Predicting Stock Returns Using Machine Learning: A Hybrid Approach with LightGBM, XGBoost, and Portfolio Optimization

Xinrong Yan

Mathematics, New York University, NY 10012, New York, USA
xy2378@nyu.edu

Abstract. Predicting stock returns presents a multifaceted challenge in the financial market, as it is influenced by a myriad of variables, including economic indicators, market sentiment, and company performance. Traditional predictive models often fall short in capturing the complexities and dynamics of contemporary markets. As a result, there has been a notable shift towards machine learning approaches, which offer advanced techniques for analyzing vast datasets. This study explores the application of two leading machine-learning models, LightGBM and XGBoost, in forecasting the future performance of S&P 500 stocks. By integrating both fundamental and technical data—such as price trends, financial metrics, and broader market conditions—these models aim to deliver accurate predictions. The insights generated can significantly inform investment strategies, enabling investors to identify favorable long and short-term positions, thereby optimizing their portfolio performance in a highly competitive and rapidly changing financial landscape. This research contributes to the evolving field of quantitative finance, highlighting the potential of machine learning in investment decision-making.

Keywords: Stock Return Prediction, Machine Learning, Finance.

1 Introduction

The prediction of stock returns is a fundamental challenge in the financial market, which is significant for investors, portfolio managers, and policymakers, as it offers investors and portfolio managers the chance to make informed decisions, optimize investment strategies, and manage risks effectively. Previously, the stock return prediction has relied on statistical models and financial theories to evaluate the company's value and estimate its future stock variation. However, the financial markets are becoming more complex and influenced by multiple factors, and the traditional models are not effective sometimes. Therefore, more advanced methods should be considered to address these limitations. There is a new method called machine learning has appeared as a powerful tool to predict stock prices more accurately. Machine learning can process a lot of historical data, uncover complex

J. Lu (ed.), *Proceedings of the International Workshop on Navigating the Digital Business Frontier for Sustainable Financial Innovation (ICDEBA 2024)*, Advances in Economics, Business and Management Research 315,
https://doi.org/10.2991/978-94-6463-652-9_94

non-linear relationships in the stock, and continuously learn new knowledge.

Nowadays, machine learning has been applied to learning models such as gradient boosting and deep learning algorithms in financial time series prediction. The research shows that the advantage of machine learning models can be superior to traditional methods [1-3], especially in predicting short-term price stocks. There are some models like XGBoost and LightGBM that achieved more attention for their power to get structured data, form accurate prediction, and provide long-term insight.

The rise of machine learning, especially ensemble learning models like XGBoost and LightGBM, has shown a marked improvement in predictive power for financial time series. Chen and Guestrin introduced XGBoost as a highly efficient and scalable gradient-boosting framework [4], which quickly gained popularity in structured data prediction tasks, including stock price forecasting. In parallel, LightGBM, introduced by Microsoft in 2017, optimizes gradient boosting by using techniques like histogram-based splitting and leaf-wise growth, making it faster and more efficient than traditional gradient boosting models.

Studies such as Fischer and Krauss used deep learning architectures, such as Long Short-Term Memory (LSTM) networks [5], for stock price prediction, showing the importance of temporal dependencies in financial time series. However, models like XGBoost and LightGBM, while not specialized in sequence learning, can achieve competitive results by effectively handling tabular data and combining both fundamental and technical factors.

This study aims to combine both fundamental and technical factors in predicting stock returns by utilizing machine learning models, specifically LightGBM and XGBoost. This paper focuses on predicting returns for the constituent stocks of the S&P 500 index, one of the most widely followed stock market indices globally, comprising 500 leading companies in various industries. By integrating fundamental analysis (which evaluates a company's financial health) with technical analysis (which focuses on price movements), this study aims to build a robust model capable of predicting whether to take a long or short position on a stock. The final output from these predictions will be used to construct an optimized portfolio of 30 stocks.

2 Method

2.1 Dataset Preparation

The dataset used for this analysis comprises historical stock prices and financial data from companies listed in the S&P 500 index, obtained from Yahoo Finance and Bloomberg. The data spans from January 2021 to July 2024, ensuring the inclusion of various market conditions, such as bull and bear markets, to provide robust training for the models.

This dataset includes major corporations like Apple (AAPL), Microsoft (MSFT), and Amazon (AMZN). The target variable for this study is the future monthly stock return, calculated as the percentage change in adjusted closing prices over a one-month period.

2.2 Stock Price Prediction

LightGBM and XGBoost were selected for their ability to handle structured datasets with numerous features [6-9]. XGBoost employs a sequential decision tree construction method, where each tree refines the errors made by previous trees. This iterative process helps the model minimize errors and avoid overfitting, particularly useful when dealing with financial data, which is often noisy.

LightGBM, on the other hand, uses a histogram-based approach and a leaf-wise tree growth strategy, allowing for faster training, especially with large datasets. This model's ability to handle high-dimensional data makes it ideal for combining fundamental and technical factors. Both models were tuned using cross-validation to optimize key parameters such as learning rate, tree depth, and the number of boosting iterations. Early stopping was employed to prevent overfitting during training.

2.3 Portfolio Optimization

Once the stock return predictions were generated, a portfolio of 30 stocks was constructed using mean-variance optimization [10, 11], a widely adopted approach in modern portfolio theory. This method aims to maximize expected returns for a given level of risk, quantified through the Sharpe ratio, which evaluates risk-adjusted performance.

