Three Machine Learning Predictions of U.S. Stock Prices

Runjia Zhang[a], Yazhe Wang[b], Jingyi Zhang[c], Lingzi Zhang[d], Jingjing Qu[e,*]

School of Mathematical and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, China

[a]RunjiaZhang22@student.xjtlu.edu.cn, [b]YazheWang22@student.xjtlu.edu.cn,
[c]JingyiZhang22@student.xjtlu.edu.cn, [d]LingziZhang22@student.xjtlu.edu.cn,
[e]JingjingQu22@student.xjtlu.edu.cn

*Corresponding author

Keywords: XGB, ANN, SHAP, RF

Abstract: In recent years, financial markets have adopted advanced machine learning techniques for stock price prediction. For example, deep learning models such as ANN, XGBoost, SHAP and RF, which have their own advantages and disadvantages in predicting stock prices. In this study, stock price prediction is carried out by 3 powerful machine learning models: XGBoost, ANN, RF and then their error and accuracy are analyzed.

1. Introduction

In recent years, the financial market has experienced a notable surge in the adoption of advanced machine learning techniques for stock price prediction. This shift is driven by the recognition of limitations in traditional methods, such as time series analysis and statistical models, which struggle to capture the intricate patterns and non-linear dynamics inherent in financial markets. The inadequacies of these conventional approaches become apparent in the face of sudden market shifts, unpredictable geopolitical events, and the growing influence of sentiment in the era of social media. These challenges have spurred researchers to explore cutting-edge technologies, including deep learning models like Artificial Neural Networks (ANN), XGBoost, and Random Forest, aiming to address the shortcomings of traditional methods. The complex and dynamic nature of financial markets demands adaptive models capable of discerning intricate patterns, handling non-linear relationships, and adjusting to evolving market conditions. This research contributes to the ongoing discourse on stock price prediction by harnessing the capabilities of three powerful machine learning models: XGBoost, Artificial Neural Networks (ANN), and Random Forest. These models offer unique strengths, presenting the potential for heightened accuracy and robustness in predicting stock prices. By delving into the application of these advanced machine learning models, our study aims to unravel the distinctive strengths and capabilities each model brings to the forefront in the realm of stock price prediction. The initial section of our investigation is dedicated to examining the application of XGBoost, a gradient boosting algorithm renowned for its efficiency and scalability. Esteemed for its proficiency in handling extensive datasets and capturing intricate relationships, XGBoost emerges as a fitting choice for the intricate task of predicting stock prices. Moving on to the subsequent section, our study delves into the application of Artificial Neural Networks (ANN), a category of deep learning models inspired by the neural structure of the human brain. Demonstrating success in capturing intricate patterns and dependencies within data, ANNs showcase a capacity for feature learning that aligns well with the dynamic and non-linear nature inherent in financial markets. In the penultimate section, our focus shifts to Random Forest, an ensemble learning method that harnesses the collective power of multiple decision trees. Through the aggregation of predictions from diverse trees, Random Forest significantly enhances overall accuracy and robustness in forecasting. This aggregation not only mitigates the impact of overfitting but also improves generalization to new data, making it a valuable asset in the realm of stock price prediction. In the final section, we concludes the paper. The article is structured to address the increasing adoption of advanced machine learning techniques in stock price prediction due to the limitations of traditional

DOI: 10.25236/iceesr.2024.046

methods. This paragraph introduces the motivation behind this shift and the challenges faced by conventional approaches. The subsequent paragraphs focus on the three advanced machine learning models—XGBoost, Artificial Neural Networks (ANN), and Random Forest—highlighting their unique strengths and applications in stock price prediction. The final section summarizes the key findings and contributions of the research, providing a cohesive conclusion to the paper.

