Quantitative Finance and Economics

QFE, 10(4): 728–752.
DOI: 10.3934/QFE.2026028
Received: 17 January 2025
Revised: 22 April 2026
Accepted: 02 June 2026
Published: 10 November 2026

https://www.aimspress.com/journal/QFE

Research article

Do pretrained models improve financial time series forecasting?

Linghan Huang[1], Yunhan Li[2] and Xuan Feng[3,*]

[1] ISCTE Business School, Lisbon, 1649-026, Portugal
[2] University of Edinburgh Business School, Edinburgh, EH89YL, UK
[3] Independent Researcher, Nice, 06200, France

* **Correspondence:** Email: xuan.feng@edhec.com.

Abstract: In this paper, we evaluated whether large pretrained time-series models improve financial forecasting across global markets. Using over 10,000 assets spanning equities, bonds, and currencies, we compared ARIMA, XGBoost, LSTM, TimeGPT, and TimesFM under a rolling out-of-sample framework. We found that large pretrained models enhance cross-market stability and consistency, particularly in structurally stable markets such as Japan. However, gains in predictive accuracy were modest and highly conditional: XGBoost remained competitive in bonds and currencies, and no model dominated in efficient markets. Overall, the evidence suggests that pretrained models improve robustness rather than deliver universally superior forecasts, and that statistical improvements do not necessarily translate into economically meaningful advantages.

Keywords: financial time-series forecasting; large pretrained models; machine learning in finance; cross-market analysis; asset-class heterogeneity

JEL Codes: C53, C22, C58, G17, G12

1. Introduction

Forecasting financial time series remains a central yet unresolved problem in quantitative finance. Despite extensive evidence of weak predictability in asset returns, a growing body of research shows that model performance varies substantially across markets, asset classes, and time periods.

Traditional financial theory, most prominently the Efficient Market Hypothesis (EMH), argues that asset prices rapidly and fully reflect all available information, implying that systematic excess returns

are difficult to achieve through forecasting (Malkiel & Fama, 1970). Nevertheless, a large empirical literature documents persistent deviations from strict market efficiency, including volatility clustering, nonlinear dependence, structural breaks, and regime shifts in financial time series (Engle, 1982; Shiller, 2003). These observations raise an important empirical question: Whether modern data-driven models can extract stable predictive signals across heterogeneous financial environments, or whether their performance remains highly context-dependent.

Early forecasting approaches relied primarily on linear statistical models, such as autoregressive integrated moving average (ARIMA) models and their extensions. These models are grounded in rigorous statistical theory and remain widely used due to their interpretability and computational efficiency. However, their reliance on linearity, stationarity, and limited flexibility constrains their ability to capture the complex, nonlinear, and time-varying dynamics that characterize modern financial markets. As financial data have grown in volume and complexity, these limitations have become increasingly pronounced, particularly in multi-asset and cross-market settings.

In response, machine learning methods have gained prominence in financial forecasting. Algorithms such as decision trees, support vector machines, and gradient boosting models can model nonlinear relationships and high-dimensional interactions without strong parametric assumptions. Among these methods, XGBoost has attracted particular attention due to its scalability, regularization mechanisms, and strong empirical performance across a wide range of prediction tasks. Empirical studies demonstrate that machine learning models often outperform traditional statistical methods in specific forecasting contexts, especially when rich feature sets and nonlinear dependencies are present. However, their empirical success is often sensitive to validation design, as improper handling of temporal dependence may lead to overfitting and overly optimistic performance estimates.

The development of deep learning further expanded the forecasting toolkit available to financial researchers and practitioners. Recurrent neural networks, and in particular long short-term memory (LSTM) architectures, were designed to address the challenges of sequential data and long-term dependencies. LSTM models have been shown to capture temporal patterns that are difficult to model using traditional approaches and have achieved notable success in equity return prediction and related tasks. Moreover, attention-based models and Transformer architectures have been introduced into time series forecasting, offering improved scalability and the ability to model long-range dependencies through self-attention mechanisms. Despite their flexibility, these models often require extensive tuning and large effective sample sizes, raising concerns about stability and generalization in noisy financial environments.

Building on these advances, a new generation of large pretrained time-series models has emerged. Advances in large pretrained time-series models (such as TimeGPT and Google Times FM) introduce a new paradigm, where models are trained on massive and diverse datasets to learn transferable temporal representations. In principle, such models may improve generalization across markets without extensive task-specific tuning. However, it remains unclear whether these gains translate into consistent improvements in financial forecasting, where signals are weak and nonstationary.

Despite rapid methodological progress, the literature exhibits several important gaps. First, most empirical studies focus on a narrow set of markets, typically the United States or other Anglo-American economies, limiting the generalizability of their conclusions. Financial markets differ substantially across countries in terms of institutional structure, sectoral composition, investor behavior, and regulatory environment. These differences may materially affect the relative performance of forecasting models, though systematic cross-market evidence remains scarce. Second, researchers often examine a

single class of models in isolation, such as comparing deep learning methods to traditional statistical benchmarks, without incorporating large pretrained models into the analysis. As a result, practitioners lack comprehensive guidance on how different generations of forecasting models compare under a unified empirical framework. Third, researchers typically focus on limited markets or model classes and often evaluate performance using static accuracy metrics without assessing cross-market stability or robustness.

In this paper, we evaluate whether large pretrained time-series models improve financial forecasting performance relative to classical statistical (ARIMA), machine learning (XGBoost), and deep learning (LSTM) approaches in a unified framework. XGBoostTimesFM Using a large-scale dataset comprising over 10,000 financial assets across equities, bonds, and currencies in multiple developed markets, we evaluate model performance from accuracy and robustness perspectives. By conducting cross-country, cross-asset, and sector-level analyses, we seek to identify systematic patterns in model performance and to clarify the conditions under which specific forecasting approaches are most effective.

The analysis is guided by three central research questions: (i) Do large pretrained models deliver more stable and consistent forecasts across markets? (ii) How does model performance vary across asset classes and market structures? (iii) Are performance differences economically meaningful, or primarily statistical?

This paper makes three contributions. First, it provides a large-scale empirical comparison of forecasting models across more than 10,000 financial assets spanning multiple markets and asset classes. Second, it emphasizes cross-market robustness, rather than pointwise accuracy, as a key criterion for model evaluation. Third, it offers new evidence on the role of large pretrained models in financial forecasting, showing that their advantage lies primarily in stability rather than uniformly superior accuracy. From a practical perspective, the results highlight that improvements in statistical accuracy do not necessarily translate into economically meaningful gains, reinforcing the need for careful evaluation of predictive models in financial applications.

The remainder of the paper is organized as follows: In Section 2, we review the related literature on financial time-series forecasting, covering statistical models, machine learning methods, deep learning approaches, and large pretrained time-series models. In Section 3, we describe the data, modeling framework, and empirical methodology. In Section 4, we present the empirical results, including cross-market robustness analysis and RMSE-based performance comparisons across asset classes and industries. In Section 5, we discuss the economic interpretation of the findings and their implications for financial institutions. In Section 6, we outline the limitations of the study and directions for future research. In Section 7, we conclude the paper.

2. Literature review

2.1. Traditional statistical models in financial time-series forecasting

Statistical models form the cornerstone of early research in financial time-series forecasting and continue to serve as essential benchmarks in empirical studies. Among these, the autoregressive integrated moving average (ARIMA) model, introduced by Box and Jenkins (1976), remains one of the most widely used approaches. ARIMA models characterize time-series dynamics through linear combinations of lagged values and past forecast errors, making them particularly effective for

stationary series with relatively stable temporal patterns. Despite their popularity, the limitations of ARIMA models in financial contexts have been well documented. Financial time series frequently exhibit nonlinearity, volatility clustering, and abrupt structural breaks, which violate the linearity and stationarity assumptions underpinning ARIMA models (Tsay, 2005; Box et al., 2015). Hyndman and Athanasopoulos (2018) emphasized that while ARIMA can perform competitively in short-horizon forecasts, its predictive accuracy deteriorates when the underlying data-generating process evolves over time. Moreover, ARIMA models are sensitive to outliers and extreme events, which are common in financial markets and may lead to biased parameter estimates and unstable forecasts.

