Journal of Statistical Sciences and Computational Intelligence
e-ISSN: 3092–9458
2(3), 792–839
DOI: https://doi.org/10.64497/jssci.309
https://jssci.org/index.php/pub

Article

Forecasting household consumption in Nigeria: a comparison of machine learning and econometric models

Nura Aliyu Kabuga
Department of Economics, Faculty of Economics and Management Sciences, Bayero University Kano, Nigeria; nakabuga.eco@buk.edu.ng

Abstract: Accurate short-term forecasts of household consumption are important for macroeconomic surveillance and policy planning, particularly in data-constrained emerging economies. This paper compares machine-learning and econometric approaches for forecasting quarterly household final consumption expenditure (HFCE) growth in Nigeria under a strict one-quarter-ahead information structure. The comparison covers simple forecasting benchmarks, autoregressive and ARIMA-based econometric specifications, penalised regression, tree-based machine learning and extreme gradient boosting. External predictors are lagged by one quarter, while model specification, feature selection and hyperparameter choice are determined exclusively through expanding-window validation. A nine-quarter holdout sample is reserved for final out-of-sample evaluation. Forecast performance is assessed using RMSE, MAE, sMAPE, MASE and RMSSE, complemented by small-sample-adjusted Diebold–Mariano–Harvey–Leybourne–Newbold tests, empirical validation-residual prediction intervals, leave-one-quarter-out sensitivity analysis and residual diagnostics. Among the validation-selected model-family representatives, XGBoost records the lowest RMSE (9.646) and MAE (7.772), followed by Gradient Boosting (10.591; 9.235), the classical AR(4) model (11.179; 10.231) and LASSO (18.216; 14.407). XGBoost also records the lowest sMAPE, MASE and RMSSE and remains ranked first by RMSE and MAE in all nine leave-one-quarter-out exercises. Formal predictive-accuracy tests show that XGBoost and Gradient Boosting significantly outperform the naïve benchmark under squared and absolute loss at the 5 per cent level, although their advantages over the historical-mean and four-quarter rolling-mean benchmarks are not statistically significant. SHAP analysis indicates that the validation-selected XGBoost model draws predictive information from lagged Fishing growth, historical HFCE growth, gross fixed capital formation and other sectoral indicators, while validation-fold analysis shows that predictor stability varies across estimation windows. The findings show that machine-learning models, particularly XGBoost, provide competitive gains in short-term HFCE forecasting, while conventional econometric models and simple benchmarks remain important comparators within a diversified and uncertainty-aware forecasting framework.

Keywords: household consumption forecasting; HFCE; machine learning; econometric models; XGBoost; expanding-window validation; Nigeria

Received: xxxx; Revised: xxxx ; Accepted: xxxx ; Published Online: xxxx
Copyright: ©2026 the Author(s). Published by JSSCI. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).
Journal Abbreviation: J. Stat. Sci. Comput. Intell.

https://doi.org/10.64497/jssci.309

1. Introduction

Household consumption expenditure constitutes a major component of aggregate demand and plays an important role in macroeconomic activity, household welfare, and business-cycle dynamics. Changes in household expenditure can accompany movements in production, employment, investment, prices, and aggregate economic growth, making reliable consumption forecasts valuable for macroeconomic surveillance and policy planning [1–5]. The forecasting problem is particularly important in emerging economies, where macroeconomic volatility, inflationary pressures, exchange-rate movements, data limitations, and structural changes can complicate the extraction of stable predictive relationships from historical observations [6–11].

Nigeria provides a relevant setting for examining this forecasting challenge. Its macroeconomic environment has experienced substantial variation in inflation, exchange rates, sectoral activity, energy and transportation conditions, investment, and aggregate economic performance. Such variation can affect household purchasing power and the economic information available for predicting consumption. More generally, evidence from macroeconomic forecasting research indicates that predictive relationships can change across periods and may involve nonlinearities, interactions, persistence, and structural instability that are difficult to represent through a single fixed specification [6,7,11], [14–16]. These characteristics make model choice an empirical question and provide a case for comparing alternative forecasting approaches under a common information set and evaluation protocol.

Classical econometric models remain an important foundation for macroeconomic forecasting because they provide parsimonious representations of temporal dependence and are comparatively transparent in small samples. Autoregressive and related time-series specifications have consequently remained prominent in forecasting applications, including studies conducted with Nigerian macroeconomic data [8], [14], [17]. Penalised regression provides an intermediate approach when the available predictor set becomes larger by shrinking coefficient estimates and controlling model complexity. Evidence from high-dimensional macroeconomic forecasting shows that penalisation can facilitate variable selection and improve forecasting when conventional unrestricted regression becomes impractical [14], [27–29]. These approaches provide useful comparators for assessing whether more flexible machine-learning models deliver meaningful forecasting gains.

Machine-learning methods extend the forecasting environment by allowing nonlinear relationships and interactions to be learned with fewer restrictions on functional form. Tree-based methods and boosting algorithms have therefore received increasing attention in economic and financial forecasting [8–13], [18–23]. Nigerian applications have examined machine-learning methods in areas including inflation, exchange rates, crude-oil production, transportation, energy systems, and other macroeconomic or financial outcomes [8–13], [15–23]. The evidence indicates that machine learning can be useful in data-rich or nonlinear forecasting settings, but it does not establish that such models will necessarily outperform simpler econometric alternatives in every application. Their relative performance depends on the target variable, information set, sample size, validation procedure, forecast horizon, and economic environment.

The expansion of macroeconomic and sectoral datasets has also increased the range of information potentially available for forecasting household expenditure. Research outside Nigeria has shown that consumption and other macroeconomic variables can benefit from information contained in alternative, behavioural, sectoral, and timely economic indicators [1–7], [24–26]. At the same time, a larger predictor set creates additional risks of overfitting and model-selection bias, particularly when the available time-series sample is relatively short. A credible comparison of forecasting methods therefore requires more than fitting competing models to the same historical observations. Predictor screening, feature selection, model specification, and hyperparameter tuning must be separated from the final evaluation sample, while

information entering each forecast should correspond to what would actually have been available when that forecast was generated.

This requirement is particularly relevant to the distinction between nowcasting and genuine prospective forecasting. A model that uses information observed during the quarter being predicted can produce useful contemporaneous estimates, but its performance is not directly comparable with that of a model restricted to information available before the forecast quarter. The present study therefore formulates the primary empirical exercise as a strict one-quarter-ahead forecasting problem. External predictors are lagged by one quarter, while historical HFCE lags are included where required by the relevant model specification. This information structure is designed to reduce look-ahead bias and to ensure that the principal results represent forecasts that could, in principle, have been generated prospectively.

Despite the growing Nigerian forecasting literature, comparatively limited evidence exists on the out-of-sample forecasting of quarterly household final consumption expenditure (HFCE) using a systematic comparison of econometric, penalised-regression, and machine-learning models. Existing Nigerian studies have concentrated more extensively on inflation [8,9], exchange rates [10,11,13], crude-oil production [12], transportation [18,19], energy-related outcomes [23], and broader macroeconomic or financial variables [15–22]. International research has examined consumption and related macroeconomic forecasting using Google Trends and other alternative information sources [1–7], [24–26], but the corresponding evidence for Nigerian household consumption remains comparatively sparse. A further issue is that forecast comparisons can become overly dependent on point accuracy measures without establishing whether observed differences are statistically distinguishable, robust to individual evaluation periods, associated with well-behaved residuals, or accompanied by credible measures of forecast uncertainty.

This study addresses these issues by comparing benchmark forecasts, classical econometric models, penalised regression, tree-based machine-learning methods, and Extreme Gradient Boosting for forecasting quarterly HFCE growth in Nigeria. The completed dataset comprises 58 quarterly observations from 2010Q1 to 2024Q2. The first 49 observations, covering 2010Q1–2022Q1, form the development sample used for estimation, feature selection, model specification, hyperparameter tuning, and expanding-window validation. The remaining nine quarters, from 2022Q2 to 2024Q2, constitute an untouched final holdout sample used exclusively for out-of-sample evaluation. Representative models within the competing families are selected using development-sample validation rather than final-test performance.

Forecast performance is evaluated using root mean squared error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (sMAPE), mean absolute scaled error (MASE), and root mean squared scaled error (RMSSE). The analysis goes beyond metric ranking by applying small-sample-adjusted Diebold–Mariano–Harvey–Leybourne–Newbold (DM-HLN) tests to selected forecast comparisons. Empirical prediction intervals constructed from validation residuals provide evidence on forecast uncertainty, while Ljung–Box diagnostics examine residual serial dependence. Leave-one-quarter-out analysis assesses the stability of model-family rankings to individual holdout observations. For the validation-selected XGBoost specification, SHAP analysis is used to examine predictive attribution, supplemented by validation-fold analysis of feature stability. SHAP values are interpreted as measures of contribution to model predictions and not as estimates of causal effects.

The study makes four principal contributions. First, it provides a comparative evaluation of household-consumption forecasting in Nigeria across benchmark, classical econometric, penalised-regression, tree-based machine-learning, and extreme-gradient-boosting approaches under a common out-of-sample framework. This design permits the forecasting value of additional model flexibility to be assessed against both simple benchmarks and conventional econometric alternatives. Second, the study imposes a strict one-quarter-ahead

information structure and confines predictor screening, model specification, feature selection, and hyperparameter choice to the development sample. The resulting separation between model development and final evaluation strengthens the credibility of the reported out-of-sample evidence.

Third, the study evaluates forecasting performance beyond a single error measure. Multiple accuracy metrics are combined with statistical forecast-comparison tests, empirical prediction intervals, residual diagnostics, and leave-one-quarter-out sensitivity analysis. This provides evidence on comparative accuracy, statistical distinguishability, forecast uncertainty, residual behaviour, and ranking stability within the same empirical framework. Fourth, the study incorporates explainable machine learning through SHAP-based predictive attribution and feature-stability analysis. This permits the information used by the strongest validation-selected machine-learning model to be examined while maintaining a clear distinction between predictive importance and causal economic interpretation.

These contributions have implications for the design of short-term macroeconomic forecasting systems in Nigeria. The central empirical question is not whether machine learning should replace econometric forecasting, but whether additional model flexibility produces sufficiently reliable out-of-sample improvements when all competing approaches are subjected to the same information restrictions, validation procedures, and untouched test sample. This comparative perspective is relevant for institutions that require timely assessments of household-demand conditions while operating with relatively short macroeconomic time series and substantial economic uncertainty.

The remainder of the paper is organised as follows. Section 2 reviews the relevant literature and develops the theoretical and conceptual forecasting frameworks. Section 3 presents the data, research design, forecasting protocol, model specifications, validation procedure, evaluation measures, statistical tests, uncertainty assessment, robustness analysis, and explainability methods. Section 4 reports and discusses the empirical findings and their policy implications. Section 5 concludes the paper, presents recommendations, and discusses the study's limitations and directions for further research.

2. Literature Review

This section reviews empirical and theoretical literature relevant to short-term household consumption forecasting. The empirical review examines the progression from conventional time-series and econometric forecasting methods to penalised regression, machine learning, and information-rich forecasting approaches. Particular attention is given to evidence from Nigeria and to studies that compare alternative forecasting methods or incorporate additional predictive information. The theoretical review subsequently develops the economic foundations of household consumption behaviour before linking those foundations to the forecasting framework adopted in this study.

2.1. Empirical Review

Macroeconomic forecasting methods have expanded considerably as the volume, frequency and dimensionality of available economic information have increased. Conventional time-series models remain important for capturing persistence and temporal dependence, while penalised regression methods provide tools for handling larger predictor sets. Machine-learning methods extend this modelling environment by allowing more flexible functional relationships, and recent forecasting research increasingly incorporates digital, behavioural and other high-frequency information sources [1–7], [14,20,21].

The empirical literature does not imply that one modelling family is uniformly superior across forecasting problems. Relative performance depends on the target variable, forecast horizon, information set, sample size, evaluation design and economic environment. This consideration is particularly important for macroeconomic

applications in which relatively short time series coexist with potentially large predictor sets. The review therefore focuses on the empirical circumstances under which different forecasting approaches have proved useful, rather than assuming in advance that greater model complexity necessarily produces greater forecast accuracy.

2.1.1. Classical Macroeconomic Forecasting Approaches

Classical time-series and econometric models remain central to macroeconomic forecasting because they provide structured methods for exploiting persistence, lagged adjustment and stochastic dependence in economic data. Autoregressive and ARIMA-type specifications are particularly useful when past values of the target variable contain information about its future path. Their parsimonious structure can also be advantageous when the available time series is relatively short and the number of observations constrains the reliable estimation of more highly parameterised models.

Evidence from Nigerian forecasting applications illustrates the continued relevance of these methods. Evans et al. [17], for example, apply an ARIMA-based time-series framework to the modelling and forecasting of petroleum motor spirit prices in Nigeria. Comparative Nigerian evidence also shows that econometric specifications continue to provide meaningful benchmarks when evaluated alongside machine-learning alternatives. Akinjobi and Omooseti [8] compare econometric and machine-learning approaches in forecasting components of Nigerian inflation, reinforcing the value of evaluating flexible models against established econometric alternatives rather than assessing them in isolation.

The wider forecasting literature similarly demonstrates that conventional statistical models remain relevant even as richer information sets become available. Marcellino and Sivec [5] examine GDP growth nowcasting in a small open economy, while [4] consider real-time GDP measurement during the COVID-19 pandemic. These studies are primarily nowcasting applications and should not be interpreted as evidence about the superiority of machine learning or conventional forecasting methods. Their relevance to the present study lies in demonstrating the importance of information timing, data availability and real-time evaluation when macroeconomic forecasts are constructed.

An important limitation of conventional econometric specifications emerges when the predictor environment becomes large relative to the available sample. Estimating numerous correlated predictors within an unrestricted linear model can increase parameter uncertainty and reduce out-of-sample stability. Smeekes and Wijler [14] show why high-dimensional macroeconomic forecasting has encouraged the use of penalised regression procedures that control model complexity while retaining information from broad predictor sets. This development does not make classical models obsolete; it establishes a motivation for comparing parsimonious time-series specifications with forecasting methods designed for larger information environments.

Classical models consequently provide two important reference points for the present study. First, autoregressive specifications assess how much one-quarter-ahead information can be extracted from the historical dynamics of household consumption itself. Second, ARIMA and exogenous-variable extensions provide conventional econometric comparators against which the incremental predictive value of penalised regression and machine-learning models can be evaluated. Their inclusion is therefore based on comparative forecasting discipline rather than an assumption that conventional econometric models must either dominate or underperform more flexible alternatives.

2.1.2. Penalised Regression and High-Dimensional Forecasting

Macroeconomic forecasting increasingly involves datasets containing numerous candidate predictors

whose information may overlap substantially. Conventional regression becomes difficult in this setting because a large predictor set relative to the available observations can produce unstable coefficient estimates, while strong correlation among macroeconomic variables can further weaken out-of-sample performance. Penalised regression addresses these problems by imposing restrictions on coefficient magnitude during estimation. The resulting shrinkage introduces some bias but can reduce estimation variance and improve forecast performance, particularly where the number of potential predictors is large or the predictors exhibit substantial collinearity.

Several forms of regularisation provide different solutions to this forecasting problem. Ridge regression applies an L_2 penalty that shrinks coefficients towards zero without normally eliminating predictors entirely, making it particularly useful where correlated variables contain overlapping predictive information. LASSO uses an L_1 penalty and can shrink some coefficients exactly to zero, thereby combining estimation with variable selection. Elastic Net incorporates both L_1 and L_2 penalties and is especially attractive where predictors occur in correlated groups because it combines the variable-selection property of LASSO with the stabilising effect of Ridge regression. These characteristics make penalised methods relevant to macroeconomic applications in which a relatively limited time-series sample must be evaluated against a broad set of potentially informative indicators.

The forecasting literature provides empirical support for this role of regularisation. Smeekes and Wijler [14] showed that penalised regression methods can be effective for macroeconomic forecasting when the predictor space is large, although forecast performance depends on the characteristics of the series being predicted and the specification of the penalty. Their evidence demonstrates that regularisation provides a practical means of exploiting information from many macroeconomic variables without estimating an unrestricted high-dimensional model. The implication is important for forecasting exercises in which dimensionality reduction must occur without discarding potentially useful information before its predictive contribution has been evaluated.

The relevance of penalisation becomes even clearer in explicitly high-dimensional settings. Uematsu and Tanaka [27] examined macroeconomic forecasting and variable selection through folded-concave penalised regression, allowing the number of potential regressors to be substantially larger than the sample size. Their empirical application used a high-dimensional monthly macroeconomic dataset to forecast US gross domestic product within a mixed-data-sampling framework. The study found that penalised regression could deliver strong forecasting and variable-selection performance in such environments. Their results extend the case for regularisation beyond conventional shrinkage by showing how penalisation can extract forecasting information from very large predictor sets while simultaneously controlling model complexity.

Time-series dependence introduces an additional consideration because regularisation procedures cannot automatically be treated as if observations were independent. Tonini et al. [28] investigated the effect of serial dependence on penalised regression methods and showed that dependence in the data can influence estimation and model-selection behaviour. This evidence is relevant to macroeconomic forecasting because economic indicators are typically serially correlated and their predictive relationships evolve through time. Consequently, the usefulness of regularisation depends not only on the choice of penalty but also on an evaluation framework that respects temporal ordering. Time-aware validation is therefore preferable to random cross-validation when penalised models are used for genuine forecasting.

Evidence from applied inflation forecasting also illustrates the potential predictive gains from regularisation. Ospina-Holguín and Padilla-Ospina [29] compared penalised linear regression models with seasonal autoregressive moving-average specifications for inflation forecasting. Their out-of-sample results favoured the penalised models across the inflation series considered, indicating that shrinkage-based specifications can outperform established autoregressive benchmarks when additional predictive information

is available. This finding is particularly relevant to macroeconomic forecasting because it demonstrates that the value of regularisation is not confined to theoretical high-dimensional settings but can translate into improvements in practical forecasting applications.

The literature nevertheless does not imply that penalised regression will invariably dominate either conventional econometric or nonlinear machine-learning methods. Regularisation controls dimensionality and estimation variance within a predominantly linear modelling structure, but its performance may deteriorate when the underlying predictive relationship contains strong nonlinearities, threshold effects or complex interactions. Conversely, the relatively constrained functional form of penalised regression can be advantageous in small macroeconomic samples because it limits model complexity. Penalised models therefore occupy an important intermediate position between conventional low-dimensional econometric specifications and more flexible nonlinear machine-learning algorithms.

