Advancing Stock Return Prediction: A Comprehensive Study of Traditional Machine Learning and Deep Learning Models

Yupeng Liu

Math, University of California, Santa Barbara, Goleta, 93106, USA
yupengliu@ucsb.edu

Abstract. Stock markets exhibit high volatility, driven by numerous factors, making accurate stock return prediction a challenging task. This paper provides a comprehensive review of machine learning models used to stock return forecasting, comparing traditional models like Linear Regression (LR), Support Vector Machine (SVM), and Random Forest (RF) with deep learning techniques such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks. Traditional methods are praised for their interpretability and computational efficiency but frequently fail to convey the non-linear and time-dependent complexities of financial data. In contrast, deep learning models, particularly LSTM, are highly effective in modeling long-term dependencies in stock data but face challenges related to overfitting, interpretability, and generalizability across different market conditions. The paper also highlights the limitations of current models, including their inability to integrate external factors such as geopolitical events and policy changes. To address these issues, potential future research avenues are explored, focusing on enhancing interpretability using interpretability techniques as well as leveraging transfer learning and domain adaptation to improve scalability and model robustness. These advancements could significantly enhance the practical application of machine learning in stock return prediction, offering more reliable and interpretable solutions for real-world financial decision-making.

Keywords: Machine Learning, Deep Learning, Stock Return.

1 Introduction

Stock markets are the foundation of the global financial system, where publicly listed corporations' shares are issued, bought, and sold. Stock prices reflect a company's perceived value and change in response to a range of factors, including economic indicators, geopolitical developments, and investor sentiment. This inherent volatility makes stock prices a subject of great interest and importance to a wide range of stakeholders, from individual investors to large financial institutions. Accurately

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_93

forecasting stock returns is therefore a vital goal, as it directly impacts investment decisions, portfolio management, and even broader economic policy.

Investors seek to maximize returns while minimizing risk by selecting stocks that outperform the market. Governments and businesses also rely on stock market forecasts to make informed decisions about economic policy and strategic planning. However, the volatility of stock prices, influenced by countless variables, presents a huge challenge to traditional forecasting models. This complexity is exacerbated by the fact that financial markets are not static; they evolve rapidly in response to new information and changing circumstances. As a result, there is an increasing need for more sophisticated models to handle this complexity and provide more reliable forecasts.

In recent years, the use of Artificial Intelligence (AI) across various fields has experienced significant advancements, especially in the area of Machine Learning (ML). AI techniques have been successfully applied in fields ranging from chemistry and biomedicine to finance, demonstrating their versatility and effectiveness. For instance, Goodell et al. conducted a comprehensive bibliometric analysis to map out the key themes and research clusters where AI and ML intersect with finance [1]. Their findings highlight significant contributions in areas like portfolio management, fraud detection, and sentiment analysis, showcasing the growing importance of these technologies in financial research. In addition, Gan et al. explored the application of machine learning in pricing financial products, specifically average options. They developed a model-free ML method that enhances the speed and accuracy of pricing, overcoming the limitations of traditional numerical approaches [2].

Several AI methodologies have become prominent in financial applications. In the context of stock return prediction, the task is notably more challenging than other financial predictions due to the high volatility and complex interactions between variables that influence stock prices. Researchers have proposed various models to address this challenge. Nguyen et al. investigated the use of three machine learning models—Support Vector Machine (SVM), Random Forest (RF), and Extra-Trees Classifier—for predicting the intraday return of the HNX index [3]. Their findings revealed that the Extra-Trees Classifier outperformed Logistic Regression, demonstrating the model's superior capacity to deal with financial data complexities. Zhu et al. presented a novel hybrid neural network model that combines layers of long short-term memory (LSTM) and convolutional neural networks (CNN) [4]. This approach, enhanced by the CEEMDAN algorithm and Savitzky-Golay filter for preprocessing, significantly improved prediction accuracy by effectively managing market noise and data complexities. Sheth and Shah explored the application of ANN, SVM, and LSTM for stock price prediction, highlighting ANN's effectiveness in capturing non-linear relationships in financial data [5], while also noting the promising potential of SVM and LSTM in handling large datasets and temporal patterns. Singh et al. focused on applying LSTM networks to technical analysis in stock market prediction [6]. Their study emphasized the model's capability in forecasting stock prices based on historical data, showcasing the effectiveness of deep learning techniques in financial applications.

