Advances in Economics, Business and Management Research, volume 215
Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development (ICSSED 2022)

Research on Investment Decisions of Open-ended Funds Based on Decision Tree, RF and LGBM during COVID-19

Ruihua Zhou

School of Mathematical Science, South China Normal University, Guangzhou, Guangdong Province, China, 510631
**Corresponding author. Email: zrhfafa@163.com*

ABSTRACT
Since the onset of COVID-19, global economic development has not been as good as it used to be, so more and more people are looking to earn more money by investing besides working. The open-ended fund has a broad market. In addition, open-ended funds have many options. It is appropriate for all types of people with varying levels of risk tolerance and investment ability. During the pandemic, the risk resistance of investment products is always one of the most important factors for people to consider whether to invest. Therefore, this paper aims to screen open-ended funds with good risk resistance during the epidemic through machine learning. In this paper, the anti-risk ability of funds is determined according to the annual return rate of each fund since the epidemic. Ten indicators such as Sharpe ratio and Treynor performance measure are used to measure the anti-risk ability of funds during the epidemic from two dimensions including funds and fund managers. Then decision tree, random forest, random forest optimization model, LGBM model are established. It uses these models to predict the fund's anti-risk ability. Finally, the author finds the best model according to the accuracy. The LGBM model has the highest prediction accuracy, 93.5% of the fund's risk resistance ability types, and the model's prediction accuracy rate of excellent funds is 71.43%.

Keywords: *open-ended funds, COVID-19, random forest, LGBM, investment advice.*

1. INTRODUCTION

COVID-19 has a certain impact on the Chinese market. Many people's investments in the stock market and fund market have been affected. The net value of many funds has declined significantly during this period of time. According to the report of Asset Management Association of China, by the end of the third quarter of 2021, there are nearly 8,000 open-ended funds in the Chinese market, with a net value of about 210,000 [1]. It is nearly ten times that of closed-end funds. This phenomenon has attracted the author's attention. Before COVID-19, China also suffered from SARS in 2003. According to the research of Zhang Yi, Liu Yanhui, Xu Shanying, and Wang Shouyang, the impact of SARS on China's stock market is limited and short-term [2]. This conclusion shows that the epidemic will have an impact on China's stock market. It will also influence the fund market. The research of Zhou Wei, Tan Lin and Ding Bingqing also confirm that the resilience of China's financial market is relatively weak, and China's fund market would be affected by the epidemic [3]. In the fund market, open-ended funds in China's fund industry were significantly less negatively impacted by COVID-19 than closed-end funds through Kuang Yicheng and Qu Bo's study about the impact of COVID-19 on China's securities investment funds [4].

There are many pieces of researches about the risk of open-ended funds, but the use of machine learning is relatively few. Peng Yanlin used support vector machine and the improved model in the research of stock open-ended funds [5], while many machine learning methods have not been used. Therefore, this paper uses more machine learning methods to enrich the field.

This paper first collected multi-dimensional information of 152 open-ended funds, and selected funds with better risk resistance based on their annual returns from the beginning of the epidemic in 2019 to the end of 2021. After that, it establishes training sets and test sets with these funds in a ratio of 95:5. The training set is input into decision tree, random forest and its improved model and LGBM model to train models. The test set is input to calculate the accuracy of the model and the

accuracy of predicting the excellent funds. At last, the best model for investment advice is selected according to the actual situation and the accuracy of the models.

This paper aims to help to complete the research gap of using machine learning to make investment decision recommendations. At the same time, Chinese open-ended fund investors are over-chasing faddish funds. It may lead to lower returns than their expected [6]. Therefore, it also allows investors to select funds with good risk resistance before investing in open-ended funds. to some extent, it would reduce the possibility of an investment loss during the epidemic.