This distinction is central to the present study. Nigerian household consumption is potentially associated with a broad set of macroeconomic and sectoral indicators, while the available quarterly time series remains comparatively small. Penalised regression consequently provides a useful test of whether systematic coefficient shrinkage and regularisation can extract predictive information from this expanded predictor set without requiring the nonlinear flexibility of tree-based machine-learning methods. Ridge, LASSO and Elastic Net are therefore evaluated as a distinct model family under the same strict one-quarter-ahead information structure and expanding-window validation framework used for the competing forecasting approaches. Their inclusion allows the empirical analysis to determine whether the forecasting gains arise primarily from regularisation and controlled dimensionality or whether more flexible nonlinear learners provide additional out-of-sample predictive value.

2.1.3. Machine Learning and Nonlinear Forecasting

Macroeconomic relationships can exhibit nonlinearities, interactions and changes in predictive structure that are difficult to represent adequately through a fixed linear specification. Household consumption is particularly susceptible to such complexity because responses to changes in prices, income conditions, investment and sectoral activity may differ across economic environments. Machine-learning methods offer an alternative forecasting framework in which these relationships can be learned from the data with fewer restrictions on functional form. Their usefulness in macroeconomic forecasting therefore rests on the possibility that additional flexibility can recover predictive patterns that conventional linear models do not fully capture.

The empirical appeal of machine learning has increased as forecasting applications have incorporated larger and more diverse information sets. Machine-learning models can process numerous candidate predictors while accommodating nonlinear responses and interactions among them. These characteristics are particularly relevant when economic activity is represented by information drawn from several sectors or indicators. The potential benefit, however, does not imply that greater model complexity necessarily produces better forecasts. Forecast performance remains an empirical question that should be established through genuine out-of-sample evaluation rather than goodness of fit within the estimation sample.

Recent forecasting evidence illustrates this point across several economic applications. Pfahler [10] evaluates advanced machine-learning methods for exchange-rate forecasting, while Qureshi et al. [11] compare artificial-intelligence and time-series approaches to real exchange-rate prediction. Nigerian applications have similarly extended machine learning to macroeconomic and sectoral forecasting. Obite et al. [12] compare classical and machine-learning approaches for modelling Nigerian crude-oil production, and Gbadebo [13] applies machine learning to exchange-rate prediction in Nigeria. These studies demonstrate the growing use of data-driven forecasting methods in economic applications, although the variables, frequencies and forecasting

https://doi.org/10.64497/jssci.309

environments differ from household consumption.

Nonlinear forecasting has also been examined through hybrid and regime-dependent models. Muhammad et al. [15] combine exponential autoregressive, FIGARCH and artificial neural-network components in a hybrid forecasting framework, illustrating how nonlinear statistical structures can be combined with machine-learning techniques. Abubakar and Akeyede [16] adopt a Bayesian regime-switching threshold autoregressive specification to model nonlinear dynamics and risk in the Nigerian stock market. These studies are not direct applications to household consumption, but they provide evidence that Nigerian economic and financial series can motivate forecasting specifications capable of accommodating nonlinear or regime-dependent behaviour. Conventional time-series approaches nevertheless remain relevant. Evans et al. [17], for example, employ an ARIMA framework to model and forecast petroleum motor spirit prices in Nigeria. The coexistence of these approaches reinforces the case for evaluating flexible forecasting methods against simpler statistical alternatives rather than presuming their superiority.

Machine learning has also been applied to problems involving economic activity outside conventional aggregate macroeconomic series. Maaouane et al. [18] employ neural-network modelling for the estimation and forecasting of transport-sector energy demand in developing countries. Imoh and Rad [19] use machine-learning models to analyse and predict traffic conditions on Ikorodu Road in Lagos State. These applications indicate that nonlinear predictive methods can exploit information embedded in sector-specific activity. This is relevant to household-consumption forecasting because sectoral indicators may contain information about changes in production, transportation, services and other components of economic activity before their effects are fully reflected in aggregate consumption measures.

More recent research has broadened the application of machine learning to aggregate economic forecasting. Ghosh [20] considers machine-learning approaches to GDP forecasting across global, regional and country-specific settings, while [22] compare time-series, machine-learning and probability models for GDP growth. Tang and Xie [21] develop a deep-learning framework for exchange-rate forecasting that combines signal decomposition with dynamic weight optimisation. Eshiemogie et al. [23] also demonstrate the application of machine-learning methods to Nigerian energy-related prediction. Although these studies address different dependent variables, they illustrate a broader movement towards comparative forecasting designs in which flexible learning algorithms are assessed alongside established statistical methods.

Evidence directly related to Nigerian macroeconomic forecasting is especially pertinent. Akinjobi and Omooseti [8] compare machine-learning and econometric models for forecasting components of Nigerian inflation. Their comparative design is relevant because it treats the relative performance of alternative model classes as an empirical issue. Bartholomew et al. [9] apply a deep recurrent neural network to Nigerian inflation forecasting and explicitly consider the role of feature-selection methods. Together, these studies indicate that machine learning has begun to occupy a more prominent position in Nigerian macroeconomic forecasting. The existing Nigerian evidence, however, is concentrated largely on inflation, exchange rates, energy, financial markets and other sector-specific outcomes. Much less attention has been directed towards quarterly household final consumption expenditure.

This distinction matters because household consumption presents a different forecasting problem. Its dynamics may reflect persistence in previous consumption growth alongside information contained in investment and sectoral activity. Tree-based ensemble methods are well suited to investigating such a setting because they can represent nonlinearities and interactions without requiring these relationships to be specified individually beforehand. Random Forest reduces the instability associated with individual decision trees by aggregating predictions across multiple trees, while boosting methods construct successive learners that attempt to improve prediction errors generated at earlier stages. Extreme Gradient Boosting extends this

principle through regularisation and additional controls over model complexity. The forecasting advantage of these algorithms must nevertheless be demonstrated on observations that played no role in estimation, feature selection or hyperparameter tuning.

The relatively small number of observations available in quarterly macroeconomic applications makes this requirement particularly important. Flexible algorithms can adapt closely to historical observations and consequently produce apparently strong in-sample performance without maintaining that advantage when forecasting unseen periods. Chronological validation is therefore more appropriate than random data splitting for time-series applications. Model selection should preserve temporal ordering and separate the observations used for tuning from those reserved for final evaluation. This distinction also provides an important connection with the penalised-regression literature discussed in Section 2.1.2, where high dimensionality, serial dependence and model selection create similar concerns [14], [27–29].

Machine-learning forecasts must also be distinguished from causal economic analysis. A variable that improves prediction does not necessarily have a causal effect on household consumption. The distinction is especially important when flexible algorithms identify complex interactions among lagged macroeconomic and sectoral indicators. Predictor importance should therefore be interpreted as evidence about predictive information, not as an estimate of a structural behavioural parameter. Explainable machine-learning methods can nevertheless make complex forecasting models more transparent by showing which variables contribute most strongly to their predictions and whether those contributions remain stable across estimation windows.

The literature consequently establishes a strong case for comparing conventional and machine-learning forecasting methods, but it also reveals an important gap in the Nigerian evidence. Existing applications have concentrated on inflation [8,9], exchange rates [13], crude-oil production [12], financial-market dynamics [15,16], energy-related outcomes [18], [23] and other sectoral variables [17,19]. Evidence on whether machine-learning algorithms improve short-term forecasts of Nigerian household consumption remains comparatively limited. The present study addresses this gap by comparing classical econometric, penalised-regression and machine-learning model families within the same forecasting experiment. Each model family is subjected to a common information structure, validation procedure and final out-of-sample test period, allowing differences in forecast performance to be attributed more credibly to predictive performance rather than unequal access to information or different evaluation samples.

2.1.4. Alternative Indicators and Information Sets for Macroeconomic Forecasting

Forecast accuracy depends not only on the choice of forecasting model but also on the information available when a prediction is generated. Traditional macroeconomic forecasting relies heavily on official economic statistics, which provide theoretically meaningful measures of economic activity but may become available only after publication delays. This creates a particular challenge for short term forecasting because changes in household behaviour and economic conditions can occur before they are fully reflected in conventional aggregate indicators. The forecasting literature has consequently examined whether broader and more timely information sets can supplement historical macroeconomic data and improve predictions of current or near term economic activity.

Internet search behaviour provides one prominent example. Carrière-Swallow and Labbé [1] demonstrate the potential usefulness of Google Trends for nowcasting in an emerging market, while [2] find that Google search information can improve forecasts of US unemployment. Niesert et al. [6] examine this question across a broader range of macroeconomic series and show that the predictive value of search information is not uniform across variables. This qualification is important. The availability of additional information does not itself guarantee an improvement in forecast accuracy because the usefulness of an indicator depends on

whether it contains predictive information beyond that already captured by conventional variables.

The relevance of alternative indicators extends directly to household consumption. Fasulo et al. [3] investigate Google Trends for nowcasting quarterly household consumption expenditure, providing evidence that search behaviour can contain information relevant to consumption dynamics. Schmidt and Vosen [24] similarly compare Google Trends with survey based indicators for forecasting private consumption. These studies are particularly pertinent because consumption decisions may respond to changes in expectations, employment prospects and economic sentiment before those responses become fully observable in subsequently released aggregate expenditure statistics. Alternative indicators may therefore complement conventional macroeconomic variables by capturing dimensions of household behaviour that are otherwise observed only with a delay.

The forecasting value of timely information became especially apparent during periods of rapid economic disruption. Koop et al. [4] examine the nowcasting of US GDP during the COVID 19 pandemic, when conventional relationships and publication schedules presented substantial difficulties for real time assessment. Marcellino and Sivec [5] study GDP growth nowcasting in a small open economy and similarly demonstrate the importance of constructing forecasting systems capable of exploiting information arriving at different times. These contributions indicate that information timing is a substantive component of forecast design rather than a secondary data issue. A model evaluated with variables that would not actually have been available at the forecast origin can appear more accurate without representing a genuinely feasible forecasting exercise.

This distinction is particularly relevant when alternative indicators are incorporated into formal time series models. Kohns and Bhattacharjee [7] combine Google Trends information with a Bayesian structural time series framework for growth nowcasting. Their approach illustrates that alternative information and statistical modelling need not be treated as competing forecasting strategies. Additional indicators can form part of the predictor set, while the forecasting model determines how their information is translated into predictions. The resulting performance therefore depends jointly on the quality of the information set and the model's capacity to extract useful predictive signals from it.

Evidence from labour market applications provides a further illustration of the policy relevance of timely indicators. Simionescu and Cifuentes-Faura [25] examine whether unemployment forecasts based on Google Trends can assist policy design in Spain and Portugal. Simionescu and Raišienė [26] investigate the relationship between Google Trends, pandemic conditions and employment expectations in European Union member states. These studies broaden the significance of alternative indicators beyond statistical forecast accuracy. Timelier estimates of changes in employment, sentiment and economic behaviour can provide information for policy assessment when official statistics are released with delays.

The implications for emerging and data constrained economies are particularly important. Forecasting systems in such settings may face shorter historical samples, data revisions, publication lags and gaps in the frequency or coverage of conventional macroeconomic indicators. Alternative information sources can potentially mitigate some of these constraints, but they introduce additional concerns. Search intensity, mobility measures and other high frequency indicators may be affected by changes in platform usage, measurement definitions and population coverage. Their predictive relationships can also vary over time. Such variables therefore require the same out of sample validation discipline as conventional macroeconomic predictors.

A related distinction concerns nowcasting and genuine one quarter ahead forecasting. Nowcasting seeks to estimate economic activity for a period that is already underway or has recently ended by exploiting contemporaneously available information. One quarter ahead forecasting imposes a stricter requirement because predictors must be restricted to information available before the forecasted quarter. Studies using Google Trends and other timely indicators demonstrate the potential value of contemporaneous information

[1], [3], [7], [24], but forecast performance obtained under that information structure should not be interpreted as equivalent to performance under a strictly prospective forecasting design.

This distinction has direct implications for the present study. Although the literature demonstrates the potential predictive value of alternative and high frequency indicators, the principal forecasting experiment in this paper deliberately adopts a strict one quarter ahead information structure. External macroeconomic and sectoral predictors are lagged by one quarter so that forecasts of HFCE growth in quarter t rely on information available through quarter t−1. This design sacrifices access to some contemporaneous information in exchange for a cleaner assessment of genuine prospective forecast performance. It also reduces the possibility that apparent improvements arise from information that would not have been observable at the actual forecast origin.

The present study consequently differs from much of the alternative indicator literature in two related respects. First, it investigates quarterly household consumption in Nigeria, an outcome for which the forecasting evidence remains limited compared with GDP, inflation, unemployment and exchange rates. Second, the emphasis is not on demonstrating that one particular alternative indicator improves prediction. The study examines whether a diversified set of lagged macroeconomic and sectoral variables contains useful predictive information when processed through competing econometric, penalised regression and machine learning models. This approach permits the contribution of the information set to be assessed within a common and operationally realistic forecasting framework.

2.1.5. Nigerian Forecasting Evidence and Research Gap

Empirical forecasting research in Nigeria has expanded across inflation, exchange rates, energy, commodity prices, financial markets and sectoral activity, yet household consumption has received comparatively limited attention. This imbalance is important because household final consumption expenditure constitutes a major component of aggregate demand and may exhibit forecasting dynamics that differ from those of prices, exchange rates or financial variables. Evidence obtained for other Nigerian macroeconomic outcomes cannot therefore be assumed to transfer directly to quarterly HFCE growth.

Inflation forecasting provides some of the clearest Nigerian evidence on comparative modelling. Akinjobi and Omooseti [8] evaluate machine-learning and econometric approaches for forecasting components of Nigerian inflation, while Bartholomew et al. [9] examine deep recurrent neural networks and the contribution of alternative feature-selection procedures to inflation prediction. These studies demonstrate that model choice and predictor selection can influence forecast performance in the Nigerian macroeconomic environment. Their focus on inflation, however, leaves unresolved whether similar gains arise when the forecasting target is household consumption, whose temporal behaviour and predictor relationships may differ substantially from those of consumer prices.

Exchange-rate applications provide another strand of evidence. Gbadebo [13] investigates machine learning for exchange-rate prediction in Nigeria, complementing international applications that compare machine-learning, artificial-intelligence and conventional time-series methods for exchange-rate forecasting [10,11,21]. Exchange rates can themselves contain useful information about macroeconomic conditions, but forecasting the exchange rate is analytically distinct from forecasting household expenditure. The evidence from this literature consequently supports the feasibility of flexible predictive methods without establishing which model family is most appropriate for Nigerian consumption growth.

Nigerian studies have also examined production, commodity and sector-specific outcomes. Obite et al. [12] compare classical and machine-learning methods for crude-oil production, whereas [17] employ an ARIMA specification for petroleum motor spirit prices. Imoh and Rad [19] apply machine learning to traffic conditions

https://doi.org/10.64497/jssci.309

in Lagos, and [23] use a machine-learning-supported framework for energy-related prediction in Nigeria. The breadth of these applications shows increasing adoption of predictive modelling across Nigerian economic and infrastructural settings. It also reveals considerable heterogeneity in target variables, data frequencies, predictor structures and evaluation procedures, which limits direct comparison of reported forecast performance across studies.

Recent Nigerian research further demonstrates interest in nonlinear and hybrid forecasting structures. Muhammad et al. [15] develop a hybrid ExpAR-FIGARCH-ANN model, while [16] investigate Bayesian regime-switching threshold autoregression for nonlinear dynamics and risk forecasting in the Nigerian stock market. These contributions provide evidence that nonlinear and regime-dependent structures can be relevant to Nigerian time series. Neither study, however, addresses household consumption or establishes that a particular nonlinear specification should dominate alternative forecasting families when all models are evaluated under the same prospective information set.

It is evident from these studies that three features of the Nigerian literature are apparent. First, forecasting research has concentrated principally on variables other than household consumption. Second, individual studies frequently emphasise a particular modelling approach or a restricted comparison of alternatives. Third, differences in data construction, forecast horizons and evaluation procedures make it difficult to determine whether reported improvements reflect genuinely stronger predictive models or differences in the information and testing environments. These issues create a need for comparative forecasting evidence in which alternative model families face the same target, forecast horizon, information restrictions and final evaluation observations.

A further gap concerns the treatment of model selection. Flexible forecasting methods introduce choices concerning predictor subsets, regularisation parameters, tree complexity and other hyperparameters. If these decisions are influenced by the same observations subsequently used to report final forecast accuracy, the resulting comparison may overstate out-of-sample performance. This concern is particularly consequential for quarterly Nigerian macroeconomic data because the available samples are relatively small. A credible comparison therefore requires model specification and tuning to be completed within the development period before forecasts are evaluated on an untouched final holdout sample.

Information timing presents a related problem. The alternative-indicator literature reviewed in Section 2.1.4 shows that contemporaneous information can improve assessment of current economic conditions [1], [3], [7], [24]. Such evidence is valuable for nowcasting but does not automatically establish performance in a genuinely prospective forecast. For one-quarter-ahead prediction, variables used to forecast quarter t should correspond to information available before that quarter. The distinction between contemporaneous and lagged information is therefore central to evaluating whether a forecasting system could have generated its predictions in real time.

Forecast evaluation itself represents another area requiring greater attention. Numerical rankings based on RMSE or MAE indicate which model records the lowest realised error in a particular test sample, but they do not by themselves establish that differences in predictive accuracy are statistically meaningful. Forecast uncertainty and ranking sensitivity are also important when the evaluation period is short. A more comprehensive comparison should consequently combine multiple scale-dependent and scale-free error measures with formal predictive-accuracy testing, forecast-uncertainty assessment, residual diagnostics and sensitivity analysis.

Interpretability creates a final gap when complex machine-learning models are used. High predictive performance alone provides limited information about the indicators contributing to a forecast. Explainability methods can improve transparency, but their outputs require careful interpretation. Feature importance and SHAP attribution describe the information used by a fitted predictive model; they do not identify causal effects

https://doi.org/10.64497/jssci.309

of economic variables on household expenditure. In addition, a variable that appears important in the final fitted model may not be selected consistently as the estimation window changes. Evaluating both predictive attribution and feature stability can therefore provide a more informative assessment of machine-learning forecasts.

The present study addresses these gaps through a unified comparison of simple benchmarks, classical econometric models, penalised regression and tree-based machine-learning methods for forecasting quarterly Nigerian HFCE growth. All competing models are evaluated at a common one-quarter-ahead horizon. External predictors are restricted to lagged information in the primary forecasting design, while model specification, feature selection and hyperparameter choice are determined through expanding-window validation within the development sample. The final nine-quarter holdout period is excluded from these decisions and is used only for final out-of-sample evaluation.

The contribution extends beyond identifying the model with the lowest realised forecast error. Predictive performance is assessed using RMSE, MAE, sMAPE, MASE and RMSSE, supplemented by small-sample-adjusted predictive-accuracy tests, empirical prediction intervals, residual diagnostics and leave-one-quarter-out sensitivity analysis. Explainable machine learning is then used to examine the information contributing to the validation-selected nonlinear forecast, while validation-fold feature frequencies provide evidence on predictor stability. The resulting framework permits a more disciplined assessment of whether additional model flexibility improves short-term Nigerian household-consumption forecasting and how confidently any observed advantage can be interpreted.