This paper aims to provide a comprehensive review of the current landscape of machine learning models used for stock return prediction, emphasizing both their predictive performance and the degree to which these models can be interpreted by users. By analyzing the strengths and weaknesses of existing approaches, the paper aims to provide a clear understanding of where these models excel and where they fall short. Additionally, it will explore potential avenues for future research that could address existing limitations, improve model robustness, and enhance their applicability in real-world financial scenarios.

The remainder of this paper is structured as follows: Section 2 reviews the most prominent machine learning models used in stock return prediction, including both traditional and contemporary approaches. Section 3 discusses the advantages and limitations of these models, as well as the challenges and future directions in this research area. Finally, section 4 wraps up the paper by summarizing the main findings and proposing possible directions for future research in stock return prediction using machine learning techniques.

2 Method

2.1 Introduction of Machine Learning Workflow

The machine learning workflow involves several key stages to develop an effective predictive model. The process starts with data collection, where relevant information is sourced from various channels, such as historical stock prices and financial reports. Once collected, the data undergoes preprocessing, which involves cleaning, normalizing, and addressing missing values to ensure the dataset is properly prepared for analysis. The next step is model building, where machine learning algorithms, such as decision trees or neural networks, are selected based on the problem at hand. Following this, the model is fitted using a portion of the dataset, learning patterns and relationships between inputs and outputs. After training, the model undergoes testing on unseen data to evaluate its performance using metrics such as accuracy, precision, recall, F1 score, and AUC-ROC, which provide a more comprehensive understanding of the model's predictive capabilities. If the results are satisfactory, the model is then deployed into a real-world environment, where it can be monitored and updated regularly to adapt to changes in the data distribution or market conditions, continuously refining its performance as new data becomes available.

2.2 Traditional Machine Learning-based Prediction

LR. Linear regression is a statistical method used to model the relationship between a target variable and a set of predictor variables by fitting a linear equation to the data. In this study, Ahangar et al. applied linear regression to predict stock prices on the Tehran Stock Exchange. They initially considered ten macroeconomic and thirty financial variables, but using Independent Component Analysis (ICA), they reduced the dataset to seven key variables. The linear regression model was then built using these selected variables to predict stock prices. The model’s performance was evaluated using Mean Square Error (MSE) and R^2 as key metrics, and it was compared against the performance of an artificial neural network [7].

SVM. Support Vector Machines (SVM) are a form of supervised machine learning algorithm that are useful for managing nonlinear patterns and high-dimensional data in regression and classification applications. Wang applied SVM to predict stock price movements by using historical stock data from the Chinese market. The model was trained to classify stock trends based on a nonlinear decision boundary. Preprocessing involved normalizing the stock data and selecting relevant features. The SVM algorithm was then used to identify optimal support vectors that separate stock movement classifications. Performance was evaluated using accuracy, precision, recall, and F1 score, with SVM outperforming other models, showcasing its strength in predicting stock returns [8].

RF. Random Forest (RF) is a machine learning model used to predict stock return rates by building multiple decision trees during training and averaging their outputs for regression tasks. To implement this model for stock return prediction, Nguyen et al. used the HNX stock index data, applying a sliding window technique to divide the dataset into training and testing sets. Kernel PCA was employed for feature extraction and dimensionality reduction to ensure only the most relevant features were used. The model was trained in historical stock data, and its accuracy was evaluated through cross-validation and metrics such as MSE, providing a more comprehensive assessment of the model's performance .

2.3 Deep Learning-based Prediction

ANN. Artificial Neural Networks (ANN) are highly effective machine learning models designed to capture complex and nonlinear relationships in data, enable them to be well-suited for predicting stock returns. ANN mimics the structure of the human brain, consisting of multiple interconnected layers of neurons that process data to recognize patterns. Sheth and Shah applied ANN to historical stock price data, passing the data through multiple layers to enable the model to predict future stock returns. The training process involved optimizing weights and biases through backpropagation to minimize prediction errors. The model's performance was evaluated using metrics like MSE, demonstrating ANN's ability to effectively forecast stock trends.

