Hindawi
Mobile Information Systems
Volume 2022, Article ID 8731261, 7 pages
https://doi.org/10.1155/2022/8731261

Research Article

Research on the Prediction of Nonbreakeven Financial Products' Yield of Commercial Banks Based on Machine Learning

Xiaoli Tong and Jiangjiao Duan

Department of Management, University of Shanghai for Science and Technology, Shanghai, China

Correspondence should be addressed to Xiaoli Tong; 193100875@st.usst.edu.cn

Received 21 August 2022; Revised 16 September 2022; Accepted 26 September 2022; Published 11 October 2022

Academic Editor: Imran Khan

Bank wealth management solutions have now become one of the most important components of the financial industry after nearly two decades of continuous development. However, there are still problems such as an imperfect pricing model and an ambiguous pricing mechanism. In this paper, we use machine learning to predict the yield of nonguaranteed financial products, and after model training and prediction, both the random forest model and the LightGBM model have high applicability; that is, machine learning can be effectively used in the yield forecasting process.

1. Introduction

2018 was a significant turning point for bank wealth management business. With the release of the "New Asset Management Regulations," bank wealth management products have shown a trend of nonguaranteed and net worth in line with policy requirements. Yueqiu and Bo believe that capital-guaranteed wealth management will gradually withdraw from the market and the main task of commercial banks in the future is to actively transform products according to policies and to transparently disclose daily net worth. In fact, nonguaranteed wealth management products are off-balance sheet businesses of banks and can be classified as shadow banking. Due to its high risk and rapid development trend in recent years, shadow banking has attracted great attention from supervisory authorities and scholars [1].

The availability of huge data and the development of a large number of scientific computing tools have fostered the use of machine learning in recent years, as computer science and technology have advanced. Although machine learning has problems, it is still a relatively new topic in asset pricing. However, it is undeniable that in a rapidly changing financial market with a huge amount of data, it is difficult to find the relationship between variables quickly and accurately with traditional forecasting models. Therefore, based on the method of machine learning, we will analyze the rate of return of wealth management products of Chinese commercial banks and predict the rate of return according to its influencing factors.

The rest of the study is organized as follows: Section 2 overviews the background. Section 3 discusses the theoretical analysis of the proposed concepts. Section 4 discusses the model, and in section 5, we explore the training and prediction of the suggested work. Section 6 concludes the article.

2. Background

With the development of bank wealth management products, scholars have begun to pay attention to the influencing variables of product yield. Pelster and Schertler [2] proposed that the most critical impact indicator is the way of capital operation. Acharya et al. [3] made an empirical analysis of wealth management products issued by 25 banks in the past seven years and proposed that the number of wealth management products due this year and the number of new issuance will have an impact on the yield. After empirical model testing, Warne and Sharma [4] determined that investors' investing patterns and investment goals will have a

role, with income being the most important element. Furthermore, freshly produced wealth management products are increasingly popular among investors. Based on the perspective of regulatory policy changes, Na [5] concluded that the introduction of regulatory policies will have a short-term inhibitory effect on the return of wealth management products, and the product term and market interest rate are the decisive factors for the return of wealth management products. Shiyang et al. [6] believe that, on the one hand, commercial banks are subject to strict loan-to-deposit ratio control, and on the other hand, due to their poor ability to absorb deposits, they cannot obtain sufficient funds to expand their business scale and usually need to issue wealth management products as a substitute for deposits. . Therefore, commercial banks with poor deposit-taking ability will tend to formulate higher expected product returns.

Some scholars have also conducted research on the prediction of the yield. Some scholars have also conducted research on the prediction of the rate of return. Ronghua et al. [7] classified financial products according to their risk levels according to the nature of bank financial products and then used a semiparametric model with random effects to construct and analyze the yield curve of financial products. Chunling et al. [8] summarized the research progress on the predictability of capital market returns and found that a machine learning method is a research hotspot in recent years.

Gan et al. [9] suggested a deep learning-based strategy for option pricing that can produce more accurate results at a faster rate. Li et al. [10] generated the stock technical index value using the stock's daily frequency price and trading volume data and then utilized the derived technical index value as an input variable to anticipate the stock price rise or decline in a few days. Pan et al. [11] used neural networks to predict the stock return rate to capture the nonlinear relationship between the three factors of the market portfolio return rate, book-to-market value ratio, and market value. Chen [12] conducted a series of studies on the pricing of catastrophe bond risk spreads and compared the effects of machine learning models and traditional regression models.