The predicted returns from both models served as inputs for expected returns, while the covariance matrix of stock returns accounted for correlations between different assets. By selecting stocks with low correlations, the portfolio could achieve better diversification, reducing overall risk.

3 Results and Discussion

3.1 Predictive Performance of Machine Learning Models

The performance of both LightGBM and XGBoost was evaluated based on their predictive capabilities in forecasting stock returns for S&P 500 companies shown in Fig. 1. Both models underwent rigorous tuning through Bayesian optimization, resulting in high accuracy for predicting stock price movements. When tested on data from January 2024 to July 2024, both models demonstrated solid performance, though there were some trade-offs between return and risk metrics.

XGBoost generated a cumulative return of 11.31%, with an annualized return of 20.81%, making it particularly attractive for strategies aimed at maximizing returns. LightGBM produced slightly lower returns, with a cumulative return of 10.74% and an annualized return of 19.73%. However, LightGBM excelled at risk management, with a maximum drawdown of 3.25%, compared to XGBoost's 2.80%.

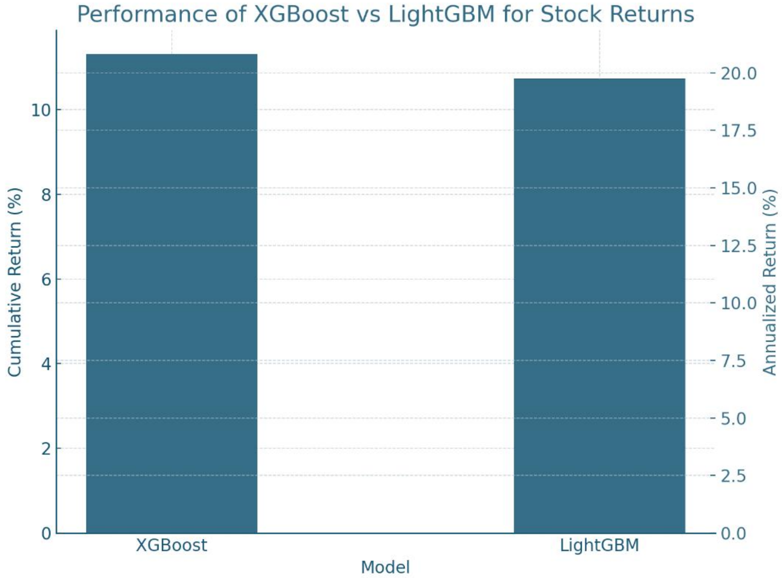

Fig. 1. The performance of two models (Photo/Picture credit: Original).

3.2 Portfolio Construction and Risk Management

Using the predictions from both models, portfolios of 30 stocks were constructed. XGBoost's portfolio was designed to maximize returns, with 50% of capital allocated to the top 10 highest predicted return stocks, and the remaining 40% invested in the next 20 stocks, leaving 10% in cash or bonds.

LightGBM's portfolio, while offering slightly lower returns, provided better protection against market volatility. This makes it particularly appealing for investors who prioritize risk management. Both portfolios incorporated risk mitigation strategies such as stop-loss orders and regular rebalancing to adapt to market fluctuations.

4 Conclusion

This study demonstrates that advanced machine learning techniques, specifically LightGBM and XGBoost, are highly effective in forecasting stock market trends, thereby aiding in the construction of more profitable and risk-adjusted investment portfolios. For investors focused on maximizing returns, XGBoost appears to be the superior choice due to its higher cumulative and annualized returns. Conversely, LightGBM is better suited for those prioritizing risk management, as evidenced by its lower maximum drawdown. The utilization of these models represents a significant improvement over traditional analytical methods. As financial markets become increasingly complex and data-driven, the relevance of machine learning in finance continues to grow. These tools not only help in adapting to rapid market changes but also in capturing subtle patterns in vast datasets, which are often invisible to traditional analyses. This capability makes machine learning an indispensable part of modern financial toolkit, offering investors a considerable edge in navigating the intricacies of the market. Moreover, machine learning's ability to process and analyze large volumes of information in real-time allows for the development of dynamic trading strategies that can adapt to new information or market conditions, further enhancing investment outcomes. As the financial industry continues to evolve, the integration of machine learning models like LightGBM and XGBoost is likely to become more prevalent, shaping the future of investment strategies and risk management.

References

1. Mahesh, B.: Machine learning algorithms-a review. International Journal of Science and Research (IJSR) 9(1), 381-386 (2020).
2. Alpaydin, E.: Machine Learning. MIT Press (2021).
3. Jordan, M. I., Mitchell, T. M.: Machine learning: Trends, perspectives, and prospects. Science 349(6245), 255-260 (2015).
4. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794 (2016).
5. Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research 270(2), 654-669 (2018).
6. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T. Y.: LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems 30 (2017).
7. Wang, D., Zhang, Y., Zhao, Y.: LightGBM: An effective miRNA classification method in breast cancer patients. In: Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, pp. 7-11 (2017).
8. Nielsen, D.: Tree boosting with XGBoost: Why does XGBoost win "every" machine learning competition? Master's thesis, NTNU (2016).
9. Ramraj, S., Uzir, N., Sunil, R., Banerjee, S.: Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications 9(40), 651-662 (2016).
10. Erlich, I., Venayagamoorthy, G. K., Worawat, N.: A mean-variance optimization algorithm. In: IEEE Congress on Evolutionary Computation, pp. 1-6. IEEE (2010).
11. Kourtis, A.: A stability approach to mean-variance optimization. Financial Review 50(3), 301-330 (2015).