ISSN 2959-6157

An Example of Machine Learning-Based Multifactor Dynamic Quantitative Stock Picking Models

Haocheng Sun

Electronic Information Engineering, Xi’an University of Posts and Telecommunications, Xi’an, 710061, China
SunHaocheng1107@163.com

Abstract:

This study utilizes the XGBoost algorithm in the field of machine learning to conduct quantitative stock picking research for CSI 300 stocks. The article firstly outlines the importance and practical application background of quantitative stock selection, and then discusses in depth the basic principle of XGBoost algorithm and its application method in quantitative stock selection. By collecting historical data of CSI 300 stocks and after data preprocessing, this study constructs a multi-factor stock prediction model based on XGBoost and conducts relevant backtesting. Comparative experiments show that the XGBoost algorithm exhibits good effectiveness and demonstrates the unique advantages and characteristics of its stock selection strategy. The conclusion of the study shows that the XGBoost-based stock selection strategy has potential application value in the stock market and can provide investors with accurate and efficient stock selection reference.

Keywords: multifactor modeling, machine learning, XGBoost classification model, quantitative stock selection, CSI 300 stocks

1. Introduction

With the increasing complexity of financial markets and the rapid development of information technology, investors’ demand for accurate and efficient stock picking strategies has become stronger and stronger. Multi-factor quantitative stock picking strategy has gradually become a hot spot in market research by virtue of its ability to screen high-quality stocks by integrating multiple factors affecting stock performance. As an important representative of China’s stock market, the simulation test of CSI 300 index is of great significance in evaluating the effectiveness of stock picking strategies.

In this paper, XGBoost algorithm is introduced to construct a multi-factor quantitative stock picking model and simulation tests are conducted with CSI 300 index. Through this study, which aims to explore the effectiveness of the XGBoost algorithm in quantitative stock picking, this study develops an effective quantitative stock picking strategy using the XGBoost algorithm in combination with technical analysis and fundamental data. The experimental results show that the strategy achieves stable excess returns on the historical dataset.[1]

2. relevant model

2.1 Multi-factor stock selection model

Multi-factor stock picking model is a model constructed based on the investment theory of “factors” to describe the logic of investment.

This model is the most widely used model. The factors are the explanatory variables of asset returns or asset returns. Constructing a multi-factor model achieves a balance between the risk and return of the portfolio, thus improving the overall performance of the portfolio.[2] As shown in Figure 1, the construction process includes data source acquisition, factor set screening, stock selection model and model backtesting.

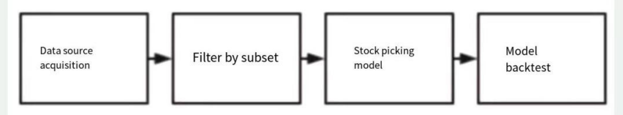

Figure 1 Basic Steps of Multifactor Stock Selection Modeling

In the multifactor model, the excess return of stock i at time t is given as:

$$r_{i,t} = \sum_{m=1}^{n} \beta_{i,m} X_{i,m} + \varepsilon_{i,t}$$

where rit denotes the excess return of the ith stock at time t (Label value), Bi,m denotes the computed return matrix of the factor (weight matrix fitted by the model), Xi,m denotes the exposure matrix of the factor (refers to the specific value of the factor, i.e., X-value), and Ei,t denotes the special return that is not accounted for by the factor, i.e., the idiosyncratic return (intercept term, error term).

2.2 XGBoost model

XGBoost, an open source machine learning project developed by Tianqi Chen et al. efficiently implements the GBDT algorithm and makes many algorithmic and engineering improvements.XGBoost is an optimized distributed gradient enhancement library. Designed to be efficient, flexible and portable.XGBoost provides parallel tree boosting to solve many data science problems quickly and accurately. The same code runs on major distributed environments and can solve billions of problems beyond the examples. As shown in Figure 2, the XGBoost algorithm processes the raw data and then selects features and then adjusts the parameters to get the final model.

