WEP
Warwick
Evans
Publishing

Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 17
WEPM 2026

Out-of-Sample Forecasting of the CSI 300 Index: AR versus Machine Learning Models

Qi Qin *

School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: qq13395197189@163.com

Abstract. This study examines the out-of-sample predictability of daily returns of the CSI 300 Index by comparing a linear autoregressive (AR) benchmark with several machine learning models within a unified rolling forecasting framework. Using one-step-ahead forecasts, we evaluate predictive performance in terms of loss-based measures, directional accuracy, and statistical significance. The results show that machine learning models generally outperform the AR benchmark in terms of mean squared prediction error, with LSTM delivering the strongest gains and XGBoost achieving superior directional accuracy. However, predictive improvements are not uniformly distributed over time. Relative cumulative error analysis and volatility-based subsample tests indicate that machine learning advantages are primarily concentrated in high-volatility periods, while the linear benchmark remains competitive in tranquil markets. These findings provide evidence that return predictability and model performance may vary across market conditions.

Keywords: Return predictability, Machine learning, Out-of-sample forecasting, CSI 300 Index.

1. Introduction

The predictability of asset returns has long been a central topic in financial economics and investment practice. Early studies suggest that stock returns may exhibit time-varying characteristics to some extent (Fama and French, 1988), yet their predictability remains controversial from both statistical and economic perspectives. In particular, under an out-of-sample forecasting framework, many variables that appear predictive in historical data often fail to maintain stable forecasting power in future periods (Welch and Goyal, 2008). Similarly, Campbell and Thompson (2008) show that under rigorous out-of-sample evaluation, predictive models rarely outperform a simple historical mean benchmark on a consistent basis. Even earlier evidence from exchange rate forecasting (Meese and Rogoff, 1983) indicates that structural models with strong in-sample explanatory power often perform poorly out of sample.

Accurately characterizing the dynamics of returns is important not only for asset pricing theory but also for risk management, portfolio allocation, and quantitative trading strategies. However, a large body of research shows that, especially at the daily frequency, return series typically display near-zero means, high noise-to-signal ratios, substantial volatility, and structural instability. These features make out-of-sample forecasting particularly challenging. Even if risk premia vary over time in theory, empirical forecasting models often struggle to consistently outperform simple linear benchmarks in real-time prediction. Therefore, improving return forecasting accuracy within a strict out-of-sample framework remains an important and open question.

Traditional return forecasting methods are typically based on linear time-series models or linear predictive regression frameworks, such as the autoregressive (AR) model and its extensions. These models are transparent and economically interpretable, and they have been widely used in early empirical asset pricing studies. However, when potential nonlinear relationships, structural changes, or high-dimensional information are present, the flexibility of linear models may be limited. In recent years, machine learning methods have been increasingly applied to return prediction. Gu et al. (2020), for example, compare a broad set of machine learning approaches in a large-feature setting and find that nonlinear models outperform traditional linear methods in cross-sectional stock return prediction.

With the development of deep learning, neural network models have also been introduced into financial forecasting. Chen et al. (2024) show that deep neural networks can extract latent factor structures and improve pricing and forecasting performance. Moreover, Bianchi et al. (2021) suggest that the predictive ability of machine learning models may vary across macroeconomic states, implying a degree of state dependence. Although machine learning methods are capable of modeling more complex dynamics in theory, their out-of-sample performance in real financial markets remains uncertain, particularly regarding whether their advantages are stable across different market environments.

Building on the discussion above, this paper focuses on the daily returns of the CSI 300 Index and systematically compares the forecasting performance of traditional linear models and several machine learning approaches within a strict rolling out-of-sample framework. Specifically, we adopt the autoregressive (AR) model as the benchmark and consider support vector regression (SVR), extreme gradient boosting (XGBoost), and long short-term memory (LSTM) networks as competing models. Their predictive performance is evaluated from three dimensions: forecast errors, directional accuracy, and statistical significance.

To better understand how relative advantages evolve over time, we further construct relative cumulative squared forecast error plots. This dynamic analysis allows us to trace the timing and persistence of predictive improvements, rather than relying solely on average loss measures over the full out-of-sample period.

In addition, given that financial markets are often characterized by state-dependent behavior, we examine whether forecasting performance varies across different volatility conditions. Using market volatility as a proxy for uncertainty, we divide the sample into high- and low-volatility subsamples and evaluate model performance separately. Compared with studies that report only full-sample average forecast errors, this regime-based analysis helps identify the specific market environments under which machine learning models tend to perform better, thereby providing a clearer interpretation of the sources of predictive gains.

