Scientific Research Publishing

Journal of Mathematical Finance, 2018, 8, 111-118
http://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

Fundamental Factor Models Using Machine Learning

Seisuke Sugitomo, Shotaro Minami

[1]Epic Partners Investment Co., Ltd., Tokyo, Japan
[2]Asuka Asset Management Co., Ltd., Tokyo, Japan
Email: sugitomo@epicgroup.jp, sminami@asuka-asset.com

How to cite this paper: Sugitomo, S. and Minami, S. (2018) Fundamental Factor Models Using Machine Learning. *Journal of Mathematical Finance*, **8**, 111-118.
https://doi.org/10.4236/jmf.2018.81009

Received: January 15, 2018
Accepted: February 9, 2018
Published: February 12, 2018

Abstract

Fundamental factor models are one of the important methods for the quantitative active investors (Quants), so many investors and researchers use fundamental factor models in their work. But often we come up against the problem that highly effective factors do not aid in our portfolio performance. We think one of the reasons that why the traditional method is based on multiple linear regression. Therefore, in this paper, we tried to apply our machine learning methods to fundamental factor models as the return model. The results show that applying machine learning methods yields good portfolio performance and effectiveness more than the traditional methods.

Keywords

Multi-Factor Model, Fundamental Factor Model, Support Vector Machine (SVM), Gradient Boosting Decision Tree (GBDT), Neural Network (NN), Artificial Intelligence Finance

1. Introduction

A typical tool of quantitative active operation (Quantz) has a multifactor model. This explains return on investment of stock with multiple factors. A general multifactor model in finance field is synonymously used with Arbitrage Pricing Theory (APT) proposed by Stephen Ross in 1976 [1], but in operation practice multifactor models such as CAPM based BARRA-type approach and Fama-French-type approach [2] [3] [4] are also widely used. The method obtaining return on equity of individual company by giving macroeconomic variable a priori, and a method that derives factor by factor analysis from the past return on equity are classified as APT-type multifactor model. On the other hand, the methods of obtaining return on equity of individual companies by using brand

attribute having the stocks of individual companies such as investment indices represented by PER and PBR are classified as BARRA-type or Fama-French-type multi-factor model. This paper is classified as multi-factor model of Fama-French type.

There is another matter that should be clarified at the time of using multifactor model.

Normally, there are two methods to use the multifactor model that explains the return on equity of individual companies in the stock attribute. In the first method, trends in market price are judged from the contribution ratio of stock attribute. This method (return model) is used for calculating the future return on equity. In second method (risk model), contribution ratio of brand attribute is regressed by market return in chronological order, fundamental beta is obtained, and portfolio attributes are analyzed [5]. In this paper, return method is assumed.

Above are the prerequisites that should be kept in mind while doing analysis. These cases are discussed in finance domain as well with confusions. The purpose of this paper is to explore the applicability of the machine learning method in the quantitative active operation, and to clarify the premise that it contributes to the development in the field of the future quantitative active operation.

The multifactor model (hereinafter referred to as fundamental factor model) mentioned in this paper is defined below.

$$R_{it} = \sum_{j=1}^{k} X_{ijt} f_{it} + \epsilon_{it} \tag{1}$$

However, R_{it} means return on equity of company i in t period, X_{ijt} is factor exposure of j factor of company i in t period, f_{jt} is factor return of j factor in t period and ϵ_{it} is an error that cannot be explained in factor.

It is a model that calculates future return on equity of individual companies by multiple regression analysis based on multiple brand attributes (factors). In this model, the relationship between return on equity and factor is linear, but considering the complexity of the financial market, relationship can be expressed more appropriate by assuming nonlinearity. In this paper, we use a typical method of machine learning that can express a nonlinear relation (support vector machine, gradient boosting, neural network) and verify the effectiveness and applicability of nonlinear methods in practical operation by comparing it with conventional linear models.

2. Related Research and Basic Concepts

2.1. Related Research

The multifactor models, which are the basis of the analysis, are classified as BARRA-type or Fama-French-type multi factor model. Bar Rosenberg has introduced Barra-type approach and Grinold and Kahn (1999), Conner *et al.* (2010) have expanded it [6] [7]. It is calculated based on cross-sectional regression analysis as it is assumed that the return on equity of traded stock at a cer-

tain point is explained by common factor. Fama-French type was first introduced by Eugene Fama and Kenneth French (1992).