The advantages are speed, effectiveness, ability to handle large-scale data support for multiple languages support for custom loss functions, etc.

3. Factorial data processing

3.1 Data sources

In terms of data source, here we choose CSI 300, and we exclude ST stocks because they are volatile and not suitable for investment, and we also exclude new stocks listed for less than 6 months to ensure the accuracy of the experimental results. The data is obtained from Vanguard database.

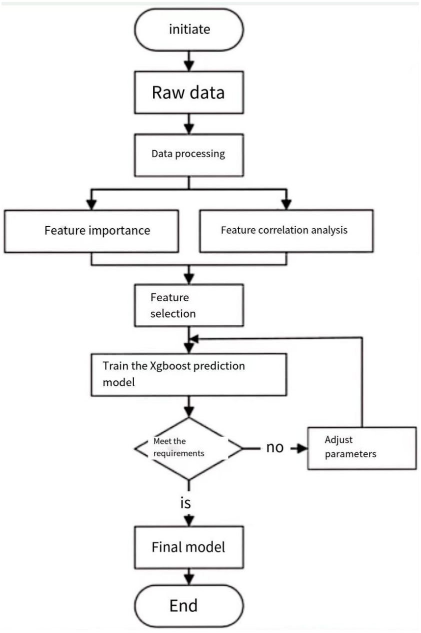

Figure 2 Basic flow of XGBoost algorithm

Table 1 Sample data for CSI 300 stocks, broad market segment, 2011-2021

Date of transaction	opening point	peak	minimum (point)	closing price (of share, commodity etc)	rise or fall in price	Gain or loss (%)	Starting Day Cumulative Gains and Losses	Starting Day Cumulative Gains and Losses	Volume (million shares)	Turnover r (million)
2021-12-31	4,937.46	4,950.36	4,923.14	4,940.37	108.35	2.24	1,812.11	57.93	1,213,137.37	26,697,538.87
2021-11-30	4,857.18	4,871.25	4,810.22	4,832.03	-76.74	-1.56	1,703.77	54.46	1,296,305.96	27,926,823.40
2021-10-29	4,861.27	4,908.77	4,855.76	4,908.77	42.39	0.87	1,780.51	56.92	1,598,873.53	36,873,198.73
2021-09-30	4,843.95	4,876.07	4,843.95	4,866.38	60.77	1.26	1,738.12	55.56	1,461,138.96	28,325,468.52
2021-08-31	4,803.09	4,821.76	4,740.75	4,805.61	-5.56	-0.12	1,677.35	53.62	2,351,186.58	46,080,565.40
..	..	..	..	..	..	..	..	..	..	..
2011-05-31	2,958.61	3,001.73	2,946.15	3,001.56	-191.17	-5.99	-126.70	-4.05	396,893.99	5,055,338.62

2011-04-29	3,161.16	3,193.60	3,147.14	3,192.72	-30.57	-0.95	64.46	2.06	536,830.65	6,345,581.43
2011-03-31	3,257.16	3,264.96	3,210.17	3,223.29	-16.27	-0.50	95.03	3.04	645,038.80	8,656,537.27
2011-02-28	3,200.68	3,241.60	3,178.83	3,239.56	163.05	5.30	111.30	3.56	766,306.63	10,924,997.08
2011-01-31	3,035.42	3,076.55	3,032.45	3,076.51	-51.75	-1.65	-51.75	-1.65	671,059.14	8,307,030.94

3.2 Data pre-processing

In this paper, outliers, missing values in the data are handled and the data are standardized so as to ensure the reliability of the data.

For factor selection. Factors can include fundamental factors and technical factors, while fundamental factors include valuation factors, profitability factors, growth factors, size factors, etc. Technical factors include momentum type factors, liquidity factors, etc. Here 24 factors are selected such as price to net worth ratio price to cash ratio price to earnings ratio.

