Prediction of Bank Product Subscription Behavior Based on Random Forest Algorithm

Tao WEI [1,a], Linying ZOU [1,b], Yanze SUN [1,c], Shengfa ZHAO [2,d]

Corresponding author (WEI Tao) Email: 534011694@qq.com

[a]534011694@qq.com, [b]2694721989@qq.com, [c]082000224@fzu.edu.cn, [d]2153225399@qq.com

[1] Fuzhou University, Mathematical finance, School of Economics and Management, Fuzhou, 350108, China

[2] Qinghai University for Nationalities, Computer Science and Technology, School of Computer Science, Xining 810000, China

Abstract: Based on the background of bank customers' purchase of products, this study uses random forest to predict whether customers buy bank products and discusses the advantages of random forest compared with other commonly used machine classification models. By selecting the product data set of Aliyun Tianchi Bank customers and using python for data processing, this study optimizes the n_estimators, max_features and other important parameters in random forest, so as to achieve the optimal effect of random forest classification. Meanwhile, by comparing KNN, logistic regression, support vector machine, Single decision tree model, the confusion matrix and ROC curve were used to evaluate the model performance. The final experimental results show that the random forest model with optimized parameters has better classification effect than other binary classification models.

Key words: Random forest, Machine learning, Bank product subscription, big data

1 Introduction

1.1 Research Background

In recent years, with the development of the financial industry, more and more attention has been paid to bank subscription products, which can effectively improve the capital structure of banks, improve their service quality and meet the needs of customers. However, due to the variety of subscribed products, the needs of customers are also different. Therefore, it is of great significance to study the customer behavior characteristics and influencing factors of subscribed products for banks to make operational decisions about subscribed products. However, China's banking industry has been established for a long time, abundant outlets, large customer volume, and accumulated a large number of customer records in the long business history [1]. With the deepening of informatization, the high speed expansion of data has brought the banking industry into the era of "big data". Reasonable use of machine algorithms to predict customers' subscription to bank products, help banks find corresponding potential target customers, reduce customer loss and cost customers, and provide guidance for banks' business decisions.

FFIT 2023, July 07-09, Chongqing, People's Republic of China

DOI 10.4108/eai.7-7-2023.2338033

1.2 Research Status and Innovation at Home and Abroad

At present, many scholars have used big data, machine learning algorithms and other technologies to study customer subscription behavior and explore its rules. Many researchers try to use a single prediction model when studying customers' purchase behaviors. For example, Tang, Wang, Xu and Li [2] proposed a purchase behavior prediction framework with optimized parameters through the SVM model based on firefly algorithm and achieved a better effect than the SVM model. Zhang, Wang, Jiao, Chen and Wang [3] used categorical regression tree (CART) and logistic regression methods to construct a user network purchase prediction model, and the results showed that CART decision tree has a high prediction accuracy.

However, a single model has weak feature interpretation ability and low accuracy in customer subscription behavior data. Therefore, the following two innovations are made in this paper. First, this paper takes the bank product subscription as the background, transforms the traditional user purchase prediction problem into a binary problem, and calculates the probability of the user's purchase behavior of the bank product through the random forest algorithm. Secondly, the data used in this paper comes from the real data of Aliyun Tianchi, which is not only consistent with the actual business scenarios, but also adopts different types of machine learning algorithms in algorith m selection, and compares and analyzes the data fitting various models, effectively improving the accuracy and robustness of the prediction [4]

1.3 Research Significance and Methodsa

The purpose of this paper is to analyze and forecast customer subscription behavior more accurately, so as to provide high-quality target groups for the precision marketing of bank products. This paper selects the data set of bank customer subscription products from Aliyun Tianchi financial Data analysis series for research. The main research content includes two aspects: data processing and model empirical analysis.

In terms of data processing, data cleaning, data coding and feature screening are mainly carried out on the data set. Label coding is carried out for category variables, and feature screening is carried out according to the feature scoring mechanism of random forest model after encoding. Then descriptive statistical analysis was carried out to analyze the collinearity between the features.

In the empirical analysis of the model, parameters in the random forest model were optimized, and appropriate evaluation indicators and ROC curve were selected to construct an evaluation system according to the possible prediction results. Then, the classification performance of KNN, logistic regression, support vector machine, single decision tree and optimized random forest models is compared and analyzed to get the experimental results.

