×10^{-8}

The residual errors of the ARIMAX model were tested for autocorrelation and conditional heteroskedasticity as shown in Table 13. The Box-Ljung test gives $X^2 = 6.8146$ and $p = 0.9973$ showing that there is no autocorrelation among the residual errors. Likewise, the results from the ARCH LM test (($\chi^2 = 12.009$, $p = 0.445$) show that there is no heteroskedasticity.

Table 13. Autocorrelation and Heteroskedasticity Test Results on the ARIMAX Residuals.

Test	Statistic	Lag	p-value
Box-Ljung	6.8146	20	0.9973
ARCH LM	12.009	12	0.445

3.2.3. ARIMAX (ARIMA(0,1,0)(0,0,2)[4]) Model Forecasting

The predictions generated by the ARIMAX model have been captured in Table 14 by considering an 80% training data set size and a 20% testing data set size. It can be observed from the table that the general trend depicted by the predictions is an increasing trend with respect to the unemployment rate, and the predictions remain at higher levels in subsequent periods. However, the prediction interval gradually widens with each period.

 doi:10.20944/preprints202604.0965.v1

Table 14. Out-of-Sample ARIMAX Forecasts with 95% Prediction Intervals.

Period	Forecast	Lower (95%)	Upper (95%)
2022 Q1	33.352	31.109	35.595
2022 Q2	33.798	30.626	36.969
2022 Q3	36.686	32.802	40.571
2022 Q4	37.021	32.535	41.506
2023 Q1	37.484	32.283	42.686
2023 Q2	37.525	31.695	43.355
2023 Q3	38.045	31.648	44.443
2023 Q4	37.688	30.770	44.607
2024 Q1	37.945	30.060	45.829
2024 Q2	38.064	29.320	46.809
2024 Q3	37.896	28.369	47.422
2024 Q4	38.021	27.771	48.271
2025 Q1	37.841	26.915	48.766
2025 Q2	38.070	26.509	49.631

Table 15 shows the forecast accuracy metrics of the ARIMAX model. From Table 15, the forecast errors seem to be relatively large, based on the RMSE and MAE, and hence there exists a difference between the forecasted values and the actual observations, and the forecast errors cannot be overlooked. The MAPE indicates an error of about 13.9%. In general, the model does not do very well when it comes to prediction.

Table 15. Out-of-sample Forecast Accuracy Measures for ARIMAX Model.

Model	RMSE	MAE	MAPE
ARIMAX	4.8171	4.5094	0.1388

Figure 13 shows the in-sample fit and forecasting results for the ARIMAX model of the unemployment rate. The in-sample fit shows a good match to the historical pattern of the series, with respect to its overall increase as well as volatility, where we observe an interesting decrease and rebound around 2020. In other words, there is an in-sample goodness-of-fit. Forecasting results imply a further increase in the value, which will eventually stabilise at higher levels. Yet, it is clear that the forecast interval is becoming wider with a longer forecast horizon, which speaks in favour of decreasing forecast accuracy.

The out-of-sample predictions made by the ARIMAX model after splitting the data into 80% training and 20% test samples can be seen in Figure 14. It can be observed that the ARIMAX model has been able to capture the overall trend of the increase in the level of unemployment but always predicts higher values than those actually observed. However, the actual observation lies within the prediction interval of 95%, which widens as time increases. The fact that the model's prediction does not match that of the actual observations indicates the model's inability to capture any current trends in the dataset. Thus, while the ARIMAX model has been successful at giving direction, it fails in terms of precision.

 doi:10.20944/preprints202604.0965.v1

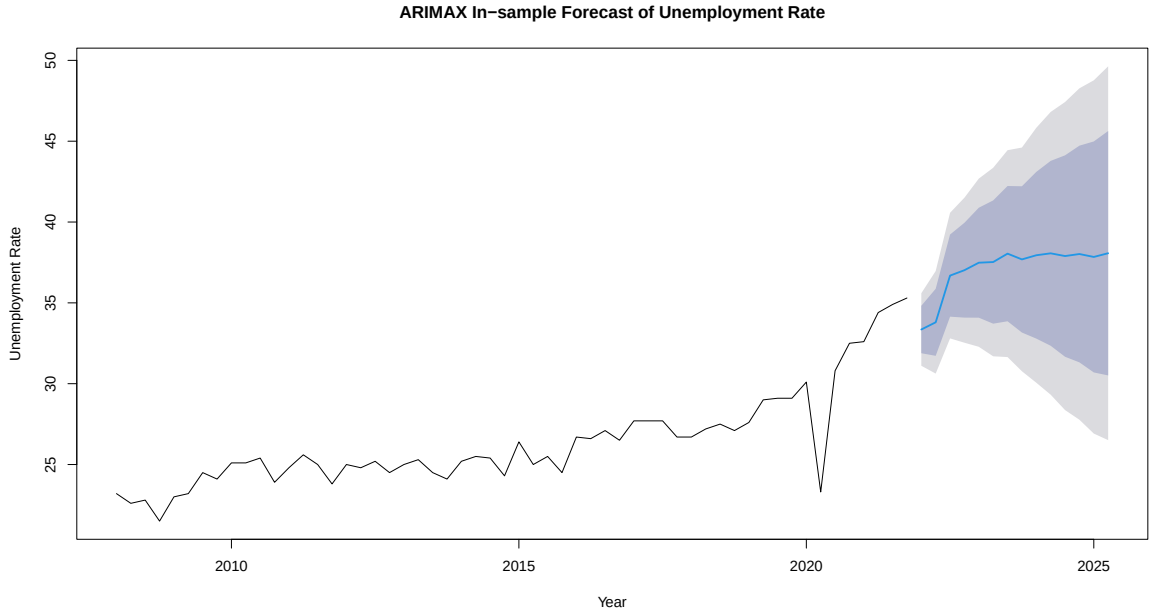

Figure 13. ARIMAX In–Sample Forecast of Unemployment Rate.

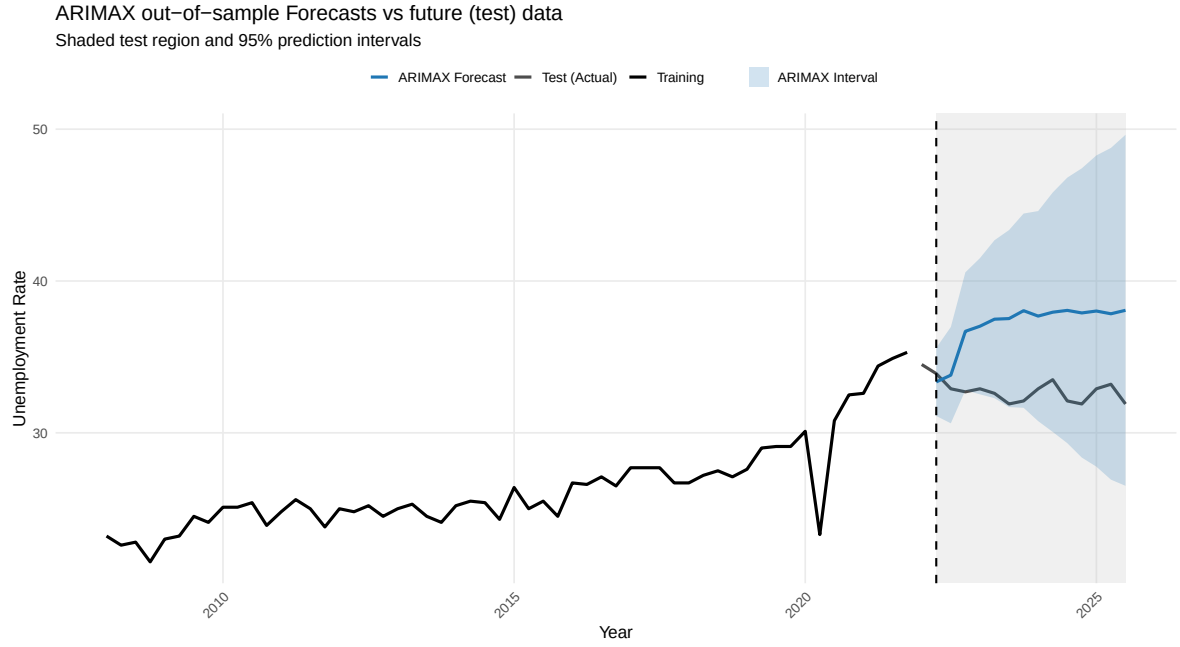

Figure 14. ARIMAX Out–of–Sample Forecasts vs Test (Actual) data.

