Dynamic Weighting Multi Factor Stock Selection Strategy Based on XGboost Machine Learning Algorithm

Li Jidong
College of Finance
Hebei University of Economics & Business
Shijiazhuang, China
Lijidong1973@126.com

Zhang Ran
Institute of Economics
Hebei University of Economics & Business
Shijiazhuang, China
Zrj6401@126.com

Abstract—**Tree boosting is a highly effective and widely used machine learning method. A dynamic weighting multi-factor stock selection strategy based on XGBoost model is constructed. XGboost machine learning method is used to predict the IC coefficients of factors. The results of back testing show that the performance of dynamic weighting strategy is superior to the equal weighting strategy and IC weighting strategy. The empirical results prove that XGBoost model is effective in predicting IC coefficients and the dynamic weighting based on XGBoost model can improve the performance of multi-factor stock selection strategy.**

Keywords—multi factor stock selection strategy, XGboost algorithm, machine learning

I. Introduction

Multi factor stock selection strategy is one of the main quantitative trading strategies in the stock market. Value factor, size factor, market factor, profitability factor and momentum factor are known to be factors which explain stock returns [1, 2]. The selection of effective factors and factor weighting are important works in the multi factor stock selection model. Factor scoring is usually used to select effective factors, which was proposed by Piotrosk (2002) firstly [3]. When factors are weighted, IC coefficients are usually used. The largest alpha returns can be obtained by maximizing the IC coefficient of portfolio [4, 5]. Although some factors can consistently generate alpha returns, but alpha is time-varying. The research of G. Nalbantov and R. Bauer（2006）found that factor style drift is related to economic and market conditions and factor timing is an important source of alpha return [6]. Many studies show that the time-varying factor weighting strategy is better than the fixed factor weighting strategy. Some machine learning methods, such as support vector machines, stochastic forests and neural networks, are used in factor dynamic weighting models. In this paper, XGBoost method is applied to IC prediction of IC coefficients in multi-factor stock selection model, and the IC are used to weigh the factors dynamically.

978-1-5386-5514-6/18/$31.00 ©2018 IEEE

II. Construction of multi factor stock selection model

A. Selection of factors

Representative factors are selected in seven categories: profitability, quality, size, growth, liquidity, valuation, momentum and reversal. Illiquidity factor is average value of the absolute value of the daily price change divided by the amount of the trading during a period of time. ILLIQ reflects the volatility of the securities price under the unit turnover. If the ILLIQ is small, the impact of securities trading on the price is small, and the liquidity of stock is good. On the contrary, if the ILLIQ is large, the liquidity is worse.

The average IC coefficients of factors from January 2007 to August 2018 are shown in Table 1. Stock pools are constituent stocks of CSI 300 Index and CSI 500 Index. According to the data in Table 1, the factors with larger IC coefficients are liquidity factor, momentum and reversal factor, size factor and valuation factor. These factors will be given greater weights than other factors in strategies.

Table 1. IC coefficients of factors

Category	Factor name	Direction	IC
Size	Total market value(SIZE)	-	-0.033
Valuation	Price-to-book ratio(PB)	-	-0.039
Profitability	Return on equity(ROE)	+	0.008
Growth	Net profit growth rate(GROWTH)	+	0.011
Quality	Flow rate(QUALITY)	+	0.007
Liquidity	Illiquidity factor (LIQUIDITY)	+	0.047
Momentum and reversal	Return of 60 days(RETURN)	-	-0.042

Notes: The adjustment period is 10 trading days, and the market benchmark is the CSI 300 index.

The correlation coefficients of the seven factors is shown in Table 2.

