

Fundamental Quantitative investment research based on Machine learning

Jiao Xu^{1,*}

¹Soochow University, Economics Department, Soochow, China

Abstract. In recent years, the status of quantitative investment in China's capital market has been improving, and fundamental quantification has emerged as a promising approach that integrates fundamental analysis and quantitative investment successfully. Hence, this kind of intelligent quantitative investment method has garnered significant attention. In this paper, eight machine learning algorithms, including Lasso regression, ridge regression, partial least squares regression, elastic network regression, decision tree, random forest, support vector machine and K-nearest neighbor method, are used to construct the stock return prediction model. The empirical results show that linear machine learning algorithm outperforms nonlinear machine learning algorithm. The annual return rate of CSI 300 index in the same term is 1.47%, while the investment strategy based on OLS model has an annualized return rate of 35.96%, and the maximum withdrawal rate is only 29.61%, showing its strong return capacity. In this paper, machine learning is introduced in the field of fundamental quantitative investment, which provides investment reference for all kinds of investors and is helpful for the country to promote quantitative investment.

1.Introduction

With the development of economic and financial theory and computer technology, all kinds of new technologies have been tried and applied in the field of quantitative investment, constantly driving the update and evolution of quantitative investment methods. As a product of deep integration of fundamental analysis and quantitative investment, fundamental quantification is an intelligent quantitative investment method that has attracted much attention in recent years [1]. It combines specific company characteristics that are indicative of excess returns with the conventional quantitative investment model, synthesizing the primary benefits of fundamental analysis, technical analysis and quantitative investment. In the industry, it starts with some special style index investments [2]. In academia, it can be traced back to the three-factor model of Fama & French [3].

At present, the main carrier of fundamental quantitative investment is multi-factor model. Foreign scholars have conducted extensive empirical studies on factor selection, stock selection method, data selection and portfolio construction, among which a large number of literatures used machine learning method to predict stock returns. K.im et al. [4] studied a technique based on genetic algorithm for feature discretization and for optimizing the weight of artificial neural network (ANN) to predict the change trend of stock index. Geoffrey Hinton et al. [5] proposed the deep learning theory, arguing that the neural network model with multiple hidden layers had stronger learning ability, which could be considered as a sign to enter the stage of deep

learning research. Albadvi and Chaharsooghi [6] used German market data to study factors related to fundamentals, and found that different industries had different effective factors, and that specific factors have varying degrees of impact on the rate of return. Buehlmaier and Zechner [7] studied text information based on machine learning and discussed the influence of text information on stock price or trading volume. Kozak et al. [8] used PCA method to extract common factors in factors and found that the model based on a few principal components could independently predict returns.

Due to the obvious style switch in the A-share market, traditional multi-factor models have met great challenges in the validity and stability of prediction. Compared with foreign countries, domestic research on quantitative investment in machine learning started relatively late. S Tang, X Xiong, M Xie [9] combined the optimization random forest algorithm with quantitative investment strategy and took the A-share market as the empirical object to show that the strategy can be effectively applied to market style investment. B Li [10] introduced machine learning into fundamental quantitative investment and adopted 12 machine learning algorithms, such as predictive combination algorithm and Lasso regression algorithm, to build stock return prediction model and portfolio. The results proved that machine learning algorithm could effectively identify the complex correlation between anomaly factors and Alpha returns to some extent. Q Chen & X Gong [11] built a multi-factor model based on decision tree, and the back-test results showed that the returns of the multi-factor model

* Corresponding author: 1808774953@qq.com

improved by decision tree were better than that of the traditional scoring model. On the whole, the integration of fundamental quantitative investment with machine learning method is advancing more rapidly than traditional quantitative investment and holds promising development prospects and research value.

2. Research method

2.1 Model design

The overall research framework of quantitative fundamentals based on machine learning is shown in Figure 1.