3.3. Fitting LSTM Model

3.3.1. LSTM Model Development and Evaluation

From the training and validation losses depicted in Figure 15, it is clear that the convergence of the model has happened quite fast in the earlier epochs during the training period. Even though the training process was set to run up to 100 epochs, it stopped automatically after the tenth epoch because of no improvement in the validation loss after this. Training loss consistently reduces and converges, showing that the model has learnt the underlying data structure well. On the other hand, the validation loss reduces at the initial epochs but reaches the minimum at the third epoch and then begins to increase. This marks the beginning of overfitting, an effect which has been mitigated through early stopping in the training.

 doi:10.20944/preprints202604.0965.v1

The values of the hyperparameters that are determined by the grid search are in accordance with the trends seen in the training and validation losses. In particular, the combination of the number of units equal to 100, the dropout rate of 0.2, and the batch size of 16 corresponds to the point when the model has the lowest validation loss during the training process. As is shown in Figure 15, the lowest value of the validation loss occurs in the beginning of the training process, after which the value increases, suggesting overfitting.

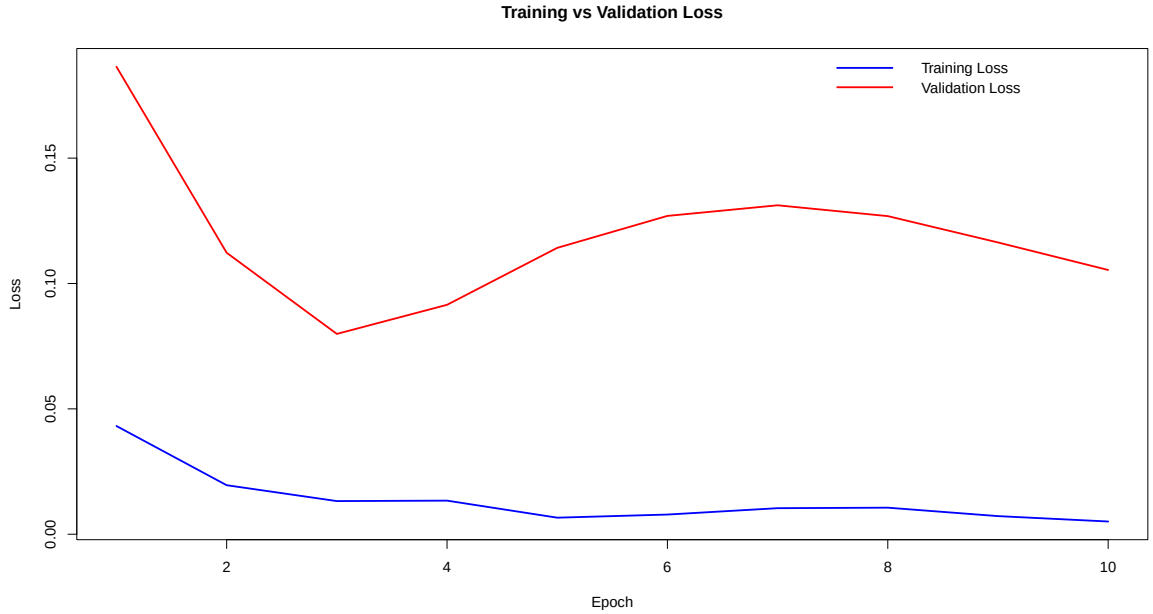

Figure 15. Training vs Validation Loss.

From the results of the diagnostics, shown in Table 16, it can be concluded that the residuals of the LSTM satisfy important assumptions associated with the adequacy of the model. In particular, for the residuals of the model, the Ljung-Box test yields a value of $p = 0.4496$, exceeding the critical value. Therefore, the lack of autocorrelation in the residuals indicates that the model adequately describes the dynamics of the data under study. In addition, using the ARCH LM test, we obtain a value of $p = 0.9994$, which indicates the absence of conditional heteroscedasticity. This means that the variance of the residuals does not change over time. Hence, we can deduce that the residuals represent white noise.

Table 16. Diagnostic Tests for LSTM Residuals.

Test	Statistic	df	p-value
Ljung-Box	9.8963	10	0.4496
ARCH LM	2.0000	12	0.9994

From the LSTM residuals diagnostic plot presented in Figure 16, it can be stated that the model fits the data adequately. The residual series shows a random pattern around zero and does not exhibit any clear trends, which implies that the model has been able to capture the inherent characteristics of the data. From the ACF and PACF plots, it is observed that almost all the autocorrelation values fall within the confidence limits, meaning that there are no signs of serial correlation in the residuals. This is in line with the assumption of white noise errors. Additionally, the plot of the histogram of residuals along with its estimated density shows an almost normal distribution of residuals, although there is some degree of non-normality in the data distribution.

 doi:10.20944/preprints202604.0965.v1

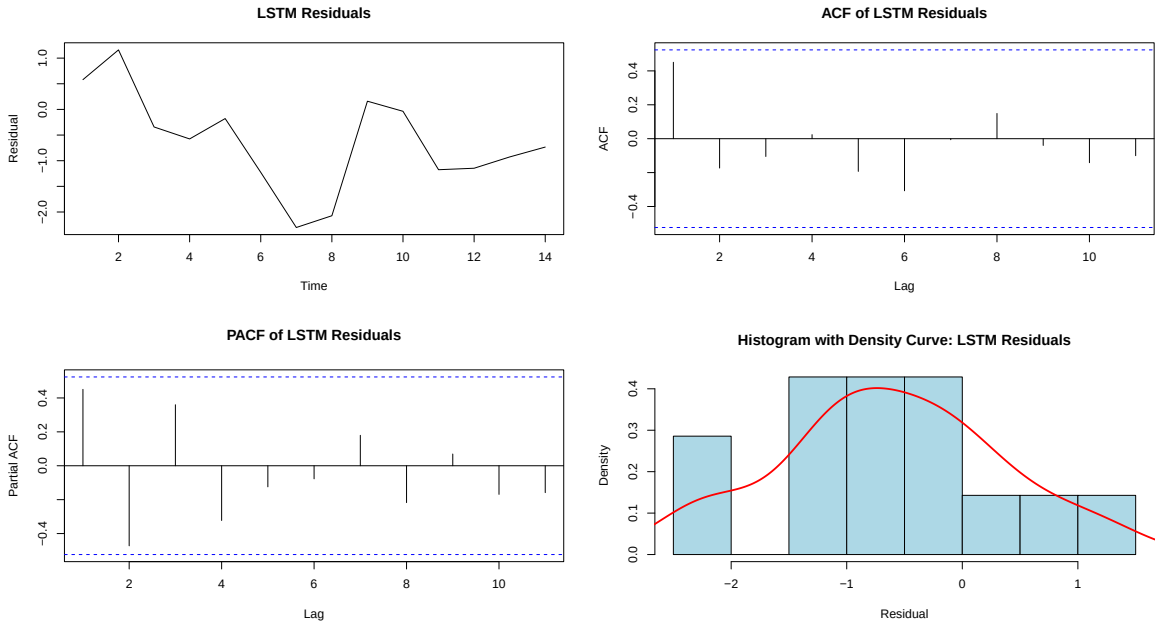

Figure 16. Time Series, ACF & PACF, and Histogram with Density Curve Plots of LSTM Residuals.