2. Iiterature Review

In recent years, the application of deep learning techniques in predicting stock prices has gained significant attention among researchers. Various studies have explored different architectures and methodologies to enhance the accuracy of stock market forecasting，thus the field of stock market prediction has witnessed significant advancements. This review examines recent studies that leverage deep learning techniques to predict stock prices and trends. Depending on the type of deep learning model used, the literature conducting research on stock price prediction can be categorized into two groups. The first strand of literature used early basic deep learning models to predict stock prices. Hiransha et al., (2018)[12] laid the groundwork by pioneering the use of Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) architectures for predicting stock prices based on historical data. Their focus on day-wise closing prices from the National Stock Exchange (NSE) of India and New York Stock Exchange (NYSE) provided a global perspective. Minh et al., (2018)[3] proposed a novel framework incorporating financial news and sentiment analysis. Their two-stream gated recurrent unit network and Stock2Vec, a sentiment word embedding, demonstrated promising results in predicting stock price directions. This approach acknowledges the impact of external factors such as news sentiment on market behavior. Cheng et al. (2018)[9] introduced attention-based LSTM models for stock price prediction, bringing attention mechanisms into the spotlight. While attention has gained traction in neural machine translation, its application to stock prediction offers a novel perspective, enhancing model interpretability and trading strategy development. Yu et al., (2019)[14] treated financial product price data as a one-dimensional series, leveraging Deep Neural Networks (DNNs) to navigate the complexities of chaotic systems. Their use of time series phase-space reconstruction highlighted the potential of DNNs in handling nonlinear financial data. He and Zhang (2022)[8] compared the effectiveness of predicting stock prices using the CNN and LSTM methods. The researchers analyze data from Tata Consultancy Services and run multiple tests to evaluate the training and test loss and errors. The results show that the CNN method outperforms the LSTM method in terms of prediction accuracy in the short term. The financial market is a complex and dynamic system, making it challenging to predict stock prices using traditional methods. Therefore, utilizing advanced techniques like CNN and LSTM can lead to more accurate forecasts with fewer inputs and a simpler model. Ren et al., (2022) Y. and Dr. Kavitha (2023) [1] developed a stock price prediction model by combining Random Forest Regression and Sentiment Analysis. The unpredictability of the stock market makes it crucial to consider public opinions and sentiments, which can be influenced by various occurrences. Exploring the relationship between public perceptions expressed in tweets and changes in stock prices. A system is developed to collect and analyze tweets using the Random Forest Classifier model to determine their sentiment as positive or negative. The accuracy of the model is evaluated by comparing the predicted stock price changes with the actual changes the next day. The results show a strong association between stock price changes and public sentiments in tweets. The Random Forest model combined with sentiment analysis achieves an accuracy of 84.86% in predicting stock prices. This research contributes to forecasting future market behavior and can be useful for investors and financial analysts. Xu (2022)[19] performed stock price prediction using RNN, LSTM, and GRU algorithms on four stocks with different fluctuation types. The regression evaluation index is used to determine the applicability of the algorithms. The results show that the fluctuation of stock price significantly affects the accuracy of the algorithms. The LSTM algorithm performs best for stocks with large cyclical fluctuations, while the GRU algorithm performs best for stocks with a slump in price. This research highlights the importance of considering the characteristics of the stock data when selecting the appropriate algorithm for stock price prediction.