To address time-varying volatility, Engle (1982) introduced the autoregressive conditional heteroskedasticity (ARCH) model, which was later generalized into the GARCH framework (Engle, 2001). GARCH-type models explicitly model conditional variance dynamics and have become standard tools for volatility forecasting and risk management. These models capture key stylized facts of financial time series, such as volatility clustering and persistence, and are widely used in asset pricing and risk assessment.

Collectively, traditional statistical models provide valuable theoretical insights and interpretable forecasting frameworks. However, their linear structure and reliance on stationarity limit their effectiveness in capturing higher-order complex and evolving patterns observed in large cross-asset datasets.

2.2. Machine learning methods for financial forecasting

Machine learning approaches offer a flexible and nonparametric alternative to traditional statistical models by relaxing strong distributional and functional assumptions. Early applications of machine learning in finance employed artificial neural networks (ANNs) to model nonlinear relationships between predictors and asset prices. Studies by Di Persio and Honchar (2016), Guresen et al. (2011), and Kim and Shin (2007) demosntrated that ANNs can outperform linear models in certain stock market prediction tasks by capturing complex, nonlinear patterns in historical data. Sureshkumar and Elango (2012) further evaluated the effectiveness of ANN-based systems in stock price prediction and highlight their ability to adapt through self-learning mechanisms.

Beyond neural networks, decision tree-based models have also been widely applied in financial forecasting. Kamble (2017) showed that decision trees can effectively model short-term and long-term stock trends by learning hierarchical decision rules from historical data. However, single-tree models are prone to overfitting and instability, particularly in noisy financial environments. To mitigate these issues, ensemble methods such as gradient boosting have been developed.

XGBoost, a scalable and regularized implementation of gradient boosting decision trees, has emerged as one of the most influential machine learning algorithms (Ke & Guestrin, 2016). XGBoost incorporates regularization terms, parallel computing, and efficient handling of missing data, resulting in strong predictive performance and computational efficiency. Empirical studies by Basak et al. (2019) and Nobre and Neves (2019) showed that XGBoost performs well in predicting stock price movements and trading strategies. Its ability to handle high-dimensional data and nonlinear relationships makes it particularly attractive for financial applications involving complex feature sets. Researchers also explore feature selection guided by large models to improve generalization in financial forecasting tasks (Mutian et al., 2025).

Support vector machines (SVMs) represent another class of widely used machine learning models in finance. Cao and Tay (2001) demonstrated that SVMs can effectively predict prices of financial products such as stocks and bonds by maximizing the margin between classes in a high-dimensional feature space. Hansen et al. (2010) further contributed to the evaluation of predictive models by proposing the Model Confidence Set framework, which enables statistically robust comparison among competing forecasting methods.

Despite their empirical success, machine learning models face inherent challenges when applied to time-series data. Most machine learning algorithms, including XGBoost and SVMs, do not explicitly account for temporal ordering and long-term dependencies. As a result, their performance relies heavily on feature engineering techniques, such as lagged variables and rolling-window statistics. Lim & Zohren (2021) argued that while machine learning methods can approximate complex nonlinear mappings, their limited ability to natively model sequential dependence constrains their effectiveness in forecasting problems with strong temporal structure. Moreover, machine learning models are highly sensitive to validation design. In financial time series, improper handling of temporal dependence may lead to overfitting and data leakage, resulting in overly optimistic performance estimates.

2.3. Deep learning models and sequential forecasting

The development of deep learning has significantly advanced the modeling of sequential and temporal data. Recurrent neural networks (RNNs) were designed to process sequential inputs by maintaining hidden states that evolve over time. However, standard RNNs suffer from vanishing and exploding gradient problems when applied to long sequences, limiting their ability to capture long-term dependencies.

To overcome these challenges, Hochreiter and Schmidhuber (1997) introduced the long short-term memory (LSTM) architecture, which employs gating mechanisms to regulate information flow across time. LSTM models have demonstrated strong performance in a wide range of sequential tasks (see Graves et al., 2013, Sutskever et al., 2014). In financial forecasting, LSTM models have been shown to outperform traditional machine learning and statistical models in predicting stock prices and market trends (Fischer & Krauss, 2018; Nelson et al., 2017).

Several studies highlight the advantages of LSTM models in capturing long-term dependencies and nonlinear dynamics inherent in financial time series. Selvin et al. (2017) demonstrated that LSTM and CNN-based sliding window models can effectively extract local patterns and trends in stock price data. Akita et al. (2016) combined numerical and textual information within a deep learning framework to improve stock prediction accuracy. Similarly, Rather et al. (2015) and Yoshihara et al. (2014) applied recurrent neural networks to predict stock returns and market trends, further illustrating the potential of deep sequential models.

However, the increased representational power of LSTM models comes at the cost of higher computational complexity and training time. Greff et al. (2017) provided a comprehensive analysis of LSTM architectures and noted that their large number of parameters can lead to substantial resource consumption, particularly when applied to large-scale datasets. Yu et al. (2019) emphasized that the computational demands of deep recurrent models may limit their practical applicability in real-time financial decision-making environments.

2.4. Transformer architectures and large pretrained time-series models

Advances in transformer-based architectures have significantly reshaped time-series forecasting by enabling the modeling of long-range dependencies and complex temporal interactions (Vaswani et al., 2017). Transformers offer superior scalability and the ability to capture long-range dependencies efficiently, making them particularly well suited for large-scale time-series forecasting tasks. Studies have shown that transformer-based architectures, including Informer-type models, significantly improve forecasting performance through multi-scale temporal attention mechanisms. Li et al. (2020) proposed the Temporal Fusion Transformer (TFT), which integrates static and dynamic features and provides interpretability through attention mechanisms. Empirical results show that TFT significantly improves forecasting accuracy across a range of time-series applications.

Building on Transformer architectures, large pretrained time-series models such as TimeGPT and TimesFM represent a new paradigm in forecasting. These models adopt large-scale pretraining strategies, learning generic temporal representations from massive collections of heterogeneous time series. TimeGPT leverages the Transformer framework to capture long-term dependencies and complex temporal patterns. TimeGPT (Garza & Mergenthaler-Canseco, 2024) is a pretrained transformer-based model tailored for time-series data, leveraging large-scale pretraining on heterogeneous temporal datasets to learn transferable representations for forecasting tasks. TimesFM further enhances performance through frequency modulation and multi-scale modeling (Oreshkin et al., 2020; Lim & Zohren, 2021).

Makridakis et al. (2022) provided comparative evidence suggesting that large-scale deep learning models can outperform traditional statistical and machine learning approaches in forecasting competitions, particularly when trained on extensive datasets. In finance, these models hold promise for cross-market and cross-asset forecasting, where data-generating processes vary substantially across regions and products. Nevertheless, their application in financial markets remains in an exploratory stage, as highlighted by concerns regarding interpretability, stability, and regulatory compliance.

Since the completion of this study, several additional pretrained time-series foundation models have been proposed, including Chronos, Lag-Llama (Rasul et al., 2024), and Moirai (Woo et al., 2024). These models further extend the paradigm of large-scale pretraining for time-series forecasting, exploring alternative architectures and training objectives. While they are not included in the empirical analysis here, they represent an important direction for future research and may provide additional benchmarks for evaluating cross-market generalization.

2.5. LLM augmented financial forecasting

Beyond transformer architectures, researchers have explored the integration of large language models (LLMs) into financial forecasting. These approaches extend traditional time-series modeling by incorporating textual information, sentiment signals, and high-level semantic representations. For example, FinSentGPT demonstrates the effectiveness of fine-tuned language models in capturing financial sentiment signals relevant for prediction (Ardekani et al., 2024). Similarly, LLM-based sentiment and feature generation frameworks have been shown to enhance stock price prediction by enriching input representations (Chen & Kawashima, 2024; Sun et al., 2025).