As observed from the Q-Q plot shown in Figure 17, the residuals of the LSTM model appear to follow a normal distribution since they are near the line for the majority of observations. In other words, the middle section of the histogram is well-conformable with the assumption of normality. The only thing one could note is that there are some deviations at the tails, especially on the lower and upper edges. Nevertheless, the deviations are rather small and do not cause much concern. It implies that the residuals can be regarded as quite well-behaved.

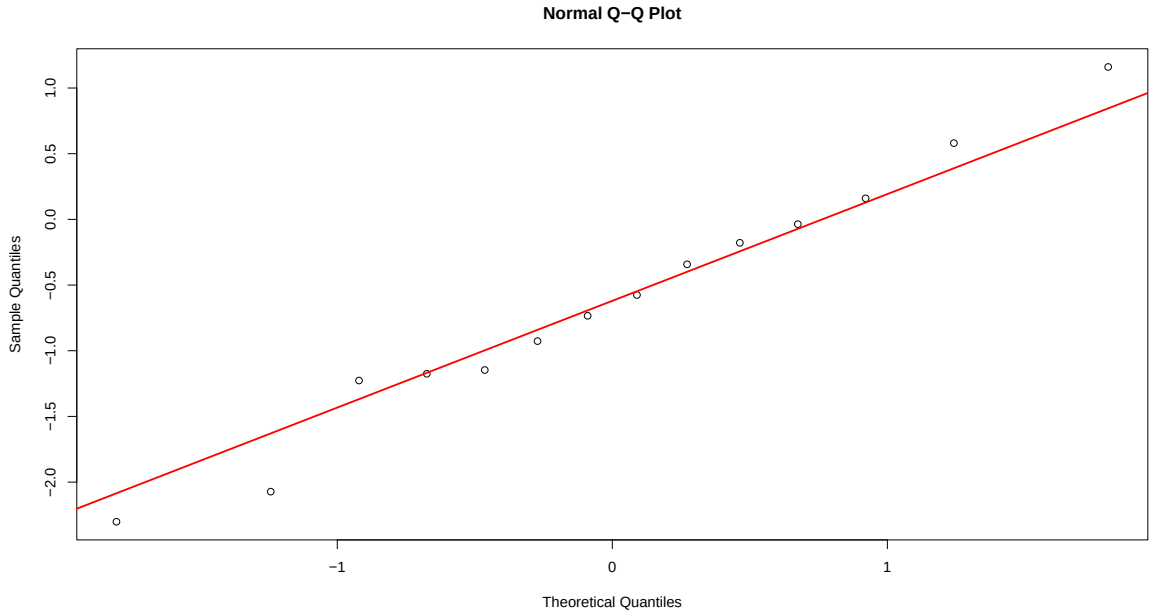

Figure 17. Normal Q–Q Plotof LSTM Residuals.

3.3.2. LSTM Model Forecasting

Table 17 depicts the out-of-sample predictions of the LSTM neural network model using the 80/20% train-test split method. It is observed that the predictions do not deviate significantly during the forecast period, which implies that there exists consistency in the unemployment level predicted by the LSTM model without high levels of variance. Moreover, the width of the prediction intervals for the

 doi:10.20944/preprints202604.0965.v1

model is narrower than that of conventional models and shows little variation throughout the forecast horizon. This means that the LSTM model performs better than ARIMAX model in out-of-sample forecasting.

Table 17. Out-of-Sample LSTM Forecasts with 95% Prediction Intervals.

Period	Forecast	Lower (95%)	Upper (95%)
2022 Q1	32.975	31.180	34.770
2022 Q2	31.919	30.124	33.714
2022 Q3	32.355	30.560	34.150
2022 Q4	32.373	30.577	34.168
2023 Q1	32.194	30.399	33.990
2023 Q2	32.870	31.074	34.665
2023 Q3	33.210	31.415	35.005
2023 Q4	33.184	31.388	34.979
2024 Q1	31.919	30.124	33.714
2024 Q2	32.619	30.823	34.414
2024 Q3	32.373	30.577	34.168
2024 Q4	32.166	30.371	33.962
2025 Q1	32.870	31.074	34.665
2025 Q2	32.963	31.168	34.758

Forecasting accuracy for the LSTM model is shown in Table 18. Low RMSE and MAE values show that the predictions made by the model are quite accurate and are almost identical to actual data. Moreover, a low MAPE value indicates that on average there is an error rate of only 2.24% in prediction. This model shows highly efficient forecasting performance and has been found to be much more efficient than conventional methods.

Table 18. Out-of-Sample Forecast Accuracy Measures for LSTM Model.

Model	RMSE	MAE	MAPE
LSTM	0.8195	0.7319	0.0224

As seen in Figure 18 , the out-of-sample forecast of the LSTM model is such that the model follows very closely the actual unemployment rate, giving an accurate representation of its overall level and movements over time. There is no sign of under- or over-prediction on the part of the LSTM model, implying that there is generalisation to the out-of-sample period by the model. Moreover, the 95% confidence intervals are reasonably small and do not widen excessively over the course of the prediction horizon.

The rolling plot for the RMSE is presented in Figure 19 and reflects the dynamics of changes in forecast errors from one period to another. For most of the initial periods, RMSE is low, which is an indication of high accuracy and adequate quality of predictions. At some point, however, the RMSE grows significantly, which might mean that the model becomes somewhat less accurate at some stage. Thereafter, the RMSE starts to decrease, meaning that the accuracy returns back to normal levels. Therefore, there is no growing pattern of RMSE, which means that the forecasting process is fairly stable over time.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0965.v1

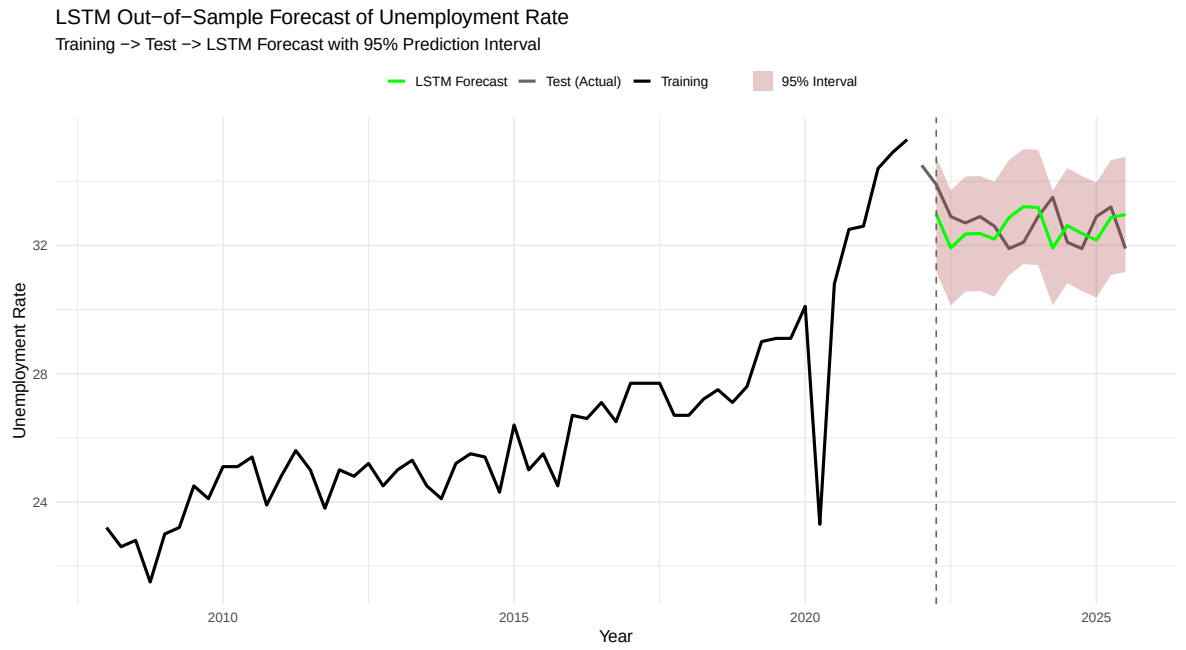

Figure 18. LSTM Out–of–Sample Forecast of Unemployment Rate vs Test (Actual) data.

