UMYU Scientifica, Vol. 4 NO. 3, September 2025, Pp 115 – 121
https://doi.org/10.56919/usci.2543.012
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina

ORIGINAL RESEARCH ARTICLE

Development of Multivariate Extreme Gradient Boost Technique for Vector Autoregressive Model

Nura Isah[1]*, Sani Ibrahim Doguwa[2], Hussaini Garba Dikko[3] and Bukar Baba Alhaji[4]
[1]Department of Statistics, Collage of Science &Technology, Jigawa State Polytechnic, Dutse, Nigeria
[2&3]Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, Kaduna, Nigeria
[4]Department of Mathematics, Nigerian Defense Academy, Kaduna, Nigeria

ABSTRACT
Machine Learning is a type of Artificial Intelligence (AI) that enables software to obtain models with good prediction results. This research aims to develop the Multivariate Extreme Gradient Boost Technique (XGBoost) for the VAR model. Simulated stationary time series data and a real-time series dataset were applied to model the VAR model to compare the forecast performance of the proposed technique with the conventional VAR Model. Augmented Dickey Fuller (ADF) was applied to test the stationarity of the time series data. The forecast performance for the proposed technique and the existing technique for VAR models for short-term and long-term forecasting would be compared based on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The proposed multivariate Extreme Gradients Boosting technique for the VAR model has been developed by applying the multivariate Extreme Gradients Boost technique to the VAR model. The results of the unit root test for real-life data indicate that all the variables are stationary without any difference. The result for simulated data in terms of forecast performance indicated that the proposed multivariate Extreme Gradients Boosting techniques for VAR model outperform existing technique (conventional VAR model) in long term forecasting base on MAE (1.008 and 1.316) and RMSE (1.357 and 1.669), while the conventional VAR model outperform proposed multivariate XGBoost technique in short term forecast using MAE (1.642 and 2.625) and RMSE (2.016 and 2.652). The results for real life dataset demonstrated that the proposed multivariate XGBoost technique for VAR model is superior in short-term forecast than that of the conventional VAR model using MAE (191.96 and 719.14) and RMSE (267.14 and 1901.49), while for long-term forecasting the result is similar to the short-term with metric value of RMSE (730.57 and 1487.53). The proposed technique for the VAR model is effective in both long-term and short-term forecasting for real-life data.

ARTICLE HISTORY
Received April 13, 2025
Accepted August 28, 2025
Published September 30, 2025

KEYWORDS
Machine Learning, XGBoost Technique, Ensemble Techniques, Development, Regularization. Vector Auto Regressive Model.

INTRODUCTION
Nowadays, machine learning techniques have remarkable success in solving prediction problems due to their robustness to overfitting and flexibility especially ensemble techniques like XGBoost, and random forest (Breiman, 2001). Despite their proven performance in univariate cases and regression problems, the application of XGBoost in the multivariate time series context, particularly the generalization of the VAR framework, remains underexplored (Chen and Guestrin, 2016).

Recent advances in machine learning, particularly ensemble methods like XGBoost, offer an alternative method for modelling nonlinear relationships without strict parametric assumptions. Despite their success in univariate time series forecasting and regression tasks, the application of EGBoost techniques to multivariate time series, particularly in the context of VAR models, remains relatively underexplored. The conventional VAR model plays important roles in modeling and forecasting multivariate time series data and has many applications in different fields. The model has an important role in finance (Tsay, 2005) and econometrics (Sims, 1980).

The development of the multivariate XGBoost technique for VAR models represents an emerging area in time series modeling, combining the strengths of traditional econometric models with advanced machine learning. Below is a list of related literature that explores this area, providing initial knowledge and recent improvements in the field of time series modeling and forecasting.

Jung *et al.* (2008) proposed a method for estimating VAR models using the LASSO technique. The performance of this technique was compared using conventional

Correspondence: Nura Isah. Department of Statistics, Collage of Science &Technology, Jigawa State Polytechnic, Dutse, Nigeria. ✉ nuraisagm@gmail.com.
How to cite: Isah, N., Ibrahim, S. D., Garba, H. D., & Baba, B. A. (2025). Development of Multivariate Extreme Gradient Boost Technique for Vector Autoregressive Model. *UMYU Scientifica*, 4(3), 115 – 121. https://doi.org/10.56919/usci.2543.012

information-based methods such as AIC and BIC, and some other existing subset selection methods with parameter constraints, such as the top-down and bottom-up strategies for simulated and real data for U. S. macroeconomic data. Based on simulation and real data, the results indicate that the LASSO method outperforms other conventional subset selection methods for small samples in terms of prediction mean squared errors and estimation errors under various settings.

Korobilis (2009) proposed a method for estimating sparse VAR models using a Bayesian approach. The technique was computationally efficient for stochastic variable selection in linear and nonlinear VAR. The performance of the proposed variable selection method is assessed in a small Monte Carlo experiment, and in forecasting real data for four UK macroeconomic series using time-varying parameter vector auto-regressions (TVP-VARs). The proposed method outperforms the unrestricted counterparts in forecasting.

