Open Access Library Journal
2023, Volume 10, e10370
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Multi-Factor Stock Selection Model Based on Categorical Prediction Model

Yufan Hu

The Affiliated High School to Hangzhou Normal University, Hangzhou, China
Email: huyf6886@outlook.com

How to cite this paper: Hu, Y.F. (2023) Multi-Factor Stock Selection Model Based on Categorical Prediction Model. *Open Access Library Journal*, **10**: e10370.
https://doi.org/10.4236/oalib.1110370

Received: June 11, 2023
Accepted: July 17, 2023
Published: July 20, 2023

Abstract

In order to reflect the concept of value investing, this paper extracts the indicators that reflect the fundamental information of listed companies such as profitability, solvency, operating capacity, growth capacity, and cash flow from the annual reports of listed companies in A-share market in 2021 as factor characteristics, establishes a multi-factor stock selection strategy based on cluster analysis model and classification prediction model respectively, and conducts an empirical study. The results show that after clustering based on the fundamental factor indicators, the investment portfolio with investment value can be classified and far outperform the performance of the SSE index in the same period, showing a high potential value of the investment. When performing classification prediction modeling, the test results on the test set show that it has a high winning rate when selecting stocks based on the prediction results.

Subject Areas

Chinese Stock Market, Stock Selection, Statistical Model

Keywords

Stock Selection Model, Multifactor Models, Cluster Analysis, Random Forest, Logistic Regression

1. Introduction

Since the establishment of the Shanghai Stock Exchange in 1990, China's securities market has grown rapidly from scratch over the past 30 years, with more than 4000 companies listed on the A-share market, the formation of two major stock exchanges, the Shanghai and Shenzhen Stock Exchanges, and the multi-level three-dimensional trading markets, such as the Main Board, Small and

Medium-sized Board, Venture Board, Science, and Technology Board and New Third Board. At the same time, the relevant policies and regulations of China's securities market are increasingly improved and open to overseas investors, and the international influence is growing. In recent years, market managers have consciously guided the trading behavior of market traders and cultivated the concept of value trading. In the eyes of value investors, the investment value of a stock is mainly reflected in the fundamental indicator data of the underlying listed company. Based on the fundamental indicator data, the relevant data analysis model is used to screen stocks and select the underlying stocks with investment value that is the main content of this paper's research.

Tushare is a free and open-source Python financial data interface package. It can provide financial analysts with fast, neat, and diverse analysis-friendly data, greatly reducing their workload in data acquisition and allowing them to focus more on researching and implementing strategies and models. Considering the advantages embodied by the Python pandas package in quantitative financial analysis, the vast majority of data formats returned by Tushare are DataFrame types, facilitating data analysis and visualization with pandas, NumPy, and Matplotlib.

The multi-factor stock selection model is one of the main quantitative stock selection models currently available. In order to make an objective and accurate assessment of the company's operating conditions, a comprehensive evaluation index system based on the main financial indicators is needed. [1] [2] constructed an index system for the comprehensive evaluation of the performance of listed companies in the information technology industry and the steel industry, respectively, considering profitability, solvency, operating capacity, and growth capacity. [3] used P/E ratio, P/N ratio, 20-day average turnover ratio, ROE, 20-day average momentum, and price-sales ratio as valid factors in the multi-factor model, clustered all stocks of CSI 300 according to the data of valid factors and identified the group with the best investment returns as potential portfolio targets through backtesting [4]. The P/E ratio, operating income growth rate, NAV growth rate, net profit growth rate, and gearing ratio were selected as the factors reflecting the information of stock ups and downs, and the response variable was whether the stock's return exceeded that of CSI 300, and a logistic regression model was established to screen the stocks that were predicted to outperform CSI 300 as potential investment targets. The backtest found that the model has a winning rate of 80%.

This paper mainly hopes to find an effective stock selection model based on the fundamental factor data provided by the annual reports of listed companies and to provide guidance for future investment by establishing corresponding data analysis methods.

