WEP Warwick Evans Publishing

Transactions on Economics, Business and Management Research
ISSN: 2960-1789, eISSN: 2960-2254 | Volume 5
BEMS 2024

Research on Multi Factor Quantitative Investment Strategies Based on Machine Learning

Maoqing Yuan

Westa College, Southwest University, Chongqing, 400799, China

2553477510@qq.com

Abstract. Training models using machine learning algorithms have shown strong capabilities in classification prediction, thus employing machine learning algorithms in quantitative research can significantly enhance the accuracy of investment decisions. Utilizing a combination of multifactor models and machine learning algorithms for quantifying investment strategies holds considerable theoretical and practical research value. The IC analysis method and the random forest algorithm are employed to select factors, which serve as input variables (explanatory variables) for linear regression models. Future 30-day returns are used as the input variable (dependent variable) for regression calculations, establishing linear regression models to identify the relationship between them. Subsequently, this linear relationship is utilized for predicting future stock returns, thereby constructing a stock investment portfolio. The research findings indicate that both sets of strategy combinations exhibit some similarities and have achieved favorable strategy returns. Additionally, they demonstrate strong performance in terms of volatility and risk control, albeit with a deficiency in success rate.

Keywords: Machine Learning Algorithms; Quantitative Research; IC Analysis; Random Forest Algorithm.

1. Introduction

With the substantial increase in computational power and the abundance of data sources, machine learning algorithms have found widespread and profound applications in various fields, particularly in the realm of quantitative investing[1]. Some scholars define machine learning as a high-dimensional model primarily utilized for predictive tasks, emphasizing its essence as prediction. In the theory of multifactor models, it is posited that the excess returns of assets are driven by exposure to numerous factors, thereby explaining the excess returns of assets. Hence, the integration of machine learning with classical factor investing theory has become a popular research area in quantitative investing[2].

In the realm of multifactor model research, scholars have found that using a univariate linear mathematical model alone to construct the relationship between asset returns and factors is insufficient. Ross (1976) introduced the Arbitrage Pricing Theory (APT), departing from the assumptions of the CAPM theory, expanding the univariate linear model into a multivariate linear model[3]. It introduced multiple factors to jointly explain asset returns, including GDP growth rate and inflation rate factors, yet failed to provide a strong explanatory multifactor model to the academic community. It wasn't until Fama and French (1992) proposed the linear three-factor model that significant progress was made. Similar to the Arbitrage Pricing Model, the mathematical model is a multivariate linear model[4]. However, what distinguishes it is the determination of three factors that explain returns: market factor, company size factor, and company value factor. These three factors are used to model asset returns, making the Fama-French three-factor model the cornerstone of multifactor models. In their subsequent research, Fama and French (2015) further expanded the three-factor model to a five-factor model by introducing additional factors such as profitability factor and style factor to explain asset returns[5]. As market behavior continues to be explored, factors based on behavioral finance and technical aspects have gradually been incorporated into multifactor models. Stambaugh and Yuan (2017) added corporate management factors and stock performance factors on

top of market factors and company size factors, providing explanations for asset returns from the perspective of behavioral finance[6]. Liu et al. (2019) included the turnover rate factor into the multifactor model, constructing a four-factor model.

In this study, factors selected through IC analysis and random forest algorithm were utilized as input variables for a linear regression model. We established a model to investigate their relationship with future 30-day returns. Subsequently, we applied this model to predict future stock returns, thus forming an investment portfolio.

2. Factor Data Processing

2.1. IC Analysis Method

The IC analysis method, originating from Information Coefficient, is an indicator widely favored in the field of active management[7]. IC measures the predictive ability of forecasting variables and is typically defined as the cross-sectional correlation coefficient between the forecasted returns at time $t+1$ and the actual returns. In practical applications, the forecasted returns at time $t+1$ are often substituted by the forecasting variables at time t. Therefore, the definition of IC becomes the cross-sectional correlation coefficient between the forecasting variables at time t and the stock returns at time $t+1$.

$$IC = corr(z_{it}, R_{it+1}) \tag{1}$$

The data acquisition and processing in this study were conducted using the BigQuant artificial intelligence quantitative investment platform, which provides access to massive datasets from multiple markets including A-shares, US stocks, Hong Kong stocks, futures, options, etc. The platform supports mainstream AI frameworks and facilitates tasks such as data retrieval and cleaning. For this study, factor data and stock returns data from the Chinese A-share market were collected from January 1, 2015, to December 31, 2020, as the training set, while data from January 1, 2021, to November 30, 2021, were used as the test set for constructing and backtesting quantitative investment strategies[8].