Nicholson *et al.* (2015) introduce the VARX-L framework, which applies structured regularization techniques to VAR models with exogenous variables, addressing high-dimensionality challenges in microeconomic forecasting using simulated and macroeconomic data. The results show that the proposed technique demonstrates superior forecast accuracy compared to other traditional VAR and other regularization approaches in low and high-dimensional data.

Billio *et al.* (2019) proposed a new Bayesian nonparametric LASSO prior (BNP-LASSO) for high-dimensional VAR models, which can improve estimation and prediction accuracy. To validate the performance of the new approach on forecasting abilities, the forecast performance was compared with that of BNP-LASSO, Elastic-Net (EN), Bayesian LASSO (B-LASSO), and SSVS using simulated data and real data. The result indicates that the proposed method BNP-LASSO outperforms Elastic-Net (EN), Bayesian LASSO (B-LASSO), and SSVS. Based on the findings, they suggest that the BNP-LASSO is useful not only for a better estimation but could also be used for forecasting purposes in macroeconomics.

Li and Chen (2020) compared the forecast performance of some ensemble techniques, such as random forest, AdaBoost, XGBoost, LightGBM, and Stacking with that of five traditional individual learners (neural network, decision tree, logistic regression, Naïve Bayes, and support vector machine) using real-world credit dataset for Lending Club in the United States. The findings indicate that the forecast performance of ensemble learning is better than individual learners.

Zhai et al. (2021) explore a hybrid approach to forecast an industrial setting by combining XGBoost and Gated Recurrent Units (GRU). XGBoost captures complex nonlinear relationships and handles structured data, while GRU is used to model temporal dependencies in data. The proposed technique was applied to predict temporal dependencies in heating finance. The results shows that the proposed technique outperformed individual models using only XGBoost or GRU.

Suotsalo *et al.* (2021) proposed a novel method called pseudo-likelihood Vector Auto-regressive model (PLVAR) by combining fractional marginal likelihood and pseudo-likelihood. The method can decide the complete VAR model structure, including the lag length, in a single run. The performance of the PLVAR method was compared with that of the smoothly clipped absolute deviation (SCAD) methods, LASSO, and unrestricted VAR models. The PLVAR method is both faster and produces more accurate estimates than the other methods based on penalized regression. The proposed technique outperforms other methods on both simulated and real data.

Sun *et al.* (2022) proposed a hybrid model by combining the informer (a transformer-based deep learning model), XGBoost, and Genetic Algorithms (GA) for multi-step time series forecasting. The informed component captures long-term dependencies, XGBoost captures non-linear relationships, and GA optimizes the ensemble weight. The proposed approach shows superior forecast accuracy compared to traditional models.

Lubbers (2023) compared the prediction performance for XGBoost and Random Forest techniques using real data on cash flow for transactions of small and medium-sized enterprises. The research intended to identify the forecast performance of the two different algorithms and assess their feasibility for practical use in their daily operations. The result of the two algorithms showed that the Random Forest technique outperformed XGBoost, but the performance of the two techniques varied depending on the training data used.

Sundari and Mahardika (2024) developed a predictive model that can predict house prices accurately based on relevant features. They adopted some ensemble learning techniques, including Gradient Boosted Regression Trees (GBRT), LASSO, and Extra Gradient Boosted Technique (XGBoost), using the Ames Housing data set. The performance of the predicted model was evaluated using Root Mean Square Error (RMSE). However, the results indicate that the combination of two ensemble techniques (GBRT and XGBoost) outperforms other methods in predicting housing prices.

This study proposed a novel extension of the XGBoost technique for VAR models by combining the structure of conventional VAR models with the multivariate XGBoost proposed by Guang (2021). The proposed technique aims to bridge the gap between the conventional VAR model and modern machine learning approaches. The technique captures linear interdependencies across multiple time series, improves forecast accuracy, and model flexibility (Rahman and Davis, 2013). The performance of the proposed technique would be validated on simulated and real-life financial datasets and compared with the conventional VAR model.

This paper contributes to the literature by developing a Multivariate XGBoost technique for the VAR model

framework that captures nonlinear cross-lagged relationships and enhances the ensemble power of XGBoost for a robust forecast.