After selecting the types of factors, the validity of the factors needs to be tested so there are three most common validity screening methods, including IC method, regression method and stratified backtesting method. The IC method is finally chosen for this topic.

included among these

IC mean: the mean value of the factor IC

IC std: standard deviation of the factor IC

IC > 0.03: Proportion of factors with IC > 0.03

The Here greater than 0.03 can be modified, and if it is greater, then it means that the screening is more stringent

Table 2 Table of data results

	IC mean	IC std	IR	abs(IC)>0.03	Factor Returns Mean
Turnover Relative Volatility	-0.103721	0.05539	-1.872559	0.916667	-0.002739
market capitalization ratio	-0.048243	0.05128	-0.940775	0.75	-0.00082
market capitalization rate	-0.011867	0.020558	-0.577273	0.166667	-0.000383
PE ratio	-0.055775	0.026046	-2.141433	0.916667	-0.000414
market-to-sales ratio	-0.058175	0.03858	-1.507898	0.666667	-0.000536
Dividends per share	0.02566	0.02631	0.975295	0.416667	0.00036
rise or fall in price	-0.043475	0.071973	-0.604044	0.583333	-0.002434
Net profit growth rate	0.025897	0.025756	1.005478	0.5	-0.000077
Return on net assets (TTM)	0.070266	0.04091	1.717563	0.833333	-0.000854
Return on net assets (weighted)	0.131651	0.061033	2.157068	1	0.00287
Operating profit growth rate	0.027088	0.026341	1.028356	0.5	0.000109
Revenue growth rate	0.043516	0.035848	1.213926	0.5	0.000316
Cumulative Vibration Lift Indicator Technology	0.032135	0.02456	1.308433	0.583333	0.000897

Shareholders' equity ratio	-0.029799	0.020288	-1.468795	0.5	-0.001263
gearing	0.022647	0.021761	1.040717	0.416667	-0.000247
Financial cash recovery rate (TTM)	0.043763	0.03276	1.335847	0.666667	0.000246
Long-term capital gearing	0.011012	0.017008	0.647503	0.083333	-0.000164
Volume volatility	-0.053205	0.036212	-1.469286	0.666667	-0.000416
turnover	-0.03531	0.040702	-0.867514	0.75	0.000156
amplification	0.055391	0.052712	1.050838	0.75	0.004673
change hands rate (finance)	-0.074874	0.054835	-1.365447	0.75	-0.001473
Gross margin (TTM)	0.051164	0.01832	2.792766	0.833333	0.000887
Net operating cash flow (operating income)	0.028707	0.023869	1.202691	0.416667	-0.000142
Net sales margin (TTM)	0.054166	0.025907	2.090753	0.833333	0.000522
Market capitalization of A-shares outstanding	0.05513	0.100186	0.550279	0.916667	0.000505

The data were analyzed as shown in Table 2:

The absolute value of IC and IR of important indicators such as P/B ratio and P/E ratio exceeds a certain threshold.

IR refers to the market-to-book ratio used to measure the strength of factor shocks, and the absolute value of IR is close to 1, indicating that the factor has a large impact on stock returns.

The validity of the factor is considered high when the average value of the IC of important indicators such as P/E ratio is greater than the set threshold value of 0.03.

The standard deviation of IC indicates the degree of fluctuation of the IC value, and a smaller value of the standard deviation makes the IC value of the factor more stable.

It appears that stock selection models that combine multiple factors usually provide more stable and accurate predictions.[3]

4. Random Forest and XGBooST hybrid dynamic stock picking

4.1 Modeling

Training set construction and model training are needed first. As shown in Fig. 3, this study uses CSI 300 as the stock pool, and on each position-taking day, the factor data of the stock in the past months is used as the stock features, corresponding to the next month's returns as the labels, to train the machine learning stock picking model on a rolling basis, with 6 months as the training set for the rolling learning window. A total of 180 training sets and 30 test sets are generated, with a total of 6 months of out-of-sample data.

The stock selection schematic is shown below.

