ISSN: 1813-4890

Construction and Empirical Study of Multifactor Stock Selection Model Based on Random Forests

--Taking the SSE 50 Index as an Example

Yuchun Cai, Hongbo Zhang, Ziyi Liu, Jing Wen

School of Anhui University of Finance and Economics, Bengbu, China;

Abstract

The effectiveness of China's stock market is weak, the market efficiency is low, and there are relatively more stocks with pricing bias in the market. The use of objective and rational quantitative investment strategies can help investors capture investment opportunities and increase the possibility of obtaining returns. Compared with traditional investment strategies, quantitative investment strategies in the construction of investment portfolios, often based on the processing and analysis of a large amount of data, the analysis of the level and angle of the more diversified, comprehensive, systematic; based on the probability of winning the historical laws of mining, construction of investment portfolios, diversification; reduce the impact of investors' own subjective factors, accurate and objective selection of stocks and trading, accuracy and discipline; the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, the market, and the investment strategy, the market, and the investment strategy. Discipline; can respond to market changes more quickly, with timeliness. Multi-factor stock selection model is one of the commonly used models for quantitative stock selection. Based on this, the random forest algorithm is chosen for modelling. In this paper, we select stocks from the SSE 50 sample stocks to construct the stock pool, and use the random forest algorithm to model the model. Finally, the model is evaluated and analysed by combining its own operation effect and backtest results. The dynamic learning model constructed in this paper reflects its timeliness, and to a certain extent, it can reflect the changes of the market. It is hoped that the research in this paper can help enrich the theories in the field of behavioural finance and quantitative investment, and provide new ideas for investors to choose quantitative trading strategies.

Keywords

Quantitative investment; multifactor stock selection models; random forests.

1. Introduction

1.1. Background and significance of the study

1.1.1. Background of the study

With the development of the financial market and the increase in the number of investors, how to make effective investment has become the focus of attention of many investors. And in the field of stock investment, stock selection is the first and most important step for investors. Traditional stock selection methods are mainly based on technical and fundamental analyses, but these methods are often affected by market conditions and short-term fluctuations, and lack stability and sustainability. Therefore, it is essential to find a stable and reliable stock selection method. The current stock selection strategy in the stock market is mainly based on

ISSN: 1813-4890

the multi-factor stock selection model, which assesses the value and risk of a stock by comprehensively analysing multiple factor variables, such as price-earnings ratio, price-net ratio, dividend yield, etc., in order to construct a quality stock portfolio. The advantage of the multi-factor stock selection model is that it can avoid the selection bias of a single factor while considering the complex relationship between multiple factor variables, thus improving the accuracy and efficiency of stock selection.

In the financial market in recent years, Random Forest, as an integrated learning method, has been widely used in financial data analysis and prediction with good results. Random forest is an integrated learning method based on decision trees, and by integrating multiple decision trees, it can substantially reduce the overfitting risk of the model and improve the accuracy and stability of the model. In addition, multi-factor stock selection model is one of the more widely used stock selection methods in recent years, which can combine multiple factors to comprehensively evaluate stocks and improve the accuracy and reliability of stock selection. The multi-factor stock selection model based on random forest can consider the complex relationship between multiple factor variables comprehensively and avoid the limitations of traditional models. At the same time, Random Forest has better generalisation ability and stability, which can improve the accuracy and stability of stock selection. Therefore, the multi-factor stock selection model based on random forest is widely used in practice.

This study takes the constituents of SSE 50 Index as an example. SSE 50 Index is an important stock index in China's securities market, which consists of 50 listed companies with large scale, good liquidity and strong representativeness. It is an important reference index in China's securities market, and investors can study this index to understand the overall trend and risk of China's securities market, so as to make more informed investment decisions. At the same time, this study will adopt the learning algorithm of random forest to select multiple factors for analysis and explore the application value of multi-factor stock selection strategy in stock investment, so as to provide investors with scientific stock selection methods and investment suggestions.

1.1.2. Significance of the study

The purpose of this study is to construct a multi-factor stock selection model based on random forests and to conduct an empirical study on the constituent stocks of the SSE 50 index as an example. The significance of the study mainly includes the following aspects:

Providing a new method for stock selection

As a powerful machine learning method, Random Forest has been widely used in the financial field, but it has not been widely used in the field of stock selection. This study applies random forest to the field of stock selection and constructs a multi-factor stock selection model based on random forest, which provides investors with a new, stable and reliable method of stock selection.

