Research on Quantitative Investment Strategies Based on Artificial Intelligence

Wenjie Sun*

Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: 2500061525@qq.com

Abstract. Artificial Intelligence (AI) has been developing rapidly in recent years, and the application of AI in various fields has gradually emerged. This paper focuses on exploring the cross-fertilization of AI with the Chinese stock market, and the study adopts 7 types of factors, totaling 29 factor indicators, covering multiple types of value, valuation, leverage, financial quality, growth, technology, and risk factors, etc. Through data preprocessing, feature engineering, and other steps on the factors, and then combined with 8 machine learning algorithms to construct corresponding quantitative trading strategies on CSI 300 constituent stocks. By comparing the results of various models, it is found that the trading strategies constructed by machine learning algorithms can obtain significant excess returns in the Chinese market. Except for Gaussian Park Bayes, the other seven model strategy returns beat the benchmark returns, among which XGBoost performs the best, achieving a 20.10% return and the smallest retraction. This paper has a positive impact on advancing the research on the intersection of machine learning and quantitative investment.

Keywords: artificial intelligence, machine learning, stock returns, quantitative investing, investment strategies.

1. Introduction

Artificial intelligence (AI) has become a strategic emerging technology field emphasized by all countries, and is regarded as a key force that can profoundly reshape national competitiveness. In order to seize the commanding heights in the development of AI, major economies have elevated it to a national key development strategy, formulated top-level plans and vigorously invested in human and material resources. The Chinese government has also made efforts to strengthen core technology research, talent training, industrial ecological construction and the construction of standardization system through policy support, aiming to promote the innovative development of AI industry and occupy a dominant position in the new round of international science and technology competition. Artificial intelligence as a disruptive technology, in all walks of life has caused a greater heat of discussion. At the same time, more and more scholars are introducing machine learning and artificial intelligence technologies into the research of their disciplines, and these innovative algorithms are driving the development process of various disciplines.

The stock market, as a barometer of the national economy, has always attracted a great deal of attention from the society, as it not only reflects the economic operation of a country, but is also closely related to the immediate interests of the majority of investors. As early as the 1990s, some quantitative investment funds began to try to combine statistical models with historical data, trying to capture the mathematical and physical patterns in stock prices and seek excess returns, which can be regarded as one of the earliest attempts to combine artificial intelligence with financial investment. With the rapid development of artificial intelligence algorithms, a large number of quantitative investment funds have begun to emerge on Wall Street, which are able to explore more investment opportunities with the help of computer science, artificial intelligence and other technologies.

In academia, scholars have long worked on predicting stock prices, trends, and returns using traditional econometric models. Some econometricians believe that the future movement of stock prices can be judged with high probability through historical information, and they believe that history repeats itself. However, others who study technical trading have found that traditional technical and fundamental analysis has significant shortcomings in explaining and predicting stock price fluctuations, making it difficult to fully grasp the complexity of the multiple factors that influence stock prices. In contrast, machine learning algorithms are able to automatically learn from massive amounts of historical data and capture the non-linear relationship between stock prices and numerous characteristic variables, thereby constructing more accurate forecasting models. Therefore, machine learning algorithms can be an important tool for stock market predictive analysis.

As the second largest economy in the world, China's huge stock market has received extensive attention from scholars and investors both at home and abroad. Compared with mature markets, China's stock market trading presents distinctive features, such as being dominated by retail individual investors, highly speculative, and significantly influenced by emotional factors. This makes the application of machine learning in Chinese stock market of great theoretical significance and practical value. On the basis of previous research, this paper adopts multiple factor indicators and combines multiple machine learning algorithms to construct a quantitative investment strategy for CSI 300 constituent stocks, to explore the prospects of machine learning algorithms in the Chinese market, and the successful application in the Chinese market will also greatly enhance the interpretability and universality of machine learning models.

2. Literature review

With the development of quantitative investment theory, more and more mature methods in machine learning that have been used in other fields have been applied to the study of the stock market. Among them, SVM model, random forest model, Adaboost model and Xgboost model are the more mature methods.

