



Machine Learning based Stock Price Prediction: A Case Study of LVMH Group

Yu Wang

Sino-French Institute, Renmin University of China, Suzhou 215123, Jiangsu, China
wangyu777@ruc.edu.cn

Abstract. In the current market, the high noise and non-stationary characteristics of stock price time series make accurate prediction extremely challenging. This study focuses on LVMH Group (MC.PA), the most representative blue chip asset in Europe, and extracts its real transaction data from January 2020 to May 2026 to construct lag features. The aim is to explore the predictive performance of machine learning models of different complexities in real financial noise environments. This study deployed both basic linear regression and complex ensemble tree models - Random Forest and Extreme Gradient Boosting (XGBoost). The empirical results reveal a counterintuitive core finding that, under the current lag characteristics, the simplest linear model achieved the lowest root mean square error (RMSE=10.3330) and the best fit ($R^2=0.9779$) in the test set, significantly better than the RMSE of the random forest and XGBoost models (11.1151 and 11.1007, respectively). This result strongly demonstrates the almost random walk nature of financial asset prices, as well as the structural flaw of ensemble tree models being prone to overfitting in high noise data.

Keywords: Stock Price Prediction, Machine Learning, Overfitting, LVMH Group.

1 Introduction

The time series prediction of stock prices has always been a core topic in the fields of quantitative finance and econometrics. According to the Efficient Market Hypothesis (EMH), asset prices have fully reflected all publicly available information, and their fluctuations exhibit an approximate "random walk" characteristic [1]. However, extensive empirical evidence from behavioral finance suggests that there are still subtle nonlinear patterns hidden in the market. In recent years, the explosion of machine learning technology has provided a new mathematical tool to break through the bottleneck of traditional financial linear models [2]. However, in the face of financial data with extremely high "signal-to-noise ratio", complex large models are prone to falling into the "overfitting" trap, mistaking historical random fluctuations for future patterns [3].

Based on the above background, this study selects LVMH Group as the subject of empirical research. As a core asset in Europe and a global luxury goods giant, LVMH's

stock price fluctuations contain rich macroeconomic information and consumer expectations. This study aims to evaluate the generalization ability of algorithm models with different complexities in the face of real high noise markets, and explore optimization paths for controlling model overfitting.

The subsequent structural arrangement of this article is as follows: Chapter 2 reviews relevant literature, Chapter 3 elaborates on data processing and model architecture, Chapter 4 presents empirical results and explores overfitting traps in depth, Chapter 5 proposes directions for extending deep learning, and finally summarizes the entire text in Chapter 6.

2 Literature Review

2.1 Traditional Financial Models

In the early research on financial time series forecasting, autoregressive moving average models (ARIMA) and generalized autoregressive conditional heteroskedasticity models (GARCH) dominated. These models perform well in predicting the mean and variance of stationary sequences, but often fall short when faced with complex nonlinear and non-stationary features in modern financial data.

2.2 Machine Learning in Stock Prediction

With the exponential growth of computing power, machine learning has fundamentally reshaped quantitative investment and empirical asset pricing [1]. Traditional linear frameworks often fail to capture the high-noise and non-stationary nature of modern financial securities, prompting a significant shift toward complex algorithmic models [4]. Ensemble learning methods, in particular, have been widely adopted to address these challenges. Random Forest (RF) effectively mitigates the high variance inherent in single decision trees through Bootstrap sampling and random feature selection [5]. XGBoost extends this by utilizing residual iteration (Boosting) to continuously optimize loss functions, demonstrating exceptional fitting capabilities in complex datasets.

Furthermore, recent literature increasingly highlights the necessity of deploying these complex models to mine deep nonlinear correlations in stock data. Comprehensive reviews indicate that advanced machine learning architectures excel in extracting hierarchical features from raw financial time series. However, as algorithms grow in complexity, the risk of misinterpreting market noise as structural signals increases. It is absolutely crucial to implement rigorous validation protocols in financial machine learning to prevent high-dimensional models from falling into the trap of overfitting to historical data [3].

2.3 Research Gap

It is generally believed that the more complex the model, the better the prediction performance. In extremely noisy micro financial time series, there is still a lack of rigorous horizontal comparison between basic linear models and complex large models in high noise environments [6]. This study aims to fill this gap and verify the limitations of complex tree models in temporal extrapolation.

3 Methodology

3.1 Data Acquisition

This study used Python's yfinance interface to extract the daily closing prices (Close Prices) of LVMH Group (MC.PA) from the Yahoo Finance database from January 2, 2020 to May 20, 2026. After removing missing values, a total of 1630 valid trading days' real samples were obtained.

3.2 Feature Engineering

To capture temporal autocorrelation, this study used the Sliding Window method to extract the closing prices of the previous three days and constructed third-order lag features (Lag_1, Lag_2, Lag_3). Finally, the closing price of the next day (Target Next Day) was used as the regression prediction target.

3.3 Data Partitioning

Before model training, the dataset is strictly truncated and divided into a training set (1304 entries) and a testing set (326 entries) in chronological order, with a ratio of 80% to 20%. The fundamental purpose of this is to eliminate the common risk of future data breaches in time series analysis [7].

3.4 Model Architectures

This study selected three models for comparative evaluation. Firstly, linear regression is used as the most fundamental statistical benchmark. Next is Random Forest, which represents the Bagging ensemble algorithm. In this study, the number of trees is set to 100. Finally, XGBoost, as a representative of Boosting algorithm, is set to have a learning rate of 0.05.

