2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST)

Research on quantitative multi-factor stock selection model based on machine learning

Ming Yang*
School of Computer Science and Technology
Shandong University of Finance and Economics
Jinan, China
* Corresponding author: 20017887@sdufe.edu.cn

2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST) | 978-1-6654-0267-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/IAECST54258.2021.9695937

Abstract—**With the popularization of computer technology, the quantitative investment business in China has been developing rapidly, but it started late compared to the developed markets, and its related theories and applications have received very high attention in the investment field in China. In this paper, we focus on the effectiveness of multi-factor stock selection model applied to A-share market and use machine learning algorithms to improve the performance of the model. First, this paper selects 10 relatively independent valid factors from 44 candidate factors by validity testing and redundancy removal; second, builds a multi-factor model and tests its validity, and introduces a hedging mechanism to further reduce the maximum retracement of the model; then changes the factor retracement period to select the model with the best effect for backtesting; finally, combines the AdaBoost algorithm with the multi-factor model and and the traditional Finally, the stock selection effect of combining AdaBoost algorithm with multi-factor model and and traditional multi-factor model is compared. The study finds that the machine learning AdaBoost [1] algorithm can enhance the stock selection effect of the traditional multi-factor model, and this study has a significant role in the future development of this field.**

Keywords-Quantitative investment, stock selection, multi-factor, AdaBoost, machine learning

I. Introduction

With the rapid development of computer technology, quantitative investment has entered people's vision. Quantitative investment theory is based on the assumption of efficient market and the principle of no arbitrage opportunities. In a broad sense, quantitative investment is the process of using computer technology and practicing financial investment concepts and methods with the help of financial mathematical models, and executing investment strategies to obtain certain returns. Subjective trading is a comprehensive consideration of investment targets and trading decisions made by investors based on actual market experience, asymmetric information, and investment preferences. Quantitative trading, on the other hand, is a data-driven process of establishing appropriate quantitative models, using computer technology and statistical principles or even artificial intelligence technology to analyze the intrinsic factors of data changes, and finally the trading signals generated by the models dominate the trading decisions. Nowadays, artificial intelligence (AI) is very effective in exploring complex non-linear laws, in terms of modeling to compensate for the singularity of the human brain's logical thinking patterns; in terms of computing power, it can achieve deep mining of massive data through "smart" algorithms. The core idea is to train different algorithms for the same training set, i.e., weak learning algorithms, and then integrate these weak learning algorithms to construct a stronger learning algorithm, which is an iterative algorithm. The empirical part of this paper is to combine the multi-factor model and AdaBoost algorithm to screen out the factors that perform better and more effectively in the A-share market, then construct a stock selection model by the effective factors and test the validity and feasibility of the model to select a more valuable stock portfolio, and finally observe the market performance of the portfolio through backtesting.

II. Overview of Multi-factor Stock Selection Models

A. Theoretical Basis of Multi-factor Models

A simple version of the multi-factor model is the single-factor model, with the advantage that the multi-factor model covers factors that have a high impact on stock returns and are highly adaptable to the market. The single-factor model captures the most effective factors for a given time period in a short period of time, but the factors are not very persistent and are prone to large retracements. Multi-factor models summarize the information of a large number of factors, which is more stable in terms of returns than single-factor models, and more sustainable in terms of time. Therefore, for the results of multi-factor model backtesting, the requirements of model setting will be more stringent. The establishment and development of multi-factor models cannot be formulated without the following three models:

1) *CAPM model:* The economists led by William Sharpe, Jan Mossin and other scholars, they proposed the Capital Asset Pricing Model (CAPM) to further simplify Markowitz's theory from an empirical point of view. The CAPM model is widely used in practice and is the core foundation of modern financial market price theory. The capital asset pricing model describes that the expected return of each security is equal to the risk-free rate plus a beta multiple of risk compensation.

2) *APT Model:* APT is an extension of the CAPM, which gives a pricing model and a CAPM model both in equilibrium, and the difference between the two is that APT is based on a factor model. The factor model is a statistical model in which the starting point of arbitrage pricing theory is the assumption that the return on a security is related to an unknown number of unknown factors. the APT is the use of the factor model to portray the determinants of asset prices and the formation of equilibrium prices.

3) *Fama-French three-factor model:* In 1992, Fama and French examined the determinants of the difference in returns

among stocks and found that beta did not explain the difference, but found that the price-to-earnings ratio, book-to-market ratio, and market capitalization of listed companies could explain the difference in returns among stocks. Fama and French-agreed that excess returns compensate for the risk not reflected in the beta of the CAPM model.