Additional studies propose hybrid architectures that combine LLM-generated signals with transformer or deep learning models, improving predictive performance through multi-modal

integration (Zhou et al., 2025; Hussain et al., 2026). Other work emphasizes the role of LLM-guided feature selection and numerical priors in improving generalization and interpretability (Guo et al., 2025; Mutian et al., 2025). These developments suggest that large pretrained models may contribute not only through sequence modeling, but also through representation learning and feature engineering.

Despite these advances, the application of LLM-based models in financial forecasting remains at an early stage, with open questions regarding robustness, stability, and their effectiveness in purely numerical time-series settings.

In contrast to LLM-augmented approaches that rely on textual or hybrid inputs, we focus on purely numerical time-series forecasting, providing a clean comparison across model classes.

2.6. Synthesis and research gaps

The literature demonstrates substantial progress in financial time-series forecasting, though several gaps remain. First, studies rarely provide unified comparisons across statistical, machine learning, deep learning, and large pretrained models. Second, many studies lack rigorous validation frameworks tailored to financial time series, raising concerns about overfitting and generalization. Third, the distinction between statistical accuracy and economically meaningful performance is often not explicitly addressed. In this study, we address these gaps by focusing on cross-market robustness, unified model comparison, and validation-aware evaluation in large-scale financial datasets.

3. Data

We employ a large-scale international financial dataset constructed from publicly available market data obtained from Yahoo Finance. The sample covers more than 10,000 financial instruments across major developed markets, including the United States, Japan, Germany, France, Italy, Spain, the Netherlands, Switzerland, and a composite G20 index. The dataset spans multiple asset classes, equities, bonds, and currencies, enabling a comprehensive cross-market and cross-asset evaluation of forecasting models. Table 1 shows the data points used in each sample class.

The dataset sample prioritizes data availability and continuity rather than investability. No survivorship bias correction is applied, as our objective is to evaluate predictive performance across observed assets rather than construct investable portfolios. Liquidity filtering and transaction cost adjustments are not implemented due to data limitations; however, their implications are discussed in Section 7. The effective sample size varies across assets due to missing observations and market-specific trading calendars, which may influence model estimation and stability.

For equities, the sample includes individual stocks classified into ten industry sectors following standard market conventions: basic materials, consumer goods, consumer services, financial services, healthcare, industrials, energy, utilities, real estate, and business services. Bond data include government and high-grade corporate bond indices, while currency data consist of major bilateral exchange rates against the U.S. dollar. All price series are sampled at a daily frequency and synchronized across markets to ensure comparability.

Following standard practice in empirical finance, all price series are transformed into continuously compounded log-returns. To mitigate the influence of outliers and market microstructure noise, extreme observations beyond the 1st and 99th percentiles are winsorized. Missing observations are handled using forward filling for short gaps and excluded otherwise. The final dataset balances

data availability with cross-market consistency, ensuring sufficient observations for model training and evaluation across regions. Table 2 shows the descriptive statistics of daily returns of the pooled sample.

The forecasting task focuses on one-step-ahead return prediction. Each model is trained using a rolling-window scheme to mimic real-time forecasting conditions and to account for potential structural changes over time. Model performance is evaluated using accuracy-based metrics, such as root mean squared error (RMSE), and robustness indicators that measure stability across markets and asset classes.

To ensure consistency with real-time forecasting settings, all data are aligned in chronological order without any forward-looking adjustments. We do not apply survivorship bias corrections, as our focus is on broad cross-sectional predictability rather than investable portfolios.

Table 1. Sample composition by region and asset class.

Region / Market	Equities	Bonds	Currencies	Total Assets	Frequency	Period
United States	2,450	420	8	2,888	Daily	2010–2023
Japan	2,100	380	15	2,495	Daily	2010–2023
Germany	1,050	210	12	1,272	Daily	2010–2023
France	980	190	10	1,180	Daily	2010–2023
Other Europe	1,850	320	20	2,190	Daily	2010–2023
Total	8,430	1,520	75	10,025		

Notes: The dataset consists of daily closing prices obtained from Yahoo Finance. Returns are computed as log differences. Europe includes major developed markets. Asset counts refer to unique instruments after filtering for data availability.

Table 2. Descriptive statistics of daily returns (pooled sample).

Asset Class	Mean	Median	Std. Dev. (%)	Skewness	Kurtosis	Min	Max	Asset Number
Equities	0.042	0.038	1.38	−0.61	6.21	−0.087	0.079	8430
Bonds	0.018	0.015	0.54	−0.12	3.89	−0.033	0.031	1520
Currencies	0.009	0.008	0.67	0.08	4.12	−0.028	0.027	75
Japan Eq.	0.047	0.044	1.21	−0.48	5.76	−0.125	0.085	2100
U.S. Eq.	0.039	0.037	1.45	−0.72	6.88	−0.092	0.105	2450

Notes: Returns are expressed in percentage terms. The median is reported alongside the mean to account for non-normality and fat tails in financial return distributions.

All data used in this study are obtained from Yahoo Finance and are publicly accessible. The dataset consists of daily closing prices for equities, bond indices, and exchange rates, as described in Table 1. Asset selection follows standard market indices and commonly used tickers within each region.

The data preprocessing pipeline is fully deterministic and can be summarized as follows. First, price series are converted to log-returns. Second, extreme observations are winsorized at the 1st and 99th percentiles on a per-series basis to mitigate the impact of outliers. Third, missing values are handled using forward filling for short gaps, while longer gaps lead to exclusion of the corresponding asset from the sample.

Model training follows a rolling-window design with a fixed 3-year estimation window and one-step-ahead forecasting horizon. All models are implemented using standard libraries and widely used default configurations, with hyperparameters fixed across assets to avoid data snooping.

While the full implementation code cannot be distributed, the above description ensures that the empirical design is transparent and can be replicated using publicly available data and standard modeling frameworks.

4. Methodology

The empirical design combines rolling-window validation, fixed hyperparameters, and cross-asset evaluation to ensure comparability across models. The framework is designed to reflect realistic forecasting conditions while minimizing overfitting and data leakage.

We adopt a rolling out-of-sample evaluation framework designed to mimic real-time forecasting. Specifically, models are trained using historical data up to time t, and predictions are generated for time $t+1$. The training window is updated sequentially, ensuring that no future information is used at any point. The empirical implementation uses a rolling window with a 3-year training period and 1-step-ahead forecasting, re-estimated at each time step. This design mitigates look-ahead bias and provides a more realistic assessment of model performance compared to static train-test splits.

To mitigate overfitting risk, several safeguards are implemented. First, all models are evaluated strictly out-of-sample using a rolling-window framework, ensuring no forward-looking information is used. Second, hyperparameters are fixed across assets and not tuned individually, preventing implicit data snooping. Third, large pretrained models (TimeGPT and TimesFM) are applied in zero-shot settings without task-specific retraining, avoiding over-optimization on financial data.

Together, these design choices ensure that performance differences reflect generalization ability rather than overfitting.

4.1. Price and return process

Let $\left\{P_t^{(i,j)}\right\}_{t=1}^{T}$ denote the observed price process of asset $i \in \mathcal{A}$ in region $j \in \mathcal{R}$, where $\mathcal{A}$ represents the set of financial assets (stocks, bonds, currencies) and $\mathcal{R}$ denotes the set of countries or regions. Following standard practice in quantitative finance, we work with log-returns:

$$r_t^{(i,j)} = \log P_t^{(i,j)} - \log P_{t-1}^{(i,j)}. \tag{1}$$

The return process is assumed to admit the decomposition.

$$r_t^{(i,j)} = \mu_t^{(i,j)} + \varepsilon_t^{(i,j)}, \tag{2}$$

where $\mu_t^{(i,j)} = \mathbb{E}[r_t^{(i,j)} \mid \mathcal{F}_{t-1}]$ is the conditional mean and $\varepsilon_t^{(i,j)}$ is an innovation, satisfying.

$$\mathbb{E}[\varepsilon_t^{(i,j)} \mid \mathcal{F}_{t-1}] = 0, \mathbb{E}[(\varepsilon_t^{(i,j)})^2 \mid \mathcal{F}_{t-1}] = \sigma_t^{2,(i,j)}. \tag{3}$$

The objective of all forecasting models considered in this study is to approximate the conditional expectation.

$$\hat{r}_{t+h|t}^{(i,j)} \approx \mathbb{E}[r_{t+h}^{(i,j)} \mid \mathcal{F}_t], \tag{4}$$

for forecast horizon $h \geq 1$.