Support Vector Machines (SVM) were introduced by Corinna Cortes and Vapnik (1995). They developed this groundbreaking machine learning technique based on neural network models.SVM classifies data by finding the optimal separating hyperplane in the high-dimensional space behind the data, and it is able to handle high-dimensional data efficiently with excellent performance.Kim (2003), on the other hand, trained the SVM model by using the daily data of the Korean stock market, and found that the hit rate of the SVM model in stock prediction was about of 56%. In addition, Huang et al. (2005) used SVM to predict the weekly data of Nikkei index and showed that the hit rate of single SVM model and combined SVM model reached 73% and 75%, respectively.

Random Forest is an Ensemble learning method proposed by Breiman (2001).Breiman constructed an Ensemble model containing multiple decision trees, i.e., Random Forest, by introducing Bagging technique and the idea of random sampling of features. The algorithm has gained wide recognition due to the advantages of being able to process high-dimensional data quickly and having strong model generalization ability. So far, many scholars have applied Random Forest to stock research in the financial field.Kumar.M (2006) earlier applied the method to the stock market, through the classification and analysis of the Standard & Poor's and emerging market stock data, and achieved good results in predicting the changes of stock indexes. Zhengfeng Cao, Hong Ji and Bangchang Xie (2014), on the other hand, combined Random Forest with the value investment strategy to construct stock portfolios and gained considerable investment returns. It is worth mentioning that random forests not only perform well in predicting stock index fluctuations, but also show excellent application prospects in actual investment decisions such as stock selection and asset allocation.

In recent years, emerging machine learning algorithms such as Adaboost and XGBoost have been widely used in stock market research.Ballings et al. (2015) predicted 5767 European listed company stocks using seven machine learning models with AUC as the evaluation index, and the results showed that Random Forest had the best prediction effect. Ye Meng et al. (2018) used KDJ as the basis for stock selection and timing, and integrated the nearest neighbor propagation clustering algorithm with Adaboost to form a classifier for fundamental stock selection. The strategy outperformed the broader market level on an annualized basis over the three-year backtesting period from 2015 to 2017.Chen and Guestrin (2017), on the other hand, used the XGBoost algorithm to build a multifactor stock picking model, and the model demonstrated better investment returns. Xiang Li

(2017) compared the XGBoost algorithm with random forest and SVM models by combining 307 metrics with the XGBoost algorithm, and the results confirmed that XGBoost performed best in terms of prediction accuracy and stability.Basak et al. (2019) used two algorithms, namely, random forest and XGBoost, to predict the estimated stock price movements, and the results found that based on the XGBoost algorithm has higher accuracy in prediction.

Although traditional machine learning algorithms such as logistic regression, KNN and decision trees have been used relatively little in stock market research, their effectiveness has been confirmed by some studies. The results of Shinong Wu and Xianyi Lu (2001) showed that the logistic method is superior in predictive effectiveness compared to multiple linear regression and Fisher model. Huaqin Wang (2018) comprehensively summarized five machine learning methods, including Support Vector Machine, KNN, Logistic Regression, Random Forest and XGBoost, and conducted a comparative analysis of stock up and down prediction and backtesting based on these algorithms, and the results showed that the prediction accuracy and return on investment of XGBoost and Random Forest were the best, followed by SVM and KNN, while Logistic Regression had the most inferior performance. Huang et al. (2008) combined the feature engineering approach with major machine models in the Korean and Taiwanese stock markets, respectively, and the results showed that the Wrapper approach was the most effective.

3. Empirical analysis

Although predicting specific stock price movements based on fundamental or technical data is a challenge, some studies have shown that accurately determining the direction of a stock's rise or fall can be profitable.Kumar and Thenmozhi (2006) conducted an empirical study on Indian stock market index data and found that the support vector machine algorithm has the highest accuracy among several machine learning models. In this paper, eight algorithms such as KNN, Decision Tree, Support Vector Machine, Logistic Regression, K Nearest Neighbors, Random Forest, XGBoost, and Adaboost are applied to construct a quantitative investment strategy by using seven types of factor indicators such as valuation, growth, and technology as input features. All data are obtained through the JionQuant platform, and the empirical process is also completed in the research environment of the platform, including machine learning model training and strategy backtesting. The study aims to explore the application prospect of machine learning in the field of stock investment and provide investors with effective quantitative investment decision support.