METHODOLOGY

Vector Autoregressive Model (VAR)

The VAR model is one of the most successful, flexible, and easy models to use for the analysis of multivariate time series. It is a natural extension of the univariate autoregressive model to dynamic multivariate time series. VAR model is useful for describing the dynamic behavior of economic and financial time series and for forecasting (Lutkepohl, 2005). Consider a column vector for k different time series variables:

$$\boldsymbol{Y}_t = (\boldsymbol{y}_{1t}, \boldsymbol{y}_{2t}, \boldsymbol{y}_{3t}, \ldots, \boldsymbol{y}_{kt})^1 \quad (1)$$

and the model is in terms of past values of the vector. The result is a vector autoregressive or VAR. The VAR(p) process is of the form:

$$\boldsymbol{Y}_t = \boldsymbol{M} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \boldsymbol{A}_2\boldsymbol{Y}_{t-2} + \ldots + \boldsymbol{A}_p\boldsymbol{Y}_{t-p} + \boldsymbol{\varepsilon}_t \quad (2)$$

where

$$\boldsymbol{A}_i = \begin{pmatrix} a_{11}^{(r)} & a_{12}^{(r)} \ldots & a_{1k}^{(r)} \\ a_{21}^{(r)} & a_{22}^{(r)} \ldots. & a_{2k}^{(r)} \\ a_{31}^{(r)} & a_{32}^{(r)} \ldots & a_{3k}^{(r)} \end{pmatrix} \qquad i, j = 1, \ldots, k; r = 1, \ldots, p \quad (3)$$

A_i is a k by k square matrix of coefficients; $\boldsymbol{M}$ is a k by 1 column vector, and $\boldsymbol{\varepsilon_t}$ is a k by 1 column vector of white noise process, with the properties that:

$E(\boldsymbol{\varepsilon_t}) = 0$ for all t

$$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_s) = \begin{cases} \boldsymbol{v}, if s = t \\ \boldsymbol{0}, if s \neq t \end{cases} \quad (4)$$

where $\boldsymbol{v}$ the covariance matrix is assumed to be positive definite. Thus, the $\boldsymbol{\varepsilon_t}$are serially uncorrelated but may be contemporaneously correlated.

For $k = 3$ and $p = 2$ Equation 3.7 becomes:

$$\begin{pmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{pmatrix} = \begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix} + \begin{pmatrix} a_{11}^1 & a_{12}^1 & a_{13}^1 \\ a_{21}^1 & a_{22}^1 & a_{23}^1 \\ a_{31}^1 & a_{32}^1 & a_{33}^1 \end{pmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \\ y_{3t-1} \end{pmatrix} + \begin{pmatrix} a_{11}^2 & a_{12}^2 & a_{13}^2 \\ a_{21}^2 & a_{22}^2 & a_{23}^2 \\ a_{31}^2 & a_{32}^2 & a_{33}^2 \end{pmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \\ y_{3t-2} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \quad (5)$$

Multivariate XGBoost Method

The multivariate XGBoost method was proposed by Guang (2021) by generalizing the XGBoost method with multi-objective functions and forms multi-objective parameters regularized tree boosting.

For a given data set: $\{\boldsymbol{D} = (\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n, \boldsymbol{x}_i \in \mathbb{R}, \boldsymbol{y}_i \in \mathbb{R}\}$, contained m features and n sample points.

For any sample point $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ consider $l(\theta_{1i}, \ldots, \theta_{li}; y_i)$ as l- variable loss function and $\theta_{1i}, \ldots, \theta_{li}$is the independent variables, the value range of each $\theta_{ji}(j = 1,2, \ldots, l)$ is a subinterval of $\mathbb{R}$. In multi-objective parameter regularized tree boosting method.

Consider $\theta_{1i}, \ldots, \theta_{li}$as loss function parameters to be estimated in the multi-objective parameter regularized tree boosting model, add K_j tree functions to obtain the predicted result of the parameter $\theta_{ji}(j = 1,2, \ldots, l)$ of $l(\theta_{1i}, \ldots, \theta_{li}; \boldsymbol{y}_i)$.

$$\widehat{\boldsymbol{\theta}}_{ji} = \phi_j(\boldsymbol{x}_i) = \sum_k^{k_j} f_{\theta_{jk}}(\boldsymbol{x}_i), \quad (6)$$

where $F = \{f(x) = \omega_{q(x)}\}$ is the space of regression trees. q represents the structure of each tree, which associates a sample with a corresponding leaf index. T is the number of leaves in the tree. Each $f_{\theta_{jk}}$is similar to an independent tree structure q and leaf weights ω. To study these tree functions in the model, the following regularization objectives are minimized:

$$l(\theta_1, \ldots, \theta_l) = \sum_{i=1}^{l} l(\theta_{1i}, \ldots, \theta_{li}; \boldsymbol{y}_i) + \sum_{k_1} \Omega_{\theta_1}(f_{k_1}) +, \ldots, + \sum_{k_l} \Omega_{k_l}(f_{k_l}) \quad (7)$$

where

$$\boldsymbol{\Omega}_{\theta_r}(f_{k_r}) = \boldsymbol{\gamma}_{\theta_r}\boldsymbol{T}_{\theta_r} + \frac{1}{2}\boldsymbol{\lambda}_{\theta_r}(\omega)^2, r = 1,2, \ldots, l \quad (8)$$

$$\Omega_{\theta_1}(f_{k_1}) = \gamma_{\theta_1} T_{\theta_1} + \frac{1}{2}\lambda_{\theta_1}(\omega)^2$$