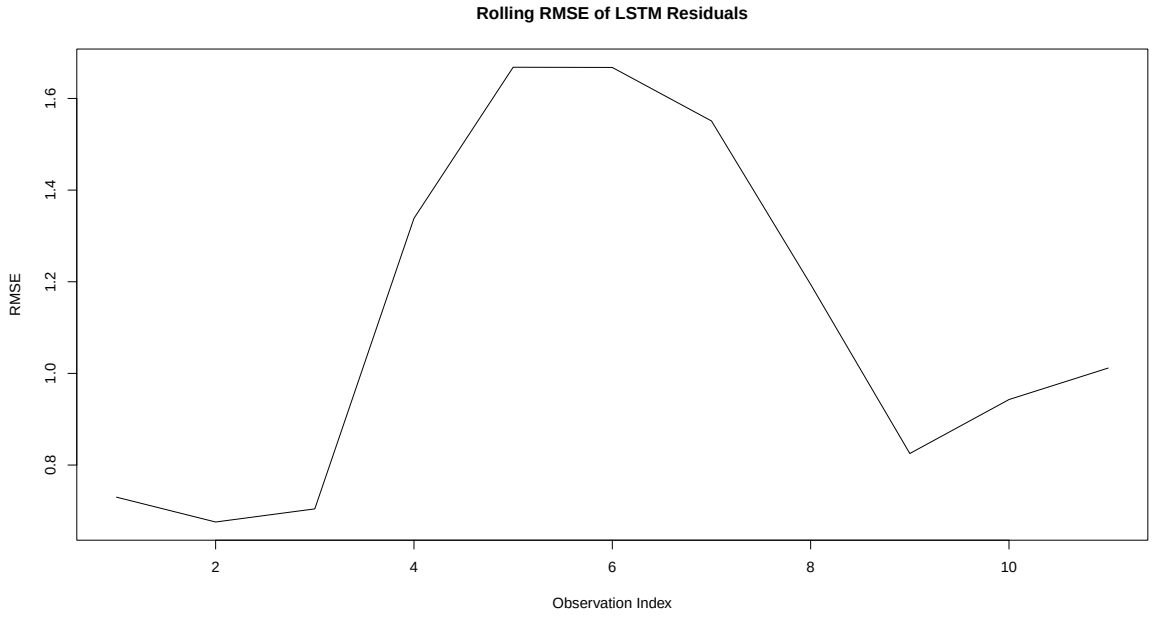

Figure 19. Rolling RMSE Plot of LSTM Residuals.

3.4. Forecast Comparison

Figure 20 shows a comparison of out-of-sample forecasting from the ARIMAX and LSTM models, where the shaded area represents the test interval and the 95% prediction intervals for both models. From the graph, it is evident that the forecasts from the ARIMAX model show an increasing trend and tend to overpredict the unemployment rate, especially in the latter intervals, while having wider prediction intervals as well. In contrast, the LSTM model's predictions appear to be more accurate, with narrower prediction intervals. The results imply that the LSTM model provides better forecasting performance than the ARIMAX model.

 doi:10.20944/preprints202604.0965.v1

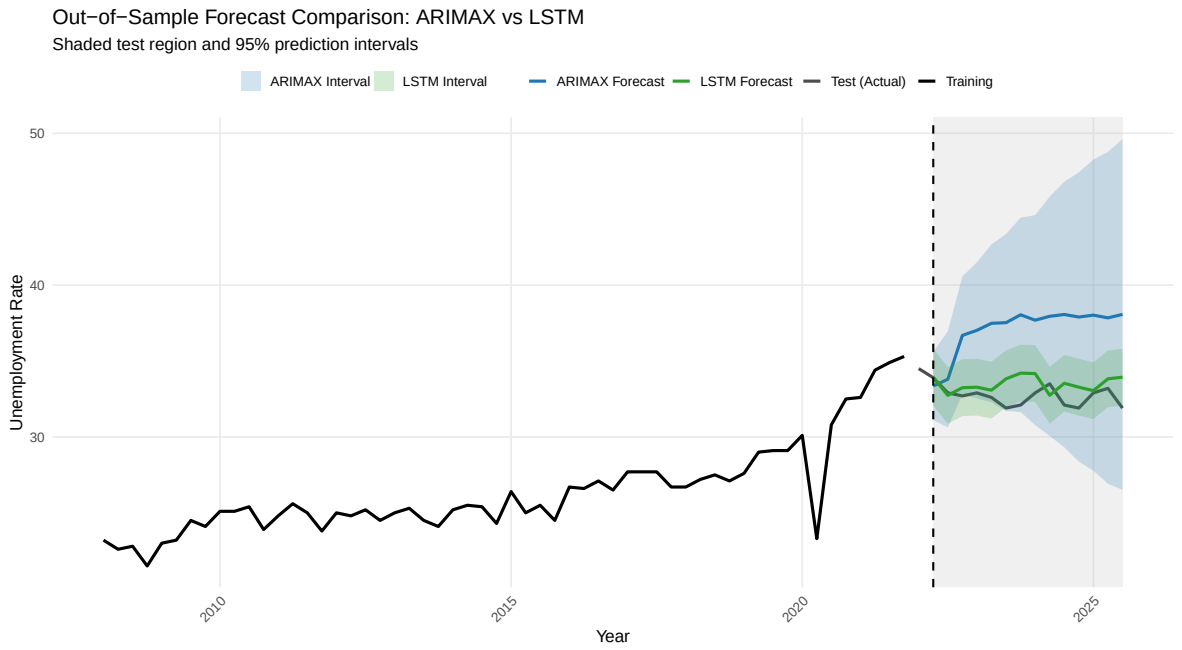

Figure 20. Out–of–Sample Forecast Comparison Plot: ARIMAX vs LSTM.

Table 19 shows the out-of-sample prediction accuracy of both the ARIMAX and LSTM models. It is quite evident that the LSTM model demonstrates a much better performance compared to the ARIMAX model, as evidenced by all performance measures considered. The LSTM model shows remarkably low values for the RMSE and MAE, implying its better accuracy in estimating the changes in the unemployment rates. Furthermore, the remarkably low MAPE value indicates that the LSTM model provides relatively accurate predictions. On the contrary, the ARIMAX model shows relatively high forecast errors, which suggest its inability to capture the nonlinear characteristics of the series, leading to less reliable predictions compared to the LSTM model.

Table 19. Out-of-Sample Forecast Performance Comparison.

Model	RMSE	MAE	MAPE
ARIMAX	4.8171	4.5094	0.1388
LSTM	1.1809	0.9469	0.0293

The results of applying the DM test for comparing the forecasting accuracies of the ARIMAX and LSTM models are provided in Table 20. It can be seen that the DM test value is positive and statistically significant since the p-value is zero. Hence, there is clear evidence of rejecting the null hypothesis concerning the equality of forecast accuracies. This means that the performance of the models differs in terms of accuracy, which supports the superiority of one of them compared to the other. The results confirm the superiority of the LSTM model based on the values of forecast error metrics shown in Table 19.