In case of Joseph [8], VIX, monthly change rate of VIX, PBR distribution, change rate of PBR distribution, factor return of PBR a month ago are considered as variable for estimating PBR factor return and verification using logistics regression analysis using the shrinkage method is carried out for parameter estimation. Similar analysis is carried out for price momentum as well.

Along with this, we have obtained the result that the forecast accuracy of the next period is significantly higher. As other machine learning methods, the Classification and Regression tree (CART) is used for verification, and it is seen that it is more effective than logistic regression analysis depending on the period.

The above analysis is attempted to apply a nonlinear machine learning method using other variables in the time series forecast of factor return, but it is not applied to the return model. In return model, future predicted return is calculated based on predicted value (called as factor weight) of return factor and latest factor exposure. This is called as predictive alpha. The above research differs from our research as multifactor of predictive model is not referred to.

The basic concept of the machine learning analysis method used in the analysis is described below [9].

2.2. Support Vector Regression

Support vector regression is a method of nonlinear regression, which performs linear regression in the feature space, considering a nonlinear mapping to the feature space of the explanatory variables. ϵ-SV R used in this research estimates the linear function $y = f(x) = \langle w \cdot x \rangle + b$ by using ϵ-insentive loss function $L^{\epsilon}(x, y, f)$ at the time of linear regression.

$$L^{\epsilon}(x, y, f) = \max\left(0, \left|y - f(x)\right| - \epsilon\right) \tag{2}$$

If predicted value exceeds actual measurement value, it is expressed as $L^{\epsilon}(x, y, f) = \xi$ and if predicted value is less than actual measurement value, it is expressed as $L^{\epsilon}(x, y, f) = \hat{\xi}$. Ultimately it will solve the following main problem.

$$\min_{w, L^{\epsilon}, b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{i} L^{\epsilon}(x_i, y_i, f) \tag{3}$$

$$s.t.\ \left(\langle w \cdot x_i \rangle + b\right) - y_i \le \epsilon + \xi_i \tag{4}$$

$$y_i - \left(\langle w \cdot x_i \rangle + b\right) \le \epsilon + \hat{\xi}_i \tag{5}$$

$$\xi_i, \hat{\xi}_i \ge 0, i = 1, 2, \cdots, l \tag{6}$$

In this case, it can be extended to nonlinear regression by mapping the original data x in the above expression to high-dimensional space by a nonlinear function $\phi(x)$. The above main problem can actually be solved as a dual problem using the dual theorem. In this case, inner product calculation appears as $\langle \phi(x_i) \cdot \phi(x_j) \rangle$ in the dual problem. In general, it is very complicated to directly

perform inner product calculation in a high dimensional space. A kernel function $K\left(x_i, x_j\right)$ that can perform this inner product calculation is applied. In this paper, Gaussian Kernel is used as Kernel function.

2.3. Gradient Boosting Decision Tree

In case of decision tree, decision is made by identifying the possible options and potential scenarios in the form of tree diagram and by comparing the expected value of each option. It is a method commonly used in the financial industry and the consulting industry. In this research, we combine the gradient boosting method which is one of ensemble learning, and a decision tree to build a more refined model. Gradient boosting is a method of repeating the recovery and extraction from the learning data, creating multiple datasets, making a weak learner for each, and seeking the final solution to take majority decision by all weak learner solutions. At the time of creating weak learner, use the result of previously created weak learner and update it so as to increase the misclassified values. In this weighting, the gradient descent method is used for gradient boosting. In this research, Gradient Boosting Decision Tree (GBDT), which uses a decision tree for this weak learner, is used as a model.

2.4. Neural Network

There are two types of neural network that is hierarchical neural network and non-hierarchical neural network, but the method used this time is a hierarchical neural network. A hierarchical neural network is a network having an input layer, an intermediate layer, and an output layer. Explanatory variables are taken in the input layer, randomly assign weights to these explanatory variables in the intermediate layer, and calculate the optimum weight so that the result gets closer to the target variable in the output layer. It is called a neural network because it is similar to the neuronal cell in the human brain which receives signal from lot of other neuronal cells, and makes the decision.

