Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026412420

Empirical Research on Daily Direction Prediction of S&P 500 ETF Based on Machine Learning and Multivariate Technical Indicators

Mingyu Qin*

University of Toronto, Toronto, Canada

**Corresponding author: Mingyu Qin.*

Abstract

This paper examines whether machine learning models can predict the next-day direction of SPY, an exchange-traded fund that tracks the S&P 500 Index. Using daily market data from 2010 to 2026, the study constructs 21 technical and cross-asset features, including returns, moving-average ratios, volatility, momentum, RSI, MACD, volume changes, and related market signals. Four supervised learning models are compared: Logistic Regression, Random Forest, XGBoost, and Support Vector Machine. The results show that XGBoost achieves the highest test accuracy among the tested models, with an accuracy of 0.5360 and a ROC-AUC of 0.5393. However, the model does not outperform the majority-class baseline accuracy of 0.5690. These findings suggest that technical indicators contain limited predictive information for next-day SPY direction, but their business value is weak as a standalone trading strategy. The model is more appropriate as a market monitoring or risk-awareness tool than as an automatic investment system.

Keywords

SPY ETF, machine learning, technical indicators, market direction prediction, time-series forecasting

1. Introduction

Predicting the short-term movement of the equity market is an important problem in finance. If traders can predict whether the market will go up or down in the next day, they would be able to make better trade decisions, portfolio construction, and manage risk effectively. However, this is still not easy because financial markets are noisy, competitive, and vulnerable to many unpredictable factors. Daily fluctuations in the financial market depend on various aspects like news, sentiment of traders, interest rate expectations, earnings of firms, and more. The recent academic literature reveals that machine learning has been increasingly used in stock market prediction; however, the performance of such models depends greatly on data, purpose, methodology, and market condition [1].

This paper focuses on SPY, an exchange-traded fund that tracks the performance of the S&P 500 Index. The value of SPY makes an excellent subject of analysis because of the instrument's liquidity, transaction frequency, and widespread adoption as an indicator of the U.S. stock market overall. Instead of trying to predict

SPY's precise price level, the paper proposes formulating the research question in terms of classification. The purpose of this classification task is to determine whether SPY will trade at a higher level the next trading day. If yes, then the target variable equals 1; otherwise, it equals 0.

The core research question to be investigated is: Are there possibilities for predicting the directional moves of SPY using machine learning based on technical indicators and their market counterparts on a day-to-day basis? This research is important from both technical and business points of view. On the technical side, this is about determining whether there is any predictive value in past price and volume behavior.

2. Literature Review

Predictions in the stock market have been thoroughly researched in both the fields of finance and machine learning. The hypothesis of technical analysis suggests that there could be some information about future movement of the market in the form of price or volume history. The indicators such as moving average, momentum, RSI, MACD, and volatility are used to evaluate trends' strength, market pressure and direction changes. Although the indicators will never produce exact price predictions, they give a good structure to describe market movement[3]. Research about technical indicators shows that the choice of indicators affects the prediction performance because different indicators provide unique information about market movement, like trend, momentum, volatility and trading activity [3].

In recent years, machine learning algorithms have been successfully applied for predicting stock market. As compared to traditional linear models, the machine learning models are able to estimate nonlinear relationships and dependencies between variables. A literature review shows that many researchers examined the use of models such as Random Forest, Support Vector Machine and gradient-boosting algorithms for prediction of stock markets [1]. However, past researches demonstrate that the performance of the models heavily depends on data set, prediction horizon, input variables and evaluation methods. It should be noted that the data about the financial market include noises and are non-stationary, and therefore short-term prediction faces specific challenges [8].

Another stream of research focuses on wider information about market beyond the target stock or index. The variables from other assets, such as bond returns, commodity prices, currency rates, and volatility indices may give additional information about investors' preferences and macro-financial environment [6]. For example, the changes in VIX may show fear on the market, while bond and gold prices may show that the investors try to protect their money. At the same time, the value of this information cannot be evaluated, since public information becomes quickly incorporated into market prices.

In continuation of these streams of literature, this paper develops new contribution by incorporating SPY technical indicators with cross-asset return variables and by testing several machine learning models under the train/test temporal split. Unlike the previous papers, which focus on the accuracy of models, the paper also considers the performance of the models against the majority class classifier. This is important for the research since SPY shows more up-days than the down-days or flat days.

