


Article

A Sustainable Quantitative Stock Selection Strategy Based on Dynamic Factor Adjustment

Yi Fu ¹, Shuai Cao ¹ and Tao Pang ^{2,*} 

¹ School of Finance and Business, Shanghai Normal University, Shanghai 200234, China; fuyi@shnu.edu.cn (Y.F.); cs13914711039@163.com (S.C.)

² Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205, USA

* Correspondence: tpang@ncsu.edu

Received: 7 April 2020; Accepted: 11 May 2020; Published: 13 May 2020



Abstract: In this paper, we consider a sustainable quantitative stock selection strategy using some machine learning techniques. In particular, we use a random forest model to dynamically select factors for the training set in each period to ensure that the factors that can be selected in each period are the optimal factors in the current period. At the same time, the classification probability prediction (CPP) of stock returns is performed. Historical back-testing using Chinese stock market data shows that the proposed CPP quantitative stock selection strategy performs better than the traditional machine learning stock selection methods, and it can outperform the market index over the same period in most back-testing periods. Moreover, this strategy is sustainable in all market conditions, such as a bull market, a bear market, or a volatile market.

Keywords: stock selection; machine learning; classification probability prediction; back-testing

1. Introduction

In modern investing, algorithmic trading is getting more and more attention from individual and institutional traders. “Algorithmic trading is a method of executing orders using automated pre-programmed trading instructions accounting for variables such as time, price, and volume” (https://en.wikipedia.org/wiki/Algorithmic_trading). It considers market observable variables such as time, price, and volume, and sends instructions to the market based on a preset algorithm. Algorithmic trading, on the one hand, can prevent traders from frequently repeating observations and manually sending trading instructions; on the other hand, it can also prevent traders’ decisions from being disturbed by subjective emotions. According to a May 2019 report from Research and Markets, “The researchers forecast the global algorithmic trading market size to grow from USD 11.1 billion in 2019 to USD 18.8 billion by 2024, at a CAGR of 11.1% during 2019–2024. The major growth drivers of the algorithmic trading market include the increasing demand for fast and effective order execution, and reducing transaction costs” (<https://www.researchandmarkets.com/reports/4770543/>).

With the development of new technologies such as machine learning, the current algorithmic trading not only includes automatic sending of transaction instructions, but also includes the automatic decision-making of the algorithm in terms of transaction time, transaction objects, and number of transactions. Quantitative stock selection, as an important part of in algorithmic trading, focuses on using various algorithms to select stock combinations in order to achieve a benchmark return rate.

Quantitative stock selection is a popular academic research area. Fama and French (1993) [1], Lakonishok (1994) [2], and Song (1994) [3] established a linear model of stock excess returns, and proposed that the excess returns can be well explained by current stock prices, book value of equity, and earnings per share. Compared with the classic linear multi-factor models, the machine learning model pays more attention to the prediction ability of the model. It can capture more detailed market signals

and obtain more stable excess returns by constructing a nonlinear relationship between the prediction target and the factors. Jigar Patel et al. (2015) [4] studied and compared the performance of the four prediction models artificial neural network (ANN), support vector machine (SVM), random forest (RF), and Naive-Bayes. Their results show that the overall performance of the random forest model is better than the other three prediction models. Liu et al. (2017) [5] proposed a convolutional neural network and long-short-term memory (CNN-LSTM) model to analyze the quantitative strategy of the stock selection. In their study, the CNN-LSTM neural network model could be successfully applied to the formulation of quantitative strategies and achieve better returns than basic momentum strategies and benchmark indexes. Li and Zhang (2018) [6] used the XGBoost model to establish a dynamic weighted multi-factor stock selection strategy. They used the XGBoost machine learning method to predict the information coefficients (ICs) of various factors. The empirical results showed that the XGBoost model is effective in predicting the ICs, and the dynamic weights based on the XGBoost model can improve the performance of multi-factor stock selection strategies. Yang and Chen (2019) [7] combined stock forecasting and stock selection to form a new hybrid stock selection method. Based on the research sample of the A-share stock market in China, they showed that the novel hybrid method is superior to the traditional methods in market returns. Chen and Ge (2019) [8] studied the stock price movement prediction based on LSTM networks, and compared the attention LSTM (AttLSTM) model with the LSTM model. Their results verify the effectiveness of the attention mechanism in the LSTM-based prediction method.