2.2. Random Forest Algorithm

Similar to the bagging ensemble algorithm, the basic construction process of the Random Forest algorithm involves creating multiple initial decision trees. When creating each decision tree, not only is the bootstrap method used to randomly select samples, but feature variables are also randomly selected. This approach helps reduce the correlation between trees. However, the final classification information is confirmed by aggregating the results of votes from each tree. Random Forest, as a specific ensemble method, is particularly suitable for addressing problems with a large number of feature variables[9]. In contrast to other ensemble algorithm models where prominent feature variables may overshadow the effects of certain variables, Random Forest randomly selects feature variables for each decision tree. This enables effective detection of the behavior of each feature variable and its contribution. Therefore, compared to a single decision tree, Random Forest significantly improves prediction accuracy and reduces errors[10].

2.3. Candidate Factor Selection

According to the multi-factor model theory, the excess returns of stocks originate from different factors. Therefore, before constructing a multi-factor quantitative investment strategy, the selection of the candidate factor pool is a crucial step. In order to comprehensively explain stock returns, academia has conducted many years of research on this topic, starting from the three-factor model[11]. Both academia and industry have continuously proposed new multi-factor models while uncovering numerous factors that can be used to explain stock returns. The candidate factors in this study are based on the research in academia and industry, selecting a total of 64 indicator factors from financial, market, and expectation categories. The financial

factors include valuation, size, growth, and quality factors; market factors include risk, liquidity, technical, momentum, and fund flow factors. Valuation factors reflect the current valuation of stocks, indicating whether the company is worth investing in. Size factors reflect the current market capitalization of the company, showing the influence of market capitalization on returns. Growth factors are used to reflect the future growth and development potential of the company, while quality factors mainly reflect the degree of financial quality of the company. Risk factors mainly reflect the volatility of asset prices over a certain period of time; liquidity factors mainly reflect the strength of asset liquidity over a certain period of time; technical factors are a collection of various technical indicators; momentum factors reflect the price momentum of stocks over a certain period of time, allowing predictions of stock price trends based on the size of the momentum effect; fund flow factors reflect the market's enthusiasm for stocks. Some candidate factors are shown in Table 1.

Table 1. Candidate factor pool

index	detailed description
EP_TTM	Net profit TTM/ Total market value
EP_LYR	Net profit (Latest annual report)/Total market value
BP_LF	Net assets TTM/ Total market value
OCF_TTM	Operating cash flow TTM/ Total market value
SP_TTM	Revenue TTM/ Total market value
SP_LYR	Operating income (Latest Annual report)/Total market

2.4. Factor Validity Test based on IC Analysis Method

Based on sample data from January 1, 2015, to December 31, 2020, this article conducts factor effectiveness analysis using the IC method. Following the content of Section 2.2 on IC analysis, all factor values are preprocessed, including handling missing values, winsorizing, standardization, and industry market value neutralization. The preprocessed factors' rank correlation coefficients with stock returns lagged by 22 trading days are calculated as IC values. The IC series of these factors across different cross-sectional periods are obtained. Subsequently, various statistical indicators such as IC, IR, and IC standard deviation are computed for assessment.

The empirical research in this paper is based on the Bigquant quantitative investment platform. Partial empirical results are presented in Table 2.

Table 2. Candidate factor base IC analysis results

	IC standard deviation	IR value	The ratio of |IC|>0.02
EP_TTM	0.06	-0.55	80.67%
EP_LYR	0.06	-0.57	78.15%
BP_LF	0.11	-0.45	88.24%
OCF_TTM	0.06	0.58	87.39%
SP_TTM	0.1	-0.36	90.76%
SP_LYR	0.1	0.36	89.08%

According to the theory of IC analysis, factors with IC values greater than 2% are considered effective factors. The proportion of factors with IC greater than 2% and the IR value reflect the stability of the factors, which enables the analysis of factor effectiveness and the selection of factors required for quantitative investment models. In this paper, the initial factors will be filtered

as follows: factors with an average IC greater than 2% and a proportion of IC greater than 2% exceeding 80% will be selected as absolute rules, while the IR value will serve as a relative rule. For factors of the same type but different periods, only one relatively effective and stable factor will be retained. Although the size factor does not meet the filtering criteria, it is retained to encompass all types of factors. The calculation results for valuation factors, size factors, and growth factors are as follows.