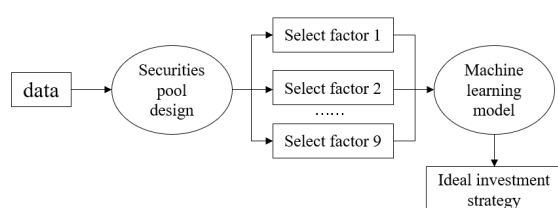


Fig.1. Fundamental quantitative research framework based on machine learning

In order to ensure the effectiveness of calculation and feasibility of investment, sliding window method is used to divide training and test data sets. The steps of model training and testing are as follows: Step1: Suppose it is in early January 2011, and the model will determine the portfolio of that month. In this paper, the factor data of the past 12 months (2010) was used as the training set to fit the machine learning model and get the model parameters. Step 2: The trained model is applied to the factor data in December 2010 to get the model's prediction of stock returns in January 2011. Step 3: To the beginning of February, this paper repeats Steps 1-2 and so on until the end of the data period. Figure 2 shows the sliding window method.

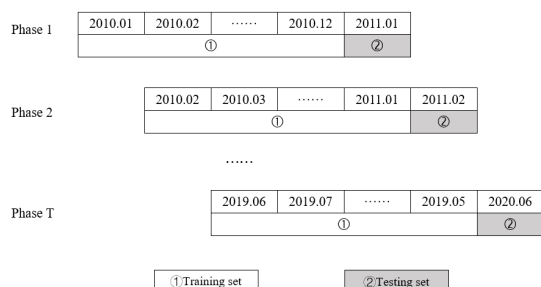


Fig.2. Sliding window method

The sliding window method is consistent with the decision-making process of investment activities in reality, and is better than the common classification methods of training set and test set (such as cross verification). Consistent with the monthly transaction frequency, the length of the test set is controlled within

one month. By fixing the training set in a fixed interval, the training time of the model can be reduced.

In order to verify the stability of the model, this paper further explores how much factor data of the past months can be used as the training set to make the machine learning model predict higher stock returns. All machine learning model parameters were selected by means of BayesSearchCV. First, the initial parameter pool was established. Secondly, each parameter was used to train the data on the training set to get the return of the portfolio, and then the optimal parameter was selected. Finally, the optimal parameters were applied to the test set to get the final investment performance. In this study, due to the high computational cost of random grid search in each phase, the optimal parameters retrieved from the first sliding window training are employed and remain constant across different time periods.

In order to evaluate the performance of the model, this paper constructs the annualized return rate, maximum retracement rate, annualized volatility, annualized return rate/maximum retracement rate, Sharpe ratio and annualized Sharpe ratio of each portfolio according to the model prediction as performance measurement indicators.

2.2 Machine learning algorithm

Machine learning is a collection of many forms of prediction functions and their various algorithms. In this paper, the traditional linear regression model is chosen as the benchmark, and eight representative machine learning algorithms are selected. Four linear machine learning models are considered in the essay, including Ridge Regression (Ridge), Lasso Regression (Lasso), Partial Least Square (PLS) and Elastic Net Regression (Elastic Net). At the same time, this paper selects four machine learning algorithms, consisting of both single and integrated learning model. Single models include Decision Tree (DT), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN), and the integrated learning model includes Random Forest (RF).

The algorithms selected in this paper are not a complete set of machine learning regression algorithms, but several representative ones have achieved good prediction performance in some other fields. The existing machine learning algorithm is used to realize the fundamental quantitative investment, so that the existing algorithm package is adopted. All data processing and machine learning implementations are based on Python language, and the machine learning algorithm is mainly implemented based on the scikit-learn algorithm package.

2.3 Data selection and pre-processing

In this paper, all listed companies in China's A-share market from January 2010 to June 2020 are selected as research samples, and the data is monthly frequency which is selected to be consistent with relevant studies (Y Hu, M Gu 2018) [12]. The data are all from Choice financial terminal. Table 1 shows the factor names and their meanings.

Table 1. Factor abbreviation and meaning

Factor	Meaning
Turn	Turnover rate
Amplitude	Amplitude
MV	Total market value
Equityadjliqmv	Floating market value
EV2	Enterprise value
Totalshare	Total capital stock
Stdevr100	Annualized volatility
TTM	Cash flow
Dividendyield	Dividend rate

The initial stock pool is for all stocks in the A-share market. ST stocks and ST* stocks are excluded to avoid their delisting risk in the market due to performance losses which may affect the research results. Additionally, stocks with abnormal trading status are also excluded to prevent any influence on the research results. To ensure the accuracy of the study, stocks listed for less than a year have been excluded to avoid any abnormal fluctuations caused by the IPO underpricing effect. Financial stocks have also been omitted in this paper due to their distinct measurement indicators compared to other listed companies.