Furthermore, machine learning can be applied in fields such as risk prediction. Fang and Luo [13] built a risk indicator alarm mechanism using the random forest algorithm to separate the risk alarm indicator variables into two categories: enterprise characteristics and business conduct.

Some scholars have also compared the application of various algorithms. Breiman [14] believes that the random forest algorithm has obvious superiority. Due to the theorem of large numbers, using this algorithm will not overfit the model. Kampichler et al. [15] compared the practical results of five machine learning algorithms: decision tree, random forest, artificial neural network, support vector machine, and rule-based fuzzy model and finally proposed that the prediction effect of the random forest is the best.

3. Theoretical Analysis

The yield of wealth management products should be impacted heavily by the nature of the product itself. The longer the money is invested, the greater the uncertainty investors bear and the higher the liquidity risk they face. In addition, the risk effect and the threshold effect will also have an impact on the time value of capital, capital liquidity, and investment risk of investment, which are reflected in the product's yield.

There are also some variations between banks. State-owned commercial banks are controlled directly by state-owned capital and are subject to greater regulation. Although wealth management-issued products are not legally guaranteed, they are actually implicitly guaranteed by state credit to investors, and banks will be more cautious when pricing assets.

In addition, smaller banks often need to broaden their funding sources to expand their business scale. The release of wealth management products in asset management business is a form. Therefore, in order to improve operational efficiency and also to absorb funds, formulating higher yields may become a way for commercial banks to increase the scale of wealth management business.

Macroeconomic changes can have a large and far-reaching impact on the financial industry. The position in other marketplaces should also be considered. The interbank market position represents the liquidity of funds in the interbank market and, thus, indirectly reflects the bank's funding channels. In addition, a considerable part of funds raised after the issuance of wealth management products will actually be invested in the fund pool for operation and management, and part of the fund pool involves the stock market.

Therefore, we discuss the independent variables that affect the yield of wealth management products from four aspects: issuing banks, product design, macroeconomics, and other markets. The variables are shown in Table 1.

4. Model

4.1. Multiple Linear Regression Model. The multiple linear regression model is usually used to study the relationship between a dependent variable and multiple independent variables, which is represented by a matrix.

We first use the multiple linear regression model to make regression predictions on the two sample sets and set up model 1 based on the analysis of influencing factors:

$$\begin{aligned}\text{rate\_max}_i = {} & \alpha + \beta_1 \text{period}_i + \beta_2 \text{risk}_i + \beta_3 \text{threshold}_i \\ & + \beta_4 \text{bank}_i + \beta_5 \text{size}_i + \beta_6 \text{quantity}_i + \beta_7 cr_i \\ & + \beta_8 lr_i + \beta_9 l\,d_i + \beta_{10} pc_i + \beta_{11} g\,dp_i + \beta_{12} m2_i \\ & + \beta_{13} cpi_i + \beta_{14} \text{fintech}_i + \beta_{15} \text{shi}_i + \beta_{16} \text{price}_i.\end{aligned} \tag{1}$$

Model 2 is as follows:

$$\begin{aligned}\text{rate\_min}_i = {} & \alpha + \beta_1 \text{period}_i + \beta_2 \text{risk}_i + \beta_3 \text{threshold}_i \\ & + \beta_4 \text{bank}_i + \beta_5 \text{size}_i + \beta_6 \text{quantity}_i + \beta_7 cr_i \\ & + \beta_8 lr_i + \beta_9 l\,d_i + \beta_{10} pc_i + \beta_{11} g\,dp_i + \beta_{12} m2_i \\ & + \beta_{13} cpi_i + \beta_{14} \text{fintech}_i + \beta_{15} shi_i + \beta_{16} \text{price}_i.\end{aligned} \tag{2}$$

Table 1: Independent variables.