2 Stochastic Forest Model Theory

A random forest model is a machine learning algorithm based on decision trees, which is an integrated learning method that combines multiple decision trees to improve the accuracy and stability of the model. It is a very efficient classification and regression algorithm that can be used to work with non-linear data. The basic principle of random forest is to further introduce a random attribute selection in the process of decision tree training to learn a subc lass as an

integration on the basis of bagging integration by using decision tree as a base learner, and finally determine the classification results by the number of votes of decision tree classification [4]. The establishment process and advantages of random forest can be used to lay a theoretical foundation for the bank's product subscription prediction.

2.1 Random Forest Algorithm

2.1.1 Process of Random Forest Algorithm

(1) Extract n samples from the training set, where n is a part of the total number of samples, and these samples are called training subsets.

(2) A decision tree is constructed on the training subset, and in the construction process, m features are randomly selected from the feature set each time the node is split, and m is a part of the total number of features.

(3) Repeat steps (1) and (2) to build k decision trees, where k is the number of decision trees.

(4) Forecast the new data and vote the predicted results of each decision tree to determine the final predicted results.

(5) Each decision tree can be pruned if needed to improve the generalization ability of the model.

(6) Finally, the algorithm is applied to actual data to solve specific problems to obtain final prediction results.

2.1.2 Generalization Error Theory of Random Forest

In the random forest algorithm, the value of generalization error is affected by the trees in the forest. When there are more trees in the forest, the generalization error will also increase and converge to a finite upper bound value. For a classification model, the main evaluation criterion of its classification performance is whether the classification model can correctly judge the data to be classified as the probability of the corresponding category [5]. Generalization performance is a good indicator to reflect the ability of constructing a model to accurately classify data.

(1) For the training set {(x, y)}, {h 1 (x), h(x), ..., hn(x)} is N tree classifiers that have been built, then the random forest interval based on x and y is defined as follows:

$$mg(x, y)=av_kI(h_k(x)=y)-max_{j\neq y}I(h_k(x)=y)$$

Where I(·) is the indicative function, avk(·) Take the average value. The minimum difference between the average number of correctly classified votes for a given sample x and the wrong number of votes for a classifier set is the interval function mg(x, y); the larger the difference between the functions, the better the classifiers are. The generalization error of the classifier can be expressed as:

$$PE^*=P_{X,Y}(mg(x,y)<0)$$

The subscript X, Y indicates that the probability is obtained in the X, Y space. In a random forest, hk(x) = h (x, thetak), if the number of numbers reaches a certain level, the equation will follow the law of strong numbers [6].

(2) After the number of numbers reaches a certain level, θ occurs in all sequence sets 1… Upper, PE*Will come out converging everywhere to:

$$PX,Y ((P\theta (h(x, \theta)=y) - \max j \neq y P\theta(h(x, \theta)=j))<0)$$

Where, the corresponding random vector of each decision tree is represented by θ, and the classifier output of x and θ is represented by h(x, θ).As the number of trees in the forest increases, the forest does not overfit, but tends to a bounded generalization error value [7].

3 Data cleaning and Feature Extraction

3.1 Data Sources

With the bank product subscription forecast as the background, predict whether customers will buy the bank's products based on historical data. The data set is derived from Aliyun Tianchi Bank customer subscription product data set. Data set is available at (accessed on 10 February 2023) found on the site. The labels are whether customers buy bank products, respectively represented by 0,1. In the original data, 86.88% of the labels that customers confirm to buy bank products and 13. 12% of the labels that customers do not buy bank products. The distribution of data samples is uneven, and the proportion of customers buying bank products is too high, so the blank accuracy rate reaches 86.88%. It shows that the accuracy of the established binary classifier learning model should reach at least 86.88%. There are 22500 pieces of original data and 20 pieces of characteristic variable information. The characteristics mainly include three categories: (1) the time information of contacting customers: the number of times of contacting customers, the duration of the last contact, the time interval of the last contact, etc. (2) Basic customer information: age, occupation, marriage, whether there is a mortgage, whether there has been a default, etc. (3) Employment, consumption information, interbank lending rate, etc.

3.2 Data Description

As shown in Table 1, according to the data provided by Tianchi, it can be found that the variable "Subscribe" is the prediction target of the model. If customers subscribe to bank products, the value of the variable is 1, and otherwise it is 2.