Table 20. Diebold–Mariano (DM) Test for Forecast Accuracy Comparison.

Statistic	p-value	Test
7.4663	0.0000	DM

Results of the MCS approach for the 5% significance level are demonstrated in Table 21. The ARIMAX model does not belong to the class of superior models according to the obtained zero p-values, meaning that its predictive capability is significantly worse than that of the best-performing model. On the other hand, the LSTM belongs to the group of superior models, having p-values of

 doi:10.20944/preprints202604.0965.v1

one, which means that this model cannot be differentiated statistically from the optimal one. The gap between average losses supports this statement as well, showing a significant advantage of the LSTM model in terms of prediction accuracy. Therefore, only the LSTM model is contained in the confidence set according to the results of the MCS test.

Table 21. Model Confidence Set (MCS) Results at 95% Confidence Level.

Model	Average Loss	p-value (H_{0,M_k})	MCS p-value
ARIMAX	23.2040	0.0000	0.0000
LSTM	1.3945	1.0000	1.0000

Figure 21 illustrates the comparison of forecast accuracy between the ARIMAX and LSTM models through the Murphy diagram for varying values of θ. It can be seen from the Figure 21 that the LSTM curve stays below the ARIMAX curve almost throughout all values of θ, which implies that the mean loss for the LSTM curve is relatively lower, making it more accurate than the ARIMAX curve. On the other hand, the ARIMAX model is associated with much larger values of losses compared to the LSTM model, especially when the value of θ is increased.

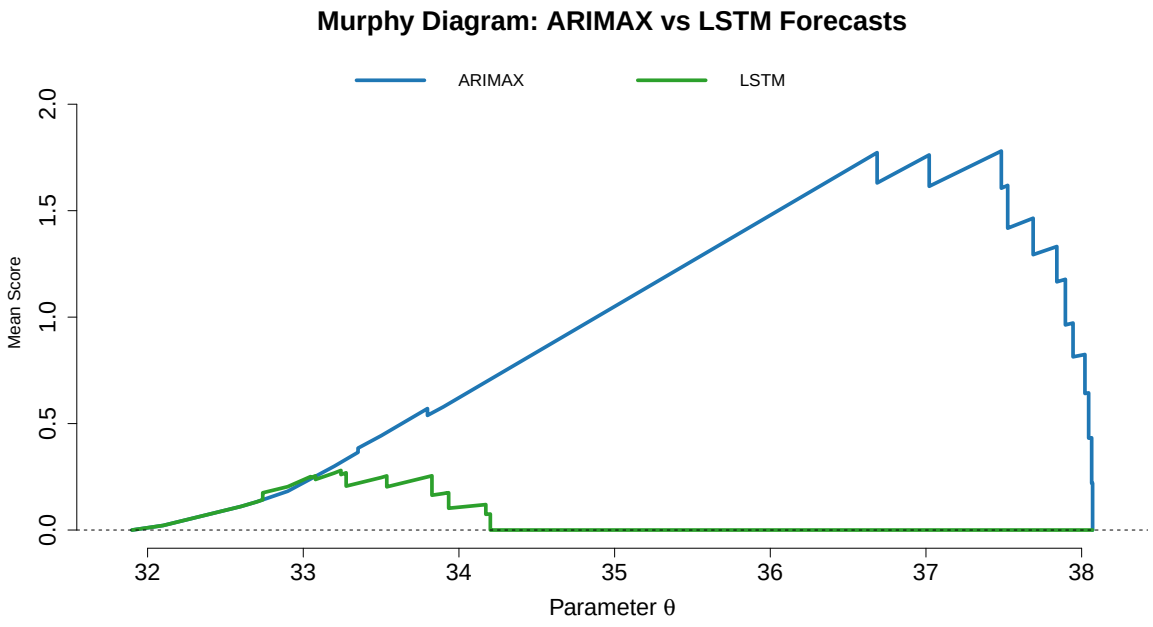

Figure 21. Murphy Diagram: ARIMAX vs LSTM Forecasts.

The bootstrapped confidence intervals for the RMSE values of the competing models are shown in Table 22. In this case, the interval span of the LSTM model is significantly lower than that of the ARIMAX model. This suggests that the LSTM model's predictions are more reliable and accurate than the ARIMAX model's. The fact that the two intervals do not cross indicates that there is a statistically significant difference between the two models' prediction capacities.

Table 22. Bootstrap Confidence Intervals for RMSE.

Model	Lower Bound	Upper Bound
ARIMAX	4.1314	5.3874
LSTM	0.8047	1.5066

Boxplots of forecast errors are shown in Figure 22 for both the ARIMAX and LSTM models. For the ARIMAX model, we observe a significantly smaller median value; most values are negative,

 doi:10.20944/preprints202604.0965.v1

implying systematic overestimation of the unemployment level by the model. In addition, the larger interquartile range and outliers imply a larger variation and instability of predictions for the ARIMAX model. On the other hand, the LSTM model makes superior predictions because the interquartile range is lower and the median error is closer to zero. The plot's findings show that the LSTM model outperforms the ARIMAX model in terms of prediction accuracy and stability.

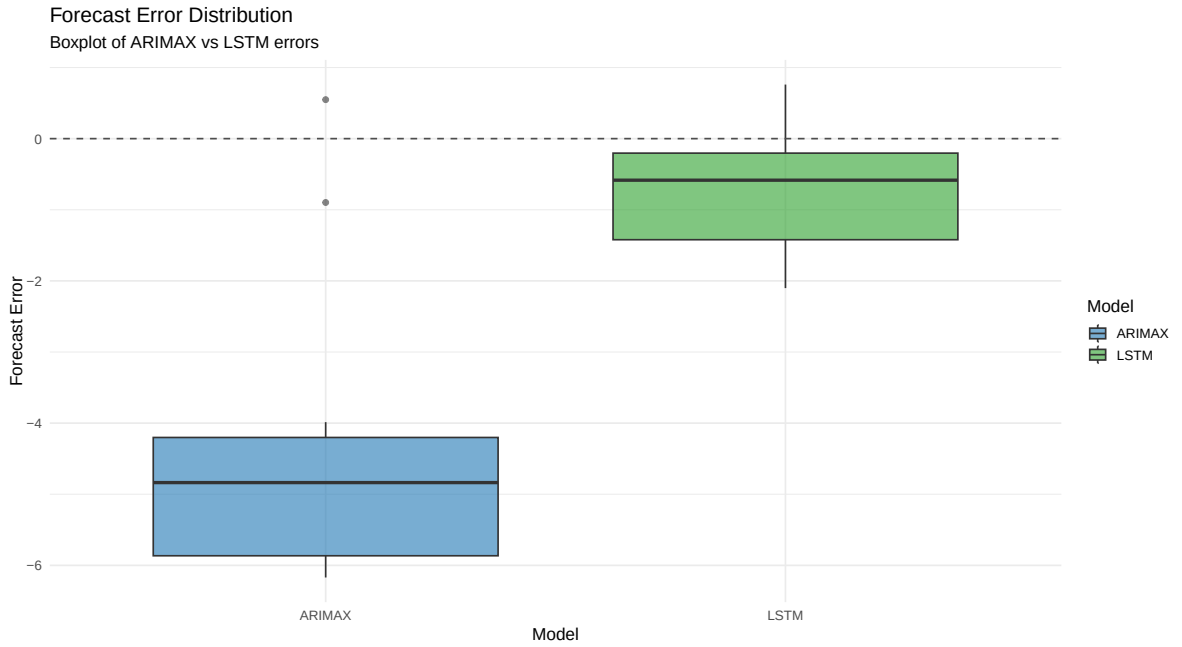

Figure 22. Forecast Error Distribution: Boxplot of ARIMAX vs LSTM errors.