B. Selection of effective factors

Most multi-factor model researchers prefer to use fundamental and trading data of listed companies as candidate factors, and this paper is no exception. The purpose of the multi-factor model is to study which factors have a greater impact on stock returns in a particular period, and use the effective factors to build a stock selection model for investment. Effective factors are mainly those factors that have a significant impact on investment returns. In general, the validity of a factor is judged by the correlation between the portfolio return and the factor score. However, the specific situation should be analyzed specifically, and other judgment conditions can be added according to the situation.

C. Elimination of redundant factors

When completing the selection of valid factors, we also need to consider an important issue: the correlation between the valid factors. If the factors are highly correlated with each other, the stock portfolios selected by the model may have a high degree of consistency in many aspects, such as returns, individual stock selection, etc. The redundant factor elimination mainly targets the factors with high correlation, and enhances the explanatory power and differentiation of individual factors by eliminating the redundant factors. In this paper, we use the correlation coefficient matrix to eliminate the factors with high correlation and keep one of them. The elimination of factors is based on the experience of some other scholars and literature.

D. Portfolio construction

After the multi-factor model is established, the portfolio construction is carried out. Generally, all the constituent stocks of the market index are used as a pool of stocks, and the top ranked stocks are selected by scoring according to the size of the effective factors. After the portfolio is constructed, the portfolio performance is tested, which is mainly to check the portfolio risk and return. In this paper, the portfolio returns are compared with the market benchmark index returns to analyze whether the portfolio selected by the multi-factor model can achieve excess returns in the backtest interval.

III. Empirical analysis of A-share market

A. Data Sources

This paper obtains empirical related data through the JoinQuant quantitative trading platform, mainly by obtaining the closing prices of all constituents of the CSI 500 Index (000905. XSHG) for each trading day from January 1, 2010 to December 31, 2017 and the monthly fundamental financial data of all constituents, and using the CSI 500 Index constituents as the pool of stocks for this paper's empirical evidence. The CSI 500 index reflects the situation of medium-capitalization listed companies, which are also the mainstay of the A-share market, and has an important mechanism of being able to exclude downward and exclude upward. The upward exclusion refers to the removal of stocks with large market capitalization, while the downward exclusion is the removal of stocks that are trading too small. The reason for not choosing CSI 300 constituents as the pool of stocks for the empirical evidence in this paper is that the update of CSI 300 constituents can only update the tail stocks, while the large weighted stocks basically remain unchanged, which severely suppresses the increase of CSI 300 index. Therefore, this paper empirically selects all CSI 500 index constituents as a basket stock pool, which better reflects the actual market situation.

B. Multi-factor selection

In this paper, 37 financial indicators were selected as fundamental factors, 2 market sentiment indicators as sentiment-based factors, 2 technical indicators as technical indicator-based factors and 3 WorldQuant Alphal01 factors. The factors were selected to follow the direction of today's economic development and to accurately reflect the laws and characteristics of the market, changing the traditional financial factors as appropriate. These factors were selected mainly based on a large number of domestic and international research reports and relevant scholarly literature, and include the following factors: Growth, Valuation, Size, Quality, Trading, Sentiment, Technical Indicator and WorldQuant Alpha101 partial factors (3 of them).

C. Factor validity test

Due to the large amount of sample data, this paper analyzes the above candidate factors using the ranking method. First, the factor size of each candidate factor at the end of each month in the test period is calculated, and the stocks of the stock pool are sorted and grouped with equal weights according to the size of this factor and held until the end of the month, and the above process is repeated at the end of the next month. The stock pool selected for this paper is the CSI 500 index, where the 500 constituent stocks are sequentially divided into 10 groups of 50 stocks each, evenly according to the size of the candidate factors. In general, the fund has a position of 60 to 70 stocks, of which 50 are long positions. This paper and the fund company have a comparable number of positions, which is more in line with the norm. Secondly, the main purpose is to compare the average return of the last group and the first group, in order to ensure that one of the two extreme combinations has a high probability of outperforming the market and one of underperforming the market. Finally, the correlation between the factors and the stock returns is tested, where monotonicity is not required only a large correlation is needed to screen out the factors and add them to the pool of valid factors. According to the general experience of other scholars and practitioners, this paper sets strict criteria for the selection of effective factors:

- The absolute value of the correlation between the factor and the return is greater than 0.7;
- The probability of winning the portfolio is greater than 60%;
- The relationship between the size of the factor and the return is divided into two types, one is the smaller the factor value, the larger the return; the other is the smaller the factor value, the smaller the return.

If all three conditions are met, the factor is considered as a valid factor.

IV. Establishment of stock selection model

A. Model validity test

According to the previous selection of valid factors and redundancy removal, we finally get 10 valid factors. We need to further consider the stock selection model comprehensively through the sample data (2010-2017), firstly, update each factor (the 10 valid factors that have been selected) of all the constituent stocks of CSI 500 index at the initial stage of model building, and weight the average of these 10 factor values with equal weights. If certain factor values are not available or are missing for the month, the factor is removed first and the remaining factors are equally weighted and averaged. The stocks in the stock pool are then ranked according to their final scores, and the top ranked stock portfolios are taken out. We divide all stocks in the pool into 10 equal parts according to their composite scores and calculate their return indicators as well as risk indicators, as shown in Table 1, Backtesting Results A.