The main findings can be summarized as follows. First, over the full out-of-sample period, machine learning models generally outperform the linear benchmark in terms of mean squared forecast error. Among them, LSTM achieves the strongest improvement in loss-function metrics and passes statistical significance tests. Second, in terms of directional accuracy, XGBoost performs relatively better and shows stronger ability in predicting market movements. Third, the cumulative error analysis indicates that predictive gains are not evenly distributed over time but are concentrated in specific periods. Finally, the regime-based results suggest that the advantages of machine learning models are primarily observed during high-volatility episodes, while the linear benchmark remains more stable in low-volatility environments. Taken together, these findings suggest that return predictability is state-dependent and that the effectiveness of more complex models may vary with market conditions.

Overall, this study provides new empirical evidence on daily return forecasting in the Chinese stock market and offers a systematic evaluation of machine learning models from both dynamic and regime-based perspectives. By examining the time evolution of predictive performance and its variation across market conditions, the results contribute to a more nuanced understanding of return predictability. The findings may also offer practical insights for model selection in investment decision-making and risk management.

The remainder of the paper is organized as follows. Section 2 reviews the related literature on linear models and machine learning approaches in return forecasting. Section 3 describes the forecasting models, the rolling out-of-sample design, and the evaluation metrics. Section 4 introduces the data and variable construction. Section 5 presents the empirical results, including out-of-sample performance, relative cumulative forecast error analysis, and regime-based comparisons. Section 6 concludes and discusses directions for future research.

2. Related Literature

2.1. Stock Index Forecasting with Linear Models

Linear models have long served as the primary tools in stock index return forecasting. Within the predictive regression framework, researchers typically employ macroeconomic variables, valuation ratios, or technical indicators as explanatory variables and use linear regressions to assess their impact on future returns. Early studies document that variables such as dividend yields, interest rate spreads, and valuation ratios exhibit statistically significant predictive power during certain periods. These studies commonly rely on univariate or multivariate linear predictive regressions to evaluate the marginal contribution of each variable to future returns.

More recent research has shifted attention toward the time variation and robustness of predictive relationships. For example, Rapach et al. (2016) systematically compare combinations of economic predictors in a multivariate framework and find that linear models can generate out-of-sample improvements in specific periods, although such gains are not persistent over time. Dangl and Halling (2012) propose predictive regressions with time-varying parameters and show that return predictability may adjust with changing macroeconomic conditions. Ignoring parameter instability, therefore, may understate the forecasting potential of linear models.

In addition, Pettenuzzo et al.(2014) examine forecast uncertainty within a Bayesian model averaging framework and highlight the role of model uncertainty in predictive regressions. More recently, researchers have sought to improve traditional linear frameworks by incorporating high-dimensional and sparsity considerations. Kozak et al.(2020), for instance, apply structured shrinkage techniques to predictor variables and demonstrate that controlling model complexity within a linear setting can also lead to more robust forecasting performance.

Overall, linear predictive regressions continue to play an important role in stock index return forecasting. However, their predictive performance is often affected by parameter instability, model misspecification, and structural changes in the underlying economic environment. As data dimensionality increases and nonlinear modeling techniques become more accessible, researchers have gradually turned to more flexible approaches in an effort to capture complex dynamic patterns that may not be fully accommodated within traditional linear frameworks.

Against this backdrop, machine learning methods have been increasingly introduced into financial forecasting research.

2.2. Machine Learning in Financial Forecasting

With improvements in computational capacity and the expansion of data availability, financial forecasting research has gradually moved beyond traditional linear frameworks toward more flexible nonlinear modeling approaches. A key advantage of machine learning methods lies in their ability to model complex relationships with fewer functional form restrictions and to handle high-dimensional settings through automatic feature selection and nonlinear interaction modeling.

Early studies began exploring the application of nonlinear methods in financial time series. For example, Cao and Tay (2003) show that support vector machines can exhibit stronger generalization ability than traditional neural networks in financial forecasting tasks. Huang et al. (2005) apply support vector regression to stock market direction prediction and find that it performs reasonably well in nonlinear environments. These studies provide an initial foundation for the subsequent application of machine learning techniques in return forecasting.