As foundational studies laid the groundwork, the second strand of literature have focused on hybrid models and optimization techniques. Agrawal et al., (2019) [13] optimized Long Short Term Memory (LSTM) networks using Correlation-Tensor built with Stock Technical Indicators (STIs). This approach aimed at refining deep learning tasks by incorporating adaptive indicators, showcasing a fusion of traditional finance and modern deep learning techniques. Ren et al., (2022) mentioned the "optimized random forest model" and describes it as a model trained using the ant colony optimization algorithm. The experimental results show that the prediction errors of the optimized random forest model are smaller than those of other models, including the gradient boosting decision tree (GBDT) model and the decision tree (DT) model. The optimized random forest model is an iterative optimization process that selects weighted random forest parameters. It uses the ant colony optimization algorithm to optimize the random forest parameters and predicts stock prices with improved accuracy. Jaiswal and Singh (2022)[6] proposes a hybrid convolutional recurrent model for stock price prediction in the financial world. The model combines the properties of 1D-CNN and GRU to leverage feature extraction and temporal regression tasks. The performance of the proposed model is evaluated and compared with existing hybrid models through experiments. The results demonstrate that the CNN-GRU model outperforms in stock price prediction. The use of deep learning models in handling large data and making accurate predictions is becoming increasingly important in the field of finance. Shi et al., (2019)[10] introduced DeepClue, a system bridging text-based deep learning models with end-users through visual interpretation. This approach emphasizes the importance of making deep learning insights accessible to non-experts, enhancing the applicability of predictive models in financial decision-making. Rezaei et al., (2021)[5] addressed the nonlinearity and volatility of financial time series by proposing hybrid algorithms—CEEMD-CNN-LSTM and EMD-CNN-LSTM. These algorithms, leveraging Empirical Mode Decomposition (EMD) and Complete Ensemble EMD (CEEMD), demonstrated effectiveness in extracting deep features for accurate one-step-ahead predictions. Li et al., (2021)[20] showcased the effectiveness of ensemble deep learning technologies in predicting future stock price trends. Their study emphasized the integration of textual, numerical, and graphical data in financial analysis, highlighting the comprehensive nature of modern machine learning approaches. Xu, X. et al. (2022) proposes a hybrid and improved LSTM-CNN model for predicting the trend of Chinese stock prices. The classical linear prediction models are not suitable for this task due to the complexity and dynamics of the stock market. Experimental results show that the LSTM-CNN model outperforms other models in terms of accuracy and risk. Friday, I.K. et al. (2022)[4] DL techniques, such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have shown significant improvements in performance compared to traditional time series models. However, these models are constantly evolving to become more efficient and effective, so propose an Integrated RNN-GRU model for stock market price prediction, using four independent features and the prices from the previous ten days, to forecast the price for the 11th day. In conclusion, the integration of deep learning techniques in stock price prediction has witnessed substantial progress. From foundational studies to hybrid models, the field continues to evolve, offering innovative solutions and contributing to more accurate forecasts, benefitting investors and financial analysts alike.

3. Methodology

3.1 ANN

Artificial Neural Networks (ANNs), also known as neural networks, are algorithms designed to imitate the behavior of neurons in the human brain. They consist of numerous interconnected nodes or neurons, each representing an output function called an activation function. The connections between nodes carry weighted values, known as weights, influencing the network's output. This output depends on the network structure, the assigned weights, and the activation function used. ANNs typically employ non-linear activation functions for their ability to model intricate relationships in data. The learning process is often supervised, where the model is trained on labeled data.(Kan,2020)ANNs exhibit characteristics such as distributed processing, self-organization, self-adaptation, and self-learning, making them particularly advantageous for tasks like predicting stock prices. The application of ANNs in stock price prediction has gained significant attention globally.(Kan,2020)Despite their effectiveness, ANNs may face challenges, particularly the risk of overfitting, especially when dealing with a large number of parameters. Overfitting occurs when the model learns noise in the training data rather than the actual underlying patterns. [7]

3.2 Random Forest

In the field of big data and artificial intelligence, random forest is a very popular machine learning algorithm. It is an ensemble learning method that improves the prediction accuracy and stability of a model by constructing multiple decision trees (The nets are static red et al., 2023) [17]. This paper will introduce the basic concepts, working principles, application scenarios and precautions of random forest.