3.1. Data acquisition

(a) Empirical time intervals and divisions

A total of 7 years of data from January 1, 2014 to December 31, 2023 are selected for empirical study, where January 1, 2014 to December 31, 2020 is the training interval of 7 years. Historical data during this interval is used to train various machine learning models and evaluate their predictive performance during this period.The period from January 1, 2021 to December 31, 2023 is a testing interval of 3 years. Data from this period will be used to test the generalization ability of the trained models on new data.

(b) Selection of the subject matter of the transaction

The empirical stock samples are all the constituent stocks of CSI 300 index after screening, and the screening rules are as follows:

(1) Filter out ST/PT stocks;

(2) Excluding stocks that cannot be traded due to suspension of trading, etc;

(3) A sample of stocks that have been on the market for less than 180 days is excluded.

(c) Factor selection

There are seven categories of indicators selected for this paper: value factors, valuation factors, leverage factors, financial quality factors, growth factors, technical indicators, and risk factors. The specific factors are shown in Table 1 below.

Table 1. Description of Selected Factors

major category	Indicator name	Calculation or key parameters
value factor	net assets	Total assets - total liabilities
valuation factor	EP	Net profit/total market capitalization
	BP	Total assets/total market capitalization
	SP	Operating income/total market capitalization
	CFP	Operating cash flow and total market capitalization
leverage factor	financial leverage	Total assets/net assets
	Debt-to-equity ratio	Non-current liabilities/net assets
	cash ratio	(Money funds + marketable securities) ÷ Current liabilities
	current ratio	Current assets/current liabilities*100%
Financial quality factor	net profit	Total income - total expenditure
	gross margin	(Sales revenue - cost of sales)/sales revenue × 100%
	return on net assets	Net profit/average net assets × 100%
	return on total assets	Net profit ÷ average total assets × 100%
	Asset turnover ratio	Sales revenue divided by average total assets
	Net operating cash flow	Sales revenue - cash costs - income tax
growth factor	Revenue growth rate	(Increase in operating income / total operating income of the previous year) × 100 percent
	profit growth rate	(Increase in profit / Total profit of the previous year) × 100%
Technical indicators	RSI	timeperiod=14
	BIAS	timeperiod=14
	DIF	timeperiod=12
	DEA	timeperiod=26
	MACD	fastperiod=12,slowperiod=26,signalperiod=20
	AR	timeperiod=6
	ARBR	timeperiod=6
	ATR14	timeperiod=14
risk factor	VOL5	timeperiod=5
	VOL60	timeperiod=60
	Skewness20	timeperiod=20
	Skewness60	timeperiod=60

3.2. Data pre-processing

After labeling and missing value processing, the dataset contains a total of 9608053 samples. After this, the data is preprocessed and the data preprocessing process in this study includes steps such as median depolarization, missing value processing, industry market value neutralization, and normalization.

Data preprocessing is carried out according to the following specific steps: first, samples with missing values in the dataset (mainly due to trading suspensions) are removed. Then, the samples are labeled according to the acquired stock returns, and the samples with the top 30% of returns are marked as 1 and the samples with the bottom 30% are marked as 0. Next, the median depolarization is performed using the MAD method, and the platform's self-contained functions are invoked to neutralize the data in terms of industry and market capitalization. Finally, the data are normalized. Table 2 demonstrates the descriptive statistics of some features of the data after preprocessing.

