Proceedings of the 4th International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/2025.BJ30479

Stock Market Prediction Using Machine Learning: A Multi-Model Ensemble

Fuyu Jin[1*], Xiangzhao Song[2], Jiajin Zhong[3]

[1]Shanghai Pinghe School, Shanghai, China
[2]Shenzhen University Normal College International High School, Shenzhen, China
[3]Taiyuan No.48 High School, Taiyuan, China
**Corresponding Author. Email: 18939822136@163.com*

Abstract. The essay presents a multi-horizon framework for S&P 500 stock price prediction, basically integrating Ridge regression and Long short-term memory. The proposed methodology aims to address the inherent challenges of financial time series prediction by combining three-dimension data sources: technical, sentiment and macroeconomics indicators. Moreover, the time phase of prediction involves 1-day, 10-day and 30-day prediction, which can be utilized in both short-term and long-term investment strategy. By testing distinct model results in the three ranges, we implement specialized and differentiated combination of models and indicators: for 1-day prediction, employing Ridge Regression processing sentiment and macroeconomic features combined with LSTM which processes the times series. The weight model is weighted averaging; the 10-day horizon incorporates technical indicators alongside sentiment and macro factors, implementing accuracy-weighted ensemble methods; The 30-day prediction leverages all the feature set with Gradient Boosting integration. Our 1-day prediction R^2 score reaches 0.002 and accuracy reaches 83.0 %; 10-day prediction R^2 score gets 0.456 and accuracy reaches 76.1%; 30-day prediction R^2 score is 0.557 and accuracy reaches 77.9%, which demonstrates a strong positive correlation between R^2 score and time phases and a negative correlation between accuracy and time phases.

Keywords: S&P 500, Prediction, Direction Accuracy, LSTM, XGBOOST

1. Introduction

As global conditions, particularly in the political and economic spheres, become increasingly volatile, financial markets have experienced heightened levels of instability and volatility. For instance, in 2025, following the extreme Tariff policy of the Trump administration, U.S. equity markets underwent a severe crash. Between 2 April and 10 April 2025, the S&P 500 declined by 14.6%, the Dow Jones Industrial declined by 10.3%, and the Nasdaq declined by 20.6% [1]. Transparenty, such drastic downturns cannot be projected without accounting for factors of political, macroeconomic, and public sentiment.

Consequently, successful participation in equity markets has become highly dependent on professional expertise and real-time monitoring and understanding of the economic situation,

DOI: 10.54254/2754-1169/2025.BJ30479

rendering investment inaccessible for many individual investors. To address this challenge, our forecasting project employs the SP 500 as a benchmark and leverages machine learning techniques as a supplementary tool to support investment decision-making. After building a specific model for the SP 500, we can employ the same mind map and a similar code for other stocks.

2. Model engineering

2.1. LSTM model

Traditional approaches include time-series models (such as ARIMA and GARCH) and factor models, but these often struggle with the non- linear patterns and regime changes inherent in financial markets. In recent years, machine learning and especially deep learning methods have gained popularity for this task. Recurrent neural networks, particularly LSTM models, have been frequently applied to capture temporal dependencies in stock data. For instance, Fischer and Krauss demonstrated that an LSTM-based strategy could slightly outperform logistic regression in predicting SP 500 daily movements, indicating the potential of deep sequence models to extract subtle pattern [2]. Generally, LSTM models tend to outperform classical models like ARIMA on financial time series. Hence, we first employ LSTM model to predict by processing the close price of SP 500.

For each trading day t, we feed the LSTM with a window of the past 20 trading days' closing prices (approximately one month of history). Based on preliminary experimentation, a lookback of 20 days provided a good balance: it is long enough to capture short-to-medium-term momentum or reversals, yet short enough to avoid excessive dimensionality given our limited data. Before input to the network, the price sequence is normalized (using z-scores) so that the LSTM sees a scaled time series with mean 0 and unit variance (computed on the training set).

Basically, our LSTM model is a relatively compact deep network, to mitigate the risk of overfitting on a not very large training set. It consists of:

• An initial LSTM layer with 64 units, configured to return sequences (so that it outputs a hidden state at each time step in the input window).

• A second LSTM layer with 32 units, which processes the sequence from the previous layer and returns only the final hidden state (i.e., this is a many-to-one architecture where the network's output corresponds to the information distilled after reading the 20-day sequence).