In recent years, deep learning methods have attracted increasing attention in financial prediction. Fischer and Krauss (2018) find that long short-term memory (LSTM) networks are able to capture temporal dependence structures and achieve improved out-of-sample performance in stock market direction forecasting. Bao et al. (2017) further show that combining LSTM with multi-scale features can enhance index prediction accuracy. In parallel, ensemble learning methods, particularly gradient

boosting algorithms such as XGBoost, have been increasingly applied in financial forecasting. Krauss et al. (2017) systematically compare several machine learning methods—including gradient-boosted trees, random forests, and deep neural networks—in a statistical arbitrage setting based on constituents of the S&P 500 Index. Their results suggest that gradient boosting models tend to perform relatively robustly out of sample and may outperform neural networks in certain periods. The study also notes that tree-based models are effective in capturing nonlinear relationships and interaction effects, although their predictive gains appear to be time-varying rather than persistent.

In addition, Zhang et al. (2020) show that ensemble learning methods exhibit strong adaptability when handling high-dimensional market features in trading strategy applications, although their performance remains sensitive to changes in market structure. Taken together, the existing evidence suggests that gradient boosting models may offer potential advantages in financial forecasting, but their stability across different sample periods and market conditions requires further examination.

Although many studies document predictive improvements from machine learning methods during certain periods, the overall evidence is not uniform. McLean and Pontiff (2016), in their re-examination of anomaly-based return predictors, find that many documented predictive relationships weaken substantially out of sample, suggesting that sample selection and data mining may play a role. Similarly, Kim and Swanson (2018) show that the apparent advantages of more complex models often depend on specific sample intervals, and their out-of-sample performance is not consistently stable.

More recent research further emphasizes the time-varying and structurally unstable nature of return predictability. Lettau et al. (2020) document pronounced time variation in factor risk premia, leading to substantial differences in model fit across subsamples. Feng et al. (2020) also find that many proposed predictive factors experience significant out-of-sample decay, indicating that forecasting relationships may evolve over time.

Within the machine learning framework, Gu et al. (2020) report that the predictive performance of high-dimensional models fluctuates noticeably in rolling samples. At the same time, Bianchi et al. (2021) highlight the close connection between risk premia predictability and macroeconomic states. Rossi (2021) provides a comprehensive discussion of the challenges faced by forecasting models in environments characterized by structural instability.

Overall, the recent literature suggests that the relative performance of forecasting models depends not only on the choice of algorithm but also on sample periods, macroeconomic conditions, and potential structural changes. Predictive gains appear to be episodic and environment-dependent rather than universally persistent. Therefore, a systematic comparison of different models within a unified out-of-sample framework, combined with an explicit examination of performance across market regimes, remains an important and relevant research question.

3. Methodology

3.1. Forecasting Models

3.1.1. Autoregressive (AR) Model

We use the autoregressive (AR) model as the linear benchmark. The AR(p) specification is given by

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t \tag{1}$$

where y_t denotes the return at time t, p is the lag order, and ε_t is a zero-mean disturbance term.

The AR model captures the linear dependence of returns on their own lags and serves as a standard benchmark for assessing the incremental predictive performance of nonlinear models in out-of-sample forecasting.

3.1.2. Support Vector Regression (SVR)

Support vector machines (SVM) are widely used supervised learning methods known for their strong generalization ability in classification tasks. Drucker et al. (1996) extend the SVM framework to regression problems, leading to support vector regression (SVR). A key feature of SVR is its ability to handle nonlinear relationships through the kernel trick. By introducing kernel functions, the original input space can be mapped into a higher-dimensional feature space, where nonlinear patterns in the data may become linearly separable. The standard SVR optimization problem is formulated as:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2}|w|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{2}$$

subject to:

$$\begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i, \forall i = 1, \ldots, n \\ \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^*, \forall i = 1, \ldots, n \\ \xi_i, \xi_i^* \geq 0, \forall i = 1, \ldots, n \end{cases} \tag{3}$$

where $|w|^2$ denotes the squared norm of the weight vector, C is the penalty parameter controlling the trade-off between model complexity and forecasting error, and ϵ defines the width of the insensitive loss region. The slack variables ξ_i and ξ_i^* allow deviations beyond the ϵ-tube.

By combining margin maximization with kernel-based nonlinear transformation, SVR provides a flexible framework for modeling potentially nonlinear relationships in financial return data.

3.1.3. Extreme Gradient Boosting (XGBoost)

XGBoost, proposed by Chen and Guestrin (2016), is an optimized implementation of the gradient boosting framework designed for efficiency, flexibility, and scalability in tree-based models. The core idea of XGBoost is to improve predictive performance by sequentially constructing and combining multiple decision trees in an additive manner. Each newly added tree is trained to correct the prediction errors of the existing ensemble. At each iteration, the model updates its predictions by minimizing a differentiable loss function using gradient-based optimization. This sequential error-correction mechanism allows the ensemble to progressively reduce overall prediction errors.