2. DATA PROCESSING AND INDEX SELECTION

The data set consists of 152 open-ended funds, and the data is sourced from a professional financial database called WIND [7]. The author considers fund age, Sharpe Ratio, Treynor Performance Measure, Jensen Ratio, beta, downside risk, annualized returns, WIND's 3-year rating and Shanghai Securities's 3-year rating, all from 2017 to 2019. In addition to the dimensions of the fund itself, the author also considers the fund manager. The fund styles formed by the portfolio investment of the same fund managers are relatively consistent. It means that the anti-risk ability of these funds may be related to the fund managers. Zhu Mu's research shows there is a correlation between the personal characteristics of fund managers and fund risks and risk-adjusted returns [8]. Therefore, the author introduced timing ability and annualized return on tenure as indicators to measure fund managers.

Before dividing the data set into test sets and training sets, funds need to be divided into excellent and average categories. According to Shen Xumin's article, there are two world-recognized well-known fund performance evaluation research institutions,called Morningstar and Standard & Poor's respectively [9]. According to the classification criteria of Morningstar, the funds having the top 32.5% of risk-adjusted fund returns are defined as excellent. In this paper, the top 45 risk-adjusted fund returns are labeled 1, accounting for 29.6%. It is recommended to invest them. The remainder is marked 0, indicating that the fund is not suitable for investment.

2.1. Annualized return of funds (logarithmic return)

$$r_t = lg\left(\frac{N_t+D_t}{N_{t-1}}\right) \times 100\% (1)$$

N_t and N_{t-1} represent the net unit value of the fund on day t and t-1 respectively. r_t represents the logarithmic return of the fund on day t. D_t is the interest from day t-1 to day t. This paper selects the emergency historical logarithmic return rate as the index to examine whether the fund is good or not.

2.2. Standard deviation of fund return rate

$$\sigma_p = \sqrt{\frac{\sum_{t=1}^{T}(r_t - E(R_p))^2}{T-1}} (2)$$

σ_p represents the volatility of the fund's return rate over the period T. It can be used to describe the total risk of a fund. In the formula, $r_t$represents the net return rate of the fund. T represents the period. $E(R_p)$ represents the average return rate of the fund's portfolio.

2.3. Fund market risk coefficient

$$\beta_p = \frac{cov(R_p, R_m)}{\sigma_m^2} (3)$$

The fund market risk coefficient β_p represents the systemic risk in the fund market. When $\beta_p<1$, it means that the risk of the stock fund are less than that of the market index, also with smaller fluctuations. At the same time, the return would be less than that of the market index too. When $\beta_p>1$, it means that the risks of the stock fund is greater than that of the market index, also with greater fluctuation. Meanwhile, it leads to more revenue. In the formula R_p is the net return rate of fund p complex rights. R_m is the market benchmark portfolio yield. σ_m^2 is the variance of the rate of return.

2.4. Sharpe Ratio

$$S_p = \frac{E(R_p) - R_f}{\sigma_p} (4)$$

Sharpe ratio is an index that comprehensively considers returns and risks. It represents the excess returns generated when the fund bears one unit of risk. In the formula, $E(R_p)$ is the average rate of return of fund products under real economic market risk. R_f represents the rate of return if there is no risk including market risk. σ_p is the standard deviation of the return rate.

According to the research conclusions of Lin Hongmei, Du Jinyan and Zhang Shaodong on Sharpe ratio, sharpe ratio is effective in describing the performance of funds' return [10].

2.5. Treynor performance measure

$$T_p = \frac{E(R_p) - R_f}{\beta_p} (5)$$

Treynor Performance Measure represents the risk premium obtained after taking each unit of risk. In the formula, β_p represents the fund market risk coefficient.

2.6. Jensen Ratio

$$J_p = (R_p - R_f) - [\beta_p \times (R_m - R_f)] (6)$$

Jensen Ratio is an indicator used to indicate the performance of a fund's portfolio. This indicator is risk-adjusted in absolute terms. In the formula, R_m represents the market benchmark portfolio rate of return.