TABLE I. Backtest results A

group	Annuity Return	Earnings Volatility	Maximum Retracement	Alpha	Beta	Sharpe Ratio	Information Ratio
1	39.33%	40.02%	35.93%	37.76%	1.04	1.07	2.41
2	27.69%	39.51%	37.58%	21.53%	1.02	0.83	1.55
3	20.91%	36.22%	39.37%	17.78%	1.06	0.52	1.29
4	19.23%	30.71%	40.59%	12.51%	1.04	0.30	1.06
5	15.60%	32.85%	42.93%	8.30%	1.07	0.23	0.96
6	9.51%	33.95%	47.61%	5.59%	1.04	0.12	0.82
7	7.22%	30.52%	45.38%	2.13%	1.08	0.10	0.61
8	30.45%	28.79%	51.47%	0.73%	0.96	0.01	0.21
9	2.74%	30.45%	52.70%	-1.52%	0.93	0.04	-0.38
10	1.18%	27.66%	52.22%	-2.44%	0.90	-0.18	-0.41

B. Introducing a model of hedging mechanism

For the one-way long trading mechanism of A-shares, we have no way to short stocks in the reverse direction in the stock market. In quantitative investment, investors usually use financing and financing securities or stock index futures for hedging. In this paper, we use CSI 500 stock index futures to hedge the risk of our investment portfolio. The fundamentals of CSI 500 stock index futures hedging risk are described below:First, we estimate the beta of the portfolio assets, which refers to the correlation between the risk of individual stocks and the degree of stock market risk. Our objective is to make the beta of the portfolio assets equal to 0. Calculate the required share of futures according to the following formula:

$$\beta_s(\frac{s}{s+NF})+\beta_f(\frac{NF}{s+NF})=0 \tag{1}$$

$$N=\frac{s*\beta_s}{F*\beta_f} \tag{2}$$

By calculating the required share of CSI 500 stock index futures for hedging, it is possible to use it to hedge the risk of the model. The results after hedging are shown in Table 2, Hedged Results A:

TABLE II. Result A after hedging

10-quartile group	Annuity Return	Earnings Volatility	Maximum retracement
1	33.56%	22.53%	19.34%
2	25.32%	18.69%	13.56%
3	18.83%	10.21%	11.02%
4	15.37%	9.02%	9.78%
5	12.71%	7.87%	8.81%
6	6.56%	8.40%	6.76%
7	4.21%	5.17%	6.09%
8	1.55%	4.02%	4.54%
9	-1.21%	7.01%	5.34%
10	-3.58%	6.38%	7.20%

Compared to the backtest results without hedging, the Annuity return is slightly lower and the maximum retracement and return volatility are substantially lower, which shows that hedging with stock index futures can reduce portfolio risk but at the expense of some portfolio return.

C. Changing the Factor Retrospective Period

This will be validated and explored below, and the factor lookback period is now set to run from 2015 to 2017, a total of three years. The selection of candidate factors, the selection criteria of valid factors and the process of redundancy removal of valid factors are exactly the same, except that the calculation of factor scores and stock portfolio returns are based on the data from 2015 to 2017. The specific process and data mining process are exactly the same as the previous empirical process, so we do not repeat the list here. The factor screening backtesting period is 2015-2017. From the comparison of each quantile group of the backtest results, the multi-factor stock selection model with a factor backtest period of 3 years is effective and the Annuity return differentiation is relatively high. A comparison with Table 1 shows that the maximum portfolio retracement is slightly lower but still large, and here it is still hedged by using CSI 500 stock index futures. The post-hedging results are shown in Table 3, Post-Hedging Results B:

TABLE III. RESULT B AFTER HEDGING

10-quartile group	Annuity Return	Earnings Volatility	Maximum retracement
1	38.14%	21.08%	16.96%
2	20.06%	12.46%	16.69%
3	13.63%	10.49%	21.08%
4	10.33%	9.08%	18.49%
5	7.46%	8.84%	18.98%
6	5.44%	6.47%	17.58%
7	2.81%	5.69%	18.71%
8	1.42%	4.12%	18. 15%
9	-l. 41%	5.15%	19.32%
10	-4.10%	8.42%	34.21%

Compared to before hedging, both the portfolio return volatility and maximum retracement are correspondingly much lower, indicating that the effect of hedging with CSI 500 stock index futures is as significant as before. After shortening the factor backwardation period, Hedged Result B has a slightly lower Annuity return and lower maximum retracement and return volatility compared to Hedged Result A. Therefore, the risk reduction is achieved at the expense of Annuity return.