3. Methodology

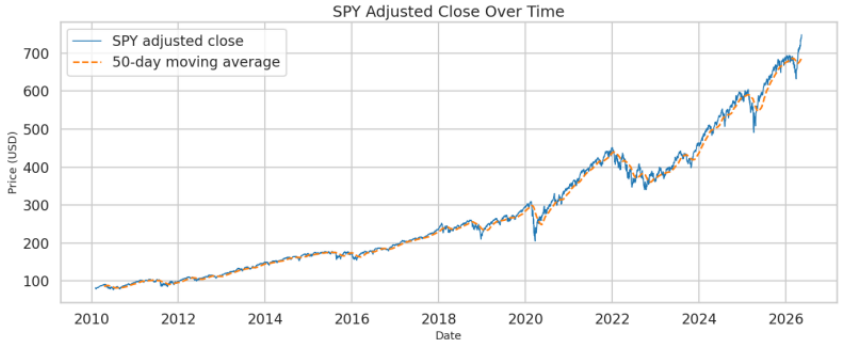
3.1 Data Source

This dataset consists of historical market data for each day from 2010 through the available portion of 2026, obtained from Yahoo Finance [2]. Given that 2026 is not a complete calendar year, only the trading data until the data collection date is included in the data set. The inclusion of data for partial 2026 maintains continuity in the latest series. The main market instrument analyzed in this project is SPY, which is used to approximate the overall U.S. stock market performance. The initial set of variables for SPY includes open price, maximum, minimum, closing prices, adjusted closing price, and volume. Besides SPY, the dataset contains other market instruments including QQQ, IWM, TLT, GLD, USO, UUP, and VIX. These financial instruments help provide additional information about the overall market performance.

QQQ is used to approximate technology-dominated U.S. stocks, whereas IWM corresponds to small capitalization stocks. TLT approximates the performance of long-term U.S. government bonds. GLD stands

for gold, USO refers to oil, and UUP estimates the performance of the U.S. dollar. Finally, VIX is used as an index of implied market volatility. All these variables allow considering not only the historical dynamics of SPY but also other financial instruments' signals. Stock market forecasting research indicates that the selection of inputs plays an important role because different market indicators capture various information. [1, 3]

Figure 1: SPY adjusted closing price from 2010 to 2026



As depicted in Figure 1, SPY showed a positive trend between 2010 and 2026 but had a non-linear path. The chart shows growth periods, declines, and high volatility. Therefore, it is crucial to assess the prediction model under different market scenarios and avoid assuming a consistent market scenario when developing the model.

For a business, the use of cross-asset information makes sense, considering that investors do not analyze the S&P 500 alone. Stock markets are associated with interest rates, commodity prices, exchange rates, and implied volatility. For example, a sharp increase in VIX may suggest that investors are more fearful, while bond prices or gold prices may reveal changes in investors' risk attitudes. Thus, cross-asset information helps predict SPY's future performance.

3.2 Variable Construction

The dependent variable that is analyzed within this research project is the movement of SPY for the next day. This measure can be calculated using the SPY next day return. If the next day's adjusted close is higher than the current day's adjusted close, then next_day_up is equal to 1, otherwise, next_day_up equals 0. Thus, the issue under analysis turns out to be a supervised binary classification problem.

The independent variables are generated based on the underlying market data. There are 21 features in the final model, such as SPY daily return, intraday return, trading range, volume change, moving average ratios, rolling volatility, momentum, RSI, and MACD. Additionally, cross-asset returns for QQQ, IWM, TLT, GLD, USO, UUP, and VIX are used in the model.

This selection of variables is made with the intention of representing different categories of information about the market that exist conventionally. Variables like moving averages and momentum are used for capturing trend behavior. Volatility is measured using rolling volatility to capture market uncertainty. RSI and MACD act as technical indicators often used to measure momentum and directional movements. Variable measuring the change in volume represents volume variation. Returns of assets other than stocks will provide an idea of the overall state of the market. Technical indicator-based stock prediction using machine learning algorithms is a recent field of research [3].

Raw price levels cannot be used directly as inputs into the model because of the lack of generalizability over time. For example, the price level of SPY in 2010 is entirely different from the price level of SPY in 2026. On the other hand, features like return, ratio, volatility, and momentum are easier to compare across various periods.