2.7. Downside Risk

$$D_r = \sqrt{\frac{1}{n}\sum_{r_t<c}^{n}(r_t - c)^2}(7)$$

Downside Risk means how much risk a fund has when it falls relative to its risk-free return. In the formula, r_t is the rate of return of the fund in period T. c is the rate of return threshold set in period T. n is the number of fund returns over the period T.

2.8. Stock Selection Ability and Timing Ability

$$R_p - R_f = \alpha_p + \beta(R_m - R_f) + \gamma(R_m - R_f)^2 + \varepsilon_p(8)$$

Stock selection ability is the ability of fund managers to select stocks with good returns. Timing ability is the ability of fund managers to grasp the market entry and exit. In the formula, α_p is stock selection ability index. γ is an indicator of timing ability.

2.9. WIND's Rating and Shanghai Securities's Rating

This paper adopts the rating of Shanghai Stock Exchange, an authoritative institution in China, and the rating of WIND Database, an industry-recognized information database for financial trading markets. Both are rated on a scale of 1 to 5, with a higher number indicating a higher overall score.

3. MODEL ESTABLISHMENT

3.1. Decision Tree

Decision tree, as its name implies, is an algorithm of decision making based on "tree" structure. It is a common classification method, and its essence is supervised learning.

Supervised learning is to give a training set. The samples in the training set have a set of attributes and a classification result. The classification result is known. Decision tree is named decision tree because the image of the classification process is tree-like.

The progress of generating decision tree consists of two steps as follows. Learning samples with their classification results known leads to these steps.

Node splitting: the rule is that if the attributes of the node couldn't be given a judgement, the node would be divided to some child nodes, maybe two or more. It depends on the kind of decision tree.

Determination of thresholds: Selecting appropriate thresholds to minimize the training error rate.

The decision tree adopted in this paper is ID3. ID3 decision tree generally uses information gain to divide attribute selection. The information gain is defined as follows. Suppose the discrete attribute A has V values: {a1,a2,⋯,aV}. In D, the sample set of aV on all attribute a is DV, and the information gain formula obtained by dividing sample set D with attribute a is as follows.

$$\mathrm{Gain}(\mathrm{D},\mathrm{x}) = \mathrm{Ent}(\mathrm{D}) - \sum_{\mathrm{v}=1}^{\mathrm{V}} \frac{|\mathrm{D}^{\mathrm{V}}|}{|\mathrm{D}|}\mathrm{Ent}(\mathrm{D}^{\mathrm{V}})(9)$$

Ent(D) is information entropy. Suppose that the proportion of the sample in class k in the current sample set D is p_k(k=1, 2, ⋯, |y|), then the information entropy formula of D is as follows.

$$\mathrm{Ent}(\mathrm{D}) = -\sum_{\mathrm{k}=1}^{|\mathrm{y}|} \mathrm{p_k} \log_2 \mathrm{p_k}(10)$$

In usual, if there is a large information gain, it always leads to a better purity gain classified by attribute x .

3.2. Random Forest

Random forest always consists of more than two decision trees. When it comes to its output, all the output of all the decision tree would obey a rule to determine the result of random forest. The steps to building each tree are as follows. Firstly, if there are some training cases and the number of them is N. At the same time, there are some features of these training cases, and the number of them is M. Secondly, there are some individual decision trees to decide the result of a node together, and m represents the number of attributes input to each decision trees. Something that needs attention is that m should be much less than M. Thirdly, a training set should be built and it consists of N samples from N training cases. The process obeys the rule of putting back sampling. The left cases are used to make predictions and evaluate the errors. In the fourth step, selecting m features to decide the decision of each node in the decision trees, and the progress of the selection should be in a random way. With these features, the best splitting mode could be calculated and get the best random forest model among all kinds of random forest.

3.3. XGboost

XGBoost stands for eXtreme Gradient Boosting. It is an optimized distributed Gradient Boosting library designed to be efficient, flexible, and portable.