2.2. Theoretical Framework

Household consumption theory provides the economic foundation for understanding why current expenditure may contain information from past consumption, income conditions, prices, wealth, investment and broader macroeconomic activity. The relevant theoretical question for the present study is not whether every predictor included in a forecasting model has a structural causal effect on consumption. It is whether established theories of consumption provide economic reasons for expecting household expenditure to respond to information arriving through different channels and over different horizons. This distinction is important because the empirical objective of the study is prediction rather than causal identification.

The theoretical framework combines intertemporal consumption smoothing with liquidity constraints and precautionary behaviour under uncertainty. These perspectives are particularly relevant to an emerging economy such as Nigeria, where household income can be volatile, formal credit access is uneven and macroeconomic shocks can alter real purchasing power. They also provide an economic basis for allowing consumption forecasts to draw information from historical HFCE dynamics and lagged macroeconomic and sectoral indicators without presuming that the underlying predictive relationships are constant or linear.

2.2.1. Intertemporal Consumption, Permanent Income and Consumption Smoothing

The starting point is the intertemporal theory of consumption. The Life-Cycle Hypothesis associated with Modigliani and Brumberg and Friedman's Permanent Income Hypothesis distinguish current income from the longer-term resources households expect to command over time [30,31]. Under these approaches, households do not determine current consumption solely from contemporaneous income. They allocate expenditure across periods according to expected lifetime or permanent resources, using saving and borrowing to reduce the effect of temporary income fluctuations on consumption.

A general intertemporal consumption problem can be represented as

$$\max_{\{C_t\}_{t=0}^{T}} E_0[\sum_{t=0}^{T} \beta^t \, u(C_t)], \tag{1}$$

https://doi.org/10.64497/jssci.309

where C_t denotes household consumption in period t, β is the subjective discount factor, u(·) is the utility function and E_0 denotes expectations conditional on information available at the initial decision point. The optimisation in Equation (1) is subject to an intertemporal resource constraint:

$$A_{t+1} = (1 + r_t)(A_t + Y_t - C_t), \quad (2)$$

where A_t represents assets or accumulated wealth, Y_t is household income and r_t is the return on assets. Equation (2) shows that current expenditure competes with saving for available resources. A temporary change in current income therefore need not produce an equivalent change in consumption when households can save during favourable periods or borrow against expected future resources during adverse periods. Consumption should respond more strongly when households perceive changes in resources as persistent than when fluctuations are regarded as temporary [30,31].

The consumption-smoothing implication is important for forecasting HFCE. If households adjust expenditure gradually as information about persistent economic conditions accumulates, historical consumption contains information about subsequent consumption growth. This provides an economic rationale for including lagged HFCE terms in forecasting models. Indicators associated with income generation, investment and aggregate economic activity may also contain predictive information when they alter expectations about future household resources.

The assumptions underlying unrestricted consumption smoothing can be demanding in developing economies because they require sufficient access to saving and borrowing instruments. When credit access is restricted or income becomes uncertain, consumption may become more sensitive to currently available resources. This limitation motivates the extension of the framework to liquidity constraints and precautionary saving.

2.2.2. Liquidity Constraints, Income Uncertainty and Precautionary Saving

Deaton's analysis of saving under liquidity constraints provides a particularly relevant extension of intertemporal consumption theory [32]. When households cannot borrow freely against expected future income, assets can serve as a buffer against adverse income realisations. Borrowing constraints and precautionary saving then influence how much current income is consumed and how much is retained to protect future consumption.

A borrowing restriction can be represented as:

$$A_{t+1} \geq \underline{A}, \quad (3)$$

where $\underline{A}$ represents the lower bound on household asset holdings or borrowing. Equation (3) limits the household's ability to finance current expenditure from future income. When the constraint binds, current consumption can become more closely related to currently available resources than predicted by an unconstrained permanent-income model [32].

Income uncertainty strengthens the relevance of this mechanism. Carroll [33] shows how uninsurable income risk, impatience and liquidity constraints can generate buffer-stock saving behaviour, with households holding assets partly to protect consumption against future adverse shocks. A conceptual representation of this relationship is:

$$C_t = g\left(Y_t^P, A_t, \sigma_{Y,t}^2, BC_t\right), \quad (4)$$

where Y_t^P denotes expected permanent income, A_t denotes available assets, $\sigma_{Y,t}^2$ represents perceived income uncertainty and BC_t captures the severity of borrowing or liquidity constraints. Equation (4) is not estimated as a structural equation in this study. It summarises the theoretical proposition that consumption can depend on expected resources, accumulated buffers, uncertainty and households' capacity to transfer resources across time.

These mechanisms are pertinent to Nigeria. Ajefu [34] reports evidence that Nigerian households experience income shocks and use informal insurance arrangements to smooth consumption, although such mechanisms do not provide complete protection. The implication for forecasting is that the same macroeconomic signal may have different predictive associations with consumption under different financial and income conditions. Such state dependence weakens the assumption that one constant linear relationship must describe consumption dynamics throughout the sample.

2.2.3. Macroeconomic Shocks, Prices and Household Consumption in Nigeria

Household consumption also operates within a macroeconomic environment shaped by movements in prices, output, investment and sectoral activity. These conditions can alter both nominal household resources and the quantity of goods and services that those resources can purchase. Inflation is especially relevant because increases in consumer prices reduce real purchasing power when nominal household resources do not adjust proportionately.

Recent Nigerian evidence supports the sensitivity of household consumption to aggregate shocks. Animashaun and Wossink [35] examine household responses to the 2009–2015 oil-price crisis and document heterogeneous consumption responses and coping adjustments. Such heterogeneity is theoretically important because aggregate shocks need not transmit uniformly across households or expenditure categories. It also supports a forecasting framework that does not impose identical responses across all economic conditions.

The purchasing-power channel can be expressed in simplified form as:

$$Y_t^R = \frac{Y_t^N}{P_t}, \tag{5}$$

where Y_t^R denotes real household resources, Y_t^N represents nominal resources and P_t is the relevant price level. Equation (5) shows why nominal income growth need not translate into equivalent growth in real consumption opportunities when prices increase. Inflation can therefore contain predictive information about household expenditure even when nominal economic activity remains strong.

Investment and sectoral activity provide additional information about the environment in which household resources are generated. Gross fixed capital formation can reflect changes in productive activity and expected economic conditions, while developments in agriculture, transport, manufacturing and services may contain information about employment, earnings, supply conditions and the availability or cost of goods and services. These indicators should not be interpreted automatically as causal determinants of HFCE merely because they enter a forecasting model. Their theoretical relevance is that they may contain information about evolving economic conditions to which household expenditure subsequently responds.

The combination of aggregate shocks, incomplete consumption smoothing, heterogeneous household exposure and changing real purchasing power also allows for asymmetric or state-dependent consumption responses. A deterioration in economic conditions may generate a stronger expenditure response among financially constrained households than an equivalent improvement generates during more favourable periods. These features provide an economic rationale for testing forecasting methods capable of representing nonlinearities and interactions.

2.2.4. From Consumption Theory to a Nonlinear Forecasting Framework

The theoretical perspectives developed above do not prescribe a particular forecasting algorithm. They establish why household consumption can depend simultaneously on its own history and on information describing expected resources, uncertainty, prices, investment and sectoral economic conditions. They also suggest that the strength of these relationships can vary across time and economic states.

A general predictive representation consistent with this framework is:

$$HFCE_t = f(HFCE_{t-1}, HFCE_{t-2}, \mathbf{Z}_{t-1}) + \varepsilon_t, \quad (6)$$

where $HFCE_t$ denotes household final consumption expenditure growth in quarter t, $HFCE_{t-1}$ and $HFCE_{t-2}$ capture historical consumption dynamics, $\mathbf{Z}_{t-1}$ denotes the vector of lagged macroeconomic and sectoral information available before the forecast quarter, f(·) represents an unknown predictive relationship and ε_t is the forecast error. Equation (6) provides the bridge between economic consumption theory and the comparative forecasting design.

Classical autoregressive models impose a parsimonious representation of temporal dependence. Penalised regression permits a larger information set while controlling coefficient complexity, whereas tree-based and boosting models allow the predictive function to accommodate nonlinearities and interactions without requiring each relationship to be specified in advance. The empirical comparison determines whether this additional flexibility improves forecast performance; the theoretical framework does not assume that it will.

The distinction between prediction and causation remains essential. Equation (6) is a forecasting relationship rather than a structural consumption equation. A variable may improve one-quarter-ahead HFCE prediction because it contains timely information about economic conditions without being an independently identified causal determinant of household expenditure. Explainability results are therefore interpreted as predictive attribution rather than causal effects.

This theoretical structure also explains the use of lagged external predictors in the primary empirical design. A forecast for quarter t should rely on information available before that quarter if it is to represent a genuine one-quarter-ahead prediction. Restricting the external information vector to $t-1$ therefore aligns the economic information set with the forecasting horizon and prevents contemporaneous information from entering the principal prospective exercise. Section 3 operationalises this restriction through expanding-window validation and an untouched final holdout sample

2.3. Conceptual Forecasting Framework

The conceptual forecasting framework integrates the economic foundations developed in Section 2.2 with the empirical forecasting design applied in this study. Its purpose is to show how information available before the forecast quarter is transformed into competing one-quarter-ahead HFCE forecasts, evaluated under a common out-of-sample protocol, and interpreted for short-term macroeconomic surveillance. The framework does not treat forecasting variables as structural causes of household consumption. It organises them as potential sources of predictive information whose usefulness is established through validation and final holdout performance.

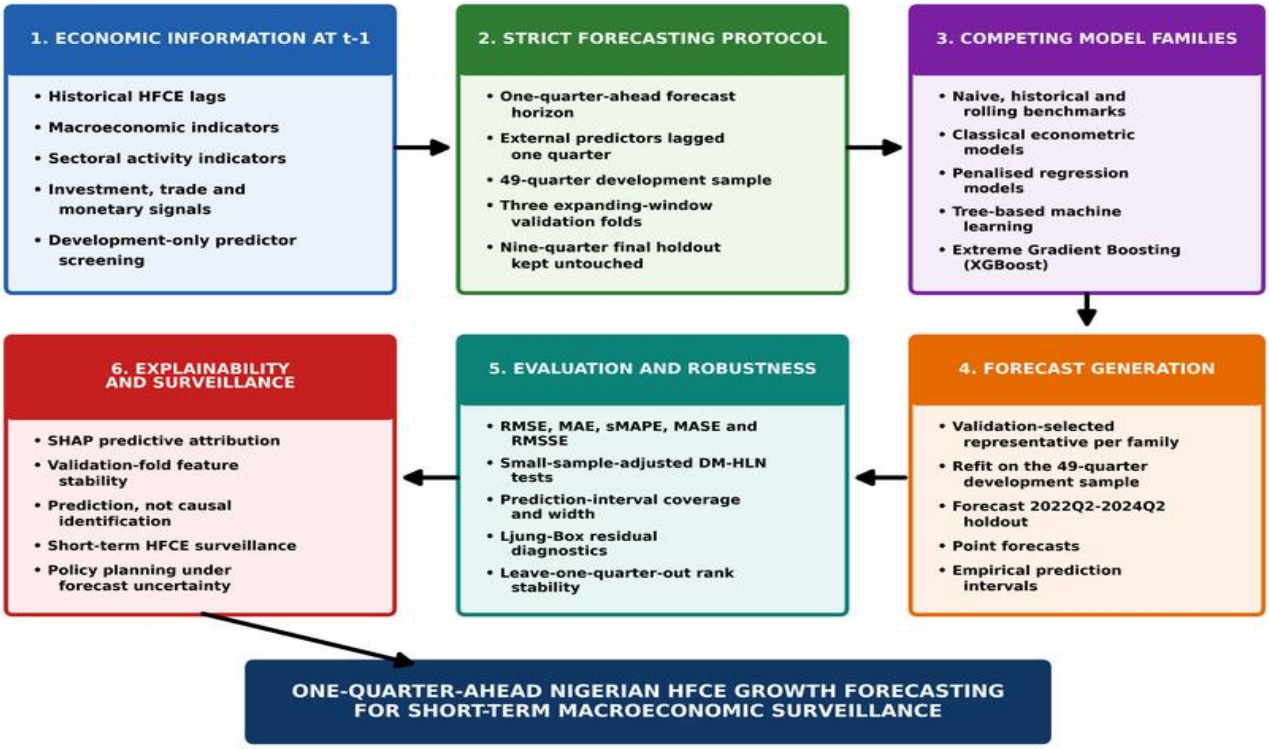

Figure 1: Conceptual forecasting framework for one-quarter-ahead HFCE growth in Nigeria.

Figure 1 presents the framework as a six-stage sequence. The first stage comprises the information set available by quarter $t-1$. It includes historical HFCE lags and eligible lagged macroeconomic and sectoral indicators relating to investment, trade, monetary conditions, production and other dimensions of economic activity. Predictor screening is conducted using development-period information only. This preserves the distinction between information that could genuinely have been available at the forecast origin and information observed later.

The second stage imposes the study's strict forecasting protocol. Each forecast targets HFCE growth one quarter ahead, while external predictors are lagged by one quarter. The 49-quarter development sample is used for estimation, feature selection, specification choice and hyperparameter tuning through three expanding-window validation folds. The final nine-quarter period from 2022Q2 to 2024Q2 is kept completely outside model development and is used only for final out-of-sample evaluation. This separation prevents the holdout observations from influencing the selection of the representative model within any forecasting family.

The conceptual relationship between the information set and each competing model family can be expressed compactly in Equation (7):

$$\hat{y}_{t|t-1}^{(m)} = f_m(\mathbf{Z}_{t-1}, y_{t-1}, y_{t-2}; \boldsymbol{\theta}_m), \tag{7}$$

In Equation (7), the term on the left-hand side denotes the one-quarter-ahead HFCE growth forecast for quarter t produced by model family m. The information vector contains lagged macroeconomic and sectoral indicators available before quarter t, while the two historical HFCE terms capture recent target dynamics. The model-specific parameter vector contains the estimated parameters or hyperparameters associated with each forecasting family. The expression is deliberately generic because the competing econometric, penalised-regression and machine-learning models impose different functional structures on the same prospective forecasting problem.

The third and fourth stages represent model comparison and forecast generation. Simple benchmarks, classical econometric models, penalised regression, tree-based machine-learning models and XGBoost are

https://doi.org/10.64497/jssci.309

evaluated under the same one-quarter-ahead objective. A representative specification is selected within each model family from validation performance and is then refitted on the full development sample before generating point forecasts over the common holdout period. Empirical validation-residual prediction intervals are constructed without using holdout observations for calibration.

The fifth stage evaluates forecast performance across complementary dimensions rather than relying on a single error measure. RMSE, MAE, sMAPE, MASE and RMSSE summarise point forecast accuracy, while the small-sample-adjusted DM-HLN procedure evaluates pairwise predictive-accuracy differences. Prediction-interval coverage and width assess uncertainty, Ljung-Box statistics examine residual dependence, and leave-one-quarter-out analysis assesses whether model rankings are disproportionately driven by individual holdout observations.

The sixth stage introduces explainability and forecasting interpretation. SHAP values are used to describe the predictive attribution of variables within the validation-selected XGBoost model, while validation-fold feature frequencies assess the stability of selected predictors across estimation windows. These outputs support interpretation of the forecasting system but are not treated as causal estimates. The final product of the framework is therefore an uncertainty-aware one-quarter-ahead HFCE forecasting system designed to support short-term assessment of household demand conditions rather than structural policy-effect identification.

3. Methodology

This section presents the empirical design used to compare benchmark, classical econometric, penalised regression, tree-based machine-learning, and extreme gradient boosting forecasts of quarterly household final consumption expenditure (HFCE) growth in Nigeria. The revised methodology is organised around strict chronological information timing, development-only model selection, an untouched final holdout sample, and reproducible expanding-window validation. All equations are numbered and explicitly cross-referenced in the text.

3.1. Research Design and Forecasting Protocol

The study adopts a quantitative comparative forecasting design. The forecasting target is quarterly growth in HFCE at current purchasers' value. The complete sample contains 58 quarterly observations from 2010Q1 to 2024Q2. Observations are ordered chronologically and are never randomly shuffled.

Table 1: Sample structure and chronological allocation.

Sample	Start	End	Observations	Role
Full sample	2010Q1	2024Q2	58	Complete quarterly series
Development sample	2010Q1	2022Q1	49	Model estimation, feature selection, tuning and validation
Final holdout test	2022Q2	2024Q2	9	Final out-of-sample evaluation only

Source: Author's computation from the final modelling dataset. The final nine quarters are not used for model specification, feature selection or hyperparameter tuning.

The primary design is a strict one-quarter-ahead forecasting exercise. For the prediction of HFCE growth in quarter t, all external predictors are lagged by one quarter, while historical HFCE lags are permitted. The forecasting relationship is defined in Equation (8).

$$y_t = f_m\left(\mathbf{X}_{t-1}, y_{t-1}, y_{t-2}, \dots, y_{t-p}; \boldsymbol{\theta}_m\right) + \varepsilon_t. \quad (8)$$

In Equation (8), y_t is HFCE growth in forecast quarter t; $\mathbf{X}_{t-1}$ denotes the vector of external information available through quarter $t-1$; y_{t-1} denotes historical HFCE growth; p is the maximum target-lag order

https://doi.org/10.64497/jssci.309

relevant to the model; $f_m(.)$ is the forecasting function for model m; $\boldsymbol{\theta}$ denotes model parameters or hyperparameters; and ε_t is the forecast error. The information restriction is designed to prevent look-ahead bias and to distinguish the primary exercise from contemporaneous nowcasting.

3.2. Data Sources, Target Construction and Predictor Universe

The master dataset is assembled from official Nigerian macroeconomic and sectoral series, principally from the Central Bank of Nigeria (CBN) and the National Bureau of Statistics (NBS). The target variable is HFCE current purchasers' value quarter-on-quarter growth. Numeric predictors cover national accounts, production and sectoral activity, investment, external trade, transport, monetary and financial conditions, prices, taxes, and related macroeconomic indicators.

To remove direct target leakage, all contemporaneous columns derived from HFCE are excluded from the external predictor pool. HFCE information is reintroduced only through explicit historical lags. All remaining external numeric predictors are shifted by one quarter before model development. Development-only availability screening then retains variables with no more than 40 per cent missing observations, at least 20 non-missing development observations, and non-zero development-sample variance. This procedure leaves 336 eligible predictors in the strict one-quarter-ahead design.

Table 2: Predictor construction and information timing.