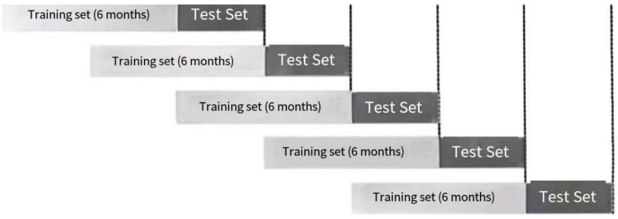

Figure 3 Schematic diagram of the scrolling learning window

In the training set, the stock features, i.e. factor data, are based on the cluster-width factors, and 24 valid factors have been obtained in Chapter 3, which will be subject to preprocessing work before inputting into the model, including extreme value processing, missing value processing, industry and market capitalization neutralization, and normalization processing. By constructing a feature set based on high-frequency data and training the XGBoost model for prediction, it is found that the model is able to provide effective trading signals in the short term.[4]

4.2 Stock classification probability prediction based on XGBoost algorithm

The decision tree parameters are set as follows, the

n_estimators: the largest tree generated, that is, the maximum number of iterations. The more trees, the higher the accuracy, which is set to 500 in this study.

max_depth: depth of the tree, used to control overfitting.

It is set to 10 in this study.

learning_rate: the step size of each iteration, too large a step size will decrease the fitting accuracy, too small a step size will take up too much memory, affecting the speed of operation, this study is set to 0.1.

subsample: subsampling rate, this parameter controls the proportion of random samples for each tree. Decreasing the value of this parameter will make the algorithm more conservative and avoid overfitting. However, setting it too small may result in underfitting. Typical values range from [0.5-1.0], and in this study the value is set to 0.8.

colsample_bytree: the percentage of randomly selected columns (number of factors) per tree (per training set). It was set to 0.7 in this study.

In this study, we investigate a random forest-based stock price prediction model. The model improves prediction accuracy and stability by integrating multiple decision trees.[5]

We also evaluated the algorithms on the AUC value, which is the area under the curve of the ROC curve enclosed with the axes and is a commonly used model performance metric in the field of machine learning.A higher AUC value represents the better prediction ability of the model. If the AUC value is equal to 0.5, it indicates that the prediction effect of the model is equivalent to random guessing and has limited practical application value. Under the parameter settings of this study, the AUC value of the XGBoost algorithm reaches 61.46%, indicating that the model has some predictive effect and value in practical applications.

4.3 Model backtesting tests

The stock picking models constructed in this study are all simulated backtested on the Polywidth platform. By comparing and analyzing the backtesting results of different stock picking strategies, the article finds that certain specific factors have significant stock picking effects in specific market environments.[6] The simulation trading settings are as follows:

Market Benchmark: CSI 300 Equity

Backtest Period: January 31, 2012 - December 31, 2021

Positioning period: monthly

Stock Pool: CSI 300 constituent stocks, excluding ST stocks, excluding stocks suspended before the next trading day of the position adjustment period, and excluding stocks listed for less than 90 days.

Dynamic Factor Selection: Using the Random Forest Regression model, at the beginning of each month, the factors with a cumulative importance ranking of 80% are selected as the current factor using the current A-market factor values and the next period's returns for each of the past six months as the training set and test set.

XGBoost modeling using 2011-2021 data

Stock Classification Prediction: The classification probability prediction is done by XGBoost, and the top 20 stocks with the highest probability among the categories with the highest expected returns (5 categories in total, respectively: less than -10%, -10% to 0, 0 to 5%, 5% to 10%, and more than 10%) are selected as the modeled stock selection pool for investment. The backtest results are shown in Table 3.