First, basic concepts

Random forest is a model composed of multiple decision trees based on random selection. During the construction of each decision tree, a part of the training data and features are randomly selected, thus reducing the overfitting risk of the model. At each node, a feature is randomly selected for splitting to improve the generalization ability of the model. Eventually, multiple decision trees output a vote to determine the final prediction.[18]

Second, the working principle

The working principle of random forest mainly includes two parts: the construction of decision tree and the output of voting results. In the process of decision tree construction, the overfitting risk of the model is reduced by random sampling of data and random selection of features. At each node, a feature is randomly selected to split, which improves the generalization ability of the model. When voting results are output, multiple decision trees output a vote result to determine the final prediction result.

When the sample features have class K, the probability of class k is denoted as p_k The expression of the Gini index is

$$\text{Gini(s)}=\sum_{k=1}^{K} p_K(1\text{-}p_k)$$

If the sample set D is divided into D_1 based on some feature A And D_2, then under the condition of feature A, the set D The Gini index is defined as

$$\text{Gini(D,A)}=\frac{D_1}{D}\text{Gini}(D_1)+\frac{D_2}{D}\text{Gini}(D_2)$$

4. Application scenarios

Random forest has a wide range of applications in various fields, such as finance, healthcare, bioinformatics, etc. In finance, random forests can be used for tasks such as risk assessment, credit scoring, and fraud detection. In the medical field, random forests can be used for tasks such as disease prediction, drug discovery, and genomics research (BOTELHO VALADARES, C. et al., 2023)[2]. In bioinformatics, random forests can be used for tasks such as gene expression analysis, protein interaction prediction, and biological network analysis.

4.1 SHAP

SHAP is an approach inspired by game theory that aims to interpret the predictions of machine learning models. The goal of SHAP is to interpret the prediction of instance x by calculating the contribution of each feature to the prediction process, calculating Shapley based on joint game theory (Lundberg and Lee, 2017).[11] SHAP is the interpretation of the Shapley value as an additional feature attribution method, a linear function. The interpretation of the shap designation is g (z') $=\emptyset_0+\sum_{j=1}^{M}\emptyset_j z'_j$

SHAP is based on the size and calculation of the feature attribute. By comparing the feature importance predicted by the model, this is a fair comparison process to show the influence of input features in the prediction process. The impact of the features of the predicted model is shown in the form of a bar chart showing the global importance of the features. In addition, the SHAP summary graph combines the importance of features with the feature effects, which shows the distribution of shape values for each feature.

4.2 XGBoost

The XGB model optimizes the structural loss function by utilizing Lasso and Ridge regularization to mitigate the degree of model fitting and minimize structural risk (Nagaraj et al, 2022)[16]. The model uses the first-order and second-order derivatives of the loss function to generate weak learners and the linear classifier is supported as the base model. In addition, the algorithm also provides parallel computation of feature dimensions to generate decision trees, which improves the speed. Parallelization is made possible by the ability to switch the base learner between internal and external. Nagaraj and Krishna Prasad (2017) [15] stated that XGB utilizes multiple kernel eigenvalues on the CPU and organizes them as a block structure for pre-ordering. This structure is then reused in subsequent iterations, resulting in a significant reduction in computational workload. XGB's ultimate objective function is

$$\mathcal{L}^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2}\left(H_j + \lambda\right) {w_j}^{2} \right] + \gamma T$$

Note:

G_j : The sum of the first-order partial derivatives of the samples contained in the leaf node is a constant.

H_j: The sum of the second-order partial derivatives of the samples contained in the leaf node is a constant.

5. Results

The data mainly reflects the analysis of stock data of different companies using neural networks, random forests, and XGBoost. The accuracy of the model for data analysis is discussed in different ways.