3.3 Model Design

These models are evaluated using Logistic Regression, Random Forest, XGBoost, and Support Vector Machine. Logistic Regression, Random Forest, and Support Vector Machine have been coded with the help of a very popular machine learning library called scikit-learn in Python [4]. The inclusion of XGBoost is because

of the fact that it can be used for scaling the gradient boosting process, which involves predictions on structured data [5].

The Logistic Regression model is used as a simple and understandable baseline method. This gives us a traditional linear classifier approach. Random Forest and XGBoost are tree-based ensemble approaches that can handle non-linear interactions between the variables. Moreover, the Support Vector Machine is considered as well since it is more flexible in defining the decision boundary between the classes. The use of tree-based methods like Random Forest and XGBoost can be seen in some recent studies regarding financial machine learning problems [6].

Model performance will be evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC score. Accuracy stands for the share of correct predictions. Precision refers to the share of positive predictions among all the predicted ups. Recall means the share of actual positives among all actual ups. The F1-score represents the balance between precision and recall. ROC-AUC measures discrimination power at different classification thresholds.

An important methodology decision lies in the split between the train and test datasets. Instead of a random split, a time-split approach will be used in the present study. The first 80% of observations will be utilized as a training dataset, while the last 20% will be used as a test dataset. It is extremely important not to use any future information when making predictions about the past. Time splitting is appropriate since financial forecasting assumes the use of past data to predict future events. Look-ahead bias minimization is especially important for machine learning tasks [7].

This project uses time-series cross-validation instead of the shuffling method of cross-validation. Time-series cross-validation ensures that time order in the data set is not violated and avoids look-ahead bias. It is important to consider this when dealing with financial data since financial information is time-dependent; that is, future observations should not be randomly interspersed among past observations.

3.4 FEATURE Interpretability & Importance Analysis

Given that 2026 is not a complete calendar year, only the trading data until the data collection date is included in the data set. The inclusion of data for partial 2026 maintains continuity in the latest series.

Aside from performance comparison between models, this study seeks to find out what indicators carry significant weight in prediction. It is necessary since the purpose of this work is not only to establish the optimal model but also to establish if certain types of market information bring extra predictive power for tomorrow's direction of SPY.

There are two methods that can help to analyze feature importance and interpretability. Correlation analysis used during the exploratory data analysis can be utilized in order to understand if any of the features are linearly related to the target variable. Another method is to use feature importance graph based on Random Forest in order to see what variables are preferentially chosen by the algorithm when creating tree partitions. Random Forest Feature importance is useful because it can detect nonlinear relationship and feature interaction. Nevertheless, it does not provide causal proof.

According to the obtained results, some of the most informative features include those that refer to volatility and recent return. VIX return, SPY recent return, ETF returns such as QQQ and IWM and short term technical indicators such as momentum and rolling volatility seem to be more informative than most other features. It is economically plausible because VIX measures changes in fear and expectations for future volatility in the market, SPY, QQQ and IWM measure recent equity market momentum and risk appetite sentiments. Momentum and rolling volatility measure the trendiness of the market or instability.

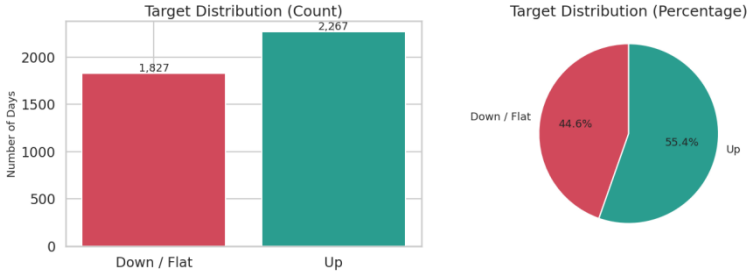
Nevertheless, the obtained results also prove that there is no single indicator that has enough signal to predict the next day SPY direction. The model needs a number of weak signals instead of a strong one. It is consistent with the main claim of the research: technical and cross asset indicators possess marginal information that is not enough in order to beat the majority class baseline. Consequently, feature importance analysis does not change anything in the interpretation of the results of the research.