Although a lot of works on quantitative models and processes have been done, there are still some areas that can be improved. First of all, in the setting of prediction targets, previous studies often used the stock return or whether the price is up or down as the prediction target, but the return rate often contains some noise, and the setting of the two classifications (up or down) does not catch much of the existing information. Secondly, in factor selection, previous studies often selected factors statically, but factors are usually valid for a certain period of time, and may not be valid after that. Therefore, the entire strategy design needs to select factors dynamically, e.g., eliminating failed ones, and introducing effective ones.

In this paper, we propose a sustainable quantitative stock selection strategy using RF to dynamically adjust the factors to predict the importance of the training set for each period. The factors are sorted in descending order. The cumulative importance of the selected factors must reach 80% to ensure that the factors selected in each period are the most important factors. Then, we use the XGBoost or RF model to classify each stock into five fixed yield ranges. For each yield range, we sort the stocks in descending order of probability, and take the top 20 most likely stocks into the stock pool for purchase. We call this a classification probability of prediction (CPP) strategy. The back-testing results from November 2013 to December 2019 show that the stock selection strategy of the XGBoost or RF CPP method can significantly outperform the Chinese Stock Index 300 (CSI 300) index. Moreover, we find that the XGBoost CPP performs better than the RF CPP method in terms of returns. Finally, the proposed strategy is a sustainable investment strategy in the sense that it works well over a long time period that consists of bear market, bull market, and volatile market periods.

2. The Basic Idea of CPP Quantitative Stock Selection Strategy Design

The general steps of the CPP quantitative stock selection strategy design were as follows (see also Figure 1).

The first step was to use all stocks in the Chinese A-share market (exclude special treated “ST” stocks and new stocks listed less than 60 days) as the stock pool, and classify the stock based on their monthly rate of return. In particular, we classified each stock into five ranges (see Table 1). We considered nine broad categories: quality factors, fundamental factors, emotional factors, growth factors, risk factors, stock factors, momentum factors, technical factors, and style factors. Then, we selected 45 factors from 9 categories as the initial factor pool. The factors in this article came

from JoinQuant’s factor library. Table 2 shows the 45 factors in the model factor pool of this article. These factors were dynamically screened into the model by the random forest (RF) model.

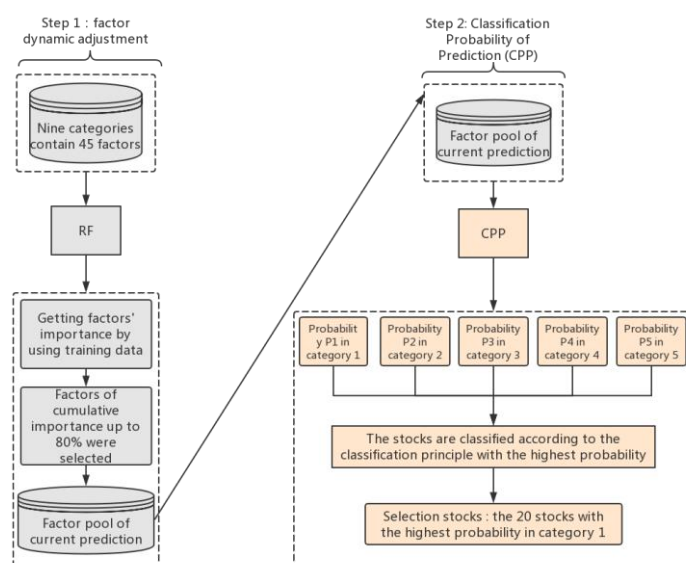


Figure 1. Flow diagram of factor dynamic adjustment and classification probability of prediction (CPP) quantitative stock selection strategy.

Table 1. Range criteria (monthly rate of return).

	Range 1	Range 2	Range 3	Range 4	Range 5
Criteria	above 10%	5–10%	0–5%	–10–0%	–10% or less

Table 2. Factors list.

