Application of machine learning algorithm in bank telephone marketing

Wenxi Zhang*

College of Science, China Agricultural University, Beijing, 100089, China

*Corresponding author: z18618158699@163.com

Keywords: Machine learning, Bank telemarketing, data analysis, xgboost

Abstract: In recent years, machine learning algorithms have been widely used in banking. This paper uses a variety of classic machine learning algorithms to model and analyze the marketing data of a Portuguese banking institution to predict whether customers will subscribe to the deposit service. Among those algorithms, the xgboost model is the best, and its prediction accuracy is **94.24%**. Using this model, we can build a reasonable product marketing system for the bank.

1. Introduction

Customer is the basis of a commercial bank's benefit. In the face of fierce industry competition, how to effectively use data for analysis is of great significance to improve the bank's efficiency and better serve the customers. In the era of big data, machine learning is an effective way to get valuable information from data. Using machine learning techniques can promote the development of business effectiveness.

Telemarketing is a traditional business of banks. In recent years, there are many cases of using machine learning algorithms to improve the success rate of telemarketing. For example, in 2016, Mingyue Li [1] proposed an improved classification decision tree algorithm, and applied it to telemarketing data sets. In 2018, Yiyan Jiang[2] used classification algorithms such as Bayesian and logistic regression to predict the best customer groups for bank telemarketing, the results of the experiment can guide the development of banking business.

The data set used in this paper is based on the telemarketing activities of a Portuguese bank, classified to predict whether customers will buy bank time deposits. In 2017, Xiaoqian Huang[3] construct five models for this data set and found that the ANN model has the best prediction effect, and the accuracy reaches 92.4%.

Therefore, based on the previous work of researchers, this paper is divided into the following parts: the first part is the introduction of the project data set; The second part is the visualization charts of the marketing result of each variable after data preprocessing; The third part is data processing, including the coding of classification features and the processing of unbalanced data, etc.; The fourth part is the model building including five classical algorithms to test the performance in this data set. The best prediction model is the xgboost model, and its prediction accuracy is 94.24%. In addition, I use the AutoGluon framework for automatic machine learning, which also achieved good results; The final part is the result analysis, I compared the models together, then put forward a reasonable suggestion to the bank's marketing strategy.

2. Introduction to datasets

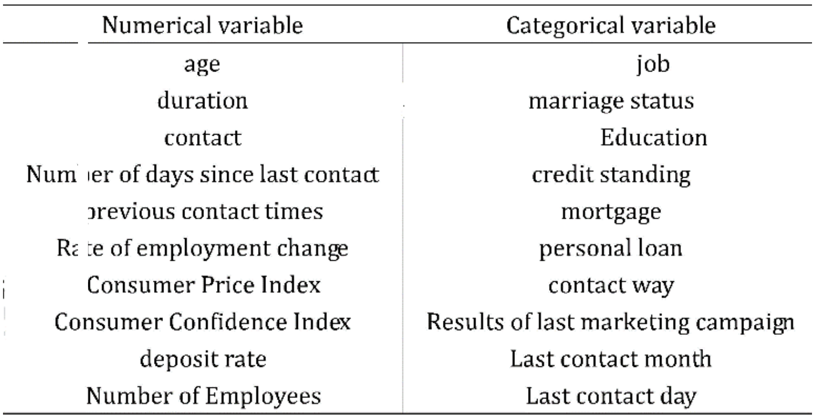
This part will give a basic introduction and analysis of the experimental environment, basic data structure, including the size of the data set, the type of variables, and the business meaning of indicators.

2.1 Basic data structure

This project is based on the Windows platform. The programming language is Python and the integrated development environment is PyCharm. The experimental data is from Alibaba Tianchi platform-bank marketing dataset.