2. Selection of Samples, Indicators, and Data Preparation

In this paper, all companies with ST and *ST are excluded when selecting examples, so to make the data as realistic and credible as possible; the data of newly listed IPO and sub-IPO stocks this year are excluded because the stock prices of IPO and sub-IPO stocks are easy to be manipulated and speculated, so are not representative. Finally, the remaining 4234 A-share listed companies were included in the study sample. In selecting the indicators for comprehensive evaluation of the fundamental performance of listed companies, we comprehensively referred to [1] [2] and relevant information in the financial platform of Flush and Tushare financial data interface package, and finally selected 36 indicators from five aspects of profitability, solvency, operating capacity, growth capacity, and cash flow, combined with stock-based indicators after comprehensive consideration to measure the fundamentals reflecting the listed companies level, as shown in Table 1.

In this paper, we extracted the data of 38 indicators in the above table from the data of 4234 sample listed companies' 2021 annual reports through the Python financial data interface package Tushare (http://tushare.org/). In order to eliminate the influence of the magnitude, we first standardized the data before modeling and analysis. The standardization method is shown in Equation (1):

$$x_{ij}^{*} = \frac{\left(x_{ij} - \overline{x}_i\right)}{s_i}, i = 1, 2, \cdots, 43; j = 1, 2, \cdots, 3083 \tag{1}$$

where x_{ij} is the original value of the ith indicator of the jth listed company, $\overline{x}_i$ is the sample mean of the ith indicator, s_i is the sample standard deviation of the ith indicator, and x_{ij}^{*} standerlized vale of the ith indicator of the jth listed company.

3. Empirical Study of Multi-Factor Stock Selection Strategy

3.1. Selection of Effective Factor Indicators

In order to eliminate noise interference as much as possible and screen out the

Table 1. Fundamental comprehensive rating index system of A-share listed companies.

Basic index	Profitability index	Operational capacity index	Growth ability index	Debt solvency index	Cash flow index
pe	roe (%)	arturnover (time)	revenue growth rate	currentratio	cf_sales
Outstanding (100 million)	net_profit_ratio (%)	arturndays (day)	nprg (%)	quickratio	ocf
totals (100 million)	gross_profit_rate (%)	inventory_turnover (time)	nav	cashratio	cf_nm
gross revenue gross revenue	net_profits (ten thousand yuan)	inventory_days (day)	targ		cf_liabilitie
net assets	eps (in the basic table)	currentasset_turnover (time)	Operating profit growth rate		
bvps	roa		tds		
pb	bips		nfc		
total assets			Revenue growth rate		
total liabilities					

