Emerging Markets Finance and Trade

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/mree20

Yield Forecasting by Machine Learning Algorithm: Evidence from China's A-share Market

Yue Hu, Haosheng Guo, Wenli Huang & Yueling Xu

To cite this article: Yue Hu, Haosheng Guo, Wenli Huang & Yueling Xu (2023) Yield Forecasting by Machine Learning Algorithm: Evidence from China's A-share Market, Emerging Markets Finance and Trade, 59:6, 1767-1781, DOI: 10.1080/1540496X.2022.2148464

To link to this article: https://doi.org/10.1080/1540496X.2022.2148464

EMERGING MARKETS FINANCE AND TRADE
2023, VOL. 59, NO. 6, 1767–1781
https://doi.org/10.1080/1540496X.2022.2148464

Routledge
Taylor & Francis Group

Yield Forecasting by Machine Learning Algorithm: Evidence from China's A-share Market

Yue Hu, Haosheng Guo, Wenli Huang, and Yueling Xu

China Academy of Financial Research, Zhejiang University of Finance & Economics, Hangzhou, China

ABSTRACT

This study uses five machine learning algorithms (Stochastic gradient descent (SGD), Decision tree, Random forest, Gradient boosting decision tree (GBDT), and Convolutional neural networks (CNN)) to explore their prediction effects on China's stock market. It constructs a monthly rolling model for stock return prediction. Selecting stocks of the CSI 300 index from January to June 2021 as specific monthly samples and classifying three factors – fundamentals, volatility(risk) and technical indicators, the results demonstrate that (1) machine learning brings favorable investment returns in simulated quantitative trading of China's A-share market (2) the technical indicator factor is the most valuable, with the momentum technical factor having greatest influence, followed by the volatility (risk) factor and fundamental factors. Therefore, this study has a critical reference value and is significant in guiding yield forecasting in intricate stock markets.

KEYWORDS

Machine learning; A-share market; yield forecasting; factor investing

1. Introduction

With the advancement of the local capital market, improvement of residents' income levels, and increase in real estate regulations, Chinese residents are now seeking more investment channels under the low-interest-rate environment. As a result, their funds are accelerating into the stock market; therefore, finding the accurate prediction model for China's A-share market can generate considerable economic value for investors. More than half a century of development of finance has allowed for the discover and accumulation of multiple factors that can effectively predict stock returns. Yet the paradigm of traditional finance based on these multiple assumptions has gradually faced limitations. This is because it is almost impossible to extract useful information from a large set of factors given the rambling order and low signal-to-noise ratios Moreover, the relationship between the stock return and its variables is difficult to derive directly. For example, standard linear models fail to depict the possible linear relationship between variables and dependent variables, increasing search complexity in the form of predictive functions. Thus, finding a non-linear relationship between stock returns and their factors is currently one of the research priorities of academia and industry.

From the early CAPM model, the Fama-French three-factor model, and the ARIMA model, scholars have explored the use of machine learning methods to solve the technical challenges brought on to the traditional research methods by the emergence of many factors. Akita et al. (2016) have used a combination of Deep learning and Long-short term memory models (LSTM) to predict the time series of financial data. By detecting 50 companies listed on the Tokyo Stock Exchange, they have found that the LTSM model's prediction accuracy of the text information within the input data can significantly increase profits. Gu et al. (2020) have systematically verified the effect of the underlying machine learning algorithm in stock return forecasting in the US market, showing that machine learning methods perform observably better than traditional linear

CONTACT Wenli Huang huangwenli@zufe.edu.cn China Academy of Financial Research, Zhejiang University of Finance & Economics, Hangzhou, 310018, China

regression models. Specifically, generalized linear model becomes infeasible without preconditions of multiway interactions, while machine learning algorithm such as regress tree is nonparametric and can perform better (Gu et al. 2020). Moreover, linear regression model has a weak performance when predicting the chaotic and high-noises factors such as stocks, which performs better when treated as a classified problem rather than regression problem (Basak et al. 2019). Markus et al. (2021) have used different machine learning models to develop a set of complex predictive indicators to test the Chinese stock market. They have found that liquidity indicators are equally important in different machine learning models, while the fundamental indicators which reflect the company's value are more secondary. In this study, we found out technical indicators are more influential compared to liquidity and fundamental indicators, meanwhile, Neural network performs outstandingly in predicting China's stock markets has been verified. Moreover, the presence of private investors makes stock prices easy to predict in the short term; while in the long run, large-cap stocks and state-owned enterprises are more predictable, and only long strategies can achieve significant gains after considering the transaction costs. In general, machine learning (the research hotspot in recent years) has been the primary method used to improve the accuracy of returns forecasting. However, even though machine learning methods have been widely used in the field, the literature of finance research is still in its infancy. Moreover, in the long run, uncertainty probably appears in any predictive model making the parameters of the model unstable, resulting in multiple problems. Is machine learning effective in predicting stock yields? Which algorithm can achieve better prediction results? If machine learning can improve investment returns, what factors matter? Can machine learning bring real economic value and benefit? This study takes data publicly disclosed by listed companies in China and the JoinQuant database to construct three significant factors – fundamental, volatility (risk), and technical indicators – and enters the processed data separately to build a monthly rolling model based on five different machine learning algorithms (SGD, GBDT, CNN, Decision tree and Random forest). It then applies the monthly frequency adjustment, and finally records the out-of-sample performance of A-share individual stocks while analyzing the effective factors. More importantly, a simple investment model is designed for testing, to demonstrate that machine learning and financial research can produce actual economic value.

The remainder of this study is arranged as follows: the second part gives a brief review of relevant literature; the third part introduces the research design and machine learning model of this study and explains the selection of factor data; the fourth part presents the empirical results and analysis; and the fifth part concludes the study.

2. Empirical and Theoretical Development of Equity Return Prediction

2.1. Literature Review

The question of predicting stock returns has been controversial for a long time. Fama (1970) proposed the efficient market hypothesis, which indicated that all the market information was already reflected in the effectiveness so that investors could not get excess returns from historical analysis. However, Campbell (2000) noted that many economic variables, including valuation ratios, interest rates and inflation, can achieve significant predictions of stock returns in the samples. On this basis, Jiang et al. (2018) constructed different portfolios based on factors such as the size of listed companies, stock concentration, and industry characteristics, to analyze the predictability of stock returns under China's stock markets. They showed that while the stock returns could be well predicted, there was no absolute contradiction between the predictability of stock returns and the efficient market hypothesis. Rapach and Zhou (2013) concluded through a study of U.S. stock premiums that the predictability of stock yields remains even in an efficient market; however, costs must be considered if excess returns are to be achieved – specifically, only the existence of trading frictions and risk adjustment. However, the expected return not equaling zero can be considered contradictory under the efficient market hypothesis.

