Stock Selection Model Based on Random Forest

Chenyao Ma(✉)

Tongji University, Shanghai, China
machenyaotj@163.com

Abstract. Nowadays, the stock selection has become increasingly significant in financial field with the rapid development of Quantitative Investment. At first, as we all known, traditional style of equity investment involves personal scrutiny of available data on a company, including subjective assessments of the company's operating and financial situation John, Miller, & Kerber [4]. However, with the continuous development of the financial industry, the number of shares has increased rapidly, which lead to a mass of data. At the same time, the limited calculation and quantitative ability of quants lead to inadequate and incomplete stock selection strategy. Besides, there are also some boundedness of the traditional type such as the subjective assume of the quants and the low recoverability. Using the machine learning, we establish a new model used to predict the stock's return rank of next term based the factors from Barra Equity Model Lu, & Lu, [7]. We mainly make use of Random forest to establish the model. And it has been divided into two models while the one of it is the regression model to predict the rank of the stock's return rate, the other is the classification model to predict if the rank of the stock portfolio can exceed 50% of all. Finally, we take the back test on the basis of the model to verify the accuracy of the model, and finally we concluded that the model we built was effective.

Keywords: Stock Selection · Quantitative Model · Random Forest · Machine Learning

1 Introduction

With the reason that the investment of stocks in a portfolio is able to bring investors excess revenue, investing in stocks has long been a favorite of investors. To the fund managers, the stocks are great choices as investment products. The alpha value, the imbalance between the actual risk return and the average excepted risk return, which is defined as the absolute return of an investment portfolio also originate from the stocks. However, how to select appropriate stocks in the stock market and forecast the yield is a problem that has long puzzled investors. For instance, in China's A share market, most investors are in the red. The traditional stock selection method is based on the quants' analysis of various data, such as the operation situation of the enterprise, the profit situation of the enterprise, the financial statements of the enterprise and the transaction data. However, as Andrew Metrick pointed out that there is no evidence that newsletters have superior stock-selection skill, either over short- or long-term horizons [10]. And,

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 654–663, 2023.
https://doi.org/10.2991/978-94-6463-010-7_67

due to the uncertainty of the stock market, the increasing number of the list companies and the expansion of the fields involved by companies, there are a mass of data and information is related to stock selection. In this condition, investors always make a subjective judgment of stock returns based on personal experience only, just like X. Yuan, J. Yuan refer that most of the investors have limited information and ability to predict the stock price trend well [18].

In general, this traditional way of stock selection has its limitations, which are reflected in the incomplete collection of market information and the unrepeatability of investors' experience. Therefore, it is meaningful to construct a new more objective and effective stock selection model. The machine learning is a popular model being able to analyze a large amount of data which can make the analysis more effective and cover more comprehensive information. Before our research, there have been some studies on this kind of problems, such as Qiyang Sun's paper: Neural Network Algorithm Strategy Based on Multi-factor Stock Selection [14]. In this paper, a regression model is constructed based on multifactor model and natural neural network algorithm. In Runhuan Liu's paper: Stock selection strategy based on support vector machine [8] and Canran, Xiao, Liwei, Hou, Jun, & Huang's paper: Research on Multi-factor Stock Selection Strategy based on Improved Particle Swarm Support Vector Machine [17], they also build stock selection model based on multi factor model and support vector machine. However, most of these articles only consider the traditional multi factor model, the lack of selected factors leads to large errors in the model. In addition, we also improved the selection of the machine learning model, combined the regression model and classification model, and compared the prediction effect of the two prediction models. Thus, in this paper, we established the predicted model based on machine learning.

The data we used were the ten factors form the Barra risk factor model, such as the Size, Beta, Residual volatility and so on. Our job is to predict the ranking of the standardized return rates of stocks in China's A-share market and whether the standardized return rates can rank in the top 50%. We made use of the classical random forest algorithm in machine learning to construct regression model and classification model. On the same time, to measure the effectivity of this model, we used the predicted value to back test the model. We calculated the correlation between the predicted value' rank and the practical value' rank, and the relevant risk indicators such as: Information Coefficient (IC), Information Ratio (IR) and Sharpe Ratio, as well as net value curves of portfolios grouped by predicted values.