Variable	Feature
Length of the commission period	Period
Risk	Risk
Threshold	Threshold
Bank nature	Bank
Asset size	Size
Number of wealth management products issued in the previous year	Quantity
Capital adequacy ratio	Cr
Deposit ratio	Ld
Liquidity ratio	Lr
Provision coverage	Pc
Gross national product	GDP
Broad money supply	m2
National consumption	CPI
Fintech index	Fintech
Interbank market	Shi
Stock market	Price

Table 2: The parameters of the random forest model.

Parameters	Descriptions and settings
n_estimators	The number of classifiers, also called the number of iterations, is the number of decision trees in the forest
Criterion	The standard used for splitting nodes, the default is gini
Max_depth	Maximum depth of trees
Min_samples_split	Minimum number of samples required to split a node inside the tree, defaults to 2
Min_samples_leaf	Minimum number of samples required at leaf nodes, defaults to 1
Min_weight_fraction_leaf	Minimum weighted score in the sum of weights at all leaf nodes, defaults to 0
Max_features	The number of features to consider when finding the best segmentation, the default is none; that is, all features are considered
Max_leaf_nodes	The maximum number of leaf nodes, which must be an integer, the default is none
Min_impurity_decrease	If the decrement of the split index is greater than this value, then split, default is 0
Bootstrap	Whether there is a randomly selected sample to put back, the default is true
Oob_score	Whether to use out-of-bag samples to evaluate the quality of the model, set to true
N_jobs	The number of parallel calculations, the default is none
Random_state	Controls the randomness of bootstrap and randomness of the selected samples. In order to facilitate the adjustment of other parameters, this parameter adopts the default value
Verbose	Controls verbosity when fitting and predicting, default is 0
Warm_start	Whether to use the trained model and add more base learners to it, set to false

4.2. Random Forest Model. The random forest model is constructed based on the bagging ensemble learning method, so the process of training and constructing the random forest model basically follows the basic process of bagging ensemble learning. Specifically, for a data set D containing k samples, we first perform k random self-sampling on D, collect k training sample subsets D1, D2,. . ., Dk, and then select D1, D2,. . ., Dk to trainand construct$k$ decision trees, then we can combine these decision trees to obtain a random forest model.

Table 2 shows the names, descriptions, and settings of important parameters in the random forest model.

4.3. LightGBM Model. The LightGBM model is partially optimized on the traditional boosting algorithm. Since the traditional boosting algorithm needs to scan all the sample points for each feature to select the best segmentation point, it is very time consuming, while the LightGBM algorithm is based on the histogram. The decision tree algorithm greatly reduces the time complexity. The histogram algorithm first performs binning processing on the eigenvalues. For continuous features, binning processing is to discretize continuous data, and then, there is no need to scan each feature as in the traditional algorithm but only need to press the bins that are scanned, which speeds up training.

Table 3 shows the names, descriptions, and settings of important parameters in the LightGBM model.

5. Training and Prediction

5.1. Data. We collected 201,572 nonguaranteed wealth management products issued by 20 commercial banks (five each of state-owned commercial banks, joint-stock commercial banks, city commercial banks, and agricultural commercial banks) from January 2017 to December 2020. The dependent variable is the upper and lower bounds of the yield set at the time of issuance.

Table 3: The parameter of the random forest model.

Parameters	Descriptions and settings
Task	Train
Objective	Model training target, which can be selected as a regression model or binary classification model
Boosing_type	Base learner, for gbdt
Metric	Metrics as a function of evaluating the optimal model
Leaning_rate	Learning rate, the default is 0.1, the smaller the value, the more accurate the learning
n_estimators	The number of iterations in the classifier
Num_leaves	The number of leaves in the tree, the indicator should be less than 2 to the power of max_depth
Max_depth	Tree depth
Feature_fraction	The feature sampling ratio for building the tree, which ranges from 0 to 1
Bagging_fraction	The sample sampling ratio for building the tree, the data range is 0 to 1
Max_bin	The maximum bin value, generally equal to the number of features
Min_data_in_leaf	The minimum number of samples for each leaf node, when the leaf node is smaller than this value, the tree will no longer be split
Min_gain_to_split	The tree stops splitting when the leaf node is smaller than this value