4. Results and Discussion

From the exploratory analysis results, one can conclude that the time series under study demonstrates a positive trend during 2010 – 2026. However, this does not mean that the trajectory is stable: there were several instances of strong growth, severe decline, and volatility. It influences the development of a predictive model since it means that the model will have to take into account different market regimes. The patterns observed in a calm market are unlikely to work when market volatility grows.

In addition, it was found out that the target variable in the SPY dataset is unbalanced since more up days than down or sideways days occur during the test period. A simple classifier that assumes all predictions to belong to the largest class achieves accuracy of 0.5690. As a result, a machine learning model should be assessed not only against 50% chance performance.

Figure 2: Distribution of next-day SPY direction



As seen from Figure 2, the dependent variable shows slight class imbalance, with SPY having more days with upward movement compared to downward and no movement days. Hence, the baseline for the majority class becomes an important point of reference. Without this benchmark, the model would appear effective just because there are more days of increase than decrease in the market.

In the list of considered models, XGBoost performs the best during testing, achieving an accuracy of 0.5360 and ROC-AUC of 0.5393. Performance metrics of Random Forest, SVM, and Logistic Regression do not differ substantially, being somewhat inferior to those of the best-performing model but still showing relatively similar performance. This indicates that all four models perform slightly above the random accuracy baseline, but their predictive power remains weak and none of them outperforms the majority-class baseline.

Table 1: Final model performance comparison

Model	Accuracy	Time-Series CV Accuracy	Precision	Recall	F1-score	ROC-AUC
XGBoost	0.5360	0.5075	0.6108	0.5086	0.5550	0.5393
SVM	0.5189	0.5057	0.5857	0.5279	0.5553	0.5076
Random Forest	0.5177	0.5024	0.5757	0.5794	0.5775	0.5253
Logistic Regression	0.5177	0.4914	0.5760	0.5773	0.5766	0.5112
Majority-Class Baseline	0.5690	—	—	—	—	—

Table 1 below contains the performance metrics of four machine learning models in terms of accuracy, time-series cross-validated accuracy, precision, recall, F1-score, and ROC-AUC. Out of four models, XGBoost is the one that has the best test accuracy (0.5360) and ROC-AUC (0.5393). Nevertheless, its time-series cross-validation accuracy is equal to 0.5075; therefore, this model can be characterized by instability of the predictive performance across different temporal segments.

Moreover, precision and recall allow us to learn about something more than accuracy does. The best precision is achieved by XGBoost (0.6108); it means that if the algorithm claims that it is going to be an up day, it is quite likely that it will really happen. Nevertheless, recall is equal to 0.5086, so the algorithm fails to recognize all-up days. Both Random Forest and Logistic Regression models have somewhat better recall and F1-scores, but their accuracy and ROC-AUC are lower in comparison with others.

F1-scores of four algorithms under consideration also have similar values and vary from 0.555 to 0.578, which means that there is no model that shows some advantages regarding this aspect. All ROC-AUCs have rather small values and are close to 0.5, which indicates weak discriminative capability to recognize up days. What is important, majority class baseline gives accuracy equal to 0.5690, which is higher than for all machine

learning models. Therefore, it is possible to say that models detect some weak signals from the market but not more than that.

Table 1 shows that XGBoost achieves the best result in terms of accuracy compared to other machine learning approaches used. Yet, the baseline based on the class of the majority continues to outperform all the approaches when considering accuracy. The finding reveals its importance in interpreting the results, as the algorithms can detect the weak market signal but still fail to produce a sufficient level of predictive power to overcome a simple baseline.

From the economic point of view, it means that technical indicators as well as cross-asset signals carry additional marginal information. However, such information is still not enough to build a reliable short-run forecast. The finding corresponds to the specifics of financial markets, as daily stock-price changes are characterized by high levels of noise and competitiveness. Market information becomes incorporated quickly into stock prices, which makes it difficult to use only technical indicators in order to create a reliable forecast. Finally, an analysis of the literature related to stock market prediction demonstrates that machine learning approaches show varying results [1].