The objective function of XGBoost consists of two components: a loss function and a regularization term, which together balance goodness of fit and model complexity. The objective can be written as

$$Obj = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{4}$$

where $L\left(y_i, \hat{y}_i\right)$ measures the discrepancy between the actual value y_i and the predicted value $\hat{y}_i$, and $\Omega(f_k)$ denotes the regularization term for the k-th tree. The regularization term is typically defined as

$$\Omega\left(f_k\right) = \gamma T + \frac{1}{2}\lambda|w|^2 \tag{5}$$

where T represents the number of leaf nodes in the tree, w denotes the vector of leaf weights, and γ and λ are tuning parameters controlling model complexity. By explicitly penalizing the number of leaves and the magnitude of leaf weights, XGBoost mitigates overfitting while retaining flexibility in modeling nonlinear interactions.

3.1.4. Long Short-Term Memory (LSTM)

The Long Short-Term Memory (LSTM) network, proposed by Hochreiter and Schmidhuber (1997), is an improved recurrent neural network (RNN) architecture designed to address the vanishing and exploding gradient problems encountered in training standard RNNs over long sequences. Compared with conventional RNNs, LSTM introduces a gating mechanism that enables selective memory and forgetting of information, thereby improving the modeling of long-term dependencies.

The core components of LSTM include the forget gate, input gate, output gate, and a cell state that runs through the entire sequence. The cell state serves as a channel for transmitting long-term information across time steps, while the gates regulate how information is retained, updated, or discarded. The main computational steps are given by:

(1) Forget gate

$$f_t=\sigma\left(W_f \cdot [h_{t-1},x_t]+b_f\right) \tag{6}$$

(2) Input Gate

$$i_t=\sigma\left(W_i \cdot [h_{t-1},x_t]+b_i\right) \tag{7}$$

$$\tilde{c}_t= tanh\left(W_c \cdot [h_{t-1},x_t]+b_c\right) \tag{8}$$

(3) Cell State Update

$$C_t=f_t \odot C_{t-1}+i_t \odot \tilde{C}_t \tag{9}$$

(4) Output Gate

$$o_t=\sigma\left(W_o \cdot [h_{t-1},x_t]+b_o\right) \tag{10}$$

$$h_t=o_t \odot tanh\left(C_t\right) \tag{11}$$

where x_t denotes the input at time t, h_{t-1} is the hidden state from the previous period, and C_{t-1} is the previous cell state. The function $\sigma(\cdot)$ represents the sigmoid activation function, and $\odot$ denotes element-wise multiplication.

Through this gated structure, LSTM can preserve relevant long-term information while filtering out less informative signals. This property makes it suitable for modeling dynamic patterns in financial time series, where temporal dependence and nonlinear dynamics are often present.

3.2. Rolling Forecast Design

Given the time-varying nature of financial markets, we implement a fixed-length rolling window framework for out-of-sample forecasting. Let the total sample size be n and the estimation window length be T. Each model is recursively re-estimated using the most recent T observations to generate one-step-ahead forecasts, producing a sequence of out-of-sample predictions.

We focus on one-step-ahead forecasting to avoid error accumulation in multi-step settings and to ensure that all predictions are based solely on information available at the time of forecasting.

3.3. Evaluation Metrics

To evaluate the out-of-sample forecasting performance, this paper employs several commonly used loss functions, including mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). These metrics are defined as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} \left| y_t - y_{i,t}^F \right| \tag{12}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n} \left(y_t - y_{i,t}^F \right)^2 \tag{13}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \left(y_t - y_{i,t}^F \right)^2} \tag{14}$$

where y_t denotes the actual value at time t, $y_{i,t}^F$ is the forecast of model i, and n is the number of out-of-sample observations.

MAE measures the average magnitude of forecast errors, while MSE and RMSE place greater weight on large deviations. RMSE is expressed in the same unit as the original series and is widely used in financial forecasting.

To further assess predictive accuracy, we employ the out-of-sample R^2(denoted as R_{oos}^2), which compares each model with the benchmark AR model:

$$R_{oos}^2 = 1 - \frac{\sum_{t=1}^{n} \left(y_t - y_{i,t}^F \right)^2}{\sum_{t=1}^{n} \left(y_t - y_t^B \right)^2} \tag{15}$$

where y_t^B is the forecast from the benchmark model. A positive R_{oos}^2 indicates that the competing model outperforms the benchmark.