The dataset is 5.56 Mb in size and contains 41,188 pieces of data, with 20 metric variables. Categorize variables by data type, including 10 numeric variables and 10 categorical variables, as shown in the following table:

Table.1. Variable types



Numerical variable	Categorical variable
age	job
duration	marriage status
contact	Education
Number of days since last contact	credit standing
previous contact times	mortgage
Rate of employment change	personal loan
Consumer Price Index	contact way
Consumer Confidence Index	Results of last marketing campaign
deposit rate	Last contact month
Number of Employees	Last contact day

2.2 Target data

Y(Deposit): whether the customer has purchased a time deposit (category: 1 Yes; 0 No)

Among the 41,188 customers, only 4,640 have fixed deposits and 36,548 don’t have. We can see that the sample distribution is not balanced.

3. Data pre-processing

This section will visualize the data after preliminary processing. I have made a large number of visual charts, which can help readers to have a more intuitive understanding of the relationship between variables, but also more conducive to the screening and processing of follow-up features.

3.1 Data pre-processing

In the original dataset, the dataset is complete and no data padding is required. But all the indicators are stacked in one column and need to be broken down first. After it, I import the new mybank.csv file.

3.2 Data description analysis

Generating a visualization chart, can fully depict the customer group portrait, the results are as follows:

- **Age**

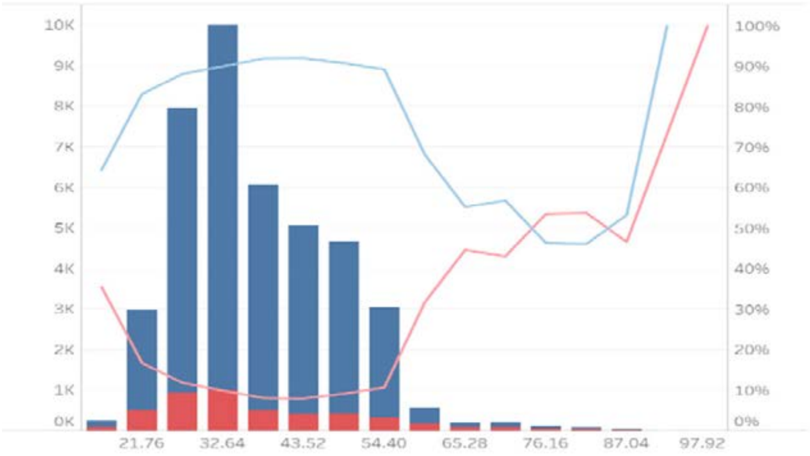


Figure 1. Number of clients of all ages

As shown in Figure 1, the histogram shows the number of customers in each age group, and the line chart shows the percentage of deposit business. Blue represents a non-fixed deposit; red represents the fixed deposit business. It was found that most of the customers were aged from 25 to 55, and the number of customers in the older age group dropped precipitously, which required further data processing Looking at the line chart, we can see those younger people under the age of 20 and older people over the age of 60 have a higher proportion of time deposits, so they can be considered as a key marketing group.