Table 1 Raw data characteristic variables

Fields	Instructions
age	Age
job	Career: admin, unknown, unemployed, management.
marital	Marriage: married, divorced, single
default	Whether the credit card is in default: yes or no
housing	Have a mortgage: yes or no
contact	Contact: unknown, telephone, cellular
month	Month last contacted: jan, feb, mar....
day_of_week	Day of the week last contacted: mon, tue, wed, thu, fri

duration	Duration of last contact (in seconds)
campaign	The number of times clients were contacted during the campaign
pdays	Number of days since last contact with the customer
previous	The number of times clients were contacted prior to this marketing campaign
poutcome	Results of previous marketing campaigns: unknown, other, failure, success
emp_var_rate	Rate of change in employment (quarterly indicator)
cons_price_index	Consumer Price Index (monthly measure)
cons_conf_index	Consumer Confidence Index (monthly indicator)
lending_rate3m	Interbank rate 3-month interest rate (daily indicator)
nr_employed	Number of employees (quarterly indicator)
Subscribe	Whether the customer makes a purchase: yes or no

3.3 Feature Engineering

3.3.1 Data Coding

The characteristic variables in the data set used in this paper include 9 continuous variables and 11 categorical variables. In this paper, it was found that the AUC of ROC curve was abnormally high when performing one-hot encoding for class variables, and when using decision tree, random forest, logistic regression and support vector machine to make binary prediction. It was presumed that the excessive category features caused by one-hot encoding led to too complex model, which was difficult to explain and easy to overfit. In this case, the model may well fit the data in the training set, but the prediction of the new data may not be accurate. Therefore, this paper decides to use label encoding for category variables. After label encoding, the AUC of each model has returned to normal.

3.3.2 Feature Screening

Random forest itself has a feature scoring mechanism, which can be applied to feature selection to achieve the purpose of screening. This paper is to establish stochastic forest classification model based on CART algorithm, and the basic principle is the minimum mean square error (MSE). For discrete response variables, CART selects the value that can optimally divide characteristic variables, and then compares data with this value to achieve the purpose of tree splitting. The basic idea is as follows: for each variable, each decision tree in random forest can measure the decline of the splitting criterion function (residual sum of squares and Gini coefficient) caused by the variable. Then, according to this decline, each decision tree can be averaged, that is, the measure of the importance of the variable. The importance of each characteristic variable is ranked in order and plotted as the importance chart of variables. The chart of characteristic importance variables obtained in this paper is shown as follows:

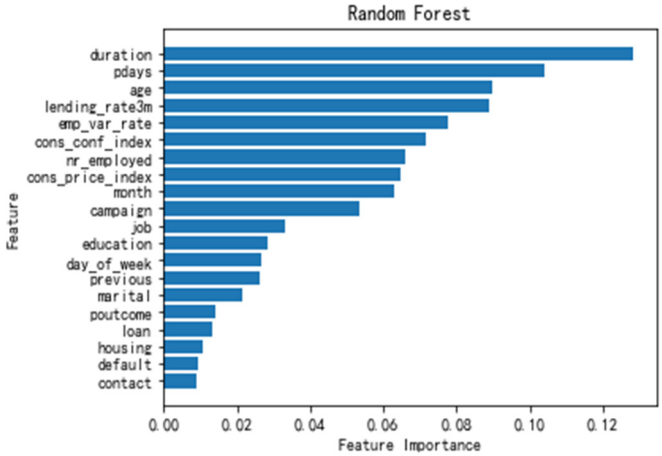

Figure 1. Feature importance graph of the products subscribed by bank customers

As shown in Figure 1, duration has the highest feature importance, with its feature importance score reaching 12.82%, followed by pdays, age and lending_rate3m, which are 10.43%, 8.97% and 8.89%, respectively. The rest are all below 8%. Using model training, this paper deletes the corresponding factors in the order of feature importance from low to high, and then tests the difference between the selected factors and the correct predicted value. After this process, it is found that even if the factors of lower importance are deleted, the accurate value of prediction will be significantly smaller. Therefore, this paper decides to select all the features in the figure as the main features to predict the products subscribed by bank customers.