Finally, the Clark-West (CW) test is applied to examine whether the competing model significantly improves forecasting performance relative to the benchmark. The adjusted loss differential is defined as:

$$f_{t+1} = e_{1,t+1}^2 - \left(e_{2,t+1}^2 - (\hat{y}_{1,t+1} - \hat{y}_{2,t+1})^2 \right) \tag{16}$$

where model 1 denotes the benchmark and model 2 denotes the competing model. The null hypothesis is $E(f_{t+1}) = 0$. A significantly positive test statistic indicates that the competing model outperforms the benchmark.

4. Data

The data used in this study consist of daily observations of the CSI 300 Index. The sample period spans from January 2, 2020 to December 31, 2025. All data are obtained from the CSMAR database.

Daily closing prices are employed to construct the return series. Specifically, the continuously compounded return is calculated as:

$$r_t = 100 \times \ln\left(\frac{P_t}{P_{t-1}}\right) \tag{17}$$

where P_t denotes the closing price of the index on day t. Returns are expressed in percentage terms.

Table 1. Results of descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max	Skewness	Kurtosis	ADF	JB
Return	1455	0.008	1.1944	-8.2088	8.142	-0.2711	9.2839	-37.1581***	2392.2028***

Note: The Augmented Dickey–Fuller (ADF) test examines the null hypothesis that the series contains a unit root, while the Jarque–Bera (JB) test evaluates the null hypothesis of normality.***, **, * indicate statistical significance at 1%, 5%, or 10% levels, respectively.

Table 1 reports the descriptive statistics of daily returns for the CSI 300 Index. The sample consists of 1,455 observations. The mean return is 0.008, which is small relative to the standard deviation of 1.1944, indicating that daily return fluctuations substantially exceed the unconditional mean. This pattern is consistent with the typical "high volatility, low mean" characteristic of financial asset returns.

In terms of extreme values, the minimum return during the sample period is −8.2088, while the maximum reaches 8.142, suggesting the presence of pronounced market swings and extreme movements. The skewness is −0.2711, indicating a mildly left-skewed distribution with limited asymmetry. In contrast, the kurtosis is 9.2839, implying significant excess kurtosis and fat-tailed behavior. The Jarque–Bera test statistic is significant at the 1% level, strongly rejecting the null hypothesis of normality. This result confirms the presence of non-normal features in the return distribution, particularly fat tails, which are commonly observed in financial time series. In addition, the Augmented Dickey–Fuller test statistic equals −37.1581 and is significant at the 1% level, rejecting the null hypothesis of a unit root. This finding indicates that the return series is stationary in a statistical sense, consistent with the standard assumption in financial economics that asset returns follow a stationary process.

Overall, the daily returns of the CSI 300 Index exhibit high volatility, fat-tailed distribution, and significant non-normality. These empirical features motivate the use of forecasting models capable of accommodating nonlinearities and complex dynamic structures.

5. Empirical Results

5.1. Out-of-Sample Forecasting Performance

Under the out-of-sample rolling framework, we compare the predictive performance of different models for daily returns of the CSI 300 Index. Taking the AR model as the linear benchmark, we examine whether SVR, XGBoost (XGB), and LSTM provide improvements in forecast errors, directional accuracy, and statistical significance.

To ensure comparability, all models are estimated using the same rolling window length and one-step-ahead forecasting design.

Table 2 reports the out-of-sample forecasting performance of the competing models for the CSI 300 Index. In terms of the relative out-of-sample goodness-of-fit measure R^2_{oos}, all machine learning models yield positive values, indicating improvements over the AR benchmark under the MSE criterion. Among them, LSTM achieves the highest R^2_{oos} (approximately 2.22%), followed by XGBoost (1.99%), while SVR delivers a more modest gain (0.51%). These results suggest that nonlinear modeling contributes to improved return predictability.

Table 2. Out-of-Sample Forecast Performance

Models	R^2_{oos}	CW	MAE	MSE	RMSE	DA
AR	-	-	0.7732	1.3985	1.1784	0.4581
XGB	1.9892	2.8960***	0.7579	1.3609	1.1665	0.5044
SVR	0.5145	1.6640*	0.7606	1.3813	1.1753	0.4933
LSTM	2.2209	2.9332***	0.7443	1.3576	1.1652	0.4911

Note: This table reports the one-step-ahead forecasting results for the four models, including R^2_{oos} (in percentage), MAE, MSE, RMSE, and directional accuracy (DA). Statistical significance is assessed

using the MSFE-adjusted test proposed by Clark and West (2007). *** and * denote significance at the 1% and 10% levels, respectively.