- **Job**

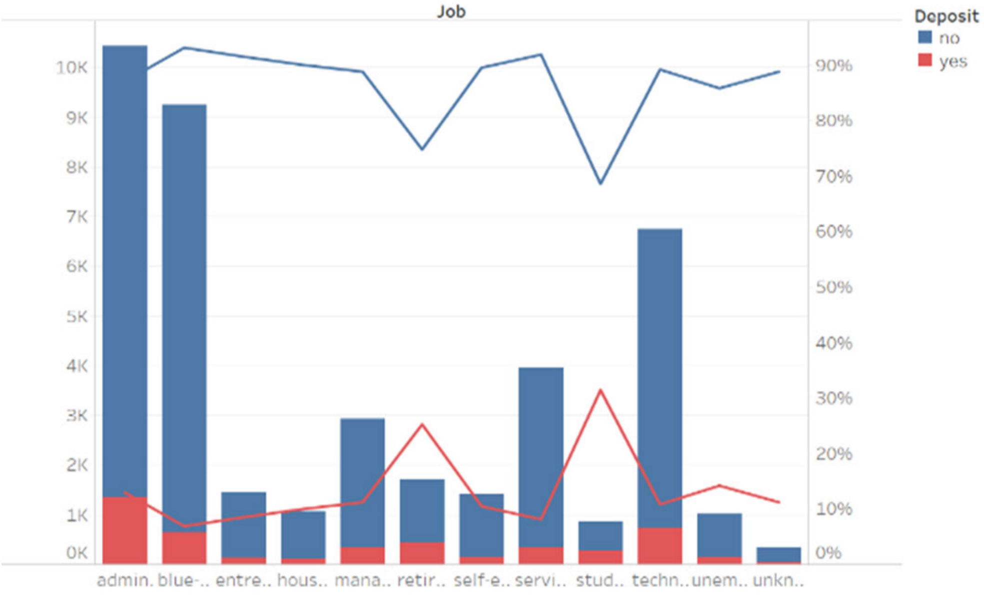


Figure 2. Number of clients for each job

From the chart, we can see the proportion of students and retired people handling savings is higher, So the bank can mainly market to these two groups.

- **Duration**

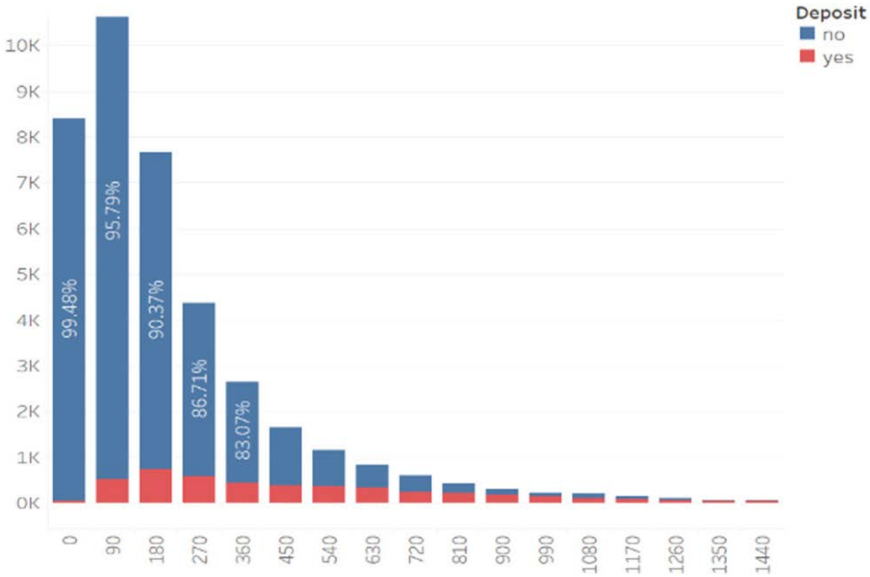


Figure 3. Historical Call Duration

In figure 3, the horizontal coordinate is the different length of the historical call time, we can see: the longer the call time, the higher the proportion of customers with a fixed deposit. It can be used as the key classification index of modeling.

- **Poutcome**

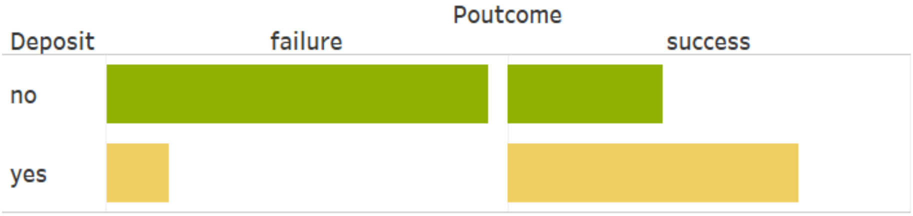


Figure 4. The impact of marketing results

Looking at Figure 4, it is found that successful customers are more likely to transact business if the activity results in "success". Therefore, we can consider the successful customers as the marketing target group, at the same time, we can work out a reasonable marketing strategy to improve the success rate of marketing activities.

By doing a descriptive analysis of these 20 indicators, in turn, readers can gain a deeper understanding of the relevance of the indicators and results: for example, loan status, marital status, and so on, do not have a greater impact on whether customers transact business, these indicators can be removed later, and age, education, occupation, the last marketing time and results have a strong correlation with whether customers deposit.