Table 1 Stock data of different companies using neural networks

	APPLE	Amazon	TESLA	NIKE
Random Forest	7.397770600351396	10.22978432650263	46.29803313764423	7.078491476513637
XGBoost	5.773883385039272	8.10999126827231	37.90903619802106	5.375461628473902
Neural Network	4.557059533374314	6.872626895856549}	29.644441948306458	4.232815651866981
MSE	5.242945831128608	7.326139326033313	62.0566721050493	6.494901577115461

From the table 1, it can be seen that the accuracy of stock prediction for each company was analyzed using three models: neural network, random forest, and XGBoost. The accuracy value of analyzing Apple using neural network models is approximately 4.56, which is lower than the accuracy values of the random forest model and XGBoost model, which are 7.4 and 5.77, respectively. The comparison results of accuracy values for the other three companies are the same. Among the results obtained, the neural network model had the smallest numerical value. Therefore, using the neural network model for stock analysis of Apple, Amazon, Tesla, and Nike was the most accurate. From the perspective of the accuracy and error of the neural network model, the accuracy analyzed by this model is very high. However, Tesla has the largest error value, with a value of about 62.06, which is much higher than Apple, Amazon, and Nike, with values of 5.24, 7.33, and 6.49, respectively. Therefore, this model cannot be used solely to analyze Tesla. For the Random Forest and XGBoosts models, Nike has the highest accuracy in analyzing Tesla, with values of 7.08 and 5.38, followed by Apple with values of 7.4 and 5.77, followed by Amazon with values of 10.23 and 8.11, respectively. Moreover, the accuracy of analyzing Tesla is much higher than the other three companies, reaching 46.3 and 37.91.

In contrast, the neural network model is more prominent in the processing of unstructured data, images, speech, and other fields. Neural network models have powerful learning and modeling capabilities and can capture nonlinear relationships and complex patterns in data. They have achieved remarkable success in tasks such as image recognition and natural language processing.

However, there are some challenges and limitations to neural network models. They require large amounts of data to train and can be prone to overfitting. In addition, the structure and hyperparameter selection of neural network models are crucial to the performance of the models and need to be properly adjusted and optimized.

Therefore, when selecting the method of data analysis, the appropriate model should be selected according to the characteristics of the specific problem and the nature of the data. Random Forest and XGBoost are generally good choices when dealing with structured and high-dimensional data, while neural network models may be more beneficial when dealing with unstructured data and complex tasks.

6. Conclusion

To sum up, the neural network model is more prominent in the processing of unstructured data, images, speech and other fields. Neural network models have powerful learning and modeling capabilities to capture nonlinear relationships and complex patterns in data, and have achieved remarkable success in tasks such as image recognition and natural language processing. However, neural network models also have some challenges and limitations. They require large amounts of data to train and are prone to overfitting. In addition, the structure and hyperparameter selection of neural network models are crucial to the performance of the models and need to be properly adjusted and optimized. Therefore, when selecting the data analysis method, the appropriate model should be selected according to the characteristics of the specific problem and the nature of the data. Random forests and XGBoost are generally good choices when dealing with structured and high-dimensional data, while neural network models may be better suited for dealing with unstructured data and complex tasks.

References

[1] A1, Y., & Dr. Kavitha. (2023). Building a stock price prediction model using random forest regression ... Retrieved from https://www.researchgate.net/publication/369483092_Building_a_Stock_Price_Prediction_Model_using_Random_Forest_Regression_and_Sentimental_Analysis

[2] BOTELHO VALADARES, C. et al. Genome-enabled prediction through quantile random forest for complex traits. Ciência Rural, [s. l.], v. 53, n. 10, p. 1–6, 2023. DOI 10.1590/0103-8478cr20220327. Disponível em: https://search-ebscohost-com-s.elink.xjtlu.edu.cn:443/login.aspx?direct=true&db=asn&AN=163202191&site=eds-live&scope=site. Acesso em: 9 dez. 2023.

[3] Dang Lien Minh; Abolghasem Sadeghi-Niaraki; Huynh Duc Huy; Kyungbok Min; Hyeonjoon Moon; "Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network", IEEE ACCESS, 2018.