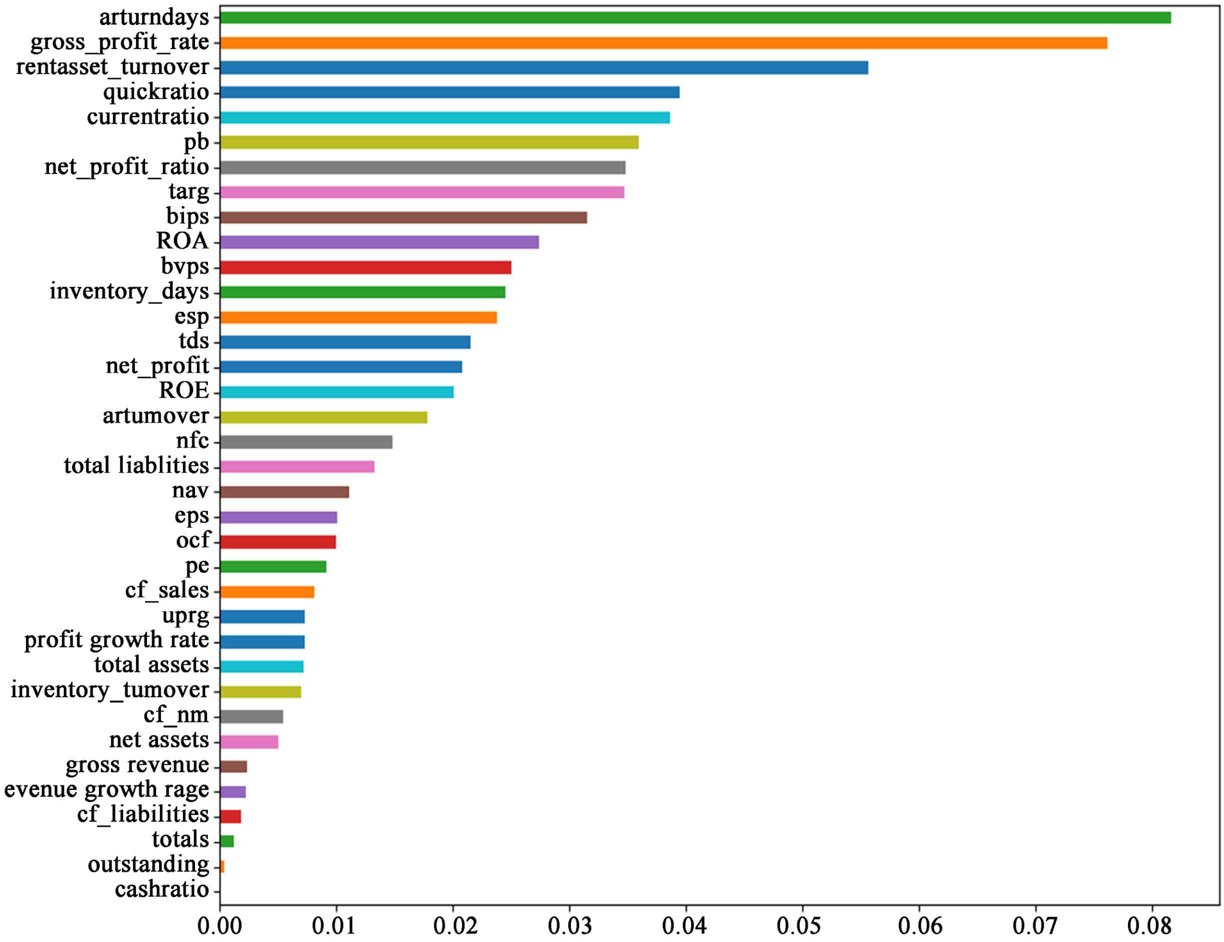

Figure 1. The absolute value of the correlation coefficient between each fundamental factor index and the cumulative benefit of stocks after the annual report.

underlying investment value more effectively, we selected some effective indicators based on the correlation between each annual report indicator and the cumulative stock return for the period of March 31-June 30, 22 years, three months after the annual report, *i.e.*, some indicators with the strongest correlation with the cumulative stock return for the three months after the annual report.

Figure 1 shows the absolute correlation coefficient value between each fundamental factor index and the cumulative benefit of stocks after the annual report. Through observation and comparison, we chose the top 16 fundamental factor indexes. The names of these indexes and their correlation coefficients with the cumulative benefit after the annual report are as follows (Table 2).

3.2. Empirical Research Based on Cluster Analysis Model

Based on the purpose of value investing, we apply K-means clustering to the sample subject according to the fundamental index data published in the 2021 annual report. Because of the large number of stocks, we divide 4234 stocks into 10 categories and then compare the investment performance of various stocks in the three months after the publication of the annual report. The results are shown in Table 3.

From the results, except for one stock in Group 4, the combination of six stocks in Group 8 has the most investment value, with a cumulative benefit of 13.17% in the three months after the annual report of the combination, while the cumulative increase of the Shanghai Composite Index in the same period is

Table 2. Sixteen factors that have the strongest correlation with the accumulated income after the annual report.

factors	correlation coefficient
arturndays	0.081662
gross_profit_rate	0.076149
currentasset_turnover	0.05564
quickratio\n	0.039467
currentratio	0.038637
pb	0.035949
net_profit_ratio	0.034814
targ	0.034769
bips	0.031505
ROA	0.027427
bvps	0.025038
inventory_days	0.024497
esp	0.02378
tds	0.021511
net_profit	0.020797
ROE	0.020105

Table 3. All stock clustering and backtesting results.