3.4. Extra-Trees

ET consists of some decision trees, resembling the random forest. Nevertheless, these two algorithms could differ by two pivotal points. Firstly, the random forest uses the bagging model, while ET uses all training samples to obtain each decision tree, that is, all the same training samples are applied in each decision tree. Secondly, the random forest gets the best bifurcation attribute in a random subset. In contrast, the bifurcation attribute in ET would be get in random and eventually get the decision tree.

3.5. Gradient Boosting

Gradient Boosting is a machine learning method that integrates weak learning models. In Gradient Boosting, the method of hierarchical learning is adopted to get the final model through m steps. In the step m, it learns a weaker model F_m. At step m+1, Instead of optimizing F_m directly, learns a basic model $h(x)$. The fitting residual term is $y - F_m$. This will make the model predicted value $F_{m+1} = F_m + h(x)$ closer to the real value. So the goal becomes how to find $h(x) = F_{m+1} - F_m$. Finally, the author wants to find a set of $h(x)$ in the space of some kind of function that makes $F(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + a$. Here, a is constant value. The algorithm adopts the idea of gradient descent and uses the negative gradient as the residual to learn the basic model.

3.6. Histogram Algorithm

The basic ideas of the histogram algorithm are as follows. Firstly, continuous floating-point eigenvalues are converted to discrete values, and the discrete values are integers, resulting in k integers. According to the discrete value, a histogram composed of K bars is obtained. The length of each bar is determined by the number of all data after discretization. After one walk through all the data, a complete histogram is obtained. Finally, the discrete values in the histogram are used to find the optimal segmentation points.

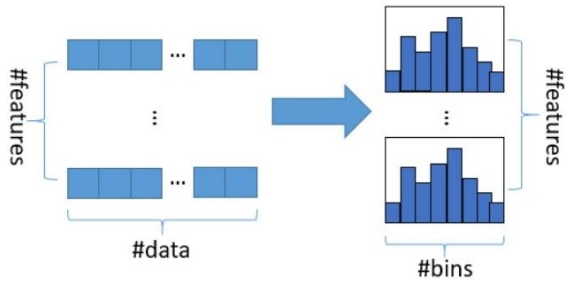

Figure 1 Histogram algorithm

3.7. LGBM

The optimizations of LightGBM are as follows. Firstly, compared to most GBDT tools, the strategy of decision tree, level-rise strategy is not used in LightGBM. Instead of it, LightGBM uses the leaf- wise strategy with the restriction of depth in decision trees. Apart from it, the differential acceleration of Histogram is another one. The histogram of a leaf can be easily obtained if the histogram of the parent node is known, and the histograms of the other child nodes under it are also known. The histogram could be get only through differentiating the histogram of the parent node from that of the other child nodes. This simple truth in histogram couldn't be doubted. However, in the traditional method, all the data on the leaf should be input. It's really time-consuming. With this principle, LightGBM can construct a histogram of a leaf and get a histogram of its leaf family at a fraction of the cost, doubling its speed.

5 models are used in this paper, including decision tree, random forest, random forest blending model 1, random forest blending model 2 and LGBM model. The random forest blending model 1 is a mixture of random forest and XGBoost. Random forest blending model 2 is a mixture of random forest, extra-trees and Gradient Boosting. The author inputted the training set into each model in turn. After the model is well trained, each index in the test set is input into the model to judge whether each fund in the test set is worth investing. Finally, the accuracy of each model is obtained.

4. RESULTS AND ANALYSIS

Table 1. Accuracy of each model

Model	Total accuracy	Accuracy of type 1	Accuracy of type 0
Decision Tree	75%	66.67%	76.92%
Random Forest	93.75%	66.67%	100%
Blending Model 1	81.25%	0%	100%
Blending Model 2	87.5%	33.33%	100%
LGBM	94.64%	71.43%	100%

It can be found from the table that only random forest and LGBM have a total accuracy higher than 90%. Except for the decision tree, all models can achieve 100% accuracy for funds of type 0. But the first recommendation of investment advice should be the fund with good risk resistance. Therefore, this paper selects the optimal model according to the accuracy of type 1. According to the table, LGBM has the highest accuracy rate of 71.43% for type 1.

