Empirical Research About Quantitative Stock Picking Based on Machine Learning

Zheng Zhongbin[1, a], Fang Jinwu[1, b]

[1]Hanghui Building, No.600 Yunjin Road, Xuhui District, Shanghai, China, 200232

[a] zhengzhongbin@caict.ac.cn, [b] fangjinwu@caict.ac.cn

Keywords: Machine learning; *Random Forest*; *XGBoost*; Multifactor stock selection.

Abstract. This study mainly uses artificial intelligence and machine learning technology to build stock selection models to help investors choose stocks reasonably. In this paper, six machine learning models were constructed for comparison and backtesting based on the framework of the machine learning stock selection. By comparing the model classification accuracy, AUC, and other index, *XGBoost* and *Random Forest* were selected, and the portfolio was constructed. According to the analysis, the portfolio could obtain an above-average rate of return, and the portfolio obtained a net value of about 1.5 times that of the benchmark portfolio during the two-year investment test period.

1. Introduction

Stock selection is a very important and challenging topic for investors and researchers. However, due to the uncertainty of the stock market, stock forecasting is often difficult, quantitative investment is increasingly attracting investors. In recent years, the development of artificial intelligence technology has brought new opportunities and research directions to the field of quantitative investment. Among them, quantitative stock selection is an important part of quantitative investment theory. It uses mathematical statistics methods and public historical data according to the needs of investors to screen out a number of stocks from large stock pools and allocate assets to win.

2. Calculation method

In this paper, a total of six stock selection classification models were constructed, including the most commonly used support vector machines and neural networks, naive Bayesian model based on probability statistics, decision tree-based reinforcement algorithm stochastic forest and *XGBoost*, and *Logistic Regression* based on linear learning.

2.1 Principal Component Analysis model

Principal Component Analysis (PCA) is to try to replace the original indicators with a number of original correlations that are regrouped into a new set of mutually independent indicators.

2.2 Logistic Regression model

Logistic Regression model is used to estimate the probability of a response variable based on one or more predictors. The basic form of *Logistic Regression* is:

$$\mathrm{P}(y=1|x)=\frac{e^{w_0+w_1x_1+w_2x_2+\cdots+w_px_p}}{1+e^{w_0+w_1x_1+w_2x_2+\cdots+w_px_p}} \quad (1)$$

2.3 Naive Bayesian model

Naive Bayesian model is a classification method based on Bayes' theorem and the independent assumption of feature conditions.

$$\mathrm{P}(\mathrm{y}|\mathrm{X})=\frac{P(X|y)*P(y)}{P(X)} \quad (2)$$

2.4 Artificial Neural Network model

Artificial Neural Network (ANN) is a nonlinear complex network formed by a large number of neurons connected through each other.

2.5 Decision Tree and Random Forest model

Decision Tree method is a tree-structured machine learning algorithm, which is continuously classified according to the characteristics of the feature factors.

Random Forest model utilizes the Bagging method, which is equivalent to the parallel connection of some decision tree classifiers.

2.6 XGBoost model

XGBoost combines multi-choice decision trees in tandem to improve the Gradient Boosting Decision Tree (*GBDT*).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

3. Model test results and comparison

3.1 The model comparison and selection

In addition to the simple models of *Logistic* and *Nbayes*, the classification accuracy of the other four models is 55%. Among them, the *Random Forest* performed best, the correct rate of cross validation set reached 68%, and the test set reached 63%. The AUC situation for each model classification result is similar to the classification accuracy rate.

Table 1. The comparison of correct rate and AUC of different models

Accuracy Rate			AUC		
Model	Cross validation set	Test set	Model	Cross validation set	Test set
Logistic	0.62	0.53	*Logistic*	0.62	0.55
Nbayes	0.52	0.56	*Nbayes*	0.52	0.52
ANN	0.68	0.59	*ANN*	0.68	0.59
RF	0.68	0.63	*RF*	0.68	0.62
XGBoost	0.65	0.58	*XGBoost*	0.65	0.59
SVM	0.68	0.56	*SVM*	0.68	0.57

Considering the average performance and stability, this paper chooses two models, *Random Forest* and *XGBoost*, to construct the strategy combination, and carries on the back-test analysis.