TABLE 2. CORRELATION COEFFICIENT OF SEVEN FACTORS

	LIQUIDITY	SIZE	ROE	GROWTH	RETURN	QUALITY
SIZE	0.87					
ROE	-0.50	-0.70				
GROWTH	-0.32	-0.32	0.58			
RETURN	0.30	0.35	-0.19	-0.13		
QUALITY	0.56	0.58	-0.04	0.11	0.08	
PB	-0.206	-0.24	-0.24	-0.35	0.07	-0.79

Each factor is standardized by quantile transformation and factor ranking information is retained only. The advantage of standardization method is that extreme values can be avoided. For each factor, each stock is sorted according to the factor value to get *Rank* ($s_{i,}$ f_j), so that the score of stock *i* on the factor *j* is:

$$Score(s_i, f_i) = 100 \times \frac{Rank(s_i, f_i) - 1}{N}$$

Here, N represents the total number of stocks in the ranking.

B. Factor weighting

In the dynamic weighting multiple factor model in this paper, the factor weight of each period is weighted by the predicted IC coefficient, which predicted by the XGBoost algorithm. The input data of XGBoost model are the 10-day IC coefficients, 5-day IC coefficients and 1-day IC coefficients of factors, the 20-day returns, volatility rates and turnover rates of the CSI 300 index and the CSI 500 index.

III. XGBOOST MODEL

A. The principle of XGBoost model

Boosting tree is a boosting method based on classification tree or regression tree. It is considered to be one of the best methods in statistical learning. The linear combination of multiple trees can well fit the training data and describe the complex nonlinear relationship between input and output data. CART classifier is a popular decision tree model, which is called classification and regression tree. It is often used as the basic classifier of gradient boosting tree. Each node of CART is divided into two sub nodes, which is a two binary tree.

1) Decision tree model

We review the process of decision tree generation for dealing with regression problems. Assuming that *x* and *y* are input and output variables, and *y* is continuous variable, the given training set is

$$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

where $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$, x_{ij} represents the feature *j* of the sample *i*.

A regression tree divides the input space into K spaces(R_1, R_2, ⋯ , R_K). There is an output value c_k in every space, so the regression tree model can be expressed as:

$$f(x) = \sum_{k=1}^{K} c_k I(x \in R_k)$$

If and only if $x \in R_k$, $I(x \in R_k) = 1$, otherwise is 0.

Choosing the feature *j* as segmentation variable and taking the segmentation point as *s*, two regions can be obtained.

$$R_1(j,s) = \{x | x^j \le s\} \text{ and } R_2(j,s) = \{x | x^j > s\}$$

In order to find *j* and s, the following optimization problems are solved.

$$\min_{j,s} \left[\min_{c_1} \sum_{x_j \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_j \in R_2(j,s)} (y_i - c_2)^2 \right]$$

For the input variable *j*, we can find the optimal cut point *s* by traversing the possible cut points. For output values, obviously

$$\hat{c}_1 = ave(y_i | x_i \in R_1(j,s)), \hat{c}_2 = ave(y_i | x_i \in R_2(j,s))$$

Traverse all the input variables, find the optimal splitting variable *j*. The input space is divided into two regions according to variable (*j*,*s*). Then repeat the segmentation process for each area until the stop condition is satisfied.

2) XGBoost model

The boosting method based on decision tree is called boosting tree. XGBoost model is an efficient method of boosting tree model. The boosting tree model can be interpreted as the addition model of the decision tree. First, determine the initial lifting tree $\hat{y}_i^0 = f_0(x_i) = 0$. The solution model of step t is:

$$\hat{y}_i^t = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{t-1} + f_t(x_i)$$

Minimize the objective function when solving decision tree t:

$$Obj^t = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i) = 0$$

$$= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) + constant$$

Here l is a differentiable convex loss function that measures the difference between the prediction and the target. The term Ω represents the complexity of decision trees. Second-order approximation can be used to quickly optimize the objective in the general setting.

$$Obj^t = \sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{t\text{-}1}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) + cons\tan t$$

Where:

$$g_i = \partial_{\hat{y}_i^{t-1}}(l(y_i,\hat{y}_i^{t-1}))$$

$$h_i = \partial^2_{\hat{y}_i^{t-1}}(l(y_i,\hat{y}_i^{t-1}))$$

Set the T leaves of each decision tree are $w_1,w_2,\dots,w_T$, the complexity of decision trees can be expressed as:

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{t} w_j^2$$

λ and γ are the penalty coefficients. Therefore, by removing the constant term, the objective function can be rewritten to:

$$Obj^t = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)$$

$$= \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{t} w_j^2$$

$$= \sum_{i=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)w_i + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) + w_j^2\right] + \gamma T$$

Definition:

$$G_i = \sum_{i\in I_j} g_i$$

$$H_i = \sum_{i\in I_j} h_i$$

The extreme value can be solved by unary two times function when:

$$w_j^* = -\frac{G_i}{H_i+\lambda}$$

The objective function is the minimum:

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_i+\lambda} + \gamma T$$

When each decision tree splits, one branch splits into two branches. The information gain of the split can be calculated by comparing the target function before and after splitting.

$$Gain = \frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} - \gamma$$

Each possible splitting point of each feature is searched so that the feature and the splitting point with the largest gain value are the optimal splitting variable and splitting point of the splitting.

B. XGBoost model prediction

The weights of factors is calculated according to IC coefficients in dynamic weighting strategy. If the IC coefficient of a factor is bigger, the weight is bigger. Information coefficients (IC) which are defined as correlation between alpha of factors and stock return, are used to express return predictability of alpha. The IC coefficients are predicted by the XGBoost model in dynamic weighting strategy. The input features of the XGBoost model are historical IC coefficients of factors and market variables.

In this paper, the alternative stock pools are constituent stocks of CSI 300 Index and CSI 500 Index. The in-sample training data is the historical data from January 2007 to December 2013, and the out-of-sample test data is the market data from January 2014 to August 2018. The rolling forecasting model is used in each sample forecasting. The sample data are updated and the number of samples is consistent with each time of the stock position changes. In the actual training, the data in the sample is divided into training set and verification set, and the model parameters are determined by the results on the verification set. The final model parameters are shown in Table 3.

TABLE 3. MAIN PARAMETERS OF XGBOOST MODEL

Parameter name	Parameter value
Learning_rate	0.02
Max_depth	4
Subsample	0.8
Colsaple_bytree	1
Lambda	10
Gamma	20
Min_child_weight	3
Colsample_bylevel	1

IV. Empirical analysis

The results of three strategies will be compared to verify the effectiveness of XGBoost algorithm in dynamic weighting multi-factor stock selection. The strategies are equal weight multi factor stock selection strategy (equal weight strategy), IC weighting multi factor stock selection strategy(IC weighting strategy) and dynamic weighting multi factor stock selection strategy (Dynamic Weighting strategy).

A. Rules for policy back testing

Different factor weighting methods are used in three strategies. In the equal weight strategy, the weights of factors are equal, and they are 1. In the IC weighting strategy, the weights of factors are determined by the IC coefficients of the sample training period, $w_i = IC_i/\Sigma IC_i$, and the weights unchanged during the test period. In the dynamic weighting strategy, the factor weights are recalculated again when the stock position is adjusted. The specific weighting method is:

(1)In time t, we predict the IC value of the factor in the next period by the XGBoost model and get the IC_{it}.

(2) For the positive factors, if $IC_{it}>0$, then $w_{it}=IC_{it}$; Otherwise $w_{it}=0$.

(3) For the negative factors, if $IC_{it}<0$, then $w_{it}=\mathrm{IC}_{\mathrm{it}}$;Otherwise $w_{it}=0$.

(4) Normalize the weight w_{it} and make their sum equal to 1.

The other parameters of three strategies are as follows. The cycle time of adjust stock position is 10 trading days. Transaction costs is two-way three thousandths. Stock pool are constituent stocks of CSI 300 Index and CSI 500 Index. The index benchmark is CSI 300 Index. The sample training period is from January 2007 to December 2013 and sample back test period is from January 2014 to August 2018.