Since there is a certain proportion of missing values in the database, the processing methods adopted can be mainly divided into two categories. If the earnings data of a certain stock is missing in month t , all the data of the stock in month t will be deleted. If a factor value of a stock is missing, it is filled with 0.

Because there are extreme values in some factors, in order to exclude the interference of extreme values and improve the accuracy of prediction, MAD method (mean absolute deviation) is applied to de-extremum of the training set data. The formula for de-extremum is:

$$\text{MAD}(X) = \text{median}(|X_i - \text{median}(X)|) \quad (1)$$

Since the values of different factors are significantly different in order of magnitude and distribution, which may lead to the deviation of prediction, this paper standardizes the training set data, making the processed data a sequence distributed within 0~1, so as to ensure that each factor index has a uniform dimension. The standardized formula is:

$$X_{scale} = \frac{X - \bar{X}}{\sigma_X} \quad (2)$$

$\bar{X}$ and σ_X are the mean and standard deviation of variable X respectively.

Industry neutralization can remove the alpha bias generated by the industry, so that different industries have the same or no impact on alpha. In the essay, the training set data is neutralized in the industry. This is achieved by subtracting the original value from the mean value and dividing the result by the standard deviation. By doing so, the new distribution mean value is centered at 0, and values larger than the original value become 1, indicating their distance from the mean in terms of one standard deviation.

3 Empirical results and analysis

3.1 Performance of changing the sliding window

Assuming that PCA method (dimensionality reduction is 3) is used, factor data of different months in the past are taken as the training set, and stock returns are shown in Table 2.

Table 2. Training set length and model performance results.

Sliding window (month)	rate of return (%)	Maximum drawdown (%)	volatility (%)	Sharpe ratio
4	19.94	39.35	28.94	0.1930
5	19.68	37.23	26.83	0.1995
6	19.54	32.34	26.22	0.2013
7	21.13	33.29	27.90	0.2072
8	23.85	36.06	28.87	0.2251
9	24.13	33.41	29.85	0.2221
10	24.03	33.41	28.97	0.2259
15	22.22	33.30	28.45	0.2131
19	26.42	31.05	29.13	0.2438
20	28.20	34.31	29.50	0.2555

According to the results of the model, when the training set length is 20 months, the annual return rate is the highest and the maximum pullback rate is relatively low, that is, the risk control ability is good, so the length of the training set selected in this paper is 20 months.

3.2 Performance of changing the PCA reduced dimension

Since the data comprises multiple dimensions, it is imperative to process the features to extract vital information. Constantly, Principal Component Analysis (PCA) is used to reduce the dimensions of the data. The paper compared the performance obtained with and without PCA, while varying PCA dimension reduction, which were shown in Table 3.

Table 3. PCA dimension and its model performance results.

PCA dimension	rate of return (%)	Maximum drawdown (%)	volatility (%)	Sharpe ratio
No dimension reduction	35.96	29.61	29.84	0.3106
1	9.51	44.91	30.98	0.1000
3	28.20	34.31	29.50	0.2555
5	29.66	34.83	28.15	0.2757

According to the performance results of the model, it can be found that the annualized return rate and sharpe ratio without PCA dimension reduction are much higher than those without PCA dimension reduction, and the maximum retracement rate without PCA dimension reduction is also the lowest. Therefore, PCA is not used in this paper.

3.3 Performance of fundamental quantitative investment model based on machine learning in A share market

This paper examines the empirical performance of the fundamental quantitative investment model based on machine learning in the A-share market. Table 4 shows the risk-benefit scenarios of eight machine learning methods in a 20-month sliding window.

Table 4. Performance of fundamental quantitative investment model based on machine learning in A share market.