In statistics, covariance is often used to describe the degree of similarity between two random variables. Using the formula, the covariance matrix between the indicators can be plotted. From it, the following conclusions can be drawn:

- The correlation between economic indicators is strong;
- There is a strong correlation between pre-existing, contact, and economic indicators, so it is reasonable to assume that bank deposit marketing will improve in good times.

4. Data Processing

Data processing is a very important step before model building, it will greatly affect the speed, accuracy, and other performance of the model. The specific handling methods are as follows.

4.1 Elimination of weakly correlated variables and outlier samples

From the data description analysis, it can be seen that the housing loan and the last contact day have very little effect on the result, so they can be removed. Besides, the outliers of age and duration are also eliminated to avoid disturbance to the model.

4.2 handling of categorical variables

The type variables in the data set are out of order, so I first converted them to data types, which is convenient for subsequent processing algorithms. Such as age, education, job, marital status, can be manually encoded, or the Label Encoder function can be used to assign a label between $0\sim\mathrm{n}_{\mathrm{class}}$-1 encoding ($n_{class}$ represents the number of variable categories).

4.3 handling of dense numerical variables

For the length of the call and some economic indicators, the numerical values are relatively concentrated and can be inter-regionalized appropriately. Specifically, by calculating the four-equilateral points of the variable, for example, the value is less than 0.25 quantiles, the notation code is 1, between 0.25~ 0.5 quantiles, coded as 2, and so on.

4.4 handling of unbalanced data

In machine learning, many models have a basic assumption about the distribution of data. If the distribution of data is too different from the assumption distribution, the model can't perform well[4]. To solve the problem of unbalanced sample distribution in the dataset, this paper uses the Technique of synthetic Minority Oversampling to balance the sample weight. After processing, the number of training set samples is 48764, the ratio of positive and negative samples is 1:1, and the remaining 18 features.

5. Model building

In this part, the author has experimented with many kinds of machine learning methods and has given a concrete graph demonstration of the model key parameter optimization. By adjusting the parameters, the performance of each model has been greatly improved.

5.1 KNN algorithm

5.1.1 Introduction to models

KNN algorithm is a common classification algorithm in supervised learning. Its core idea is that if a sample belongs to the same class in most of the K most adjacent samples in the feature space, then the sample also belongs to this class and has the features of the samples in this class. In the KNN algorithm, the similarity between samples is measured by calculating the distance between samples.

5.1.2 K value selection

The key step in KNN algorithm is to determine the number of "Neighbors", that is, the selection of the K value. Therefore, this paper compares the prediction accuracy of KNN models under each K value, and draws a line graph between different K values and average prediction accuracy, as shown in Figure 5:

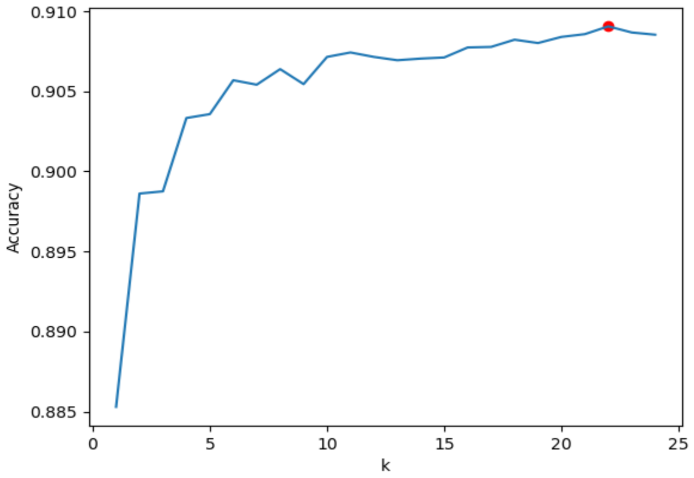


Figure 5. The effect of K value on accuracy

It is found that when K=22, the accuracy is the highest.

5.1.3 Model assessment

When K = 22 is substituted into KNN model, the prediction accuracy on the test data set is 93.76%. The training time is 21.40 seconds.