The accuracy rate for type 1 of over half of the models is more than 50%. These results are ideal. The high accuracy rate may be due to the small variety and number of funds involved in this paper. In practical applications, due to the increase of the amount of data and the increase of fund types, whether the accuracy will decrease remains to be discussed. In addition, the accuracy rate calculated in this paper is based on the rate of return from COVID-19 to the end of 2021. In future practical applications, the time period may be longer, and the accuracy rate may also fluctuate.

5. CONCLUSION

In recent years, the number of investors in China's fund market has grown rapidly, and many novice investors have begun to invest in funds. But many novice investors don't know how to avoid risky portfolios, which can lead to losses. In addition, since the end of 2019, the global epidemic situation has intensified, and

more investors do not have sufficient trial and error costs. Therefore, this paper focuses on providing investment advice through machine learning models. It helps more investors select funds with strong risk resistance, so that investors can avoid losses to a greater extent after investment.Prior to this paper, the research methods in this area were relatively limited, and few machine learning methods were applied. Therefore, this paper adopted more machine learning models to apply them.

In this paper, a total of 5 models are established to give fund investment suggestions during the epidemic period. The best model is LGBM to predict the accuracy of funds with good risk resistance ability. This model can help investors select funds with good risk resistance and invest during the epidemic. The paper still has some limitations. The machine learning methods in this paper are limited to decision tree and random forest and their extension methods, while there are many other methods in machine learning that can be used in decision investment. For future research on investment advice, more different types of funds can be added to the data set, more machine learning methods can be tried, and more indicators can be added to improve the accuracy of investment advice.

REFERENCES

[1] Asset Management Association of China. Data statistics of public fund in China market. 2021.

https://www.amac.org.cn/researchstatistics/datastatistics/mutualfundindustrydata/

[2] Zhang Yi, Liu Yanhui, Xu Shanying, Wang Shouyang. Empirical analysis of SARS impact on Chinese stock market. Management Review(05), 2003, pp.3-63 doi:0.14120/j.cnki.cn11-5057/f.2003.05.001

[3] Zhou Wei, Tan Lin & Ding Bingqing. Research on anti-risk ability of index fund from the perspective of bubble analysis. Audit & Economy Research(05) (2017) 108-118.

[4] Kuang Yicheng & Qu Bo. Research on the impact of COVID-19 on China's securities investment funds. China Price (02) (2021) 74-77.

[5] Peng Yanlin. Research on investment decision of stock open-ended Fund based on improved support Vector Machine. University of Electronic Science and Technology of China. 2021. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFDTEMP&filename=1021749069.nh

[6] Guan Qifan. Is open-ended Fund Investors' Behavior Rational or Irrational?—An Empirical Study on China's open-ended Funds. Beijing Foreign Studies University. 2021. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202102&filename=1021063152.nh

[7] Zhang Ming, Li Chenxi & Wang Zhe. A comprehensive analysis of the heterogeneity of Corporate leverage in China: size, cycle and profitability — based on comparative analysis of three major databases. Financial Review (02) (2020) 1-123.

[8] Zhu Mu. Research on the influence of fund manager's personal characteristics on the performance of open-ended partial equity fund Guangxi Normal University. 2021. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFDTEMP&filename=1021617852.nh

[9] Shen Xumin. The Research of the Effeective Factors to the open-ended Fund's Performance. East China University of Science and Technology. 2021. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD2012&filename=1012309894.nh

[10] Lin Hongmei, Du Jinyan & Zhang Shaodong. Sharpe Ratio:Estimation Method, Applicability and Empirical Analysis (06) (2021)73-88.