The rest of the article is organized as follows: In Sect. 2, we have a brief description of the data we used and introduce the process of data preprocessing. In Sect. 3, we use the random forest algorithm in machine learning to construct regression model and classification model to predict stock return rate. In Sect. 4, we conduct back test with the constructed models and further evaluated the models based on various risk indicators. We also analyzed the results of the model. Finally, we summarized the whole modelling process and put forward the improvement plan.

2 Date Description

2.1 Date Discription

In order to build the model, we selected data from April 8, 2005 to August 31, 2020, including the market value, daily return, industry classification and ten Barra risk factors of 300 constituent stocks of CSI 300. The CSI 300 index is composed of 300 stocks with larger market value and stronger liquidity in Shanghai and Shenzhen stock exchanges. The reason why we choose it as the data resource of the model is that it can better reflect the overall performance of China's A-share market, those stocks is more representative and can make the model more stable.

We download those data from https://www.joinquant.com/. Barra risk factors are 10 major style factors proposed in Barra structural risk model, which are used to explain the volatility of stock prices. Barra structural risk model is based on the (APT) Arbitrage Pricing Model, which is used to predict the yield. Arbitrage pricing model is an extension of Capital Asset Pricing Model [13], which is based on multi-factor model. It is a cross-sectional linear model that the stock return is linear with a group of factors [16]. The reason is that in the structured model, various factors are correlated with the stock characteristics, so the yield can be predicted. We refer to the method of Morgan Stanley Capital International to calculate these ten factors in The Barra China Equity Model [11] First, CNE5 model first calculate 21 secondary factors and then combined them into 10 major style factors according to the given weight. The meanings of the 10 risk factors and 21 secondary factors are listed in Table 1.

2.2 Date Processing

Considering the huge amount of our data, we preprocess the data in order to reduce the amount of calculation, save the time of calculation, and make our data better adapt to our prediction model and prediction target. We found that the types of variables in the original data were complex and had few missing values, followed by some stocks that were not even listed earlier in the year. In order to deal with missing values in the data, we deleted records with missing values. Firstly, we have two forecasting objectives, one is to predict the rank of yield in the future 20 trading days, the other is to predict whether a stock's forecast yield is in the top 50% of all stocks. In order to achieve the predict goal, we combine daily return into the return in the next 20 days. At the same time, we also transform the industry classification into dummy variables. In order to avoid the inconsistency of cross-section distribution which means that the data on different cross-sections are not comparable, we need to standardize the factors. The standardized formula is as follows:

$$\mathrm{Z} = \frac{X - U}{\sigma}$$

The first problem is the regression problem and the other is the classification problem. For the second problem, we define a categorical variable as our prediction target, that is, if the yield ranking is in the top 50%, we define its value as 1, otherwise it is 0.

Table 1. The meanings of the 10 risk factors and 21 secondary factors

First-grade factor	Second-grade factors	Factor meaning
Size	Lncap	Logarithm of total market value
Beta	Beta	History beta
Momentum	Rstr	Historical abnormal return series value
Residual Volatility	Dastd	Historical abnormal return series volatility
	Cmra	Cumulative excess return spread
	Hsigma	Volatility of historical residual return series
Decapitalization	Nlsize	The cube of size
Book-to-price	Btop	Book value / Market value
Liquidity	Stom	Monthly turnover rate
	Stoq	Seasonal turnover rate
	Stoa	Annual turnover rate
Earnings yield	Epfwd	Price to earnings ratio
	Cetop	Operating cash flow / Market value
	Etop	Net profit / Market value
growth	Egrlf	Analysts forecast long-term profit growth rate
	Egrsf	Analysts forecast short-term profit growth rate
	Egro	Profit growth rate
	Sgro	Growth rate of operating income
leverage	Mlev	Market leverage
	Blev	Book leverage
	Dtoa	Asset liability ratio

3 Model Building

In order to build the model, we need to consider that stock returns change with time, which is a time series. At each time point of position adjustment, we first train our model with the data from the last 252 trading days, and use this model to give a forecast of the next 20 days' return rank, and build a portfolio according to the forecast value. We roll forward 20 days at a time to do the cycle calculation. The process of model construction is shown in Fig. 1.