To account for cross-market dependence, we model returns jointly as a panel stochastic process. Panel return vector:

$$\mathrm{r}_t = (r_t^{(1)}, \ldots, r_t^{(N)})^\top \tag{5}$$

Conditional mean vector:

$$\mu_t = \mathbb{E}[\mathrm{r}_t \mid \mathcal{F}_{t-1}] \tag{6}$$

Conditional covariance matrix:

$$\Sigma_t = \mathbb{E}[(\mathrm{r}_t - \mu_t)(\mathrm{r}_t - \mu_t)^\top \mid \mathcal{F}_{t-1}] \tag{7}$$

Correlation decomposition:

$$\Sigma_t = \mathrm{D}_t \mathrm{R}_t \mathrm{D}_t \tag{8}$$

Volatility scaling:

$$\mathrm{D}_t = \mathrm{diag}(\sigma_{1t}, \ldots, \sigma_{Nt}) \tag{9}$$

The following subsections describe the models we study in this research.

4.2. Predictive models

In this research, we mainly compare the performance of following five models. The mathematical mechanics for these models are detailed in the Appendix.

ARIMA is a classic time series prediction model that builds an autoregressive and moving average model after differential and stationary time series data.

XGBoost is a machine learning model based on the gradient improvement framework, with high efficiency and strong scalability, and can process large-scale data and perform well in a variety of prediction tasks.

LSTM is a recurrent neural network specializing in processing sequence data with long-term dependencies and is widely used in fields such as time series prediction.

TimesFM is a pretrained decoder-only Transformer model for time-series forecasting (Das et al., 2024). It uses a patch-based tokenization scheme to process long temporal sequences efficiently and is trained on a large corpus of heterogeneous time-series data. In this study, TimesFM is applied in a zero-shot setting without task-specific fine-tuning.

TimeGPT is another large model based on the GPT architecture specifically for time series data, using the ability of large-scale pre-trained language models to predict and analyze time series.

To ensure comparability across models, hyperparameters are fixed across assets and follow standard configurations commonly used in the literature, ensuring consistency and avoiding implicit

data snooping. This avoids implicit data snooping and reflects realistic deployment constraints. Large pretrained models (TimeGPT and TimesFM) are used without task-specific fine-tuning, consistent with their intended zero-shot or few-shot application.

4.3. Evaluation metrics

In the prediction analysis of financial market, choosing the right model is crucial to accurately grasp the market dynamics. Root mean square error (RMSE), as a key indicator to measure the accuracy of the model prediction, can directly reflect the deviation degree between the predicted value and the actual value of the model. Next, a detailed analysis of the three radar maps will explore the performance differences of the five models of LSTM, ARIMA, TimeGPT, TimesFM, and XGBoost in different scenarios.

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_t - \hat{r}_t)^2}. \tag{10}$$

The consistency of using RMSE as a risk measure can be seen by the law of large numbers; if $\{(r_t, \hat{r}_t)\}$ is ergodic, then

$$\frac{1}{T}\sum_{t=1}^{T}(r_t - \hat{r}_t)^2 \xrightarrow{a.s.} \mathbb{E}[(r_t - \hat{r}_t)^2]. \tag{11}$$

Thus, RMSE consistently estimates the L^2 prediction risk.

While RMSE measures pointwise accuracy, it does not capture the cross-sectional and temporal stability of forecasts. In financial applications, where signals are weak and regime-dependent, robustness across assets and time is often as important as raw predictive accuracy. To capture this aspect, we introduce a cross-market robustness index (RI) defined as follows:

$$RI_m = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{|e_{i,t}^{m} - \bar{e}_i| + \epsilon} \tag{12}$$

where $e_{i,t}^{m}$ denotes the forecast error of model m for asset i at time t, and $\bar{e}_i$ is the average error for asset i. The index captures the stability of model performance across assets and time, with higher values indicating more consistent forecasts.

This robustness index depends on a small positive constant ε, introduced to avoid numerical instability when deviations approach zero. In this study, ε is set to a sufficiently small value such that it does not materially affect the ranking of models. The role of ε is purely technical, and the results are not sensitive to reasonable variations in its magnitude. Conceptually, the proposed robustness index is related to dispersion-based measures of forecast stability, such as the coefficient of variation of RMSE across assets or the variance of forecast errors over time. Compared to these alternatives, the RI emphasizes consistency across the cross-sectional and temporal dimensions, providing a unified measure of stability in large-scale panel settings. While alternative stability metrics could be considered, the qualitative conclusions of this study are unlikely to depend on the specific choice of robustness measure, as the observed differences across models are systematic and persistent across markets. Therefore, this metric complements RMSE by emphasizing robustness rather than isolated

accuracy improvements. While formal statistical tests such as the Diebold–Mariano test are commonly used in forecast comparison, we focus on descriptive differences in RMSE and robustness due to the large cross-sectional setting and the relatively small magnitude of differences. The interpretation therefore emphasizes economic relevance rather than statistical significance. Although RMSE and robustness metrics quantify statistical accuracy, their economic interpretation requires caution. In financial markets, small improvements in prediction error may not translate into economically meaningful gains once transaction costs, turnover, and implementation constraints are considered. Therefore, the results are interpreted as evidence on statistical predictability rather than direct investment profitability.

5. Empirical results and discussion

In this section, we present the out-of-sample forecasting results under the rolling evaluation framework. We focus on two dimensions: (i) cross-market robustness, and (ii) predictive accuracy across asset classes. Particular emphasis is placed on interpreting performance differences in relation to market structure and model characteristics.

5.1. Cross-market robustness analysis

We evaluate the performance of the five models of ARIMA, XGBoost, LSTM, TimeGPT, and TimesFM in different regions and financial assets. Table 3 summarize the corresponding RI averaged across regions, and Figures 1–4 present heat maps of model performance across regions and asset classes, providing a visual comparison of cross-market robustness under the rolling out-of-sample evaluation framework. Darker colors indicate higher robustness (RI), corresponding to more stable forecasting performance.

Several key patterns emerge from the analysis. First, large pretrained models, particularly TimesFM, consistently achieve the highest robustness across regions, indicating more stable forecasting performance across assets and time. As shown in Table 4, TimesFM exhibits the highest robustness across all regions, with a substantial margin relative to competing models (Δ vs Best = 45–109). The confidence intervals further indicate that these differences are economically and statistically meaningful. This suggests that pretrained architectures are effective in capturing transferable temporal structures across heterogeneous financial environments.

Second, model performance varies substantially across regions. In particular, Japan exhibits significantly higher robustness across all model classes compared to the United States and Europe. This pattern is visible in Figures 1–4 and is reflected in the higher RI values reported in Table 4. Such differences highlight the importance of cross-market heterogeneity in financial forecasting and suggest that model performance is strongly influenced by underlying market characteristics.

Third, despite the overall advantage of pretrained models, no single model dominates uniformly across all regions and asset classes. Machine learning and deep learning models such as XGBoost and LSTM remain competitive in several cases, while traditional models such as ARIMA continue to perform reasonably well in certain settings. This reinforces the view that forecasting performance is context-dependent rather than universally determined by model complexity.

The observed regional variation can be interpreted in light of differences in market structure. The relatively strong performance across models in Japan may reflect more stable macroeconomic

conditions and sector concentration, which give rise to more persistent and predictable temporal patterns. In contrast, the lower and more dispersed robustness observed in U.S. and European markets may be associated with higher informational efficiency, greater sectoral diversity, and increased exposure to global shocks, all of which reduce the persistence of exploitable signals.

Overall, the results suggest that large pretrained models improve robustness primarily through enhanced stability rather than uniformly superior accuracy. However, their relative advantage depends critically on market structure, emphasizing the need to account for cross-market heterogeneity when evaluating forecasting models.

