

Progresses of Machine Learning in Stock Market Prediction: A Comprehensive Model Comparison

Zheng Qin^a

Zhengzhou No.9 High School, Zhengzhou, China

Keywords: Artificial Intelligence, Machine Learning, Stock Price Prediction.

Abstract: Stock price prediction is a crucial technique in investment, but traditional approaches can hardly get high accuracy and precision on predicting as they cannot consider about the various factors in the complex market. As a result, more advanced method is desired by the market. In this review, several major methods of machine learning of artificial intelligence are referred, such as Linear Regression, Random Forest, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). And the basic workflow behind the application of these methods is also introduced. Based on the cases of different models in predicting stock prices, the advantages and disadvantages of various machine learning models in forecasting are illustrated and compared. Traditional models like Linear Regression offer simplicity but are limited by their inability to capture non-linear relationships. In contrast, Random Forest models improve prediction accuracy through ensemble methods but face challenges like overfitting and lack of temporal modelling capabilities. Advanced models such as RNNs and LSTMs excel in capturing complex temporal dependencies, making them more effective for stock price forecasting. This review can provide a good reference for scholars in the intersection of finance and AI.


1 INTRODUCTION

The stock market provides a platform where individuals can buy and sell shares of companies. Each share sold in the market represents ownership in a company, traded under numerous factors including economic indicators, company performance, investor behavior, and global events that influence the price. For investors, predicting stock prices becomes important because it influences various investment decisions, the management of risks, and financial planning.

Stock price prediction is the procedure of estimating future stock prices based on historical data. This has traditionally been done with the help of financial models like the Capital Asset Pricing Model (CAPM) and Efficient Market Hypothesis (EMH). However, in general, these kinds of models are based on the assumptions of market behavior but mostly fail to capture the ever-changing complexity and dynamism of financial markets. This would mean

very low predictive accuracy in facing settings where not only changes but also those happening fast should be taken into account. Advanced computing has facilitated the emergence of advanced methods for predicting stock prices. These methods are data-driven and incorporate strong algorithms to analyze past movement in price and possibly capture any patterns that old models may not capture. Artificial Intelligence has been regarded as a promising direction for better prediction of stock price. As a whole computer science, AI can process very big volumes of data, to identify trends and even make predictions with very little human intervention.

AI has taken a giant leap in recent years, and it can be observed that there are many algorithms that show potential for predictive tasks. Examples include decision trees, random forests, and logistic regression, which are well known for their inherent characteristics in dealing with different data and prediction problems. Furthermore, the advent of large language models like Chat Generative Pre-Trained Transformer (ChatGPT) epitomizes the vastness of

^a  <https://orcid.org/0009-0003-7651-1253>

the boundaries crossed by AI into natural language processing and, further, decision-making.

AI has not been staying within the world of finance but has touched the other major fields of biology, chemistry, and medicine as well. Due to the somewhat general capabilities gained by AI, these have found applications in finance, including being applied to the problem of prediction of stock price.

There are a great amount of cases of applying machine learning models in the financial field. Zeinalizadeh et al. created an application using a neural network model to accurately forecast financial service customer behavior (Zeinalizadeh, 2015). Zhang et al. applied an SVM model to stock price prediction with great success (Cao, 2019). A random forest model was also used to further investigate the prediction of market trends by Chen et al. (Chen, 2021), while Singh et al. introduced a deep learning approach using Recurrent Neural Networks (RNNs) and Long Short-term Memory (LSTM) networks to improve stock price predictions (Singh, 2022). The results of these studies all bring into focus the interest in using AI to tackle financial prediction problems. There is an incessant urge given the importance of predicting stock prices, especially with the rapid advances in AI technology, to explore the transformation that these technologies are bringing about in the financial markets. As a result, this review was aimed to assess the state of the art in AI-based stock price prediction, effectiveness, and identification of further research and innovation possibilities. The more AI develops, the more it is integrated into the fabric of financial markets, the better it gets at predictive accuracy, benefitting investors, and in a far greater sense, the economy.

2 METHOD

2.1 The Introduction of Machine Learning Workflow

Usually, the development process of a machine learning-based stock price prediction model is executed in a structured workflow that involves several critical steps aim to ensure the robustness and accuracy of the models, which will deliver reliable predictions based on financial data.

2.1.1 Data Collection

The first step is gathering relevant data, it is a quite crucial part for building an effective model. In the context of stock price prediction, data is gained from

many kinds of sources such as stock market databases and financial platforms like the website Yahoo Finance. The data consist of different stocks in different markets and key attributes such as opening prices, closing prices, trading volumes, and other market indicators. It is essential to collect all-sided data because it forms the foundation of the model.