Category label	quantity	Cum_exchange (%)	Ex_return (%)	Std_return (%)
4	1	17.01	0.23	2.93
8	6	13.17	0.18	4.14
1	145	4.36	0.02	3.01
2	2239	3.07	0.01	3.28
6	1	1.01	0.01	3.69
0	1213	0.75	-0.03	3.31
3	95	-1.04	-0.05	3.49
9	532	-2.25	-0.08	3.38
7	1	-8.11	-0.15	3.63
5	1	-18.77	-0.37	3.26

4.04%. Because the Shanghai and Shenzhen 300 stocks include 300 representative stocks in Shanghai and Shenzhen, the business development of the target company is relatively more stable. Next, we use the Shanghai and Shenzhen 300 stocks as potential targets for modeling and analysis. Similarly, according to the fundamental index data published in the 2021 annual report, the Shanghai and Shenzhen 300 underlying stocks are divided into 10 categories by K-means clustering, and the performances of various stocks after the publication of the annual report are as follows (Table 4).

From the results, the 14 stock portfolios in Group 4 have more investment value than other groups, and the cumulative benefit of the portfolio in the three months after the annual report reaches 16.75%, which is higher than other categories and 13.17% of all stock cluster group 8 in Table 3. At the same time, the standard deviation of Group 4's daily return of 2.96% is lower than that of all stock cluster 8's 4.14%, and the relative investment value is higher.

3.3. An Empirical Study Based on Classified Forecasting Model

In this part, we set the state 0 (underperforming index) or 1 (underperforming index) as the target response varies according to whether the cumulative increase of stock prices outperformed the Shanghai Composite Index in the three months after the annual report, and take the standardized data of 36 fundamental index variables in Table 1 as the explanatory variables, trying to establish random forest and logistic regression models to predict the future state of stock prices. Training models and evaluating the prediction effect of the model are used to identify the stock status after the annual report data comes out and choose to invest in those stocks that are predicted to outperform the index.

In order to reduce the risk and increase the stability of the model strategy, we directly model the Shanghai and Shenzhen 300 stocks as potential targets, and divide the training data and test data according to the ratio of 4:1. Firstly, we

Table 4. Cluster analysis and backtest results of CSI 300.

Category label	Quantity	Cum_exchange (%)	Ex_return (%)	Std_return (%)
4	14	16.75	0.20	2.96
1	57	12.71	0.14	2.71
6	1	12.21	0.16	2.84
7	32	8.03	0.09	3.34
3	37	7.19	0.07	2.89
5	24	2.85	-0.01	2.78
0	65	2.12	-0.02	2.65
9	3	1.73	-0.03	2.83
8	5	1.58	-0.03	2.17
2	1	-18.77	-0.37	3.26

established a logistic regression model, and the prediction accuracy of the model on the test data was 64.58%. The confusion matrix is shown in Figure 2(a).

From the confusion matrix, the model predicts that there are 34 stocks that outperform the Shanghai Composite Index, among which 24 stocks do outperform the index, with a winning rate of 70.58%.

Next, we trained the random forest model and optimized the model parameters by 10-fold cross-validation. The prediction accuracy and investment success rate of the model on the test samples is close to those of the logistic regression model, as shown in Figure 2(b). From the confusion matrix, the model predicts that there are 35 stocks that outperform the Shanghai Composite Index, among which 24 stocks actually outperform the index, with a winning rate of 68.57% and about 70%.

4. Conclusions

This paper selects the index data of listed companies in A-share market in 2021, which reflects the fundamental information such as profitability, solvency, operating ability, growth ability, and cash flow, and establishes multi-factor stock selection strategies based on cluster analysis model and classification prediction model respectively, and makes an empirical study.

In Section 3.2, after clustering based on the fundamental factor indicators, the investment performance of all the stock classification categories 8 and 4 of the Shanghai and Shenzhen 300 stock classification is higher than that of the Shanghai Stock Exchange Index in the same period, showing high investment potential value.