In the construction of the prediction model, we adopted the random forest model. Random forest [6] Random Forest is a model based on decision tree model. It establishes multiple decision trees and integrates them to gain a more accurate and more stable model. Decision tree model is a classic machine learning model, which can be used for classification and regression problems [15]. There are many algorithms for decision tree model, such as ID3 algorithm and C4.5 algorithm, etc. [12]. They determine the depth of the tree by information gain and information gain ratio, respectively. We

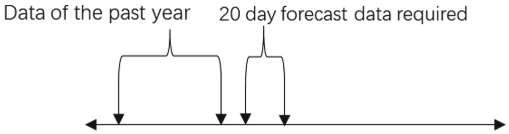

Fig. 1. Forecast method

choose ID3 algorithm to construct the decision tree. In this algorithm, the construction steps of decision tree include the following steps: First, the root node is created and all possible information entropy gains are calculated for the node. After calculating the information entropy, the maximum information entropy is selected as the node feature, the optimal feature is divided into datasets to construct leaf nodes, and the above method is used recursively on the leaf nodes. However, we know that a single decision tree has an unstable prediction effect on the model. Therefore, we use the random forest model, which is a representative ensemble learning method, to improve the accuracy of prediction. Random forest algorithm is a relatively new machine learning algorithm proposed by Breiman (2001) [1]. The randomness of random forest model is reflected in the construction of decision tree through random sampling and random features. Its principle is to build multiple decision trees which are based on samples in the bag and count the prediction results of each decision tree on the sample data, and finally select the best result from all the results by voting method [18]. The advantage of random forest in this model is that it can directly calculate the data of high latitude, reduce the information loss caused by data dimension reduction, and improve the accuracy and stability of the model. In the process of building the model with the data of the past 252 trading days, 70% of the data is used as the training set and 30% as the test set, and the optimal parameters are selected according to the test error.

We can get the stock yield predictive value in the next 20 days by taking the data into the random forest model. As mentioned above, our forecasting problems can be divided into classified forecasting and regression forecasting. From the specific calculation steps, we first ranked the stock yield from large to small, and divided it into 5 groups on average, with the stocks in each group holding equal weight. After holding for 20 days, we build a model to exchange positions, and then continue to hold for another 20 days, which is a cyclical process.

4 Back Test

In this paper, the calculation is carried out by using the methods introduced in the section of model building. We adopted the time window rolling method mentioned above, train our model with data in the past year, and then obtained our prediction. At the same time, in order to make the model more accurate, we build random forest with different number of trees and compare the results. By calculating the testing accuracy, we selected the random forest with minimum test error as our model, and further analyzed and studied them.

We analyze the results by back testing the constructed model. For the regression problem, we rank directly by the projected yield and get the portfolio. According to the ranking value, we divide stocks into five portfolios evenly. For the classification problem, we are able to rank according to the probability of top 50% and construct the portfolio. In order to analyze the results, we first need to sort all the stocks into 5 groups according to the forecast yield and calculate the net value of each group. In addition, we calculated a number of risk return indicators, including total yield, annualized yield, annualized volatility and SHARP ratio. The SHARP ratio Koldanov, Kalyagin, & Pardalos [5] measures the excess return for each additional unit of risk taken, so the higher the value is, the better. The formula is as follows:

$$SharpeRatio = \frac{\mathrm{E(R_P)} - R_f}{\sigma_p}$$

In order to more accurately measure the accuracy of the model and evaluate the prediction effect of the model, we also calculated the Pearson's and the Spearman's correlation coefficient (IC) between the predicted value and the real income, and calculated the mean IC, standard deviation of IC and IR (information ratio) of different panels. The Pearson correlation coefficient is used to measure the degree of correlation between the 2 variables, and the Spearman correlation coefficient is used to measure the degree of rank correlation between the 2 variables. At the same time, the IC and IR values calculated here are commonly used to measure the prediction ability in quantitative investment, in which IC can judge the prediction ability of the factor value on the next stock selection yield. The IR value takes into account the stock selection ability and the stability factors of stock selection prediction.