2.1.2 Data Preprocessing

After the collection of the data is finished, it has to experience a preprocessing to make sure its quality and suitability is high enough for model training. It concludes disposing missing value with imputation, identifying and managing those error values affecting the effectiveness of prediction, and sort the dataset into two sets, one is for training and one is for testing. Normalization is also applied to the data in this step, it can scale the data by making sure the contribution of all features are equal and rational. This part is able to reduce the biases and the performance in great extent.

2.1.3 Model Building

For building the model, suitable machine learning or deep learning models are selected and designed. Random Forests, Decision Trees, and Neural Networks are the most common ones, each of them provides unique advantages for stock price prediction. In this phase, the model's architecture is carefully constructed, and hyperparameters are tuned to optimize its ability to learn from the data.

2.1.4 Model Training

The training process involves feeding the preprocessed training data into the chosen models. During training, the models learn patterns within the data, then fitting the parameters to avoid the occurrence of errors. This process is iterative, and it gives the model the ability to generalize effectively to new, unseen data.

2.1.5 Model Testing

After training, the models are evaluated using the testing dataset. Measurement numbers such as Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are worked out to assess prediction accuracy. These metrics help in identifying the strengths and weaknesses of the model.

2.1.6 Model Deployment

After the model passes the test with metrics which are lower than the threshold and show its satisfying performance, the deployment is ready. During the deployment the model is integrated into a real-world system where it can make predictions based on live data. Continuous monitoring and updating are taken in order to remain the model is accurate while the market condition evolves.

2.2 Linear Regression

Linear regression is a basic statistical technique which is always utilized building the connection among a dependent and at least one independent variable. The main concept behind the model is to fit the linear equation to the gathered data such that the total of squared variances between the gathered and the predicted values is minimized.

Linear regression can be utilized to forecast stock prices in coming days by inputting the past data from the source. For instance, a study carried out by Cakra et al., employed linear regression to analyze stock price movements and developed a model that demonstrated how past price trends could be leveraged to predict future performance (Cakra, 2015). It is realized easily by fitting the model and then asking the model to generate the forecasting result. As detailed in their findings, the potential of linear regression for making initial predictions is shown, although it also acknowledged the limitations of linear models in capturing complex market dynamics.

2.3 Random Forest

Random Forest is a method for group learning it works by combining multiple decision trees for both higher accuracy and less serious overfitting. Creating a "forest" of decision trees is the main concept of the model, every tree is formed from a subset of the data and features randomly. After integrating the result from all the trees, the overall prediction is worked out, commonly using voting for classification or averaging for regression.

In stock price prediction, Random Forest enhance prediction performance by leveraging its ability to handle large datasets and complex interactions between features. A study by Khaidem et al. employed Random Forest to predict stock prices and demonstrated its effectiveness in capturing intricate patterns within the data (Khaidem, 2016). Their review illustrated how Random Forest's robustness against overfitting and analyzed how its ability of

disposing nonlinear relationships results in more accurate stock price forecasts.

2.4 Recurrent Neural Networks (RNN)

An artificial neural network class called RNNs is specifically made for sequence prediction applications. They are quite different from traditional feedforward networks, because connections of it can form cycles, so they have unique ability to hold the memory for a while. That is the reason why RNNs are particularly suitable for time series data, where past information is crucial for predicting future values.

As for stock price prediction, RNNs can be used to simulate temporal dependencies and trends. For example, Lu et al developed an RNN-based model: Time-series Recurrent Neural Network (TRNN) to analyze stock prices, and their work successfully demonstrates how the network work in capturing temporal relationships between past prices could be utilized to forecast future stock trends (Lu, 2024). Their research emphasized the advantages of RNNs in modeling sequential data and their creative solution for improving prediction accuracy by transforming the one-dimension price-volume relationship into two dimensions ones.

2.5 Long Short-Term Memory (LSTM)

LSTM networks are improved version of RNN because it is designed to address the drawbacks of original RNNs, like their difficulty in learning long-term dependencies. Improved models have an outstanding capability that enables the network to be able to remember and make use of information over longer sequences by introducing memory cells and gating mechanisms for controlling the information flow. Because of their capacity to handle intricate patterns and long-term dependencies in time series data, LSTMs have shown efficacy in stock price prediction. For instance, Achyut et al used LSTM networks to forecast stock prices and develop a predictive model that take the advantage of the network's capacity for remaining information across extended periods (Ghosh, 2019). Their findings showcased LSTM's superior performance in capturing long-term trends and enhancing prediction accuracy.

3 DISCUSSIONS

A number of machine learning models have distinct strengths and weaknesses in predicting stock prices. This section compares the advantages and limitations

of Linear Regression, Random Forest, RNNs, and LSTMs, and concludes the key challenges and future directions in this field.

3.1 Comparison of Machine Learning Models

Linear Regression: Linear Regression is straightforward and easy to realize. It gets a superior performance when the relationships between variables are approximately linear. However, in contrast it almost does not work in capturing patterns present in stock price movements, when they are complex and non-linear. So, its effectiveness is strongly limited when facing more intricate prediction tasks.