CNN. Convolutional Neural Networks (CNN), a type of deep learning model, are built to capture spatial patterns in data, typically used in image processing but also applied in financial time-series prediction. In the paper by Cao, CNN is used for stock return prediction by processing historical stock price data. The authors applied a Sliding-Window model, where stock data was divided into smaller windows to predict future prices. The model is composed of convolutional layers, which extract local patterns from the stock data, and pooling layers, which down-sample the data to reduce its dimensionality. The output is then passed through fully connected layers for final predictions. Performance was evaluated using MSE and MAE, showing CNN's capacity to capture complex market dynamics [9].

LSTM. Long Short-Term Memory (LSTM), a variant of Recurrent Neural Network (RNN), is particularly well-suited for stock price prediction due to its ability to effectively capture long-term dependencies in time series data. In their approach, Singh et al. used LSTM to predict stock return rates by processing historical stock price data with multiple time lags. They constructed an embedded layer to reduce the dimensionality of input data, followed by an LSTM layer to capture temporal dependencies. The final output layer provided predictions based on stock price movements. The model was trained using the Bombay Stock Exchange (BSE) dataset, with performance evaluated through metrics like MSE. The structure allowed the model to efficiently learn patterns from past stock data, providing accurate predictions for future stock trends.

3 Discussion

The application of ML in stock return prediction presents both significant advantages and considerable challenges. As the field evolves, the choice of models—ranging from traditional methods like LR and SVM to more advanced deep learning techniques—has sparked considerable discussion in both academic and practical contexts. This section discusses the strengths and limitations of traditional machine learning models in comparison to deep learning methods and examines the challenges these models face, such as interpretability, applicability to different data distributions, and sensitivity to external factors like policy changes and news events. Additionally, it explores future prospects for overcoming these challenges through techniques such as interpretability-enhancing tools, transfer learning, and the incorporation of external data. Traditional machine learning models like LR, SVM, and RF offer advantages in stock return prediction. LR is highly interpretable, providing clear linear relationships between input variables and stock returns, enabling investors to assess factor influences. SVM handles high-dimensional and non-linear data, offering robust classification of stock trends, while RF reduces overfitting through its ensemble approach. However, LR struggles with the non-linearity of financial markets, and though SVM and RF capture non-linear relationships, they are computationally intensive, particularly for huge datasets. Additionally, these models often fail to fully account for time-dependent patterns in stock data, limiting their accuracy in predicting future returns.

On the other hand, deep learning models, such as ANN, CNN, and LSTM networks, have gained prominence due to their superior ability to capture complex and non-linear relationships in financial data. LSTM, in particular, excels at modeling long-term dependencies and temporal patterns, making it ideal for stock price prediction based on historical data. While originally used in image processing, CNNs have been modified for use in financial time series data analysis, where they are effective at capturing local patterns in stock movements. However, despite their enhanced accuracy and predictive power, deep learning models introduce new challenges, particularly in terms of interpretability and overfitting, which can undermine their real-world applicability.

One of the major limitations of current machine learning models, particularly deep learning methods, is their lack of interpretability. While models like LR offer clear insights into the factors driving predictions, deep learning models are frequently regarded as "black boxes," offering limited insight about how or why a particular prediction is made. This lack of transparency challenges financial applications, as investors may hesitate to trust models that cannot explain their predictions, especially in high-stakes decisions.

Another challenge is the applicability issue, where models trained on specific stocks or indices often struggle when applied to different stocks or sectors due to variations in data patterns. Financial markets are highly diverse, and training a model on every individual stock is resource-intensive. For instance, a model trained on HNX index data may perform poorly on other indices with different volatility or market structures, reducing the model's generalizability across various financial environments and limiting its scalability for stock return prediction.

The third significant challenge lies in the models' inability to account for external factors, such as government policies, geopolitical events, or major news reports, all of which can profoundly influence stock prices. AI models, especially deep learning algorithms, typically rely solely on historical price data and technical indicators. However, these external factors, which are often unstructured and unpredictable, can shift market sentiment and thus stock prices, in ways that are not captured by the training data.