Table 2. Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
EP	35126	0.05	0.06	-1.32	0.02	0.04	0.08	0.72
BP	35129	0.53	0.42	0.01	0.22	0.41	0.72	3.86
SP	35126	0.82	1.30	-0.20	0.16	0.41	0.89	18.38
CFP	35126	0.02	0.16	-2.14	-0.01	0.01	0.05	2.64
FLA	35127	3.47	3.34	0.99	1.61	2.31	3.64	33.68
DQR	29238	0.42	0.57	-0.17	0.06	0.22	0.55	9.64
CHR	29118	0.59	1.53	-5.16	0.11	0.27	0.60	74.55
CTR	29347	2.31	17.03	-60.96	1.06	1.42	2.15	1883.35
NI	34828	0.17	0.28	-1.92	0.05	0.13	0.26	25.35
GPM	30268	0.32	0.21	-2.75	0.16	0.29	0.43	1.13
ROE	35111	0.12	0.21	-32.87	0.07	0.11	0.17	1.52
ROA	35107	0.06	0.06	-0.52	0.02	0.04	0.09	0.66
AT	34819	0.54	0.46	-0.66	0.21	0.47	0.75	5.70
NOCF	35040	3.34	147.45	-1647	-0.72	0.95	2.65	11340.61
SG	33793	0.51	21.72	-80.72	0.01	0.12	0.27	2651.63
PG	33807	1.45	71.04	-914.3	-0.14	0.11	0.36	5758.76
RSI	35129	49.17	9.97	11.78	42.16	48.50	55.71	92.56
BIAS	35129	-0.32	6.17	-46.72	-3.59	-0.57	2.65	81.59
PSY	35129	46.91	10.55	5.00	40.00	45.00	55.00	85.00
DIF	35129	-0.05	2.92	-163.5	-0.28	-0.03	0.19	83.99
DEA	35129	-0.03	2.61	-97.67	-0.24	-0.02	0.19	106.97
MACD	35129	-0.04	2.58	-137.7	-0.22	0.00	0.18	106.70
AR	35113	113.22	41.03	0.00	85.62	107.24	134.25	2635.29
ARBR	35112	125.61	235.00	-6.10	-9.81	7.13	23.10	4419.00
ATR14	35129	620.75	731.21	0.00	0.63	2.95	9.70	9528.49
VOL5	35129	1.23	1.49	0.00	0.40	0.76	1.48	42.58
VOL60	35129	1.26	1.31	0.00	0.47	0.85	1.57	22.42
Sk20	35129	0.19	0.85	-4.47	-0.31	0.18	0.69	4.47
Sk60	35129	0.25	0.80	-7.75	-0.17	0.24	0.69	7.75

3.3. Feature engineering

A total of 29 features are selected in this paper. However, too many features may lead to feature redundancy, which not only increases the running time cost, but also sometimes reduces the performance of the trained model. Therefore, this paper uses the Wrapper packing method to select the most important features in the dataset and transforms the original feature matrix into a new feature matrix that contains only the selected features. The following table shows the ranking of the features.

Table 3. Feature Importance Ranking

hallmark	Rank	hallmark	Rank
EP	1	Profit_G_q	7
BP	1	RSI	1
SP	1	BIAS	1
CFP	1	PSY	8
financial_leverage	4	DIF	1
debtequityratio	1	DEA	1
cashratio	1	MACD	1
currentratio	6	AR	2
NI	5	ARBR	1
GPM	9	ATR14	1
ROE	1	VOL5	1
ROA	1	VOL60	1
asset_turnover	1	Skewness20	10
net_operating_cash_flow	1	Skewness60	3
Sales_G_q	1		

3.4. Model Evaluation

In this paper, we use GridSearchCV to perform a grid search for eight models to find the optimal parameters, and then evaluate the predictive performance of the models by the mean value of accuracy and the mean value of AUC.

Fig. 1 Comparison of Accuracy Means and AUC Means of Models

The above figure shows that XGBoost model has the best performance with a mean accuracy value of 84.76% and a mean AUC value of 0.84, in addition, Random Forest and Adaboost models also perform better with a mean accuracy value and a mean AUC value of more than 0.8. The performance of KNN, Logistic Regression, Support Vector Machines, and Decision Tree models among the remaining five is also relatively robust, and the performance of Gaussian Park Bayes has the worst model performance. Combining the above indicators, the performance of each model is ranked in descending order as follows: XGBoost, Random Forest, Adaboost, Support Vector Machine, KNN, Logistic Regression, Decision Tree, and Gaussian Spartan Bayes.The prediction ability of XGBoost, Random Forest, and Adaboost is significantly better than that of other algorithms, mainly because they all belong to the category of integrated learning. Integration learning algorithms typically improve prediction accuracy and generalization performance by combining multiple base learners (e.g., decision trees) to form a more powerful prediction model. Among them, Adaboost is a boosting algorithm that gradually improves the overall model performance by iteratively training new base learners and focusing on samples based on the performance of previous base learners. Random Forest, on the other hand, introduces randomness (e.g., randomly selecting a subset of features) during the decision tree training process, which enhances the generalization ability of the model and avoids overfitting. XGBoost, on the other hand, introduces regularization terms in the decision tree training, effectively controlling the complexity of the model and thus improving the generalization performance. These integrated learning algorithms are able to significantly improve the prediction ability and generalization while ensuring the model complexity through multiple strategies, such as boosting, randomness, and regularization, thus achieving better prediction results than a single base learner.