3.3.3 Descriptive Statistical Analysis

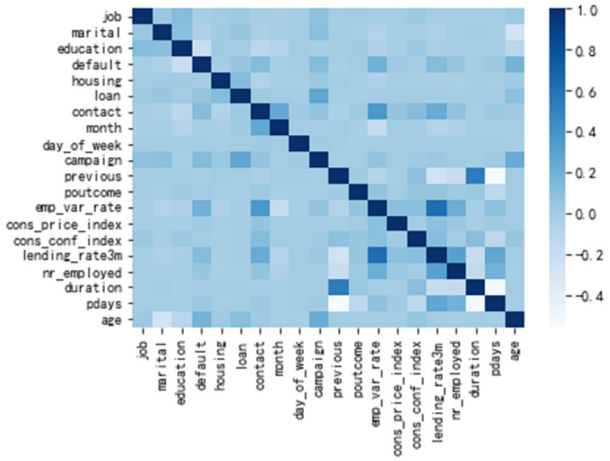

Figure 2. Correlation coefficient matrix heat map

As shown in Figure 2, pdays presents a moderate negative correlation with duration and previous, and its correlation coefficients are -0.52 and -0.55, respectively. lending_rate3m is positively

correlated with emp_var_rate and duration and previous, and its correlation coefficients are 0.62 and 0.54, respectively. The correlation coefficients of the other variables are all less than 0.4. In general, the collinearity between variables is small.

4 Empirical Analysis of Bank Product Subscription

4.1 Model Construction Based on Random Forest

4.1.1 Parameter Tuning

Random forest mainly uses the following parameters: n_estimators: the number of decision trees; max_features: the number of subfeature variables extracted; max_depth: maximum depth corresponding to the decision tree; min_samples_spilt: The minimum number of samples needed for the internal node to classify again.

This paper mainly optimizes two parameters, n_estimators and max_features, in the parameters of random forest. n_estimators are the decision tree established in random forest, and max_features is the number of sub-feature variables that can be extracted from all feature variables for each tree. After determining the number of selectable features of a single tree and the number of model sub-processors, the model is roughly set up, and then the optimal result is obtained by traversing the remaining parameters [5]. The bootstrap sampling method was used to extract training data with put back.

4.1.2 Selection of n_estimators

The generalization error is stabilized by increasing the number of decision trees in the forest.The values of other parameters are fixed, and the value of max_features is the default value of the model parameter, namely auto. Observe the change of the error rate with n_estimors, and the n_estimators value when the error rate tends to steady state are obtained. Taking MSE as the evaluation standard, comparing the single decision tree, the results of bagging method and random forest are shown in the figure below.

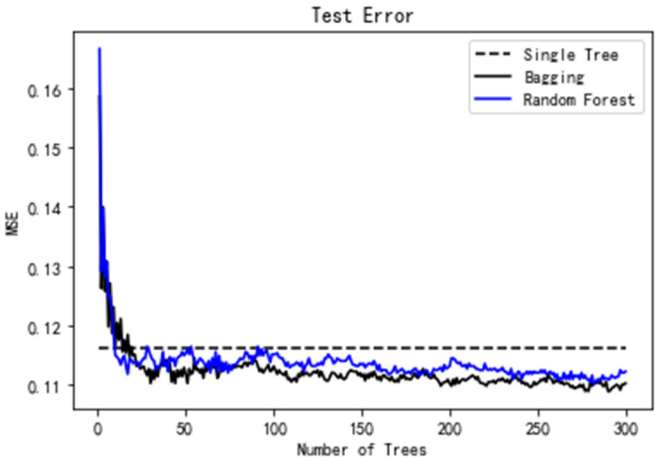

Figure 3. Variation of test error with n_estimatorst

As shown in Figure 3, when the decision tree reaches 300, the test error of random forest has converged and is close to bagging method, while the test error of both methods is significantly lower than that of the optimal decision tree.