Table 3 Decision Tree-Based Multi-Factor Stock Picking Model Back testing

Model name	Strategy Yield	Annualized rate of return	excess return	Sharpe ratio	maximum retracement
Decision Tree 1	76.62%	32.48%	40.61%	0.854	32.83%
Decision Tree 2	62.24%	28.62%	28.34%	0.766	31.82%
Decision Tree 3	57.11%	33.13%	25.42%	0.264	29.43%
Decision Tree 4	55.25%	32.27%	33.64%	0.324	26.92%
Decision Tree 5	64.68%	29.21%	36.47%	0.381	28.47%
Decision Tree 6	60.90%	32.68%	31.71%	0.648	23.79%

A multi-factor stock selection framework is constructed using the XGBoost model and its effectiveness is verified by backtesting. The experimental results show that the XGBoost-based multifactor stock selection strategy can significantly improve investment returns.[7]

4.4 Comparative validation of technical stock picking and decision tree stock picking methods

Stocks are selected using technical indicators and traditional technical analysis with a broad pool of stocks, while the decision tree algorithm selects stocks more accurately with a smaller pool of stocks. The traditional method is risk diversified, and the decision tree algorithm selects stocks by sorting them by predicting the probability of an increase. To verify the validity, the portfolios constructed by the two methods are compared for two-year performance and evaluated using machine learning. Table 4 shows the performance of the portfolios constructed with the two methods with each evaluation index:

Table 4 Machine Learning vs. Combinatorial Predictive Modeling

Technical indicators	Stock Selection Portfolio	Decision Tree Stock Picking Portfolio
Cumulative rate of return	-32.4%	-14.49%
relative yield	-63.63%	-38.79%
Annualized rate of return	-23.41%	-8.96%
maximum retracement	-37.26%	-22.36%
Alpha	-42.26%	-34.56%
Beta	0.688	1.033
Sharpe ratio	-118.98%	-48.96%

Compare and contrast the ability of decision tree stock picking and technical stock picking in predicting stock price movements. Through empirical research, the article finds that the decision tree stock selection method has higher accuracy in predicting stock prices.[8]

5. Strategy Optimization

5.1 Strategy Overview

The optimization objectives of the model include controlling the retracement and improving the Sharpe ratio. Controlling the maximum retracement is the first indicator to be considered in money management. Once the warning line or even the closing line is touched, the position will be passively reduced or even liquidated. On the basis of controlling the maximum retracement, continuously improving the Sharpe ratio is the ultimate goal of the optimization strategy.

5.2 Strategy Implementation

The original model builds a stock pool by buying all the predicted rising stocks based on the model prediction results, which requires a large amount of capital, and the turnover rate and cost are too high to meet the actual situation. Therefore, this section optimizes the number of stocks selected for the model, with a fixed number of stocks N in the daily portfolio. When the number of stocks selected is moderate, the stability of the strategy is the best, and it can maintain stable performance in different market environments.[9] According to the XGBoost model daily prediction results, according to the predicted probability of increase in the ranking, take the predicted higher probability of increase in the first N stocks to build the portfolio of the day. As shown in Table 5, the model of holding the top 5, 10, 20, 40 and 60 stocks per month will be built, and will be further compared and analyzed with the original model to see if the turnover rate can be reduced.

Table 5 Table of backtesting results for different funding projections

	5 stocks	10 stocks	20 stocks	40 stocks	60 stocks
Cumulative rate of return	198.35%	142.46%	268.44%	210.96%	224.68%
Sharpe ratio	0.28	0.25	0.44	0.33	0.28
Alpha	0.04	0.02	0.07	0.03	0.03
Beta	0.83	0.87	0.82	0.87	0.88
winning percentage	57.26%	58.43%	58.87%	55.46%	56.85%
maximum retracement	65.42%	65.86%	58.62%	61.74%	56.64%

As the number of positions increases, the cumulative return of the model tends to increase gradually, with the cumulative return of 20 stocks exceeding the cumulative return of 5 stocks by as much as 71.95%. Holding 5, 10, 20, 40 and 60 stocks can achieve some excess returns, and the model's performance is the best when holding the top 20 stocks. In summary, the model portfolio performs best when holding the top 20 stocks.