5.2 Logistic regression models

5.2.1 Introduction to models

Logistic Regression is a classical classification model in machine learning. Because of its simple and efficient algorithm, it is widely used in practice. The LR model can be thought of as a linear regression model normalized by the Sigmoid function, which assumes:

$$P(y = 1 \mid x; \theta) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T * x}} \tag{1}$$

When we assume the function $P(y = 1 \mid x; \theta) \geq 0.5$, we predict a positive class and a negative class.

5.2.2 Model assessment

Considering the small training data set of this paper, using solver = ‘liblinear' as the optimization algorithm, substituting the LR model, the prediction accuracy on the test data set is 94.06%, the training speed is very fast, only 0.13s, the model works well.

5.3 Support vector machine

5.3.1 Model description

The support vector machine is a binary classification model, the idea is to find a hyperplane to segment the samples. The most classical SVM model for solving the maximum separation hyperplane problem can be expressed as the following constrained optimization problem[5]:

$$\min_{w,b} \frac{1}{2} |w|^2 \tag{2}$$

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1 \mathrm{i} = ,2 \dots , N \tag{3}$$

However, the practical problems are generally nonlinear and indivisible, so researchers hope that nonlinear transformation can be used to transform nonlinear problems into linear problems[6].

The general approach is to map the training sample from the original space to a higher-dimensional space. In this process, the selection of the SVM kernel function is very important. The kernel function is used to make the data separable in the feature space by mapping the linear indivisible data into a high-dimensional feature space.

5.3.2 Model assessment

In this paper, three common kernel functions are used to test the performance of SVM model:

- Gauss, Radial basis function

$$K(x, xi) = \exp\left(-\frac{|x - xi|^2}{\delta^2}\right) \tag{4}$$

This is a local kernel function, which can map a sample to a higher dimensional space. This kernel function has good performance on different size data sets.

Using the RBF kernel function, the prediction accuracy is 89.58% and the training time is 206.28s.

- Nonlinear functional kernel Sigmoid of a neuron

$$K(x, x_i) = \tanh(\eta < x, x_i > +\theta) \tag{5}$$

Using the Sigmoid kernel, the Support vector machine implements a multilayer neural network. The prediction accuracy on the test set was 88.77%, and the training time was 18.66 s.

- SVM algorithm based on RBF kernel function using pipeline

The overall performance of the model is improved compared with the previous two methods. The prediction accuracy on the test set is 94.13% and the training time is 19.47 s.

5.4 Random forest models

5.4.1 Introduction to models

Random Forest is an algorithm that integrates many trees by ensemble Learning, its basic unit is the decision tree. From an intuitive point of view: each decision tree is a classifier (assuming that this is now the classification problem), then for an input sample, n trees will have n classification results. The random forest integrates all the voting results of the classification and specifies the category with the most votes as the final output.

5.4.2 Parameter adjustment

The number of base evaluators is considered to be the number of trees in the forest. The bigger the number, the better the model tends to be. So, we first draw a learning curve for N to find the best parameter, as shown in Figure 6:

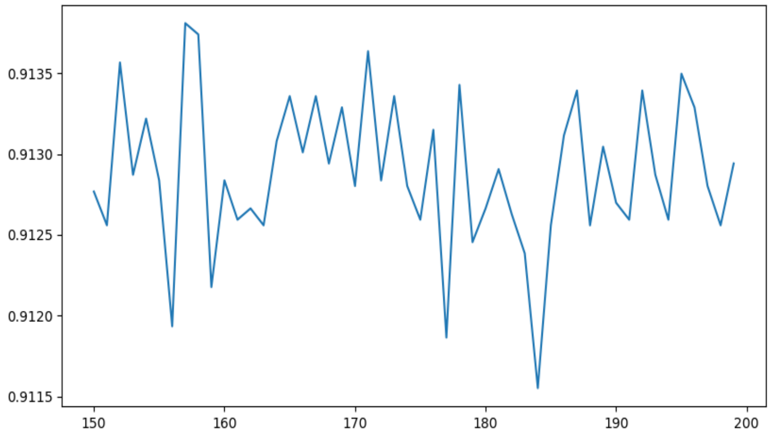


Figure 6. The effect of N on accuracy

It is found that the prediction accuracy is maximized when N is used. But this way of finding the optimal parameters takes a long time.