Table 3. Factor calculation results based on IC analysis method

	IC standard deviation	IR value	The ratio of |IC|>0.02
EP_TTM	0.06	-0.55	80.67%
BP_LF	0.11	-0.45	88.24%
OCF_TTM	0.06	0.58	87.39%
SP_TTM	0.1	-0.36	90.76%
SP_LYR	0.1	0.36	89.08%
LN_MV	0.17	-0.04	90.76%
SALES_GR_TTM	0.06	0.4	80.67%

2.5. Factor Importance Ranking based on Random Forest Algorithm

This paper conducts factor importance analysis using the Random Forest machine learning algorithm based on sample data from January 1, 2015, to December 31, 2020. According to the theoretical content of the Random Forest machine learning algorithm, only missing value processing is performed on the data. Subsequently, the factor data is used as the feature input variable, and stock returns are categorized into two classes as the output variable. The Random Forest machine learning algorithm is then utilized for training and outputting factor importance. The parameters of the Random Forest machine learning algorithm are as follows:

Table 4. Random forest machine learning algorithm input parameters

Number of trees	Maximum depth of data	Minimum number of samples per leaf node	Parallelism degree	Random seed
10	30	200	1	0

The empirical research in this paper is based on the Bigquant quantitative investment platform. Partial empirical results are provided below, with detailed information presented in Figures 1, 2, and 3.

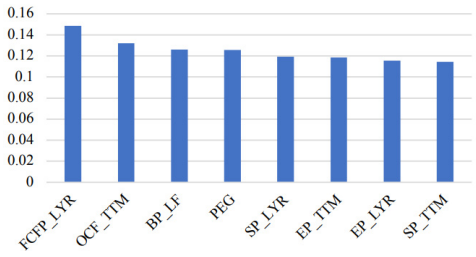

Fig 1. Ranking the importance of valuation factors

The higher the importance score of a factor, the more important it is for classifying stock returns, indicating its better explanatory power for stock returns. After obtaining the factor importance scores, this paper will select factors with relatively higher scores to contrast with those selected by the IC analysis method. This selection ensures consistency in quantity and type with the factors

identified through IC analysis. Therefore, partial results of the factors selected through the Random Forest machine learning algorithm are presented in the table below.

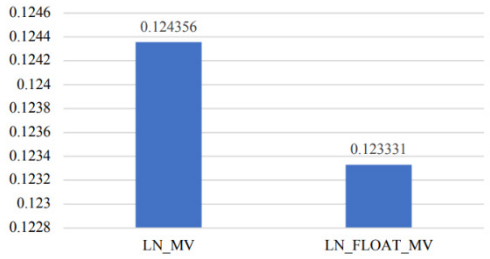

Fig 2. Scale class factor importance ranking

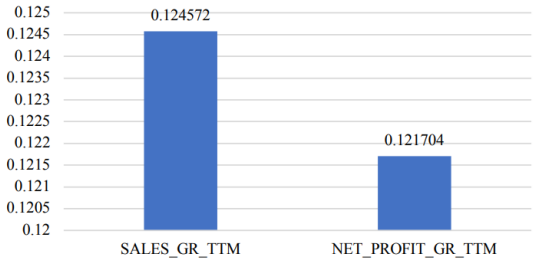

Fig 3. Ranking the importance of growth factors

Table 5. Comparison of factor screening results between IC analysis method and random forest algorithm

Financial factor	Importance score
FCFP_LYR	0.148612
OCF_TTM	0.131956
BP_LF	0.126081
PEG	0.125612
SP_LYR	0.119159

Table 6. Comparison of factor screening results between IC analysis method and random forest algorithm

Factor screening results of random forest algorithm	IC analysis factor screening results
FCFP_LYR	EP_TTM
OCF_TTM	BP_LF
BP_LF	OCF_TTM
SP_LYR	SP_LYR
LN_MV	LN_MV
SALES_GR_TTM	SALES_GR_TTM

Comparing the results obtained through the Random Forest machine learning algorithm and IC analysis, it is evident that there are significant differences in the factors selected by these two methods. For instance, in the financial factor category, valuation factors such as FCFP_LYR, OCF_TTM, and PEG are excluded from effective factors in IC analysis due to their relative

inefficacy. However, according to the importance ranking method of Random Forest machine learning algorithm, FCFP_LYR and OCF_TTM are considered relatively important factors within the valuation category. Furthermore, the results of factor selection differ completely between the two methods for quality factors, liquidity, and fund flow factors. Despite the substantial disparities in the results of factor selection between the two methods, identifying effective factors remains a crucial step in constructing quantitative investment strategy models for this paper's quantitative investment strategy.