$\boldsymbol{\gamma}_{\theta_j}$and $\boldsymbol{\lambda}_{\theta_j}$are respectively the regularization parameters of $\boldsymbol{T}_{\theta_j}$and $\phi_j = (\omega)^2$

The objective function for t th iteration is

$$l_t(\theta_1, \ldots, \theta_l) = \sum_{i=1}^{l} l(\theta_{1i}, \ldots, \theta_{li}; \boldsymbol{y}_i) + \sum_{i=1}^{l}[g\{f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_l)\} + \frac{1}{2}h\{f(\boldsymbol{x}_1)^2, \ldots, f(\boldsymbol{x}_l)^2\}] + \sum_{i=1}^{l} \boldsymbol{\gamma}_{\theta_i}\boldsymbol{T}_{\theta_i} + \frac{1}{2}\boldsymbol{\lambda}_{\theta_l}(\omega)^2 \quad (9)$$

A maximum of l trees could be trained simultaneously in each iteration of training, that is, all the parameters could be estimated simultaneously. Each tree corresponds to one parameter to be estimated and has its own independent hyperparameters.

The Proposed Multivariate XGBoost Technique for the VAR Model

The proposed technique would be developed by hybridizing the multivariate XGBoost method proposed by Guang (2021) and the conventional VAR model to estimate the Models. Thus, by estimating all regression equations simultaneously. The model estimation and forecasting can be more efficient compared to solving each regression equation independently, as noted by Evgeniou and Pontil (2004). The limitation of the proposed technique would only apply to multivariate data, and the data should be a time series. For a given train data set: $\{\boldsymbol{D} = (\boldsymbol{y}_{it}, \boldsymbol{y}_{jt-r}), i, j = 1, \ldots, k; r = 1, \ldots, p; \boldsymbol{y}_{it} \in \mathbb{R}, \boldsymbol{y}_{jt-r} \in \mathbb{R}\}$contained k features.

For any sample point$(\boldsymbol{y}_{it}, \boldsymbol{y}_{jt-r})$, consider $l(\boldsymbol{y}_{1t}, \ldots, \boldsymbol{y}_{kt}; \boldsymbol{y}_{it})$ as l- variable loss function and$\boldsymbol{y}_{1t-r}, \ldots, \boldsymbol{y}_{kt-r}$, are the independent variables, Consider $\theta_{1t}, \ldots, \theta_{kt}$ as loss function parameters to be estimated in the multi-objective parameter regularized tree boosting model, add K_j tree functions to obtain the predicted result of parameter $\boldsymbol{\theta}_{it}(i = 1, \ldots, k)$of $l(\boldsymbol{y}_{1t}, \ldots, \boldsymbol{y}_{kt}; \boldsymbol{y}_{it})$.

$$\widehat{\boldsymbol{\theta}}_{it} = \phi_{it}^{r}(\boldsymbol{y}_{ji-r}) = \sum_{j=1}^{k}\sum_{r=1}^{p} f\,(\boldsymbol{y}_{jt-r}) \quad (10)$$

where $\{f(\boldsymbol{y}_{ijt-r}) = \omega_{q(\boldsymbol{y}_{ijt-r})}\}$ was the space of regression trees. q represents the structure of each tree, which associates a sample with a corresponding leaf index. T was the number of leaves in the tree. Each $f_{\theta_{jk}}$ is similar to an independent tree structure q and leaf weights ω. To study these tree functions in the model, minimize the following regularization objectives.

$$\theta_{1t}, \ldots, \theta_{kt} = \sum_{j=1}^{k}\sum_{r=1}^{p} l[(\boldsymbol{y}_{1t-r}, \ldots, \boldsymbol{y}_{kt-r}; \boldsymbol{y}_{jt-r})] + \sum_{j=1}^{k}\sum_{r=1}^{p} \Omega\, f(\boldsymbol{y}_{jt-r}) \quad (11)$$

where Ω represents the regularization term, a factor used to measure the complexity of the tree $f(\boldsymbol{y}_{jt-r})$.