Consistent evidence emerges from traditional loss measures (MAE, MSE, and RMSE). All machine learning models outperform the AR benchmark, with LSTM attaining the lowest forecast errors across all three metrics, indicating the strongest overall predictive accuracy. This finding is consistent with the presence of nonlinear dynamics in financial returns.

Statistical significance is assessed using the MSFE-adjusted test of Clark and West （2007）[30]. The results show that LSTM and XGBoost are significant at the 1% level, while SVR is significant at the 10% level. This indicates that the observed forecasting improvements are statistically meaningful rather than driven by random variation.

Regarding directional forecasting performance, machine learning models generally achieve higher directional accuracy than the AR benchmark. XGBoost ranks first, with a directional accuracy exceeding 50%, suggesting a relative advantage in predicting market movements. Although LSTM performs best in terms of loss minimization, its directional accuracy is slightly lower than that of XGBoost, implying potential trade-offs between error minimization and direction prediction.

Overall, LSTM delivers the strongest gains in forecast accuracy, whereas XGBoost exhibits superior performance in directional prediction. These findings indicate that machine learning and deep learning methods are more effective than the linear benchmark in capturing nonlinear structures in CSI 300 return dynamics, leading to significant out-of-sample improvements.

5.2. Relative Cumulative Forecast Errors

While average loss measures such as MSE and MAE summarize overall predictive performance, they do not reveal how relative forecasting advantages evolve over time. To address this limitation, we construct the Relative Cumulative Squared Prediction Error (RCSPE) with respect to the Autoregressive model benchmark. This measure accumulates the period-by-period differences in squared forecast errors, allowing us to trace the dynamic path of relative performance during the out-of-sample period. An upward (downward) trajectory indicates cumulative improvement (deterioration) relative to the benchmark.

Fig. 1 presents the RCSPE results using the AR model as the reference. Overall, the curves of the three machine learning models remain above zero for most of the evaluation period, consistent with the positive R^2_{oos} results reported earlier. However, their temporal patterns differ.

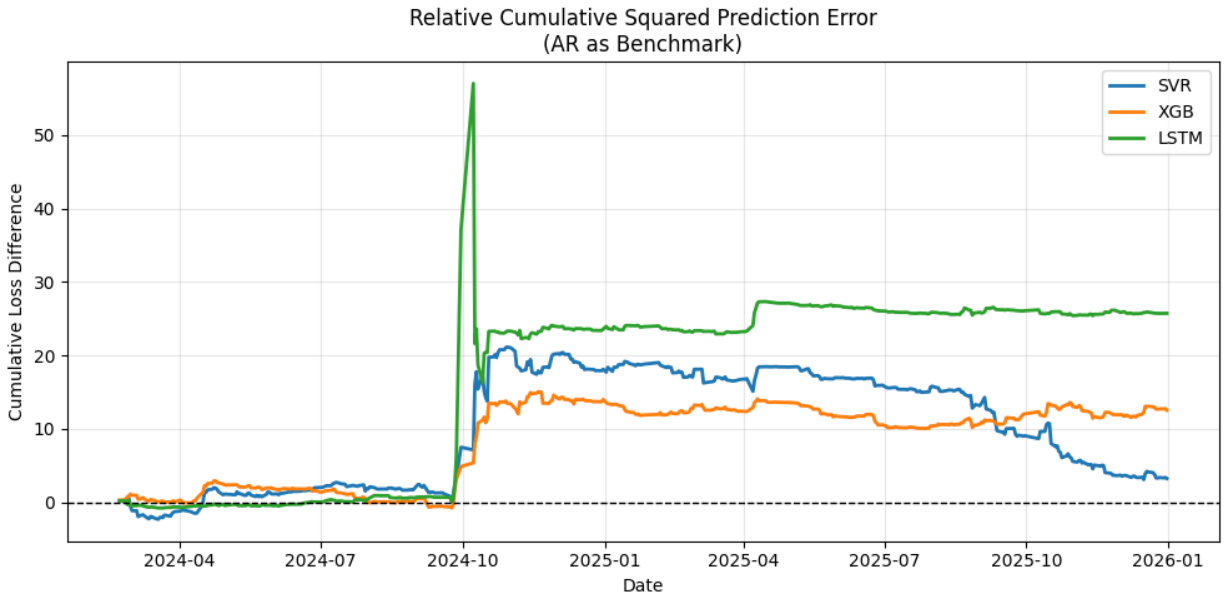

Fig. 1 Cumulative Relative Forecast Errors

SVR and XGBoost exhibit relatively smooth and gradually increasing trajectories, suggesting that their forecasting gains accumulate steadily over time. In particular, XGBoost displays comparatively stable improvements with limited fluctuations.