At this time, the prediction accuracy on the test data set is 93.50%, the single training time is 2.76 s, and the model effect is good.

5.5 xgboost optimization algorithm

5.5.1 Introduction to models

Boosting is an ensemble algorithm that combines many weak classifiers $f_i(x)$ to form a strong classifier $F(X)$, it iterates over the new learner via the gradient descent. In this paper, the xgboost optimization algorithm is based on GBDT (gradient boosting decision Tree) based on the improvement of boosting algorithm, can achieve parallel processing, compared to GBDT has a speed leap. The core idea is:

- Adding one tree at a time, corresponds to learning a new function $f(x)$ to fit the last forecast residuals;
- After training k trees, we need to predict the score of a sample. According to the characteristics of this sample, each tree will fall to a corresponding leaf node, each leaf node corresponds to a score;
- Finally, the predicted value of the sample is simply the sum of the scores corresponding to each tree[7].

Our goal is to make the predicted value $\widehat{y_l}$ as close as possible to the real value y_i, and has a good generalization ability.

The following formula defines the target function for xgboost:

$$L(\phi) = \sum_i l(\widehat{y_l} - y_i) + \sum_k \Omega(f_t) \tag{6}$$

The above expression $\widehat{y_l}$ is the output of the whole summation model and the sum $\sum_k \Omega(f_t)$ is the function of the complexity of the tree.

5.5.2 Parameter adjustment

The xgboost algorithm involves many parameters. First, we optimize the performance of the trainer by adjusting the parameters.

By adjusting the parameters in order of importance, an optimal set of parameters can be obtained:

$$\text{learnin}g\_r\text{ate} = 0.1 \quad \text{n\_estimators} = 320$$

$$colsample\_bytree = 0.9 \quad \text{subsample} = 0.6$$

$$max\_depth = 3 \quad min\_child\_weight = 5$$

$$\text{reg\_alpha} = 3 \quad reg\_lambda = 0.5$$

5.5.3 Model assessment

After updating the optimal parameters, the model is established and predicted, and the prediction accuracy on the test set is 94.24%, training time is 4.56s. It can be seen that the accuracy of the algorithm is improved compared with other models, and the training speed is very fast.

5.6 AutoGluon framework

5.6.1 Technical principles

AutoGluon describes the framework on its website as follows: AutoGluon supports an easy-to-use extended automatic machine learning, meaning that the technology can automatically extract features from the data and train the model.

The main techniques it uses are:

- Technique 1: Stacking

Different models can be trained independently on the same data, such as simple KNN, tree model, kernel method, or complex neural network. The outputs of these models are put into a linear model to get the final output, which is the weighted sum of these outputs.

- Technique 2: K Cross bagging

Bagging means training multiple models of the same type, which may use different initial weights or different data blocks, and then averaging the outputs of these models.

- Technique 3: multilayer stacking

The principle is to combine the output and data of each model to do stacking again, that is to say, based on this, to train more models, and finally to do the output with a linear model.

5.6.2 Model assessment

Using AutoGluon, we get the prediction accuracy of 94.26% on the test set., the total training time of the model is 72.38s. The training is good.

6. Results analysis

For this project, we used five classical machine learning algorithms to test the prediction accuracy and training time on each model, and the results are shown in the follow table:

Table.2. Effects of models

model	accuracy	The training time
KNN	93.76%	21.40s
Logistic regression	94.06%	0.13s
SVM(RBF kernel)	94.13%	19.47s
Random forests	93.50%	2.76s
Xgboost algorithm	94.24%	4.56s
AutoGluon framework	94.26%	72.38s

The comparison shows that the xgboost optimization algorithm and the AutoGluon framework work best, each with its own merits: For the boost optimization algorithm, the single training speed is very fast, but adjusting parameters is complicated; and the AutoGluon framework can help us automatically optimize, but each training time is long. In addition, the support vector machine and logistic regression models added to the pipeline performed better on the dataset.