Table 3. Average cross-market robustness index by model.

Model	Japan	Europe	US	Overall	Δ vs Best	95% CI (Overall)	Std. Error	Rank
ARIMA	160.2	48.3	52.1	86.9	84.3	[77.5, 96.3]	4.8	4
XGBoost	220.5	76.1	81.4	126.0	45.2	[115.8, 136.2]	5.2	2
LSTM	214.3	69.5	74.2	119.3	51.9	[109.5, 129.1]	5.0	3
TimeGPT	98.4	42.1	46.5	62.3	108.9	[54.6, 70.0]	3.9	5
TimesFM	309.8	105.2	98.6	171.2	0.0	[159.2, 183.2]	6.1	1

Notes: The robustness index (RI) measures the stability of predictive performance across assets and time. Higher values indicate more consistent forecasting performance. "Overall" is computed as the simple average across regions."Δ vs Best" denotes the difference relative to the best-performing model (TimesFM). 95% confidence intervals are constructed using ±1.96 × standard error. Standard errors are computed across asset-level observations. Rank is based on the overall RI.

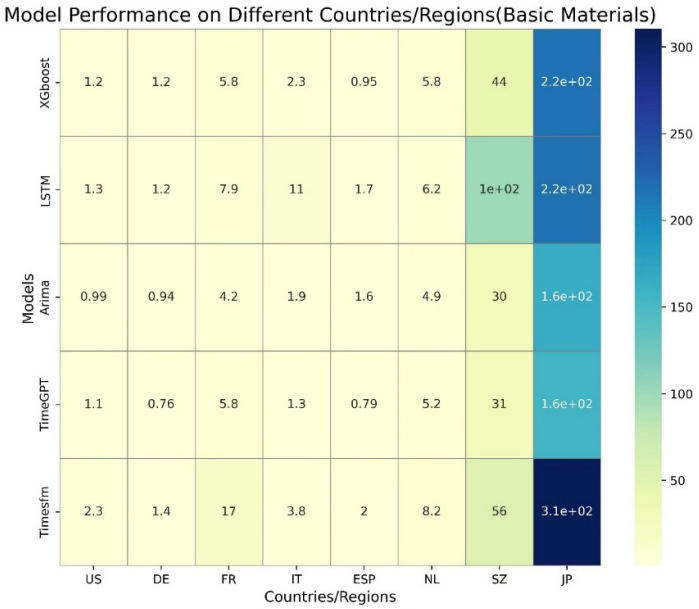


Figure 1. Cross-market robustness (RI) for basic materials sector across countries. Notes: Results are based on rolling out-of-sample forecasts. Higher RI indicates greater stability across assets and time.

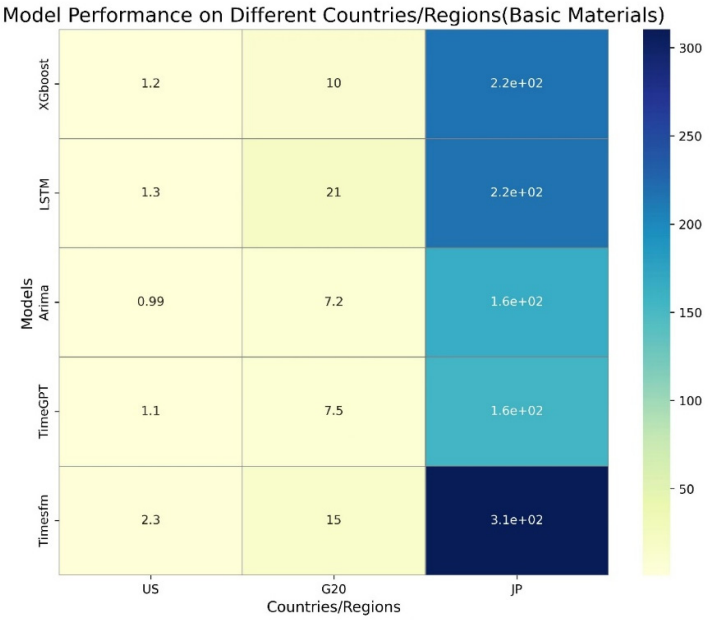


Figure 2. Cross-market robustness (RI) for basic materials sector across broader regions. Notes: Results are based on rolling out-of-sample forecasts. Higher RI indicates greater stability across assets and time.

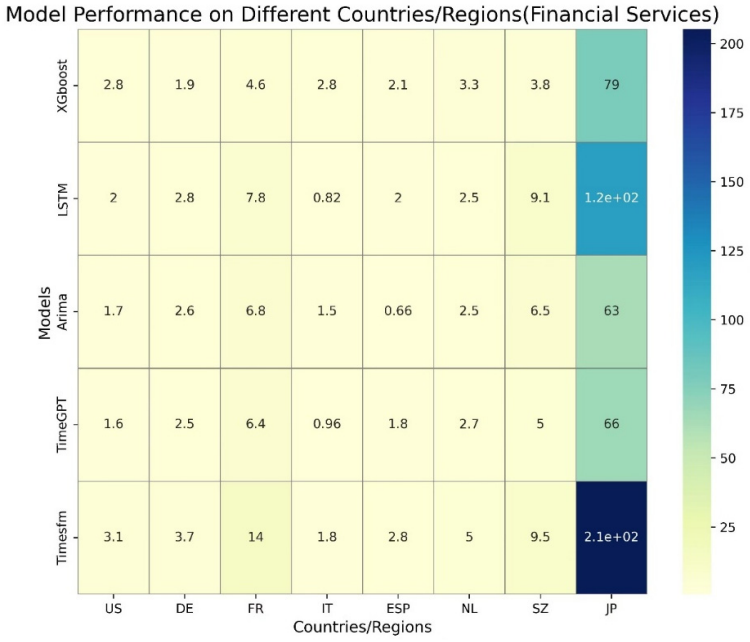


Figure 3. Cross-market robustness (RI) for financial services sector across countries. Notes: Results are based on rolling out-of-sample forecasts. Higher RI indicates greater stability across assets and time.

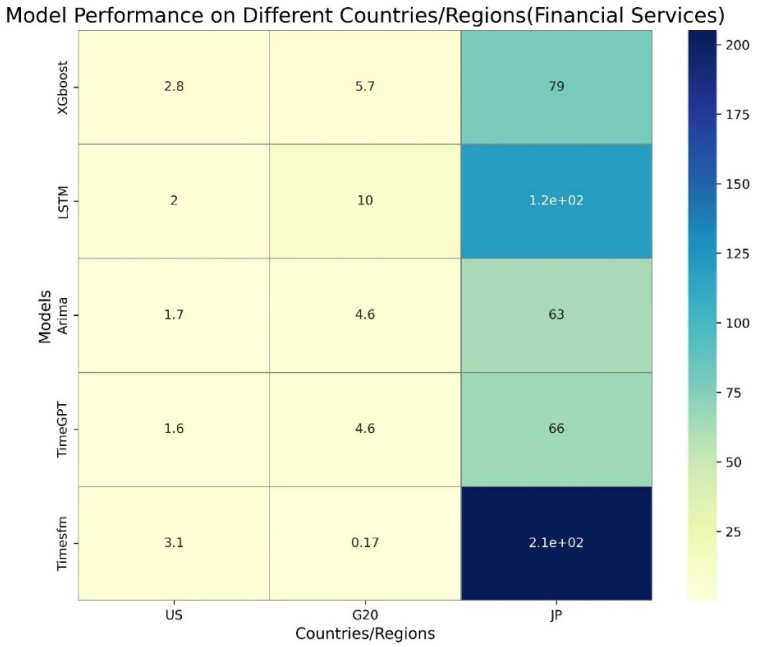


Figure 4. Cross-market robustness (RI) for financial services sector across broader regions. Notes: Results are based on rolling out-of-sample forecasts. Higher RI indicates greater stability across assets and time.

5.2. Predictive accuracy across asset classes

Table 5 reports the out-of-sample forecasting accuracy, measured by RMSE, across major asset classes, including equities, bonds, and currencies. Figures 5 and 6 further extend this analysis by examining RMSE across industry sectors in Germany and the United States, respectively, providing a more granular view of model performance within equity markets.