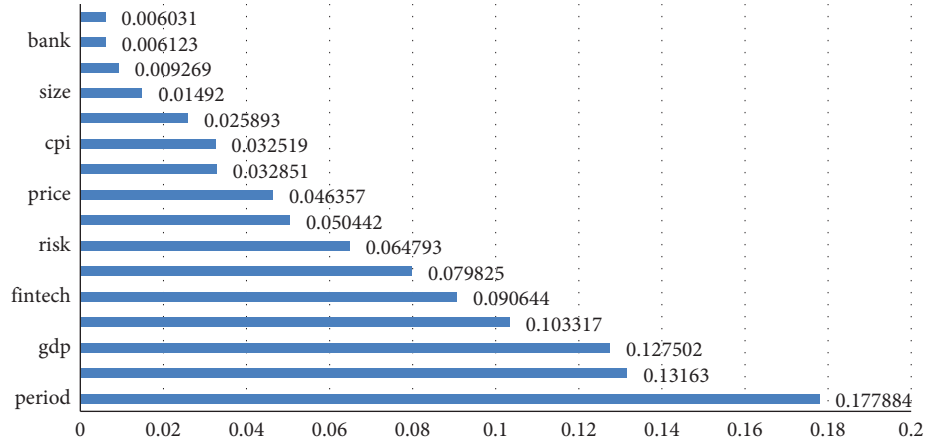

Figure 1: Eigenvalue importance of yield cap samples.

At the same time, we collected data such as the minimum purchase threshold amount and the period set in the product description, as well as the asset size of the issuing bank, the number of wealth management products issued, and the nature of the bank. The sample banks were chosen based on their issuance size in recent years, and the data are all obtained from the wind database. In addition, we have gathered the Shibor Index, Fintech Index, and Consumer Price Index in terms of macroeconomic operation. Because of the short publishing cycle, GDP and m2 only have annual data. At the same time, since the 20 banks in the sample are all listed banks, the closing price is taken as one of the factors to be considered in other markets.

We take the capped expected yield and the floored expected yield of the financial products in the sample as dependent variables, named rate_max and rate_min.

After data preprocessing, we divide samples after removing missing values and extreme values into the training set and the test set. The divided test set ratio is set to 0.2, and the division standard is randomly divided by using software.

5.2. Importance of Features. The random forest algorithm can measure the relative importance of each feature value for prediction, that is, the average contribution of each feature to each tree in the random forest model. An evaluation index is used to calculate the mistake rate. The difference value of out-of-bag data is calculated by randomly adding noise interference to a feature value. We utilize the random forest approach to examine and rank the importance of the sixteen eigenvalues in the sample after processing characteristics one by one, as shown in Figures 1 and 2.

According to the eigenvalue importance ranking results output by the machine learning algorithm, in both samples, the period of entrust is the most important eigenvalue, and the importance of risks and thresholds are also in the upstream position. It can be seen that the design of the wealth management product has a crucial impact on the progress of predicting.

However, whether or not it is a state-owned bank has little bearing. In the two samples, bank assets and the number of issuances in the preceding year have diametrically opposed importance and influence.

Among the macroeconomic operating variables, the gross national product and m2 in the two sample sets are both factors that will have a greater impact on the expected yield.

For other markets, the stock market affects yields more than the interbank market. On the one hand, the stock

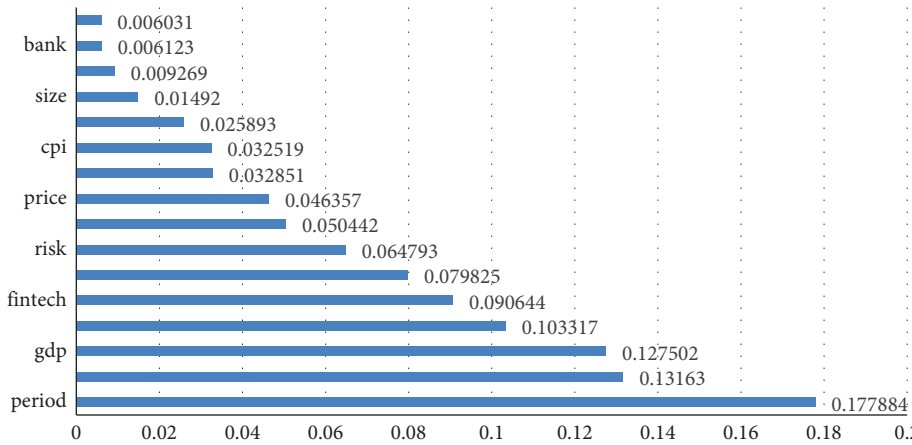

Figure 2: Eigenvalue importance of yield floor samples.