Improve the accuracy and stability of stock selection

Compared with traditional stock selection methods, the random forest-based multi-factor stock selection model can combine multiple factors to make a comprehensive assessment of stocks, avoiding the misjudgement and risk caused by a single factor. In addition, as an integrated learning method, Random Forest can effectively reduce the overfitting risk and improve the accuracy and stability of the model, thus improving the accuracy and stability of stock selection and reducing investment risk.

Explore the scope of application of multi-factor stock selection model

At present, multi-factor stock selection model has become a hot topic in the field of stock selection, but different factor combinations may have different applicability to different stocks and markets. Therefore, this study explores the scope of applicability of multi-factor stock selection models under different market conditions by conducting an empirical study using the

ISSN: 1813-4890

constituents of the SSE 50 Index as an example, so as to provide investors with targeted stock selection suggestions.

Provide practical experience and reference value

This study applies the multi-factor stock selection model based on Random Forest to the empirical study of the constituents of the above SEC 50 index, which provides a feasible stock selection methodology and practical experience, and has a certain reference value for investors in stock selection. In addition, the methodology and ideas of this study can also provide reference for other stock selection and financial data analysis fields.

2. Investment design

2.1. Setting of investment objectives

In this study, we take the constituents of the SSE 50 Index as our sample and use regression analysis to determine a multi-factor stock selection strategy, aiming to achieve excess returns relative to the market while controlling investment risk. Specifically, our investment objectives are set as follows:

Short-term objective: to achieve a significant positive quarterly return relative to the CSI 300 Index.

2. Long-term objective: to achieve a stable and significantly higher investment return than the market average, with long-term cumulative investment returns significantly higher than the market average.

In determining the investment objectives, we need to consider the market situation and the economic situation, and at the same time, we need to formulate a specific investment plan based on the investor's risk appetite and asset allocation. In respect of the tourism industry, China's tourism market is growing and has broad market prospects and investment opportunities. Therefore, the investment objective of this study is based on clear grounds and prospects.

In order to reach our investment objectives, a series of investment strategies and risk control mechanisms need to be developed, which will be elaborated in detail in the following.

2.2. Parameter optimisation and testing of a multi-factor stock selection model based on random forests

2.2.1. Factor characterisation test

In this paper, we next select the stock data from 31 March 2013 to 31 December 2018 as a sample, by using spss for feature test. The following table demonstrates the importance ratio of each feature, based on the feature importance, we removed the factors that are less than 3%: earnings per share (diluted) ($/share), operating profit per share ($/share), net assets per share ($/share), basic earnings per share, net assets per share, attributed net profit, and net profit after deductions.

Table 1 Proportion of importance of each characteristic

Feature name	Characteristic importance
Quarterly turnover (%)	27.50%
PB_PB	11.60 per cent
Market Capitalisation Ratio_PCF	4.70 per cent
Market_Sales_Ratio_PS	3.80 per cent
Earnings per share (diluted) ($/share)	1.50 per cent
Return on net assets (diluted)	3.50 per cent

ISSN: 1813-4890

Operating profit per share (RMB/share)	1.90 per cent
Net assets per share (yuan/share)	2.10%
Operating income per share (yuan/share)	3.10%
basic earnings per share	2.90 per cent
net asset per share	2.30 per cent
return on net assets	6.40 per cent
return on total assets	4.00 per cent
Return on invested capital	5.50 per cent
gearing	4.30 per cent
total revenue	5.20 per cent
Net profit attributable to	2.80 per cent
Non-deductible net profit	2.50 per cent
gross margin	4.50 per cent

2.2.2. Random Forest Parameter Optimisation

Random Forest is a machine learning model based on integrated learning, which consists of multiple decision trees, and the result of each decision tree is integrated into the final prediction result. In Random Forest, each decision tree is constructed by sampling the original data with put-back, which reduces the risk of overfitting while ensuring the randomness of the data. In addition, the splitting evaluation criteria for each node are commonly known as Gini coefficient and mean square error (mse). mse is a commonly used splitting criterion for regression trees because it can effectively measure the magnitude of error between the predicted value and the true value.

In this study, we performed data shuffling to ensure that the samples used for each decision tree are independent. We used mse as the splitting evaluation criterion for the nodes and tuned the parameters such as the maximum depth of the tree, the maximum number of leaf nodes, and the number of decision trees. Specifically, we chose the maximum depth of the tree to be 10, the maximum number of leaf nodes to be 50, and the number of decision trees to be 500.These parameters were chosen based on our exploratory analyses of the dataset and experience from previous studies. Among them, the maximum depth of the tree and the maximum number of leaf nodes are parameters that limit the complexity of the decision tree, which can effectively avoid the risk of overfitting; the number of decision trees is a key parameter to control the complexity and accuracy of the model, and 500 trees can effectively improve the predictive ability of the random forest; there is a put-back sampling to increase the randomness and reduce the risk of overfitting; and conducting out-of-bag data testing can effectively assess the generalisation ability of the model.