In terms of the research on stock returns models, Markowitz (1952) established the mean-variance model to measure returns and risks; this was the first time that mathematical models were introduced into the financial investment field. Besides, Devpura et al. (2017) proves the stocks time-varying predictability on U.S stock markets sampling from 1927 to 2015. After that, the three-factor model (Fama and French 1993), the four-factor model (Jegadeesh and Titman 1993), and the five-factor model (Fama and French 1993) are all categorized under the asset pricing framework of the linear relationship. Later, Narayan et al. (2016) proposed the GARCH model to test US stock markets within the large samples, showing the stocks predictability features that stable prices stocks perform better than others. Later, Awartani and Corradi (2005) used the GARCH model, based on the time series characteristics of stock data, showing different time-before-and-after relationships and randomness to make regression predictions on future indices. They reported that the prediction model could accurately describe the volatility changes of the stock market on information pairing, albeit the effect was faint under asymmetric backgrounds. Recently, Li, Shao, and Li (2019) used 12 machine learning algorithms to predict stock prices based on 96 anomalies in the A-share market and demonstrated that portfolio strategies based on machine learning prediction could produce better investment returns.

Predictability performance resonates with the industry characteristics (Phan et al. 2015). By generating yield forecasting indicators from constructing models, Jiang, Qi, and Tang (2018) found that the lower the degree of investment friction, the higher the expected return of the stock in the company. Additionally, various indicators such as the gross margin and return on assets can also improve the prediction performance of the stock market yield. Specifically, Han et al. (2016) constructed the trend factor based on the moving average price, which performs well in interpreting cross-section stock returns, arguing that stock returns can be predicted based on reasonable technical indicators. Further, using seven international markets at a daily frequency, Chen, Jiang, and Tong (2017) found a negative correlation between the overnight yield of China's stock market and the international volatility. Obviously, stock returns can be predicted well within machine learning models, and more relevant forecasting indicators lead to more accurate results.

2.2. Machine Learning Algorithms for Prediction

This study uses five machine learning models for its predictions – Stochastic gradient descent, Decision tree, Random Forest, Gradient boosting decision tree, and Convolutional neural networks. Their functions and features are explained below.

Stochastic gradient descent, derived from the stochastic gradient descent regression, has been chosen to better accommodate the high-dimensional problems in this study and is added to the objective function with the Huber loss function at first. The traditional linear unbiased estimation $\left(X^{T}X\right)^{-1}$ places excessive demands on computer performance due to the large amount of data processes in this study, and the properties of the Huber loss function are not as simple as the sum of squares of the residuals, thereby making search inversion matrices less intuitive. That is why this study uses the Stochastic gradient descent algorithm.

Decision Tree, a sophisticated computer data analysis technology, has derived various modified algorithms including carrying out predictions on nonlinear data sets (Panigrahi and Mantri 2015). Nair, Dharini, and Mohandas (2010) have used the decision tree model for feature selection and combined it with an adaptive network-based fuzzy inference system to predict stock trends, eventually revealing that the predictive capabilities of hybrid models are significantly higher than those of models without feature selection. More importantly, Wang et al. (2019) have improved the model's fit by boosting the cascading decision trees, showing that the new model can reduce the mean square error and retain a decent stock prediction capability.

Random forest is an integrated learning algorithm based on the bagging algorithm (Wang et al. 2016). Wang, Cao, and Chen (2016) have predicted a stock's highs and lows through it, claiming that the prediction of stock selection investments has a good return in China's stock market. Moreover,

Zhang and Wei (2018) have selected 16 technical indicators to predict the stock market using Random Forest, indicating that it could cope with a large amount of data and generate high prediction accuracy.

Gradient boosting decision tree is a robust learning method based on the boosting algorithm using gradient optimization (Friedman 2001), which can be applied to prediction scenarios and is suitable for financial scenarios with nonlinear structures. Krauss, Xuan, and Huck (2017) have used Random Forest, Gradient boosting decision trees, and deep learning to predict the S&P 500, illustrating that the stock selection model constructed by the Gradient boosting decision tree is better than the other two. Moreover, it has been applied with technical indicators to predict stock returns, verifying that its prediction capability in the CSI 300 is significantly higher than those of the Linear regression and Random forest models (Zhang and Wei 2018).

Convolutional neural networks (CNN) can autonomously learn and extract features from massive amounts of data while generalizing the results into anonymous data of the same type. It has been used to train and predict reversal points according to the stock market's k-line chart, corroborating valid prediction results (Lin 2016). Furthermore, Wang (2019) have normalized financial time series into images based on this, confirming a valid prediction model in the financial field.

3. Research Design

3.1. The Overall Design of the Predictive Model

This study utilizes the sliding window shown in Figure 1 to categorize the data set for preserving the time series characteristics of the stocks. The samples are split into three parts – January 2006 to December 2014 as the training set, January 2015 to November 2016 as the verification set, and December 2016 as the test set – to predict individual stock results during January 2017. The three parts rotate for sliding training according to the time window.

Unlike the typical training set and test set division, the sliding window method is consistent with the actual investment decision process and preserves the time series characteristics of the dataset. Additionally, exhausting the grid tuning with a large amount of data is limited by the computer hardware and can lead to excessive training time. That is why this study uses the random search for model tuning. As the time window slides, the optimal parameter for each period of each model changes as well, which is selected according to the objective function specified in the verification set.

To evaluate the model's prediction accuracy, the out-of-sample R^2 (suggested by the de-averaged denominator) replaces the traditional one, which usually leads to additional invalid information loss because the historical mean of the stock includes a lot of noise (Gu et al. 2020).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} y_i^2}$$

$\hat{y}_i$is the predictive value and y_i is the actual value.