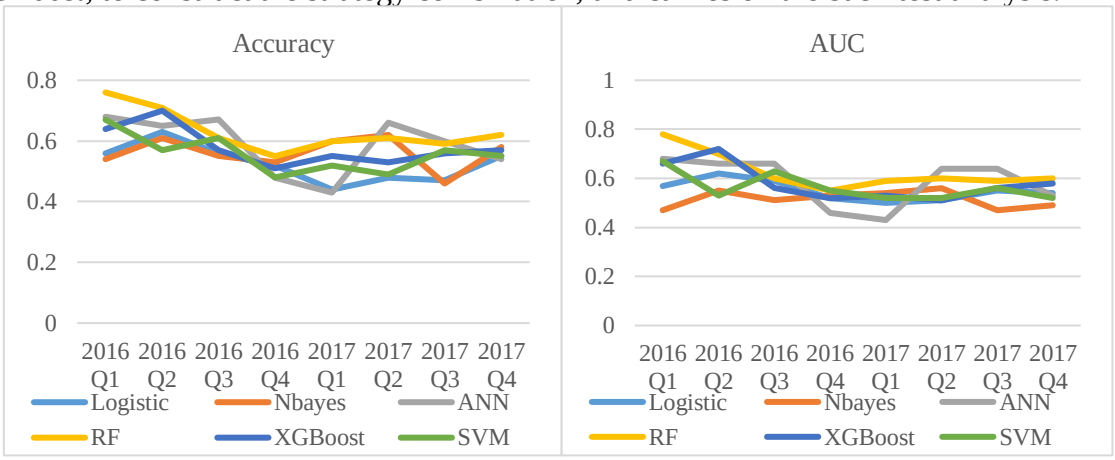

Fig. 1. Different model test set correct rate and AUC change graph

3.2 The portfolio construction and backtesting analysis

After comparing the performance of the six models, this paper chose *Random Forest* and *XGBoost* model to select stocks and construct portfolio with equal weight. At the end of each quarter of the test period from 2016 to 2017, we chose the ten stocks with the highest probability of high returns next month to adjust our portfolio.

The portfolio obtained from the two stock selection models is compared with the benchmark portfolio (average level of Shanghai and Shenzhen 300 stocks). It can be found that the portfolio constructed in this paper can obtain higher than the average return. In addition, in the two-year investment test period, we can get a portfolio with a net value of about 1.5 times that of the CIS300 benchmark portfolio.

Fig. 2. *Random Forest* and *XGBoost* portfolio yield and portfolio net value divided by the benchmark portfolio net value

4. Summary

The stock selection range of this paper is the constituent stock of the Shanghai and Shenzhen 300 Index. The quarterly data for 2014-2015 is used as a training sample, and the quarterly data for 2016-2017 is used as a test sample.

In this paper, from the classification accuracy rate and the average value of AUC, the classification results of *Logistic Regression* and *Naive Bayesian* model are the worst. On the contrary, *Random Forest*, *Neural Network* and *XGBoost* classification perform better. From the data changes of the test set, the *Neural Network* model is relatively unstable, and the performance of the period is fluctuating greatly. The investment portfolio constructed by the *Random Forest* and *XGBoost* stock selection can obtain higher than average yield.

In this paper, the research of machine learning stock selection model provides new methods and ideas for quantifying the field of stock selection. However, the real influential factors also need investors to judge and choose according to their own understanding of the market, and make adjustments according to the real-time changes of market conditions.

References

[1] Abraham, Hybrid intelligent systems for stock market analysis, *In Computational science-ICCS*, vol.15, pp. 337–345, 2001.

[2] Chen, A.-S., Application of neural networks to an emerging financial market: Forecasting and trading the taiwan stock index, *Computers & Operations Research*, vol.30, pp. 901–923, 2003.

[3] Cox, DR, The regression analysis of binary sequences (with discussion), *J Roy Stat Soc B*, vol. 20, pp. 215–242, 1958.

[4] Eduardo A. Gerlein, Evaluating machine learning classification for financial trading: An empirical approach, *Expert Systems With Applications*, vol.54, pp. 193–207, 2016.

[5] Göçken, M., Integrating metaheuristics and artificial neural networks for improved stock price prediction, *Expert Systems with Applications*, vol.44, pp. 320–331, 2016.

[6] Gong, X.,Si, Financial time series pattern matching with extended URC suite and support vector machine, *Expert Systems with Applications*, 55, 284–296. 2016.

[7] Hassan, M. R., A fusion model of *HMM*, *ANN* and *GA* for stock market forecasting, *Expert Systems with Applications*, vol.33, pp. 171–180, 2007.