• We include Dropout regularization (dropout rate of 0.2) and batch normalization after each LSTM layer. Dropout randomly drops a fraction of the units during training, which helps prevent overfitting, and batch normalization stabilizes training and helps the network generalize.

• After the LSTM layers, we add a fully connected (dense) layer with 16 neurons and ReLU activation. This intermediate dense layer allows the model to learn a further non-linear combination of the LSTM's output features.

• Finally, we have an output layer with a single neuron and linear activation, which produces the predicted return (as a continuous value). This is trained to approximate the target log-return for the given horizon.

The LSTM model is trained using a regression loss (mean squared error) between its prediction and the actual target return. We use the Adam optimizer for efficient gradient-based optimization. We also apply early stopping based on validation loss to prevent over-training.

Moreover, it is important to highlight that the LSTM model by itself is trying to predict returns using only the index price history (and possibly its internal memory of patterns). This means it does not directly know about macroeconomic or sentiment conditions – those factors might indirectly

DOI: 10.54254/2754-1169/2025.BJ30479

influence prices, but the LSTM would have to infer them from the price behavior. This justifies our addition of the second model that explicitly uses those external features.

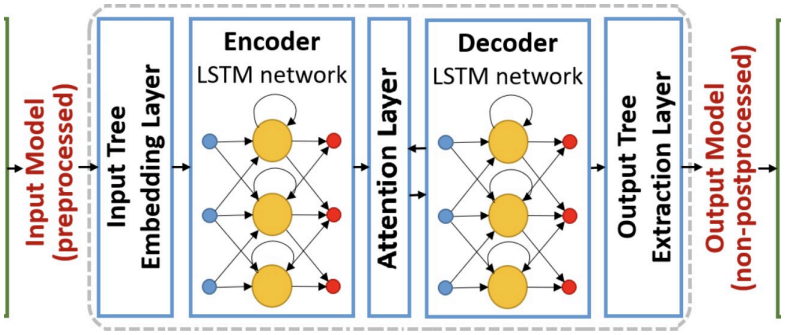

Figure 1. LSTM-based neural network

Figure 1 describes the complete picture of the network of LSTM. Indeed, the LSTM is a complex, complete and efficient model; however, the result of prediction by only LSTM was lower than anticipated, in which the R ∗ ∗2 is -7.079 and accuracy is only 53.18 % ,close to random prediction. As such, this model is necessary to be improved, taking into account more factors to improve the accuracy and robustness of the system.

2.2. Indicators selection

First of all, Sentiment indicators can significantly benefit the prediction of stock market because it demonstrates market confidence, which represents as an aspect of demand and supply. In particular, FinBERT – a BERT-based language model fine-tuned for financial text – has emerged as a powerful tool to quantify sentiment in finance-specific language. Prior studies have found that sentiment scores extracted from news using FinBERT or similar models can enhance prediction performance when combined with price data. For example, Zeng and Jiang integrated FinBERT-derived news sentiment with an LSTM model and reported that adding sentiment analysis "significantly enhances the model's ability to anticipate market fluctuations", relative to using price features alone [3].

Meanwhile, the role of macroeconomic variables in explaining and forecasting stock returns has been well documented in the asset pricing literature. Macroeconomic indicators such as GDP growth, inflation, unemployment rate, and interest rates reflect the overall health of the economy and can influence investors' expectations of future corporate earnings. Therefore, they can also serves as a determinant of demand and supply of a certain stock, given that SP 500 is almost representative to the overall stock market condition is U.S. Economic theory suggests that macroeconomic conditions affect stock prices through several channels:

• Strong GDP growth and low unemployment increase corporate revenues and profits, supporting higher stock valuations.

• Higher interest rates increase the discount rate applied to future cash flows, typically leading to lower stock prices.

• Rising inflation can erode real returns and create uncertainty, often causing downward pressure on equity markets.

• Macroeconomic instability increases risk aversion among investors, reducing demand for equities.

For instance, Chen, Roll, and Ross demonstrated that systematic economic forces—including inflation, industrial production, and interest rates—have significant explanatory power for equity

DOI: 10.54254/2754-1169/2025.BJ30479

market movements [4]. Their findings suggest that macroeconomic fundamentals are important drivers of expected returns, especially over medium- to long-term horizons, where such factors have sufficient time to influence corporate earnings and investor risk appetite.