2.5. Implementation

It is based on stock analysis system which we have built. Python is used for the basic database and calculation system and fundamental factor model, and R and "nnet", which is a machine learning package, are used for machine learning calculation [10].

3. Verification

3.1. Concept of Verification

Normally, it is necessary to verify the effectiveness and stability of the factor return for deciding whether the future return on equity calculated by the fundamental factor model can be utilized in actual operation. However, it is difficult to specify the coefficient corresponding to the factor return for the machine learning method (support vector machine, gradient boosting, neural network).

Therefore, the following verification is carried out. First, sort the future return on equity required as a result of applying the machine learning method, in descending order of individual stocks, and divide it into a group of five quantiles. Consider the largest group as long portfolio, smallest group as short portfolio and measure the portfolio for the next period respectively. Calculate the difference (spread/return) between long portfolio and short portfolio. Repeat calculations over the analysis period and compare the results of the conventional linear model and the machine learning method. If the predictive capability of the future rate of return is higher, the difference between the realized future return on equity and the projected rate of return will be smaller, and the cumulative return of the long and short portfolio must be larger. For verification, RMSE and MAE are also calculated using realized values and predicted values.

3.2. Verification Procedure

Universe is TOPIX500 constitutive brand which is the top 500 stocks with high market capitalization and liquidity of the TOPIX adopted stocks. Regarding the factor to be used, five commonly used investment indices such as PER, PBR, ROE, logarithmic market capitalization and 3 months β are considered. PER is the ratio for valuing a company that measures its current share price relative to its per-share earnings. PBR is the ratio used to compare a stock's market value to its book value. It is calculated by dividing the current closing price of the stock by the latest quarter's book value per share. ROE is the amount of net income returned as a percentage of shareholders equity. Market capitalization is a variable representing company size. Logarithmic transformation is performed so that the distribution is close to the normal distribution. β is the measure of the volatility, or systematic risk, of a security or a portfolio in comparison to the market as a whole. This is calculated by dividing the covariance the stock's returns and the benchmark's returns by the variance of the benchmark's returns over a specified period. Considering the settlement period, 3 months is selected for the calculation period for β. However, PER and PBR are converted to reciprocal numbers. The analysis period is from the end of January 2000 until June 2017. Regarding the portfolio, return is measured based on monthly rebalancing. In addition, to eliminate the influence due to the difference in the level of the factor value by the industry belonging to the individual stock, first performed the standardization in TSE 33 industry (Z-score), and then, again performed the overall standardization in TOPIX500 [11].

For the multiple regression model which is the standard for comparison, regression coefficient (factor return) is calculated by using the monthly factor value of each stock of the past 1 year (before the t-phase) as the explanatory variable and by using the end of the next month (t-phase) return as the target variable. We have calculated the future stock return rate (expected return) of each stock by multiplying the resultant regression coefficient (factor return) with the factor value (factor weight) of each stock at the end of the test period. According

to the above idea of the verification, we considered first quantile of the five quantiles portfolio made on the basis of the resultant expected return as long portfolio and fifth quantile as short portfolio, and calculated the difference in return (spread return) of the long portfolio and the short portfolio when held until the next month. We rolled this on a monthly basis and accumulated the obtained returns. The process of the analysis is similar for the machine learning method (support vector machine, gradient boosting, and neural network) which is the target for comparison. We used machine learning method when calculating the expected return.

Two patterns of comparison and prediction accuracy of portfolio performance are shown to verify effectiveness. In addition, for portfolio performance, monthly average return, monthly standard deviation, and sharp ratio are shown. The prediction accuracy is RMSE and MAE accumulated over the calculation period.

3.3. Verification Result (Table 1)

In comparison of portfolio performance, the monthly average return was the highest in neural networks model. The same result was also obtained for the sharp ratio considering the volatility. In comparison of prediction accuracy, the cumulative values of RMSE and MAE, which were obtained from the actual and predicted values, was the best result for the SVM. Regarding GBDT, though the performance of portfolio was inferior to the multiple regression analysis, the cumulative RMSE and cumulative MAE results were highly accurate.