Figure 3: Model accuracy comparison with random and majority-class baselines

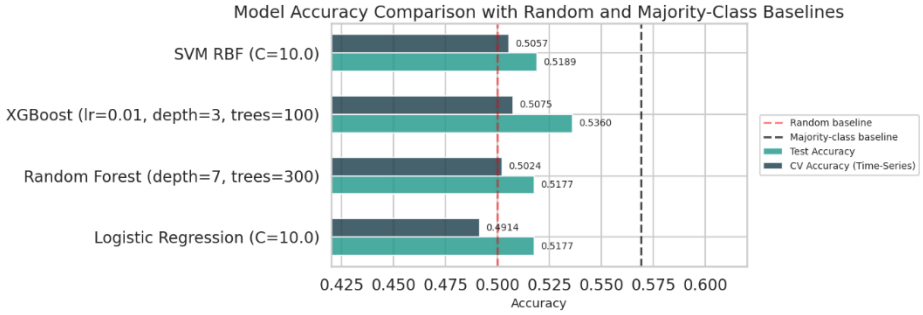


Figure 3 provides a comparison of the models in relation to both the random and majority-class baselines. While the machine learning models perform at least as well as the random baseline, they do not outperform the majority-class baseline. Consequently, the models lack practical applications as independent trading models.

The baseline comparison is especially important from a business perspective. If the model reaches approximately 53.6% accuracy, it seems like a decent result compared to random predictions; however, since a "predict-up" strategy achieves 56.9% accuracy, the machine learning model fails to provide enough value. This doesn't mean the model is useless; it just can't be used as a trading strategy.

This example demonstrates the necessity of evaluating models carefully. While the model can seem beneficial when compared only to the 50% accuracy of random guesses, it should be noted that the stock market experiences positive drift over time. As such, the majority-class baseline is quite powerful. Thus, while the model might outperform random predictions, the question becomes whether it can overcome realistic competitors.

As far as a business perspective is concerned, the study highlights both strengths and weaknesses of machine learning in making financial decisions. In other words, the model might still be valuable when used in conjunction with other tools, e.g., as a way of checking the state of the market, comparing various risk indicators, and so on. However, it should not be considered an independent trading algorithm. Before using the model in actual trading, investors need to account for factors such as transaction fees, slippage, tax implications, position sizes, and risk-adjusted returns. According to recent studies on machine learning for the S&P 500, it is also important to evaluate the model from an investment perspective, i.e., how well it performs in terms of portfolios and risks, rather than just classification accuracy [6].

Moreover, one additional aspect to be noted here is that all the factors used by the model belong only to the market side. The model does not incorporate macroeconomic factors, FED announcements, earnings reports, and even sentiments from financial news. Recent studies have found that financial news sentiments, together with implied volatility and trading volume, can contribute additional information for market classification [8]. Therefore, the current model can be viewed as a preliminary test rather than a full-fledged investing strategy.

5. Business Implications

The main implication for practice is the proper usage of machine learning for market prediction purposes. According to the experiment outcomes, the technical indicators do not give enough predictive power to predict future market moves based on the SPY index. From all models under investigation, XGBoost shows the best results; nonetheless, even in this case its accuracy does not surpass the accuracy of the majority-class baseline. As one can see, machine learning does not give enough predictive power to make a trading decision automatically.

The model can be applied in practice in two different cases: institutional investment and individual trading. As far as the case of institutional investor is concerned, then the model can be used as an additional tool for monitoring the market. The tool can help investors to assess the current technical situation on the market, changes in volatility and signals from other assets before the portfolio decision making process. If the model gives poor signal about the market, the portfolio manager can lower the risk, postpone the portfolio adjustments or compare the model output with other macroeconomic and fundamental indicators. In this respect, the model is not supposed to replace professionals' decision-making but to assist in assessing risks and researching portfolios.

Individual investors face limited possibilities. The model is not supposed to be used as a trading signal since it does not surpass the majority-class baseline in accuracy. Nonetheless, the model can be used by the individuals to understand technical indicators, to compare momentum and volatility signals and not to rely only on intuition. Without considering transaction cost, stop losses, position sizes and other risk management procedures, the usage of the model for trading will be very risky.

Moreover, the model itself can contribute to enhancing market awareness. Though it does not predict market movements successfully, it can help organize market data. Instead of using only intuition for predicting market direction, an investor can use various engineered indicators for summarizing such factors as trends, volatility, momentum, volume, and correlation between assets.

Thus, it is necessary to underline that the practical importance of the proposed models largely depends on implementation practices. In particular, a relatively modest model does not provide much value for trading directly but can be applied in scientific research and monitoring processes.

