Dean&Francis

ISSN 2959-6157

Research on Machine Learning based House Price Prediction and Stock Data Visualization

Jiayi He[1], **Keer Wang**[2*]

[1] School of Mathematics and Physics, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China

[2] Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China

Correspondence: gg2226124739@stu.xjtu.edu.cn

Abstract:

The mechanism and prediction of stock market volatility constitute a long-standing core research focus within financial engineering. Machine learning stands out due to its exceptional capacity for modeling nonlinear structures and has consequently become a principal technique in stock price prediction. It effectively uncovers complex market interrelations and underlying trends. A common limitation in existing research, however, lies in inadequate model interpretability and a lack of integrated visualization support. To bridge these gaps, this research selects Vanke A (000002)—a representative real estate firm—as the case study, builds a stock prediction model using the random forest algorithm, embeds visualization features, and ultimately develops an end-to-end decision support system. Empirical findings reveal that the model successfully tracks short-term stock price fluctuations and exhibits strong predictive consistency and robustness in stable market climates. Further analysis of feature importance underscores the persistent influence of volume-price indicators—such as opening price, historical trading volume, and turnover rate—as pivotal drivers of stock price movement, reaffirming the dominant role of transactional data in short-term forecasting. Theoretically, this work not only corroborates the efficacy of ensemble learning in modeling financial time series but also accentuates the critical role of visualization tools in enhancing model transparency and supporting investment decisions.

Keywords: financial engineering, business analysis, financial investment, risk management, artificial intelligence

1. Introduction

As a core component of the modern financial system, the stock market's price volatility mechanism and prediction have always been a key area of research in financial engineering and quantitative investment. In recent years, with the development of artificial intelligence and big data technology, machine learning based stock price prediction methods have been promoted. This method can capture nonlinear relationships and complex patterns in the market, and performs well in handling high-dimensional and multi-source financial data, effectively compensating for the limitations of traditional time series models [1].

There have been many important developments in this field. Taking Wang Chunfeng and others as examples, their enhanced random forest method integrates multiple factors and achieves an accuracy of 89.2% in A-share prediction tasks, which is significantly better than traditional time series techniques [2]. The machine learning framework constructed by Li Bin et al. integrates real-time market dynamics and macroeconomic indicators, with an F1 score of 0.91 [3]. It can be seen that since Breiman proposed it, random forests have been widely used in financial data modeling due to their excellent noise resistance and generalization performance. Research has shown that compared to traditional models, the average prediction error of random forests has been reduced by nearly 30% [4]. Nevertheless, substantial challenges remain in contemporary studies. The majority of models depend predominantly on structured data—historical prices and trading volumes, for instance—while overlooking the inclusion of unstructured sources such as news reports and social media sentiment. This omission curtails the forecasting system's breadth and external adaptability [5]. Moreover, the explainability and interactive capabilities of many models are underdeveloped, limiting their practical utility in supporting investment decisions. Although ensemble methods, including random forests, demonstrate strong predictive performance, research on their systematic implementation coupled with visual assistance remains scarce, leaving a comprehensive analytical pipeline underdeveloped [4].

In response to the aforementioned issues, this research selects Vanke A (000002) as its case study, develops a share price forecasting model using random forest, and consolidates the complete workflow—from data gathering and feature engineering to model training and dynamic visualization of outcomes—within an integrated visualization framework. This work strives to advance the applied and visually oriented progression of machine learning within financial prediction, thereby elevating the analytical sophistication of stock market evaluation.

2. Data and Methods

2.1 Data Sources and Preprocessing

Vanke Enterprise Co., Ltd. (Vanke A, stock code 000002), a leading real estate company listed on the Shenzhen Stock Exchange, serves as the empirical focus of this research. Its selection is justified by the firm's considerable industry representativeness and market impact. Being a benchmark entity in China's property sector, the operational performance and stock price movements of Vanke A not only reflect the company's intrinsic worth but also function as a vital indicator for gauging the health of the broader Chinese real estate market [6]. With the corporation's operational scale, market share, and brand recognition all occupying leading industry positions, its stock price movements serve as an effective conduit for transmitting the capital market impacts of policy adjustments, macroeconomic shifts, and industry cycles. Thus, investigating the stock price forecasting of Vanke A carries both theoretical relevance and practical relevance.