Predictor group	Examples / definition	Primary timing	Treatment
Target	HFCE current purchasers' value QoQ growth	Quarter t	Forecast variable
Target history	HFCE growth lag 1; HFCE growth lag 2	$t-1;\ t-2$	Explicit autoregressive information
National accounts	GDP, GFCF, exports, taxes and related aggregates	Lagged one quarter	External predictive information
Sectoral activity	Agriculture, fishing, wood products, transport, education, services and related sectors	Lagged one quarter	Disaggregated real-activity information
Monetary/financial	Treasury-bill rates and other macro-financial series	Lagged one quarter	Financial-conditions information
Other eligible numeric series	Additional non-HFCE variables passing development-only availability screening	Lagged one quarter	Candidate predictor pool

The full machine-readable variable dictionary is retained in the reproducibility outputs. Feature selection is performed inside each validation fold, so the final model does not use all 336 predictors simultaneously.

Table 2 summarises the construction and information timing of the variables entering the forecasting framework. Historical HFCE lags provide autoregressive information, while external macroeconomic, sectoral and financial predictors are lagged by one quarter in the primary specification. This treatment ensures that external information used to forecast quarter t would have been available by the forecast origin.

3.3. Data Preprocessing, Leakage Control and Feature Selection

Preprocessing is fitted within each training fold rather than on the full sample. Missing predictor values are imputed with the median estimated from the relevant training fold. Supervised feature selection uses

https://doi.org/10.64497/jssci.309

SelectKBest with the univariate F-regression score. The candidate feature counts are k = 6, 10 and 15 and are themselves selected by validation performance.

For penalised linear models, the fold-specific pipeline is median imputation, supervised feature selection, standardisation and model fitting. StandardScaler is therefore estimated only on the training observations within each fold. For tree and boosting models, the corresponding pipeline uses median imputation and supervised feature selection without standardisation. The final selected pipeline is refitted on the full 49-quarter development sample only after model selection is frozen.

Development-only availability screening is performed before supervised ranking. The final holdout sample is excluded from missingness thresholds, feature ranking, choice of k, model specification and hyperparameter selection.

3.4. Expanding-Window Validation and Model Selection

Model selection uses three expanding-window validation folds implemented with a fixed eight-quarter validation window. The training set expands through time while each validation block occurs strictly after its corresponding training block.

Table 3: Expanding-window validation design.

Fold	Training period	Training n	Validation period	Validation n
1	2010Q1-2016Q1	25	2016Q2-2018Q1	8
2	2010Q1-2018Q1	33	2018Q2-2020Q1	8
3	2010Q1-2020Q1	41	2020Q2-2022Q1	8

The final holdout period, 2022Q2-2024Q2, is outside all validation folds.

For each candidate specification, validation RMSE is calculated in every fold and averaged across the three folds. Mean validation RMSE is the pre-specified selection criterion. Validation MAE, sMAPE and the standard deviation of fold RMSE are retained as supporting diagnostics. The best specification within each subfamily is selected by mean validation RMSE, after which one representative per broad model family is selected by the same validation criterion. The resulting registry is frozen before the nine-quarter final test sample is evaluated.

3.5. Benchmark Forecasting Models

Three simple benchmarks establish the minimum predictive hurdle. The naïve forecast in Equation (9) uses the most recently observed HFCE growth rate.

$$\hat{y}_t^N = y_{t-1}. \tag{9}$$

The historical-mean benchmark in Equation (10) uses the mean HFCE growth rate calculated from the development sample only.

$$\hat{y}_t^H = \frac{1}{T_D}\sum_{s=1}^{T_D} y_s. \tag{10}$$

The four-quarter rolling mean in Equation (11) uses only the four most recent HFCE observations available before each forecast origin.

$$\hat{y}_t^{R4} = \frac{1}{4}\sum_{j=1}^{4} y_{t-j}. \tag{11}$$

For the rolling benchmark, the forecasting history is updated sequentially after each realised test observation becomes available. No future test value is used to construct an earlier forecast.

https://doi.org/10.64497/jssci.309

3.6. Classical Econometric Forecasting Models

The classical comparison includes autoregressive (AR), ARIMA and SARIMAX specifications. Candidate orders are pre-specified before test evaluation. AR lag candidates are p = 1, 2 and 4. ARIMA candidates are (1,0,0), (2,0,0), (4,0,0), (1,1,1), (2,0,1) and (1,0,1). SARIMAX candidates are (1,0,0), (2,0,0), (4,0,0) and (1,0,1).

The autoregressive specification is written in Equation (12).

$$y_t = c + \sum_{j=1}^{p} \phi_j \, y_{t-j} + \varepsilon_t. \tag{12}$$

AR models include a constant and deterministic time trend. The ARIMA representation is given in Equation (13), where B is the backshift operator, d is the differencing order, φ(B) is the autoregressive polynomial and θ(B) is the moving-average polynomial

$$\phi(B)(1-B)^d y_t = c + \theta(B)\varepsilon_t. \tag{13}$$

SARIMAX extends the ARIMA structure with exogenous predictors as in Equation (14).

$$\phi(B)(1-B)^d y_t = c + \boldsymbol{\beta}'\mathbf{X}_{t-1} + \theta(B)\varepsilon_t. \tag{14}$$

Within each SARIMAX training fold, three exogenous variables are selected using the absolute correlation with the target computed on the training fold only. Missing values are filled using training-fold medians. SARIMAX estimation sets enforce_stationarity=False and enforce_invertibility=False, while order selection remains based on expanding-window validation RMSE rather than final-test performance.

Stationarity is assessed using complementary Augmented Dickey-Fuller (ADF), Kwiatkowski-Phillips-Schmidt-Shin (KPSS) and Phillips-Perron (PP) tests on the development sample. ADF and PP test a unit-root null, whereas KPSS tests a stationarity null. The tests are interpreted jointly. The revised diagnostics support stationarity of HFCE growth and provide majority or concordant evidence for the development-sample regressors used in the classical diagnostics. AIC, BIC and HQIC are also reported for classical candidates as supporting in-sample diagnostics, but they do not replace the forecast-validation selection rule. Auto-ARIMA is not used.

3.7. Penalised Regression Models

Ridge, LASSO and Elastic Net provide linear high-dimensional comparators. Ridge minimises the squared-error criterion with an L2 penalty, as shown in Equation (15).

$$\widehat{\boldsymbol{\beta}}_{\text{Ridge}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p}\beta_j^2\right\}. \tag{15}$$

LASSO uses the L1-penalised objective in Equation (16), allowing coefficients to shrink exactly to zero.

$$\widehat{\boldsymbol{\beta}}_{\text{LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|\right\}. \tag{16}$$

Elastic Net combines L1 and L2 penalties as shown in Equation (17).

$$\widehat{\boldsymbol{\beta}}_{\text{EN}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda\left[\alpha \sum_{j=1}^{p}\left|\beta_j\right| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right]\right\}. \tag{17}$$

Table 4: Penalised-regression validation search space.

Model	Tuned quantity	Candidate values
Ridge	alpha	0.01, 0.1, 1, 10, 100
LASSO	alpha	0.001, 0.01, 0.1, 1
Elastic Net	(alpha, l1_ratio)	(0.01, 0.2/0.5/0.8); (0.1, 0.2/0.5/0.8); (1, 0.2/0.5/0.8)
All penalised models	Selected features k	6, 10, 15

LASSO and Elastic Net use a maximum of 20,000 iterations. Random state is fixed at 42 where supported.

https://doi.org/10.64497/jssci.309

3.8. Tree-Based Machine Learning and XGBoost

The nonlinear comparison includes Decision Tree, Random Forest, Gradient Boosting and XGBoost. Decision trees recursively partition the predictor space. Random Forest averages predictions across bootstrap-based trees to reduce variance. Gradient Boosting constructs an additive sequence of weak learners. A general boosting representation is shown in Equation (18), where η denotes the learning rate and h_m(.) denotes the learner added at boosting stage m.

$$F_M(\mathbf{x}) = F_0(\mathbf{x}) + \sum_{m=1}^{M} \eta\, h_m(\mathbf{x}). \quad (18)$$

XGBoost minimises a regularised boosting objective represented by Equation (19), where ℓ(.) is the predictive loss and Ω(.) penalises model complexity.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \ell\left(y_i, \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i)\right) + \Omega(f_t). \quad (19)$$

Table 5: Tree-based and boosting validation search space.

Model	Tuned quantities	Candidate configurations
Decision Tree	Maximum depth	2, 3, 4, 5
Random Forest	Trees x maximum depth	50/100/200 x depth 3; 50/100/200 x depth 5
Gradient Boosting	Trees, learning rate, depth	(50,0.01,2); (100,0.05,2); (100,0.10,2); (100,0.05,3); (200,0.05,2)
XGBoost	Trees, learning rate, depth	(50,0.03,2); (100,0.05,2); (100,0.03,3); (200,0.03,2)
All tree/boosting models	Selected features k	6, 10, 15

A pre-specified grid search is evaluated through expanding-window validation. Random/Bayesian search is not used. Random state is fixed at 42; parallel execution uses n_jobs = -1 where supported.

3.9. Validation-Frozen Model Registry and Final Test Evaluation

Table 6: Validation-selected representatives frozen before final testing.

Broad family	Frozen representative	Mean validation RMSE	Role
Classical econometric	AR(4)	7.964	Validation-selected family representative
Penalised regression	LASSO, alpha = 0.001, k = 10	10.660	Validation-selected family representative
Tree-based machine learning	Gradient Boosting, 100 trees, learning rate 0.05, depth 3, k = 15	8.345	Validation-selected family representative
Extreme gradient boosting	XGBoost, 100 trees, learning rate 0.03, depth 3, k = 15	8.383	Validation-selected family representative

These configurations are selected before final-test evaluation. Their final-test errors do not alter model selection.

After the registry is frozen, each selected specification is refitted using the complete 49-quarter development sample and is evaluated once over the common nine-quarter final holdout period. Final-test performance is therefore an evaluation result, not a tuning criterion.

3.10. Forecast Accuracy Measures

Forecast performance is evaluated using root mean squared error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (sMAPE), mean absolute scaled error (MASE), and root mean

https://doi.org/10.64497/jssci.309

squared scaled error (RMSSE). Conventional mean absolute percentage error (MAPE) is not used as a principal evaluation criterion because HFCE growth contains negative and near-zero observations, conditions under which percentage-based errors can become unstable. Directional accuracy is not reported because the nine-quarter final test sample yields only eight quarter-to-quarter transitions, a sample size too small to draw statistically meaningful conclusions.

RMSE is defined in Equation (20) and places greater weight on large forecast deviations

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}. \tag{20}$$

MAE is defined in Equation (21) and measures average absolute forecast error.

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|. \tag{21}$$

sMAPE is defined in Equation (22) and provides a symmetric percentage-based error measure.

$$\text{sMAPE} = \frac{100}{n}\sum_{t=1}^{n}\frac{2|y_t-\hat{y}_t|}{|y_t|+|\hat{y}_t|}. \tag{22}$$

MASE is defined in Equation (23), where the scaling denominator is the mean absolute one-step naïve difference calculated from the development sample.

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t-\hat{y}_t|}{\frac{1}{T_D-1}\sum_{s=2}^{T_D}|y_s-y_{s-1}|}. \tag{23}$$

RMSSE is defined in Equation (24), using the corresponding development-sample squared-difference scale.

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{\frac{1}{T_D-1}\sum_{s=2}^{T_D}(y_s-y_{s-1})^2}}. \tag{24}$$

3.11. *Formal Predictive-Accuracy Testing*

Descriptive differences in RMSE and MAE are supplemented by the Harvey-Leybourne-Newbold small-sample-adjusted Diebold-Mariano test. For each model-benchmark pair, the loss differential is formed under squared-error and absolute-error loss. Equation (25) reports the DM statistic and its HLN adjustment for forecast horizon h = 1.

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\text{Var}}(\bar{d})}}, \qquad DM_{\text{HLN}} = DM\sqrt{\frac{T+1-2h+h(h-1)/T}{T}}. \tag{25}$$

Two-sided p-values are computed using a t distribution with T-1 degrees of freedom, where T = 9 for the final holdout comparisons. The null hypothesis is equal expected predictive loss. The test is interpreted cautiously because the short evaluation window limits statistical power.

3.12. *Forecast Uncertainty, Ranking Robustness and Residual Diagnostics*

Prediction uncertainty is assessed using a common empirical validation-residual method across model families. For each validation-selected family representative, out-of-fold expanding-window residuals are pooled before the final test is evaluated. The absolute-residual quantiles use a conservative finite-sample rank rule based on ceil[(n+1)(1-alpha)]. Symmetric prediction intervals are then constructed according to Equation (26).

$$PI_{1-\alpha,t} = [\hat{y}_t - q_{1-\alpha},\ \hat{y}_t + q_{1-\alpha}]. \tag{26}$$

The 80 per cent and 95 per cent intervals are calibrated exclusively from pre-test validation residuals. They are therefore described as empirical validation-residual prediction intervals, not exact Gaussian confidence

intervals. Observed coverage and mean width are reported descriptively on the nine-quarter holdout.

A leave-one-quarter-out sensitivity analysis recomputes RMSE and MAE rankings after omitting each final-test quarter in turn. This diagnostic does not create a new model-selection stage; it examines whether the descriptive ranking is disproportionately driven by one test observation.

Residual serial dependence is assessed with exploratory Ljung-Box tests at lags one and two. Because the final test contains only nine observations, these diagnostics are interpreted cautiously and are not used to claim that residuals are universally white noise.

3.13. Explainable Machine Learning and Feature-Stability Analysis

Explainability is applied only after the forecasting specification has been frozen through validation. For the primary strict one-quarter-ahead framework, SHAP is computed for the validation-selected leading nonlinear XGBoost model. SHAP decomposes the fitted prediction as shown in Equation (27), where phi_0 is the baseline prediction and phi_j is the attribution assigned to predictor j.

$$f(\mathbf{x}) = \phi_0 + \sum_{j=1}^{M} \phi_{j.} \quad (27)$$

Mean absolute SHAP values are used to rank predictive attribution. SHAP values are interpreted as model-specific predictive contributions and not as causal effects. Feature-selection stability is assessed separately by refitting the frozen pipeline in each expanding-window fold and recording the frequency with which each predictor is selected.

Table 7: Predictors retained in the final validation-selected XGBoost attribution analysis

Predictor	Information timing
FISHING QoQ GROWTH (t-1)	External predictor lagged one quarter
HFCE Growth (t-2)	Historical target lag
Gross Fixed Capital Formation, current purchasers' value QoQ growth (t-1)	External predictor lagged one quarter
WOOD AND WOOD PRODUCTS QoQ GROWTH (t-1)	External predictor lagged one quarter
Gross Fixed Capital Formation, 2010 constant purchasers' value QoQ growth (t-1)	External predictor lagged one quarter
AGRICULTURE TOTAL Lag1	External predictor lagged one quarter
ROAD TRANSPORT QoQ GROWTH (t-1)	External predictor lagged one quarter
MOTION PICTURES, SOUND RECORDING AND MUSIC QoQ growth (t-1)	External predictor lagged one quarter
EDUCATION (t-1)	External predictor lagged one quarter
POST AND COURIER SERVICES QoQ GROWTH (t-1)	External predictor lagged one quarter
LIVESTOCK QoQ GROWTH (t-1)	External predictor lagged one quarter
OTHER SERVICES QoQ GROWTH (t-1)	External predictor lagged one quarter
WATER TRANSPORT QoQ GROWTH (t-1)	External predictor lagged one quarter
EXPORTS, current purchasers' value (t-1)	External predictor lagged one quarter
TRANSPORTATION AND STORAGE QoQ GROWTH (t-1)	External predictor lagged one quarter

The selected-feature set is model- and sample-dependent. Validation-fold selection frequency is reported separately from final-model SHAP magnitude.

4. Results and Discussion

This section presents and discusses the empirical results from the one-quarter-ahead forecasting analysis of quarterly household final consumption expenditure (HFCE) growth in Nigeria. The analysis compares benchmark forecasts, classical econometric models, penalised regression methods, tree-based machine-learning

models, and extreme gradient boosting within a common forecasting framework. External predictors are lagged by one quarter to ensure that the information set corresponds to variables available prior to the forecast target period. Model specification and hyperparameter selection are conducted using expanding-window validation, while the final nine-quarter test sample is reserved exclusively for out-of-sample evaluation.

The empirical analysis proceeds in stages. The first subsection examines the statistical characteristics and stationarity properties of household consumption growth and the variables entering the classical forecasting framework. The subsequent analysis evaluates benchmark forecasting performance, classical model specification, and the comparative predictive accuracy of the validation-selected representatives from the principal model families. Formal forecast-comparison tests are then used to determine whether observed differences in predictive performance are statistically meaningful. Prediction uncertainty, model-ranking sensitivity, residual behaviour, feature-selection stability, and SHAP-based predictive attribution are also examined to provide a broader assessment of forecasting reliability and model behaviour.

Forecast performance is evaluated using root mean squared error (RMSE), mean absolute error (MAE), symmetric mean absolute percentage error (sMAPE), mean absolute scaled error (MASE), and root mean squared scaled error (RMSSE). Conventional mean absolute percentage error (MAPE) is not used as a principal evaluation criterion because HFCE growth contains negative and near-zero observations, conditions under which percentage-based errors can become unstable.

In this paper, numerical forecast rankings are interpreted alongside small-sample-adjusted Diebold–Mariano–Harvey–Leybourne–Newbold tests, empirical prediction intervals, leave-one-quarter-out ranking sensitivity, residual diagnostics, and model-attribution evidence. This combination of point forecast accuracy, statistical testing, uncertainty assessment, and robustness analysis provides a more comprehensive basis for comparing the forecasting performance of nonlinear and conventional modelling approaches.

4.1. Preliminary Data Characteristics and Stationarity Diagnostics

Table 8 presents the descriptive statistics for quarterly household final consumption expenditure (HFCE) growth over the study period. The series records a mean quarterly growth rate of 2.815 per cent and a median of 4.288 per cent. The standard deviation of 12.697 percentage points indicates substantial variation in household consumption growth across quarters. This variability is also evident from the observed range, with quarterly growth varying from a minimum of −19.526 per cent to a maximum of 38.336 per cent. The relatively small positive skewness coefficient of 0.124 suggests an approximately symmetric distribution, while the kurtosis value of −0.212 indicates a distribution that is slightly flatter than the normal distribution.

Table 8: Descriptive statistics of quarterly household consumption growth in Nigeria.

Statistic	Value
Mean	2.815
Median	4.288
Standard deviation	12.697
Minimum	-19.526
Maximum	38.336
Skewness	0.124
Kurtosis	-0.212

Source: Author's computation based on the study dataset.