By contrast, the LSTM curve shows a pronounced upward jump in the early part of the out-of-sample period (around October 2024), indicating a phase in which its forecast errors were substantially lower than those of the AR model. As squared errors are sensitive to large deviations, this jump likely reflects concentrated performance differences during a high-volatility episode rather than a uniformly persistent advantage. After this adjustment, the LSTM curve continues to rise more gradually, implying that its relative gains are maintained but not further amplified.

Overall, the figure suggests that forecasting improvements are not uniformly distributed across time but instead display clear phase-dependent characteristics. Performance differentials tend to widen during certain periods and stabilize or narrow in others, highlighting the time-varying nature of model advantages. To further investigate this heterogeneity, the next subsection examines model performance under different volatility regimes.

5.3. Performance under Different Volatility Regimes

To examine whether predictive performance is state-dependent, we measure market volatility using the 20-day rolling standard deviation of returns and split the sample into low- and high-volatility regimes based on the 30th and 70th percentiles of full-sample volatility. Within each subsample, we recompute out-of-sample forecasting metrics, including RMSE, directional accuracy (DA), and R^2_{oos} relative to the Autoregressive model benchmark. This design allows us to assess whether predictive gains vary with the level of market uncertainty without altering model specifications.

Table 3 reveals clear volatility dependence in model performance. During low-volatility periods, the AR benchmark achieves the lowest RMSE, and all machine learning models produce negative R^2_{oos}, indicating no improvement over the linear specification. This suggests that when return fluctuations are subdued, dynamics may be closer to linear structures, limiting the incremental value of nonlinear models. Although SVR shows a slight improvement in directional accuracy, this does not translate into lower forecast errors.

Table 3. Forecast Performance across Volatility Regimes

Volatility Regime	Model	RMSE	DA	R^2_{oos}
Low Volatility	AR	0.6045	0.4351	-
	SVR	0.6211	0.4961	-0.0556
	XGB	0.6223	0.4503	-0.0599
	LSTM	0.6086	0.4274	-0.0137
High Volatility	AR	1.8416	0.3816	-
	SVR	1.8209	0.4732	0.0222
	XGB	1.8094	0.5190	0.0346
	LSTM	1.7846	0.5038	0.0609

Note: Volatility regimes are defined based on rolling standard deviation percentiles.

In contrast, during high-volatility periods, machine learning models outperform the AR benchmark. All R^2_{oos} values turn positive, with LSTM achieving the largest improvements in RMSE and R^2_{oos}, while XGBoost delivers relatively stable gains in directional accuracy. These findings indicate that nonlinear models are better suited to capturing complex return dynamics when market uncertainty intensifies.

Overall, the results suggest that predictive gains are not uniform across market conditions. Machine learning models tend to exhibit advantages primarily during periods of elevated volatility, whereas their benefits are limited in tranquil environments. This state-dependent pattern implies that model selection or combination strategies may benefit from incorporating information about prevailing market conditions.

6. Conclusion

This study employs a rolling out-of-sample forecasting framework to compare the predictive performance of a traditional linear model and several machine learning approaches for daily returns of the CSI 300 Index. Using the Autoregressive model (AR) as the benchmark, we evaluate models along three dimensions: forecast errors, directional accuracy, and statistical significance, and further examine their performance across different volatility regimes.

The empirical results show that, over the full out-of-sample period, machine learning and deep learning models outperform the AR benchmark in terms of mean squared prediction errors. LSTM achieves the strongest gains in loss-based measures and R^2_{oos}, with improvements confirmed by the Clark and West test. In contrast, XGBoost delivers the highest directional accuracy, indicating a relative advantage in predicting market movements. These findings suggest potential trade-offs between point forecast accuracy and directional performance.

The relative cumulative squared prediction error analysis further indicates that forecasting gains are time-dependent rather than uniformly distributed. Improvements tend to concentrate in specific periods. Consistent with this pattern, regime-based results show that machine learning models exhibit advantages primarily during high-volatility episodes, while the linear benchmark remains competitive in low-volatility environments. This evidence supports the view that return predictability is state-dependent, and that the benefits of nonlinear models become more pronounced when market uncertainty intensifies.