We can obtain the optimum $f(\boldsymbol{y}_{jt-r})$ by adding first and second order for gradian statistics for each loss function to minimized the objective functions.

$$\theta_{1t}, \ldots, \theta_{kt} = \sum_{j=1}^{k}\sum_{r=1}^{p} l[(\boldsymbol{y}_{1t-r}, \ldots, \boldsymbol{y}_{kt-r})] + \sum_{j=1}^{k}\sum_{r=1}^{p}[g\{f(\boldsymbol{y}_{1t-r}), \ldots, f(\boldsymbol{y}_{kt-r})\} + \frac{1}{2}h\{f(\boldsymbol{y}_{1t-r})^2, \ldots, f(\boldsymbol{y}_{kt-r})^2\}] + \sum_{j=1}^{k}\sum_{r=1}^{p} \Omega f(\boldsymbol{y}_{jt-r}) \quad (12)$$

where *h* and *g* are first and second order of loss functions.

and $\Omega f(\boldsymbol{y}_{jt-r}) = \boldsymbol{\gamma}_{jt-r}\boldsymbol{T}_{jt-r} + \frac{1}{2}\boldsymbol{\lambda}_{jt-r}\boldsymbol{\omega}_{jt-r}^2$

where $\boldsymbol{\gamma}_{jt-r}$and $\boldsymbol{\lambda}_{jt-r}$are the degrees of regularization. T_{jt-r} and $\boldsymbol{\omega}_{jt-r}$ are the numbers of leaves and the vector of values attributed to each leaf, respectively.

By removing the constant terms, we have:

$$\theta_{1t}, \ldots, \theta_{kt} = \sum_{j=1}^{k}\sum_{r=1}^{p}[g\{f(\boldsymbol{y}_{1t-r}), \ldots, f(\boldsymbol{y}_{kt-r})\} + \frac{1}{2}h\{f(\boldsymbol{y}_{1t-r})^2, \ldots, f(\boldsymbol{y}_{kt-r})^2\}] + \sum_{j=1}^{k}\sum_{r=1}^{p}\{\boldsymbol{\gamma}_{jt-r}\boldsymbol{T}_{jt-r} + \frac{1}{2}\boldsymbol{\lambda}_{jt-r}\boldsymbol{\omega}_{jt-r}^2\} \quad (13)$$

This is the derivation for the proposed Multivariate XGBoost technique for the VAR model to obtain the best model based on the forecast performance.

RESULTS

Simulated Data Results

The simulated data contained 12 features, each with 100 samples. Data points were obtained using the Python package. The conventional VAR model was estimated using information-criteria-based methods, such as AIC and BIC, and the proposed VAR model (using multivariate XGBoost techniques).

