The Effectiveness of Metamodeling in Predicting the Sign of Asset Returns: An Empirical Assessment

Paul Cristian Donoiu[1]

Abstract

The study evaluates the effectiveness of metamodeling in predicting the sign of next-day returns for 30 assets between 2001 and 2024. In essence, we compared using 10-year rolling windows five single models (ARIMA, Logistic Regression, Random Forest, XGBoost and LSTM) with a metamodel designed to predict the sign using two rules – majority, if four out of five models predict the same sign, or a fallback mechanism if there is no consensus among individual models. The performance is evaluated using 3 indicators: sign accuracy, Sharpe ratio and cumulative returns. A separate analysis was carried out for the period of the global financial crisis. Results indicate that the Metamodel generally provides robustness and performance similar to, but slightly worse than that of the best single model. The situation is the same when we take in consideration the global financial crisis period. The performance of the Metamodel is better when analyzing equities from the technology sector or stock indices. Thus, metamodeling is useful as a stabilizing instrument and to reduce the model selection risk, but the benefits depend significantly on market regimes or asset classes.

Keywords: metamodeling; equity return sign prediction; machine learning; financial time series

JEL Classifications: G11; G15; G17; C45

DOI: 10.24818/REJ/2026/92/08

1. Introduction

Predicting the next-day return sign remains a central issue in finance, having major implications for the selection of trading strategies and financial risk management. Researchers have studied extensively the performance of both classical econometric models and modern machine learning algorithms and have obtained various results depending on the assets that they used, the market conditions or the analyzed interval. A parallel line of work, based on forecast combination and, more recently, metamodeling, has been proposed to combine two or more models in order to obtain superior performance.

This study aims to investigate the effectiveness of metamodeling in predicting the sign of returns, using a large number of diversified assets with global coverage (30

[1] PhD candidate, Bucharest University of Economic Studies, Bucharest, Romania; paul_donoiu@yahoo.com

assets – major global indices and equities from different economic sectors) and a long period of time (2001-2024). The benchmark was the performance of 5 single models, respectively ARIMA, Logistic Regression, Random Forest Classifier, XGBoost Classifier and LSTM. The Metamodel was designed to use a majority rule to predict next-day sign and, if it is not the case, a fallback mechanism.

The results suggest that metamodeling tends to offer robustness and competitive performance in predicting next-day return sign, but, in most cases, the Metamodel failed to be better than the best single model. The difference in performance between them varies depending on market regime or asset class.

The contribution of this study to the existing literature consists of several aspects. Firstly, we provide a large scale, cross-sectoral empirical analysis of metamodeling, evaluating the performance by using a wide range of financial assets. Secondly, we evaluate the performance of the models during different market regimes, with a particular focus on the global financial crisis. Thirdly, we compare the Metamodel with the best single model in every analyzed window, which offers a better perspective on performance than simply comparing it with the best single model over the whole sample.

2. Review of the scientific literature

The prediction of the sign of asset returns has been extensively studied, being one of the main focuses of researchers and practitioners worldwide. A wide range of methods, from classical econometric models, such as Logistic Regression and ARIMA, to modern artificial intelligence algorithms, such as Random Forest, XGBoost or neural networks (e.g., LSTM) has been used to identify trends in market evolution. However, studies such as Sonkavde et al. (2023) show that despite the rapid development and increasing complexity of these methods, the individual models have fluctuating performance depending on the markets and the analyzed interval.

Metamodeling is a technique used to combine the predictive power of two or more methods in order to create a metamodel designed to outperform single models in predicting market dynamics. This concept builds on the foundation of forecast combination, a concept introduced several decades ago in financial literature (Bates and Granger, 1969; Clemen, 1989). For example, Yu et al. (2009) show that a neural-network metamodel approach achieves better performance than individual methods in forecasting financial time series. In a more recent study, Gambetti et al. (2022) prove that using a metamodel can increase the prediction accuracy for bond

recovery rates. Similarly, Abir et al. (2025) construct a metamodel using LSTM, Random Forest, Gradient Boosting and SVM, which, according to the authors, has a better predictive power than individual models for detecting the market dynamics in BRICS countries. An additional argument for combining forecasts is highlighted by Hendry and Hubrich (2010) which analyze different ways to estimate the dynamics of a variable and show that it is safer to aggregate information from multipe models rather than from only one when there is model uncertainty and errors may occur. The authors emphasize that the performance differences are not caused only by regime shifts, but also from the estimation of the model and its variance. All of these are arguments for combining forecasts in order to enhance performance.