4.1.3 Selection of max_features

The total number of decision trees in the random forest is fixed to 300, the values of other parameters are fixed, and 10-fold cross validation is performed for all possible values of max_features. The training data set of bank customers' subscription products is divided into 10 different groups, each group is trained, and the forecast results of each group are averaged to reduce the possible errors caused by different data subsets and groups. Taking the negative mean square error as the evaluation criterion, GirdSearchCV grid search was used to obtain the cross verification graph of max_features, as shown below:

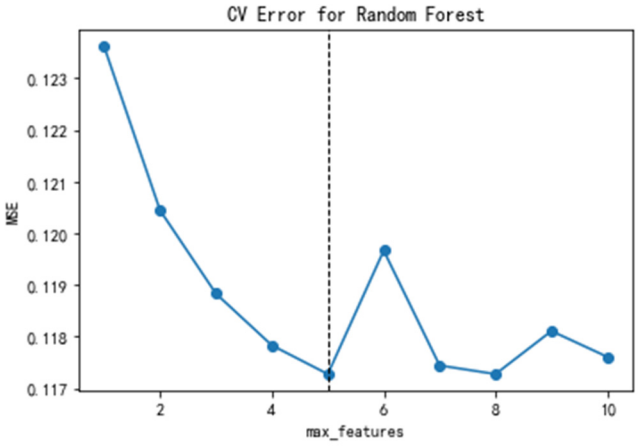

Figure 4. Cross verification graph of max_features

As shown in Figure 4, you can see from the figure above that max_features have an optimal value of 5. The optimal maximum number of split features equal to 5 and the number of decision trees equal to 300 are input into the new random forest construction parameter, and the other two parameters are automatically tuned using GirdSearchCV grid search. The final two parameters max_depth and min_samples_spilt are 9 and 12, respectively.

4.2 Analysis of Experimental Results

By using python software to construct a binary model of random forest based on experimental data, the research results are as follows:

4.2.1 Prediction Results

Let P represent the total number of samples purchased by the original customer and N represent the total number of samples not purchased by the original customer. Comparing the predicted result of the optimized Random Forest with the real result of the customer, there are roughly four scenarios. As shown in the following table:

Table 2 Four scenarios for predicting results

True Category			
		Customer purchases	Customers don't buy
Predicted Categories	Customers buy	TP	FN
	Customer not buying	FP	TN

As shown in Table 2 , TP represents the number of samples that were actually purchased by customers and predicted correctly by the model, and FN represents the number of samples that were actually not purchased by customers but predicted incorrectly by the model. Similarly, FP represents the actual number of samples that customers did not purchase and were correctly predicted by the model; TN represents the actual number of samples that customers purchased but were incorrectly predicted by the model. In accordance with the above situation, the following evaluation indicators are selected to analyze the results:

(1) True rate (TPR) represents the proportion of customer purchases that are correctly classified: TPR=TP/(TP+FN)

(2) True negative rate (TNR) represents the proportion of customers who do not buy items that are correctly classified: TNR=TN/ (TN+FP)

(3) accuracy: represents the percentage of all samples that were correctly predicted: accuracy= (TP+TN) /(P+N)

4.2.2 Experimental Results

After the evaluation indexes are selected, the values of the three evaluation indexes are finally obtained according to the random forest model constructed above, as shown in the table below:

Table 3 Three indexes of experimental results

TPR	TNR	accuracy
97.12%	28.13%	89.18%

As shown in Table 3, the accuracy rate of the test set sample of the randomized forest algorithm with optimized parameters is 89. 18%, and the true rate is 97. 12%, indicating that 97. 12% of customers in the test set are predicted to subscribe to bank products and eventually choose to purchase them, while 2.88% of customers in the rest test set are predicted to subscribe to bank products but do not purchase them. The true negative rate is 28.13%, indicating that 28.13% of the customers in the test set were predicted not to subscribe to the products and did not purchase them, and the remaining 71.87% of the customers were predicted not to subscribe to the bank products but did purchase them. According to these indicators, the random forest constructed in this paper has good performance.

4.2.3 ROC Curve

This paper also chooses to measure the performance of the model with an ROC curve, which is a graphical method of plotting true rate (TPR) on the vertical axis and false positive rate (FPR=1 - TPR) on the horizontal axis by changing the critical values used to create the confusion matrix

between 0 and 1 [4]. The ROC curve can be used to select a threshold for the classifier that maximizes the true rate and minimizes the false positive rate, and plots the trajectory of TPR and FPR under different thresholds. The area under the ROC curve is the value of AUC, and the higher the value of AUC, the higher the accuracy of model classification. The ROC curve drawn according to the experimental results is shown in the figure below:

As shown in Figure 5, the optimized random forest has a high AUC value of 0.89 on the test set, and the overfitting phenomenon is reduced after the tuning, indicating that the optimized random forest algorithm has a better classification effect.