The coverage of the 95% prediction intervals for both models is presented in Table 23. Coverage of the prediction intervals for ARIMAX is quite satisfactory since it approaches the nominal level, meaning that prediction intervals are adequately wide and provide a satisfactory fit for the data. Coverage of the prediction intervals of the LSTM model is much lower, implying that these intervals are relatively narrow and understate uncertainty. Despite being more accurate, the intervals of the LSTM seem to be less effective as probabilistic forecasts compared to those of ARIMAX.

Table 23. Prediction Interval Coverage at 95% Confidence Level.

Model	Coverage (%)
ARIMAX	92.86
LSTM	78.57

The graph presented in Figure 23 represents the cumulative sum (CUSUM) of errors in forecasting for the ARIMAX and LSTM models. It is clear that the CUSUM plot of the ARIMAX model depicts a steep decline towards the negative side. This reveals the presence of a significant number of negative errors in forecasting, which indicates that there may be some bias in the forecasting results generated by the ARIMAX model. On the other hand, the CUSUM plot of the LSTM model stays relatively stable at around the zero line, indicating that there were fewer negative errors in forecasting.

Figure 24 shows how the ARIMAX and LSTM models' predicted errors are distributed. Consistent overestimation and increased prediction variability are suggested by the ARIMAX errors, which are more concentrated in the negative region and have a larger spread. In comparison, the LSTM errors show better accuracy and consistency, as they are more firmly packed around zero, have a narrower distribution, and fewer extreme deviations. Additional evidence of less bias in comparison to the ARIMAX model is the LSTM error distribution's relative symmetry and concentration. The histogram shows that the LSTM model makes better, more accurate predictions.

 doi:10.20944/preprints202604.0965.v1

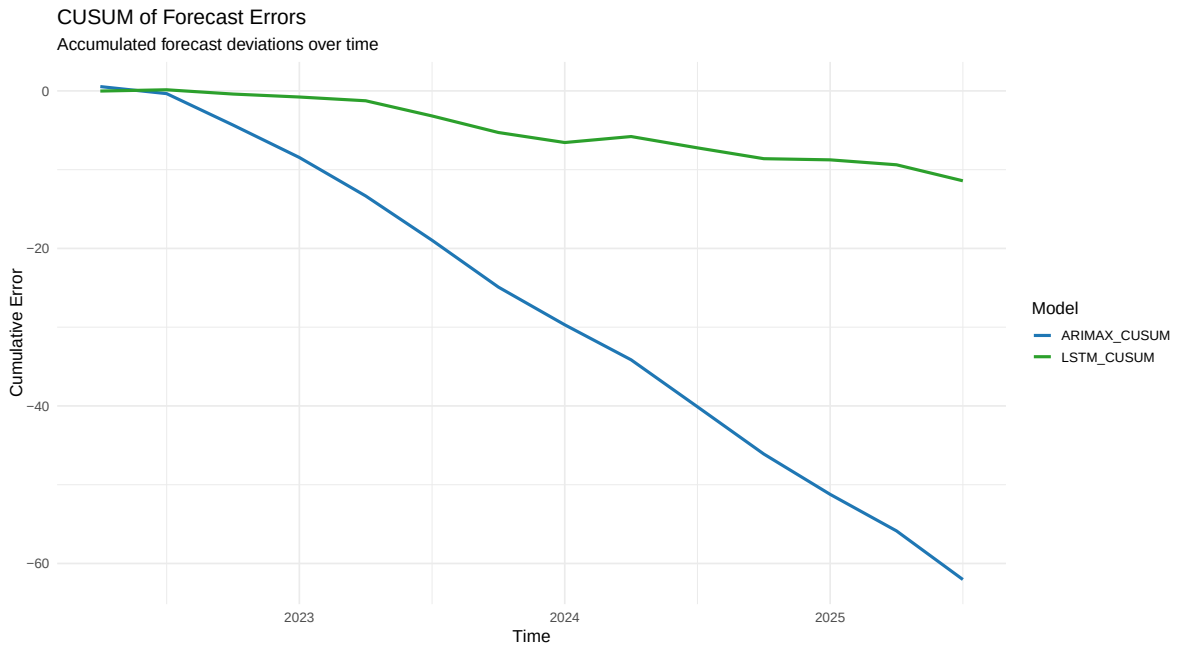

Figure 23. Accumulated forecast deviations over time.

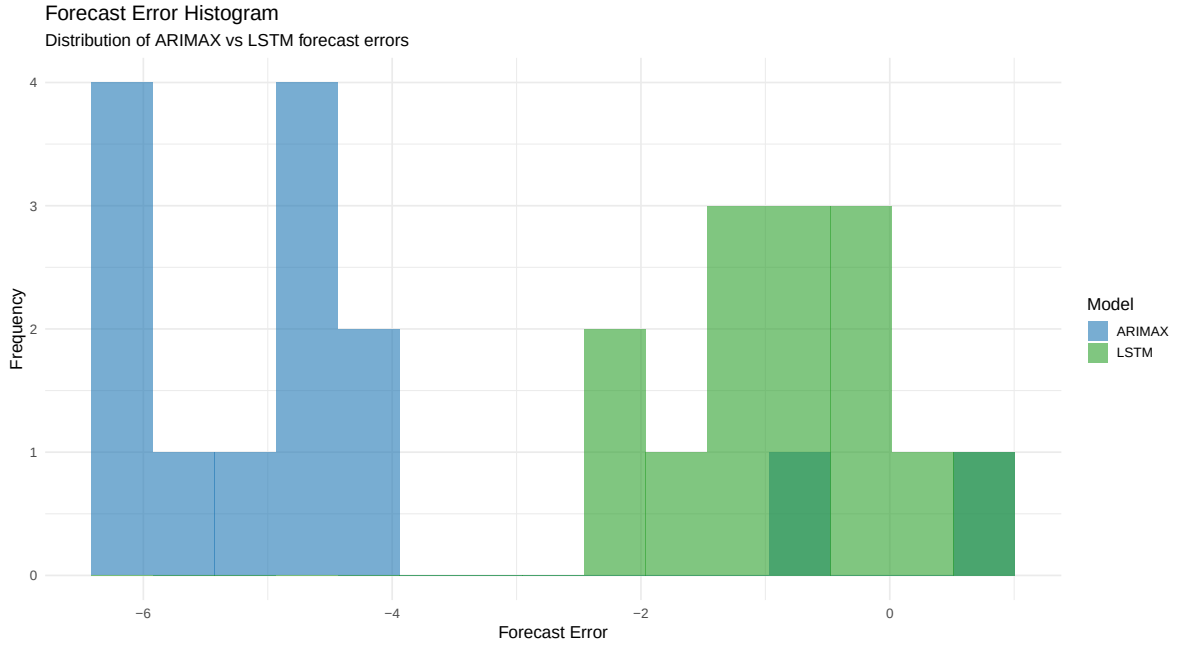

Figure 24. Forecast Error Histogram Plot: Distribution of ARIMAX vs LSTM forecast errors.

Figure 25 shows the comparison of the forecasted value and the actual value for both the ARIMAX and LSTM models. It is observed that the forecasted value using the ARIMAX model is mostly above the 45-degree line, which shows a consistent bias towards an overestimation of the unemployment rate. However, the forecasted value using the LSTM model is comparatively close to the actual value on the 45-degree line, showing reduced bias and high consistency with the actual unemployment rate.