One of the most important advantages of metamodels is that they can remain more stable under some regime shifts. In contrast, individual models tend to learn from historical trends, which makes them vulnerable to rapid market changes. For example, Wang and Lera (2024) develop a metamodel that learns the relationships between market conditions and future returns and outperforms a wide range of individual models in forecasting market dynamics.

In this strand of the literature, Ang and Timmermann (2011) show that financial markets go through different phases, in which the expected return, volatility and correlations change and remain at those level for a period of time. A direct consequence is that models calibrated on a single market regime tend to lose their forecast accuracy when the regime changes. To prevent this, it is recommended to use models that can incorporate and select among multiple specifications. More and more recent studies on financial series propose a mechanism based on combining forecasts and metamodeling. For example, Su et al. (2025) combine multiple model architectures and adjust their weights in order to better capture the market regime shifts. According to the authors, this approach offers gains in accuracy and a better forecasting stability than using a single model.

In summary, prior evidence indicates that metamodeling can enhance the predictive power of models under certain market conditions or time frames. However, the effectiveness of metamodeling can be weaker when using long historical periods and a wide range of globally distributed assets for the analysis.

3. Research methodology

The purpose of this study is to evaluate whether metamodeling improves the accuracy of predicting the next-day return sign.

The study uses daily return data from the Yahoo Finance database for a mix of 30 assets – global indices (S&P 500, NASDAQ, DAX 30, FTSE 100, NIKKEI 225, HANG SENG INDEX) and equities across major sectors (technology, financial, energy, industrials, consumers and healthcare). The analyzed interval is 01/01/2001 – 31/12/2024. For the purpose of this study, we use logarithmic returns. Predictors are the lagged log returns *(R_{t-1}, R_{t-2}, R_{t-3}, R_{t-4}, R_{t-5})* and the target is the sign of the next-day logarithmic return *(R_{t+1})*.

To capture the evolution of model performance over time, we constructed rolling windows of 10 years with a 6-month step. The data were split chronologically 70/30 into train and test. This methodological approach ensures no data leakage and a comprehensive analysis of various market regimes.

The models used in this analysis for forecasting are ARIMA, Logistic Regression, Random Forest Classifier, XGBoost Classifier and LSTM. They were applied in Python, using a wide range of libraries, namely *statsmodels* for ARIMA, *scikit-learn* for Logistic Regression and Random Forest, *xgboost* for XGBoost Classifier, *tensorflow/keras* for LSTM.

The Metamodel was designed to combine all five models by the following procedure. If the sign predicted by at least 4 models is the same, the metamodel adopts that sign (majority). The threshold was chosen in order to strike a balance between the strength of cross-model consensus and the availability of trading signals. Imposing a four-model agreement requirement reduces the likelihood that the metamodel's decision is driven by a weak or unstable signal. Lower thresholds, such as 3/5 voting rule, would increase the frequency of predictions but could lead to situations characterized by more fragile agreement across models.

If there is not a clear majority, the Metamodel uses a fallback mechanism: the sign comes from the individual model with the highest accuracy on training data. Because of the fact that the fallback decision is based solely on in-sample training accuracy, the metamodel may be exposed to a degree of overfitting, aspect that should be taken into consideration when interpreting the results.

The trading strategy used involves multiplying the sign predicted by the six models with the actual return on day t+1. The performance of the models is evaluated by 3 indicators: 1) sign accuracy *(%);* Sharpe ratio (using the returns of the strategy); cumulative return (%). The performance metrics are used for comparative purposes rather than to demonstrate the profitability of a trading strategy, and the results

should therefore be interpreted with caution given the absence of transaction costs and other market frictions.
Moreover, to evaluate model performance under adverse market conditions, we identify rolling windows for which the test sample overlaps the global financial crisis period (December 2007–June 2009) for a minimum duration of six months.

4. Results and discussions

Table 1 presents the overall performance of the 6 models, measured on the whole analyzed interval. As shown, the best model in predicting sign of returns was ARIMA, with a mean sign accuracy of 51.3%, followed by Logistic Regression (51.25%) and the Metamodel (51.08%). Random Forest, XGBoost or LSTM, which are theoretically more complex than linear models, show weaker performance. ARIMA also achieves the best Sharpe ratio and cumulative return, followed closely by Logistic Regression and the Metamodel. Thus, we can observe that although metamodeling does not achieve the best performance, it offers robustness and balanced results.