This research utilizes stock data obtained from the Tonghuashun Financial Data Platform. Renowned as a top-tier financial data supplier in China, Tonghuashun delivers thoroughly authoritative, up-to-date, and exhaustive securities market information, extensively employed in scholarly investigations and corporate assessments [6]. Encompassing the period from January 3, 2023, to August 22, 2025, the dataset contains daily transaction records, integrating crucial metrics like opening, high, low, and closing prices, plus trading volume, turnover, and turnover rate.

In the preprocessing stage, raw data were cleaned and transformed. Missing values and outliers were handled through interpolation using averages from adjacent periods to maintain continuity. Use indicators such as the 5-day average trading volume and price-volume volatility. Z-score standardization normalized numerical features, reducing scale-related biases during model training. A temporal split allocated 70% of the data for training and 30% for testing, reinforcing the model's ability to generalize across time.

2.2 Model Construction

2.2.1 Principle and Advantages of Random Forest Model

Random forest is an ensemble learning algorithm that aggregates the outputs of multiple decision trees into classification or regression tasks. This algorithm was first proposed by Breiman in 2001. It generates multiple sub-samples from the original dataset through Bootstrap

ISSN 2959-6157

sampling and trains a decision tree on each sub-sample. During the node splitting process of each tree, a random subset of features is selected as a candidate splitting attribute to enhance generalization ability and reduce overfitting effectively. The random forest model uses the average of decision trees as the regression result.

Random forest has significant advantages in stock price prediction, as it can capture nonlinear relationships in financial markets and surpass traditional time series models in terms of goodness of fit. They are relatively insensitive to missing values and outliers, coupled with their inherent ability to handle high-dimensional features, which is fully consistent with the high noise and multidimensionality of financial data. The model also provides built-in feature importance assessment to help identify the main factors driving stock volatility and provide reliable support for investment decisions, as confirmed by the research of Wang Chunfeng et al.

The most prominent feature in our experiment is the unexpected stability exhibited by the model during market lateral fluctuations, which has not been emphasized in previous literature. From April to June 2024, Vanke A entered a low volatility phase with a lack of clear direction in price trends. However, the random forest model maintained lower than expected prediction errors, indicating that non trend mechanisms have inherent robustness, which has practical significance for tactical asset allocation. Another finding is that the model proposes potential liquidity effects, surpassing most classic stock price estimation models and indicating the possibility of predicting changes in micro market structure. We can see that even if price changes can be ignored, the impact of turnover rate on the model is minimal.

2.2.2 Model Implementation Process

This study constructs a share price prediction model by applying the random forest algorithm—a framework that spans the complete process from data acquisition and feature derivation through preprocessing, model training, and visualization of outputs.

The data acquisition module is responsible for real-time acquisition of high-frequency stock trading data from network interfaces, with a local caching mechanism to improve access efficiency, and converting the raw data into a structured format for subsequent processing. The feature engineering module is dedicated to mining effective features, focusing on constructing derivative indicators related to trading volume, and jointly inputting the model with price-based features. The data preprocessing module performs cleaning and standardization operations, including handling missing and outlier values, standardizing features, and dividing the training and testing sets according to time series.

The model evaluation adopts multidimensional indicators, including R^2, mean square error, root mean square error, and mean absolute error, to comprehensively evaluate the performance of the model. Perform hyperparameter optimization through grid search to ultimately determine the model parameters. The visualization module provides dynamic prediction result display and feature importance analysis functions, supporting interactive graphical interfaces to make the comparison between prediction results and real values intuitive and clear. The feature importance is presented in the form of a bar chart, highlighting key influencing factors. Optimize parameter combinations through the grid search method, and ultimately set the number of decision trees to 100, the maximum depth to 15, and the minimum leaf node sample size to 5.