Table 2 Projected evaluation indicators

	MSE	RMSE	MAE	MAPE	R^2
training set	0.016	0.128	0.101	712.073	0.696
test set	0.047	0.216	0.145	537.70	0.084

In the above table, the prediction evaluation metrics of the cross-validation set, the training set and the test set are demonstrated to measure the prediction effect of the random forest through quantitative metrics. Where the evaluation metrics of the cross-validation set allow for constant adjustment of the hyperparameters to obtain a reliable and stable model. Where MSE is the expected value of the squared difference between the predicted and actual values. The smaller the value is, the higher the model accuracy is; RMSE is the square root of MSE, the smaller the

ISSN: 1813-4890

value is, the higher the model accuracy is; MAE is the mean of absolute error, which can reflect the actual situation of the prediction value error. The smaller the value, the higher the model accuracy; MAPE is the deformation of MAE, which is a percentage value. The smaller the value, the more accurate the model is; R^2 compares the predicted value with the mean value only, the closer the result is to 1 the more accurate the model is. It can be seen that the model works better.

2.2.3. Model testing

The data for the training set is from 31 March 2013 to 31 December 2018, and the data for the test set test period study period 31 March 2019 to 30 September 2021. In order to better show the test, set data, whether the difference between the real value and the test value is large, so we show the test data graph to see whether the predicted results are close to the actual results.

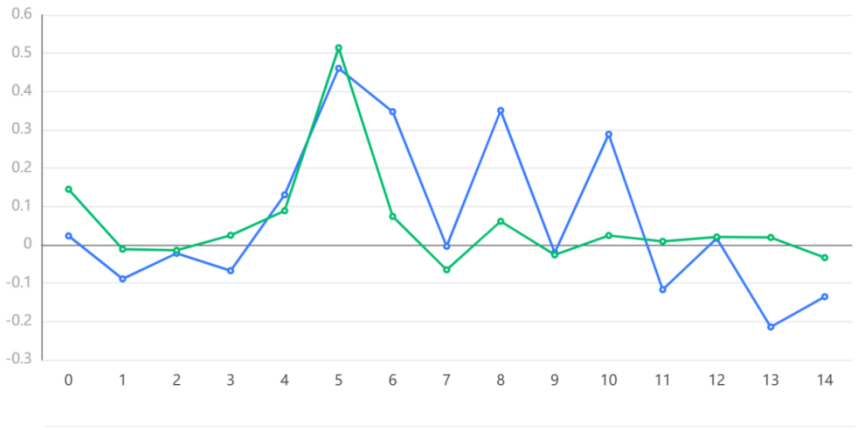

Fig. 1 Test data graph

As you can see from the graph the test is acceptable. In this way the method of stock selection with random forest machine learning can be carried out. In order to further validate the effectiveness of the model as well as to select better stock portfolios, we will weight the stocks, and in doing so, the optimisation objective and constraints should be considered. The optimisation objective should be to maximise the expected return of the portfolio while minimising the risk of the portfolio. The constraints should be that the weight of each stock should be greater than 0 and the sum of all stock weights should be equal to 1. In addition, the upper limit of the individual stock weights should be set with a moderate degree of certainty, and it should not be set too large or else it will concentrate on buying the highest expected stocks, which is a very risky investment portfolio. Of course, it should not be set too small either, as the number of stock portfolios is limited and there is a need to ensure the yield of the stock portfolios.

Here we use Principal Component Analysis (PCA) method for empowerment. Principal Component Analysis (PCA) is a dimensionality reduction technique that extracts the main features and reduces the feature dimensions from the raw data, and is suitable for situations where high dimensional data is being processed. In stock investing, the PCA method can be used to identify stocks with high composite scores and then assign different weights based on the level of scores.

Table 3 Results of PCA algorithm

rankings	line index	aggregate score	Principal component 1
1	Guizhou Maotai town in Renhuai county, Guizhou	2.998493031	8.022333675

ISSN: 1813-4890

7	COSCO Sea Controls (company)	1.928433416	2.258974388
16	China National Petroleum Corporation	1.333324455	0.419466898
22	Wanhua Chemical	1.172689689	1.842679645

The weights of individual stocks should be set moderately, not too large or you will be concentrating on buying the stocks with the highest expectations, and certainly not too small or the portfolio's yield will not be guaranteed.