Finally, technical indicators are also necessary for stock prediction. Indeed, all professional investors determine the rise and fall direction of stock price by looking at and analyzing such technical indicators such as MACD and RSI. Specifically, they are derived from historical price series to capture momentum and trend dynamics. They are computed through transformations of moving averages and relative price ratios, allowing the model to incorporate information about short-term fluctuations and long-term market tendencies.

2.3. Data source

This essay considers three indicators including technical, macroeconomic and sentiment indicators. The technical indicators is calculated by the SPY ETF data that can be found online. Python can read historical data of SP 500, and calculate such factors: RSI, MACD, Bollinger Bands, Moving averages (5, 10, 20, 50 days), Price momentum, volatility measures, Volume ratios, volume anomalies of SP 500, etc. Meanwhile, the essay also integrate VIX volatility index includes VIX close price, VIX change percentage, VIX moving averages (5-day, 20-day), VIX volatility measures. The essay collect those datas from Yahoo Finance API (yf.download('VIX') [5].

For Sentiment data, the essay directly utilizes the result from an report, in which the sentiment of SP 500 has already be collected and categorized. Using Finbert, a language processed model to process those news, we gain a positive, neutral, negative sentiment score of publicity. These sentiment features are intended to reflect the market's tone – optimism or fear – which can lead or lag price movements. By including them, we allow our model to account for the information content of news that might drive investor behavior and thus index returns [6].

For macroeconomic indicators: We collected a set of macroeconomic time series from the Federal Reserve Economic Data (FRED) repository. These include:

- GDP Growth (GDP)
- Unemployment Rate (UNRATE)
- Non-farm Payrolls (PAYEMS)
- Industrial Production (INDPRO)
- Money Supply (M2SL)
- USD-EUR Exchange Rate (DEXUSEU)
- Crude Oil Price (DCOILWTICO)

These macro series come at mixed frequencies (daily for some financial series like exchange rate and oil, monthly or quarterly for others). We aligned them to the daily index data by forward-filling the most recent value until an update occurs. While this introduces some staleness for low-frequency indicators, it allows the model to at least hold the last known economic conditions as features on each day. We recognize the limitation that, for example, the unemployment rate feature remains constant throughout a given month, but we mitigate this by engineering lagged versions of these features (described below) to capture any delayed effects on the stock market after their release.

2.4. Feature processing and selection

After getting the raw data of all the materials, the essay process the features involved the 3 indicators. First of all, the Return Horizons are determined which includes 1-day, 10-day and 30-day ahead return. Specifically, for each day t, we define:

DOI: 10.54254/2754-1169/2025.BJ30479

• target-1d: the log return from day t to day t+1 (next trading day).
• target-10d: the log return from day t to day t+10.
• target-30d: the log return from day t to day t+30.

These are the outcomes that our models aim to predict. Alongside, we derive binary direction labels: e.g., direction-1d is 1 if target-1d > 0 (index went up the next day) or 0 if target-1d 0 (went down or stayed flat), and similarly for 10d and 30d. These direction labels are used for evaluating classification accuracy of the model's predictions (up vs down). It's important to note that for the last few days of our dataset, we cannot compute a 30-day future return (due to data ending in July 2023); those are left blank and not used in training or evaluation.

Secondly, the model integrate lagged features into all the features.To capture the possibility that certain indicators influence the market with a delay, we augmented the feature set with lagged versions of key macroeconomic and sentiment features. For each macro and sentiment feature (denoted X), we created lagged features X(t-1), X(t-2), X(t-3), and X(t-5) representing the value of that feature 1, 2, 3, and 5 trading days prior. The inclusion of lags allows the model to learn relationships where, for example, a spike in news sentiment might not impact returns immediately on the same day but could have a measurable effect a few days later. It also helps align monthly macro releases with daily market movement – e.g., the unemployment rate announced on the first Friday of a month might influence the market not just on that day but in subsequent days as investors digest the news. By lagging, we ensure our model can utilize such delayed effects. After adding lagged features, our feature space expanded significantly (each original macro/sentiment feature contributes up to 4 additional lagged features).