The heatmap in Figure 26 presents the temporal distribution of forecast errors for both the ARIMAX and LSTM models, where blue shades denote underprediction and red shades indicate overprediction. The ARIMAX model exhibits predominantly negative errors throughout the forecast horizon, as evidenced by the concentration of darker blue tones, suggesting a systematic tendency to underestimate unemployment levels with increasing magnitude over time. In contrast, the LSTM model shows lighter and more neutral shades, indicating smaller and more balanced forecast errors.

 doi:10.20944/preprints202604.0965.v1

Although minor instances of overprediction are observed, these are infrequent and less pronounced. The results suggest that the LSTM model provides more stable and less biased forecasts, whereas the ARIMAX model demonstrates a persistent downward bias.

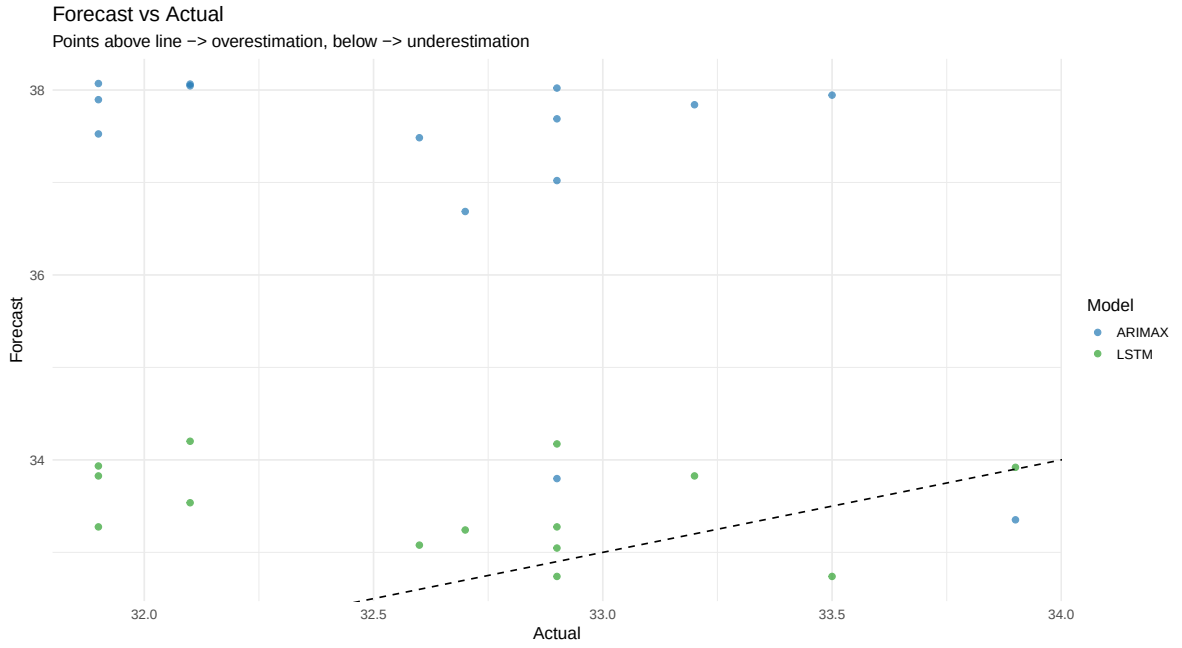

Figure 25. Forecast vs Actual Values for ARIMAX and LSTM Models.

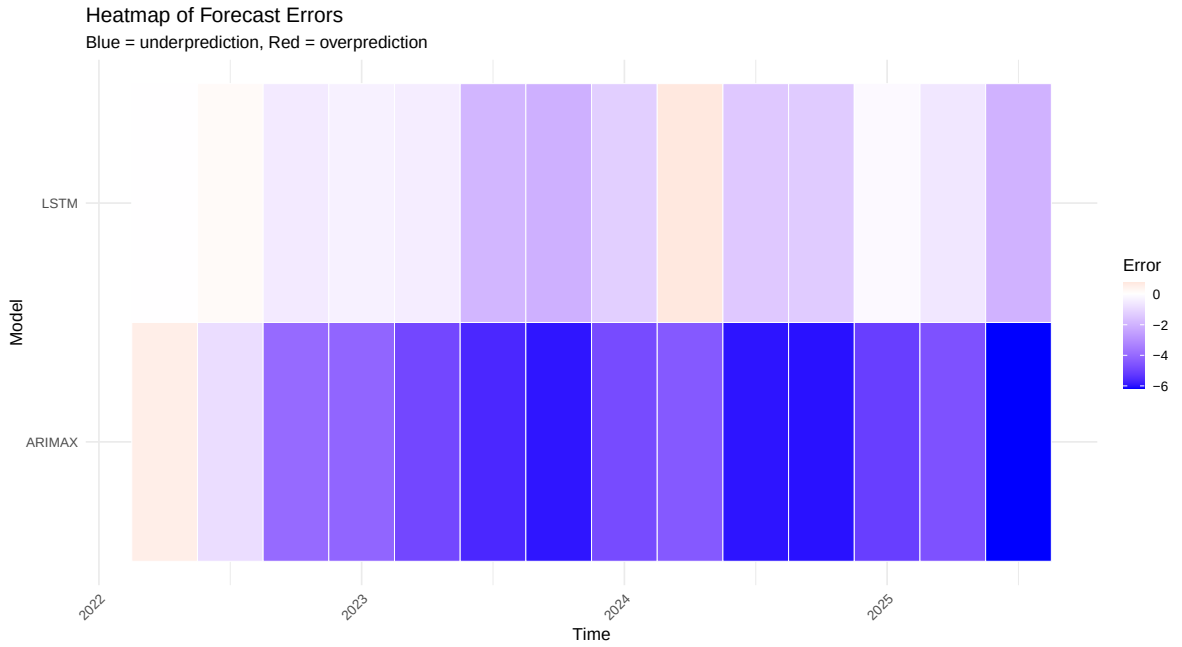

Figure 26. Heatmap of Forecast Errors for ARIMAX and LSTM Models.

The results from non-parametric test statistics for forecast errors as shown in Table 24 clearly show statistically significant differences between forecast errors from ARIMAX and LSTM models. In the first case, the results of the sign test confirm statistical significance with a p-value of 0.0018, meaning that there exists a significant difference in the median value of differences between forecast errors in both approaches. As observed, the negative value of the median means that differences between forecast errors obtained from ARIMAX and LSTM are positive, meaning that ARIMAX forecast error values are higher compared to those from LSTM. The second result comes from the Wilcoxon signed-rank test

 doi:10.20944/preprints202604.0965.v1

with a p-value of 0.0002, implying a significant difference in the distribution of forecast errors between both models.

Table 24. Non-parametric Comparison of Forecast Errors: ARIMAX vs LSTM.

Test	Statistic	p-value	Median / Shift	95% CI
Sign Test	$s = 1$	0.0018	-3.9287	$(-4.5432, -3.5038)$
Wilcoxon Signed-Rank Test	$V = 1$	0.0002	–	–