[4] Friday, I. K., Godslove, J. F., Nayak, D. S. K., & Prusty, S. (2022). IRGM: An Integrated RNN-GRU Model for Stock Market Price Prediction. In *2022 International Conference on Machine Learning, Computer Systems and Security (MLCSS)* (pp. 129–132). [https://doi-org-s.elink.xjtlu.edu.cn:443/10.1109/MLCSS57186.2022.00031]

[5] Hadi Rezaei; Hamidreza Faaljou; Gholamreza Mansourfar; "Stock Price Prediction Using Deep Learning and Frequency Decomposition", EXPERT SYST. APPL., 2021.

[6] Jaiswal, R., & Singh, B. (2022). A Hybrid Convolutional Recurrent (CNN-GRU) Model for Stock Price Prediction. In *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 299-304). [https://doi.org/10.1109/CSNT54456. 2022.9787651]

[7] Jiawei Long; Zhaopeng Chen; Weibing He; Taiyu Wu; Jiangtao Ren; "An Integrated Framework of Deep Learning and Knowledge Graph for Prediction of Stock Price Trend: An Application in Chinese Stock Exchange Market", APPL. SOFT COMPUT., 2020.

[8] Kexin, H., & Zhijin, Z. (2022). Retrieved from https://www.clausiuspress.com/article/5231.html

[9] Li-Chen Cheng; Yu-Hsiang Huang; Mu-En Wu; "Applied Attention-based LSTM Neural Networks in Stock Prediction", 2018 IEEE International Conference On Big Data (Big Data), 2018.

[10] Lei Shi; Zhiyang Teng; Le Wang; Yue Zhang; Alexander Binder; "DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction", IEEE Transactions On Knowledge And Data Engineering, 2019.

[11] Lundberg, S. M., & Lee, S. -I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768–4777).

[12] M Hiransha; E. A. Gopalakrishnan; Vijay Krishna Menon; K. P. Soman; "NSE Stock Market Prediction Using Deep-Learning Models", Procedia Computer Science, 2018.

[13] Manish Agrawal; Asif Ullah Khan; Piyush Kumar Shukla; "Stock Price Prediction Using Technical Indicators: A Predictive Model Using Optimal Deep Learning", International Journal Of Recent Technology And Engineering, 2019.

[14] Pengfei Yu; Xuesong Yan; "Stock Price Prediction Based on Deep Neural Networks", Neural Computing And Applications, 2019.

[15] P Nagaraj and A.V. Krishna Prasad, "A Cloud Computing Emerging Security Threats and Its Novel Trends in Knowledge Management Perception", International Journal of Emerging Technology and Advanced Engineering, vol. 7, no. 2, December 2017.

[16] P. Nagaraj, M. V. Dass, E. Mahender, and K. R. Kumar, "Breast Cancer Risk Detection using XGB Classification Machine Learning Technique," 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET), Bhopal, India, 2022, pp. 1-5, Doi: 10.1109/CCET56606.2022.10080076.

[17] Zhang Zhenyu; Deng P. A method of Los/Nlos base station identification based on random forest. Telecommunication Engineering, [s. l.], v. 63, n. 10, p. 1596–1602, 2023. DOI 10.20079/j.issn.1001-893x.220320001. Disponível em:https://search-ebscohost-com-s.elink.xjtlu.edu.cn:443/login.aspx?direct=true&db=bsu&AN=173327394&site=eds-live&scope=site. Acesso em: 9 dez. 2023.

[18] Wanjawa, B. W., & Muchemi, L. (2014). Retrieved from https://arxiv.org/abs/1502.06434

[19] Xu, X., Yang, M., Liu, H., & Zhang, D. (2022). A hybrid improved LSTM-CNN model for Chinese stock price trend prediction. In *2022 IEEE 4th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (pp. 76-83). [https://doi.org/10.1109/ICCASIT55263.2022.9986705]

[20] Yang Li; Yi Pan; "A Novel Ensemble Deep Learning Model for Stock Prediction Based on Stock Prices and News", International Journal Of Data Science And Analytics, 2021.