Random Forest: Random Forest gets satisfying performance by aggregating multiple decision trees, which enhances robustness and reduces overfitting compared to individual decision trees. But it may still be overfit with too many trees or irrelevant features. Random Forest is less suitable for capturing stock trends because it does not inherently take account for temporal dependencies in time series data.

Recurrent Neural Networks: RNNs are designed to handle sequential data and can capture temporal dependencies, driving them applicable for time series forecasting. But they may struggle when facing vanishing and exploding gradient problems because they increase the difficulty of learning because of greater complexity when facing long-term dependencies.

Long Short-Term Memory: It can address many of the limitations of RNNs by incorporating mechanisms to manage long-term dependencies and complex temporal relationships. That is the reason why they get much better performance in predicting compared with RNNs. But they are also not perfect, LSTMs can be very demanding and costly in resources because they are computationally intensive and require careful tuning.

3.2 Challenges in Stock Price Prediction

AI Model Interpretability: Many advanced models all have a very low interpretability which means it is very hard to understand the algorithm behind the model and how do they work, such as deep learning networks, they can be quite similar to "black boxes", This lack of transparency can reduce trust in the results, complicate efforts to improve or adjust the model when predictions are inaccurate.

Model Generalizability: Models trained on one stock often struggle to generalize to other stocks

because of significant differences in data characteristics. So most models are not applicable to every piece of stock. But training separate models for each stock is impractical due to high costs and data collection challenges, so it is necessary to find a way to improve the adaptability of models to enable them get satisfying performance in prediction of various stocks.

3.3 Future Directions

Expert Systems and Explainability Methods: In order to make it easier to understand how the prediction is made, it is essential to adopt some approaches such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). SHAP values help quantify the impact of each feature on predictions, while LIME provides simplified, interpretable approximations of complex models, improving transparency and understanding.

Transfer Learning and Domain Adaptation: As for the poor generality of most models, some techniques can offer promising and practical solutions. Transfer learning leverages knowledge from one domain to enhance performance in a related domain (Weiss, 2016; Zhuang, 2020), while domain adaptation adjusts models to perform well across different data distributions. Taking full advantages of these can save a lot of effort which is originally wasted on retraining models and facilitate more effective predictions across diverse stocks.

In summary, while traditional models like Linear Regression and Random Forest have their uses, advanced models such as RNNs and LSTMs are better suited to handle the complexities of stock price prediction. Addressing challenges related to model interpretability and generalizability through innovative methods will be crucial for advancing the field and achieving more accurate predictions.

4 CONCLUSIONS

This review summarizes the application and progress of the AI and machine learning in stock price prediction and mentions many practical application cases. And introduced the machine learning workflow, different models such as linear regression, random forest and their application in predicting stock prices. In this paper, the advantages and disadvantages of various model algorithms have been listed and compared. Different models all have some

specific shortages. These findings highlight the importance of deep learning models in financial forecasting, especially as markets become more complex and data driven. The successful application of these models could lead to more accurate forecasts. And this paper can provide a good overview reference for computing and finance. This paper only focuses on AI models but does not investigate traditional financial models or sequential models such as Arima, which will be further considered and analyzed in the future.

REFERENCES

- Cakra, Y. E., & Trisedya, B. D. 2015. Stock price prediction using linear regression based on sentiment analysis. 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE.
- Cao, H., et al. 2019. Stock price pattern prediction based on complex network and machine learning. *Complexity*, 2019.1: 4132485.
- Chen, Q., et al. 2021. Mapping China's regional economic activity by integrating points-of-interest and remote sensing data with random forest. *Environment and Planning B: Urban Analytics and City Science*, 48.7: 1876-1894.
- Ghosh, A., et al. 2019. Stock price prediction using LSTM on Indian share market. *Proceedings of the 32nd International Conference on*. Vol. 63.
- Khaidem, L., Saha, S., & Roy Dey, S. 2016. Predicting the direction of stock market prices using random forest. arXiv preprint arXiv:1605.00003.
- Lu, M., & Xu, X. 2024. TRNN: An efficient time-series recurrent neural network for stock price prediction. *Information Sciences*, 657: 119951.
- Singh, N., Mohan, B. R., & Naik, N. 2022. Hybrid model of multifactor analysis with RNN-LSTM to predict stock price. *Advanced Machine Intelligence and Signal Processing*. Singapore: Springer Nature Singapore.
- Weiss, K., Khoshgoftaar, T. M., & Wang, D. 2016. A survey of transfer learning. *Journal of Big Data*, 3: 1-40.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., ... & He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43-76.
- Zeinalizadeh, N., Shojaie, A. A., & Shariatmadari, M. 2015. Modeling and analysis of bank customer satisfaction using neural networks approach. *International Journal of Bank Marketing*, 33.6: 717-732.