4. Discussion and Conclusion

4.1. Discussion

The statistical characteristics and predicting behaviour of the South African unemployment rate are amply demonstrated by the empirical results. In line with normal macroeconomic labour market statistics, the exploratory data analysis shows that the series has a strong rising trend, moderate variability, and a minor positive skewness. As seen in Figure 6, the existence of a structural break around 2020 indicates a major external shock that permanently changed the unemployment rate's trajectory. The series' non-stationarity is confirmed by this structural instability as well as the outcomes of the ADF and KPSS tests. Strong persistence is also revealed by the ACF and PACF patterns, indicating an integrated process of order one. Furthermore, as shown in Table 4, the rejection of the null hypotheses in both the Ljung–Box and ARCH LM tests suggests the existence of conditional heteroskedasticity and serial correlation in the raw series. These characteristics support the use of models that can account for both time-varying volatility and dependence. The results of the ARIMAX model show that labour market involvement is essential to understanding the dynamics of unemployment. Unemployment fluctuations are directly and favourably impacted by changes in economic inactivity, as evidenced by the highly significant differenced *NEA* variable. On the other hand, if other factors are taken into account, the *Discouraged* workers variable is not statistically significant, indicating minimal additional explanatory ability. Although residual diagnostics show that the ARIMAX model effectively represents serial dependence, the existence of extreme values and the violation of normality draw attention to the model's limits when it comes to representing complex data aspects. The ARIMAX model's out-of-sample performance exposes significant flaws. The model consistently overestimates unemployment rate and generates progressively wide prediction ranges, showing increased uncertainty over longer horizons, even though it catches the overall upward trend. Its poor predictive ability is confirmed by the comparatively high prediction error metrics (RMSE = 4.8171, MAE = 4.5094, MAPE $\approx$ 13.9%). The LSTM model, on the other hand, performs better in every evaluation dimension. In order to avoid overfitting, the training procedure exhibits quick convergence and efficient regularisation by early stopping. Diagnostic tests verify that the residuals exhibit white noise behaviour and show no signs of heteroskedasticity or autocorrelation. Additionally, reliable forecasts with comparatively small prediction intervals are produced by the LSTM, suggesting increased precision. The LSTM model is heavily favoured by the comparison analysis. It outperforms the ARIMAX model by achieving much reduced forecast errors (RMSE = 0.8195, MAE = 0.7319, MAPE $\approx$ 2.24%). The DM test and the MCS method are two statistical procedures that verify this performance difference is statistically significant rather than the result of chance. The LSTM model regularly produces forecasts that are more accurate, stable, and less biased, according to additional results from bootstrap intervals, Murphy diagrams, and error distributions. Nonetheless, a significant trade-off is noted. Although the LSTM outperforms ARIMAX in point forecasting accuracy, its prediction intervals show less empirical coverage, indicating that it might underestimate forecast uncertainty. Despite great point prediction accuracy, this reveals a weakness in probabilistic forecasting.

 doi:10.20944/preprints202604.0965.v1

4.2. Conclusions

This research compares a traditional econometrics model (i.e., ARIMAX) with a deep learning technique (i.e., LSTM) for predicting the unemployment rate in South Africa. It is found that the unemployment rate quarterly data is non-stationary, unstable, and vulnerable to exogenous shocks; therefore, it requires stronger models for prediction purposes. The ARIMAX model offers interpretability and highlights meaningful economic relationships, especially the importance of the non-economically engaged population. But its forecasting accuracy is restricted because it fails to fully reflect nonlinear dynamics, leading to systematic bias and increasing uncertainty in forecasts. On the other hand, the LSTM model significantly improves predicting accuracy by capturing complicated temporal dependencies and nonlinear patterns in the data. Several statistical tests and robustness checks consistently prove the superiority of the LSTM model, indicating its dependability for short- to medium-term prediction. The outcomes have a significant impact for policymakers and scholars. For policymakers, the increased accuracy of LSTM forecasts offers a more reliable foundation for labour market planning, policymaking, and economic decision-making. The diminished reliability of LSTM prediction intervals suggests that subsequent research ought to focus on improving making decisions. For researchers, the outcomes highlight the limitations of linear models in complex economic systems and the importance of incorporating machine learning approaches into macroeconomic forecasting frameworks. The lower reliability of LSTM prediction intervals indicates that future research should concentrate on enhancing uncertainty quantification in deep learning models. Hybrid techniques that combine the interpretability of econometric models with the predictive capacity of machine learning may provide a viable avenue for future research. This research shows that nonlinear, data-driven techniques are a more effective framework for predicting and forecasting unemployment dynamics in changing economic settings, particularly when compared to traditional linear models that may not capture the complexities of economic fluctuations.

Author Contributions: Conceptualization, I.M. and L.M.; methodology, I.M. and L.M.; software, I.M.; validation, I.M. and L.M.; formal analysis, I.M.; investigation, I.M. and L.M.; data curation, I.M.; writing-original draft preparation, I.M.; writing-review and editing, I.M. and L.M.; visualization, I.M.; supervision, L.M.; project administration, L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data were obtained from the Statistics South Africa (Stats SA) website https://www.statssa.gov.za (accessed on 20 February 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller test
ARIMA	Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with Exogenous Variables
DM	Diebold-Mariano test
EDA	Exploratory Data Analysis
KPSS	Kwiatkowski-Phillips-Schmidt-Shin test
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCS	Model Confidence Set

 doi:10.20944/preprints202604.0965.v1

MSE	Mean Squared Error
NEA	Not Economically Active
PACF	Partial Autocorrelation Function
PICP	Prediction Interval Coverage Probability
RMSE	Root Mean Squared Error
VAR	Vector Autoregression
VIF	Variance Inflation Factor

References

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons. [Google Scholar]
2. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. [CrossRef]
3. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. [CrossRef]
4. Nelson, D. M. Q., Pereira, A. C. M., & De Oliveira, R. A. (2017, May). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 1419–1426. lee. [CrossRef]
5. Medeiros, M., Vasconcelos, G., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119. [CrossRef]
6. Dixon, R. (2024). Unemployment entry, exit and okun's law: An analysis with Australian data. *Australian Journal of Labour Economics*, *27*(2), 161–181. [Google Scholar]
7. D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801–816. [CrossRef]
8. Yurtsever, M. (2023). Unemployment rate forecasting: LSTM-GRU hybrid approach. *Journal for Labour Market Research*, *57*(1),18. [CrossRef]
9. Mero, K., Salgado, N., Meza, J., Pacheco-Delgado, J., & Ventura, S. Unemployment rate prediction using a hybrid model of recurrent neural networks and genetic algorithms. *Applied Sciences*, *14*(8), 3174. [CrossRef]
10. Annastasya, T., Passarella, R., & Yamani, Z. (2025). Unemployment rate forecasting in Indonesia using macroeconomic indicators with a machine learning approach. *Discover Analytics*, *3*(1), 15. [CrossRef]
11. Li, Y. (2025). Comparative Analysis of Time Series Models for Forecasting the U.S. Unemployment Rate: A Study of ARIMA, LSTM, and Intervention Approaches. In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE)*, 372–378. SciTePress. [CrossRef]
12. Nkoane, S. S., & Seeletse, S. M. (2021). Forecasting unemployment rate in South Africa with unexpected events using robust estimators. *International Journal of Economics and Financial Studies*, *13*(2), 199–222. [Google Scholar]
13. Mulaudzi, R., & Ajoodha, R. (2020, December). Application of deep learning to forecast the South African unemployment rate: a multivariate approach. In *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, 1–6. [Google Scholar]
14. Sibarani, A. N. N., Bastiaan, M. E., Ferdinand, F. V., & Edbert, J. S. (2024). Assessing Unemployment Rate Forecasting Accuracy During COVID-19 Using Machine Learning. In *2024 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 744–748. [Google Scholar]
15. Adu, W. K., Appiahene, P., & Afrifa, S. (2023). VAR, ARIMAX and ARIMA models for nowcasting unemployment rate in Ghana using Google trends. *Journal of Electrical Systems and Information Technology*, *10*(1), 12. [CrossRef]
16. Rygh, T., Vaage, C., Westgaard, S., & Lange, P. E. d. (2025). Inflation Forecasting: LSTM Networks vs. Traditional Models for Accurate Predictions. *Journal of Risk and Financial Management*, *18*(7), 365. [CrossRef]
17. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. [CrossRef]
18. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. [CrossRef]

 doi:10.20944/preprints202604.0965.v1

19. Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497. [CrossRef]
20. Bernardi, M., & Catania, L. (2018). The model confidence set package for R. *International Journal of Computational Economics and Econometrics*, *8*(2), 144–158. [CrossRef]