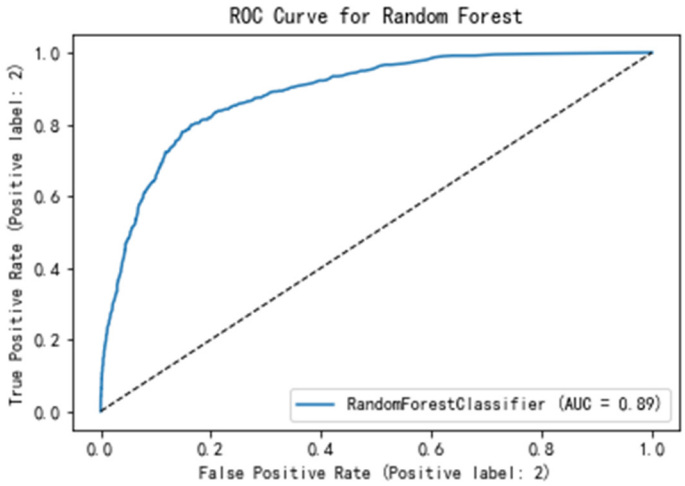

Figure 5. ROC curve of random forest tuning

4.3 Comparative Analysis of Empirical Results of Different Algorithms

4.3.1 Classification Accuracy

In order to verify the advantages of the optimized random Forest in the product dichotomy of bank customers, four other models are selected in this paper. Including KNN algorithm, logistic regression, support vector machine and single decision tree model, the experimental results are shown in the table below:

Table 4 Comparison of classification effects of different algorithms

	TPR	TNR	accuracy
KNN	96.85%	14.23%	86. 15%
Logistic	99.08%	6.86%	87. 13%
SVM	98.80%	9.78%	87.26%
Single tree	96.50%	25. 18%	87.91%
RandomForest	97. 12%	28. 13%	89. 18%

As shown in Table 4, the TPR values of these five models are all high, above 96%, indicating that they have high classification accuracy in predicting that customers will buy bank products, among which the Logistic model has the highest TPR, reaching 99.08%.However, the TNR of

these five models is relatively low, all less than 30%, indicating that their classification accuracy of predicting that customers will not buy bank products is low, among which the TNR of random forest model is the highest, with a TNR value of 28. 13%.

The accuracy rate of the random Forest algorithm reached 89. 18%, which was the highest compared with the other four models, indicating that the accuracy rate of the random forest model on the whole sample was basically 90%, while the accuracy rate of the KNN model was 86. 18%, which was the lowest compared with the other four models, but also more than 85%, indicating that its prediction accuracy of the bank product subscription was good. SVM and Single tree are close, 87.26% and 87.91% respectively, ranking the second and third in accuracy. Therefore, random forest algorithm has a good accuracy rate in the research of bank product subscription prediction. It can effectively process complex data, accurately predict the trend of customers' subscription of bank products, and provide high-quality services for banks.

4.3.2 Comparative Analysis of ROC Curve

Similarly, in order to further judge the performance of each model, for KNN algorithm, logistic regression, support vector machine and single decision tree binary classification model, ROC curve is also selected in this paper for comparative analysis with random forest. The resulting ROC curve pair is shown in the following figure:

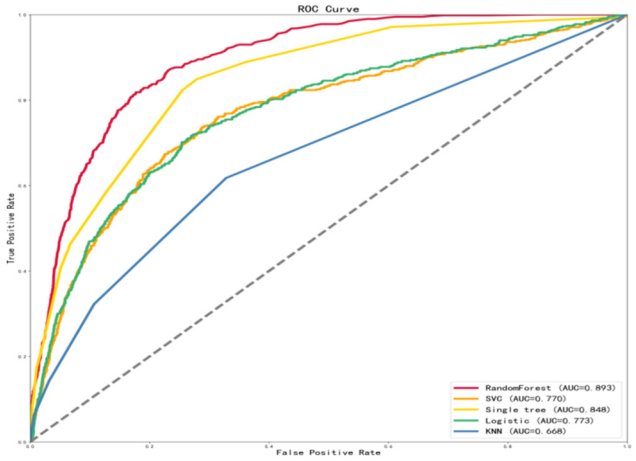

Figure 6. Comparison of ROC curves of different model

As shown in Figure 6, Random Forest has the largest AUC value of 0.893, while KNN has the smallest AUC value of 0.668. It indicates that the random forest algorithm has the best classification effect, high accuracy and strong reliability, and can better predict whether customers will subscribe or buy bank products. On the contrary, KNN algorithm has the worst classification effect, low accuracy, and low performance of classifier in predicting customers' research on bank products. The AUC values of Logistic algorithm, SVC and Single tree are all above 0.7, and the accuracy of prediction is high, which can show the feasibility and effectiveness of different types of machine learning algorithms in the classification model of the financial field.