No.	Classification	Factors	No.	Classification	Factors
1	Quality factor	net_profit_to_total_operate_revenue_ttm	24	Risk factor	Skewness20
2	Quality factor	DEGM	25	Risk factor	sharpe_ratio_60
3	Quality factor	roe_ttm	26	Stock factor	net_asset_per_share
4	Quality factor	GMI	27	Stock factor	net_operate_cash_flow_per_share
5	Quality factor	ACCA	28	Stock factor	eps_ttm
6	Fundamental factor	financial_liability	29	Stock factor	retained_earnings_per_share
7	Fundamental factor	cash_flow_to_price_ratio	30	Stock factor	cashflow_per_share_ttm
8	Fundamental factor	market_cap	31	Momentum factor	ROC20
9	Fundamental factor	net_profit_ttm	32	Momentum factor	Volume1M
10	Fundamental factor	EBIT	33	Momentum factor	TRIX10
11	Emotional factor	VOL20	34	Momentum factor	Price1M
12	Emotional factor	DAVOL20	35	Momentum factor	PLRC12
13	Emotional factor	VOSC	36	Technical factor	MAC20
14	Emotional factor	VMACD	37	Technical factor	boll_down
15	Emotional factor	ATR14	38	Technical factor	boll_up
16	Growth factor	PEG	39	Technical factor	MF14
17	Growth factor	net_profit_growth_rate	40	Style factors	size
18	Growth factor	operating_revenue_growth_rate	41	Style factors	beta
19	Growth factor	net_asset_growth_rate	42	Style factors	momentum
20	Growth factor	net_operate_cashflow_growth_rate	43	Style factors	book_to_price_ratio
21	Risk factor	Variance20	44	Style factors	liquidity
22	Risk factor	sharpe_ratio_20	45	Style factors	growth
23	Risk factor	Kurtosis20			

In the second step, the training and test sets were constructed by recombining the factors and yield intervals of each period. In particular, the period $i-3$ factor was combined with the monthly rate of return for period $i-2$, the period $i-2$ factor was combined with the monthly rate of return for period $i-1$, and the period $i-1$ factor was combined with the monthly rate of return of the period i . All together were combined to construct the training set of the period i . The factors of the period i and the monthly rate of return of the period $i+1$ constructed the test set of period i . See Figure 2 for illustration.

	i-3	i-2	i-1	i	i+1	i+2	...
Yield Rate							
Factors		Train			Test		

Figure 2. Training data and test data construction.

In the third step, we used an RF model to predict the importance of factors for each training set, and sort the importance in descending order. We chose the most important factors to ensure that the cumulative importance of the selected factors reached 80%. As the factors had their own validity periods, the IC values of the factors in different periods were not completely unchanged. As shown in Figures 3 and 4, the IC values of the factors ATR14 and EBIT have changed in different periods. Therefore, the factors applicable to different time periods are also different. For this reason, we used dynamic factor selection to select the most important factor in the current period and improve the accuracy of stock selection.

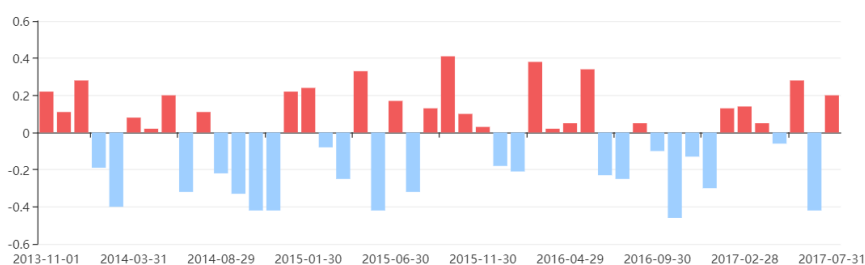


Figure 3. IC in ATR 14 (Data Source: JoinQuant platform).

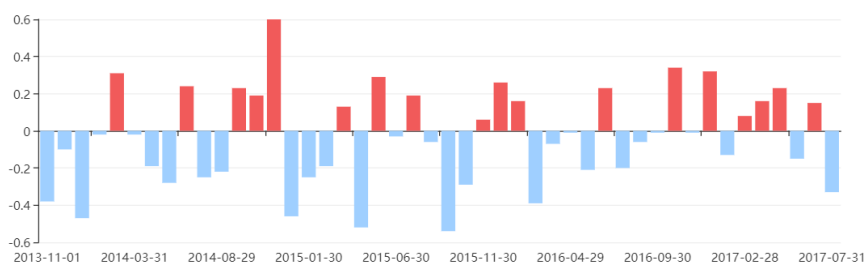


Figure 4. IC in EBIT (Data Source: JoinQuant platform).

The fourth step was to use XGBoost CPP method to predict the classification (the previous month's factor predicts the monthly yield range), and classify each stock into five yield ranges based on the factors dynamically selected in the third step. The stocks in the group yield range were sorted in descending order of probability, and the top 20 stocks with the highest probability were taken into the buying stock pool. On the last trading day of each month, the position was adjusted. When the position was adjusted, the stocks that were not in the buying stock pool are sold, and new stocks in the buying stock pool were bought. Then, we looped into the training set for the next period.

