Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 10, No. 2, 2024

Research on Predicting Stock Market Profitability Changes Based on Machine Learning

Gaofeng Xu [1], Xinchen Lu [2]

[1] Shanghai Urban Construction Vocational College, Shanghai, China
[2] Shanghai Shangde Experimental School, Shanghai, 200120, China

Abstract: The expansion of the market in our country has led to rapid economic growth, and many companies have begun to seek foreign investment for themselves. What investors are given is an invisible but very useful thing - stocks. However, stocks are unpredictable and subject to many influences, making it difficult for investors to predict whether the stock trend will decline or rise. It is uncertain whether investors will make a profit or lose money. The development of artificial intelligence technology has played a crucial role in driving stock model prediction. With the development and gradual maturity of machine learning technology, machine learning models are used for efficient and accurate analysis of massive data in the stock market, thereby achieving data mining in the stock market, analyzing changes in stock returns, and predicting them to improve investor returns.

Keywords: Artificial Intelligence; Machine Learning; Data Mining.

1. Introduction

Since AI was first proposed in 1950, it has caused a huge uproar. In the past stock market, it was often difficult for people to calculate when they would sell or hold. This dilemma lasted for hundreds of years. Newton has a famous saying: I can explore the mysteries of nature, but I cannot explore the dangers of human nature. What describes this is that people will only continue to have various events in the free market. After all, in the capitalist market, the government rarely intervenes in the development of the market. For example, the real estate foam in Miami in 1928 led to the great crisis in 1929, which made countless people homeless. This makes people extremely hopeful for the increase in their stock prices. In recent times, AI has been rapidly developing, especially in Google's open AI, which is becoming increasingly intelligent and capable of computing massive amounts of data [1-3], allowing people to calculate the profit margin of their own stocks. We will conduct research and exploration based on this development. The application of machine learning models [4-6] in quantitative investment strategies [7-9] has promoted the development and innovation of the fintech field. This not only promotes the application of new technologies in the financial sector, but also provides an example of technological innovation for other industries. Not only does it promote interdisciplinary integration, but it also combines machine learning with financial investment, promoting the cross integration between computer science, statistics, and economics. It can also drive innovation, and the rapid development of financial technology [10-12], especially in the fields of big data, cloud computing, and artificial intelligence, provides new ideas and tools for quantitative investment strategies, promoting innovation and development of the entire industry.

2. Related Work on Predicting Stock Market

2.1. Machine Learning in the Field of Investment Portfolios

In the field of investment portfolios, machine learning has shown great potential in stock price prediction, portfolio optimization, and has become an important tool for many financial institutions and investors Based on this, existing literature research on investment decision-making and predicting future stock returns can be roughly divided into the following three aspects: firstly, traditional econometric models represented by autoregressive moving average models Traditional investors heavily rely on historical data from financial markets and accumulate experience in investment decisions by establishing rigorous and complex econometric models The second is classical machine learning models represented by support vector machines. Machine learning methods have broken the limitations of traditional methods such as high cost and optimized stock prediction capabilities, which can further improve prediction accuracy and stability. The third is a deep learning model centered around neural networks Due to the success of deep neural networks in financial data prediction modeling, deep learning has become a promising financial prediction method.

2.2. SPtock Return Prediction Model based on Improved SVM

Based on multiple comprehensive considerations, including timeliness, representativeness, and computational cost of data, the study selected the stocks of the Shanghai and Shenzhen 300 Index as the research sample. There is a lot of useless information in the collected sample data, which leads to excessive redundancy of the sample data, increases the computational complexity of the model, and affects the prediction accuracy of the model. Therefore, data preprocessing is needed. By manually filtering, some enterprises with missing data can be deleted to reduce the complexity and redundancy of the data. In addition, there are also a large number of noise signals in the data samples, which seriously affect the quality of the data and significantly reduce the prediction accuracy and efficiency of the model. Therefore, it is necessary to filter and denoise them. The study uses wavelet transform algorithm to denoise the sample data.