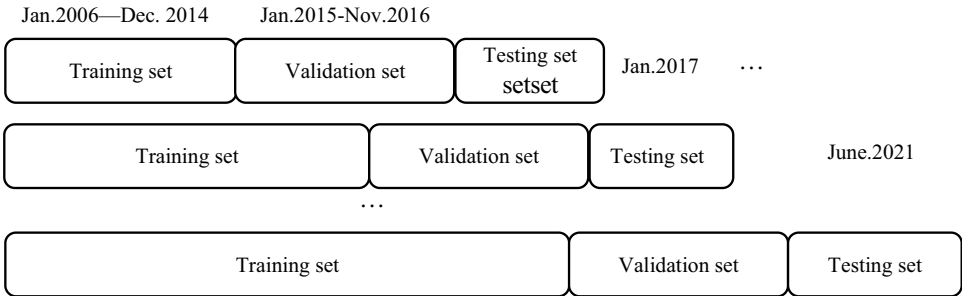

Figure 1. Sliding window.

3.2. Data Source and Sample Selection

The A-share market and financial statement data used in this study are from the Wind financial database and the JoinQuant database. They include 837 samples with 4.82 million pieces of data in total. Considering the size of the data and the computer hardware availability, this study takes the CSI 300 shares (spanning January 2006 to June 2021) as the stock pool, based on a monthly frequency. It is filled by the announcement data of listed companies, which is released quarterly.

This study selects 31 stock indexes to form three categories according to the factor attributes – fundamental, volatility (risk), and technical index factors. Firstly, due to the specific stock circumstance in China, personal investors are dominant, so we consider fundamental factors as one of the important factors. Secondly, volatility factors such as skewness and kurtosis can increase predictability performance, and Chinese stock markets share the similar feature with US stock markets (Mei et al. 2017). Moreover, technical factors, one of the most popular factors, have been proved an efficient forecasting factor in stock markets (Neely et al. 2014; Sezer, Ozbayoglu, and Dogdu 2017). Hence, we selected 14 fundamental factors, 4 volatility factors and 13 technical factors from database. Please refer to the factor construction instructions in Table 1 for the details of the construction of specific indicators.

From the descriptive statistics in the above table, the apparent differences of factors, both by orders of magnitude and distribution, will manifest in the error analysis. Specifically, a slight error is illustrated by the larger impact on the prediction results; slower predicted algorithm iterations converge, resulting from the larger order of magnitude factors, uneven distributions, and orders of magnitude. Therefore, this study standardizes the design of the above model, assuming that these samples come from a random but unified variable with a mean of 0 and a variance of 1.

4. Empirical Results and Analysis

4.1. Overall Predictability Performance

This section presents the out-of-sample prediction accuracy of different machine learning models under the out-of-sample R^2 metric, shown in Table 2. During the forecasting process, the window is moved monthly for 54 months, with the units in the table represented as a percentage. The results imply that the average value of R^2 of the out-of-sample models is positive, indicating that all five models have a prediction effect; even the linear regression model with a stochastic gradient descent performs well. Specifically, the neural networks performed best with an R^2 averaging at 0.62%, followed by the gradient boosting decision tree and Random Forest. However, the mean $R^2$of the decision tree is smaller than that of the linear regression model because the decision tree model is prone to overfitting in the more intricate circumstances, biasing the algorithms toward certain features.

4.1.1. Effective Predictability Performance of Models

Figure 2 shows the value of the monthly R^2. When R^2 is smaller than 0, the forecast months of the negative results are not statistically closer to the actual value than simply using 0 as the forecast result. According to Table 4–1, although the monthly R^2 of both the stochastic gradient descent model and the decision tree model are greater than 0, the maximum values are only 1.0335% and 0.5862%, respectively, indicating that the overall prediction performance is still far behind that of the other three machine learning methods. Additionally, the gradient boosting decision tree and random forest models have a high R^2that is greater than 0; however, the valid forecasting months only account for 53.7% and 55.56%. Moreover, the prediction performance of the neural networks among the five methods has significantly improved, shown by the high proportion of effective prediction months at 83.33%.

Table 1. Factor construction instructions.

Factor_Name	Calculation method
A.fundamental factors	
net_asset_growth_rategt	Shareholders' equity for the current quarter/Shareholders' equity prior to the three quarters) − 1
operating_revenue_growth_rate	this year's operating income (TTM)/last year's operating income (TTM)) − 1
net_operate_cashflow_growth_rate	Net cash flow from operating activities this year (TTM)/Net cash flow from operating activities (TTM) last year) −1
current_ratio	Total current assets/total current liabilities
LVGI	Asset-liability ratio for the current period (annual report)/Asset-liability ratio for the previous period (annual report)
MLEV	Total non-current liabilities/(total non-current liabilities + total market capitalization)
total_asset_turnover_rate	Net sales/Total assets
current_asset_turnover_rate	Total operating income for the past 12 months/average current assets for the past 12 months
cash_rate_of_sales	Net Cash Flow from Operating Activities (TTM)/Operating Income (TTM)
net_profit_ratio	Net Profit (TTM/Operating Income (TTM)
total_asset_growth_rate	Total Assets of the Current Period/Total Assets of the Previous Period −1
natural_log_of_market_cap	The natural logarithm of the total market capitalization of a company
net_working_capital	Current assets - current liabilities
operating_assets	Total Assets - Financial Assets
B. Volatility（risk）factor	
Kurtosis20	Kurtosis of individual stock returns
sharpe_ratio_20	（Rp-Rf）/Sigmap, Rp is the annualized rate of return for individual stocks, Rf is the risk-free rate (0.04), and Sigmap is the yield volatility (standard deviation) of individual stocks.
Skewness20	The skewness of individual stock returns
Variance20	Monthly annualized variance of returns
C. Technical index factor	
BIAS20	(Closing price - N-day average of closing price)/N-day average of closing price *100, monthly n is 20
CCI20	TYP-MA(TYP,N))/(0.015*AVEDEV(TYP,N)); N:=20
ROC20	AX = today's closing price - the closing price 20 days ago; BX = closing price 20 days ago: ROC=AX/BX*100
TVMA20	The moving average of the monthly transaction amount
TVSTD20	The standard deviation of the monthly transaction amount
VOL20	Mean monthly turnover rate (%)
VSTD20	The standard deviation of the monthly transaction volume
EMAC20	Exponential Moving Average
MAC20	Moving Average
Price1M	The closing price of the day/(the average of the closing prices of the past month −1
Rank1M	1-The ratio of yield ranking to total stocks over the past month
share_turnover_monthly	The logarithm of the sum of the stock turnover rates in the past month
money_flow_20	The average of the closing, high and low prices * the volume of the day can get the capital flow for that trading day

TTM: The rolling P/E ratio is the P/E ratio for the last 12 months.
TYP: (highest prices +lowest prices+ closing prices)/3
MA: Average prices
AVEDEV: Average absolute deviation
CCI20: Measure whether the stock price has exceeded the normal distribution range and if it belongs to the category of overbought or oversold
BIAS20: Calculation of the percentage difference between the market index or closing price and a certain moving average
The following table shows the descriptive statistic of factors.