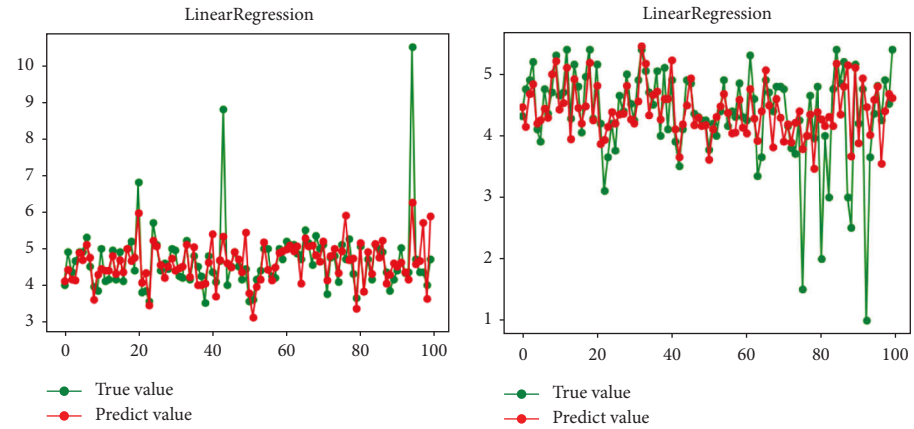

Figure 3: The result of the multiple linear regression model.

market situation of listed commercial banks represents the bank's reputation and the investor's confidence, and on the other hand, the funds obtained by banks through the issuance of wealth management products will also enter the stock market.

5.3. Prediction. The prediction results of the multiple linear regression model are shown in Figure 3. Due to the large sample size, only the results of 100 pieces of data are shown in the graph. The horizontal axis is the serial number of the test set, the vertical axis is the expected return value, and the two lines are the true value and the predicted value obtained by the multiple linear regression model. It can be seen from Figure 3 that the two lines do not fit, and the prediction made by the multiple linear regression model does not coincide with the real value.

We use the mean absolute error (MAE) index, root mean square error (RMSE) index, and R-square to score and analyze the model regression results.

The mean absolute error is used to measure the mean absolute error between the predicted value and the true value. The smaller the MAE, the better the model. It is defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - y_i^*\right|. \tag{3}$$

The root mean square error is also used to indicate the error that the model will produce in prediction. The smaller the indicator is, the better the model is. It is defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}. \tag{4}$$

R-square represents the fit of the model. The closer the value to 1, the better the model. It is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2}. \tag{5}$$

In the random forest, n_estimators is the maximum number of iterations of the weak learners of the random forest model, also known as the maximum number of weak learners in the random forest. We set the range of this parameter from 1 to 200, and other parameters are temporarily set to default values; then, we use the grid search command to adjust the parameters for the upper and lower limits of the expected rate of return of sample financial products.

Table 4: Random forest model-adjusted parameter results.

Parameters	Yield cap samples	Yield floor samples
n_estimators	120	110
Max_depth	51	50

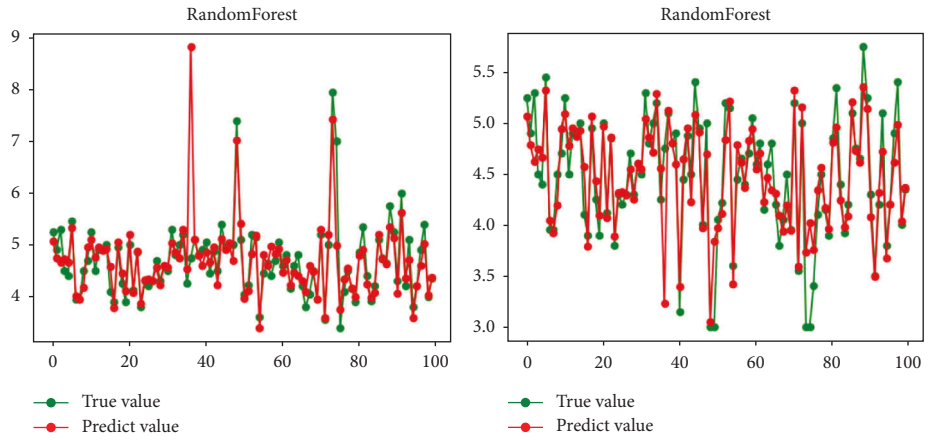

Figure 4: The result of the random Forest model.