The performance of regression model and classification model is shown in Table 2. From the data in Table 2, we can see that the IC and IR values of Pearson and Spearman coefficients of our classification model are greater than those of the regression model. We can conclude that the effect of the classification model is better than that of the regression model. The reason is that the dependent variable of the classification model is the classification variable only considering the values of 0 and 1, while the dependent variable of the regression model is a continuous variable, so the prediction is more difficult, resulting in the prediction effect will be lower than the classification model.

Finally, we will draw the relative net value curve and net value curve of the selected stocks of each group calculated by the regression model and the classification model in the Figs. 2 and 3. For the classification model, the net value curves of the 5 combinations are monotonic and separable, indicating the effectiveness of the classification model. However, the regression model is not as effective as the classification model because there are some cross in the net worth curve. However, we have observed that the time period where the net worth curve obviously intersects is around 2008 and 2015 [3]. We

Table 2. The performance of regression model and classification model.

	Reg-Pearson	Reg-Spearman	Clf-Pearson	Clf-Spearman
IC	0.03	0.05	0.05	0.07
Std	0.18	0.19	0.19	0.21
IR	0.19	0.26	0.26	0.32

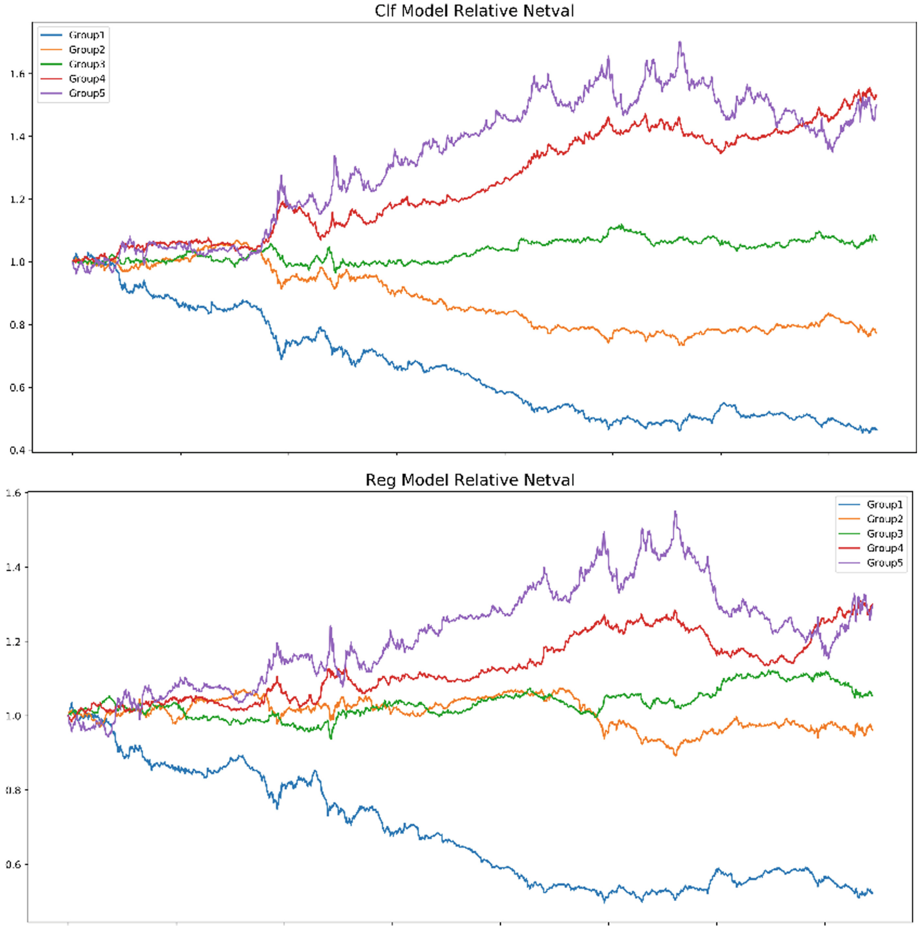

Fig. 2. Relative net value of regression model and classification model

can know that in these 2 time periods, the world's financial markets have experienced tremendous turbulence, which has also greatly affected the Chinese stock market. Due to the occurrence of the black swan event or small probability events, it is beyond the scope of the original prediction model, resulting in a large deviation and low accuracy of prediction. In addition, most of the forecast results of our model are more accurate. We have also drawn the relative net worth curves of 5 portfolios, which can more clearly reflect the forecast effect. It is more obvious from the figure that the prediction effect is poor when the time is earlier, which is due to the less training amount of data in the initial prediction stage, resulting in the lower accuracy of the prediction. At the same time, by drawing the relative net value curve, we can find that the prediction effect of the classification model in our model is better than that of the regression model.