D.Multi-factor stock selection using AdaBoost algorithm

Previously, we used the traditional method of factor ranking for multi-factor model construction, and now we use the AdaBoost algorithm in machine learning for factor selection. Here, we need to pay special attention to the fact that since the machine learning algorithm is sensitive to the input data, the required data range is (0,1), so we sort the stock data by factor and divide the sorted ranking by the total number of stocks, so that the factor values are normalized. Then, we rank the stock returns of the next period from the largest to the smallest, and use the top 30% as the strong portfolio, the bottom 30% as the weak portfolio, and the middle ones as the noise data to eliminate, with the strong stocks marked as +1 and the weak stocks marked as -1. In order to compare with the previous multi-factor model, find the stable and effective factors and ensure the stability of the algorithm, we use the panel data of the past 12 months to build the training set In order to compare with the previous multi-factor model and to ensure the stability of the algorithm, the training set is constructed using the past 12 months' panel data. Due to the limited code capacity, only the final backtest results are shown here, with the same backtest period from January 1, 2010 to December 31, 2017, with an initial capital of 1 million and the number of selected stocks of 10. The backtest results are shown in Table 4:

TABLE IV. MULTI-FACTOR STOCK SELECTION BACKTEST RESULTS BASED ON ADABOOST ALGORITHM

Strategy Benefits	Strategy Annuity Return	Benchmark earnings
551.75%	27.26%	39.36%
Alpha	Beta	Sharpe
0.23	0.686	0.97
Winning percentage	Profit/Loss ratio	Maximum backtest
0.623	2.421	25.31%

From the empirical results in the above table, it can be obtained that the multi-factor stock selection model based on the AdaBoost algorithm is significantly effective, and the strategy return reaches 551.75%, which is much higher than the strategy return of the traditional multi-factor model in the previous section.The Alpha and Beta values are also higher than the previous model, which explains the improved profitability of the model. The Sharpe ratio is 0.97, which is also higher than the previous model, but still not at the level required by the fund management company. The win/loss ratio and P/L ratio have also improved accordingly. A point of interest is the maximum retracement of 25.31%, which is lower compared to the traditional model. This achieves a reduction in risk with essentially little volatility in returns. The overall return-risk indicators of the model are both better than the traditional multi-factor model, indicating that the multi-factor stock selection model based on the AdaBoost algorithm is quite effective.

V. CONCLUSION

This study aims to construct a multi-factor stock selection model with controlled risk range and higher return, so the hedging mechanism of stock index futures is introduced to verify the validity of the model, and to explore the effective factors with stronger market sensitivity while controlling the risk. The stock selection performance with shortened factor lookback period is compared with historical data to test whether the validity of the factors is enhanced by the shortened factor lookback period with such a perspective. The results are also remarkable that we can indeed enhance the validity of the effective factors with market sensitivity by shortening the factor lookback period, and thus select the effective factors that are more compatible with the general market environment. Finally, based on the machine learning AdaBoost algorithm to construct a multi-factor stock selection model, the results of the model backtest are still excellent, and the stock selection effect and performance are improved compared with the traditional multi-factor stock selection model.

REFERENCES

[1] Luo, Z. N.. Research on multi-factor stock selection model with machine learning based on Stacking integration. China Price,2021(11):77-78+81.

[2] Luo, Z. N.. Research on quantitative stock selection strategy of Stacking based on integrated tree model. China Price,2021(02):81-84.

[3] Chen-Yang Liu. Research on stock selection based on K-means clustering and relative valuation method. Journal of Changchun University of Finance, 2021(01):34-42.

[4] Liu Jiaqi,Zhang Jian. A multi-factor stock selection model based on machine learning. Time Finance,2020(17):99-103.

[5] Ge Yua Mo,Zhou Xian. Multi-factor stock selection model based on XGBoost. Information Technology and Standardization,2020(05):36-41.

[6] Wu ZX, Zhang X, Zhang XEF. Research on the application of quantitative stock selection. Think Tank Times,2020(01):290-292.

[7] Wang L,Li L. A multi-factor quantitative stock selection strategy based on gcForest. Computer Engineering and Applications,2020,56(15):86-91.

[8] Le B,Cai ZJ,Hu WC. A multi-factor stock selection model for machine learning based on factor context. Mathematical Modeling and its Applications, 2019,8(04):10-19.

[9] Tang S. J.,Xiong X.,Xie M.,Ding L.,Zhang Shang. A multi-factor model for optimizing stocks based on machine learning. Information and Computer(Theory Edition),2019,31(23):30-32.

[10] Wang Z. Selection of quality stocks based on machine learning[J]. Electronic Production,2018(07):60-62.