Table 5: LightGBM model-adjusted parameter results.

Parameter	Yield cap samples	Yield floor samples
n_estimators	57	90
Max_depth	32	79

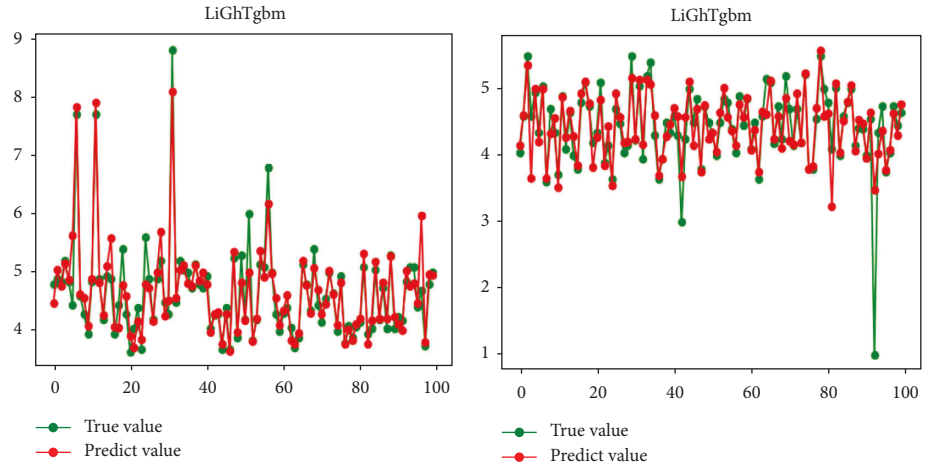

Figure 5: The result of the random forest model.

In the random forest model, max depth is the maximum depth of the tree. If it is not altered, the default value is none, which means that the decision tree will not limit the depth of the subtree when it is built. When the sample size is small, this parameter will not have an excessive impact on the regression process. Our sample size is relatively large, so we adjust the maximum depth parameter for the two sample sets, respectively, and set the max_depth parameter range from 1 to 100. After adjustment, the results of the parameters are shown in Table 4.

After applying the adjusted parameters to the regression prediction model, the specific prediction results are shown in Figure 4, and only 100 results are displayed. It can be seen from Figure 4 that although some points of the two discounts do not overlap, the trends are basically the same.

We used the grid search approach to optimise the parameters of the LightGBM model as we did for the random forest model and continuously changed the interval and step size until we achieved the ideal parameter values, which are displayed in Table 5. Among them, Num_leaves is the

Table 6: The comparison between three models.

	Yield cap samples		
	Multiple linear regression	Random forest	LightGBM
MAE	0.24	0.23	0.21
RMSE	0.28	0.23	0.26
R2	0.39	0.46	0.67
	Yield floor samples		
	Multiple linear regression	Random forest	LightGBM
MAE	0.27	0.18	0.17
RMSE	0.30	0.18	0.20
R2	0.51	0.72	0.81

number of leaves in each tree, and this parameter has an important influence on the complexity of the model tree.

We apply the optimized parameters to the regression model to obtain the prediction result graph. Due to the large sample size, only 100 results are shown in the graph. As shown in Figure 5, the two polylines have a high degree of fit and have the same trend.

5.4. Comparison. To forecast the upper and lower bounds of the predicted return rate of the financial products in this sample, we utilize the multiple linear regression approach, random forest, and LightGBM model, respectively. The comparison results of MAE, RMSE, and R-square values of each model are shown in Table 6.