Several key findings emerge. First, predictive performance varies substantially across asset classes. TimesFM achieves the lowest RMSE in equities, indicating relatively stronger performance in capturing complex dynamics in stock markets. In contrast, XGBoost performs best in bonds and currencies, suggesting that tree-based models remain highly competitive in markets characterized by smoother dynamics and stronger macroeconomic structure.

Second, no single model consistently outperforms others across all asset classes. While TimesFM leads in equities, its performance deteriorates in bonds and currencies, where simpler models such as XGBoost and ARIMA remain competitive. This lack of uniform dominance highlights the context-dependent nature of financial forecasting and suggests that model effectiveness is closely tied to asset-specific characteristics. This heterogeneity becomes even more pronounced at the sector level. As shown in Figures 5 and 6, model rankings vary substantially across industries, with no model maintaining consistently superior performance across all sectors within either Germany or the United States.

Third, the magnitude of performance differences across models is relatively modest. Although certain models outperform others in specific asset classes, the overall variation in RMSE is limited.

This indicates that improvements in statistical accuracy are incremental rather than transformative, and that model selection should not rely solely on marginal gains in pointwise prediction accuracy.

The observed patterns can be interpreted in light of differences in market structure. Equity markets typically exhibit higher levels of nonlinearity, cross-sectional heterogeneity, and regime dependence, which may favor flexible models such as deep learning and large pretrained architectures. In contrast, bond and currency markets are often more influenced by macroeconomic factors and exhibit smoother temporal dynamics, enabling models such as XGBoost to perform effectively without requiring complex representation learning. At the sector level, this variation may reflect differences in industry-specific dynamics, such as sensitivity to macroeconomic conditions, cyclicality, and firm-level heterogeneity, which affect the predictability of returns and the relative effectiveness of different modeling approaches.

Overall, the results suggest that predictive accuracy in financial markets is inherently conditional, and that no single modeling approach can be expected to deliver superior performance across all asset classes. This reinforces the importance of adopting a context-specific approach to model selection in practical forecasting applications.

Table 4 summarizes the RMSE by Asset class and model.

Table 4. RMSE by asset class under rolling out-of-sample forecasting.

Model	Equities	Δ vs Best	Bonds	Δ vs Best	Currencies	Δ vs Best
ARIMA	21.4	4.1	1.98	0.16	2.06	0.19
XGBoost	19.6	2.3	1.82	0.00	1.87	0.00
LSTM	18.9	1.6	2.04	0.32	3.11	1.24
TimeGPT	20.7	3.4	1.96	0.24	2.02	0.15
TimesFM	17.3	0.0	2.09	0.37	2.14	0.27

Notes: RMSE values are computed from rolling out-of-sample forecasts. Lower values indicate higher predictive accuracy."Δ vs Best" denotes the difference relative to the best-performing model within each asset class. Differences in RMSE are reported descriptively without formal statistical significance testing, consistent with the focus on economic magnitude and cross-sectional robustness.

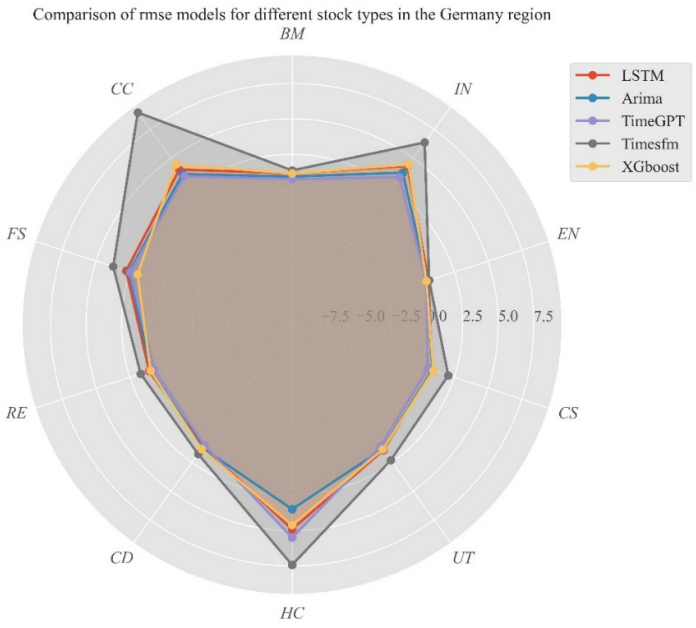


Figure 5. Sector-level RMSE comparison across models (Germany). Notes: Lower RMSE indicates higher predictive accuracy.

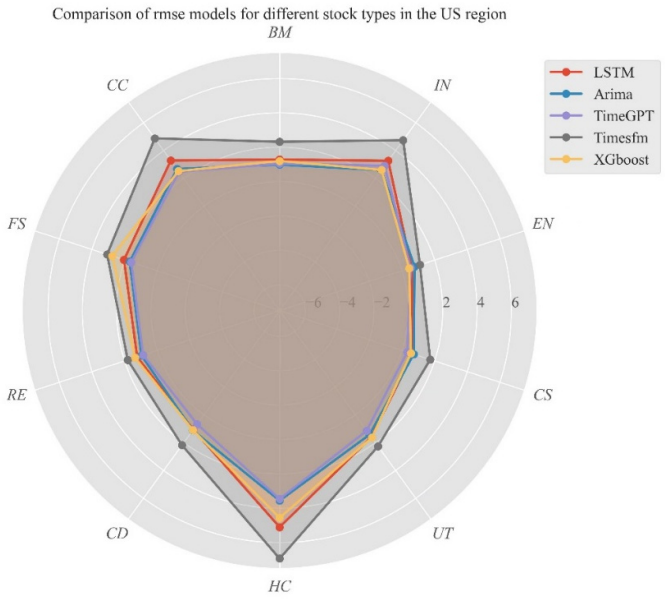


Figure 6. Sector-level RMSE comparison across models (United States). Notes: Lower RMSE indicates higher predictive accuracy.

5.3. Interpretation

The empirical findings highlight pronounced heterogeneity in forecasting performance across markets and asset classes. As shown in Figures 1–4 and summarized in Table 4, large pretrained models, particularly TimesFM, tend to achieve higher cross-market robustness, indicating more stable performance across assets and time. However, this advantage is not uniform and varies significantly depending on the underlying market environment.

A notable pattern is the strong performance observed in the Japanese market. Across multiple model classes, robustness levels are consistently higher in Japan compared to the United States and Europe. One possible explanation lies in the structural characteristics of the Japanese market, including sector concentration in manufacturing and basic materials, as well as relatively stable macroeconomic conditions over the sample period. These features may lead to more persistent temporal patterns, which are more effectively captured by models with strong representation learning capabilities, such as large pretrained architectures.

In contrast, forecasting performance in U.S. and European markets appears more fragmented, with no single model consistently dominating across regions or sectors. These markets are typically characterized by greater sectoral diversification, higher exposure to global macroeconomic shocks, and potentially higher informational efficiency. As a result, predictive signals are weaker and less persistent, limiting the extent to which any model can achieve consistently superior performance. This observation is consistent with the broader view that financial predictability is highly context-dependent and varies across market structures.

Importantly, the results also indicate that improvements in statistical accuracy do not necessarily translate into economically meaningful gains. As shown in Table 5, differences in RMSE across models are relatively small, suggesting that the practical value of selecting one model over another may be limited when transaction costs, model complexity, and implementation constraints are taken into account. In this sense, robustness and consistency across markets may be more relevant criteria than isolated improvements in predictive accuracy.

Taken together, the findings emphasize that model selection in financial forecasting should be guided by statistical performance and economic considerations. While large pretrained models offer advantages in terms of stability and cross-market generalization, their benefits are not universal. In practical applications, particularly in multi-asset or cross-region settings, prioritizing robustness over marginal accuracy gains may lead to more reliable and scalable forecasting outcomes.