The CPP quantitative stock selection strategy with dynamic factor adjustment has some obvious advantages. The core of quantitative investments is the model, and the core of the model is the factor. This is particularly true in the neutral Alpha strategy with huge market capacity. Therefore, how to find a stable and effective factor becomes the first step in developing a mature profitable quantitative strategy. The random forest (RF) model is an ensemble learning method for classification, regression, and other tasks (https://en.wikipedia.org/wiki/Random_forest). The RF model can not only effectively correct the overfitting problem in the decision tree model, but also give the importance of each input variable (importance). In 1995, Ho proposed the RF algorithm [9], and some scholars extended the algorithm and conducted subsequent research (see, e.g., Breiman [10] and Lin and Jeon [11]). In this

paper, we used the RF model to predict the importance of the factors in the training set, and rank the importance of the factors in descending order. Then, we selected the cumulative importance of the factors to reach 80%, ensuring that the factors in each period were the optimal choices. By doing that, we enhanced the impacts of the factors.

To the best of our knowledge, most quantified stock selection strategies based on machine learning use the regression method to accurately predict the future return of the stock, and then buy stocks with high predicted returns. The fitted stock selection method seems to be more accurate than the multi-class probability prediction stock selection method, but its fault tolerance is relatively low. Once a prediction error occurs, it will have a greater impact on the overall return. Moreover, the noise in the yield is usually large, and the probability of regression errors is usually high. Therefore, it is easy to cause a large maximum retracement. The proposed multi-class probability prediction stock selection strategy is not to select the stock with the highest predicted return rate, but to select the stock with the highest probability of return in this range after the determined expected return range. Although some of the benefits are sacrificed in this way, the accuracy rate and fault tolerance rate are both improved, and with the increase of the accuracy rate, some of the sacrificed benefits will also be made up.

3. Back-test Analysis of CPP Quantitative Stock Selection Strategy

In this section, we conduct 74 back-testing analyses of market data from November 2013 to December 2019. The data source was from the JoinQuant quantization platform.

The goal of the stock selection was to achieve a high return, and we did not limit the investment strategies to any particular investment style. Therefore, it was natural to use the overall market return as the benchmark. In this paper, we chose the CSI 300 index as the benchmark.

3.1. Dynamic Factor Adjustment Analysis

Among the 45 factors, the style category was most likely to be selected (see Table 3). The liquidity factor (liquidity) had a probability of being selected as high as 98.65%. The market value factor (size) was selected with probability 94.59% and the beta factor (beta) was selected with probability 68.92%. There were three growth type factors in the top ten factors, where the net asset growth rate (net_asset_growth_rate) had a selection probability of 95.95%, the net profit growth rate (net_profit_growth_rate) had a selection probability of 79.73%, and the price-earnings (P/E) ratio relative to the earnings growth ratio (PEG) had a selection probability of 71.62%. There were two risk type factors in the top ten. In particular, the 20-day annualized return variance (Variance20) was selected with a probability of 95.95%, the 20-day Sharpe ratio (sharpe_ratio_20) was selected with a probability of 74.32%. Finally, there was one emotion factor and one momentum factor among the top ten factors, where the trading volume shock (VOSC) was selected with a probability of 93.24%, and Price1M was selected with a probability of 90.54%.

Table 3. Choosing the TOP10 factors with the highest probability.

Rank	Factor	Selected Times	Total Times	Selected Probability
1	liquidity	73	74	98.65%
2	Variance20	71	74	95.95%
3	net_asset_growth_rate	71	74	95.95%
4	size	70	74	94.59%
5	VOSC	69	74	93.24%
6	Price1M	67	74	90.54%
7	net_profit_growth_rate	59	74	79.73%
8	sharpe_ratio_20	55	74	74.32%
9	PEG	53	74	71.62%
10	beta	51	74	68.92%

The market value factor considered here is not the same as the traditional market value factor. It refers to the natural logarithm of the company's total market value. The formula of liquidity factor is given by:

$$\text{Liquidity Factor} = 0.35 \times \text{STOM} + 0.35 \times \text{STOQ} + 0.3 \times \text{STOA}, \quad (1)$$

where STOM is the stock turnover rate in one month, given by the logarithm of the sum of stock turnover rates in the past 21 days; STOQ is the average turnover rate in the past three months, given by the logarithm of the average STOM in the past three months; and STOA is the average turnover rate in the past 12 months, given by the logarithm of the average STOM in the past 12 months. The formula for net asset growth rate is given by:

$$\text{Net asset growth rate} = \frac{\text{shareholder equity for the current quarter}}{\text{shareholder equity before the third quarter}} - 1. \quad (2)$$

3.2. Back-testing Revenue

In this section, we compare and analyze the benefits under different back-testings. See Table 4 for parameter settings.