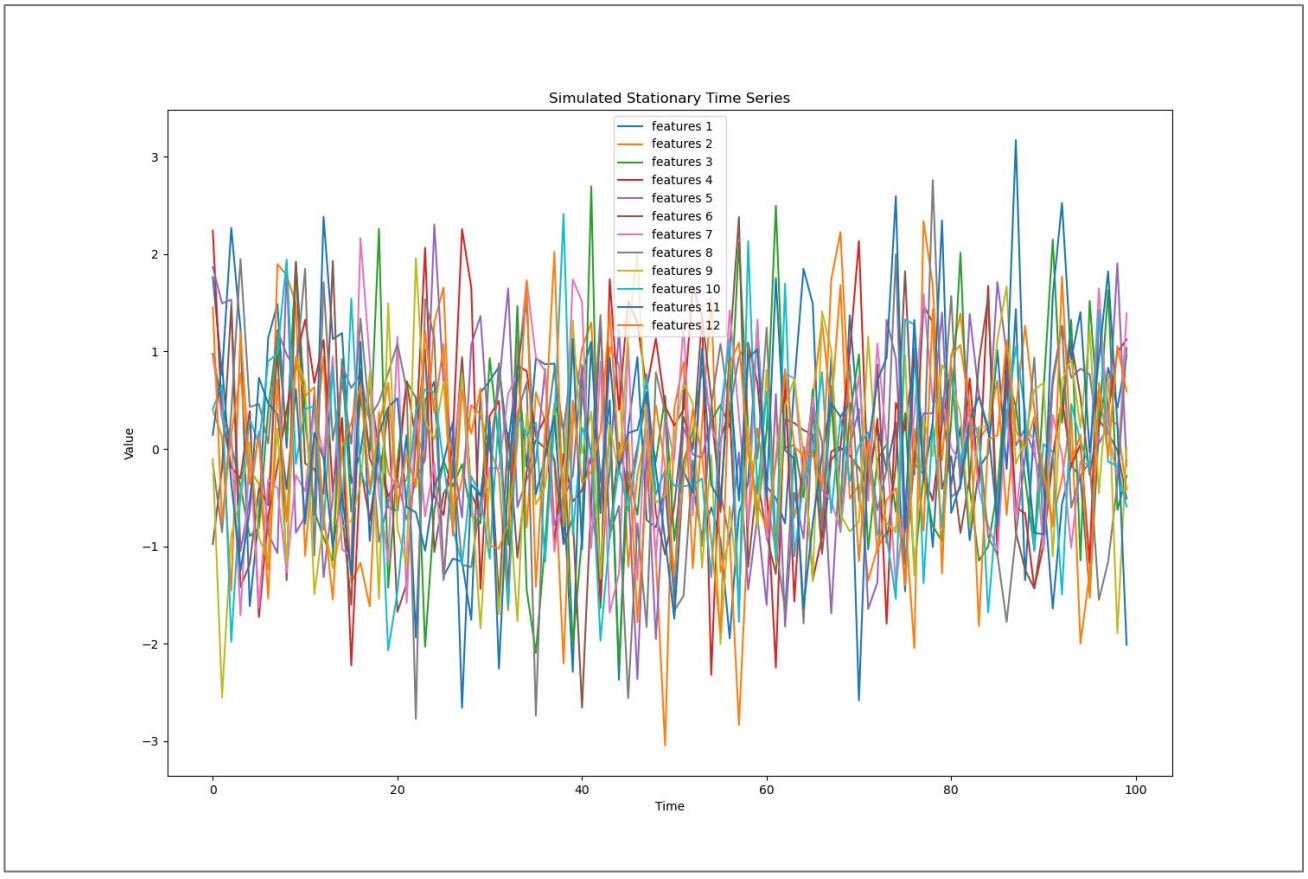

Figure. 1: Time series plot for 12 monthly simulated multivariate time series data

The forecast performance for proposed and existing techniques for estimated VAR models using MAE and RMSE would be computed and compared based on short-term and ten-step-ahead long-term forecasts.

The nature of the simulated time series graph in Figure 1 indicates the existence of stationarity in the time series data at level.

Lag order selection

First, we undertake a VAR Lag Order selection process. The results for various selection criteria are presented in Table 1.

Table 1: Lag order selection

Lag Order	AIC	BIC	HQIC
0	-1.096	-0.7249*	-0.9476
1	0.05013	5.322	2.426
2	1.732	11.00	5.434
3	2.268	15.55	7.746
4	-0.4128	17.76	6.842
5	-10.74*	11.88	-1.710*

Table 1 shows that, based on AIC and HQIC, lag 5 was selected as the optimal lag for estimating the VAR Models.

Forecast Performance for conventional VAR Models and the Proposed Multivariate XGBoost technique for the VAR model

After estimating the conventional and proposed techniques for VAR models, the forecast performance for the estimated VAR models will be estimated using short-term and long-term forecast horizons based on MAE and RMSE. The summary results are shown in Table 2.

Summary of Results for Simulated Data

From Table 2, the conventional VAR model outperforms the proposed Multivariate XGBoost technique for the VAR Model in short-term forecast, since the conventional VAR model has a smaller value of MAE and RMSE than the proposed technique. The result for long-term forecasting shows that the proposed Multivariate XGBoost technique for VAR (MXGB) model demonstrates a high forecast accuracy compared to the conventional VAR model, since the proposed technique has the least value of MAE and RMSE compared to the conventional VAR model for simulated data.

Results for Real Dataset

The real dataset consists of monthly data for Nigerian financial time series spanning the period January 2010 to July 2024, a total of 175 data points to estimate the models. The data for all the variables were obtained from the Central Bank of Nigeria website are considered in fitting the estimated models. In this research, nine years of Nigeria's trade in goods and services time series data, categorized into visible goods and services, are considered in estimating the models. The nine variables are: *Oil Sector* (OILSE), *Log of Financial Sector* (LFINS), *Industrial Sector* (INDSE), *Balance of Trade* (BOT), *Log of Commercial Services* (LCOMS), *Transport Services* (TRANS), *Tourism and Travel services* (TTS), *Health and related Social services* (HRSS), and *Mining Sector* (MINSE).