B. Back testing results of strategies

The cumulative return curve of three strategies in the back testing period is shown in Figure 1. All the three strategies can overcome market benchmarks. Among them, the cumulative return of dynamic weighting strategy is significantly higher than that of equal weight strategy and IC weighting strategy.

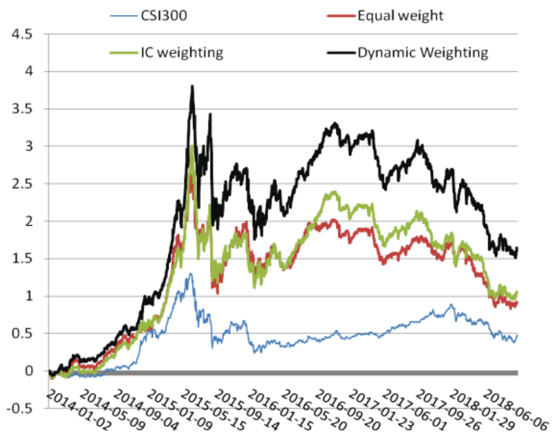

Fig 1. Cumulative return curve of three strategies

The 12 periods average IC coefficients of the three strategies are shown in Figure 2. According to the IC coefficient of figure 3, the dynamic weighting strategy is better than the other two strategies.

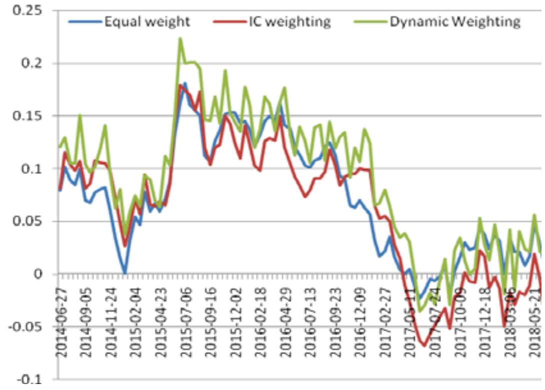

Fig 2. Twelve periods average IC coefficient of three strategies

The weights of seven factors during each period in the dynamic weighting strategy are shown in Figure 3.

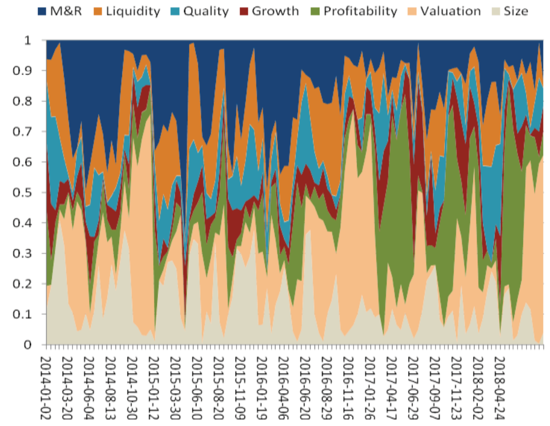

Fig 3. Factor weight of dynamic weighting model

The annual returns of piecewise ranking portfolios in the dynamic weighting multi-factor stock selection strategy are shown in Figure 4. The time period is from 2014 to June 2018. Bar charts show significant monotonicity. In the 10% group with the highest stock ranking score, the annual stock return is 16.82%. On the contrary, the 10% group with the lowest ranking score, the annual return is - 5.99%.

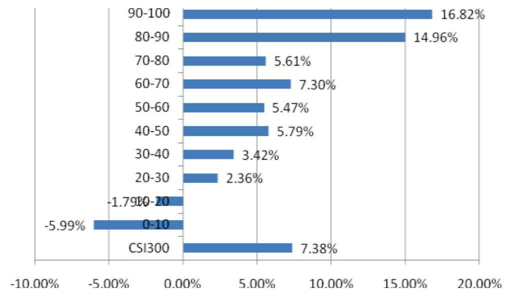

Fig 4. Annual return of piecewise ranking portfolios

The cumulative yield curves of the 10 division long-short strategy with dynamic weighting multi-factor selection method are shown in Figure 5.