2.3. Deep Learning Models

The core of deep learning lies in artificial neural networks,

which mimic human thinking processes to analyze and learn the technology of data. This process is called deep learning. Deep learning has demonstrated outstanding capabilities in feature extraction, capable of extracting deeper essential features from raw data, which is difficult to achieve with traditional machine learning methods. In traditional machine learning, features need to be manually designed, which is a complex and time-consuming process. Deep learning techniques can automatically identify features from data and require optimizing the learning process by setting appropriate network hierarchy and number of neurons. The multi-layered hidden layer structure of deep learning enables models to learn more complex abstract features, which are processed through multiple linear transformations with the ultimate goal of improving the prediction accuracy of the model.

2.4. Long Short-Term Memory

Long short-term memory network is a special type of recurrent neural network designed specifically to solve the long-term dependency problem encountered by standard RNNs when processing long sequence data. In traditional RNNs, due to the problem of vanishing or exploding gradients, the model finds it difficult to maintain and learn long-term dependencies in time series. LSTM solves this problem by introducing special structural units - LSTM units. Each LSTM unit contains three gate structures: input gate, forget gate, and output gate. These gate structures allow LSTM units to selectively store, update, or delete information, effectively maintaining and transmitting long-term state information. The input gate controls the entry of new information, the forget gate determines which information should be discarded, and the output gate controls the flow of information from the unit state to the output. The working principle diagram of the LTSM model is shown in Figure 1:

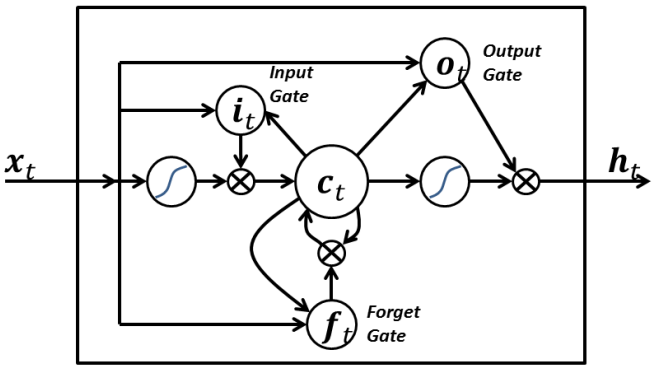

Figure 1. Working principle diagram of LTSM model.

2.5. Quantitative Investment

The importance of data in quantitative investment is self-evident. Investors will collect information including stock prices, trading volumes a large amount of historical and real-time data, including company financial data and market economy indicators. These data are not limited to traditional numerical data, but may also include unstructured data from social media, news reports, and other sources. Through comprehensive analysis of these data, quantitative investment aims to identify market trends and investment opportunities, providing scientific basis for investment decisions. The core of quantitative investment strategy is to establish and apply complex mathematical models and statistical algorithms. These models and algorithms are capable of processing and analyzing large amounts of data, helping investors predict market trends, assess risks, and build investment portfolios. In practical operation, quantitative investment often uses algorithmic trading, which involves executing buy and sell orders according to preset strategies through automated trading systems. This method not only improves transaction efficiency, but also helps reduce transaction costs and market shocks. However, quantitative investment is not without risks. The complexity and unpredictability of the market mean that even the most advanced models cannot completely eliminate investment risks. In addition, errors in the model and data may result in unexpected losses. Therefore, risk management has become an important component of quantitative investment, including the use of multiple risk control models to assess and manage potential risks.

Quantitative investment strategies guide investment decisions through the application of mathematical models, computer algorithms, and statistical analysis, and can be divided into multiple types, each operating based on different theories and market behaviors. The main classifications of quantitative investment strategies include momentum strategy, mean regression strategy, arbitrage strategy, factor investment strategy, event driven strategy, machine learning strategy, high-frequency trading, etc. Different strategies have their own characteristics and are suitable for different market environments and investment objectives. Quantitative investors typically choose and adjust suitable investment strategies based on market conditions, risk preferences, and investment periods. With the development of financial technology, these strategies are also constantly evolving to adapt to market changes.