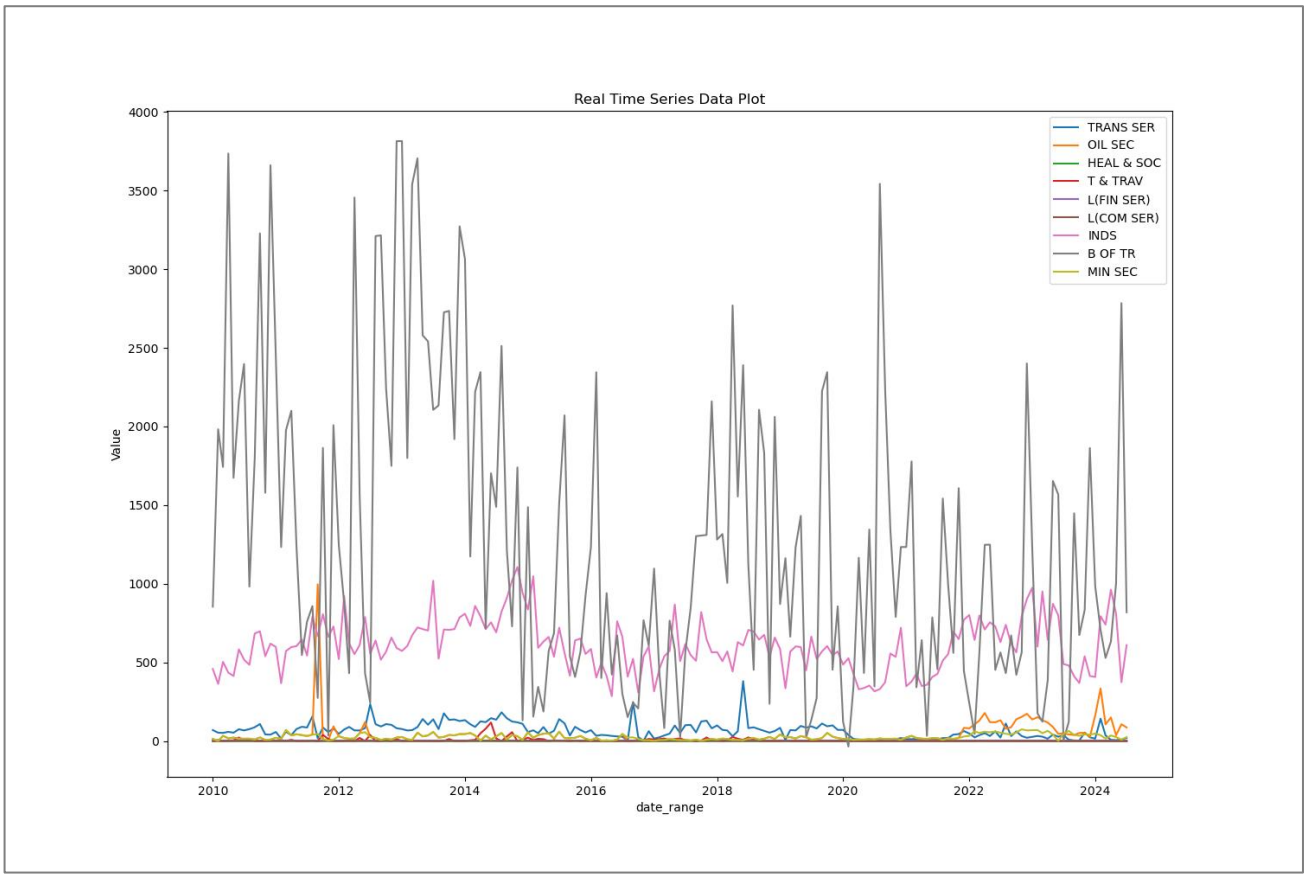

Figure 2: Time series plot for 9 monthly Trade in Goods and Services data

Table 2: Forecast performance for existing and proposed techniques based on short-term forecast

Models	Short-term Forecast		Long-term Forecast	
	MAE	RMSE	MAE	RMSE
Conventional VAR Model	1.642	2.016	1.316	1.669
VAR (MXGB) Model	2.625	2.652	1.008	1.357

Table 3: Augmented Dickey Fuller Unit root test for real dataset

Variables	**ADF Statistic**	**p-value**	**Order of Integration**
MINSE	-3.87861	0.002198 **	I(0)
OILSE	-5.516953	1.915×10^{-6} **	I(0)
HRSS	-12.33116	6.422×10^{-23} **	I(0)
TTS	-4.08723	0.001017 **	I(0)
INDSE	-3.07069	0.02880 **	I(0)
BOT	-3.41569	0.010436 **	I(0)
MINSE	-3.52964	0.007258**	I(0)
FINS	-3.52964	0.007258 **	I(0)
COMS	-2.34250	0.158626 *	I(0)
TRANS	-2.32547	0.163885 *	I(0)
LFINS	-3.87670	0.002198 **	I(0)
LCOMS	-3.98152	0.001510 **	I(0)

Note: ** indicates the variable is significant at 5% and * indicates the variable is significant at the 10% level. The null hypothesis of a unit root test is rejected, implying that all the variables are stationary. The null hypothesis states that there is the existence of unit root in the data.

Table 4: Lag order selection

Lag Order	**AIC**	**BIC**	**HQIC**
0	44.25	44.47*	44.34
1	43.54	45.77	44.46
2	43.97	48.15	45.67
3	44.55	50.70	47.04
4	45.01	53.14	48.31
5	45.27	55.38	49.37
6	44.98	57.06	49.88
7	45.55	59.61	51.25
8	44.91	60.94	51.41
9	43.00	61.02	50.31
10	39.11	59.10	47.22
11	30.34*	52.31	39.25*

The lag order selection results show that, based on AIC and HQIC, lag 11 is selected as the optimal lag for estimating the VAR Models.

Table 5: Comparative Analysis on the forecast performance for existing and proposed techniques based on short-term and long-term forecasts

Models	Short-term Forecast		Long-term Forecast.	
	MAE	RMSE	MAE	RMSE
Conventional VAR Model	719.14	1901.49	428.81	1487.53
VAR (MXGB) Model	191.96	267.14	451.23	730.57

Figure 2 presents the monthly plot of the financial time series data, while Table 3 reports the corresponding unit root test results. Both the visual inspection of the time series plot and formal diagnostic testing confirm stationarity for all nine time series at level, as evidenced by ADF Test result meeting conventional significance thresholds ($p < 0.05$).