Table 2. Descriptive statistics of R^2out-of-sample of each machine learning model.

	Mean	Min	0.25	Median	0.75	Max
SGDR	0.1283	0.0019	0.0157	0.035	0.1170	1.0335
TREE	0.0732	0.0018	0.0120	0.0222	0.0620	0.5862
FOREST	0.2480	−1.2987	−0.1951	0.0443	0.5115	3.5461
GBRT	0.4369	−4.2485	−1.0719	0.0674	1.0652	9.5384
CNN	0.6240	−1.5177	0.0259	0.1807	1.1032	3.9354

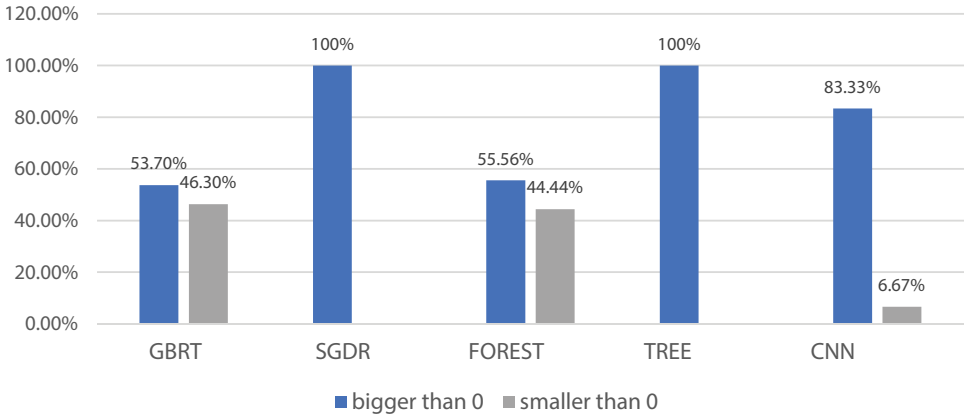

Figure 2. Monthly R^2 (the proportions of positive value and negative) of different machine learning models.

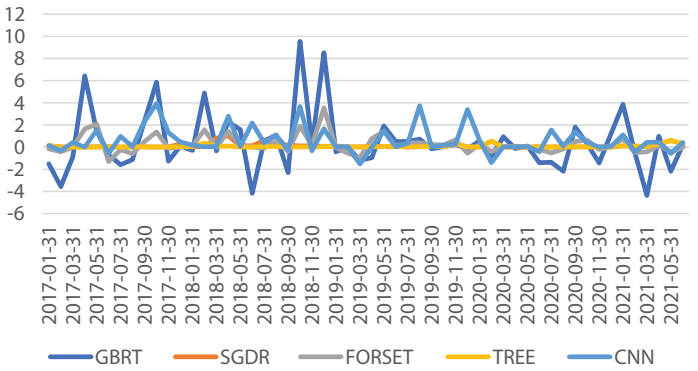

Figure 3. Monthly R^2 trends for different machine learning models.

4.1.2. Predictability Fluctuations of Models

Figure 3 shows the monthlyR^2trend for each model. Notably, the average $R^2$of the gradient boosting decision tree and the decision tree is higher, but the monthly R^2fluctuations of the two are larger as their stability is lower. Moreover, the $R^2$of the stochastic gradient descent linear regression model and the decision tree model is not statistically significant, indicating that the predictability results are less volatile. Surprisingly, the peaks and troughs of the neural networks, gradient boosting decision trees, and random forest models during the fluctuation process are consistent with the time of occurrence, showing that the prediction performances of the models are basically in the same direction.

On the whole, compared to the linear regression with stochastic gradient descent model, the machine learning algorithms can improve the model's out-of-sample prediction performance with better prediction results. These results corroborate the superiority of machine learning methods in capturing the complex interactions among predictors, thereby verifying their effectiveness in China's stock market.

4.2. Analysis of Feature Importance

This study utilizes the Feature Importance from the scikit-learn library, the computational feature of Python, to explain the reasons behind the models' predictions of stock returns for the decision tree, random forest, and gradient boosting decision tree models. Thus, it reveals the essential characteristics

that drive the model to make predictions for individual and overall samples and explain how these characteristics affect the prediction results.

4.2.1. Feature Importance Ranking

The results in the heat map (Figure 4) show the importance of the factors at the stock level, suggesting the overall feature importance of each model based on the complete sample collection. The feature importance ranking reflects the overall contribution of the features to all models, with each column corresponding to one model, where the color gradient represents the specific importance of the model, from highest to lowest importance (i.e., from darkest to brightest).

This graph distinctly implies that the importance of the factor features varies across machine learning models. Specifically, the three-factor characteristics – including standard deviation of the monthly transaction amount, deviation rate, and the homeopathic index – significantly influence the three models. Subsequently, the fundamental factors, such as the growth rate of net assets and operating income growth rate, have a limited impact on the model's predictions, suggesting scarcely any impact on the fit of the monthly rolling model in this study; it is caused by the quarterly data publishing of the listed companies.