3. Quantitative Strategy Construction based on Linear Regression Model

The basic idea of the model based on linear regression is to use the factors selected through IC analysis and the Random Forest algorithm as input variables (explanatory variables) for the linear regression model. The future 30-day returns are used as the output variable (dependent variable) for regression calculation to establish the linear regression model and identify the relationship between them. Subsequently, this linear relationship is utilized for predicting future stock returns, thereby constructing a stock investment portfolio. The specific procedure involves selecting factor data from January 1, 2015, to December 31, 2020, and calculated stock return data as in-sample data for training the linear regression model. Then, factor data from January 1, 2021, to November 30, 2021, and calculated stock return data are used as out-of-sample data for predicting future 30-day returns. The factor data serves as explanatory variables to forecast the future returns. Stocks with higher ranked returns are then selected for buy-and-hold operations to construct the investment portfolio. Relevant data on portfolio returns are obtained through simulated backtesting. The basic process of constructing a quantitative investment strategy model based on linear regression is outlined as follows.

Fig 4. Construct flow chart of quantitative investment strategy based on linear regression model

Data Acquisition: Retrieve stock returns data and factor values data for all stocks in the Chinese A-share market. Feature and Label Extraction: Calculate 23 predetermined factor values as feature variables or explanatory variables, and calculate the future 30-day returns of stocks as labels or dependent variables. Feature Variables Preprocessing: Perform winsorization and missing value handling. Strategy Backtesting Steps: Utilize data from January 1, 2015, to December 31, 2020, as in-sample data for model training. Use data from January 1, 2021, to November 30, 2021, as out-of-sample test data for predicting stock returns. Purchase the top 5 stocks ranked by predicted values daily, holding them for at least 30 days. Specifically, the higher the ranking, the more funds are allocated, with the maximum capital allocation not exceeding 3%. Initially, allocate funds evenly over the first 30 days, and thereafter, use remaining funds as much as possible.

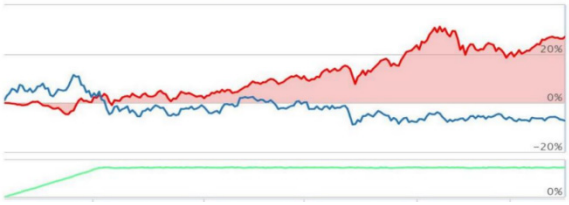

Fig 5. Strategy yield chart during backtest

Based on the Bigquant quantitative platform, the backtesting results are as follows: There are two sets of results in this backtest. One set of results uses the factors selected by IC analysis as input factors

for backtesting, while the second set uses the factors selected by the Random Forest machine learning algorithm as input factors for backtesting. The results of the first set of backtesting are shown in Fig 5..

The first set of results represents a quantitative investment strategy based on factor selection using IC analysis and prediction through a linear regression model. From the trend charts of strategy returns and benchmark returns, it is evident that the strategy outperforms the benchmark returns for the majority of the time. Moreover, during the backtesting period, the strategy achieves an absolute return of 27.1%, while the benchmark return is -7.28%, resulting in an excess return of 33%. The strategy's annualized return exceeds 30%, indicating excellent profitability. In terms of other indicators, stability is satisfactory, with a volatility of returns of only 16.34%. Additionally, for each unit of systematic risk undertaken, the strategy generates a return of 1.58 units, demonstrating exceptional risk-return control. Furthermore, the maximum drawdown of the strategy is only 9.67%, kept within 10%, which is a challenging target for most investors in stock investments. However, the strategy's success rate is only 0.52, implying that profitable trades occur only half of the time. Overall, the strategy portfolio proves to be effective and performs exceptionally well.