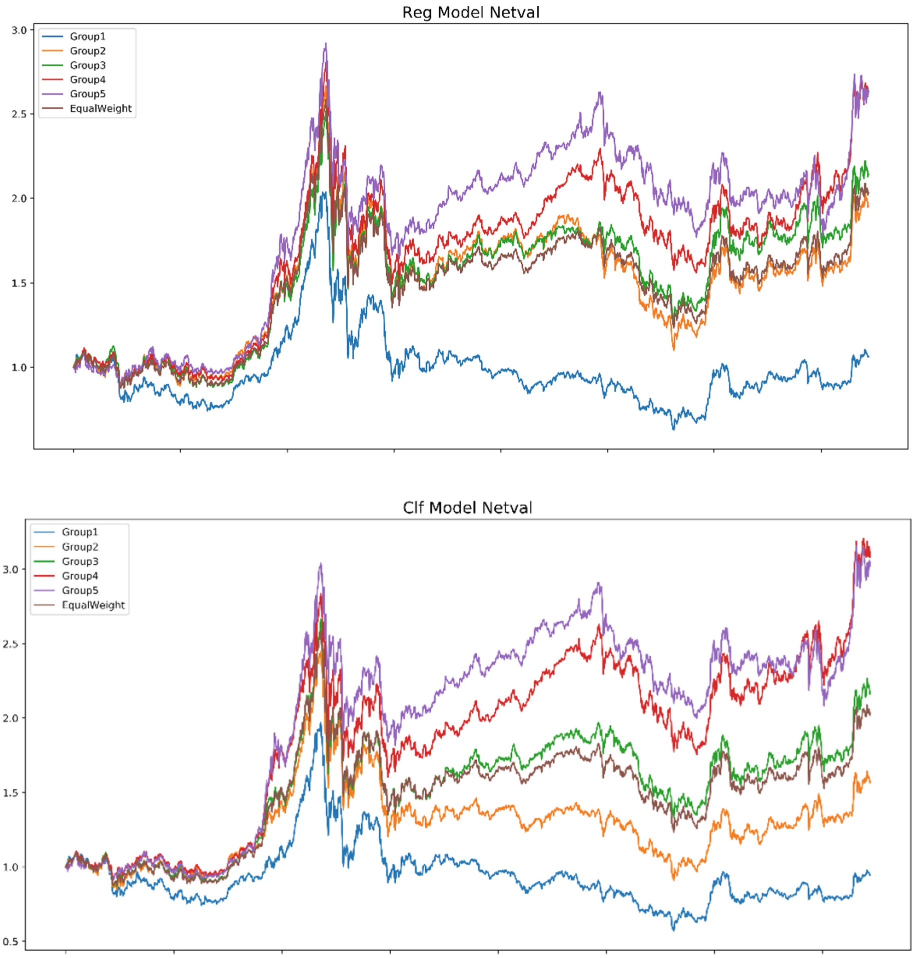

Fig. 3. Net value of regression model and classification model

5 Conclusions

Quantitative investment is an important field of financial industry, which involves financial engineering, financial mathematics and other disciplines. With the development of quantitative investment technology, more and more investors invest in the stock market. Therefore, this paper hopes to build a quantitative investment model with the help of the popular machine learning algorithm to improve the accuracy of investors' investment and reduce their losses. We forecast the stock return objectively through a large amount of data. We collected data on various factors of the 300 stocks included in the CSI 300 index of the A share market from August 2,005 to 31 August 2,020. We combine the most popular random forest algorithm to build regression model and classification model. We combined the Barra structure risk model to predict the yield of each stock, and constructed the required portfolio through the prediction results. Not only that, in order to measure the accuracy of the model, we carried out back testing on the model, and we used Sharp Ratio and IC and other indicators to measure the accuracy of the model. Finally, we conclude that the investment portfolio we constructed is meaningful for investment, and the accuracy of the model is higher when the market is relatively stable. However, there are still many areas for our model to improve. Firstly, we can find more effective factors through data mining to reduce model bias. Secondly, we can use different algorithms such as Xgboost [2] and short-term and long-term memory [9] to predict the yield.

References

1. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
2. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 785–794, August 2016
3. Hall AB, Yoder J, Karandikar N (2017) Economic distress and voting: evidence from the subprime mortgage crisis
4. John GH, Miller P, Kerber R (2002) Stock selection using rule induction. IEEE Expert 11(5):52–58