The MAE and RMSE indicators in the cap sample of the multiple linear regression model are 0.24 and 0.28, respectively and in the floor sample, they are 0.27 and 0.30, respectively, all of which show high errors. However, the error values of the random forest model and the LightGBM model are not much different, and they neither exceed 0.26 in the upper sample set nor 0.20 in the lower sample set.

R-square is a fitness index. In the cap sample set, the R-square of the three models are 0.67, 0.46, and 0.39, respectively. In the floor sample set, they are 0.51, 0.72, and 0.81, respectively. It can be seen that the fitting degree of the LightGBM regression model is the best, followed by the random forest model, and finally the multiple regression model. The prediction efficiency of the multiple linear regression model and the other two models is quite different.

6. Conclusion

We use the multiple linear regression model, random forest, and LightGBM in ensemble learning to predict the yield of the nonguaranteed wealth management products issued by twenty commercial banks in the past four years.

The empirical research on forecasting is typically separated into two components, and the first of which is an examination of eigenvalue importance ranking. Then, based on the importance of the eigenvalues of the two sample sets, we built a machine learning model and made value predictions for returns. Regardless of the sample set, the fitting degree of the LightGBM regression model is the best, followed by the random forest model, and finally the multiple regression model.

Data Availability

The datasets used during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Z. Yueqiu and Z. Bo, "Transformation and development of commercial bank wealth management business in the era of asset management 2.0," *Financial Forum*, vol. 24, no. 1, pp. 3–11, 2019.

[2] M. Pelster and A. Schertler, "Pricing and issuance dependencies in structured financial product portfolios," *Journal of Futures Markets*, vol. 39, no. 3, pp. 342–365, 2019.

[3] V. V. Acharya, J. Qian, Y. Su et al., "In the shadow of banks: wealth management products and issuing bank's risk in China," *Social Science Electronic Publishing*, Social Science Research Network, Rochester, NY, USA, 2020.

[4] D. P. Warne and P. Sharma, "Pattern of investment: behavioral study regarding financial products," *Asia Pacific Journal of Research in Business Management*, vol. 3, no. 12, pp. 13–17, 2012.

[5] L. Na, "Research on the influencing factors of commercial bank wealth management product returns——based on the perspective of regulatory policy changes," *Economic Jingwei*, vol. 36, no. 2, pp. 149–157, 2019.

[6] H. Shiyang, Z. Jigao, and L. Zhengfei, "Research on the ability of commercial banks to absorb deposits, issue wealth management and their economic consequences," *Financial Research*, vol. 468, no. 6, pp. 94–112, 2019.

[7] L. Ronghua, L. Huazhen, and Z. Lihong, "Construction and analysis of the yield curve of bank financial products - method based on random effects semiparametric model," *Financial Research*, no. 07, pp. 99–112, 2013.

[8] Z. Chunling, J. Fuwei, and T. Guohao, "Research progress on the predictability of capital market returns," *Economics Dynamics*, no. 02, pp. 133–148, 2019.

[9] L. Gan, H. M. Wang, and Z. J. Yang, "Machine learning solutions to challenges in finance: an application to the pricing of financial products," *Technological Forecasting and Social Change*, vol. 153, Article ID 119928, 2020.

[10] B. Li, Y. Lin, and W. M. L.-T. E. A. Tang, "A set of quantitative investment algorithms based on machine learning and technical analysis," *System Engineering Theory and Practice*, vol. 37, no. 5, pp. 1089–1100, 2017.

[11] S. Pan, J. Liu, and Y. Wang, "Research on stock return prediction based on neural network," *Journal of Zhejiang University (Science Edition)*, vol. 46, no. 5, pp. 550–555, 2019.

[12] H. Chen, "Pricing of catastrophe bond risk spread based on machine learning algorithm," *Practice and Understanding of Mathematics*, vol. 50, no. 20, pp. 71–81, 2020.

[13] R. Fang and P. Luo, "Research on third-party payment violation risk early warning based on random forest," *Technology and Economics*, vol. 39, no. 9, pp. 11–21, 2020.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 3, pp. 261–277, 2001.

[15] C. Kampichler, R. Wieland, S. Calme, H. Weissenberger, and S. Arriaga-Weiss, "Classification in conservation biology: a comparison of five machine-learning methods," *Ecological Informatics*, vol. 5, no. 6, pp. 441–450, 2010.