3.5 Evaluation Metrics

This study uses root mean square error (RMSE) and coefficient of determination (R^2) to measure the performance of the model on the test set. Root mean square error is mainly used to measure the absolute error between the predicted values and the true

values of a model. This indicator penalizes larger prediction biases, and the smaller the value, the higher the prediction accuracy of the model. The coefficient of determination is used to evaluate the explanatory power of the model on data variability, and its value range is usually between 0 and 1. The closer the value is to 1, the better the fitting effect of the model on the target variable.

4 Empirical Results and Discussion

4.1 Exploratory Data Analysis

Data statistics show that LVMH's stock price experienced severe fluctuations during this period. Its average closing price was 575.38 euros, hitting a low of 256.94 euros and climbing to a maximum of 837.37 euros. The high standard deviation of 130.13 highlights the highly dispersed yield distribution of the underlying asset during the pandemic shock and recovery cycle, demonstrating strong high-noise characteristics.

4.2 Performance Comparison

The evaluation results of the models on the independent test set are shown in Table 1.

Table 1. Performance comparison of models on the test set

Algorithms	Test Set RMSE	Test Set R ² Score	Performance
Linear Regression	10.3330	0.9779	Best
XGBoost	11.1007	0.9745	Sub-optimal
Random Forest	11.1151	0.9744	Poor

The above results reveal a profound and counterintuitive phenomenon, where the simplest linear regression defeated the most cutting-edge ensemble large models [8]. In the short-term price prediction of financial assets, the market is highly close to a random walk, and tomorrow's price is largely anchored to today's price. Complex ensemble tree models have extremely high degrees of freedom and branching capabilities. They overexert themselves during the training process, mistakenly recording meaningless fluctuation noise in historical data as signals in nodes. When these models face unknown random shocks in the test set, their generalization ability inevitably decreases due to overfitting, leading to an expanded error [9]. Linear regression, relying on its minimalist structure, instead stripped away the noise and firmly captured the main trend of the sequence.

5 Future Work Deep Learning Extension

5.1 Limitations of Tree based Models

This stage of the experiment proves that decision tree based models have structural limitations when dealing with time series problems [10]. They essentially still treat the input as a combination of cross-sectional features, ignoring the long-range memory effect and dynamic evolutionary logic between time steps. In addition, the ensemble tree model is prone to over segmentation when faced with high-frequency financial noise, resulting in a large number of meaningless random fluctuations being incorrectly recorded in nodes and lacking the ability to smooth out temporal trends.

5.2 Introducing LSTM

To overcome this limitation, the next stage of research will introduce Long Short Term Memory (LSTM) networks in deep learning. LSTM, through its unique forget gate and memory gate architecture, can actively filter high-frequency financial noise and retain core temporal signals that have a profound impact on price prediction. This network topology structure with self feedback loops will provide a truly time series compliant technical solution for stock price prediction in high noise environments.

5.3 Integration of Alternative Data

In addition to upgrading the model architecture, future research can also expand horizontally from the dimension of feature engineering. Due to the significant impact of global macroeconomics and consumer sentiment on LVMH Group's stock price, a single price volume characteristic often has obvious information bottlenecks. The follow-up research plan is to introduce natural language processing technology to extract emotional scores from financial news and social media texts, construct multimodal deep learning inputs, and more comprehensively capture market fluctuations.

5.4 Real world Trading Strategy Evaluation

The current model evaluation is limited to statistical error indicators such as RMSE, and in real quantitative investment practice, smaller prediction errors do not necessarily equate to higher strategy returns. Therefore, the expansion research of deep learning models will closely integrate with financial reality, introduce risk adjusted return indicators such as maximum drawdown and Sharpe ratio, and comprehensively test the cross cycle robustness of the model in different market cycles such as bull and bear markets.

6 Conclusion

This study takes LVMH Group as the empirical subject and conducts empirical comparison of short-term stock price prediction based on lag characteristics. The core conclusion indicates that in noisy micro financial sequences, simple linear regression unexpectedly outperforms complex machine learning ensemble models with its robust ability to track main trends. This discovery has important warning significance for quantitative investment practice, that is, in financial feature engineering, the more complex the algorithm, the better. How to effectively control the overfitting of the model and respect the basic logic of market random walk is the core of building practical value strategies.

References

1. Gu, S., Kelly, B., Xiu, D.: Empirical asset pricing via machine learning. *The Review of Financial Studies* 33(5), 2223-2273 (2020)
2. Jiang, W.: Applications of deep learning in stock market prediction: Recent progress. *Expert Systems with Applications* 184, 115537 (2021)
3. Lopez de Prado, M.: *Advances in Financial Machine Learning*. John Wiley & Sons, New York (2018)
4. Drobetz, W., Otto, T.: Empirical asset pricing via machine learning: evidence from the European stock market. *Journal of Asset Management* 22(7), 507-538 (2021)
5. Andri, F., Nur, R.: A Comparative Analysis of XGBoost and Random Forest for Time Series Based Stock Price Prediction with Directional Movement Evaluation. *Terapan Informatika Nusantara* 4(1), 15-25 (2024)
6. Fieberg, C., Metko, D., Poddig, T., Loy, T.: Machine learning techniques for cross-sectional equity returns' prediction. *OR Spectrum* 45(1), 289-323 (2023)
7. Kaufman, S., Rosset, S., Perlich, C., Ori, O.: Leakage in Data Mining: Formulation, Detection, and Avoidance. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 556-563. ACM, New York (2012)
8. Marsi, A.: Predicting European stock returns using machine learning. *SN Business & Economics* 3(7), 118 (2023)
9. Babiak, M., Barunik, J.: Deep Learning, Predictability, and Optimal Portfolio Returns. *arXiv preprint arXiv:2012.14442* (2020)
10. Sezer, O. B., Gudelek, M. U., Ozbayoglu, A. M.: Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing* 90, 106181 (2020)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