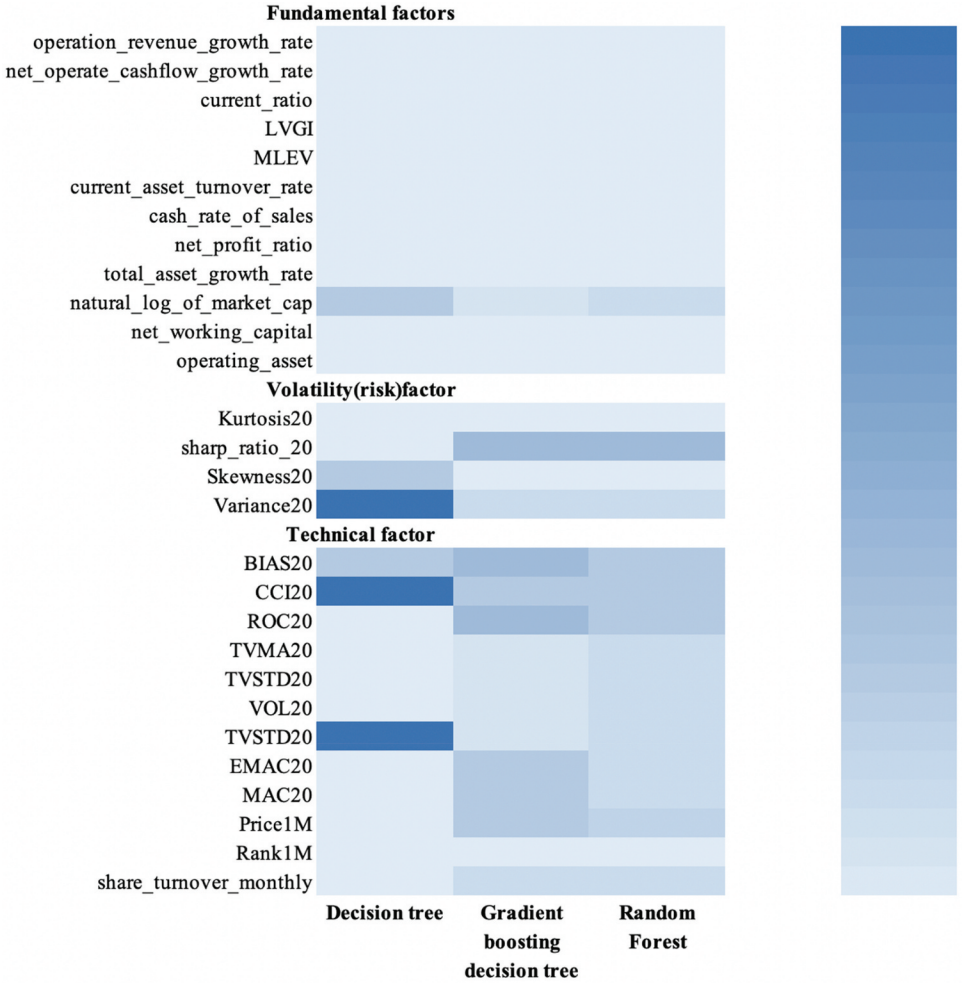

Figure 4. The factor features importance of Decision tree, Random forest and Gradient boosting decision tree.

4.2.2. Factor Importance Comparison

Figure 5 summarizes the variable importance of the fundamental factors in each model. The fundamental factors play a tremendous role in the decision tree model, but only the logarithmic market value matters, which is the most influential fundamental factor when predicting monthly returns in the Chinese stock market. Therefore, the influence of each factor regarding the fundamental factor class creates a great discrepancy in the random forest and gradient boosting decision tree models.

Figure 6 presents the variable importance of volatility (risk) factors in the decision tree, random forest, and gradient boosting decision tree models. The volatility (risk) factor has a more significant impact on the decision tree model when predicting the monthly return of the Chinese stock market, while it has almost the same impact on the random forest and gradient boosting decision tree models. Moreover, the volatility (risk) factors have a significantly greater influence on the decision tree, rather than the random forest as well as the gradient boosting decision tree models.

Figure 7 states the variable importance of technical factors in each model, showing that the decision tree, random forest, and gradient boosting decision tree models have roughly similar influences among these factors. On the contrary, the difference in the influence of the various technical factors in the random forest model is markedly smaller than that of the decision tree and gradient boosting decision tree models, sharing similar features with Figure 6.

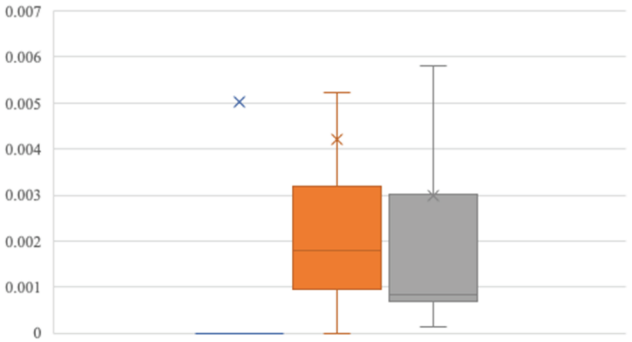

Figure 5. The importance of the three models' fundamental factor features comparison (Decision tree, random forest, and gradient boosting decision tree).

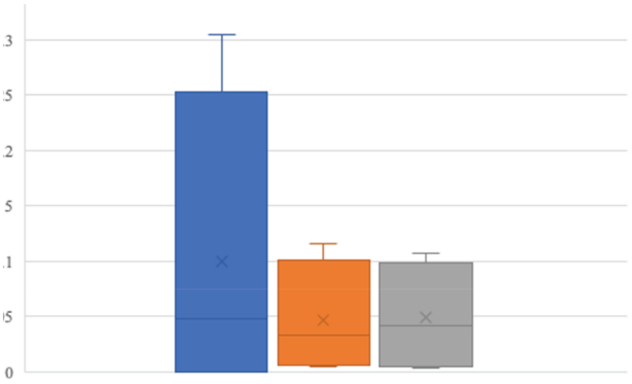

Figure 6. The importance of the three models' volatility(risk) factors features comparison (Decision tree, random forest, and gradient boosting decision tree).

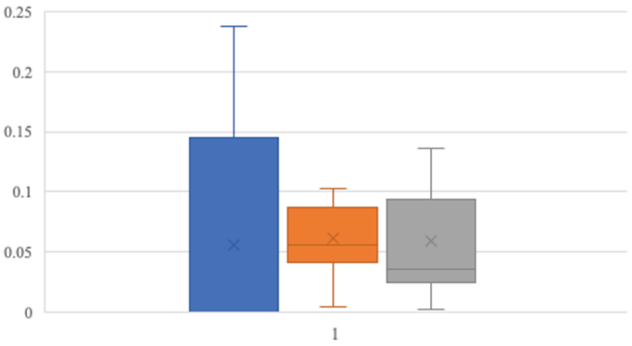

Figure 7. The importance of the three models' technical index factors features comparison (Decision tree, random forest, and gradient boosting decision tree).

4.2.3. Factor Importance Analysis

This study identifies the most influential technical factors and least influential fundamental factors from the perspective of behavioral finance, such as investor bias, investor psychology, and information dissemination. First, the BSV model suggests that finitely rational investors have behavioral biases that lead to underreacting and overreacting to information (Barberis et al. 1998). What's more, Barberis et al. (2001) claim that stock prices continue to move forward with the insufficient response of investors, eventually shown by the momentum effects. More precisely, the HS theory (Hong and Stein 1999) holds that as information in the market fluctuates, stock prices spread by degrees among investors, and the order of obtaining this information leads to underreacting and overreacting phenomena.