3.5. Quantitative strategy testing

The above results suggest that machine learning models do have some ability to predict Chinese stock returns. In order to further explore the performance of these models in quantitative investment, this paper analyzes each of the eight models with stratified backtesting. Stratified backtesting is a commonly used method for evaluating quantitative investment strategies, which divides the prediction results into multiple equal subgroups and then compares the return performance of each group with the benchmark strategy. Through stratified backtesting, we can more intuitively observe the performance of machine learning models in quantitative investment. If the return of the first group is significantly higher than that of the benchmark strategy, it indicates that the model does have some predictive ability in identifying low-yielding stocks and can be applied to actual investment decisions. On the contrary, if the returns are lower than the benchmark, it indicates that the model may have less value for application in quantitative investment. Specifically, this paper adopts the quintile method, i.e., the prediction results are evenly divided into five groups from low to high, and then the returns of the first group are compared with those of the benchmark strategy, as shown in Figure 2.

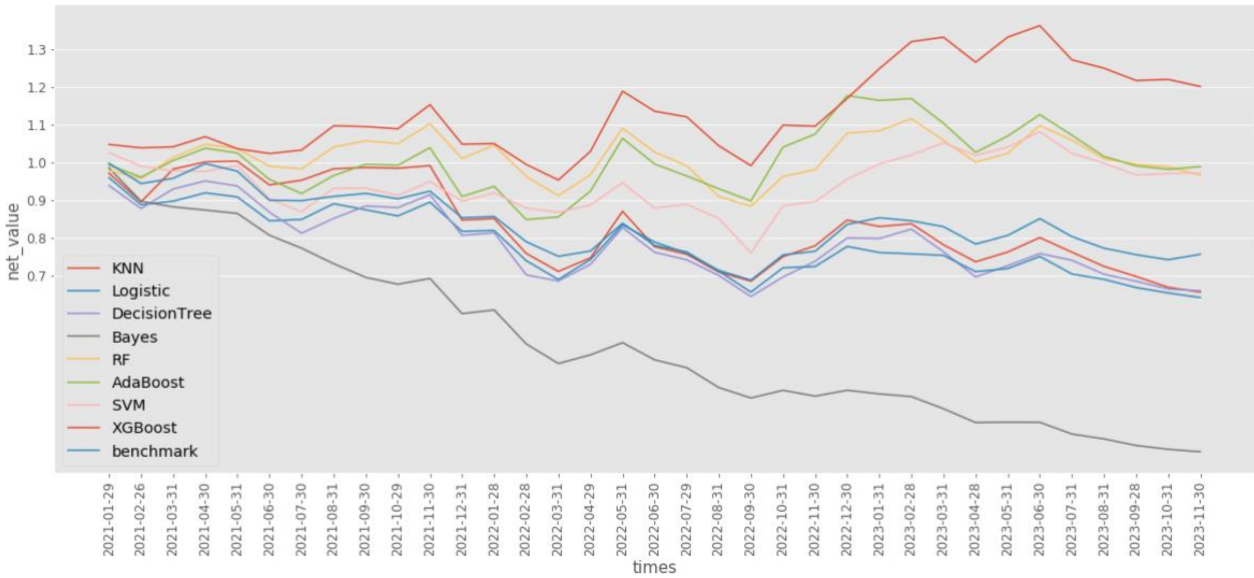

Fig. 2 Cumulative rate of return on quintiles of investment income

In addition, performance evaluation indicators such as annualized return, maximum retracement, and Sharpe ratio of each group of each model were statistically evaluated, as shown in Tables 3-4.