To address the issue of interpretability, researchers are exploring the application of techniques such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These approaches offer post-hoc explanations for model predictions, making it easier to understand the contribution of individual features in complex models like LSTM and CNN. Integrating these tools into stock return prediction models could significantly enhance user trust and facilitate the practical deployment of these models in real-world scenarios [10].

The problem of limited applicability across different stocks and market environments may be mitigated through techniques such as transfer learning and domain adaptation. Transfer learning enables a model trained on one dataset to be refined and adapted for use on a different dataset, reducing the computational cost and improving generalizability. Domain adaptation further refines this approach by adjusting models to work with data that come from different distributions. These techniques could allow stock return models to better handle the diversity of financial markets without requiring extensive retraining.

Finally, to improve stock return predictions, it is essential to integrate external factors like news reports and policy changes into machine learning models. Recent developments in Natural Language Processing (NLP) and sentiment evaluation enable the extraction of meaningful information from unstructured sources, such as social media and news report, allowing this data to be incorporated into stock return prediction models. This would create a more holistic and comprehensive prediction model that better reflects real-world market dynamics.

4 Conclusion

This paper comprehensively reviews machine learning models used in stock return forecasting, focusing on traditional methods such as LR, SVM, and RF, as well as more advanced deep learning techniques such as ANN, CNN, and LSTM. This paper explored the strengths of these models, especially in capturing intricate nonlinear correlations in financial data, while also addressing their limitations such as scalability, interpretability, and difficulty in interpreting time-dependent patterns and external factors such as government policies or market news.

However, despite their advanced capabilities, deep learning models face challenges related to scalability, interpretability, and overfitting. Through the analysis, several limitations were identified, including the difficulty of generalizing models to different stock markets and the inability to integrate external factors like news or policy changes. Future research should focus on enhancing model interpretability through techniques like SHAP and LIME, and improving scalability with transfer learning and domain adaptation. Addressing these challenges could help researchers develop more powerful and applicable stock prediction models.

References

1. Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D.: Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, *32*, 100577 (2021).
2. Gan, L., Wang, H., & Yang, Z.: Machine learning solutions to challenges in finance: An application to the pricing of financial products. *Technological Forecasting and Social Change*, *153*, 119928 (2020).
3. Nguyen, T. N., Ho-Phuoc, T., Nguyen, D. T., & Mac, M. N.: Stock Return Prediction using Machine Learning-Based Techniques. *Journal of Science and Technology: Issue on Information and Communications Technology*, *18*(12.2), 49-56 (2020).
4. Zhu, R., Zhong, G. Y., & Li, J. C.: Forecasting price in a new hybrid neural network model with machine learning. *Expert Systems with Applications*, *249*, 123697 (2024).
5. Sheth, D., & Shah, M.: Predicting stock market using machine learning: best and accurate way to know future stock prices. *International Journal of System Assurance Engineering and Management*, *14*(1), 1-18 (2023).
6. Singh, A., Bhardwaj, G., Srivastava, A. P., Bindra, A., & Chaudhary, P.: Application of neural network to technical analysis of stock market prediction. In *2022 3rd International Conference on Intelligent Engineering and Management (ICIEM)* (pp. 302-306). IEEE (2022).
7. Ahangar, R. G., Yahyazadehfar, M., & Pournaghshband, H.: The comparison of methods artificial neural network with linear regression using specific variables for prediction stock price in Tehran stock exchange. *arXiv preprint arXiv:1003.1457* (2010).

8. Wang, H.: Stock price prediction based on machine learning approaches. In *Proceedings of the 3rd International Conference on Data Science and Information Technology* (pp. 1-5) (2020).
9. Cao, R.: Stock Price Prediction Using Deep-Learning Models: CNN, RNN, and LSTM. In *SHS Web of Conferences* (Vol. 196, p. 02004). EDP Sciences (2024).
10. Gramegna, A., & Giudici, P.: SHAP and LIME: an evaluation of discriminative power in credit risk. *Frontiers in Artificial Intelligence*, *4*, 752558 (2021).