5 Conclusion

In this study, the random forest model is used to forecast the subscription of bank products. According to the data set of Aliyun Tianchi Bank customer subscription products and related literature, the data is processed by python. The random forest algorithm is mainly used for the data set, and the Logistic algorithm, KNN algorithm, SVC algorithm and single decision tree are used to construct the bank product subscription prediction model. By comparing the performance differences of different types of machine learning algorithms in the same data set, this study provides a reliable method for banks to forecast customer subscription products.

The research conclusions of this paper are as follows:

(1) It can be seen from the experimental results that with AUC as the model performance evaluation index, the stochastic forest algorithm has the highest model classification performance, with a value of 0.893, 0.285 points higher than Logistic algorithm, and the model has the best classification performance. Random forest can process large scale data effectively and with high accuracy. It can improve the accuracy of classification and regression by randomly sampling the characteristics of the data set and then classifying and regression through a series of decision trees. Its classification accuracy is higher than traditional decision trees, and it is not easy to be affected by noisy data.

(2) Random Forest has the highest accuracy in predicting the subscription of bank products. It can comprehensively consider multiple attributes, learn from large amounts of data, and make intelligent analysis and prediction. It can effectively predict the subscription of bank products, provide accurate reference for enterprises, improve the subscription rate and realize the growth of enterprises. It can train multiple different decision trees in a sample set many times, and get the most accurate results from it, which is higher than other machine learning algorithms.

(3) Logistic algorithm has the highest TPR value of 99.08%, indicating that after the same data is preprocessed, there are differences in different types of machine learning algorithms. In practical business applications in the financial field, appropriate machine learning algorithm models should also be selected according to different problems.

The advent of the era of big data brings challenges to the information capability of banks. The potential value of big data to banks ranks first in all industries, which also shows that the arrival of big data has brought greater potential for commercial banks to "gold mine". In order to make a fortune in the era of big data, banks need to apply processing methods suitable for big data, such as selecting appropriate machine learning algorithm models, which will provide scientific guidance and reference for banks to further make operational decisions on subscribed products.

Author Contributions

This paper was jointly completed by Wei Tao, Zou Linying, Sun Yanze and Zhao Shengfa. In the research of this subject, Wei Tao, Zou Linying and Sun Yanze made the same contribution to the paper, and the three people wrote it together. Zhao Shengfa is the second producer. It is hereby explained.

References

[1] Huiyu Wang. Predicting the Impact of Marketing Activities on Customer Ordering Time Deposits Based on Machine Learning: Nankai University, 2021 (in Chinese)

[2] Ling Tang, Anying Wang, Zhenjing Xu, et al. Online-Purchasing Behavior Forecasting with a Firefly Algorithm-Based Svm Model Considering Shopping Cart Use. Eurasia Journal of Mathematics, Science and Technology Education, 2017, 13(12): 7967～7983

[3] Pengyi Zhang, Dan Xue Wang, Yi Fan Jiao, et al. Research on Mobile Purchase Prediction Based on User Browsing Log. Journal of Data Analysis and Knowledge Discovery, 2018, 2(1): 51～63 (in Chinese)

[4] Yue Hu. An Empirical Study of Financial Credit Risk Control Model Based on Machine Learning Algorithm: Northern University for Nationalities, 2022. (in Chinese)

[5] Yi Sheng Yuan. Small and Medium-Sized Enterprise Credit Risk Assessment Based on Random Forest Algorithm: Shan Dong University, 2021. (in Chinese)

[6] Leo Breiman. Random Forests. Machine Learning, 2001, 45(1): 5～32

[7] Ya Wen Kang. Grade Evaluation Model of Medical Industry Suppliers Based on Random Forest: Anhui University, 2017. (in Chinese)