Table 1. Overall performance of the models

Model	LR	ARIMA	LSTM	RF	XGB	META
Acc(%)_mean	51.2482	51.3033	50.2492	49.9789	50.0486	51.0803
Acc(%)_median	51.1936	51.5272	50.1326	49.9336	50.0000	51.0610
Acc(%)_std	2.4071	2.7265	2.8340	1.9151	1.9346	2.4717
Sharpe_mean	0.0188	0.0229	0.0103	0.0022	0.0061	0.0187
Sharpe_median	0.0207	0.0286	0.0108	0.0025	0.0067	0.0233
Sharpe_std	0.0386	0.0391	0.0422	0.0352	0.0354	0.0394
CumRet(%)_mean	36.7513	34.5143	17.3735	0.4755	6.6161	33.8094
CumRet(%)_median	14.0802	24.3711	2.2753	-6.3660	-2.4333	17.9603
CumRet(%)_std	146.72	121.8015	85.0149	47.2102	52.1372	131.1544

Source: data: author's calculations

When comparing the Metamodel with the best single model, which most of the time is Logistic Regression or ARIMA *(Table 2)*, we can observe that in approximately 82% of the analyzed cases the single model is more efficient. Moreover, on average, the Metamodel compared with the best single model obtained poorer performance in terms of sign accuracy (-1.58 p.p.), Sharpe ratio (-0.02) and cumulative return (-37.29 p.p.).

Table 2. Comparative performance of the Metamodel vs. the best single model

Win% (Meta1 > BestSingle)	8.0460
Tie% (Meta1 = BestSingle)	10.3448
Δ(Meta-BestSingle) Acc_mean (p.p.)	-1.5836
Δ(Meta-BestSingle) Acc_median (p.p.)	-1.1936
Δ (Meta-BestSingle) Sharpe_mean	-0.0214
Δ (Meta-BestSingle) Sharpe_median	-0.0136
Δ (Meta-BestSingle) CumRet_mean (p.p.)	-37.2914
Δ(Meta-BestSingle) CumRet_median (p.p.)	-17.1112

Source: data: author's calculations

During periods of adverse market conditions *(Table 3)*, Logistic Regression and the Metamodel obtain better performance than ARIMA, taking into consideration all three indicators: sign accuracy, Sharpe ratio and cumulative return. The performance of these two models in crisis periods is similar to their performance over the entire analyzed interval, while in the case of the other models the performance has deteriorated significantly. This means that while the Metamodel does not achieve the results of Logistic Regression, it represents a good compromise by avoiding the performance deterioration recorded by complex models, probably caused by the difficulties of adapting to rapid market shifts.

Table 3. Performance of the models during the global financial crisis

	LR	**ARIMA**	**LSTM**	**RF**	**XGB**	**META**
Acc(%)_mean	51.1375	50.9053	49.2468	49.6401	49.7931	50.9301
Acc(%)_median	51.4608	50.837	48.9376	49.668	49.4024	51.1952
Acc(%)_std	2.1383	2.1816	2.2648	1.8921	2.1144	2.1778
Sharpe_mean	0.03	0.0108	0.0027	-0.0082	0.001	0.0166
Sharpe_median	0.0288	0.0107	0.0014	-0.0098	-0.0023	0.0175
Sharpe_std	0.0344	0.0199	0.0304	0.0318	0.032	0.0298
CumRet(%)_mean	71.7307	-1.0213	-1.1582	-18.3691	1.4124	31.976
CumRet(%)_median	35.1735	-1.7002	-22.0786	-30.4497	-19.5087	8.1318
CumRet(%)_std	177.26	38.4092	67.757	49.4802	79.1395	137.4421

Source: data: author's calculations

Figure 1 (vertical axis: assets, horizontal axis: 10-year rolling windows) synthesizes the difference in sign accuracy between the Metamodel and the best single model. It shows a slightly negative aggregate pattern, with green as the dominant color, which means that most of the figure cells are between -2 and 0 p.p. Moreover, we can observe that all cells in the figure associated with indices and with companies

from technological sectors have almost a neutral color, which indicates that the Metamodel tends to perform well. The same happens when analyzing the figure cells associated with periods of adverse market regimes. This emphasizes that metamodeling works as a stabilization mechanism, allowing to obtain competitive results, even if they are not the best.