China's A-share market has two characteristics. First, individual investors in China's A-share market account for 33.27%, far greater than the total shareholding market value of domestic professional and institutional investors (at 16.59%), according to the calculation of current market caliber. On the one hand, institutional investors obtain company information more efficiently in the A-share market than individual investors, who prefer to "pick by names" in the stock markets. On the other hand, a lag in the response of individual investors to information leads to a herd effect, which is the psychological follow-up and the herd mentality of economic individuals. The second, characteristic of the A-share market – price limit – is an interfering system that hinders the transmission of information and affects the liquidity and effectiveness of the market. More importantly, price limits make investors move toward profits solely by overly paying attention to the short-term impacts on the stock price. This leads to the unique limit-up and limit-down phenomenon of China's stock market, which then superimposes the herd effect described above, eventually resulting in the emergence of the momentum phenomenon. The above explanation verifies that the deviation rate and trend indicators generated by the technical indicators of momentum in this study can indeed have a considerable impact on the three models.

4.3. Stock Picking Strategy Performance of Machine Learning

This study constructs a simple stock selection investment strategy to structure a stock portfolio of each month, sorting the stock yield forecasts of the monthly rolling models from high to low. Moreover, each model builds the portfolio based on the weights of the top 20 stocks in terms of returns, according to the monthly rolling forecasts, to simulate the investment requirements for risk diversification in real markets. The out-of-sample test period is from 2017 to June 2021.

To simulate the actual market trading situation, this study sets 3/10,000 as the buying commission during the stock trading process, 3/10,000 plus 1/1,000 stamp duty when selling, and the minimum

commission for each transaction is 5 yuan. However, during the actual process, the accurate transaction price usually has a specific deviation from the expected price when placing an order. That is why this study adds slippage to simulate the performance of the actual market, setting the slippage default at 0.00246.

4.3.1. Cumulative Returns of Models

Figure 8 indicates the monthly and cumulative yields of a weighted portfolio of assets constructed by the prediction results of each model. The benchmark yield curve, yield of the Shanghai Securities Composite Index during the out-of-sample test period, is used for comparative studies. In short, all models share a near consistent trend with the A-share market, shown by the synergy of the portfolio building strategies. Notably, when only considering the returns of the out-of-sample portfolio, the portfolio strategy of the neural network has the most significant cumulative returns followed by the random forest and gradient boosting decision tree models. However, the cumulative returns of the investment strategies based on the gradient boosting decision tree are always higher than that of the random forest model. The yields of the investment strategies of the random forest model have risen rapidly since June 2021. Hence, its sustainability needs to be further observed.

4.3.2. Out-of-Sample Portfolio Predictability Performance

This study utilizes some evaluation indicators of the trading strategies to better measure risk and further compare the predictivity of each model. These include the strategy cumulative return, strategy annualized return, excess return, Sharpe ratio, win rate, maximum drawdown, volatility, and profit-loss ratio. Table 4 and 3 comprises the various analytical indicators of the portfolio strategy based on the monthly rolling forecasts built from each model and illustrates that the portfolio strategies built on the neural network models have the highest Sharpe ratios, followed by the gradient boosting decision tree and random forest models. Additionally, portfolio strategies built on neural networks have a significantly higher win rate than the other four models. Although the portfolio strategy based on the random forest model echoes with the stochastic gradient descent regression and decision tree models, the P/L ratio of the random forest model is the highest among the three, resulting in a significantly higher yield than the other two models. From the perspective of maximum back testing and volatility, the portfolio strategy based on the neural network model performs the best, which is illustrated by the small

Figure 8. Cumulative returns from portfolios out-of-samples of different machine learning models.

Table 3. Descriptive statistic of factors (from January 2006 to June 2021).

	mean	std	min	0.25	median	0.75	max
net_asset_growth_rate	0.41	89.05	−1.00E+04	0.02	0.08	0.19	1.47E+04
operating_revenue_growth_rate	6342.98	8.89E+07	−8434.55	−0.03	0.11	0.28	1.25E+08
net_operate_cashflow_growth_rate	−1.34	147.38	−2.06E+04	−0.93	−0.13	0.51	9523.06
current_ratio	2.17	10.71	−60.96	0.92	1.33	2.02	1883.35
LVGI	1.06	0.75	0.05	0.94	1.01	1.09	51.76
MLEV	0.13	0.17	−0.14	0.01	0.05	0.19	0.93
total_asset_turnover_rate	0.64	0.57	−0.90	0.27	0.52	0.82	30.19
current_asset_turnover_rate	1.58	1.54	−1.83	0.74	1.24	2.04	102.29
cash_rate_of_sales	1.20E+04	3.01E+07	−111.79	0.02	0.09	0.21	7.54E+09
net_profit_ratio	−8338.26	1.93E+06	−4.83E+08	0.02	0.08	0.17	2111.59
total_asset_growth_rate	0.83	40.24	−1.00	0.02	0.11	0.23	5796.56
natural_log_of_market_cap	23.42	1.27	18.61	22.59	23.36	24.14	29.38
net_working_capital	2.50E+09	2.04E+10	−2.78E+11	−2.30E+08	8.03E+08	2.92E+07	4.35E+15
operating_assets	1.49E+11	1.13E+12	0.00	3.31E+09	8.59E+09	2.63E+10	3.07E+13
money_flow_20	5.46 E+09	1.14E+10	0.00	9.08E+08	2.28E+09	5.51E+09	5.25E+11
TVMA20	2.73E+08	5.76E+08	0.00	4.44E+07	1.13E+08	2.74E+08	2.64E+10
TVSTD20	1.31E+08	2.75E+08	0.00	2.21E+07	5.52E+07	1.34E+08	1.49E+10
VOL20	1.87	2.11	0.00	0.58	1.19	2.41	50.87
VSTD20	3.97E+06	1.62E+07	0.00	3.78E+05	9.51E+05	2.56E+06	1.66E+09
BIAS20	0.12	8.41	−69.11	−4.09	−0.44	3.65	467.35
CCI20	−5.93	113.91	−666.67	−92.95	−14.49	82.38	666.67
ROC20	1.80	18.96	−87.25	−6.62	0.00	7.99	2288.89
Kurtosis20	0.94	2.18	−8.00	−0.33	0.33	1.51	20.00
sharpe_ratio_20	1.17E+10	4.13E+12	−109.75	−1.58	−0.15	4.07	1.46E+15
Skewness20	0.07	0.84	−4.47	−0.40	0.02	0.51	4.47
Variance20	0.30	18.29	0.00	0.07	0.14	0.27	5336.73
EMAC20	1.00	0.07	0.25	0.97	1.00	1.04	3.09
MAC20	1.01	0.08	0.18	0.96	1.00	1.04	3.24
Price1M	0.00	0.08	−0.71	−0.04	0.00	0.04	6.02
Rank1M	0.51	0.27	0.01	0.28	0.51	0.75	1.00
share_turnover_monthly	−1.39	1.03	−11.42	−2.04	−1.35	−0.66	2.01
yield_rate	0.01	0.15	−0.83	−0.06	0.00	0.08	5.12

drawdown and high return among the five models; conversely, the portfolio based on the decision tree model has large volatility but low yields.