3. Dataset

3.1. Experimental Dataset

Firstly, obtain two datasets as raw data: the historical dataset of the Shanghai Composite Index and the dataset of related financial news headlines The sampling time for both datasets is from November 1, 2012 to November 11, 2022 Preprocess the raw dataset of Shanghai Composite Index prices to obtain daily logarithmic returns and 7-day volatility, which will be used as the target variables for prediction On the other hand, calculate the sentiment score of each news headline and summarize it on a daily basis to construct an investor's accumulated sentiment index as part of the model input. The Shanghai Composite Index dataset consists of 2412 trading day observation data We further divided the dataset into two sets, namely the training set and the testing set, with a segmentation ratio of 8:2. That is, we used the first 80% of the data (1929 observations) to train the prediction model, and the remaining 20% (483 observations) as the testing dataset to evaluate the model. To ensure that the prediction model is robust and does not overfit the training data, we adopt a rolling prediction process, where the rolling window is set to 121 days There are a total of 4 rolling test cycles, with cycle 1 from November 16, 2020 to May 17, 2021, cycle 2 from May 18, 2021 to November 12, 2021, cycle 3 from November 13, 2021 to May 17, 2022, and cycle 4 from May 17, 2022 to November 10, 2022 The descriptive statistics of the daily logarithmic return, 7-day volatility, and cumulative investor sentiment index of the Shanghai Composite Index are listed in Table 1.

From Table 1, it can be seen that all three time series data exhibit high kurtosis values, indicating that they are all non-normally distributed. In addition, the enhanced Dickey G.

Fuller test (ADF) results of the three-time series are statistically significant at the 1% significance level, indicating that they are all stationary.

Table 1. Descriptive statistics of time-series data

	sample size	mean value	standard deviation	skewness
Yield rate	2412	0.62	0.09	-0.99
7-day volatility	2412	0.18	0.13	2.24

3.2. Data Processing

Through the GARCH model incorporating the cumulative sentiment index of investor news, we found that positive news shocks are negatively correlated with volatility, while negative news shocks are positively correlated with volatility, which can be understood as the game between irrational and rational investors after being impacted by different types of news This result indicates that news sentiment has an asymmetric impact on the volatility of the Shanghai Composite Index returns, with the market often reacting differently to positive and negative news.

3.3. Data Construction

The structural framework of the BERT model consists of an input layer, an encoding layer, and an output layer The input layer is used to represent text data as computer-readable embedded vectors In order to adapt to different semantics, the input of the BERT model includes three parts: word vector, segment vector, and position vector. After the text passes through the input layer of the BERT model, a tag is first added at the beginning, as well as one or more tags to separate sentences or indicate the end of the text If the text consists of several sentences, the dimension of the text vector obtained after processing will be converted to the dimension of the model's hidden layer The encoding layer is composed of multiple Transformer encoder parts stacked together, and the core structure is a self attention mechanism combined with residual connections and layer normalization, plus a feedforward neural network.

4. Experiments and Results

Accurately predicting securities returns and volatility is a prerequisite for optimizing asset allocation This article proposes an index based on news text sentiment to measure the cumulative sentiment of investors, and constructs a model that integrates the cumulative sentiment index of investors. The cumulative sentiment index of investors is incorporated into the daily logarithmic return and 7-day volatility prediction models of the Shanghai Composite Index Using event study method and a model with sentiment index, this study empirically examines the impact of investors' accumulated sentiment index on the daily logarithmic return and 7-day volatility of the Shanghai Composite Index Then, the variational mode decomposition is applied to decompose the historical data of daily logarithmic returns and 7-day volatility of the Shanghai Composite Index into various intrinsic modes The results indicate that market sentiment has a significant impact on the price fluctuations of the Shanghai Composite Index Both positive and negative news events have a significant impact on the price operation of the Shanghai Composite Index, and the impact of positive and negative shocks is asymmetric By comparing the prediction results, it was found that the deep learning model combined with the variational mode decomposition of investors' temporal and auditory cumulative sentiment index had better prediction performance than various econometric models and machine learning models This article proposes an attempt to integrate news text data into stock return and volatility prediction, demonstrating the feasibility of deep learning models in stock price volatility prediction In future research, more online data resources can be considered to assist in prediction Meanwhile, due to the characteristics of deep learning models, further optimization can be achieved by increasing data volume and other methods to improve prediction performance.