Using the PCA algorithm, we selected 4 stocks for the portfolio, which also have the best expected returns, and therefore are assigned weights based on the composite score.

Table 4 Table of weights

Guizhou Maotai town in Renhuai county, Guizhou	COSCO Sea Controls (company)	China National Petroleum Corporation	Wanhua Chemical
0.38	0.26	0.19	0.17

Next we take the portfolio of stocks selected above and run a rolling test. We bring the factors for Q1 2019 into the model for the regression to get the expected returns for Q2 2019 and hold them from Q2 2019 until Q3 2021. Details are given in Table 5 below.

Table 5 Portfolio of stock picks

Stock Name	stock code (computing)	expected return	Real average earnings	holding period
Guizhou Maotai town in Renhuai county, Guizhou	000069	3.4141784	2.2139	2019q2-2021q3
COSCO Sea Controls (company)	600009	5.2669533	4.56093	2019q2-2021q3
China National Petroleum Corporation	601888	-0.21451	-0.0787	2019q2-2021q3
Wanhua Biological	601111	1.6218	3.145174	2019q2-2021q3

As can be seen from the table, the expected returns based on the model do not differ much from the actual returns of the stock in a particularly large way, indicating that the model can be used for stock selection.

3. Performance assessment

3.1. Indicators and methods of performance assessment

Performance assessment indicators, i.e., assessment factors or assessment items. Performance assessment indicators refer to the aspects from which the performance of the subject of the assessment is measured or evaluated and address the question of what is being assessed.

In the performance appraisal process, the aspects that point to the performance of the appraisee are the appraisal indicators. Performance assessment indicators are used to measure whether or to what extent the actual results of the assessed person's behaviour meet the performance objective. For a specific performance objective, it is necessary to establish numerous relevant assessment indicators, including indicators of quantity, quality, timeliness, cost, output and so on. Performance assessment indicators are a key factor affecting the objectivity and accuracy of assessment results, and a scientific and comprehensive performance assessment indicator system should be established in order to accurately and scientifically assess the performance of the public sector.

ISSN: 1813-4890

Investors can use multiple benchmarks to monitor portfolio performance or to measure the performance of individual investment securities. The primary portfolio monitoring metric for performance is total return, which is usually measured against a benchmark. Other metrics include statistical risk methods such as Standard Deviation, Beta, R-Squared, Sharpe Ratio and Sortino Ratio.

Measuring portfolio performance is more than just total return. Investors choose appropriate benchmarks wisely. Statistical metrics such as standard deviation, Sharpe ratio and R-squared can be used to measure portfolio risk.

3.2. Performance evaluation of investment strategies based on the constituents of the above SEC 50 Index

The four stocks initially selected are held from the first quarter of 2019 until the third quarter of 2021. The weighted returns of the stock portfolios are obtained as shown in Table 7 below. Over this duration, the maximum return is 119% and the maximum retracement is 66%, indicating that the high risk of the stock selection model designed in this paper is accompanied by excess returns.

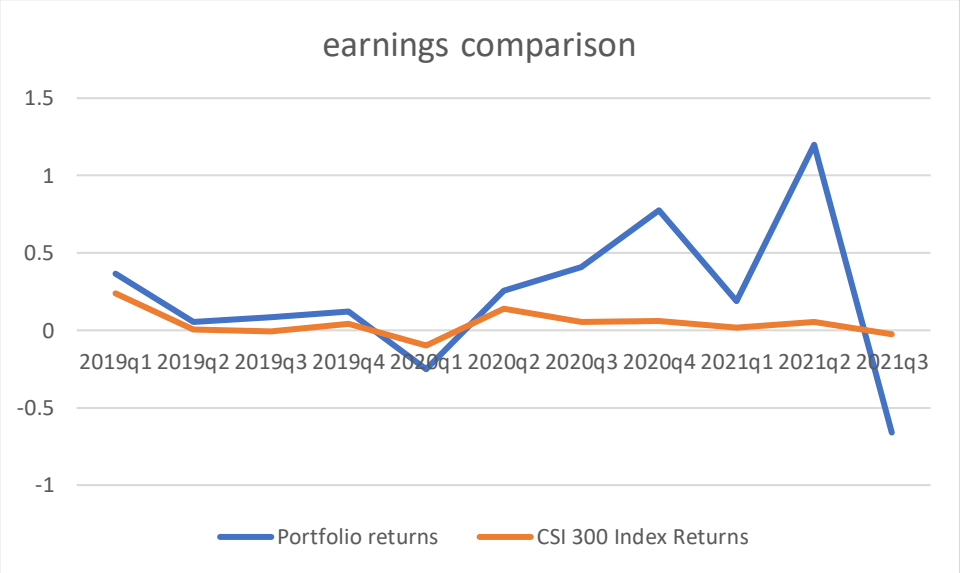

Figure 2 Comparison of returns

It can be seen that the model constructed through the multi-factor regression selected portfolio is significantly better than the performance of the CSI 300 index, even if the period is lower than the CSI 300 index, but the difference is very small, basically the same as the broader CSI 300, long-term holdings, you can get a better excess return, the model effect is more significant, the multi-factor stock selection model model has a certain degree of practicality. This model and method of stock selection is worth further research.