It is important to emphasize that statistical improvements in RMSE do not necessarily imply economically meaningful gains. In financial forecasting, small differences in prediction error often translate into negligible differences in trading performance once transaction costs, market impact, and estimation risk are taken into account. This observation is consistent with a large body of literature showing that predictive accuracy and economic value are only weakly linked in noisy and low signal-to-noise environments.

Given our focus of this study on large-scale cross-market comparison, the analysis deliberately prioritizes statistical consistency and robustness rather than constructing specific trading strategies. A full economic evaluation would require additional assumptions regarding portfolio construction, transaction costs, and risk management, which are beyond the scope of the current paper. We therefore interpret the results as evidence on statistical predictability rather than direct investment performance.

6. Practical insights

6.1. Cross-market heterogeneity in predictive performance

The empirical results highlight substantial cross-market heterogeneity in forecasting performance. In particular, the relatively strong performance observed in the Japanese market may reflect structural characteristics such as sector concentration and more stable macroeconomic conditions, which facilitate the emergence of persistent temporal patterns, thereby improving forecastability for traditional and data-driven models.

By contrast, predictive performance in European and U.S. markets is more fragmented across models. These markets are characterized by higher exposure to global macroeconomic shocks, greater sectoral diversification, and potentially higher informational efficiency, which may reduce the effectiveness of any single dominant forecasting approach. Taken together, this evidence reinforces the view that forecasting performance is highly market-specific and cautions against assuming universal model superiority across regions. This suggests that forecastability is not an inherent property of models alone but also depends critically on the underlying data-generating process of each market.

6.2. Why large pretrained models exhibit superior robustness

Among the five models examined, TimesFM demonstrates the highest overall robustness across markets and asset classes. Unlike traditional statistical models and task-specific machine learning approaches, large pretrained time-series models benefit from exposure to massive and diverse datasets during pretraining. This enables them to learn generic temporal representations that transfer effectively across markets with different data-generating processes. However, this advantage appears to manifest primarily in terms of stability rather than uniformly superior predictive accuracy.

However, the results also show that TimeGPT does not consistently outperform simpler deep learning or machine learning models, indicating that pretraining alone does not guarantee superior performance in financial forecasting. The effectiveness of pretrained models depends on their alignment with financial data characteristics, including noise levels, regime shifts, and weak signal structures. This highlights that model architecture and domain adaptation may be as important as scale in determining performance.

6.3. Asset-class–specific model dominance

Model performance exhibits clear asset-class dependence, reflecting differences in underlying market dynamics. The analysis reveals differences in model performance across asset classes. XGBoost achieves the lowest RMSE in bond and currency markets, while TimesFM performs best in equity forecasting. One possible explanation is that bond and currency markets are more closely linked to macroeconomic fundamentals and exhibit smoother dynamics, which tree-based ensemble methods can capture effectively through feature engineering. Equity markets, by contrast, display stronger nonlinearities, higher volatility clustering, and sector-specific dynamics, favoring deep learning and large pretrained architectures.

These results challenge the notion of a universally optimal forecasting model and instead reinforce the view that forecasting models should be selected conditionally, rather than assuming the existence of a universally optimal approach.

6.4. Model interpretability and diagnostics

Model interpretability varies substantially across the models considered. For tree-based models, such as XGBoost, feature importance can, in principle, be extracted, although we focus on aggregate predictive performance rather than feature-level analysis. For deep learning and large pretrained models, interpretability is more limited due to model complexity and the absence of transparent feature attribution mechanisms.

Instead of feature-level diagnostics, we evaluate model behavior through cross-market stability and sector-level performance patterns. These provide indirect evidence on model robustness across economic environments. A more detailed interpretability analysis is left for future research.

6.5. Implications for buy-side institutions

The empirical results should be interpreted with an explicit distinction between statistical and economic significance. From a practical perspective, the findings have important implications for buy-side institutions engaged in cross-market and multi-asset allocation. First, large pretrained models such as TimesFM offer substantial robustness advantages, making them suitable as baseline forecasting tools in diversified portfolios. Second, the strong performance of XGBoost in bond and currency markets suggests that computationally efficient machine learning models remain highly relevant, particularly in environments where interpretability and speed are prioritized.

More broadly, the results suggest that hybrid forecasting frameworks, combining large pretrained models with asset-specific or region-specific models, may offer a more effective balance between robustness, accuracy, and operational feasibility than reliance on a single model class. In practice, this implies that model selection should be aligned with the specific asset class and market environment, rather than relying on a single global forecasting model.

7. Limitations and future research

Despite the comprehensive scope of this study, several limitations should be acknowledged, which also point to promising avenues for future research. These considerations are particularly relevant in financial forecasting, where data quality and evaluation design critically affect model performance.

First, the empirical analysis relies on publicly available financial data sourced from Yahoo Finance. While this ensures transparency and reproducibility, such data may be subject to survivorship bias, data revisions, and differences in market microstructure across countries. Researchers could extend the analysis using proprietary or high-frequency datasets to assess whether the relative performance of forecasting models persists under alternative data environments.

Second, although consistent training, validation, and evaluation procedures are applied across all models, differences in model complexity and hyperparameter sensitivity may influence comparative performance. Large pretrained models, in particular, require substantial computational resources and careful tuning, which may not be fully comparable to classical statistical or tree-based machine

learning methods. Researchers could investigate cost-adjusted performance measures that explicitly account for computational efficiency and scalability.

Third, the evaluation framework focuses primarily on statistical accuracy metrics, such as RMSE and robustness measures, rather than economic utility. While prediction accuracy is a necessary condition for successful forecasting, it does not directly translate into investment profitability. This limitation is consistent with our empirical findings that improvements in RMSE are relatively modest across models. Researchers could incorporate portfolio-level performance metrics, transaction costs, and risk-adjusted returns to assess the economic value of competing forecasting models in realistic trading settings.

Fourth, we abstract from issues of interpretability and regulatory compliance, which are particularly relevant for large pretrained models in financial applications. As financial institutions increasingly adopt machine learning–based forecasting tools, understanding the transparency, explainability, and stability of large models becomes essential. Researchers could explore hybrid or interpretable architectures that balance predictive performance with regulatory and risk management requirements.

Moreover, A more granular subperiod analysis (e.g., pre-COVID vs post-COVID) and formal structural break testing would provide additional insights into regime-specific performance. Due to data availability constraints, such analysis is left for future research.

Finally, while the analysis covers multiple developed markets and asset classes, emerging markets and alternative assets remain unexplored. Extending the framework to fewer liquid markets or assets with different data-generating processes would provide further insights into the generalizability of large pretrained time-series models and their limitations in more challenging forecasting environments. This would further test the robustness of pretrained models under more challenging and less structured environments.

8. Conclusions

This study provides a unified empirical evaluation of traditional statistical, machine learning, deep learning, and large pretrained models for financial time-series forecasting across global markets. Using a large-scale dataset and multiple evaluation metrics, we document substantial heterogeneity in predictive performance across markets and asset types.

Our findings show that large pretrained models, particularly TimesFM, exhibit superior robustness across regions, while simpler machine learning approaches such as XGBoost remain highly competitive in specific asset classes, including bonds and currencies. No single model dominates universally, and performance differences are often modest in magnitude, underscoring the importance of conditional model selection in financial forecasting.

By offering systematic cross-market evidence, this study contributes to the literature on financial time-series prediction and provides actionable insights for institutional investors engaged in global asset allocation. As large pretrained models continue to evolve, researchers should focus on bridging statistical forecasting performance with economic value, including portfolio-level evaluation, transaction costs, and real-time decision-making.

Author contributions

Linghan Huang conducted the main research, collected data, analyzed the results, and wrote the initial and revised paper. Linghan Huang is responsible for the accuracy of all data. results (including all figures and tables), analysis, and concludion in this paper.

Yunhan Li also contributed to this research in both the draft and revision stages, including analysis, text writing, references collection, etc.

Dr. Xuan Feng, the corresponding author and supervisor, initiated the research idea, designed the initial research framework, and provided the mathematical formulas for the theoretical models before the empirical work began. He then supervised the project only in a project-management capacity (e.g., communicating with the editors) and did not participate in, or audit, the formal analysis or the manuscript writing carried out after the formal analysis began. Dr. Feng is not responsible for the accuracy of the data, results, empirical analysis, conclusions, references, or text of this paper.