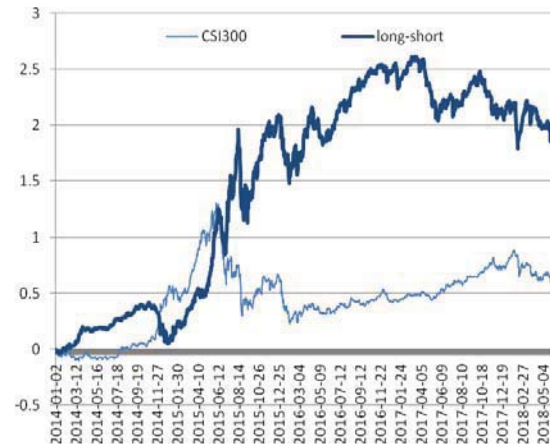

Fig 5. Cumulative yield curve of short-short strategy

The performance indicators of three strategies are shown in Table 4.

TABLE 4. THE PERFORMANCE INDICATORS OF THE THREE STRATEGIES

	Equal weight	IC weighting	Dynamic weighting
Average turnover per year	362. 07%	389. 95%	429. 63%
Winning rate of trading	59. 55%	59. 26%	66. 82%
Annualized yield	14. 56%	16. 22%	22. 54%
Sharpe ratio	0. 37	0. 42	0. 62
Maximum retracement rate	49. 49%	51. 29%	47. 74%
Information ratio	0. 35	0. 43	0. 76
Beta	0. 94	0. 95	0. 98
Alpha	6. 30%	7. 92%	14. 13%

The performance indicators of dynamic weighting strategy, such as annual return, sharp ratio, information ratio, alpha, maximum withdrawal rate, are superior to the equal weighting strategy and IC weighting strategy. The winning rate of dynamic weighting strategy is also significantly higher than that of equal weight strategy and IC weighting strategy. Although dynamic weighting increases the turnover rate of the strategy, the advantage of winning rate is sufficient to compensate for the cost loss caused by the increase of turnover.

V. CONCLUSION

Tree boosting is a highly effective and widely used machine learning method. A dynamic weighting multi-factor stock selection strategy based on XGBoost model is constructed in this paper. Seven factors are selected to represent profit, quality, scale, growth, liquidity, valuation, momentum and reversal. XGboost machine learning method is used to predict the IC coefficients of factors. According to the predicted IC coefficients, the weights of factors are adjusted dynamically. A higher weight will be given to the effective style factor, and the invalid factor will be given a lower weight.

The results of back testing show that the performance measurement of dynamic weighting strategy, such as annual return, sharp ratio, information ratio, alpha, maximum withdrawal rate, are superior to the equal weighting strategy and IC weighting strategy. The empirical results prove that XGBoost model is effective in predicting IC coefficients and the dynamic weighting based on XGBoost model can improve the performance of multi-factor stock selection strategy.

REFERENCES

[1] M. Carhart, “On Persistence in Mutual Fund Performance, ” Journal of Finance, 1997, vol. 52,pp. 57-82.

[2] E.F.Fama, K.R. French, “A five-factor asset pricing model. Journal of Financial Economics”. Apr2015, Vol. 116 Issue 1, pp. 1-22.

[3] J. D. Piotrosk, “Value investing: The use of historicalfinancial statement information to separate winners from losers”. Journal of Accounting Research. 2000, vol. 38, pp. 1-41.

[4] R. Grinold, “A Dynamic Model of Portfolio Management, ” Journal of Investment Management,2006, vol. 4, pp. 5-22.

[5] N.Garleanu, L. Pedersen, “Dynamic Trading with Predictable Returns and Transaction Costs, ” Journal of Finance, 2013, vol. 68, issue 6, pp. 2309-2340.

[6] G. Nalbantov, R. Bauer,”Equity style timing using support vector regressions”. Applied Financial Economics, 2006, vol. 16, pp. 1095–1111.