The dispersion of HFCE growth indicates substantial quarter-to-quarter variation in household

consumption expenditure during the sample period. Table 8 shows an average quarterly growth rate of approximately 2.82 per cent, compared with a median of 4.29 per cent. The lower mean suggests that periods of relatively weak or negative expenditure growth affected the average performance of the series. This variability is further reflected in a standard deviation of approximately 12.70 percentage points and a wide range of observed outcomes, from a minimum growth rate of −19.53 per cent to a maximum of 38.34 per cent. The series therefore encompasses substantial contractions and expansions in household consumption growth, presenting a demanding environment for evaluating forecasting performance.

4.1.1. Stationarity and Unit-Root Diagnostics

The time-series properties of HFCE growth and the variables entering the classical forecasting framework were examined using complementary stationarity and unit-root tests. The Augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests evaluate the null hypothesis that a series contains a unit root, while the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test evaluates the null hypothesis that the series is stationary. The use of tests with opposing null hypotheses provides a broader assessment of the stochastic properties of the variables than reliance on a single testing procedure.

Table 9: Stationarity Diagnostics for the Forecast Target and Selected Classical-Model Predictors

Variable	ADF statistic	KPSS statistic	PP statistic
HFCE growth	−3.403**	0.182	−8.746***
Net taxes growth	−2.936**	0.332	−10.228***
Net indirect taxes on products growth	−0.779	0.431	−9.866***
GDP at market price growth	−3.446***	0.145	−15.252***

Source: Author's computation (Python output). Note: For the ADF and PP tests, the null hypothesis is that the series contains a unit root. For the KPSS test, the null hypothesis is stationarity. ***, ** and * denote rejection of the relevant null hypothesis at the 1%, 5% and 10% significance levels, respectively.

The results provide consistent evidence of stationarity for HFCE growth. The ADF statistic of −3.403 has a p-value of 0.011, leading to rejection of the unit-root null hypothesis at the 5 per cent significance level. The PP test similarly rejects the unit-root null, while the KPSS test fails to reject its null hypothesis of stationarity. The agreement across the three tests supports the treatment of quarterly HFCE growth as stationary within the sample. Comparable evidence is obtained for net taxes growth and GDP at market price growth. For both variables, the ADF and PP tests reject the null hypothesis of a unit root at conventional significance levels, while the KPSS tests fail to reject stationarity. The results therefore provide consistent evidence that these series are stationary.

The evidence for net indirect taxes on products growth is less uniform. The ADF test fails to reject the unit-root null hypothesis, with a p-value of 0.825, whereas the PP test strongly rejects the unit-root null. The KPSS p-value of 0.064 also fails to reject stationarity at the 5 per cent significance level. The test results are therefore mixed, although two of the three procedures support stationarity at the adopted significance level. This discrepancy warrants a cautious interpretation of the series rather than an unequivocal classification based on a single test. The stationarity diagnostics provide the statistical basis for the subsequent estimation and comparison of the classical forecasting specifications. Model adequacy and forecasting performance are nevertheless evaluated separately using validation and out-of-sample evidence, since stationarity alone does not establish the predictive superiority of a particular specification.

https://doi.org/10.64497/jssci.309

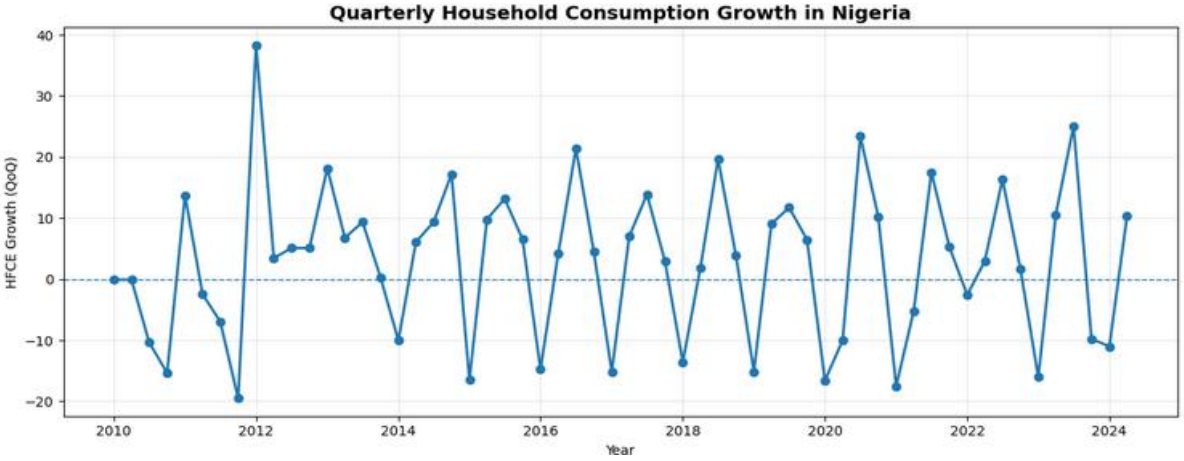

Figure 2: Quarterly household consumption growth in Nigeria.

Figure 2 presents the temporal evolution of quarterly HFCE growth over the sample period. The series exhibits substantial variation, with alternating episodes of positive and negative growth and several pronounced quarter-to-quarter reversals. Periods of contraction are frequently followed by relatively strong positive growth, while episodes of expansion are not consistently sustained across subsequent quarters. The temporal pattern therefore confirms that the dispersion reported in Table 8 reflects recurring fluctuations throughout the sample rather than a small number of isolated observations.

The series also shows limited evidence of a persistent upward or downward trajectory in quarterly growth rates over the full sample period. Positive and negative observations occur across different parts of the sample, although the magnitude of the fluctuations varies considerably over time. This irregular temporal behaviour is relevant to the forecasting exercise because predictive models must accommodate both changes in the direction of HFCE growth and substantial variation in its magnitude. The visual evidence is interpreted descriptively, while the stochastic properties of the series are examined formally through the stationarity tests reported subsequently.

4.2. Benchmark Forecasting Performance

Benchmark models provide reference forecasts against which the predictive performance of the more sophisticated modelling approaches can be evaluated. Three benchmark specifications are considered, namely, the historical mean, a four-quarter rolling mean, and a naïve forecast. Table 10 reports their out-of-sample performance over the nine-quarter test period using RMSE, MAE, sMAPE, MASE and RMSSE.

Table 10: Out-of-sample forecasting performance of benchmark models.

Model	RMSE	MAE	sMAPE	MASE	RMSSE
Historical Mean	12.929	10.865	132.726	0.719	0.693
Rolling Mean (4Q)	13.231	11.371	137.097	0.752	0.709
Naïve	19.244	16.662	154.982	1.102	1.032

Source: Author's computation (Python output). Note: RMSE = root mean squared error; MAE = mean absolute error; sMAPE = symmetric mean absolute percentage error; MASE = mean absolute scaled error; RMSSE = root mean squared scaled error. Lower values indicate better forecasting performance. Bold values indicate the lowest forecast error for each metric.

The historical-mean benchmark records the lowest forecast error across all five error measures. Its RMSE

https://doi.org/10.64497/jssci.309

of 12.929 and MAE of 10.865 are lower than the corresponding values of 13.231 and 11.371 for the four-quarter rolling mean and 19.244 and 16.662 for the naïve forecast. The same ordering is observed for sMAPE, which equals 132.726 for the historical mean, 137.097 for the rolling mean, and 154.982 for the naïve forecast. The consistency of this ranking across the principal error measures identifies the historical mean as the strongest benchmark in terms of forecast magnitude over the final test period.

The scaled-error measures provide additional evidence on benchmark performance. The historical mean records a MASE of 0.719 and an RMSSE of 0.693, while the four-quarter rolling mean records corresponding values of 0.752 and 0.709. Both measures are scaled by the mean one-step naïve forecast error computed from consecutive observations in the development sample. Values below one indicate lower absolute or squared forecast errors than the one-step naïve scaling benchmark constructed from the development sample. The naïve test forecast records a MASE of 1.102 and an RMSSE of 1.032, with both values exceeding one.

The relatively close error performance of the historical mean and four-quarter rolling mean indicates that restricting the averaging benchmark to more recent observations does not improve forecast accuracy over the test period. The naïve forecast produces substantially larger magnitude errors across the evaluation metrics. The historical mean consequently provides the strongest error-based benchmark against which the subsequent econometric and machine-learning forecasts are assessed.

4.3. Classical Econometric Forecast Performance

Classical econometric forecasting models produce substantial improvements relative to the benchmark frameworks discussed earlier. The specifications evaluated in this subsection include autoregressive, ARIMA, and SARIMAX structures designed to capture temporal dependence and cyclical behaviour within household consumption growth. These frameworks extend beyond simple persistence assumptions by incorporating lag dynamics and serial adjustment patterns across time.

Table 11: Expanding-window validation performance of classical forecasting specifications.

Family	Model	Validation RMSE	Validation RMSE SD	Validation MAE
AR	AR(4)	7.964	2.039	7.000
ARIMA	ARIMA(4,0,0)	8.761	1.478	6.954
SARIMAX	SARIMAX(4,0,0)	11.232	3.599	9.528
AR	AR(2)	11.886	0.591	9.638
SARIMAX	SARIMAX(2,0,0)	11.962	2.970	10.226
ARIMA	ARIMA(2,0,0)	12.037	0.313	9.493
ARIMA	ARIMA(2,0,1)	12.257	0.925	9.761
ARIMA	ARIMA(1,1,1)	12.315	0.937	10.081
AR	AR(1)	12.859	0.659	10.402
ARIMA	ARIMA(1,0,0)	12.940	0.697	10.323
SARIMAX	SARIMAX(1,0,1)	13.027	2.518	11.197
SARIMAX	SARIMAX(1,0,0)	13.127	2.384	11.304
ARIMA	ARIMA(1,0,1)	13.604	0.409	10.871

Source: Author's computation from the strict one-quarter-ahead expanding-window validation framework. Note: Model selection is based on mean validation RMSE. The final test sample is not used in specification selection.

The validation results identify AR(4) as the preferred classical specification. It records the lowest mean validation RMSE of 7.964, compared with 8.761 for ARIMA(4,0,0) and 11.232 for SARIMAX(4,0,0), the strongest

https://doi.org/10.64497/jssci.309

alternatives within the ARIMA and SARIMAX classes. Although ARIMA(4,0,0) produces a slightly lower validation MAE of 6.954 than the AR(4) value of 7.000, the pre-specified selection criterion is mean validation RMSE. The AR(4) specification is therefore retained as the classical-family representative before evaluation on the final test period.

The relative performance of the candidate specifications also suggests that additional model complexity does not necessarily improve validation accuracy within the available sample. In particular, the inclusion of lagged exogenous predictors in the SARIMAX specifications does not yield lower validation RMSE than the selected autoregressive model under the strict one-quarter-ahead information structure. This result concerns predictive performance within the present sample and should not be interpreted as evidence that external macroeconomic variables are generally uninformative for household consumption forecasting.

When evaluated on the previously unseen nine-quarter test period, the validation-selected AR(4) model records an RMSE of 11.179 and an MAE of 10.231. Its sMAPE is 140.350, while the scaled-error measures are 0.677 for MASE and 0.599 for RMSSE. Both scaled measures remain below one, indicating lower absolute and squared forecast errors than the corresponding one-step naïve scaling benchmark constructed from the development sample.

Relative to the simple benchmark results reported in Section 4.2, the AR(4) forecast records lower RMSE and MAE than the historical-mean, four-quarter rolling-mean, and naïve forecasts. The same pattern is evident in the scaled-error measures. These differences establish the descriptive out-of-sample ranking of the classical specification, while the statistical significance of forecast-accuracy differences is examined separately through the formal predictive-accuracy tests presented later.

4.4. Comparative Forecasting Performance across Model Families

This subsection compares the out-of-sample performance of the validation-selected representatives from the principal forecasting families. The comparison includes the AR(4) classical econometric model, LASSO penalised regression, Gradient Boosting, and XGBoost. Each specification is selected using expanding-window validation within the development sample before evaluation on the common nine-quarter final test period. Table 12 reports the resulting RMSE, MAE, sMAPE, MASE, and RMSSE values.

Table 12: Comparative forecasting performance across model families.

Model Family	Validation-Selected Model	RMSE	MAE	sMAPE	MASE	RMSSE
Extreme Gradient Boosting	XGBoost, 100 trees, learning rate = 0.03, depth = 3	9.646	7.772	96.256	0.514	0.517
Tree-Based Machine Learning	Gradient Boosting, 100 trees, learning rate = 0.05, depth = 3	10.591	9.235	120.900	0.611	0.568
Classical Econometric	AR(4)	11.179	10.231	140.350	0.677	0.599
Penalised Regression	LASSO, $\alpha = 0.001$	18.216	14.407	109.258	0.953	0.976

Source: Author's computation from the strict one-quarter-ahead forecasting framework. Note: Bold values indicate the lowest forecast error for each evaluation metric. Model-family representatives are selected using expanding-window validation only; the final test sample is not used for model selection.

Table 12 shows that XGBoost records the lowest observed forecast error across all five evaluation measures. Its RMSE of 9.646 is lower than the corresponding values of 10.591 for Gradient Boosting, 11.179 for AR(4), and

https://doi.org/10.64497/jssci.309

18.216 for LASSO. A similar ordering emerges for MAE, with XGBoost recording 7.772 compared with 9.235 for Gradient Boosting, 10.231 for AR(4), and 14.407 for LASSO. The XGBoost specification therefore occupies the strongest descriptive position within the final test sample on both squared-error and absolute-error criteria.

The percentage and scaled-error measures provide further evidence on relative forecast performance. XGBoost records the lowest sMAPE at 96.256, followed by LASSO at 109.258, Gradient Boosting at 120.900, and AR(4) at 140.350. The ranking differs on sMAPE from that obtained using RMSE and MAE, illustrating the importance of assessing forecasts across complementary error measures rather than relying on a single criterion. XGBoost nevertheless remains the lowest-error specification under sMAPE.

All four family representatives record MASE and RMSSE values below one. XGBoost again produces the lowest values, with a MASE of 0.514 and an RMSSE of 0.517. Gradient Boosting follows with values of 0.611 and 0.568, while AR(4) records 0.677 and 0.599. LASSO produces the largest scaled errors at 0.953 and 0.976. Values below one indicate that each selected specification generates lower absolute and squared forecast errors than the corresponding one-step naïve scaling benchmark derived from the development sample.

The relative ranking therefore places XGBoost first across the five reported error metrics, followed generally by Gradient Boosting and the classical AR(4) specification. LASSO performs least favourably on RMSE, MAE, MASE, and RMSSE, although its sMAPE is lower than those of Gradient Boosting and AR(4). The results provide descriptive evidence that the nonlinear boosting models achieve lower out-of-sample forecast errors within the nine-quarter evaluation period. Given the limited size of the final test sample, however, differences in point forecast errors alone are insufficient to establish statistical superiority. Formal predictive-accuracy tests are therefore considered separately in the subsequent analysis.

4.5. Formal Predictive-Accuracy Comparison

The descriptive forecast rankings reported in the preceding subsection indicate differences in out-of-sample error across the validation-selected model families. Given the limited nine-quarter final test period, however, numerical differences in RMSE and MAE do not by themselves establish that one forecasting model is statistically more accurate than another. Formal pairwise forecast-comparison tests are therefore used to assess whether the observed differences in predictive loss are statistically distinguishable.

The analysis employs the small-sample-adjusted Diebold–Mariano–Harvey–Leybourne–Newbold (DM-HLN) test under squared-error and absolute-error loss functions. Squared-error loss provides a formal comparison corresponding to the emphasis placed on RMSE, while absolute-error loss complements this assessment by evaluating differences in absolute forecast errors. The null hypothesis in each comparison is that the competing forecasts have equal expected predictive loss.

Table 13 presents formal pairwise predictive-accuracy tests for the four validation-selected model-family representatives against the naïve, historical-mean, and four-quarter rolling-mean benchmarks. The tests complement the descriptive error rankings reported in Section 4.4 by assessing whether differences in forecast loss are statistically distinguishable within the final test period.

XGBoost provides the strongest formal evidence of improvement over naïve persistence. Under squared-error loss, the comparison produces a DM-HLN statistic of −2.478 with a p-value of 0.038. Under absolute-error loss, the corresponding statistic is −3.256 with a p-value of 0.012. Both results reject the null hypothesis of equal predictive accuracy at the 5 per cent level. The negative test statistics indicate that XGBoost incurs lower average predictive loss than the naïve forecast under both loss criteria.

Gradient Boosting produces a similar result against the naïve benchmark. Its DM-HLN statistic is −2.436 under squared-error loss, with a p-value of 0.041, and −2.805 under absolute-error loss, with a p-value of 0.023. Both comparisons are statistically significant at the 5 per cent level. The formal evidence therefore indicates that

the predictive improvements of both boosting specifications over naïve persistence are supported under squared and absolute forecast-loss criteria.

Table 13: Small-sample-adjusted DM-HLN predictive-accuracy tests.

Model	Benchmark	Squared Error DM-HLN (p-value)	Absolute Error DM-HLN (p-value)
XGBoost	Naïve	−2.478 (0.038)	−3.256 (0.012)
XGBoost	Historical Mean	−0.925 (0.382)	−0.865 (0.412)
XGBoost	Rolling Mean 4Q	−1.040 (0.329)	−1.065 (0.318)
Gradient Boosting	Naïve	−2.436 (0.041)	−2.805 (0.023)
Gradient Boosting	Historical Mean	−0.673 (0.520)	−0.465 (0.654)
Gradient Boosting	Rolling Mean 4Q	−0.789 (0.453)	−0.641 (0.540)
AR(4)	Naïve	−1.941 (0.088)	−1.994 (0.081)
AR(4)	Historical Mean	−0.917 (0.386)	−0.300 (0.772)
AR(4)	Rolling Mean 4Q	−1.087 (0.309)	−0.579 (0.578)
LASSO ($\alpha = 0.001$)	Naïve	−0.308 (0.766)	−0.465 (0.654)
LASSO ($\alpha = 0.001$)	Historical Mean	1.043 (0.327)	0.666 (0.524)
LASSO ($\alpha = 0.001$)	Rolling Mean 4Q	1.028 (0.334)	0.612 (0.558)

Source: Author's computation from the strict one-quarter-ahead forecasting framework. Note: The table reports small-sample-adjusted Diebold–Mariano–Harvey–Leybourne–Newbold (DM-HLN) statistics, with two-sided p-values in parentheses. The null hypothesis is equal predictive accuracy between the selected model and the stated benchmark. Negative DM-HLN statistics indicate lower average predictive loss for the selected model relative to the benchmark, while positive statistics indicate higher average predictive loss. Bold entries denote rejection of the null hypothesis at the 5 per cent significance level.

The evidence for AR(4) is less conclusive. Its comparisons with the naïve benchmark yield p-values of 0.088 under squared-error loss and 0.081 under absolute-error loss. Although both DM-HLN statistics are negative, indicating lower average predictive loss than the naïve benchmark, neither comparison reaches the conventional 5 per cent significance threshold. LASSO also fails to demonstrate statistically significant improvement over the naïve benchmark under either loss function.