Table 4. Parameter settings for policy back-testing.

Item	Detail
Object of transaction	all stocks after screening (excluding ST shares, new shares, secondary shares, and stocks suspended within 20 days)
Returns of the benchmark	Index gains for the CSI 300
Time of back-testing	1 November 2013 (Fri.) to 31 December 2019 (Tue.)
Days of back-testing	1507 trading days
Data sources	JoinQuant quantitative investment platform
Initial funding	10 million
Overnight or not	yes
Stop's way	RSRS stop loss
Number of the position	20 stocks
Adjustable frequency	one month
Slippage	0.2%
Commission charge	0.03% commission when buying, 0.03% commission plus 0.1% stamp duty when selling, commission for each transaction a minimum deduction of 5 yuan
Software language	Python

3.2.1. XGBoost Classification Prediction and XGBoost regression Prediction

In 2015, the XGBoost model was proposed by Chen et al. [12], which is optimized for fast parallel tree construction. "It has gained much popularity and attention recently as the algorithm of choice for many winning teams of machine learning competitions (<https://en.wikipedia.org/wiki/XGBoost>)". Because of the XGBoost model's good performance, we chose the XGBoost model to predict the stock's return rate.

The core model of this paper is the XGBoost multi-class prediction model, and the model parameters are shown in Table 5. We used the XGBoost multi-class prediction model to perform back-testing from November 2013 to December 2019. A total of 74 class predictions were carried out. The comprehensive evaluation of the prediction is shown in Table 6. Among them, accuracy, sensitivity C1, and precision C1 are defined similar to those for the two-class classification. The specific formulas are given by Equations (3)–(5), where x_{ij} is given in Table 7.

$$\text{accuracy} = \frac{\sum_{i=1}^5 x_{ii}}{\sum_{i=1}^5 \sum_{j=1}^5 x_{ij}} \quad (3)$$

$$\text{sensitivity C1} = \frac{x_{11}}{\sum_{i=1}^5 x_{i1}} \quad (4)$$

$$\text{precision C1} = \frac{x_{11}}{\sum_{j=1}^5 x_{1j}} \quad (5)$$

Table 5. Main parameters of XGBoost classification.

Parameter	Value
max_depth	10
learning_rate	0.1
n_estimators	500
min_child_weight	5
colsample_bytree	0.7
reg_lambda	0.4
scale_pos_weight	0.8
subsample	0.8

Table 6. XGBoost Multi-class prediction evaluation.

Item	Mean	Stdev	Max	Min
accuracy	51.7%	7.9%	67.0%	37.5%
sensitivity C1	75.4%	7.8%	87.7%	59.3%
precision C1	62.1%	10.3%	78.2%	41.2%

Table 7. Confusion matrix of 5 classification model.

		True Condition				
		Category1	Category2	Category3	Category4	Category5
Predicted Condition	Category1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
	Category2	x_{21}	x_{22}	x_{23}	x_{24}	x_{25}
	Category3	x_{31}	x_{32}	x_{33}	x_{34}	x_{35}
	Category4	x_{41}	x_{42}	x_{43}	x_{44}	x_{45}
	Category5	x_{51}	x_{52}	x_{53}	x_{54}	x_{55}