4. Conclusions

In this study, we focused on the fundamental factor model; applied GBDT, SVM, and neural networks in addition to the conventional multiple regression analysis; and compared the accuracy of return prediction. As a result, it was observed that the cumulative RMSE, cumulative MAE, which is the applicable accuracy, improved in all nonlinear models, and improvement was also observed in some models in monthly average returns and monthly sharp ratio. This implies that the relationship between the return of the stocks in the financial market and the factor value is not a conventional linear relationship, but it is a nonlinear relationship, and a model that can capture such a nonlinear relationship is considered to be superior to the conventional model.

Table 1. Results.

	Multiple regression analysis	GBDT	SVM	NN
Average return	0.104	0.097	0.122	0.185
Standard deviation	2.839	2.862	2.892	2.728
Sharp ratio	0.037	0.034	0.042	0.068
Cumulative RMSE	1832.178	1831.618	1823.537	1832.920
Cumulative MAE	1428.715	1428.144	1418.282	1428.392

As a future perspective of this research, nonlinear analysis is also important in actual operation. For example, it is observed that the tilt factor is effective in case of large fund or a fund called smart beta that is bet on a factor, but the fund performance is often not in agreement. This is because the conventional multi-factor model is a linear model. There is a possibility of deviation due to linear evaluation of what should be evaluated essentially in non-linear form. Also, application to cross section regression analysis (BARRA type) is expected. Alternatively, profound results may be obtained by application to multivariate regression analysis considering both cross section and time series regression analysis. It is not easy to apply the machine learning method by controlling basic analytical method and ideas because the field of active management of measurement itself is a subject of deep research. However, in addition to the idea that is widely used in the actual operation, considering that the compatibility of the active management of measurement and the machine learning method is also a good aspect, application from all angles is required in the future. We would like to explore it in future.

As the limitations of research, there are limitations to prediction since it is not possible to learn features that are not found in the past data. There is also a limitation that it is difficult to decompose the return contribution degree.

Acknowledgements

This paper does not represent official views of Epic Partners Investments Co., Ltd. and Asuka Asset Management Co., Ltd. to which the authors belong. Everything is the personal opinion.

References

[1] Ross, S.A. (1976) The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory*, **13**, 341-360. https://doi.org/10.1016/0022-0531(76)90046-6

[2] Rosenberg, B. and Rudd, A. (1982) Factor Related and Specific Returns of Common Stocks: Serial Correlation and Market Efficiency. *Journal of Finance*, **37**, 543-554. https://doi.org/10.1111/j.1540-6261.1982.tb03575.x

[3] Fama, E.F. and French, K.R. (1992) The Cross-Section of Expected Stock Returns. *Journal of Finance*, **47**, 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

[4] Fama, E.F. and French, K.R. (1988) Permanent and Temporary Components of Stock Prices. *Journal of Political Economy*, **96**, 27-36. https://doi.org/10.1086/261535

[5] Chincarini, L.B. and Kim, D. (2006) Quantitative Equity Portfolio Management: An Active Approach to Portfolio Construction and Management. McGraw-Hill Library of Investment and Finance.

[6] Grinold, R.C. and Kahn, R.N. (1999) Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Selecting Superior Returns and Controlling Risk. McGraw-Hill Library of Investment and Finance.

[7] Zivot, E. (2011) Factor Models for Asset Returns. University of Washington, Seattle, Washington.

[8] Mezrich, J. (2014) Factor Forecasting with Machine Learning. Nomura Equity Re-

search Report.

[9] Minami, S. and Mitsusada, Y. (2017) Possibility and Limitations of Application of AI to Asset Management Business. *Securities Analyst Journal*, **55**, 16-26.

[10] Arratia, A. (2014) Computational Finance: An Introductory Course with R. Atlantis Press, Atlantis. https://doi.org/10.2991/978-94-6239-070-6

[11] Takaaki, Y. (2013) Introduction to Quantitative Analysis for Stock Investment. Nihon Keizai Shimbun, Inc., Tokyo.