A stronger qualification emerges when the models are compared with the historical-mean and four-quarter rolling-mean benchmarks. None of the four selected models records a statistically significant difference from either benchmark at the 5 per cent level. XGBoost, Gradient Boosting and AR(4) generally produce negative DM-HLN statistics in these comparisons, indicating lower average predictive loss, but the corresponding p-values are too large to reject equal predictive accuracy. LASSO produces positive statistics against both the historical and rolling means, indicating higher average predictive loss, although these differences are also statistically insignificant.

Across the 24 individual tests represented by the 12 model–benchmark pairs and two loss functions, only four reject equal predictive accuracy at the 5 per cent level. These are XGBoost versus naïve under squared and absolute loss and Gradient Boosting versus naïve under squared and absolute loss. The formal evidence therefore supports statistically significant predictive gains for the two boosting models over naïve persistence, but does not establish their statistical superiority over the stronger historical-mean and four-quarter rolling-mean benchmarks.

This result provides an important qualification to the descriptive ranking in Section 4.4. XGBoost remains the lowest-error model across RMSE, MAE, sMAPE, MASE and RMSSE, but its descriptive lead should not be interpreted as universal statistical superiority. The nine-quarter final test period also limits the power of the

https://doi.org/10.64497/jssci.309

DM-HLN tests, so both significant and non-significant comparisons should be interpreted cautiously.

4.6. Forecast Uncertainty and Prediction-Interval Performance

Point forecast accuracy provides only one dimension of predictive performance because it does not quantify the uncertainty surrounding individual forecasts. Forecast uncertainty is therefore evaluated using empirical prediction intervals calibrated from residuals generated during expanding-window validation. The calibration procedure uses only validation residuals available before the final test period, thereby preserving the separation between model development and final out-of-sample evaluation. Table 14 reports observed coverage and average interval width for nominal 80 and 95 per cent prediction intervals across the four validation-selected model families.

Table 14: Empirical prediction-interval performance across validation-selected model families.

Model Family	Selected Model	Validation Residuals	Observed 80% Coverage (%)	Observed 95% Coverage (%)	Mean 80% Interval Width	Mean 95% Interval Width
Classical Econometric	AR(4)	24	44.444	100.000	21.837	35.548
Extreme Gradient Boosting	XGBoost (100 trees, learning rate = 0.03, depth = 3)	24	77.778	88.889	22.520	37.213
Penalised Regression	LASSO (α = 0.001)	24	55.556	88.889	32.352	56.590
Tree-Based Machine Learning	Gradient Boosting (100 trees, learning rate = 0.05, depth = 3)	24	66.667	100.000	21.354	39.679

Source: Author's computation from the strict one-quarter-ahead forecasting framework. Note: Prediction intervals are empirical validation-residual intervals calibrated exclusively from expanding-window validation residuals before the final test period. Observed coverage is evaluated over the nine-quarter final test sample.

The interval results reveal differences in both empirical coverage and interval width across the selected forecasting models. For the nominal 80 per cent interval, XGBoost records the closest observed coverage at 77.778 per cent, corresponding to seven of the nine test observations. Gradient Boosting records 66.667 per cent coverage, while LASSO and AR(4) attain 55.556 and 44.444 per cent, respectively. The XGBoost interval therefore provides the closest empirical correspondence to the nominal 80 per cent coverage level within the final test period.

At the nominal 95 per cent level, AR(4) and Gradient Boosting cover all nine test observations, while XGBoost and LASSO each cover eight of nine observations, corresponding to 88.889 per cent. Coverage alone, however, does not provide a complete assessment of interval performance because wider intervals can increase coverage at the cost of reduced precision. The average 95 per cent interval widths range from 35.548 for AR(4) to 56.590 for LASSO. XGBoost records an average width of 37.213, while Gradient Boosting records 39.679.

The width results also distinguish the forecasting families at the 80 per cent level. Gradient Boosting produces the narrowest average interval at 21.354, followed by AR(4) at 21.837 and XGBoost at 22.520. LASSO produces a substantially wider average interval of 32.352 while attaining only 55.556 per cent empirical coverage. The penalised regression specification therefore exhibits comparatively weak interval performance

https://doi.org/10.64497/jssci.309

within this evaluation period.

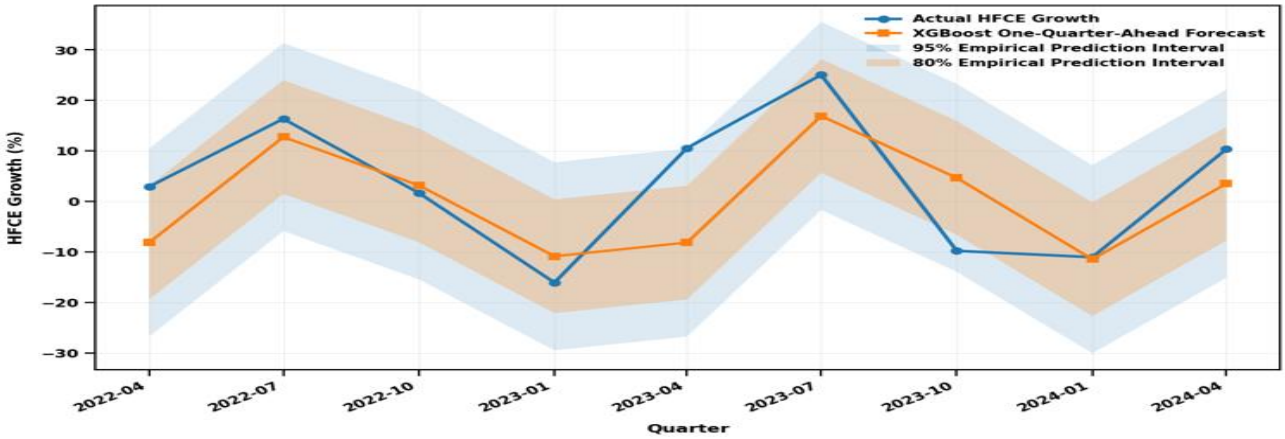

Figure 3: One-quarter-ahead XGBoost forecast with empirical validation-residual prediction intervals.

For XGBoost, which records the lowest point forecast errors in Section 4.4, the empirical intervals provide additional information about forecast uncertainty. Its 80 per cent interval contains seven of nine realised values, while its 95 per cent interval contains eight of nine. Figure 3 illustrates the one-quarter-ahead XGBoost forecast path together with the corresponding empirical 80 and 95 per cent prediction intervals. The figure shows that uncertainty bands widen around the point forecasts while retaining the same validation-frozen forecasting specification used for the point forecast evaluation.

4.7. Forecast Robustness and Residual Diagnostics

The robustness of the out-of-sample rankings is examined using leave-one-quarter-out sensitivity analysis, quarter-specific absolute forecast errors, residual autocorrelation diagnostics, and a supplementary alternative-information-set comparison. These assessments complement the point forecast measures and formal predictive-accuracy tests by examining whether the observed rankings depend disproportionately on individual test observations or particular features of the primary forecasting information set.

4.7.1. Leave-One-Quarter-Out Ranking Stability

The final test period contains only nine quarterly observations, so a model's aggregate RMSE or MAE could potentially be influenced by one unusually large forecast error. A leave-one-quarter-out sensitivity exercise therefore recalculates the model-family rankings after sequentially excluding each of the nine test quarters. Table 15 summarises the resulting ranking stability.

The sensitivity results show substantial stability in the position of XGBoost. Extreme Gradient Boosting remains first by RMSE in all nine leave-one-quarter-out iterations, producing a mean RMSE rank of exactly 1.000. The same result is obtained using MAE, with XGBoost occupying first place in all nine iterations. Its leading descriptive position in the full test sample is therefore not attributable to any single test quarter.

Gradient Boosting occupies either second or third position across the RMSE sensitivity exercises and records a mean rank of 2.111, while AR(4) records a mean rank of 2.889 and similarly varies between second and third place. LASSO remains fourth in every omission. The persistence of these rankings indicates that the broad ordering observed in Section 4.4 is comparatively insensitive to the exclusion of individual test observations. This result strengthens the descriptive robustness of the ranking but does not alter the statistical conclusions from the DM-HLN tests reported in Section 4.5.

Table 15: Leave-one-quarter-out stability of out-of-sample model rankings.

Model Family	Mean RMSE Rank	Best RMSE Rank	Worst RMSE Rank	RMSE Rank-1 Occurrences	Mean MAE Rank	MAE Rank-1 Occurrences
Extreme Gradient Boosting	1.000	1	1	9/9	1.000	9/9
Tree-Based Machine Learning	2.111	2	3	0/9	2.111	0/9
Classical Econometric	2.889	2	3	0/9	2.889	0/9
Penalised Regression	4.000	4	4	0/9	4.000	0/9

Source: Author's computation from the strict one-quarter-ahead forecasting framework. Note: Each iteration removes one of the nine final-test quarters and recomputes the ranking using the remaining eight observations. Rank 1 denotes the lowest forecast error.

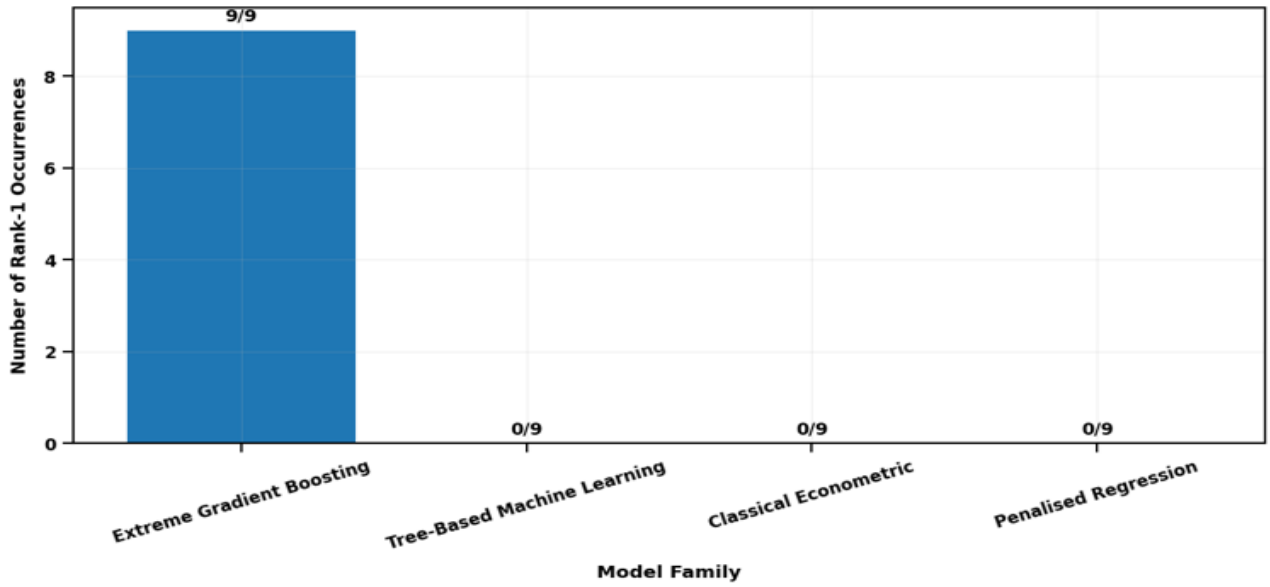

Figure 4: Leave-one-quarter-out stability of out-of-sample RMSE rankings.

Figure 4 visually represents the rank-one occurrence results. XGBoost retains the lowest RMSE in all nine omission exercises, while none of the other family representatives occupies first position. Given the short final test period, this result provides useful evidence that the leading RMSE performance is not concentrated in one or two influential quarters.

4.7.2. Quarter-Specific Forecast Errors

Aggregate error measures can also conceal substantial variation in forecast performance across individual quarters. Figure 5 therefore plots the absolute forecast errors of the four validation-selected family representatives throughout the final test period.

The quarter-specific error profiles reveal considerable variation across both models and forecast dates. XGBoost and Gradient Boosting generally record smaller absolute errors than the competing specifications, although neither model dominates every individual quarter. AR(4) also performs comparatively well in some periods but records larger errors during several quarters. LASSO exhibits the most pronounced variation, including particularly large absolute errors during parts of the later test period.

https://doi.org/10.64497/jssci.309

The figure also demonstrates why aggregate rankings should not be interpreted as uniform quarter-by-quarter dominance. XGBoost records the lowest aggregate RMSE and MAE, but its relative advantage varies across the nine observations. The leave-one-quarter-out results are therefore especially informative because they show that the model's first-place aggregate ranking persists even when each test observation is removed in turn.

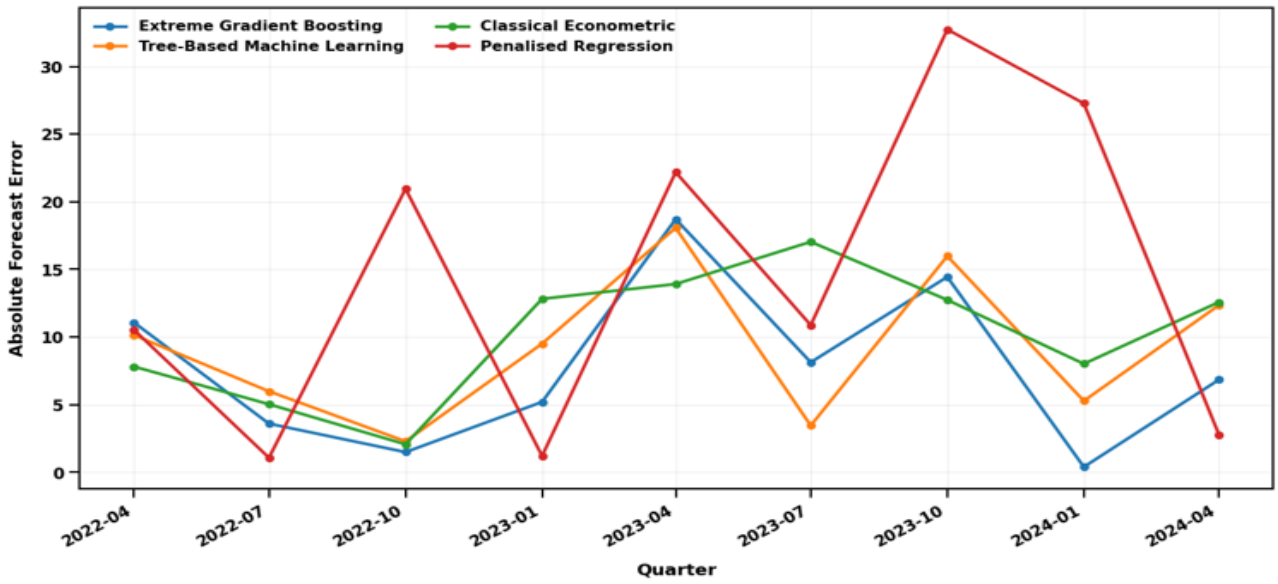

Figure 5: Absolute forecast error across the final test period.

4.7.3. Residual Autocorrelation Diagnostics

We assess residual dependence using Ljung–Box tests at lags one and two. Table 16 reports the test statistics and p-values for each validation-selected family representative.

Table 16: Ljung–Box residual autocorrelation diagnostics.

Model Family	Selected Model	Lag	Ljung–Box Statistic	p-value
Extreme Gradient Boosting	XGBoost	1	0.086	0.770
Extreme Gradient Boosting	XGBoost	2	6.536	0.038
Tree-Based Machine Learning	Gradient Boosting	1	0.063	0.802
Tree-Based Machine Learning	Gradient Boosting	2	6.408	0.041
Classical Econometric	AR(4)	1	0.124	0.725
Classical Econometric	AR(4)	2	8.345	0.015
Penalised Regression	LASSO ($\alpha = 0.001$)	1	1.259	0.262
Penalised Regression	LASSO ($\alpha = 0.001$)	2	3.465	0.177

Source: Author's computation from final-test forecast residuals. Note: The null hypothesis of the Ljung–Box test is absence of residual autocorrelation up to the stated lag. The diagnostics are interpreted cautiously because the final test contains only nine observations.

None of the four models shows statistically significant first-order residual autocorrelation. The lag-one p-values range from 0.262 for LASSO to 0.802 for Gradient Boosting. At lag two, however, the results change for three specifications. XGBoost records a p-value of 0.038, Gradient Boosting 0.041, and AR(4) 0.015, providing

https://doi.org/10.64497/jssci.309

evidence of residual dependence at the 5 per cent level. LASSO does not reject the null at lag two, with a p-value of 0.177.

These findings caution against describing the forecast residuals of the leading models as white noise. In particular, XGBoost combines the lowest observed point forecast errors and highly stable leave-one-quarter-out rankings with evidence of remaining second-order residual dependence. Forecast accuracy and residual serial independence therefore provide distinct dimensions of model evaluation. The Ljung–Box results should also be regarded as exploratory because the nine-observation test sample provides limited information for residual diagnostic testing.

4.7.4. Robustness to the Forecast Information Set

The primary results use the strict one-quarter-ahead specification in which external predictors are lagged by one quarter. A contemporaneous-information specification is examined separately as a supplementary robustness exercise. Figure 6 compares RMSE under the two information structures across the validation-selected model families.

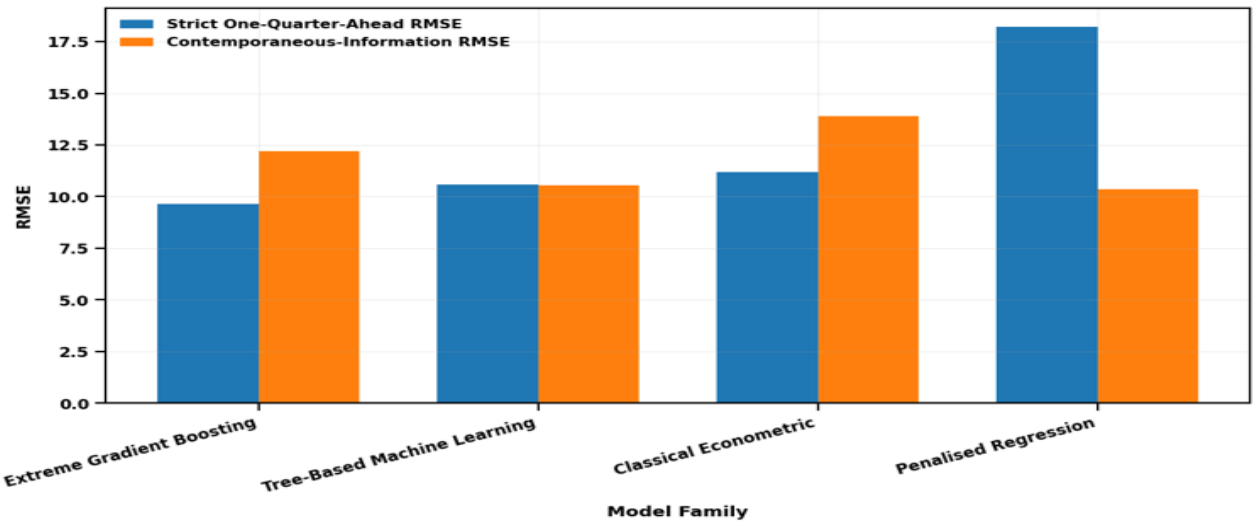

Figure 6: Forecast performance under alternative information sets.