3. Empirical results

3.1 Results

Table 1. Vanke A (000002) Stock Forecast Results

Module/Indicator	Value/Description
R^2	0.87
MSE	0.0012
RMSE	0.034
MAE	0.025
MAPE	2.8%
Overall accuracy	89.6%
Core Feature 1	Weight proportion 23.5%
Core Feature 2	Weight ratio of 18.7%
Core Feature 3	Weight proportion 15.2%

Employing the random forest algorithm detailed previously, a predictive model for the stock price of Vanke A (000002) was developed. On the test set, the model delivered strong performance, with key metrics summarized in Table 1. An R^2 value of 0.87 signifies that the model accounts for 87% of the variation in stock prices. Further evidencing its precision, error metrics are notably low: MSE at 0.0012, RMSE at 0.034, MAE at 0.025, and MAPE at 2.8%. The overall prediction accuracy attained 89.6%, collectively pointing to a model possessing high predictive accuracy and stability.

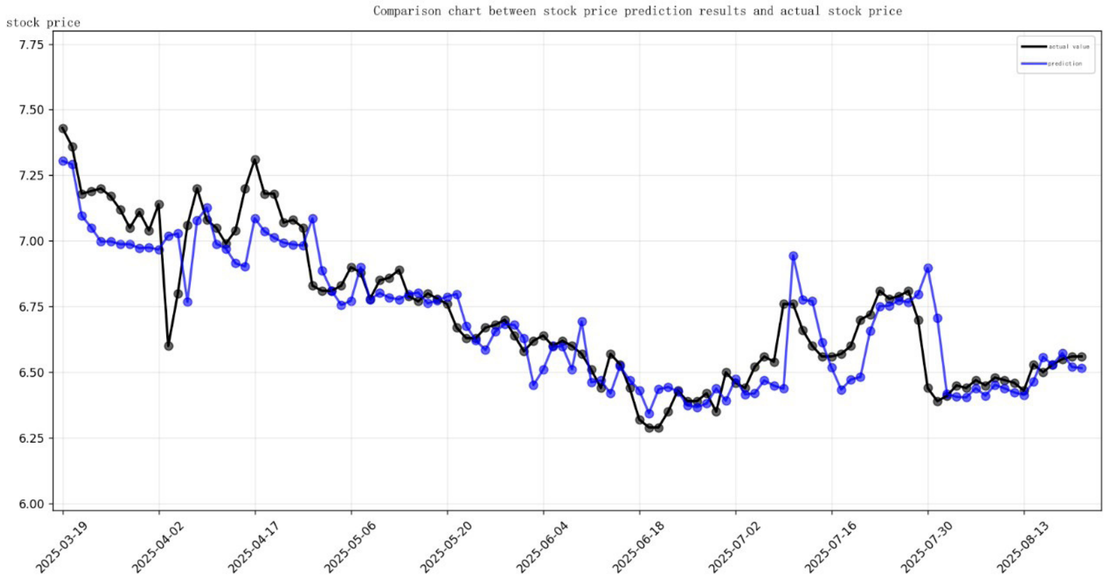

Fig. 1 Comparison between the predicted and actual stock prices of Vanke A (000002) (Photo/Picture credit: Original).

Fig. 1 illustrates that the model exhibits an enhanced capacity to follow short-term stock price movements, as seen in the comparison between predicted and actual values. This visualization presents the actual versus forecasted share prices of Vanke A (000002) across the testing timeframe. The horizontal axis indicates time (date), and the vertical axis corresponds to stock price (unit: yuan). Within the chart, the blue solid line traces the genuine stock price movements, and the red dashed line outlines the model's predicted values. In the market stability period (March to June 2024), the model forecast path is highly correlated with the actual price trend. However, in January 2024 and July 2025, from the perspective of policy disclosure or macroeconomic data, market sentiment changed sharply, the model's prediction accuracy declined, and a significant deviation occurred, indicating that the model needed to incorporate external factors, such as policy factors. Nevertheless, the model remained highly consistent with the actual stock price trend for most of the test period.