Table 4. Model Strategy Evaluation

	KNN	logistic	decision tree	Bayes	RF	Adaboost	SVM	XGBoost	standard of reference
Cumulative rate of return	-34.50%	-24.40%	-34.10%	-76.80%	-3.40%	-1.20%	-2.90%	20.10%	-35.90%
Annualized rate of return	-13.50%	-9.14%	-13.32%	-39.40%	-1.18%	-0.41%	-1.00%	6.48%	-14.14%
maximum retracement	34.64%	28.34%	32.25%	76.43%	19.88%	18.35%	25.89%	17.33%	35.71%
Sharpe ratio	-2.71	-2.42	-2.75	-8.14	-0.96	-0.71	-0.97	0.46	-3.87
Annualized excess return	1.33%	5.90%	1.37%	-29.07%	14.98%	16.34%	14.91%	23.58%	0.00%
Maximum excess return for the month	8.25%	5.80%	5.48%	3.07%	4.92%	6.13%	6.64%	8.86%	0.00%
Outperforms benchmark as a percentage	51.43%	57.14%	57.14%	17.14%	65.71%	74.29%	62.86%	68.57%	0.00%
Percentage of months with positive returns	45.71%	42.86%	40.00%	20.00%	42.86%	45.71%	54.29%	48.57%	37.14%
Information ratio	0.2	2.31	0.25	-9.59	5.76	4.84	4.85	7.37	-

As can be seen from the above chart, the Gaussian Plain Bayes model has the worst performance, with a model return of -76.80%, and its maximum retraction is also large, while the strategy returns of the other machine learning models exceed the benchmark return, of which XGBoost has the best performance, with a strategy return of 20.10%, and it is also the only one of the eight model strategies that has achieved positive returns, in addition to the Random Forest, Support Vector Machine and In addition, Random Forest, Support Vector Machine and Adaboost did not achieve positive returns, but compared to the benchmark return is also achieved a great improvement, the performance of the remaining three models is average, and their strategy returns are about the same as the benchmark return. From the perspective of risk evaluation indicators such as maximum retracement and Sharpe ratio, the performance of XGBoost, Adaboost, and Random Forest models are also better among the eight models, with maximum retracement of 17.33%, 18.35%, and 19.88%, and Sharpe ratios of 0.46, -0.71, and -0.96, respectively.From the perspective of the other ratios, it is also the case that XGBoost, Adaboost, and Random Forest are the three models with better performance, which also confirms that these modeling strategies have relatively low risk while achieving higher returns.

4. Summary

In this paper, we have constructed a quantitative investment strategy using machine learning algorithms by researching and learning the knowledge related to quantitative investment and machine learning. The main work and conclusions of this paper are as follows:

(a) Data processing and feature engineering: In this paper, we collect Chinese A-share market data for 10 years from January 1, 2014 to December 31, 2023, including stock prices of CSI 300 constituent stocks and seven types of factor indicators, and carry out the necessary data cleaning to prepare for the subsequent model training. In addition, considering the effect of feature redundancy, this paper refers to Huang et al. (2008) who chose the Wrapper method, which evaluates the performance of each subset by trying different subsets of features, and then gives a ranking of the importance of each feature and selects the best subset of features.

(b) Model Evaluation. The eight algorithms of KNN, logistic regression, decision tree, Gaussian plain Bayes, support vector machine, Adaboost, random forest, and XGBoost were used to evaluate the prediction performance of the eight models trained using accuracy and AUC area after the data within the sample were tuned by GridSearchCV, and the prediction performance of each model was ranked in descending order as follows:XGBoost, Random Forest, Adaboost, Support Vector Machine, KNN, Logistic Regression, Decision Tree, and Gaussian Plain Bayes. The XGBoost model has the best performance with the mean accuracy of 84.76% and the mean AUC of 0.84, while the Random Forest and Adaboost models also perform better with the mean accuracy and the mean AUC greater than 0.8. The performance of the remaining five models, namely, KNN, Logistic Regression, Support Vector Machines, and Decision Tree, is also relatively robust, and the performance of Gaussian Spartan Bayes is the poorest. model has the worst performance.