As shown in Table 4, the cumulative return of the portfolio strategy based on the random forest model is significantly higher than that of the gradient boosting decision tree model; however, its win rate and P/L ratio are relatively smaller. As the final yield is mainly from June 2021, the effectiveness of its portfolio strategy needs to be confirmed by further observation. Further, the portfolio strategy based on machine learning is associated with the impact of the A-share market, suggesting that the maximum drawdown of each model is synchronized with the A-share market, which fell for 11 consecutive months, reflecting in the large drawdown of each portfolio strategy during 2018.

In summary, the neural network model has a distinctive advantage over the other machine learning algorithms discussed in this study when it comes to predicting stock returns in China's A-share market. Theoretical analysis shows that the decision tree model is relatively simple, is liable to be biased toward some characteristics, and generates overfitting problems resulting in a weak capability of generalization. Thus, the deviation in the out-of-sample yield predictions of this study is relatively large, and the strategic return and retained risk are not ideal. According to this, the random forest and gradient boosting decision tree models use an integrated algorithm based on the decision tree, so that the proposed method has a better prediction performance, reducing the overfitting problem, and improving its ability for generalization. Thus, they improve the yield prediction results and the accuracy of risk indicators compared to the decision tree model.

Table 4. Details of different machine learning out-of-sample portfolio strategies.

Indicators	Stochastic gradient descent	Decision tree	Random forest	Gradient boosting decision tree	Neural networks
Strategic cumulative return	22.16%	4.20%	41.19%	31.8%	46.14%
Strategic annualized return	4.69%	0.95%	8.22%	6.52%	9.07%
Excess return	5.58%	−9.95%	22.02%	13.9%	26.3%
Sharpe ratio	0.0038	−0.158	0.218	0.148	0.32
Win-ratio	0.471	0.475	0.476	0.507	0.526
Maximum drawdown	31.64%	34.14%	32.19%	34.7%	29.2%
Volatility ratio	0.183	0.194	0.172	0.171	0.159
Profit-loss ratio	1.208	1.102	1.250	1.362	1.523

5. Conclusions and Implications

5.1. Conclusions

This study analyzes the performance of each model in predicting the A-share yield, demonstrates the influence level of different factors on the predictive behavior of different models. It compares the empirical performance of portfolio strategies based on the machine learning models in China's A-share market as well as the investment return and risk levels of different models, eventually drawing the following conclusions: 1. Machine learning is effective in predicting the yield of China's A-share market, with the out-of-sample test of the neural network model performing best within the database of this study, followed by the gradient boosting decision tree and random forest models; 2. By sorting the factors according to the importance and analyzing the influence of each type of factor in the decision tree, random forest, and gradient boosting decision tree models, this study reveals that the technical index factor is the most valuable, followed by the volatility (risk) factor, and the fundamental factor. Additionally, the technical factors of biased momentum, such as the deviation rate and homeopathic indicators, have a greater impact on the forecasting performance from the perspective of behavioral finance, considering the structure of China's investment markets and investors' characteristics. 3. By simulating the yield results obtained from simulated trading using a portfolio strategy with high stock prediction yields, it is found that the neural network model still performed best, followed by the random forests and gradient boosting decision tree models. Therefore, machine learning models can bring huge economic benefits in practice when predicting the yield of A-share markets.

5.2. Implications

5.2.1. Theoretical Significance

These results have rich implications for the applied research of finance. By systematically comparing the empirical performance of five machine learning algorithms, this study provides a typical example of introducing sophisticated machine learning models in predicting the yield of China's A-share market in financial research. Specifically, combining theoretical research with innovative technologies allow a better exploration of the effective stock characteristics to predict China's A-share market during the analysis of finance-related forecasts. Regarding this, stockholders can have a more rounded insight into the operating characteristics of the stock market by understanding the predictive information of different trading anomalies. Besides, the empirical performance of this study demonstrates that the machine learning model is applicable to the Chinese market. However, further research is needed based on local circumstances, to be innovative and eventually develop a set of research data specific to the market.

5.2.2. Practical Significance

Applying machine learning algorithms can bring considerable benefits for securities investments. For the securities investment industry, machine learning algorithms can improve the efficiency and effectiveness of securities asset management by providing new types of asset analysis tools. For asset management companies, the research process can provide insights to further explore asset management practices; for example, companies can consider combining computer technologies with others to improve the effectiveness of asset management. For individual investors, technical factors are valid in predicting the behavior of China's A-share market; more specifically, technical analysis can help investors achieve profitability in the A-share market.

The findings of this study also have rich implications for regulators. Recently, investment circumstances have changed under the slightly overweight real estate regulations, along with the deepening of reforms and opening of the capital market as a result of the increase in residents' wealth. On the one hand, artificial intelligence echoes machine learning, as a more mature computer technology that can be effectively applied to the securities industry and all aspects of finance. However, according to the findings in this study, imperfect models still cause adverse effects. When applying new technologies in the financial field, regulators can actively guide financial experts, computer talents, and other multidisciplinary specialists to build more comprehensive models. These models are conducive to preventing the impact of problems that arise from artificial intelligence models in China's financial market, consequently controlling the potential risks and improving the stability of the financial system. On the other hand, quantitative investment, similar to machine learning algorithms, can help the China's capital market increase significantly. Additionally, China's regulatory authorities can ensure an orderly liberalization of quantitative investment policies, gradually optimizing hedging mechanisms, and effectively improving the efficiency of resources in China's financial market to ensure the sustained and healthy operation of China's economy.

5.3. Limitations

This study has some limitations, which further research can help address First, due to the limitations of the scikit-learn database, this study only analyzes the feature importance among the three machine learning models – decision tree, random forest, and gradient boosting decision tree. Second, in addition to the five machine learning algorithms, future studies should consider other machine learning algorithms to form more comprehensive insights. Therefore, it is necessary to continue further studies that can fine-tune the results by incorporating more machine learning algorithms and investing factor features.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Akita, R., A. Yoshihara, T. Matsubara, and K. Uehar. 2016. Deep learning for stock prediction using numerical and textual information. *Computer Science* 15:1–6.