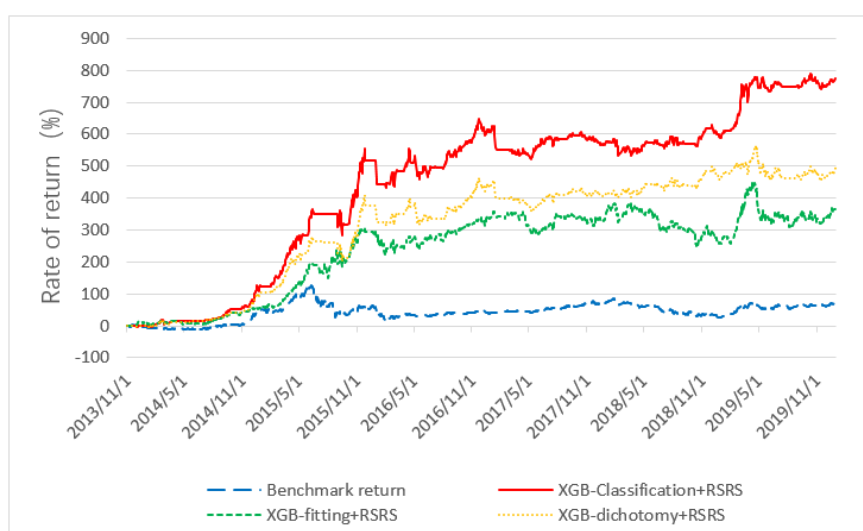
The stock selection criterion is to hold stocks that are predicted to be in the first category and are ranked in the top 20 in probability. Therefore, sensitivity C1 and precision C1 are more important for evaluating the prediction ability. Among them, sensitivity C1 represents the proportion of stocks that can be correctly predicted in the first category of stocks, and precision C1 represents the proportion of stocks that are truly in the first category. In the 74 predictions, the mean value of sensitivity C1 was 75.4% and the standard deviation was 7.8%; the mean value of precision C1 was 62.1% and the standard deviation was 10.3%. The average accuracy of the 74 predictions was 51.7% and the standard deviation was 7.9%. Although the overall accuracy was not very high, this indicator had little effect on the overall performance in terms of back-testing returns. We believe that the precision C1 indicator is the most important of the three indicators. The higher value of this indicator indicates that the model can screen out high-yield stocks with a high probability.

Next, the comparison between XGBoost classification prediction and XGBoost regression prediction was performed. In XGBoost classification prediction, we used the XGBoost model to predict the return rate range of each period of the back-testing stage; that is, to carry out multi-class prediction. In XGBoost regression prediction (parameters are given in Table 8), we predicted the return rate value of each period of the back-testing stage, that is, regression the yield, and holding the 20 stocks with the highest predicted returns. Both methods use the RSR index (relative strength of resistance support) stop-loss module to stop the loss.

Table 8. Main parameters of XGBoost regression.

Parameter	Value
max_depth	10
learning_rate	0.3
gamma	0.1
min_child_weight	3
colsample_bytree	0.7
lambda	3
subsample	0.5

As shown in Figure 5 and Table 9, the performance of the quantitative stock selection strategy based on the XGBoost multi-class prediction was much better than the CSI 300 Index in the back-testing interval from November 2013 to December 2019. In terms of the annualized yield, Sharpe ratio, maximum retracement, and Calmar ratio, the performances of the XGBoost multi-class prediction method were significantly better than the quantitative stock selection strategy based on XGBoost regression and XGBoost two-class classification in the same period. Therefore, we believe that the quantitative stock selection strategy of XGBoost multi-class probability prediction has a better back-testing performance.

**Figure 5.** Back-testing earning chart of regression and classification stock selection.**Table 9.** Stock selection strategy back-testing indicators of XGBoost-regression and classification.

+	Model and Stop-Loss		
	XGBoost-Regression + RSRS	XGBoost-Classification + RSRS	XGBoost-Dichotomy + RSRS
Annual yield rate	0.26	0.57	0.36
Accumulated yield rate	3.65	7.76	4.91
Annualized Volatility	0.25	0.23	0.24
Sharpe Ratio	0.62	2.21	1.42
Calmar Ratio	0.90	2.71	1.33
Stability_of_timeseries	0.71	0.84	0.57
Maximum Drawdown	0.29	0.21	0.27
Sortino Ratio	0.90	3.03	1.81
Information Ratio	0.96	1.92	1.05
Alpha	0.16	0.54	0.31
Beta	0.70	0.36	0.58

3.2.2. Back-testing Revenue of Different Models

Next, in order to compare the combined back-testing effects of different models and stop-loss modules, we compared the performances of different combinations of the XGBoost and random forest decision-making models (parameters of the RF model are given by Table 10) with the RSRS index (relative strength of resistance support) stop-loss module and the MACD (moving average of similarities and differences) stop-loss module. The back-testing results are given in Figure 6 and Table 11.

Table 10. Main parameters of random forest.