(c) Investment strategy evaluation and application. In this paper, eight models are separately backtested in a hierarchical manner, and the evaluation indexes such as annualized return, maximum retracement, Sharpe ratio, etc. of each group of each model are counted, and finally the first group of each model is compared with the benchmark. The performance of each model strategy is listed in descending order as follows: XGBoost, Adaboost, Support Vector Machine, Random Forest, Logistic Regression, Decision Tree, KNN, Gaussian Plain Bayes. Among them, the Gaussian Plain Bayes model has the worst performance, with a model return of -76.80%, and its maximum retracement is also as high as 76.43%, while the strategy returns of the other machine learning models exceed the benchmark return, of which the XGBoost has the best performance, with a strategy return of 20.10%, and a Sharpe ratio of 0.46, which is the only one of the eight models that has achieved a positive return, and in addition, the random forest, support vector machine, and Adaboost have the best performance. Forest, Support Vector Machines and Adaboost did not achieve positive returns, but compared to the benchmark return is also achieved a great improvement, the remaining three models performance is average.

This paper provides investors with a strategy construction method that integrates artificial intelligence into quantitative investment, and provides a valuable reference for subsequent investment decision optimization and practical application.

References

[1] Vapnik V N. The nature of statistical learning theory [M].Springer,1995.

[2] Cao L J,Tay Feh, ,Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting [C] IEEE Transaction on Neural Networks, 2003 ,14(6).

[3] Kim.K, Han.I, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index[J],. Expert Systems with Applications, 2000, (19): 125-132.

[4] Huang W, Nakamori Y, Wang S Y. Forecasting Stock MarketMovement Direction with Support Vector Machine [J].Compu ters & Operations Research, 2005, 32 (10):2513 ~ 2522.

[5] Breiman. L. Random Forests [J], Machine Learning, 2001, 45:5-32.

[6] Kumar, Manish and Thenmozhi, M. Forecasting Stock Index Movement: a Comparison of Support Vector Machines and Random Forest [Z]. Indian Institute of Capital Markets 9th Capital Markets Conference Paper. 2005.

[7] Cao Zhengfeng, Ji Hong, Xie Bangchang. Using Random Forest Algorithm to Realize the Selection of High Quality Stocks [J], Journal of Capital University of Economics and Business, 2014, 2:21-27.

[8] Ballings, Michel, Van den Poel Drik, Nathalie Hespeels, and Ruben Gryp, 2015, Evaluating multiple classifiers for stock price direction prediction , Expert Systems with Applications, 42, 7046-7056.

[9] MENG Ye, YU Zhongqing, ZHOU Qiang. A safe quantitative stock picking strategy - Empirical evidence from A-share market[J]. Financial Theory and Practice,2018(08):102-107.

[10] Chen T Q,Guestrin C.XGBoost:A scalable tree boosting system[C].CA:22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD),San Francisco,CA,2016,785-794.

[11] Basak, Suryoday, Saibal Kar, Snehanshu Saha, and Luckyson Khaidem, 2019, Predicting the direction of stock market prices using tree-based classifiers, North American Journal of Economics and Finance, 47, 552-567.

[12] Laitinen E K, Laitinen T. Bankruptcy prediction: application of the Taylor's expansion in logistic regression [J]. International Review of Financial Analysis,2001,9(4) : 327-349.

[13] FANG Yi, CHEN Yuzhi, WEI Jian. Artificial intelligence and China's stock market - A quantitative study of investment portfolios based on machine learning prediction[J]. Industrial Technology and Economics,2022,41(08):83-91.

[14] WU Shi-nong,LU Xian-yi. Research on the prediction model of financial distress of listed companies in China[J]. Economic Research,2001,(06):46-55+96.

[15] Huaqin Wang. Quantitative stock selection based on machine learning and evaluation of its effect[D]. University of Electronic Science and Technology,2020.DOI:10.27005/d.cnki.gdzku.2020.003254.

[16] Huang C J, Yang D X, Chuang Y T. Application of Wrapper Approach and Composite Classifierto the Stock Trend Prediction [J]. Expert Systems with Applications, 2008, 34:2870-2878.

[17] PENG Li-Fang, MENG Zhi-Qing, JIANG Hua et al. Application of time series-based support vector machine in stock forecasting[J]. Computing Technology and Automation, 2006, 25(3):70-93.

[18] Fang Kuangnan, Zhu Jianping, Xie Bangchang. Research on fund return direction prediction and trading strategy based on random forest method[J]. Economic Jingwei, 2010(02):61-65.

[19] Zhou Liang. Research on multi-factor investment in stocks based on random forest model[J]. Financial Theory and Practice,2021,(07):97-103.