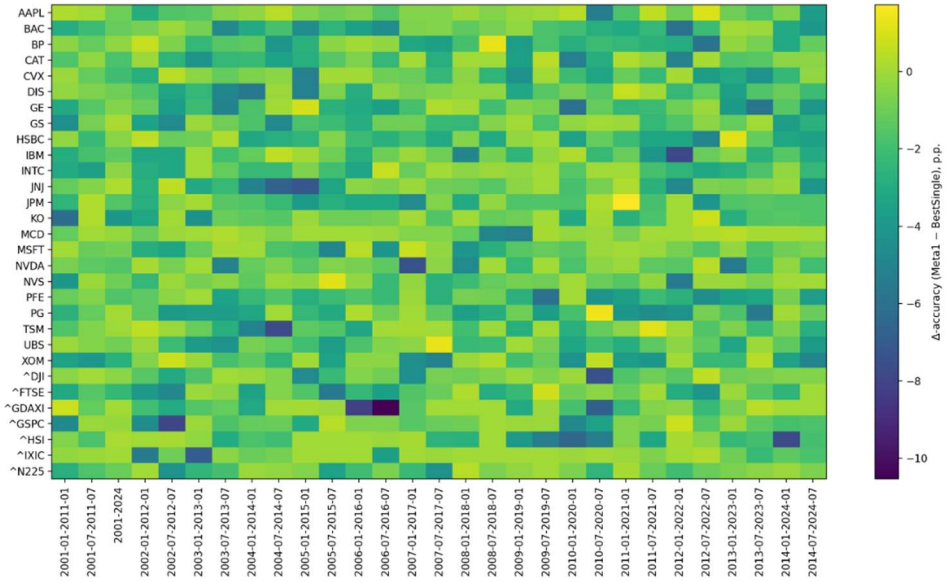

Figure 1. Difference in sign accuracy between the Metamodel and the best single model
Source: data: author's calculations

Table 4 presents the models' performance across different economic sectors. Similar to the overall performance, ARIMA and Logistic Regression offer the best results, while machine learning models fail to predict next-day returns sign when using daily data. The performance of the Metamodel is robust across all economic sectors, with better results in predicting next-day return sign of companies in technology industry. The poor performance related to energy and industrial sectors, both for individual models and the Metamodel, is probably caused by the exogenous volatility, which made it difficult for the model to learn patterns.

Table 4. Performance of the models across asset classes

Asset Class	Model	Acc(%)_mean	Sharpe	CumRet(%)
Consumer& Healthcare	ARIMA	52.0220	0.0375	35.7751
	LR	51.3893	0.0205	21.7573
	META	51.2625	0.0230	22.6909

Asset Class	Model	Acc(%)_mean	Sharpe	CumRet(%)
	LSTM	50.2830	0.0138	13.5276
	RF	49.9052	0.0027	1.1277
	XGB	49.8679	0.0040	3.7648
Energy& Industrials	LR	50.0370	-0.0001	-3.8362
	ARIMA	49.8666	0.0041	5.1599
	META	49.7892	0.0005	-3.0123
	XGB	49.6176	-0.0020	-6.7723
	LSTM	49.5196	-0.0027	-9.1128
	RF	49.4126	-0.0074	-10.5693
Financials	LR	50.1398	0.0142	48.1326
	META	49.8611	0.0103	26.7532
	ARIMA	49.7738	0.0062	4.3630
	LSTM	49.7641	0.0077	16.1048
	XGB	49.5535	0.0071	7.6371
	RF	49.3300	-0.0003	-3.5092
Tech	ARIMA	51.3970	0.0352	100.3282
	META	51.3388	0.0321	98.8072
	LR	51.1186	0.0296	92.9061
	LSTM	50.6435	0.0196	55.7710
	XGB	49.8705	0.0106	16.5923
	RF	49.8343	0.0053	4.5722
Index	LR	52.8750	0.0246	24.4741
	ARIMA	52.6231	0.0231	19.3458
	META	52.4694	0.0217	20.5569
	RF	51.0447	0.0079	7.0470
	XGB	51.0433	0.0094	9.7503
	LSTM	50.7452	0.0100	8.1321

Source: data: author's calculations

The observation linked to the indices and the technology companies is in alignment with the literature stating that in those segments where is a significant tendency component or in those where the information is much better reflected in prices, the aggregation of many prediction rules tends to generate much more concrete results. In contrast, regarding those assets highly influenced by external shocks - energy, some industrial companies - the signal derived only on lagged returns is weaker, therefore the metamodel has less information to combine.