The comparison indicates that forecast performance is sensitive to the information set and that the direction of the change differs across model families. XGBoost performs better under the strict one-quarter-ahead specification than under the contemporaneous alternative, while AR(4) also records lower RMSE under the primary specification. Gradient Boosting displays relatively similar performance across the two information structures. The largest change occurs for penalised regression, for which the contemporaneous-information specification records substantially lower RMSE than the strict one-quarter-ahead specification.

These differences reinforce the importance of defining the forecasting horizon before model evaluation. The contemporaneous-information results do not replace or re-rank the primary models because the study is specified as a one-quarter-ahead forecasting exercise. The strict specification therefore remains the basis for the principal empirical conclusions, while the contemporaneous-information results provide supplementary evidence on the sensitivity of forecast performance to information availability.

4.8. Explainable Machine-Learning Evidence and Feature Stability

The preceding analysis identifies XGBoost as the lowest-error forecasting specification within the final test sample. Explainable machine-learning analysis is therefore applied to the validation-selected strict XGBoost

https://doi.org/10.64497/jssci.309

model to examine the relative predictive contributions of the variables entering its forecast function. SHapley Additive exPlanations (SHAP) are used for this purpose. Because the study is predictive rather than causal, SHAP values are interpreted as measures of predictive attribution and not as estimates of causal economic effects.

Figure 7: Predictive attribution in the validation-selected XGBoost model.

Figure 7 reports mean absolute SHAP values for the principal predictors in the frozen XGBoost specification. Fishing quarter-on-quarter growth at (t-1) records the largest mean absolute SHAP value, at 5.351, and therefore has the greatest average predictive attribution within the fitted model. The magnitude is considerably larger than that of the next-ranked variable, HFCE growth at (t-2), which records a mean absolute SHAP value of 2.441. The prominence of the historical HFCE lag indicates that information contained in earlier household consumption growth remains important to the model's one-quarter-ahead forecasts.

Gross fixed capital formation measures also occupy prominent positions in the attribution ranking. Quarter-on-quarter growth in gross fixed capital formation at current purchasers' value records a mean absolute SHAP value of 1.445, while the corresponding measure at 2010 constant purchasers' value records 1.029. Wood and Wood Products quarter-on-quarter growth at (t-1) records a value of 1.177, and the lagged Agriculture Total indicator records 0.909. These results indicate that the fitted XGBoost forecasts draw predictive information from a mixture of investment, sectoral activity and historical consumption variables rather than from a single broad macroeconomic aggregate.

Road Transport quarter-on-quarter growth at (t-1) records a mean absolute SHAP value of 0.554. Several additional sectoral indicators contribute more modestly to forecast formation, including Motion Pictures, Sound Recording and Music growth and Education, both at approximately 0.322; Post and Courier Services growth at 0.321; Livestock growth at 0.278; Other Services growth at 0.269; and Water Transport growth at 0.259. Exports at (t-1) and Transportation and Storage growth at (t-1) appear further down the final-model SHAP ranking, with mean absolute values of 0.171 and 0.150, respectively.

The SHAP ranking should be interpreted as a description of the fitted XGBoost model's predictive allocation rather than as a ranking of structural determinants of household consumption. A high mean absolute SHAP value indicates that a variable contributes materially to variation in the model's forecasts over the evaluated observations. It does not establish that changes in that predictor cause subsequent changes in household consumption growth, nor does a lower SHAP value imply that the variable is economically unimportant outside the fitted forecasting model.

4.8.1. Validation-Fold Feature-Selection Stability

Final-model attribution is complemented by an assessment of feature-selection stability across the expanding-window validation folds. This distinction is important because a variable may receive substantial

attribution after final model estimation even though its selection is sensitive to the particular validation window. Table 17 summarises the most consistently selected predictors for the XGBoost specification.

Table 17: Validation-fold feature-selection stability for the XGBoost model.

Predictor	Validation Folds Selected	Selection Frequency (%)	Stability Classification
Exports, current purchasers' value (t-1)	3/3	100.0	Stable across all folds
Gross fixed capital formation measure (t-1)	3/3	100.0	Stable across all folds
Net indirect taxes on products (t-1)	3/3	100.0	Stable across all folds
91-day Treasury bill rate (t-1)	3/3	100.0	Stable across all folds
GDP at basic price measure (t-1)	2/3	66.7	Recurrent
GDP at market price measure (t-1)	2/3	66.7	Recurrent
Gross fixed capital formation, current purchasers' value (t-1)	2/3	66.7	Recurrent
HFCE growth (t-2)	2/3	66.7	Recurrent
Information and Communication growth (t-1)	2/3	66.7	Recurrent
Net taxes, current purchasers' value (t-1)	2/3	66.7	Recurrent
Oil shock (t-1)	2/3	66.7	Recurrent
Telecommunications and Information Services growth (t-1)	2/3	66.7	Recurrent

Source: Author's computation from expanding-window validation feature-selection outputs. Note: Selection frequency is the proportion of the three validation folds in which a predictor was retained. Predictors selected in all three folds are classified here as stable across validation windows, while those selected in two folds are described as recurrent. Additional predictors selected in only one of the three folds are not shown in the compact table.

Four predictors exhibit complete validation-fold stability: exports, a gross fixed capital formation measure, net indirect taxes on products, and the 91-day Treasury bill rate. Each is selected in all three validation folds. A second group, including GDP measures, current-value gross fixed capital formation, lagged HFCE growth, Information and Communication growth, net taxes, the oil-shock indicator, and Telecommunications and Information Services growth, appears in two of the three folds. Other variables are selected in only one validation window, indicating greater sensitivity to the composition and timing of the development sample. The underlying validation output therefore shows a combination of persistent and fold-specific predictive information rather than complete invariance in the selected feature set.

An important distinction emerges when the validation-stability results are compared with the final-model SHAP ranking. Fishing growth at (t-1), for example, records the largest final-model mean absolute SHAP value, yet it is selected in only one of the three validation folds. Conversely, exports is selected in every validation fold but occupies a relatively modest position in the final SHAP ranking. These results are not contradictory. Feature-selection frequency measures the persistence with which a predictor enters the forecasting specification across alternative validation windows, whereas mean absolute SHAP measures the magnitude of its predictive attribution once included in the final frozen model.

HFCE growth at (t-2) provides an intermediate case. It is selected in two of the three validation folds and

subsequently records the second-largest mean absolute SHAP value in the final XGBoost specification. Gross fixed capital formation also appears prominently in both analyses, with investment-related measures displaying high validation-selection frequencies and relatively large final-model SHAP contributions. The correspondence for these predictors provides some evidence that their predictive relevance is not confined solely to the final fitted sample.

The combined evidence therefore supports a qualified interpretation of model explainability. The XGBoost forecast relies on predictive information distributed across historical consumption, investment and a broad range of lagged sectoral indicators. Some variables display substantial final-model attribution but limited validation-fold stability, while others are selected consistently across validation windows despite more modest final SHAP contributions. The explainability results should consequently be interpreted as model-specific predictive evidence whose importance can vary across estimation windows, rather than as a fixed ordering of the economic determinants of Nigerian household consumption.

4.9. Discussion of Findings and Policy Implications

The empirical results provide several insights into the forecasting of quarterly household consumption growth in Nigeria. The principal finding is that XGBoost records the lowest observed forecast errors under the strict one-quarter-ahead information structure, followed by Gradient Boosting and the classical AR(4) specification. The ranking is consistent across RMSE, MAE, sMAPE, MASE and RMSSE for XGBoost, while the leave-one-quarter-out analysis places the model first by RMSE and MAE in every test-quarter omission. The formal DM-HLN tests, however, provide a more qualified result. XGBoost and Gradient Boosting significantly outperform the naïve forecast under squared and absolute loss, but their numerical advantages over the historical-mean and four-quarter rolling-mean benchmarks are not statistically significant at the 5 per cent level. The evidence therefore supports a predictive advantage over simple persistence without establishing universal superiority over all competing forecasting rules.

The comparatively strong performance of the boosting models is consistent with a wider forecasting literature in which flexible machine-learning methods perform well when the predictor environment is multidimensional and relationships may vary across observations. Nigerian evidence provides similar results in other macroeconomic applications. Akinjobi and Omooseti [8] report gains from machine-learning methods in inflation forecasting, Obite et al. [12] find that machine-learning specifications perform competitively against classical methods in modelling Nigerian crude-oil production, and Gbadebo [13] documents useful forecasting performance from Random Forest and Extreme Gradient Boosting in exchange-rate prediction. International evidence also reports growing use of nonlinear and machine-learning structures in GDP, transport and energy forecasting [18]–[23]. The present results extend this evidence to quarterly household consumption forecasting, an area in which Nigerian empirical applications remain comparatively limited.

The results should not, however, be interpreted as evidence that flexible machine-learning models invariably dominate conventional time-series approaches. AR(4) performs considerably better than the naïve benchmark on the principal error measures and remains relatively close to the two boosting specifications in the descriptive ranking. This indicates that the historical dynamics of HFCE growth retain substantial forecasting information even when a broader macroeconomic predictor set is available. The importance of lagged HFCE is also reflected in the XGBoost attribution results, where HFCE growth at (t-2) is the second most important predictor by mean absolute SHAP value. Persistence in consumption growth therefore remains useful for prediction, while the lower errors obtained from the boosting models suggest that additional information contained in lagged sectoral and macroeconomic variables can complement the target series' own history.

https://doi.org/10.64497/jssci.309

The relatively weak performance of LASSO provides a further insight into the forecasting structure. Penalised regression is valuable in high-dimensional settings because shrinkage and variable selection can reduce estimation instability and manage correlated predictor sets [14]. In this study, however, LASSO records substantially larger RMSE and MAE than the validation-selected XGBoost, Gradient Boosting and AR(4) models. One possible interpretation is that variable shrinkage alone is insufficient to reproduce the predictive relationships contained in the available information set when these relationships are not well represented by a common linear specification. This interpretation is consistent with the broader distinction between penalised linear forecasting and more flexible forecasting systems discussed by Smeekes and Wijler [14]. It remains a model-specific predictive result rather than proof that the underlying economic relationships are structurally nonlinear.

The forecast-uncertainty and residual evidence further qualifies the point-forecast rankings. XGBoost achieves observed coverage of 77.8 per cent for the nominal 80 per cent empirical prediction interval and 88.9 per cent for the nominal 95 per cent interval. At the same time, its Ljung–Box result at lag two indicates remaining residual dependence. Gradient Boosting and AR(4) display a similar second-order residual-autocorrelation signal. These findings show that lower aggregate forecast errors do not imply that all predictable structure has been eliminated from the residuals. The practical implication is that model evaluation should not rely exclusively on a single leaderboard measure. Point accuracy, formal forecast comparisons, uncertainty intervals, ranking sensitivity and residual diagnostics provide complementary information about forecast reliability.

The alternative-information-set analysis reinforces this point. Forecast performance changes when contemporaneous predictors replace the strict one-quarter-ahead information structure, but the direction and size of the change differ across model families. This sensitivity is economically relevant because macroeconomic forecasting performance depends not only on modelling flexibility but also on what information would genuinely have been available at the forecast origin. The strict specification therefore provides the more appropriate basis for prospective one-quarter-ahead forecasting, whereas the contemporaneous specification is more closely related to an information-rich current-period assessment. The distinction is important for institutional forecasting because apparent gains obtained from information unavailable at the forecast date would provide an unrealistic indication of operational predictive performance.

The SHAP results offer additional insight into the informational content used by the validation-selected XGBoost model. Fishing quarter-on-quarter growth at (t-1) records the largest mean absolute SHAP value, followed by HFCE growth at (t-2), gross fixed capital formation measures, Wood and Wood Products growth, Agriculture, and Road Transport growth. These variables span household-consumption persistence, investment activity, primary production and transport-related economic activity. Their presence suggests that one-quarter-ahead HFCE forecasts benefit from information distributed across several dimensions of real economic activity rather than from a single aggregate indicator. This broad informational structure is consistent with evidence that alternative, sectoral and higher-frequency indicators can improve macroeconomic forecasting when they contain information not fully captured by conventional aggregates [1–3], [6], [7], [18], [19], [24].

The economic interpretation of these SHAP results requires caution. A large SHAP value indicates that a predictor makes a comparatively large contribution to variation in the fitted XGBoost forecasts; it does not show that the predictor causes household consumption to change. For example, the high attribution assigned to Fishing growth may reflect its informational association with broader agricultural activity, rural income conditions, food-sector production or other contemporaneously related features embedded in the dataset. Similarly, the prominence of gross fixed capital formation may reflect information about broader investment

and economic activity that is useful for anticipating household expenditure conditions. These are plausible economic interpretations of predictive associations and should not be treated as identified causal channels.

The validation-fold stability analysis also demonstrates why final-model SHAP rankings should not be read as permanent structural rankings of economic importance. Exports, a gross fixed capital formation measure, net indirect taxes on products and the 91-day Treasury bill rate are selected in all three validation folds, whereas Fishing growth, despite having the largest final-model SHAP attribution, appears in only one validation fold. This difference indicates that predictive importance can vary with the estimation window. For forecasting institutions, stable recurring variables may be particularly useful for maintaining a core information set, while fold-specific variables may provide additional episodic information when economic conditions change.

The broader finding that sectoral and activity indicators contribute to forecasting is consistent with the increasing use of non-traditional information in macroeconomic prediction. Carrière-Swallow and Labbé [1] show that Google Trends can strengthen short-term forecasting in an emerging-market setting, while [6] and [7] demonstrate the forecasting usefulness of search-based information. Fasulo et al. [3] apply Google Trends directly to quarterly household consumption expenditure, and [24] compare survey indicators with Google Trends for private-consumption forecasting. Although the information sources differ from those used in this paper, the common implication is that consumption forecasting can benefit from information beyond a narrow set of conventional macroeconomic aggregates.

The results have several implications for macroeconomic monitoring in Nigeria. First, institutions responsible for forecasting household demand need not replace conventional econometric models with a single machine-learning system. The evidence supports a forecasting architecture in which strong benchmark and autoregressive models remain part of the comparison set, while validation-selected machine-learning models provide additional predictive capacity. The lack of statistically significant superiority over the historical and rolling means reinforces the value of retaining simple benchmarks as continuing performance hurdles rather than treating them as disposable preliminary models.

Second, the results support greater use of timely and disaggregated economic information in household-consumption forecasting. The predictive relevance of investment, agricultural, transport and other sectoral indicators suggests that forecasting systems may benefit from integrating data across the National Bureau of Statistics, Central Bank of Nigeria and other official statistical sources within a common real-time monitoring framework. The purpose would not be to infer causal effects from each variable, but to exploit their combined information content for earlier detection of changes in household expenditure conditions.

Third, prediction intervals should accompany point forecasts used for policy surveillance. The final results show that uncertainty around one-quarter-ahead HFCE forecasts is substantial and varies across models. Reporting a point forecast without an associated uncertainty range could therefore give policymakers a false impression of precision. Empirical prediction intervals calibrated without using the final evaluation period provide a more transparent indication of the range within which future consumption growth may plausibly fall.

Fourth, forecast-monitoring systems should incorporate periodic revalidation rather than assume that the predictor set remains permanently stable. The difference between SHAP attribution and validation-fold selection frequency indicates that predictors can gain or lose forecasting relevance across estimation windows. A practical forecasting system should therefore monitor feature stability, residual dependence, forecast-error accumulation and performance relative to simple benchmarks as new quarterly observations become available.

The policy value of the forecasting framework consequently lies primarily in improving economic surveillance rather than identifying policy-treatment effects. More timely recognition of a probable weakening

or strengthening in household consumption growth could assist fiscal and monetary authorities in assessing aggregate demand conditions, preparing revenue and expenditure scenarios, and evaluating whether existing macroeconomic projections remain plausible. Decisions about specific interventions would still require complementary structural and causal evidence. The forecasting results provide information about the likely path and uncertainty of household expenditure, not evidence that changing any individual SHAP-ranked predictor would mechanically produce a corresponding change in consumption.

In summary, the evidence supports a balanced conclusion. XGBoost provides the strongest descriptive point-forecast performance and a highly stable leave-one-quarter-out ranking, while formal testing establishes significant gains over naïve persistence but not universal superiority over the stronger benchmark alternatives. AR(4) retains useful forecasting information through consumption persistence, and the weaker performance of LASSO shows that high-dimensional variable selection alone does not guarantee stronger forecasts. The explainability evidence further indicates that predictive information is distributed across historical consumption, investment and sectoral activity, with substantial differences between final-model attribution and validation-fold stability. These results support the use of diversified, validation-based and uncertainty-aware forecasting systems for Nigerian household-consumption surveillance.

5. Conclusion and Recommendations

5.1. Conclusion

This study compares the forecasting performance of benchmark, classical econometric, penalised-regression and machine-learning approaches for quarterly household final consumption expenditure growth in Nigeria. The analysis is implemented as a strict one-quarter-ahead forecasting exercise in which external predictors are lagged by one quarter and model specification, predictor selection and hyperparameter tuning are confined to the development sample. The completed dataset comprises 58 quarterly observations from 2010Q1 to 2024Q2, with the first 49 observations used for model development and expanding-window validation and the remaining nine quarters, covering 2022Q2–2024Q2, reserved exclusively for final out-of-sample evaluation.

The comparative evidence shows that forecasting performance differs materially across model families. Among the validation-selected representatives, XGBoost records the lowest RMSE, MAE, sMAPE, MASE and RMSSE over the final holdout period. Gradient Boosting ranks second, followed by the classical AR(4) specification, while LASSO records the largest RMSE and MAE. The leave-one-quarter-out analysis reinforces the stability of the principal ranking: XGBoost remains first by both RMSE and MAE in all nine omission exercises. The evidence therefore indicates that nonlinear boosting methods can provide useful forecasting gains for Nigerian household consumption when model development is subjected to disciplined time-series validation.