Unit Roots test

A stationary series must be obtained before it can be used to specify and fit a model. The unit roots test, which has as the null hypothesis of existence of a unit root versus the alternative hypothesis of the nonexistence of a unit root, will help us to determine the stationarity of a series. The

Augmented Dickey-Fuller (1981) was used to test for the stationarity of the series. The results for the unit root test presented in Table 3 indicate that all the variables are stationary at the level, except FINS and COMS, which are stationary using a log transformation. Thus, all the 9 variables are stationary.

Lag order selection

First, we undertake a VAR Lag Order selection process. The results for various selection criteria are presented in Table 4.

Forecast Performance for conventional VAR Models and the Proposed Multivariate XGBoost technique for the VAR model

After estimating the conventional VAR model based on information criteria and proposed technique for VAR models, the forecast performance for the estimated VAR models would be compared using short-term and long-term forecast horizons based on MAE and RMSE. The summary results are shown in Table 5.

Summary Result for Real Data

Result in Table 5, indicate that the proposed Multivariate XGBoost technique for VAR model outperform the conventional VAR Model in short-term forecast, since the proposed technique has smaller value of MAE and RMSE than that of conventional VAR model in both short-term, the result for long-term forecasting indicated that the proposed technique outperformed the conventional VAR model, since the metric value of RMSE for proposed technique is smaller than that of conventional VAR model for real dataset.

CONCLUSION

In this study, the proposed multivariate Extreme Gradients Boosting technique for the VAR model was developed by hybridizing the multivariate Extreme Gradients Boost technique to the VAR model. The results for simulated data indicated that the proposed multivariate Extreme Gradients Boosting techniques for VAR model outperform the existing (conventional VAR model) in long-term forecasting, while the conventional VAR model is superior in terms of forecast accuracy than the proposed multivariate XGBoost technique in short-term forecasting based on MAE and RMSE. The results for the real dataset indicated that the proposed multivariate XGBoost technique for the VAR model outperformed the conventional VAR model in terms of forecast accuracy in both short-term and long-term forecasting, as measured by MAE and RMSE.

RECOMMENDATION

Based on the findings of this study, it is recommended that:

1. The multivariate XGBoost technique is the best technique for short-term and long-term forecast in real-life time series data for the VAR model.
2. The proposed techniques are to be considered as alternative techniques for VAR models in modelling financial time series data.

REFERENCES

Billio, M., Casarin, R., & Rossini, L. (2019). Bayesian nonparametric sparse VAR models. *Journal of Econometrics*, 212, 97–115.[crossref]

Chen, T., & Guestrin, C. (2016). A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. .[crossref]

Evgeniou, T., & Pontil, M. (2004). Regularized multitask learning. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 109–117. [crossref]

Guang, Y. (2021). *Generalized XGBoost method*. Blue Print.

Jung, N. H., Lin, H. H., & Mei, Y. C. (2008). Subset selection for vector autoregressive processes using LASSO. *Journal of Computational Statistics and Data Analysis*, 52(7), 3645–3657. [crossref]

Korobilis, D. (2009). *VAR forecasting using Bayesian variable selection*. University of Strathclyde and Rimini Center for Economic Analysis Review. [crossref]

Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Tianjin University Review, China*. [crossref]

Lubbers, L. (2023). *Comparing the performance of XGBoost and random forest models for predicting company cash position* [Master's thesis, Utrecht University].

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer. [crossref]

Nicholson, W., Matteson, D., & Bien, J. (2015). *Structured regularization for large VAR with exogenous variables*. arXiv preprint. arXiv:1508.07497.

Rahman, M. G., & Davis, D. N. (2013). Addressing the data imbalance problem in software defect prediction using cost sensitive learning. *International Journal of Machine Learning and Computing*, 3(4), 333–336. [crossref]

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1), 1–48. [crossref]

Sun, C., Chain, Z., Qin, Y., & Want, B. (2022). Multi-steps time series forecasting based on Informer - XGBoost - GA. *Journal of Physics: Conference Series*, 2333(1), 012009. [crossref]

Sundari, P. S., & Mahardika, K. P. (2024). Optimization house price prediction model using gradient boosted regression trees (GBRT) and Xgboost algorithm. *Journal of Student Research Exploration*, 2(1), 1–10. [crossref]

Suotsalo, K., Xu, Y., Corander, J., & Pensar, J. (2021). High-dimensional structure learning of sparse vector autoregressive models using fractional marginal pseudo-likelihood. *Columbia University Review*. [crossref]

Tsay, L. S. (2005). *Analysis of financial time series*. John Wiley & Sons. [crossref]

Zhai, Y. (2021). Multivariate time series forecast in industrial process based on XGBoost and GRU. *2020 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, 9. [crossref]