Machine learning algorithm	rate of return (%)	Maximum drawdown (%)	volatility (%)	Sharpe ratio
000300.SH	1.47	40.56	23.51	0.0143
OLS	35.96	29.61	29.84	0.3106
Lasso	32.80	28.93	29.62	0.2890
Ridge	35.96	29.61	29.84	0.3106
PLS	32.58	29.72	30.07	0.2842
Elastic	34.16	29.83	29.81	0.2974
DT	16.02	39.94	29.24	0.1590
RF	28.12	31.24	33.46	0.2346
SVM	35.55	30.46	31.33	0.2965
KNN	16.93	50.77	32.64	0.1580

According to the results of the model, it can be found that the results of linear machine learning algorithm are better than those of other machine learning algorithms. Among them, the investment strategy based on OLS model has the highest return capacity, with the annual return rate reaching 35.96%, while the maximum retreat rate just reaching 29.61%. The reason may be that the number of factors is not enough, and the linear relationship between factors is obvious, while the machine learning model is more complex, better at showing the nonlinear relationship between factors. Although the traditional linear model is simple, it is still a comprehensive model, which can automatically find the relatively accurate relationship between the predicted object and the influencing factors. That is why the traditional linear regression model still has a high frequency of application in quantitative investment.

4 Conclusion and suggestion

4.1 Research conclusion

In this paper, all the listed companies in China's A-share market from January 2010 to June 2020 are selected as research samples. Based on fundamental analysis, machine learning algorithms are introduced to build eight kinds of machine learning-based algorithms, and the empirical performance of machine learning-based models and linear regression models in China is systematically compared. The empirical findings demonstrate that the traditional linear regression model exhibits superior performance when compared to other algorithms, while the linear machine learning algorithm proves to be more effective overall than its nonlinear counterpart.

This result shows that there is a strong linear relationship between the selected fundamental factors, and the traditional linear regression model can recognize the linear pattern in order to obtain better prediction effect and portfolio returns.

4.2 Policy suggestion

This paper studies quantitative investment by machine learning and provides investment reference for various investors. For retail investors, quantitative investment decisions are made entirely on the basis of objective data, avoiding the impact of individual subjective emotions on trading behaviors. For institutional investors, stock selection strategies based on machine learning are provided.

The findings of this study shed light on the promotion of quantitative investment and suggest that financial authorities should gradually ease some restrictions on quantitative investment practices. Quantitative investment is a technical tool to help capital operate efficiently. It can be applied in a wide range of fields with a good prospect. However, prevailing limitations on the growth and decline of the A stock market and hedging mechanisms hinder its full potential. Therefore, it is recommended that China incrementally remove current constraints on quantitative investment in a systematic manner, thereby improving the allocation of financial market resources.

References

1. X. Hou, B. Wang, Based on the fundamental analysis of quantitative investment: the review and prospect [J]. Journal of northeast normal university , No. 309 (01) : 124-131 + 141. (2021)
2. C. Shi, Y. Liu, X. Lian, Factor investment: Method and Practice [M]. Publishing House of Electronics Industry, 2020. Luigi T. De Luca, *Propulsion physics* (EDP Sciences, Les Ulis, 2009)
3. E.F. Fama, K.R. French, The cross - section of expected stock returns[J]. the Journal of Finance, 47(2): 427-465.(1992)
4. K. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for prediction of stock index, Expert Syst. Appl.(19):125- 132.(2000)
5. G.E. Hinton, S. Osindero, Y. Teh, A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 18, 1527-1554. (2006)
6. A. Albadvi, S.K. Chaharsooghi, Risk:Reward Views:Unlocking the Full Potential of Fundamental Analysis[J]. Morgan stanley Global Research, 2013(12):216-227.(2013)
7. M. Buehlmaier, J. Zechner, Slow-Moving Real Information in Merger Arbitrage[J]. Social Science Electronic Publishing. (2012)
8. S. Kozak, S. Nagel, S. Santosh, Shrinking the Cross Section[J]. CEPR Discussion Papers. (2017)

9. S. Tang, X. Xiong, M. Xie, Optimization Stock Multi-factor Model based on Machine Learning [J]. Information and Computer, 31(23):30-32.(2019)
10. B. Li, X. Shao, Y. Li, Machine learning to drive the fundamentals of quantitative investment research [J]. China industrial economy, No. 377 (8) : 61-79. (2019)
11. Q. Chen, X. Gong, Research on Stock Multi-factor Optimization Model Strategy Based on Decision Tree [J]. Information Technology and Informatization, No.238(01):209-211.(2020)
12. Y. Hu, M. Gu, Buffett's Alpha: An Empirical Study from China's stock Market [J]. Management world, 2018 (8) : 41-54 + 191. (2018)