5. Koldanov AP, Kalyagin VA, Pardalos PM (2015) Step down and step up statistical procedures for stock selection with sharp ratio. In: Pardalos P, Pavone M, Farinella G, Cutello V (eds) Machine Learning, Optimization, and Big Data. MOD 2015. LNCS, vol 9432, pp 26–36. Springer, Cham. https://doi.org/10.1007/978-3-319-27926-8_3
6. Lee S, Kim J (2020) Prediction of nanofiltration and reverse-osmosis-membrane rejection of organic compounds using random forest model. J Environ Eng 146(11):04020127
7. Lu S, Lu C (2018) Barra risk model based idiosyncratic momentum for Chinese equity market. Available at SSRN 3140113
8. Liu R (2020) Stock selection strategy based on support vector machine. In: 2020 the 3rd International Conference on Machine Learning and Machine Intelligence, pp 10–13, September 2020
9. Meng X, Liu M, Wu Q (2020) Prediction of rice yield via stacked LSTM. Int J Agric Environ Inf Syst (IJAEIS) 11(1):86–95
10. Metrick A (1999) Performance evaluation with transactions data: the stock selection of investment newsletters. J Financ 54(5):1743–1775
11. Orr DJ, Mashtaler I (2012) Supplementary Notes on the Barra China Equity Model (CNE5)
12. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674
13. Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. J Financ 19(3):425–442
14. Sun Q (2021) Neural network algorithm strategy based on multi-factor stock selection. In: Abawajy JH, Choo K-K, Xu Z, Atiquzzaman M (eds) ATCI 2020, vol 1244. AISC. Springer, Cham, pp 295–300. https://doi.org/10.1007/978-3-030-53980-1_44
15. Sun J (2021) Decision tree classification algorithm in college PE teachers' score analysis. In: Atiquzzaman M, Yen N, Xu Z (eds) Big Data Analytics for Cyber-Physical System in Smart City. BDCPS 2020. AISC, vol 1303, pp 780–786. Springer, Singapore. https://doi.org/10.1007/978-981-33-4572-0_112
16. Trzcinka C (1986) On the number of factors in the arbitrage pricing model. J Financ 41(2):347–368

17. Xiao C, Hou L, Huang J (2019) Research on multi-factor stock selection strategy based on improved particle swarm support vector machine. In: 1st International Symposium on Economic Development and Management Innovation (EDMI 2019), pp 437–441, August 2019. Atlantis Press
18. Yuan X, Yuan J, Jiang T, Ain QU (2020) Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market. IEEE Access 8:22672–22685