Moreover, the model employs Granger Causality Feature Selection to avoid useless features considered which might negatively impact the accuracy of our model: we applied a filtering step to reduce dimensionality and exclude features that were not predictive. In particular, the essay tried Granger causality tests on the training data for each candidate feature to determine that whether past values of that feature had a statistically significant ability to forecast future results. Therefore, only features that passed these tests were retained and included for model training.

Finally, the essay normalizes the features: all continuous features were standardized appropriately before modeling. For features used in the Ridge Regression, we applied z-scores standardization method using the training set statistics, which is subtracting the mean and dividing by the standard deviation of each feature.

The LSTM model was also standardized directly by StandardScaler. Therefore, this ensures that the optimization of both models is well-behaved and not dominated by large-magnitude variables. Importantly, scaling parameters (means and stds) were computed on training data and applied to test data in a forward-looking manner to avoid any lookahead bias in normalization.

Through these steps, we prepared a comprehensive and refined feature matrix for the regression model and a normalized price sequence for the LSTM. Crucially, we took care to avoid data leakage: all feature engineering for a given day uses only data from that day or earlier, and the model training process (described next) uses a strict train-test split in chronological order to simulate real-time prediction without peeking into future data.

3. Model upgrade

3.1. Model selection

Now the three indicators should be processed by particular model for an output, and essay select six models for comparison: XG Boost, Light GBM, Random Forest, Liner regression, Support Vector

DOI: 10.54254/2754-1169/2025.BJ30479

Regression, Ridge regression.

According to Figure 2, by testing the performance of differentiated models, we conclude that ridge regression outperforms any other model, demonstrating its appropriate ability for processing the three indicators.

Moreover, because the R^2 of technical indicators in 1-day prediction is negative with a significantly low, the technical indicators are excluded in this horizon, while the other two horizons remain the same.

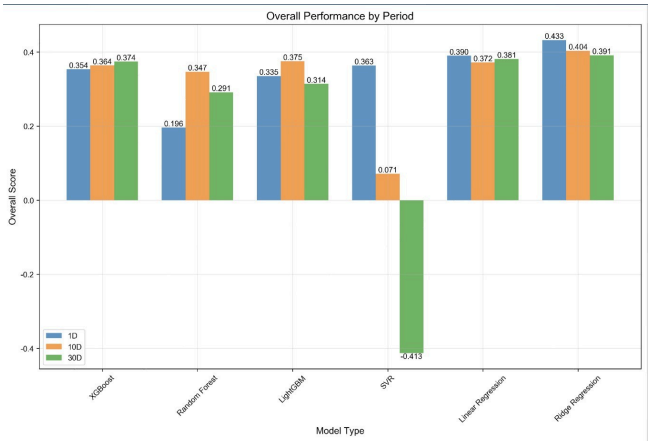

Figure 2. Distinct model comparision

3.2. Integrating models

Same as the process of model chosen for indicators, the stacking model also test several models including Simple Average, Weighted Average, Accuracy-Weighted, Median Ensemble, Best Model, Majority Voting, Random forest and gradient Boosting. The performance of the stacking models are demonstrated in the following diagram:

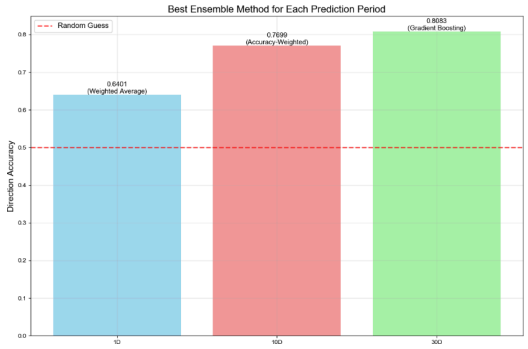

Figure 3. Stacking models comparision

The diagrams demonstrate that for different horizions, there should be different stacking models: utilizing Weighted average for 1-day prediction, Accuracy-Weighted for 10-day prediction and Gradient Boosting for 30-day prediction.

4. Results evaluation

We now present the performance results of our models on the test dataset, broken down by prediction horizon (1-day, 10-day, 30-day). We compare the Ridge regression, LSTM, and Ensemble

DOI: 10.54254/2754-1169/2025.BJ30479

for each horizon to assess the value added by each component and the ensemble combination. The key evaluation metrics are R^2 (for return magnitude prediction) and directional accuracy (for predicting the sign of the return). Table 4 summarizes these results, and we discuss them in detail below.