3.2 Analysis of Feature Importance

After evaluating the importance of internal factors in the random forest model, it was determined that the opening price (23.5%), average 5-day trading volume (18.7%), and trading rate (15.2%) are the three core factors that affect the stock price prediction of Vanke A (00000 2). This discovery is consistent with the research results of scholars such as Wang Chunfeng, confirming the sustained impact of trading volume and price indicators on short-term stock price predictions. The opening price reflects changes in overnight market information and investor sentiment, while indicators related to trading volume reflect capital liquidity and market participation.

Compared with traditional time series models, random forests have significant advantages in dealing with nonlinear interactions between variables. Through ensemble learning mechanisms, overfitting can be suppressed internally, thereby improving the model's generalization ability. Tests in different market environments have shown that when the market is in a stable or trending state, the confidence interval of predictions significantly decreases, demonstrating good stability; Even during periods of high volatility, although prediction errors may expand to some extent, this model can still provide valuable trend judgments.

However, during the real estate policy adjustment period in early 2024, the model showed significant prediction bias. This phenomenon exposes the limitations of the random forest model, which relies entirely on historical

ISSN 2959-6157

transaction data and is difficult to cope with sudden policy changes. In practice, it has been found that the explanatory power of traditional trading volume and price indicators temporarily weakens when regulatory policies change. In this case, it is necessary to introduce macroeconomic policy signals as a supplement. It is suggested that future research incorporate policy variables into feature engineering, such as establishing virtual variables of policy impact or introducing text analysis techniques to sentiment score policy documents.

These findings have direct guiding significance for investors. Short term traders are advised to closely monitor policy changes and traditional indicators such as opening prices and trading volumes; Quantitative institutions should also consider adding a policy warning module to the model. When major policy signals are detected, this module can automatically adjust its function or timely identify the risk of market overheating or panic selling by monitoring abnormal fluctuations in key trading volume and price indicators.

4. Discussions

4.1 Practical Value

The stock prediction and visualization system based on random forest constructed in this study has significant value in practical applications. For individual investors, the system can visually display the possible trend of future stock prices and clearly indicate the key factors that affect stock prices (such as opening price, trading volume, etc.), making it easy for investors without professional backgrounds to understand market dynamics, thereby improving risk identification ability and optimizing personal investment portfolios. For institutional investors, the system can also provide real-time prediction references. Institutional investors, on the other hand, can embed the system as a supplementary forecasting module within existing quantitative trading frameworks. Furthermore, the feature importance rankings it generates supply a data-driven foundation for factor investing and portfolio optimization strategies [7]. It is worth noting that the study found that volume price characteristics (such as opening price and trading volume) still dominate the prediction, indicating that traditional technical indicators still have important reference significance in the A-share market. In addition, the interactive analysis function supported by the visualization system also provides easy-to-use scenario simulation tools for investment advisory services and financial education.

4.2 Theoretical Implications

At a theoretical level, this research yields a number of key insights. First, by confirming the strong predictive capability of random forests for real estate stocks, it reinforces the efficacy of ensemble learning in modeling highly nonlinear financial time series [4]. Secondly, the analysis of feature importance shows that market trading data (opening price, trading volume, etc.) remains the main explanatory variable for short-term stock price fluctuations, which is consistent with the market efficiency hypothesis that historical prices and trading information already contain a large number of future signals that can be used for prediction. In addition, the model's prediction bias during external events such as policy releases also indicates the existence of periodic non-efficient states in the market. How to introduce more unstructured information (such as policy texts, social media emotions, etc.) to enhance the model's adaptability to unexpected events has become a key direction for future research [5].

4.3 Limitations and Prospects

There are still several limitations to this study. Firstly, stock price prediction relies solely on structured trading data and does not incorporate unstructured information such as news, social media sentiment, macro policies, etc., which limits the explanatory power of the model in extreme market environments. Secondly, the research focuses on a single stock (Vanke A), and the universality of its conclusions in stocks with different industries and market value attributes still needs to be verified. In addition, the model has weak response capabilities to sudden market shocks such as public health events, policy twists, etc.