The second set of results represents a quantitative investment strategy based on factor selection using the Random Forest algorithm and prediction through a linear regression model. When considering this strategy independently, it achieves an absolute return of 25.83% and an excess return of 31%, far surpassing the benchmark return. Additionally, it demonstrates good stability, with a volatility of returns of 14.14% and a risk-return ratio of 1.72. Moreover, the maximum drawdown is only 6.57%, indicating strong risk control capabilities of the strategy portfolio. However, the success rate of this portfolio is also only 51%, but this does not significantly affect the profitability of the strategy.

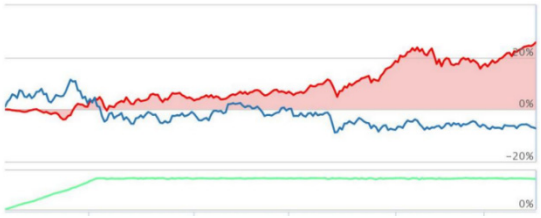

Fig 6. Investment strategy during the return chart

In summary, the results of the two sets of strategy portfolios exhibit certain similarities and both achieve favorable strategy returns. Additionally, they demonstrate strong performance in terms of volatility and risk control. However, they fall short in terms of success rate, with profitable trades occurring only half of the time. Nevertheless, as a medium to long-term holding strategy portfolio, the results indicate that this does not significantly impact the profitability of the strategy portfolio. Therefore, factor selection through IC analysis and the Random Forest algorithm, coupled with constructing quantitative investment strategies using linear regression models, proves to be effective and worthy of further practical implementation.

4. Conclusion

T The empirical analysis in this paper demonstrates that using IC analysis and the Random Forest machine learning algorithm for screening candidate factors, followed by constructing quantitative investment strategies based on linear regression models, can achieve higher excess returns. However, there is still much work to be done to further advance this research.

(1) Factor selection is the foundation of the quantitative investment strategy construction in this paper. Many factors were not included in the candidate factor library when established. Therefore, optimizing the factor library can be achieved by adding many relevant factors.

(2) The use of a fixed-length static training period for training and testing in this paper has certain limitations and complexities. Subsequent research can explore dynamic hyperparameter tuning and rolling forecasting techniques.

References

[1] Oh J S, Shong I, “A case study on business model innovations using Blockchain: focusing on financial institutions,” Asia Pacific Journal of Innovation & Entrepreneurship, vol. 11, no. 3, pp. 335-344, 2017.

[2] W. Luo, “Innovation and application of blockchain technology in the financial field,” Technoeconomics & Management Research, no. 8, pp. 90-95, 2018.

[3] Firdaus A, Razak M F A, and Feizollah A, et al, “The rise of “blockchain”: bibliometric analysis of blockchain study,” Scientometrics, vol. 120, no. 3, pp. 1289-1331, 2019.

[4] H. Cheng, and Y. Yang, “The development trend of block chain and commercial banks should study the policy,” Financial Regulation Research, no. 6, pp. 73-91, 2016.

[5] Schutz A, Fertig T, and Weber K, et al, “Vertrauen ist gut, Blockchain ist besser – Einsatzmöglichkeiten von Blockchain für Vertrauensprobleme im Crowdsourcing,” HMD Praxis der Wirtschaftsinformatik, vol. 55, no. 6, pp. 1155-1166, 2018.

[6] J. Zhang, “Construction and application of enterprise financial sharing service under cloud computing environment,” Friends of Accounting, vol. 597, no. 21, pp. 136-140, 2018.

[7] N. Li, and J. C. Mitchell, “RT: a Role-based Trust-management framework,” Proceedings DARPA Information Survivability Conference and Exposition, vol. 1, pp. 201-212, 2013.

[8] Ouaddah A, Abou Elkalam A, and Ait Ouahman A, “FairAccess: a new Blockchain-based access control framework for the Internet of Things,” Security and Communication Networks, vol. 9, no. 18, pp. 5943-5964, 2016.

[9] Aitzhan N Z, and Svetinovic D, “Security and Privacy in Decentralized Energy Tradingthrough Multi-signatures, Blockchain and Anonymous Messaging Streams,” IEEE Transactions on Dependable & Secure Computing, pp. 1-3, 2016.

[10] Wadud M, and Ali Ahmed H J, “Factors affecting delinquency of household credit in the U.D. Does consumer sentiment paly a role,” The North American Journal of Economis and Finance, no. 52, pp.101132-101134, 2021.

[11] Patil S, Nemade V, and Soni P K, “Predictive modelling for Credit card fraud detection using data analytics,” Procedia Computer Science, no. 132, pp. 385-395, 2021.