Overall, the results suggest that machine learning and deep learning methods offer potential advantages in capturing nonlinear dynamics in return series, but their predictive gains are not uniformly stable across market conditions. Future research may explore the economic mechanisms underlying these performance differences and examine model combination strategies for practical investment applications.

References

[1] Fama E F, French K R. Dividend yields and expected stock returns [J]. Journal of Financial Economics, 1988, 22 (1): 3-25.

[2] Welch I, Goyal A. A comprehensive look at the empirical performance of equity premium prediction [J]. The Review of Financial Studies, 2008, 21 (4): 1455-1508.

[3] Campbell J Y, Thompson S B. Predicting excess stock returns out of sample: Can anything beat the historical average? [J]. The Review of Financial Studies, 2008, 21 (4): 1509-1531.

[4] Meese R A, Rogoff K. Empirical exchange rate models of the seventies: Do they fit out of sample? [J]. Journal of International Economics, 1983, 14 (1-2): 3-24.

[5] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning [J]. The Review of Financial Studies, 2020, 33 (5): 2223-2273.

[6] Chen L, Pelger M, Zhu J. Deep learning in asset pricing [J]. Management Science, 2024, 70 (2): 714-750.

[7] Bianchi D, Büchner M, Tamoni A. Bond risk premiums with machine learning [J]. The Review of Financial Studies, 2021, 34 (2): 1046-1089.

[8] Rapach D E, Strauss J K, Zhou G. International stock return predictability: what is the role of the United States? [J]. The Journal of Finance, 2013, 68 (4): 1633-1662.

[9] Dangl T, Halling M. Predictive regressions with time-varying coefficients [J]. Journal of Financial Economics, 2012, 106 (1): 157-181.

[10] Pettenuzzo D, Timmermann A, Valkanov R. Forecasting stock returns under economic constraints [J]. Journal of Financial Economics, 2014, 114 (3): 517-553.

[11] Kozak S, Nagel S, Santosh S. Shrinking the cross-section [J]. Journal of Financial Economics, 2020, 135 (2): 271-292.

[12] Cao L J, Tay F E H. Support vector machine with adaptive parameters in financial time series forecasting [J]. IEEE Transactions on neural networks, 2003, 14 (6): 1506-1518.

[13] Huang W, Nakamori Y, Wang S Y. Forecasting stock market movement direction with support vector machine [J]. Computers & operations research, 2005, 32 (10): 2513-2522.

[14] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions [J]. European Journal of Operational Research, 2018, 270 (2): 654-669.

[15] Bao W, Yue J, Rao Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory [J]. PloS one, 2017, 12 (7): e0180944.

[16] Krauss C, Do X A, Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500 [J]. European Journal of Operational Research, 2017, 259 (2): 689-702.

[17] Zhang Z, Zohren S, Roberts S. Deep learning for portfolio optimization [J]. arXiv preprint arXiv:2005.13665, 2020.

[18] McLean R D, Pontiff J. Does academic research destroy stock return predictability? [J]. The Journal of Finance, 2016, 71 (1): 5-32.

[19] Kim H H, Swanson N R. Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods [J]. International Journal of Forecasting, 2018, 34 (2): 339-354.

[20] Lettau M, Pelger M. Factors that fit the time series and cross-section of stock returns [J]. The Review of Financial Studies, 2020, 33 (5): 2274-2325.

[21] Feng G, Giglio S, Xiu D. Taming the factor zoo: A test of new factors [J]. The Journal of Finance, 2020, 75 (3): 1327-1370.

[22] Rossi B. Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them [J]. Journal of Economic Literature, 2021, 59 (4): 1135-1190.

[23] Campbell J Y, Thompson S B. Predicting excess stock returns out of sample: Can anything beat the historical average? [J]. The Review of Financial Studies, 2008, 21 (4): 1509-1531.

[24] Drucker H, Burges C J, Kaufman L, et al. Support vector regression machines [J]. Advances in neural information processing systems, 1996, 9.

[25] Breiman L. Random forests [J]. Machine learning, 2001, 45 (1): 5-32.

[26] Chen T, Guestrin C. Xgboost: A scalable tree boosting system [C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016: 785-794.

[27] Tibshirani R. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996, 58 (1): 267-288.

[28] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997, 9 (8): 1735-1780.

[29] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. arxiv preprint arxiv:1406.1078, 2014.

[30] Clark T E, West K D. Approximately normal tests for equal predictive accuracy in nested models [J]. Journal of econometrics, 2007, 138 (1): 291-311.