5. Conclusions

This paper evaluates whether metamodeling improves the accuracy and the quality of results in predicting the next-day sign of returns. Overall, the performance of the Metamodel is very close to the one of the best single model across all analyzed windows, but, on average, is weaker by about 1-2 p.p. The Metamodel registers solid performances across certain classes of assets, such as equity of companies from technology sectors or major global indices.

From an applied perspective, metamodeling act as an insurance against picking the wrong model at a given moment of time, providing robustness throughout the interval. However, the Metamodel is rarely better than the best individual model, recording most of the time the second or the third-best performance. The benefit of using metamodeling increases in periods where there is consensus between single models and decreases in timeframes dominated by strong autoregressive patterns.

From an operational perspective, the results suggest that the metamodel should be considered as a tool for regularising the forecasting process rather than a method for improving accuracy. For a manager who runs multiple parallel signals and has to choose daily which ones are trustworthy, the fact that the metamodel is constantly positioned in proximity to the most tested models is an advantage: it reduces decision costs by lowering the possibility of selecting the least adaptive model.

The limitations of this study include the simplicity of the parameters used (only lagged returns) and the lack of technical and fundamental variables that could improve the performance of the Metamodel. In this context, the results should be interpreted as evidence on the behaviour of individual models and the metamodel under restricted information, rather than as a general assessment of the effectiveness of machine learning methods in forecasting returns. Some future research directions could imply evaluating the performance of metamodeling by constructing a model that uses weights depending on the performance of single models during the training period, combines single models depending on market regimes or uses more input parameters.

References

Abir, S.I., Saimon, S.I., & Saha, T.R., 2025. Comparative analysis of currency exchange and stock markets in BRICS using machine learning to forecast optimal trends for data-driven decision making. *Journal of Economics, Finance and Accounting Studies*, 7(1). Available at: https://al-kindipublishers.org/index.php/jefas/article/view/8566 [Accessed 26 Oct. 2025].

Ang, A., & Timmermann, A., 2011. Regime Changes and Financial Markets. NBER *Working Paper Series*, 17182. Available at: https://www.nber.org/papers/w17182 [Accessed 26 Oct. 2025].

Bates, J.M., & Granger, C.W.J., 1969. The combination of forecasts. Operational Research Quarterly, 20(4), pp. 451-468. Available at: https://www.jstor.org/stable/3008764 [Accessed 26 Oct. 2025].

Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4), pp. 559-583. Available at: https://doi.org/10.1016/0169-2070(89)90012-5 [Accessed 26 Oct. 2025].

Gambetti, P., Roccazzella, F., & Vrins, F., 2022. Meta-learning approaches for recovery rate prediction. Risks, 10(6), p. 124. Available at: https://www.mdpi.com/2227-9091/10/6/124 [Accessed 26 Oct. 2025].

Hendry, D.F., & Hubrich, K., 2010. Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *ECB Working Paper Series*, 1155. Available at: https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1155.pdf [Accessed 26 Oct. 2025].

Sonkavde, G., Dharrao, D.S., Bongale, A.M., Deokate, S.T., Doreswamy, D., & Bhat, S.K., 2023. Forecasting Stock Market Prices Using Machine Learning and Deep Learning Models: A Systematic Review, Performance Analysis and Discussion of Implications. *International Journal of Financial Studies*, 11(3), 94. https://doi.org/10.3390/ijfs11030094 [Accessed 26 Oct. 2025].

Sun, Y., Qu, Z., Zhang, T., & Li, X., 2025. Adaptive ensemble learning for financial time-series forecasting: A hypernetwork-enhanced reservoir computing framework with multi-scale temporal modeling. *Axioms*, 14, 597. Available at: https://doi.org/10.3390/axioms14080597 [Accessed 26 Oct. 2025].

Timmermann, A., 2006. Forecast combinations. Handbook of Economic Forecasting, 1, pp. 135-196. Available at: doi:10.1016/j.ijforecast.2005.05.004 [Accessed 26 Oct. 2025].

Wang, Y., & Lera, S.C., 2024. Meta-learning for return prediction in shifting market regimes. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5022829 [Accessed 26 Oct. 2025].

Yu, L., Wang, S., & Lai, K.K., 2009. A neural-network-based nonlinear metamodeling approach to financial time series forecasting. Applied Soft Computing, 9(2), pp. 563-574. Available at: https://www.sciencedirect.com/science/article/pii/S156849460800118X [Accessed 26 Oct. 2025].