Use of AI tools declaration

AI tools, including ChatGPT 5.6 and 6.0, were only used for language enhancement for this paper.

Acknowledgments

The research was conducted for academic study, and there is NOT any recommendations for investment.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

All the finance assets' data is publicly available from the Yahoo Finance website.

Funding

No funding was received.

References

Akita R, Yoshihara A, Matsubara T, et al. (2016) Deep learning for stock prediction using numerical and textual information. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science*, 1–6. Okayama, Japan. https://doi.org/10.1109/ICIS.2016.7550882

Ardekani AM, Bertz J, Bryce C, et al. (2024) FinSentGPT: A universal financial sentiment engine?. Int Rev Financ Anal 94: 103291. https://doi.org/10.1016/j.irfa.2024.103291

Athanasopoulos G, Hyndman RJ, Kourentzes N, et al. (2017) Forecasting with temporal hierarchies. *European Eur J Oper Res* 262: 60–74. https://doi.org/10.1016/j.ejor.2017.02.046

Basak S, Kar S, Saha S, et al. (2019) Predicting the direction of stock market prices using tree-based classifiers. *North Am J Econ Finance* 47: 552–567. https://doi.org/10.1016/j.najef.2018.06.003

Box GEP, Jenkins GM (1976) *Time series analysis: Forecasting and control*. Holden-Day.

Box GEP, Jenkins GM, Reinsel GC (2015) *Time series analysis: Forecasting and control*. John Wiley & Sons.

Cao L, Tay FEH (2001) Financial forecasting using support vector machines. *Neural Comput Appl* 10: 184–192. https://doi.org/10.1007/s005210170010

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

Chen Q, Kawashima H (2024) Stock price prediction using llm-based sentiment analysis. *In 2024 IEEE International Conference on Big Data (BigData),* 4846–4853. https://doi.org/10.1109/BigData62323.2024.10825946

Das A, Kong W, Leber A, et al. (2024) A decoder-only foundation model for time-series forecasting. *ICML* 235: 10148–10167.

Di Persio L, Honchar O (2016) Artificial neural networks architectures for stock price prediction: Comparisons and applications. *Int J Circuits Syst Signal Process* 10: 403–413.

Engle RF (2001) GARCH 101: The use of ARCH/GARCH models in applied econometrics. *J Econ Perspect* 15: 157–168. https://doi.org/10.1257/jep.15.4.157

Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50: 987–1007. https://doi.org/10.2307/1912773

Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. *Eur J Oper Res* 270: 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Garza A, Mergenthaler-Canseco M (2024) TimeGPT: Foundation models for time series forecasting. arXiv preprint.

Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947

Greff K, Srivastava RK, Koutník J, et al. (2017) LSTM: A search space odyssey. *IEEE Trans Neural Netw Learn Syst* 28: 2222–2232. https://doi.org/10.1109/TNNLS.2016.2582924

Guresen E, Kayakutlu G, Daim TU (2011) Using artificial neural network models in stock market index prediction. *Expert Syst Appl* 38: 10389–10397. https://doi.org/10.1016/j.eswa.2011.02.068

Guo H, Kwok PY, Guo Y, et al. (2025) FinLSPM: Large stock predict model via numerical prior knowledge from LLM. *Expert Syst Appl*, 130294. https://doi.org/10.1016/j.eswa.2025.130294

Hansen PR, Lunde A, Nason JM (2010) The model confidence set. *Econometrica* 78: 453–497. https://doi.org/10.3982/ECTA5771 https://doi.org/10.3982/ECTA5771

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

Hussain SA, Qiu-Shi C, Hussain SA, et al. (2026) Improving financial forecasting with a synergistic LLM-transformer architecture: A hybrid approach to stock price prediction. *arXiv preprint arXiv:2601.02878*. https://doi.org/10.62762/TIS.2025.976754

Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and practice*. OTexts.

Kamble RA (2017) Short and long term stock trend prediction using decision tree. In *2017 International Conference on Intelligent Computing and Control Systems*, 1371–1375. https://doi.org/10.1109/ICCONS.2017.8250694

Kim HJ, Shin KS (2007) A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Appl Soft Comput* 7: 569–576. https://doi.org/10.1016/j.asoc.2006.03.002

Lim B, Zohren S (2021) Time series forecasting with deep learning: A survey. *Philos Trans R Soc A* 379: 20200209. https://doi.org/10.1098/rsta.2020.0209

Li Y, Zhu Z, Kong D, et al. (2020) Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int J Forecast* 37: 1748–1764. https://doi.org/10.1016/j.ijforecast.2020.07.006

Malkiel BG, Fama EF (1970) Efficient capital markets: A review of theory and empirical work. *J Finance* 25: 383–417. https://doi.org/10.2307/2325486

Makridakis S, Spiliotis E, Assimiakopoulos V (2022) Statistical, machine learning and deep learning forecasting methods: Comparisons and ways forward. *J Oper Res Soc* 73: 840–859. https://doi.org/10.1080/01605682.2022.2118629

Mutian O, Thomas JJ, Tianzhou Y, et al. (2025) LLM-guided semantic feature selection for interpretable financial market forecasting in low-resource financial markets. *Franklin Open*, 100359. https://doi.org/10.1016/j.fraope.2025.100359

Nelson DM, Pereira AC, de Oliveira RA (2017) Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks*, 1419–1426. https://doi.org/10.1109/IJCNN.2017.7966030

Nobre J, Neves RF (2019) Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. *Expert Syst Appl* 125: 181–194. https://doi.org/10.1016/j.eswa.2019.01.076

Oreshkin BN, Carpov D, Chapados N, et al. (2020) N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *International Conference on Learning Representations*.

Rather AM, Agarwal A, Sastry V (2015) Recurrent neural network and a hybrid model for prediction of stock returns. *Expert Syst Appl* 42: 3234–3241. https://doi.org/10.1016/j.eswa.2014.12.003

Rout AK, Dash P, Dash R, et al. (2015) Forecasting financial time series using a low complexity recurrent neural network and evolutionary learning approach. *J King Saud Univ Comput Inf Sci* 29: 536–552. https://doi.org/10.1016/j.jksuci.2015.06.003

Selvin S, Vinayakumar R, Gopalakrishnan EA, et al. (2017) Stock price prediction using LSTM, RNN and CNN-sliding window model. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 1643–1647. https://doi.org/10.1109/ICACCI.2017.8126078

Shiller RJ (2003) From efficient markets theory to behavioral finance. *J Econ Perspect* 17: 83–104. https://doi.org/10.1257/089533003321164967

Sun Y, Gao P, Yan Y, et al. (2025) Stock price prediction with attention-based framework by integrating LLM-generated features. In *International Conference on Neural Information Processing*,562–576.https://doi.org/10.1007/978-981-95-4445-5_38

Sureshkumar KK, Elango NM (2012) Performance analysis of stock price prediction using artificial neural network. *GJCST* 12: 19–25.

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. *NeurIPS* 27: 3104–3112. https://doi.org/10.48550/arXiv.1409.3215

Tsay RS (2005) Analysis of financial time series (2nd ed.). Wiley. https://doi.org/10.1002/0471746193

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In *Advances in Neural Information Processing Systems* 30: 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Yoshihara A, Fujikawa K, Seki K, et al. (2014) Predicting stock market trends by recurrent deep neural networks. In *Proc. Pacific Rim International Conference on Artificial Intelligence*, 759–969. https://doi.org/10.1007/978-3-319-13560-1_60

Yu Y, Si X, Hu C, et al. (2019) A review of recurrent neural networks. *Neural Comput Appl* 31: 1233–1255. https://doi.org/10.1007/s00521-018-3628-5

Zhou L, Zhang Y, Yu J, et al. (2025) Llm-augmented linear transformer–cnn for enhanced stock price prediction. *Mathematics* 13: 487. https://doi.org/10.3390/math13030487



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)