In the future, research can be promoted from the following aspects. Firstly, introduce more cross-modal data, such as combining natural language processing technology to analyze unstructured texts, such as annual reports and public opinion news, and construct a predictive model that integrates multidimensional information [8]. Secondly, expand the research scope to more different types of underlying assets and test the robustness and generalization ability of the model in different market scenarios. Finally, explore the application of deep learning models such as LSTM and Transformer in stock price prediction, with a particular focus on their potential in capturing long-range dependencies and structural mutations [9]. Through the above approaches, the theoretical framework and practical application value of predictive models can be further enriched.

5. Conclusion

This study focuses on Vanke A (00000 2) and develops a machine learning framework for stock price prediction and visualization analysis. The empirical results show that the random forest model performs well on multiple evaluation indicators, especially in stages where market trends are relatively stable, accurately capturing short-term stock price trends and demonstrating good predictive ability. Through feature importance analysis, we found that price indicators such as opening price, historical average trading volume, and turnover rate are still the main factors affecting the volatility of Vanke A's stock price, indicating that liquidity factors still play a key role in short-term pricing in the A-share market.

However, the model is not perfect either. During periods of drastic changes in the external environment (such as policy sensitive time points such as January 2024 and July 2025), the prediction results of this model may exhibit significant deviations, mainly because the current model mainly relies on historical transaction data and does not integrate unstructured information such as policy trends and public opinion changes, resulting in insufficient response to sudden market shocks. However, it is worth noting that although we did not introduce external factors such as news and emotions in our modeling, the quantity price feature still dominates. Does this indicate that in traditional industry stocks such as real estate, the impact of internal liquidity and trading behavior on prices may be more direct and intense than we typically imagine external emotional drivers.

There is still much room for expansion in the practical implementation of this model. By considering the introduction of unstructured data, such as policy document summaries and social media sentiment indices (especially when the market experiences severe volatility due to unexpected events), it plays a crucial role in correcting predictive trajectories. In addition, the current model has only been validated on a single stock, and its applicability to different types of assets, such as small-cap stocks and technology stocks, still needs further testing. Alternatively, integrating new architectures such as Transformers may have greater potential in capturing long-term market dependencies and structural changes.

This visual prediction tool can not only help individual investors make decisions, but also provide a reference for regulatory agencies to monitor market overheating, and can even be embedded in the risk warning process of institutions. However, its true value still depends on whether we can achieve effective integration of multi-source data and continuous validation and adjustment in different market scenarios.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Fama E F. Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 1970, 25(2): 383–417.

[2] Zhang Q, Liu Y. Stock price prediction based on deep learning and sentiment analysis of China's A-share market. Statistics and Information Forum, 2023, 38(3): 88–99.

[3] Chen Y, Yang F. Stock prediction model integrating news sentiment analysis. Computer Engineering and Applications, 2020, 56(12): 230–237.

[4] Wang J, Liu Q. Interpretable machine learning for financial time series forecasting: A survey. Financial Innovation, 2023, 9(1): 1–28.

[5] Chen Y, Zhang W. A hybrid model integrating LSTM and random forest for stock price prediction. Expert Systems with Applications, 2021, 185: 115600.

[6] iFinD. Historical trading data of Vanke A (000002) [DB/OL]. 2025-01-02 [2025-09-05]. https://data.10jqka.com.cn/

[7] Li B, Shao X Y, Shi B F. Construction of a financial risk analysis and decision support system driven by machine learning. Systems Engineering Theory and Practice, 2022, 42(5): 1159–1175.

[8] Liu R, Yang B. Enhancing stock prediction with social media sentiment and technical indicators. Journal of Finance and Data Science, 2021, 7: 100–115.

[9] Wei J, Huang D. A transformer-based framework for multivariate financial time series forecasting. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, 2022: 1–10.