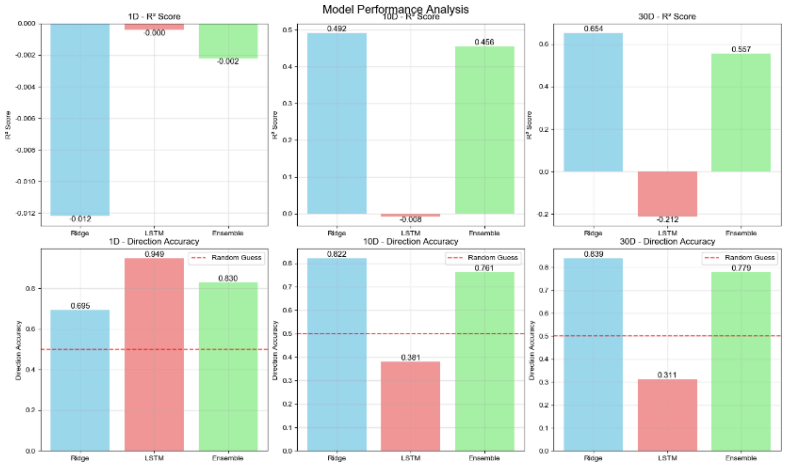

Figure 4. R^2 and directional accuracy for each model

Predictability clearly increases with horizon in our findings. The 1-day horizon was essentially unpredictable in magnitude (R^2 0), though we managed to get direction right quite often (which might partly be attributed to the test period's characteristics). The 10-day and 30-day horizons had substantial predictable components, with our model achieving $R^2 = 0.45$ and 0.56 respectively for those horizons, and high directional accuracies (76% and 78%). These results affirm the benefit of integrating macroeconomic and sentiment information: a pure price-based model (LSTM alone) would not reach these R^2 levels; by adding the extra features via Ridge and combining, we greatly improved performance. It also validates our adaptive ensemble strategy: each horizon's best ensemble method yielded superior results relative to any single approach.

5. Conclusion

5.1. Limitations

• Macroeconomic Data Frequency. Macroeconomic indicators are low-frequency and were forward-filled to match daily returns, which may introduce staleness. While lag features partly mitigate this, static values are unlikely to aid very short-term (1-day) forecasts. Future work may use higher-frequency proxies such as daily bond yields.

• Limited Data and Overfitting. The dataset covers only 3.3 years (850 trading days), a period marked by extraordinary events (COVID, monetary stimulus, inflation). Models may have learned regime-specific patterns, limiting generalizability. Regularization reduced risk, but further validation on longer samples is needed.

• Sentiment Coverage. We relied solely on news-based FinBERT sentiment. Social media sentiment and more granular distinctions (e.g., macro vs earnings news) were not included, potentially limiting predictive power at short horizons.

DOI: 10.54254/2754-1169/2025.BJ30479

5.2. Future improvement

• Extend the framework, with the ultimate goal of developing a prediction mo=

• Adjust the hyperparameter of distinct models to further enhance predictive accuracy and stability.

• Use higher-frequency or more informative macro inputs for short-term predictions

• Expand sentiment sources to social media for a more comprehensive sentiment measure.

5.3. Summarize

In this paper, we presented a comprehensive approach to forecast SP 500 returns over short, medium, and longer-term horizons by integrating macroeconomic indicators, financial news sentiment, and technical price patterns in a hybrid ensemble model. Our framework combined a Ridge regression model (to leverage a wide array of engineered features) with an LSTM neural network (to capture sequential price dynamics), and crucially, used an adaptive ensemble strategy tailored to each prediction horizon (1-day, 10-day, 30-day). This multi-tier architecture allowed us to exploit the strengths of different modeling techniques and data sources in a unified prediction system.

The essay demonstrates the feasibility of stock market prediction. While the stock market will always retain a degree of randomness and surprise, approaches like ours push the boundary of what can be forecasted, offering valuable tools for investors and researchers to understand and anticipate market movements. We envision that as more data (especially unstructured data) becomes available and machine learning techniques continue to advance, hybrid models such as presented here will become increasingly important in financial forecasting, providing better decision support while respecting the uncertainty that inherently remains.