The comparison nevertheless provides an important qualification to any general claim of machine-learning superiority. Small-sample-adjusted DM-HLN tests show that XGBoost and Gradient Boosting significantly outperform the naïve benchmark under squared-error and absolute-error loss at the 5 per cent level. Their advantages over the historical-mean and four-quarter rolling-mean benchmarks are not statistically significant at the same threshold. Numerical superiority in the final holdout sample should therefore be distinguished from statistically established predictive superiority. The evidence supports significant forecasting gains relative to naïve persistence, but it does not demonstrate that machine learning universally dominates every simpler forecasting alternative.

The performance of AR(4) reinforces this conclusion. Although it does not match the forecast accuracy of

the two boosting models, the classical econometric specification remains comparatively competitive and outperforms the simple benchmarks on several principal accuracy measures. Historical HFCE dynamics therefore retain substantial forecasting information even when a broader set of macroeconomic and sectoral predictors is available. Conversely, the weaker performance of LASSO indicates that coefficient shrinkage and variable selection do not automatically translate into superior forecasts within a relatively short quarterly macroeconomic sample.

Forecast uncertainty and residual diagnostics provide further qualifications. Empirical validation-residual prediction intervals reveal considerable uncertainty around quarterly HFCE forecasts, while Ljung–Box tests identify evidence of second-order residual dependence for XGBoost, Gradient Boosting and AR(4). Forecasting performance should consequently be evaluated across several dimensions rather than through point-error rankings alone. Accuracy metrics, formal predictive-accuracy tests, uncertainty estimates, residual diagnostics and sensitivity analysis provide complementary evidence on the reliability of competing models.

The explainability analysis further shows that XGBoost draws predictive information from several dimensions of the Nigerian economy. Fishing quarter-on-quarter growth at t−1 records the largest mean absolute SHAP attribution, followed by HFCE growth at t−2, gross fixed capital formation measures and selected production and transport indicators. The validation-fold evidence, however, shows that final-model attribution and predictor-selection stability are distinct concepts. Some variables with large SHAP contributions are not consistently selected across validation windows, while exports, gross fixed capital formation, net indirect taxes on products and the 91-day Treasury bill rate exhibit relatively strong recurrence.

These results should be interpreted as evidence about forecasting relationships rather than causal economic effects. SHAP values indicate how predictors contribute to the fitted model's forecasts; they do not establish that changing a particular sectoral or macroeconomic variable would cause a corresponding change in household consumption. The economic interpretation of predictor importance should therefore remain consistent with the predictive objective of the study.

The central conclusion is consequently not that machine learning should replace econometric forecasting of household consumption in Nigeria. The evidence supports a comparative forecasting strategy in which model complexity must justify itself through genuine out-of-sample performance. XGBoost provides the strongest forecasts in the present application, but Gradient Boosting, AR(4) and simple benchmarks provide essential comparators for determining whether those gains are persistent, statistically meaningful and operationally useful. This finding supports a diversified forecasting architecture combining econometric discipline, flexible machine learning, realistic information timing, explicit benchmark comparison and systematic assessment of forecast uncertainty.

5.2. Policy and Forecasting Recommendations

The empirical findings support several recommendations for institutions responsible for macroeconomic forecasting and household-demand surveillance in Nigeria.

First, forecasting institutions should adopt a multi-model comparison framework rather than rely permanently on either econometric or machine-learning specifications. XGBoost provides the strongest descriptive performance in this study, but the DM-HLN evidence does not establish statistically significant superiority over every benchmark. Classical econometric models and simple benchmarks should therefore remain integral components of the forecasting system and should be evaluated alongside machine-learning challengers as new quarterly observations become available.

Second, institutions such as the Central Bank of Nigeria and the National Bureau of Statistics could employ a champion-challenger forecasting architecture in which the currently strongest model is continuously

https://doi.org/10.64497/jssci.309

evaluated against alternative econometric, benchmark and machine-learning specifications. Model status should depend on prospective forecasting evidence rather than methodological preference. Such an approach would permit machine-learning methods to complement established econometric forecasting without presuming that greater model complexity necessarily produces superior forecasts.

Third, operational forecasts should preserve realistic information timing. Predictors used to forecast HFCE growth in quarter t should correspond to information that would have been available when the forecast was generated. The sensitivity of forecasting performance to alternative information sets demonstrates that apparent accuracy can change when contemporaneous information is introduced. Release calendars, predictor lags, data vintages and forecast origins should therefore be documented explicitly in operational forecasting systems.

Fourth, forecasting institutions should continue developing integrated databases containing timely macroeconomic and disaggregated sectoral indicators. The results show that useful predictive information is distributed across historical consumption, investment and several areas of sectoral activity. A broader information environment can therefore complement aggregate indicators when its contribution is established through genuine out-of-sample validation. Predictor inclusion should remain evidence-based and should not be interpreted as proof of causal policy channels.

Fifth, institutional forecasts should report prediction intervals alongside point forecasts. The observed uncertainty around quarterly HFCE forecasts demonstrates that a single numerical forecast provides an incomplete representation of the likely range of outcomes. Interval forecasts can improve macroeconomic surveillance, fiscal planning and scenario analysis by communicating the uncertainty surrounding projected household-demand conditions.

Sixth, all forecasting models should undergo periodic revalidation. Predictor relevance, forecast errors and model rankings can change as additional observations become available or macroeconomic conditions evolve. Forecasting institutions should monitor point accuracy, DM-HLN comparisons where sample size permits, residual dependence, prediction-interval performance and feature stability. New observations should expand the evaluation evidence without being repeatedly recycled for retrospective model optimisation.

Seventh, explainability should accompany institutional use of complex machine-learning models. SHAP and related attribution methods can assist analysts in determining which information a forecasting model is using and whether predictive reliance shifts over time. Explainability should support model governance, diagnostic assessment and economic interpretation, but causal policy decisions should continue to rely on appropriate structural or causal-identification methods.

The recommended institutional framework is therefore neither exclusively econometric nor exclusively machine-learning based. It is a comparative, continuously evaluated forecasting system in which simple benchmarks establish minimum performance standards, econometric models provide transparent temporal structure, and machine-learning models compete on their ability to generate genuine prospective forecasting improvements.

5.3. Study Limitations and Directions for Future Research

Several limitations define the scope within which the findings should be interpreted.

The first concerns the relatively short final evaluation sample. Although the complete dataset contains 58 quarterly observations, strict separation between model development and final evaluation leaves nine quarters for the untouched holdout period. This limits the precision of estimated forecast-performance differences and the statistical power of formal predictive-accuracy tests. The leave-one-quarter-out analysis shows that XGBoost's ranking is not attributable to a single holdout observation, but it cannot substitute for evaluation

over a substantially longer independent forecasting period.

The limited test sample is particularly relevant to the DM-HLN results. Failure to reject equal predictive accuracy should not be interpreted as evidence that two forecasting methods are equivalent. The absence of statistical significance against the historical-mean and rolling-mean benchmarks may partly reflect limited statistical power. Conversely, the significant results against the naïve benchmark should be interpreted within the same small-sample constraint.

Second, the study examines a single national HFCE growth series for Nigeria. Forecasting relationships may reflect characteristics of the Nigerian economy, the quarterly HFCE measure and the historical period covered by the sample. The resulting model rankings cannot automatically be generalised to other countries, subnational consumption series or alternative measures of household expenditure. Future research could apply the same comparative validation framework to other African economies and related household-demand indicators.

Third, the predictor environment remains constrained by data availability, publication timing and potential revisions to Nigerian macroeconomic statistics. Lagging external predictors by one quarter substantially reduces the risk of look-ahead bias, but the dataset is not a complete historical real-time vintage database reproducing exactly the information released at every original forecast date. Future studies could construct vintage datasets that explicitly incorporate preliminary releases, subsequent revisions and asynchronous publication schedules.

Fourth, predictor-selection and SHAP results are model and sample dependent. The validation-fold analysis demonstrates that variables with high final-model predictive attribution are not necessarily selected consistently across estimation windows. Fishing growth provides a prominent example. Predictor rankings should therefore not be regarded as permanent structural characteristics of Nigerian household consumption. Longer samples could permit investigation of whether predictor relevance changes across inflationary, exchange-rate, growth or other macroeconomic regimes.

Fifth, residual diagnostics show that additional temporal structure may remain exploitable. Evidence of second-order residual autocorrelation for XGBoost, Gradient Boosting and AR(4) suggests scope for investigating alternative lag structures, forecast combinations or hybrid econometric-machine-learning specifications. Such extensions should preserve development-only specification choice and genuinely independent final evaluation.

Sixth, the empirical prediction intervals are calibrated from expanding-window validation residuals rather than derived from complete parametric predictive distributions. Coverage estimates are also sensitive to individual observations because the final test sample contains only nine quarters. Future research could compare alternative uncertainty-quantification procedures, including conformal prediction, quantile-based machine learning and probabilistic forecasting, using longer evaluation samples.

Seventh, the study evaluates a one-quarter-ahead forecasting horizon. The relative performance of econometric and machine-learning models may change as the forecasting horizon increases. Future work should therefore extend the comparative design to two-quarter-ahead, four-quarter-ahead and multi-horizon forecasts to determine whether the advantages observed for boosting methods persist over longer horizons.

Finally, the study is predictive rather than causal. Forecast accuracy, feature-selection frequencies and SHAP values identify information useful for forecasting HFCE growth but do not estimate structural economic effects. Future research concerned with the effects of monetary conditions, investment, sectoral activity or other variables on household consumption should complement forecasting models with structural econometric or credible causal-identification strategies.

Author Contributions

The author solely conceptualised the study, collected and processed the data, developed the forecasting framework, implemented the econometric and machine-learning models, conducted the empirical analysis, interpreted the findings, and prepared the manuscript.

Funding Statement

The author received no specific funding for this research.

Data Availability

The data used in this study were obtained from publicly available macroeconomic and statistical sources, including the Central Bank of Nigeria (CBN) and the National Bureau of Statistics (NBS). Processed data and forecasting codes are available from the author upon reasonable request.

Acknowledgments

The author appreciates the valuable contributions of researchers whose studies and datasets supported the development of this work.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Y. Carrière-Swallow and F. Labbé, "Nowcasting with Google Trends in an emerging market," *Journal of Forecasting*, vol. 32, no. 4, pp. 289–298, 2011, doi: 10.1002/for.1252.

[2] F. D'Amuri and J. Marcucci, "The predictive power of Google searches in forecasting US unemployment," *International Journal of Forecasting*, vol. 33, no. 4, pp. 801–816, 2017, doi: 10.1016/j.ijforecast.2017.03.004.

[3] A. Fasulo, A. Guandalini, and M. D. Terribili, "Google Trends for nowcasting quarterly household consumption expenditure," *Rivista Italiana di Economia Demografia e Statistica*, vol. 71, no. 4, pp. 5–14, 2017.

[4] G. Koop, S. McIntyre, J. Mitchell, and A. Poon, "Nowcasting 'true' monthly U.S. GDP during the pandemic," *National Institute Economic Review*, vol. 256, no. 1, pp. 44–70, 2021, doi: 10.1017/nie.2021.8.

[5] M. Marcellino and V. Sivec, "Nowcasting GDP growth in a small open economy," *National Institute Economic Review*, vol. 256, no. 1, pp. 127–161, 2021, doi: 10.1017/nie.2021.13.

[6] R. F. Niesert, J. A. Oorschot, C. P. Veldhuisen, K. Brons, and R.-J. Lange, "Can Google search data help predict macroeconomic series?," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1163–1172, 2020, doi: 10.1016/j.ijforecast.2018.12.006.

[7] D. Kohns and A. Bhattacharjee, "Nowcasting growth using Google Trends data: A Bayesian Structural Time Series model," *International Journal of Forecasting*, vol. 39, no. 3, pp. 1384–1412, 2023, doi: 10.1016/j.ijforecast.2022.05.002.

[8] S. O. Akinjobi and T. O. Omooseti, "Forecasting Nigeria's inflation components using machine learning and econometric models: A comparative analysis," *Research Square*, 2025, doi: 10.21203/rs.3.rs-8160354/v1.

[9] D. C. Bartholomew, H. C. Iwu, and I. P. Ibemere, "A deep recurrent neural network approach to inflation forecasting in Nigeria: Evaluating the impact of feature selection methods," *Franklin Open*, vol. 13, p. 100443, 2025.

[10] J. F. Pfahler, "Exchange rate forecasting with advanced machine learning methods," *Journal of Risk and Financial Management*, vol. 15, no. 2, pp. 1–17, 2022.

[11] M. Qureshi, N. Ahmad, S. Ullah, and A. R. Mustafa, "Forecasting real exchange rate using artificial intelligence and time series models," *Heliyon*, vol. 9, no. 5, p. e16335, 2023, doi: 10.1016/j.heliyon.2023.e16335.

[12] C. P. Obite, A. Chukwu, D. C. Bartholomew, U. I. Nwosu, and G. E. Esiaba, "Classical and machine learning modeling of crude oil production in Nigeria," *Energy Reports*, vol. 7, pp. 3497–3505, 2021.

[13] A. D. Gbadebo, "Leveraging machine learning for exchange rate prediction: A business and financial management perspective in Nigeria," *Reviews of Management Sciences*, vol. 6, no. 2, pp. 37–52, 2024.

[14] S. Smeekes and E. Wijler, "Macroeconomic forecasting using penalized regression methods," *International Journal of Forecasting*, vol. 34, no. 3, pp. 408–430, 2018, doi: 10.1016/j.ijforecast.2018.01.001.

[15] A. B. Muhammad, O. O. Ishaq, B. B. Janet, M. A. Alhassan, S. Manzo, and S. Shehu, "A hybrid ExpAR-FIGARCH-ANN model for time series forecasting," *Journal of Statistical Sciences and Computational Intelligence*, vol. 1, no. 4, pp. 283–293, 2025, doi: 10.64497/jssci.68.

[16] A. U. Abubakar and I. Akeyede, "Bayesian regime-switching threshold autoregression for nonlinear dynamics and risk forecasting in the Nigerian stock exchange," *Journal of Statistical Sciences and Computational Intelligence*, vol. 2, no. 2, pp. 313–325, 2026, doi: 10.64497/jssci.184.

[17] P. O. Evans, M. O. Abifarin, and E. A. Ali, "Modeling and forecasting petroleum motor spirit (PMS) prices in Nigeria: An ARIMA-based time series analysis," *Journal of Statistical Sciences and Computational Intelligence*, vol. 1, no. 4, pp. 485–494, 2025, doi: 10.64497/jssci.82.

[18] M. Maaouane *et al*., "Using neural network modelling for estimation and forecasting of transport sector energy demand in developing countries," *Energy Conversion and Management*, vol. 258, p. 115556, 2022.

[19] U. U. Imoh and M. M. Rad, "Analysis and prediction of traffic conditions using machine learning models on Ikorodu Road in Lagos State, Nigeria," *Infrastructures*, vol. 10, no. 5, p. 122, 2025.

[20] N. Ghosh, "Advancing GDP forecasting: Machine learning approaches across global, regional, and country-specific contexts," 2025, doi: 10.22541/au.176402164.41325037/v1.

[21] X. Tang and Y. Xie, "Exchange rate forecasting: A deep learning framework combining adaptive signal decomposition and dynamic weight optimization," *International Journal of Financial Studies*, vol. 13, no. 3, p. 151, 2025.

[22] I. Elbatal, M. Sarwar, F. Jamal, M. Daniyal, Z. Hussain, and A. Ben Ghorbal, "Modelling on gross domestic product annual growth rate data by using time series, machine learning, and probability models," *Journal of Radiation Research and Applied Sciences*, vol. 18, p. 101481, 2025.

[23] S. A. Eshiemogie, P. P. Aielumoh, T. Okamkpa, M. C. Jude, L. Efe, A. N. Amenaghawon, H. Darmokoesoemo, and H. S. Kusuma, "A machine learning-supported framework for predicting Nigeria's optimal energy storage and emission reduction potentials," *Renewable Energy Focus*, vol. 53, p. 100677, 2025.

[24] T. Schmidt and S. Vosen, "Forecasting private consumption: Survey-based indicators vs. Google Trends," *Ruhr Economic Papers*, no. 155, 2009.

[25] M. Simionescu and J. Cifuentes-Faura, "Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal," *Journal of Policy Modeling*, vol. 44, no. 1, pp. 1–21, 2022, doi: 10.1016/j.jpolmod.2021.09.011.

[26] M. Simionescu and A. G. Raišienė, "A bridge between sentiment indicators: What does Google Trends tell us about COVID-19 pandemic and employment expectations in the EU new member states?," *Technological Forecasting and Social Change*, vol. 173, p. 121170, 2021, doi: 10.1016/j.techfore.2021.121170.

[27] Y. Uematsu and S. Tanaka, "High-dimensional macroeconomic forecasting and variable selection via penalized regression," *The Econometrics Journal*, vol. 22, no. 1, pp. 34–56, 2019, doi: 10.1111/ectj.12117.

[28] S. Tonini, F. Chiaromonte, and A. Giovannelli, "*On the impact of serial dependence on penalized regression methods*," LEM Working Paper Series, no. 2022/21, Laboratory of Economics and Management, Sant'Anna School of Advanced Studies, Pisa, Italy, Jul. 2022.

[29] J. H. Ospina-Holguín and A. M. Padilla-Ospina, "Penalised regressions vs. autoregressive moving average models for forecasting inflation," *Económicas CUC*, vol. 41, no. 1, pp. 65–80, 2020, doi: 10.17981/econcuc.41.1.2020.Econ.3.

[30] M. Friedman, *A Theory of the Consumption Function. Princeton*, NJ, USA: Princeton University Press, 1957.

[31] F. Modigliani and R. Brumberg, "*Utility analysis and the consumption function: An interpretation of cross-section data*," in Post-Keynesian Economics, K. K. Kurihara, Ed. New Brunswick, NJ, USA: Rutgers University Press, 1954, pp. 388–436.

[32] A. Deaton, "Saving and liquidity constraints," *Econometrica*, vol. 59, no. 5, pp. 1221–1248, 1991.

[33] C. D. Carroll, "A theory of the consumption function, with and without liquidity constraints," *Journal of Economic Perspectives*, vol. 15, no. 3, pp. 23–45, 2001.

[34] J. B. Ajefu, "Income shocks, informal insurance mechanisms, and household consumption expenditure," *International Journal of Social Economics*, vol. 44, no. 12, pp. 1818–1832, 2017, doi: 10.1108/IJSE-04-2015-0094.

[35] J. Animashaun and A. Wossink, "How do households cope during aggregate shocks? Evidence from the 2009–2015 oil crisis in Nigeria," *Resources Policy*, vol. 95, Art. no. 105170, 2024, doi: 10.1016/j.resourpol.2024.105170.

Disclaimer/Publisher's Note: The views, opinions, and content expressed in all articles are solely those of the respective author(s) and contributor(s) and do not necessarily reflect those of the JSSCI, its editors, or the publisher. JSSCI and its editorial team assume no responsibility for any harm or damage resulting from the use of information, methods, or products mentioned in the publication.