In Section 3.3, we respectively established logistic Regression and Stochastic Forest Model [5] to predict whether the stock price will outperform the Shanghai Composite Index after the annual report according to the fundamental information

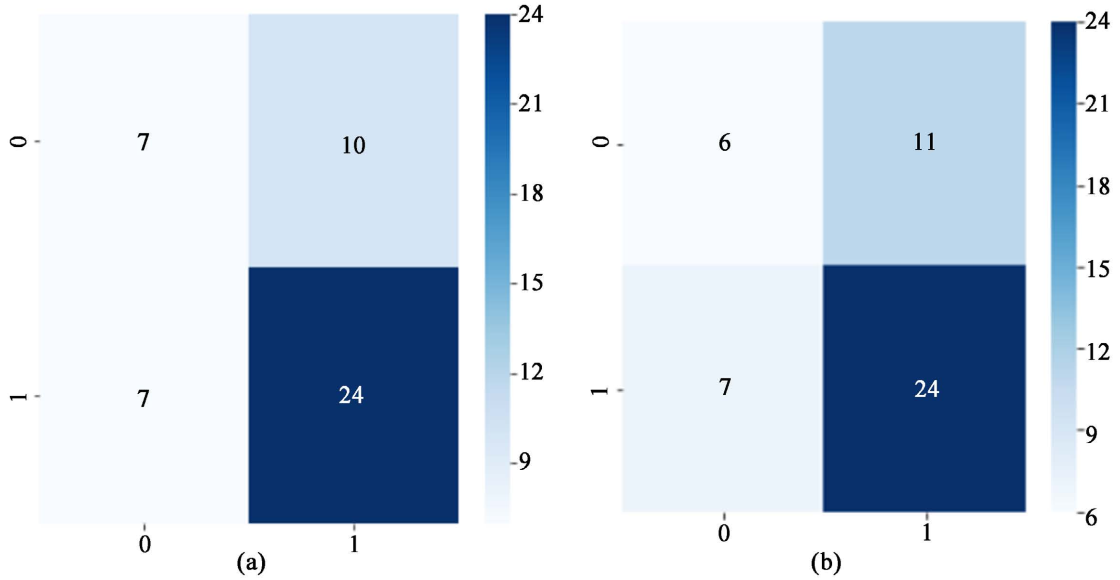

Figure 2. Prediction confusion matrix on test samples. (a) Logistic regression model; (b) Random forest model.

and selected the stocks predicted to outperform the Shanghai Composite Index as potential investment targets. Judging from the prediction effect of the model on the test set, when selecting stocks based on the two classification models in Shanghai and Shenzhen 300 constituent stocks, they all have a winning rate of about 70%, which has good empirical application value.

It can be seen that the multi-factor stock selection strategy based on the concept of value investing has a certain investment reference value in China market. It should be noted that since only one year's annual report data of the sample stocks is used in the study, the stability of the model remains to be discussed. Future studies will consider using several years of sample data to train the model and analyze the reliability of the model. At the same time, we can also consider introducing market trading index factors such as volume and alpha factor into the multi-factor stock selection model to improve the accuracy and success rate of stock selection.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] He, Y. (2016) Research on Performance Evaluation of Listed Companies in China's Information Technology Industry Based on Principal Component Analysis and Cluster Analysis. *Modern Business Trade Industry*, **37**, 74-76.

[2] Li, L. and Chu, X. (2009) Performance Evaluation of Listed Companies in the Steel Industry Based on Factor Analysis. *Journal of Shenzhen University* (*Science and Technology Edition*), **26**, 217-220.

[3] Huang, R. (2016) Multi-Factor Stock Selection Model Based on K-Means Clustering. *Business Information*, No. 34, 231.

[4] Wang, W. and Cai, W. (2020) Research on the Probability Prediction of Stock Price Rising Based on Logistic Regression. *China Market*, No. 6, 7-8.

[5] Wang, L. (2019) Stock Selection Strategy Based on Deep Forest. *Economic Research Guide*, No. 27, 78-79.