4. Conclusion and Recommendations

4.1. Advantages and shortcomings of the investment strategy

4.1.1. Advantages of the investment strategy

Multi-factor stock selection strategy can select stocks by considering multiple factors, which is more accurate and comprehensive compared with single-factor stock selection strategy. Based on regression analysis, we propose a multi-factor stock selection strategy that can use historical data and models to predict the future performance of stocks, improving the accuracy and efficiency of investment decisions. By setting stop-gain and stop-loss points and making regular adjustments to asset allocation, it can effectively control risks and maximise returns. On special

occasions, if an epidemic or other unfavourable factors cause an impact on the market, temporarily casting back to avoid the risk, observing the surrounding environment, and then buying can reduce the investment risk during a period of high risk.

4.1.1.1 Disadvantages of the investment strategy

The effectiveness of an investment strategy depends on the accuracy and validity of the stock selection factors, and the factors may also vary from market to market, making it difficult to implement them consistently over time. Proposing an investment strategy based on historical data and model forecasts has certain risks and limitations with respect to future market uncertainties. Investors should have certain investment experience and knowledge, and a deep understanding of the stock market in order to properly implement this investment strategy and make relevant decisions. In some special occasions, such as emergencies, investors need to react in time and adjust their strategies in order to effectively reduce the risk and gain returns, otherwise there is a risk of loss.

4.1.1.2Suggestions for future research and direction of improvement

Future research and practice could be refined and deepened in the following ways.

1. Improve the diversity and flexibility of stock selection factors. A multi-factor stock selection strategy based on random forests takes into account more of a company's financial data and other basic factors, and in the future technical factors, the impact of market sentiment and so on can be added to think about stock performance and risk control from multiple perspectives.

2. Enhance macro research on markets, industries and other factors. The performance and risk of a stock are affected by internal factors of the company, as well as by the macroeconomic and policy environment, industry competition and other factors. In the future, research and analysis of these factors should be strengthened and integrated into stock selection strategies to make them more comprehensive and provide a precise judgement on the performance and risk of individual stocks.

3. Strengthen risk management and control. Although the multi-factor stock selection strategy based on regression analysis can control risks, risk management and control is a dynamic process that requires continuous monitoring and adjustment. In the future, more flexible take-profit and stop-loss strategies and more detailed asset allocation programmes can be adopted to meet different market environments and investment needs.

In summary, although the multi-factor stock selection strategy based on random forest proposed in this paper has certain limitations and risks, it can be improved and innovated in many aspects in the future research and practice process to improve the efficiency and precision of their decision-making, so that it can better serve the majority of stock investors.

References

[1] Yang Li-Xin. A comparative study of multi-factor stock selection models in China's A-share market[D]. Capital University of Economics and Business,2021.DOI:10.27338/d. cnki.gsjmu. 2021.000574.

[2] Qiu Chengshi. Application research on multi-factor stock selection model based on integration algorithm[D]. Zhongnan University of Economics and Law,2021.DOI: 10.27660/ d.cnki. gzczu. 2021.001683.

[3] Hu B. Construction and application of multi-factor stock selection model based on random forest[D]. Sichuan University,2021.DOI:10.27342/d.cnki.gscdu.2021.000406.

[4] Li J. Research on multi-factor stock selection model based on random forest algorithm[D]. Harbin Institute of Technology, 2019.DOI:10.27061/d.cnki.ghgdu.2019.002033.

[5] Liu Q. Z. Construction of Multifactor Stock Selection Model Introducing Investor Attention and Random Forest Algorithm[D]. Shanghai Normal University, 2021.DOI: 10.27312/d. cnki.gshsu. 2021.002188.

[6] Su Jingyu. An empirical study of multi-factor stock selection model in A stock market [D]. Anhui University,2018.
[7] Zhou Zhan. Empirical research on multi-factor stock selection model based on SVM algorithm[D]. Zhejiang Gongshang University,2017.