6. Limitations

The current study involves certain limitations at data, model, and evaluation levels. At data level, the limitation refers to the absence of macroeconomic indicators such as inflation, unemployment rate, interest rates, etc., as well as to the absence of company-specific news like earnings announcements, analyst opinions, press releases of Federal Reserve System, sentiment analysis of news about financial market in general, etc. Although the current data includes price trend, volatility, momentum, trading activity, and general market movement, there is no information regarding how SPY may behave in response to certain sentiment or volatility change [8]. Thus, the current data set is likely to be incomplete and may miss relevant information that affects the movement of SPY in the short term.

Model-level limitation relates to evaluating just four supervised learning algorithms – Logistic Regression, Random Forest, XGBoost, and Support Vector Machine. Such a choice of models allows comparing linear regression, decision trees-based algorithm, boosting technique, and margin-based classifier; however, more advanced models (e.g., recurrent neural networks or LSTM) can potentially provide better forecasts of nonstationary processes. Financial markets are nonstationary, which means that a model built on a certain period of time might be less accurate on another one due to changes in monetary policy, inflation, market crisis, investor behavior, etc.

The limitation in the evaluation level is that it relies on classification accuracy, precision, recall, F1-score, and ROC-AUC to measure the forecasting ability of each model. Although these measures provide valuable information on which algorithm is better at classification, they do not show any information about trading ability of the model – even if the model shows good results in classification, its use during highly volatile periods may lead to losses. Moreover, no backtest is carried out in the current research.

As for the prediction target, it equals daily SPY market direction. However, daily return is quite volatile and can be affected by unexpected news or noise, so weekly or monthly prediction would be more stable and beneficial.

7. Conclusion

The present study aims to explore whether ML models are capable of predicting the next day direction of SPY based on technical indicators and market-related signals. Specifically, this paper examines 21 engineered features created from SPY and corresponding market assets using daily data from 2010 through 2026. There are four supervised learning models used for analysis: Logistic Regression, Random Forest, XGBoost, and Support Vector Machine.

It can be concluded that forecasting the next day's direction of SPY is difficult. The results show that XGBoost demonstrates the most accurate performance by producing a test accuracy score of 0.5360 with ROC-AUC equal to 0.5393. However, it should be noted that the baseline for majority class achieves an accuracy level of 0.5690, which means that even though the models perform above the random baseline, they do not outperform the simple majority-class baseline.

From the business point of view, these outcomes can be interpreted as follows. First of all, the model cannot be viewed as a trading strategy. From this standpoint, it seems likely that technical indicators and cross-market signals contain only weak predictive value and cannot be used as an independent trading strategy. Thus, a potential application of the model may consist in its utilization as a support for the decisions made by analysts during market surveillance, risk management, or portfolio research.

To conclude, the current study supports a skeptical attitude towards using machine learning in finance. Even though it allows for organizing data and comparing signals, it does not guarantee successful pattern discovery and trading. To implement the model in practice, better predictive performance, a more informative data source, and costs and risks analysis are required.

References

[1] Kumbure, M. M., Lohrmann, C., Luukka, P., and Porras, J. "Machine Learning Techniques and Data for Stock Market Forecasting: A Literature Review." Expert Systems with Applications, vol. 197, 2022, Article 116659.

[2] Yahoo Finance. "Historical Daily Market Data for SPY, QQQ, IWM, TLT, GLD, USO, UUP, and VIX." Accessed 2026.

[3] Mostafavi, S. M., et al. "Key Technical Indicators for Stock Market Prediction." Intelligent Systems with Applications, 2025.

[4] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[5] Chen, T., and Guestrin, C. "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[6] Caparrini López, A., Arroyo, J., and Escayola Mansilla, J. "S&P 500 Stock Selection Using Machine Learning Classifiers: A Look into the Changing Role of Factors." Research in International Business and Finance, vol. 70, 2024, Article 102336.

[7] Lopez de Prado, M. Advances in Financial Machine Learning. Wiley, 2018.

[8] Davidovic, M., and McCleary, J. "News Sentiment and Stock Market Dynamics: A Machine Learning Investigation." Journal of Risk and Financial Management, vol. 18, no. 8, 2025, Article 412.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.