Parameter	Values
max_depth	5
min_samples_leaf	2
n_estimators	200
min_samples_split	2
criterion	gini

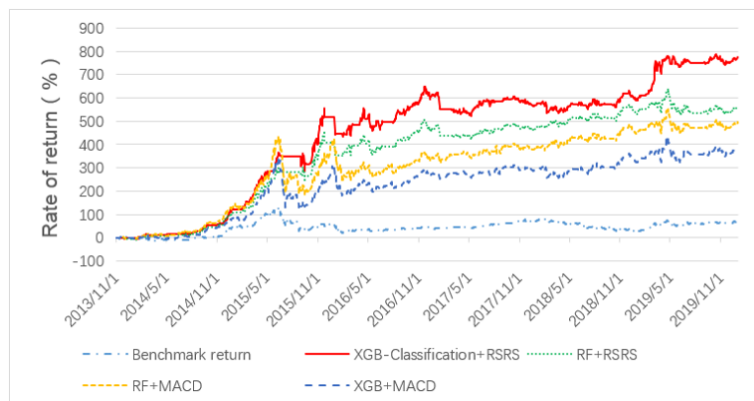


Figure 6. The back-testing for return rate.

Table 11. The back-testing index of different models.

Index	Model and Stop-Loss			
	XGBoost + RSRS	RF + RSRS	RF + MACD	XGBoost + MACD
Annual yield rate	0.57	0.41	0.37	0.28
Accumulated yield rate	7.76	5.63	4.96	3.84
Annualized Volatility	0.23	0.24	0.33	0.35
Sharpe Ratio	2.21	1.54	1.02	0.65
Calmar Ratio	2.71	1.86	0.71	0.51
Stability_of_timeseries	0.84	0.82	0.72	0.70
Maximum Drawdown	0.21	0.22	0.51	0.55
Sortino Ratio	3.03	2.57	1.01	0.92
Information Ratio	1.92	1.26	1.18	0.98
Alpha	0.54	0.41	0.24	0.17
Beta	0.36	0.48	0.83	0.79

As shown in Figure 6 and Table 11, the back-testing benefit of the combination of the XGBoost model and the RSRS index stop loss module was higher than that of the random forest model. This indicates that, under the timing given by the RSRS index stop loss module, the XGBoost multi-class probability prediction is more accurate than the random forest model. However, under the timing given by the MACD stop loss module, the return of the XGBoost model was lower than that of the random forest model. In the case of the same machine learning model, the effect of the RSRS index

stop loss module is significantly stronger than the MACD stop loss module. Therefore, we decided to choose the combination of XGBoost model and RSRS index stop loss module as the main model of CPP quantitative stock selection strategy.

For the CPP quantitative stock selection strategy proposed in this paper, the annualized return reached 57%, the Sharpe ratio was 2.21, the maximum drawdown was 21%, the Calmar ratio was 2.71, and the win rate was 63.5%. The return of the strategy reached the lowest value of -3.85% on 10 January 2014, and reached the highest point on 14 October 2019 when cumulative gain of the strategy was 788.52%. Since 19 December 2013, the cumulative returns of CPP's quantitative stock selection strategy have been better than the CSI 300 Index over the same period.

3.2.3. CPP Quantitative Stock Selection Back-Testing Income

After determining that the main model is a combination of the XGBoost multi-class forecast and the RSRS index stop loss module, this paper conducted back-testing in the back-testing interval from 1 November 2013 to 31 December 2019, and the results were given in Figure 7 and Table 12.

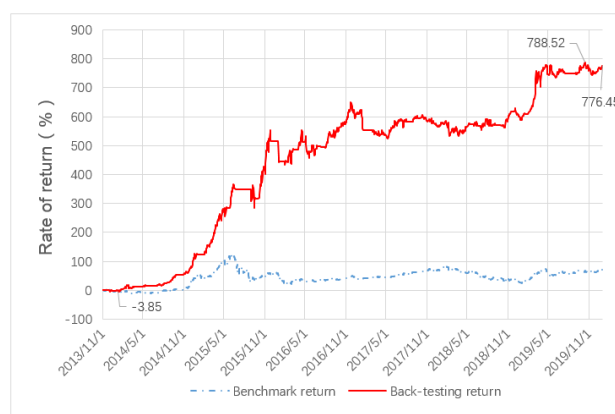


Figure 7. Chart of CPP stock selection back-testing return rate.

Table 12. Excess returns of CPP quantitative stock selection strategy at different time periods.