Awartani, B., and V. Corradi. 2005. Predicting the Volatility of S&P-500 Stock Index via GARCH Models: The Role of Asymmetries. *International Journal of Forecasting* 21:167–83. doi:10.1016/j.ijforecast.2004.08.003.

Barberis, N., and M. Huang. 2001. Mental Accounting, Loss Aversion, and Individual Stock Returns. *The Journal of Finance* 56 (4):1247–92. doi:10.1111/0022-1082.00367.

Barberis, N., A. Shleifer, and R. Vishny. 1998. A Model of Investor Sentiment. *Journal of Financial Economics* 49:307–43. doi:10.1016/S0304-405X(98)00027-0.

Basak, S., S. Kar, S. Saha, L. Khaidem, and S. R. Dey. 2019. Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance* 47:552–67. doi:10.1016/j.najef.2018.06.013.

Campbell, J. Y. 2000. Asset pricing at the millennium. *The Journal of Finance* 55 (4):1515–67. doi:10.1111/0022-1082.00260.
Chen, J., F. Jiang, and G. Tong. 2017. Economic policy uncertainty in china and stock market expected returns. *SSRN Electronic Journal*. doi:10.2139/ssrn.2808862.
Devpura, N., P. K. Narayan, and S. S. Sharma. 2017. Is stock return predictability time-varying? *Journal of International Financial Markets, Institutions&money* 52:152–72. doi:10.1016/j.intfin.2017.06.001.
Fama, E. F. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25 (2):383–417. doi:10.2307/2325486.
Fama, E. F., and K. R. French. 1993. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33:3–56. doi:10.1016/0304-405X(93)90023-5.
Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29 (5):1189–232. doi:10.1214/aos/1013203451.
Gu, S. H., B. Kelly, D. C. Xu, and A. Karolyi. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies* 33 (5):2223–73. doi:10.1093/rfs/hhaa009.
Han, Y. F., G. F. Zhou, and Y. Z. Zhu. 2016. A trend factor: Any economic gains from using information over investment horizons?. *Journal of Financial Economics* 122 (2):352–75. doi:10.1016/j.jfineco.2016.01.029.
Hong, H., and J. C. Stein. 1999. A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *The Journal of Finance* 54 (6):2143–84. doi:10.1111/0022-1082.00184.
Jegadeesh, N., and S. Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance* 48 (1):65–91. doi:10.1111/j.1540-6261.1993.tb04702.x.
Jiang, F. W., X. L. Qi, and G. H. Tang. 2018. Q-theory, mispricing, and profitability premium: Evidence from China. *Journal of Banking and Finance* 87:135–49. doi:10.1016/j.jbankfin.2017.10.001.
Krauss, C., A. D. Xuan, and N. Huck. 2017. Deep neural networks, gradient-boosted trees, random forests. Statistical arbitrage on the S&P 500. *European Journal of Operational Research* 259 (2):689–702. doi:10.1016/j.ejor.2016.10.031.
Lin, X. 2016. Convolutional Neural Networks based detection of outlier and turning points in stock trading. Master diss., Huazhong University of Science and Technology
Li, B., X. Y. Shao, and Y. Y. Li. 2019. Research on machine learning driven quantamental investing. *China Industrial Economics* 8:61–79.
Markowitz, H. M. 1952. Portfolio Selection. *The Journal of Finance* 7:77–91. doi:10.1111/j.1540-6261.1952.tb01525.x.
Markus, L., Q. Wang, and W. Y. Zhou. 2021. Machine learning in the Chinese stock market. *Journal of Financial Economics* 145 (2):64–82. doi:10.1016/j.jfineco.2021.08.017.
Mei, D. X., J. Liu, F. Ma, and W. Chen. 2017. Forecasting stock market volatility: Do realized skewness and kurtosis help? *Physica A: Statistical Mechanics and Its Applications* 481:153–59. doi:10.1016/j.physa.2017.04.020.
Nair, B., N. M. Dharini, and V. P. Mohandas. 2010. A stock market trend prediction system using a hybrid decision Tree-Neuro-Fuzzy System. Advances in Recent Technologies in Communication and Computing (ARTCom) International Conference, 381–85. doi:10.1109/ARTCom.2010.75.
Narayan, P. K., R. P. Liu, and J. Westerlund. 2016. A GARCH model for testing market efficiency. *Journal of International Financial Markets, Institutions and Money* 41:121–38. doi:10.1016/j.intfin.2015.12.008.
Neely, C. J., D. E. Rapach, J. Tu, and G. F. Zhou. 2014. Forecasting the equity risk premium: The role of technical indicators. *Management science* 60 (7):1772–91. doi:10.1287/mnsc.2013.1838.
Panigrahi, S. S., and J. K. Mantri. 2015. Epsilon-SVR and decision tree for stock market forecasting. International Conference on Green Computing & Internet of Things, Greater Noida, Delhi, 761–66, October 8-10. 10.1109/ICGCIoT.2015.7380565
Phan, D. H. B., S. S. Sharma, and P. K. Narayan. 2015. Stock return forecasting: Some new evidence. *International Review of Financial Analysis* 40:38–51. doi:10.1016/j.irfa.2015.05.002.
Rapach, D., and G. F. Zhou. 2013. Forecasting stock returns - ScienceDirect. *Handbook of Economic Forecasting* 2 (A):328–83.
Sezer, O. B., M. Ozbayoglu, and E. Dogdu. 2017. An artificial Neural network-based stock trading system using technical analysis and big data framework. ACM Southeast conference ACMSE 2017, GA, U.S.A., Kennesaw State University, April 13-15.
Wang, Y. X. 2019. Stock prediction based on Convolutional Neural Networks. Master diss., Tiangong University.
Wang, S. Y., Z. F. Cao, and M. Z. Chen. 2016. Research on application of random forest in the quantitative stock selection model. *Operations Research and Management Science* 25 (3):163–68.
Wang, Y., D. Y. Chen, and Y. X. Tang. 2019. A Stock Prediction Model Based on Cart and Boosting Algorithm. *Journal of Harbin University of Science and Technology* 24 (6):99–103.
Zhang, X., and Z. X. Wei. 2018. Predicting the direction of stock market prices using random forest. *China Management Informationization* 3:120–23.