References

[1] Wikipediacontributors, “2025StockMarketCrash, ”Wikipedia, TheFreeEncyclopedia, https: //en.wikipedia.org/wiki/2025stockmarketcrash, accessedSeptember 2, 2025.

[2] T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions, ” European Journal of Operational Research, vol. 270, no. 2, pp. 654–669, 2018. https: //doi.org/10.1016/j.ejor.2017.11.054, accessed October 10, 2025.

[3] Qingyun Zeng, Tingsong Jiang, “Financial Sentiment Analysis: An Empirical Comparison of BERT, LSTM, and Hybrid Models, ” arXiv preprint arXiv: 2306.02136, https: //arxiv.org/pdf/2306.02136, accessed September 6, 2025.

[4] N.-F. Chen, R. Roll, and S. A. Ross, “Economic forces and the stock market, ” Journal of Business, vol. 59, no. 3, pp. 383–403, 1986.

[5] Yahoo Finance, “Historical Market Data, ” https: //finance.yahoo.com/, accessed September 6, 2025.

[6] Federal Reserve Bank of St. Louis, “FRED Economic Data, ” https: //fred.stlouisfed.org/, accessed September 6, 2025.

DOI: 10.54254/2754-1169/2025.BJ30479

Appendix

Additional diagrams

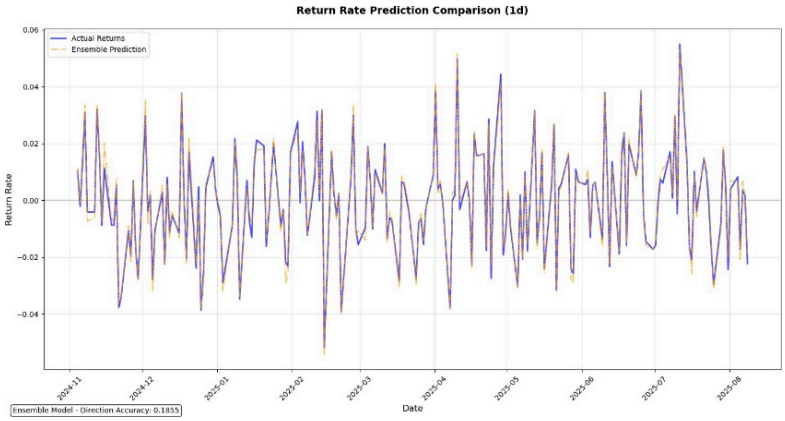

Figure 5. 1-day horizon results

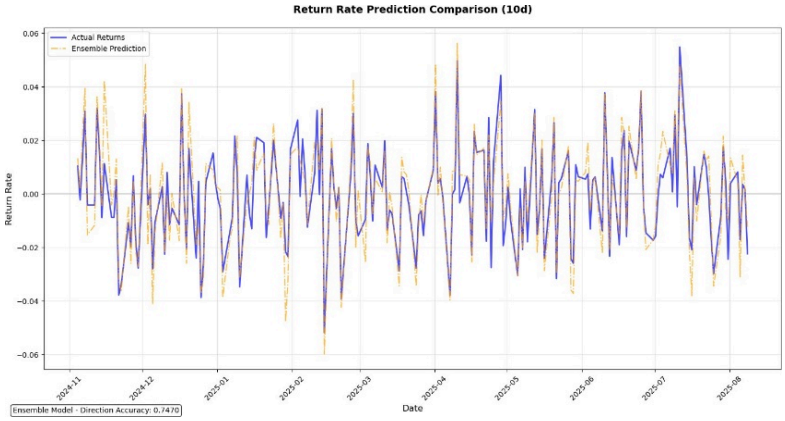

Figure 6. 10-dayhorizon results

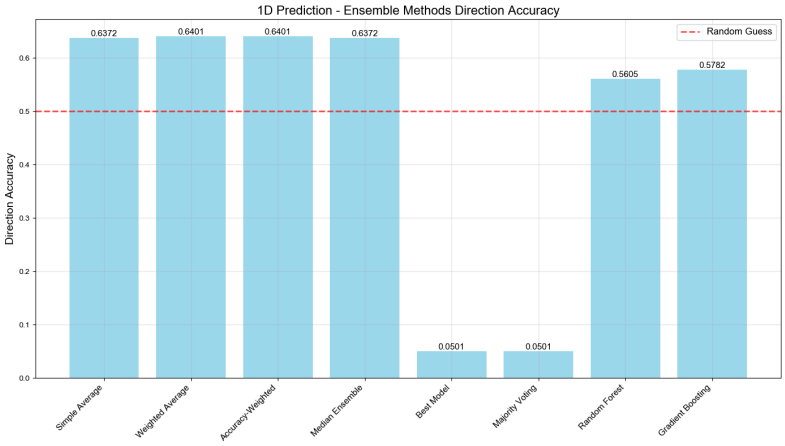

Figure 7. 1-day horizon distinct stacking models comparision

DOI: 10.54254/2754-1169/2025.BJ30479

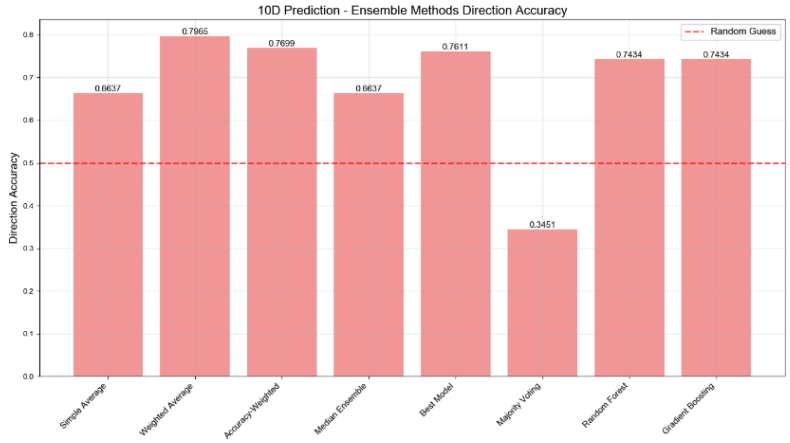

Figure 8. 10-day horizion distinct stacking models comparision

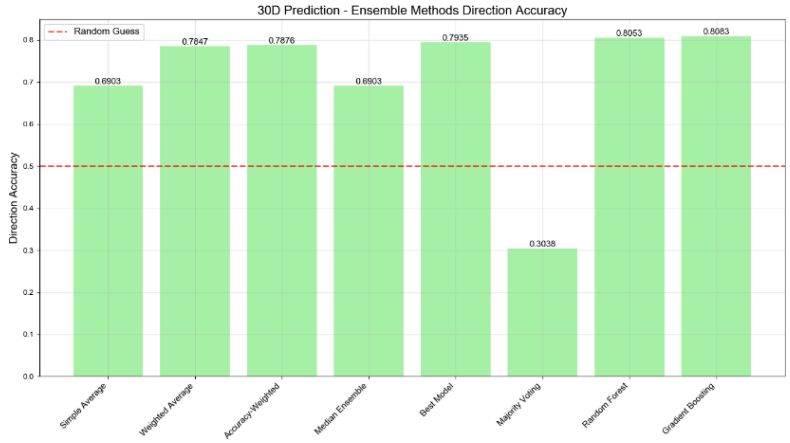

Figure 9. 30-day horizon distinct stacking models comparison

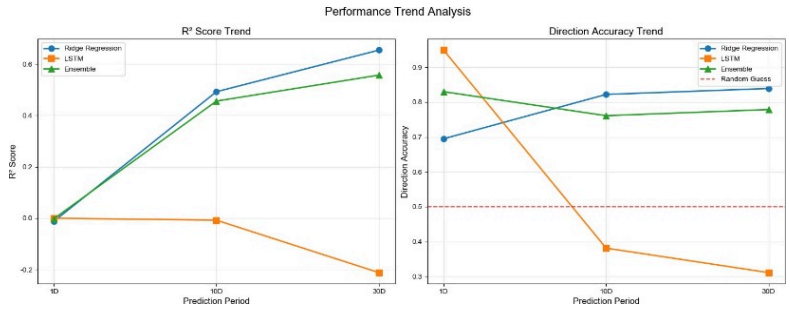

Figure 10. Performance trend analysis