Evaluate the comprehensive performance of the model through the working characteristic curve of the subjects. It can be seen that the AUC value of the IFOA-SVM model reaches 0 954, which is 0. higher than the ECGNN model and IRF model, respectively 012 and 0 023. In summary, the constructed IFOA-SVM stock return prediction model can efficiently and accurately make intelligent predictions of stock returns, provide data support for investors' investment decisions, improve investors' investment returns, and promote the development of China's stock market.

5. Conclusion

The evaluation metrics of feedforward neural networks were found to be superior to the other six models. Then we further discussed the impact of the first fully connected layer, activation function, second fully connected layer, output layer, and optimizer on the prediction results of the feedforward neural network and optimized the model accordingly. Based on the prediction results of the feedforward neural network, we constructed specific investment strategies and analyzed the returns of the investment strategies. The experimental results show that feedforward neural networks are not only suitable for stock price prediction, but also a powerful and versatile network model, especially in terms of prediction accuracy.

Predicting stock returns is beneficial for providing data support to investors, improving their returns, and stimulating the development of China's stock market, which has a positive significance for the development of China's market economy. There have been many research achievements in predicting stock returns based on machine learning models, but the accuracy and efficiency are still not ideal.

By constructing a Shanghai and Shenzhen stock selection strategy based on feedforward neural networks and evaluating its performance using indicators such as total return, Sharpe ratio, and annualized return, the effectiveness of the feedforward neural network model was further explored. This strategy has performed well in both total and annualized returns, and has a high Sharpe ratio, which proves the effectiveness of feedforward neural networks and demonstrates their potential in cross-sectional stock selection applications.

References

[1] Cheng K , Su Y , Guan H ,et al.Mapping China's planted forests using high resolution imagery and massive amounts of crowdsourced samples[J].ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 196:356-371.DOI:10.1016/j.isprsjprs.2023.01.005.

[2] Huang D , Zhou J , Mi B ,et al.Key-Based Data Deduplication via Homomorphic NTRU for Internet of Vehicles [J].IEEE

Transactions on Vehicular Technology, 2023. DOI:10. 1109/ TVT. 2022.3205627.

[3] Agapito G , Cannataro M .An Overview on the Challenges and Limitations Using Cloud Computing in Healthcare Corporations [J]. Big Data and Cognitive Computing, 2023. DOI: 10. 3390/bdcc7020068.

[4] Carlos,Fernandez-Lozano,Rubén,et al.Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models[J].Journal of Theoretical Biology,2015.

[5] Montazeri M , Montazeri M , Montazeri M ,et al.Machine learning models in breast cancer survival prediction[J]. Technology & Health Care Official Journal of the European Society for Engineering & Medicine, 2015, 24(1):31.

[6] Faber F A , Hutchison L , Huang B ,et al.Fast machine learning models of electronic and energetic properties consistently reach approximation errors better than DFT accuracy[J]. 2017.DOI: 10. 48550/arXiv.1702.05532.

[7] Harvey C R .THE MASTERS SERIES Campbell R. Harvey, PhD: Examining Quantitative Investment Strategies[J]. 2022.

[8] Kou Z , Yu H , Peng J ,et al.Automate Strategy Finding with LLM in Quant investment[J]. 2024.

[9] Juddoo K , Malki I , Mathew S ,et al.An impact investment strategy[J].Review of Quantitative Finance and Accounting, 2023, 61(1):177-211.DOI:10.1007/s11156-023-01149-0.

[10] Kaiwen Z , Sen G , Guolei L .The impact of financial technology on China's low-carbon transformation of energy and mechanism analysis[J].Environmental Science and Pollution Research, 2024(4):31.

[11] Hakim A L , Zaerofi A , Mulyana R .Analysis of Small Medium Enterprise's Sukuk Investment Intention Through Financial Technology Securities Crowdfunding[J].Tazkia Islamic Finance and Business Review, 2023.DOI: 10.30993/ tifbr. v16i2. 301.

[12] Xuan J , Liu Y .Research on the Path of Financial Technology Enabling Wealth Management[J].Modern Economics & Management Forum, 2023, 4(3):55-57.DOI:10. 32629/memf. v4i3. 1374.