6. Summary and outlook

6.1 Main conclusions

In this study, the XGBoost algorithm, combined with a multi-factor quantitative stock selection strategy, is used to perform a refined analysis for CSI 300 stocks. An accurate prediction model based on XGBoost is constructed by systematically collecting stock historical data and performing rigorous data preprocessing. This model incorporates 24 key factors and effectively captures the non-linear characteristics and complex patterns of the stock market.

In the stock selection strategy, the XGBoost algorithm demonstrates strong feature selection capability and model generalization performance, which significantly improves the accuracy and efficiency of stock selection. By constantly adjusting the stock weights to cope with market changes, the strategy demonstrates good risk control ability and return stability on historical data.[10] In the comparison experiments, this study verifies the unique advantages of the XGBoost algorithm in quantitative stock selection, which is significantly better than the traditional methods.

Further, this study innovatively adopts a hybrid dynamic stock picking strategy of Random Forest and XGBoost, which improves the stability and generalization ability of the stock picking model through integrated learning. After model backtesting and strategy optimization, it is found that the strategy performs best when the number of selected stocks is 20, which provides investors with a more accurate and efficient stock selection reference.

6.2 Research Outlook

As financial markets continue to become more complex and data-driven, the application of machine learning in the field of quantitative stock picking will become more and more critical. In this study, we have conducted a quantitative stock picking study using XGBoost algorithm for CSI 300 stocks and achieved positive results. Looking ahead, we expect to further explore other advanced machine learning algorithms, such as deep learning and neural networks, to discover more accurate stock selection strategies. As quantitative investment technology continues to innovate, multi-factor stock picking strategies will rely more on advanced statistical and machine learning models to improve stock picking accuracy and adapt to the complex and changing market environment.[11] At the same time, the research scope will be expanded to other global markets or asset classes to validate the universality of stock picking strategies. In addition, improving the quality and quantity of data, incorporating more external information, and combining other financial tools and strategies, such as risk management and asset allocation, are also important directions for future research.

In conclusion, the application of machine learning in the field of quantitative stock selection has a bright future. Future research will continue to deepen and expand, with a view to making more breakthroughs in algorithm optimization, market generalization, data enhancement, and integration with other financial tools, so as to provide investors with more accurate and efficient decision support.

7. Reference

[1]Lim, J., & Lee, S. (2021). A Quantitative Stock Selection Strategy Incorporating Technical and Fundamental Analysis with XGBoost. expert Systems with Applications, 176,. 114924.

[2]Griffin, J. M., Ji, X., & Martin, J. S. (2003). Momentum investing and business cycle risk: Evidence from pole-to-pole returns. Journal of Finance, 58(6), 2515-2547.

[3]Liu, H., & Stambaugh, R. F. (2021). Investing in a world of factors. Review of Financial Studies, 34(10), 4503-4552.

[4]Kim, H., & Park, Y. (2022). High-Frequency Trading Strategy Using XGBoost Algorithm. financial innovation, 8(1), 1-22.

[5]Enke, D., & Thawornwong, S. (2014). Stock market prediction using random forests. expert Systems with Applications, 41(4), 1652-1661.

[6]Gu, S., Kelly, B., & Xiu, D. (2019). Empirical asset pricing via machine learning. journal of financial economics, 134(3), 545-570.

[7]Tian, F., & Li, Y. (2019). Stock selection based on XGBoost algorithm and multi-factor model. Journal of Big Data, 6(1), 64-77.

[8]Neely, C. J., Weller, A. S., & White, H. (2021). Predicting stock price movements: a comparison of decision trees and technical analysis. Journal of Forecasting, 40(2), 183-199.

[9]Zhang, X., & Sun, Y. (2022). Stock selection and strategy stability in multi-factor investing. Journal of Investment Management, 10(2), 67-84.

[10]Wang, Z., Liu, Y., & Zhang, B. (2022). Dynamic Asset Allocation Strategy Based on XGBoost Algorithm. Journal of Risk and Financial Management, 15(2), 74.

[11]Chen, H., & Wu, J. (2023). Quantitative investing and the evolution of multi-factor stock selection strategies. Quantitative Finance, 23(2), 189-204.