Different Period	State of Market	Excess Rate of Return
1 November 2013–31 August 2014	volatile market	30.29%
1 September 2014–31 May 2015	bull market	94.40%
1 June 2015–31 December 2015	bear market	80.58%
1 January 2016–31 December 2019	volatile market	86.63%

In different periods of the market, the applicable strategies will be different, and it is difficult for a strategy to perform well in all periods. The CPP quantitative stock selection strategy has different levels of excess returns at different time periods. As shown in Table 12 and Figure 7, from 1 November 2013 to 31 August 2014, a horizontal price movement period (volatile market) before the bull market, the CPP quantitative stock selection strategy achieved an excess yield of 30.29% during this 10-month period. From 1 September 2014 to 31 May 2015, the CPP quantitative stock selection strategy achieved an excess return of 94.4%. From 1 June 2015 to 31 December 2015, after the stock market crashed sharply, the CPP quantitative stock selection strategy achieved an excess return of 80.58%. From 1 January 2016 to 31 December 2019, another horizontal price movement period (volatile market), the CPP quantitative stock selection strategy achieved an excess return of 86.63%. As we can see, the proposed CPP quantitative stock selection strategy is a sustainable investment strategy that works well over an extensive period that covers bull market, bear market, and volatile market states.

4. Conclusions

In this paper, we used a random forest model to dynamically select factors for the training set in each period to ensure that the factors that could be selected in each period were the optimal factors in the current period. At the same time, the classification probability prediction (CPP) of stock returns was performed. This method can effectively take into account the accuracy of income prediction and avoid the interference of noise in the rate of return. Historical back-testing shows that the CPP quantitative stock selection strategy based on dynamic factor adjustment performs better than the traditional machine learning stock selection methods, and can outperform the CSI 300 Index over the same period in most back-testing periods. It is a sustainable investment strategy in the sense that, no matter in a bull market, a bear market, or a volatile market state, the CPP quantitative stock selection strategy based on dynamic factor adjustments can achieve better excess returns.

It should be noted that all the results in this article were derived from historical data back-testing, and the results may be different from the results of actual investments. As we used the historical data for back-testing, we did not consider the impacts of the market liquidity, and the impacts of this strategy on the decisions of other market participants, etc. Therefore, there is no guarantee that the strategy works for real market investments. We are not responsible for any loss caused by implementing the strategy.

Author Contributions: Conceptualization, Y.F. and T.P.; writing—original draft preparation, Y.F., T.P. and S.C.; writing—review and editing, Y.F. and T.P.; software, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by MOE (Ministry of Education in China) Youth Project of Humanities and Social Sciences (Project No. 17YJCZH044), MOE (Ministry of Education in China) Project of Humanities and Social Sciences (Project No. 18YJAZH127) and The 10th Key Discipline of Shanghai Normal University: Quantitative Economics.

Acknowledgments: Data and back-testing are based on JoinQuant Quantization Platform (<https://www.joinquant.com/>).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fama, E.F.; French, K.R. Business Conditions and Expected Returns on Stocks and Bonds. *J. Financ. Econ.* **1989**, *25*, 23–49. [[CrossRef](#)]
2. Lakonishok, J.; Shleifer, A.; Vishny, R.W. Contrarian Investment, Extrapolation, and Risk. *J. Financ.* **1994**, *49*, 1541–1578. [[CrossRef](#)]
3. Song, F.M. A Two-Factor ARCH Model for Deposit-Institution Stock Returns. *J. Money Credit Bank.* **1994**, *26*, 323–340. [[CrossRef](#)]
4. Patel, J.; Shah, S.; Thakkar, P.; Kotecha, K. Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques. *Expert Syst. Appl.* **2015**, *42*, 259–268. [[CrossRef](#)]
5. Liu, S.; Zhang, C.; Ma, J. CNN-LSTM Neural Network Model for Quantitative Strategy Analysis in Stock Markets. In Proceedings of the International Conference on Neural Information Processing, Guangzhou, China, 14–18 November 2017; pp. 198–206.
6. Li, J.; Zhang, R. Dynamic Weighting Multi Factor Stock Selection Strategy Based on XGboost Machine Learning Algorithm. In Proceedings of the 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), Chongqing, China, 10–12 December 2018.
7. Yang, F.; Chen, Z.; Li, J.; Tang, L. A Novel Hybrid Stock Selection Method with Stock Prediction. *Appl. Soft Comput.* **2019**, *80*, 820–831. [[CrossRef](#)]
8. Chen, S.; Ge, L. Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction. *Quant. Financ.* **2019**, *19*, 1507–1515. [[CrossRef](#)]
9. Ho, T.K. Random Decision Forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995.
10. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]

11. Lin, Y.; Jeon, Y. Random Forests and Adaptive Nearest Neighbors. *J. Am. Stat. Assoc.* **2006**, *101*, 578–590. [[CrossRef](#)]
12. Chen, T.; Guestrin, C. XGBoost: Reliable Large-Scale Tree Boosting System. In Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 13–17.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).