The xgboost optimization algorithm is also interpretable, providing automatic estimates of the importance of features from well-trained prediction models[8]. So, this article uses a built-in function provided in the xgboost library to draw the following bar graph in order of importance:

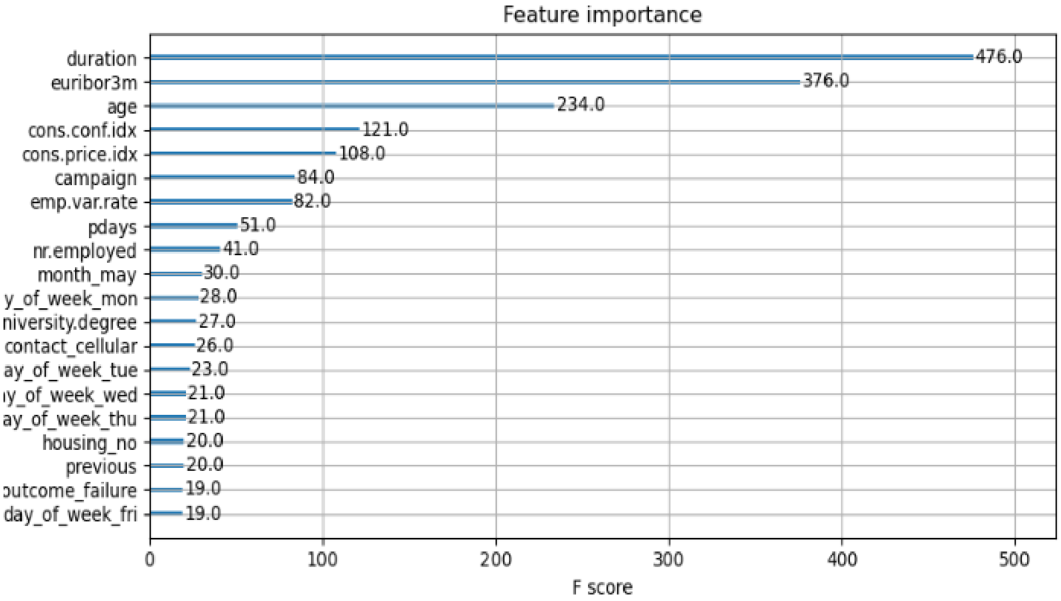


Figure 7. Feature importance ranking

With the help of the graph above, it is found that duration is the feature that has the greatest impact on the results, which is consistent with the conclusion of data description. On top of that, cpi, cci, job, education also have a great impact on whether customers will buy bank deposit products. It is concluded that the current economic situation of society has a greater correlation with the result, and it can be inferred that the higher the customer's education, the higher the degree of attention to financial management, and the higher the probability of purchasing products.

To sum up, the bank can build up their product marketing systems, such as the smart push of deposit products to highly educated and stable customers during periods of economic prosperity. On the one hand, the bank employees can improve their work efficiency, and save marketing time; on the other hand, the target customers can also better meet their financial needs and promote the sustainable and healthy development of the social economy.

References

[1] Mingyue Li. The application of decision tree algorithm in bank telephone marketing. Huazhong University of Science and Technology, 2016.

[2] Yiyan Jiang. Using Logistic Regression Model to Predict the Success of Bank Telemarketing [C]. International Journal on Data Science and Technology. April 1,2018:7-9

[3] Xiaoqian Huang. Using data mining technology to improve bank promotion success rate [J]. Information and computers (theoretical edition), 2017(09): 133-135 + 140.

[4] Junchen Wang. Bank user credit risk analysis based on machine learning algorithm [D]. Nankai University, 2021.

[5] Statistical Learning Methods [M]. Tsinghua University Press, Hang Li, 2012

[6] Machine Learning [M]. Tsinghua University Press, Zhihua Zhou, 2016

[7] Tianqi Chen et.al, XGBoost: A Scalable Tree Boosting System, 2016.

[8] Jason Brownlee. Feature Importance and Feature Selection with XGBoost in Python,2016.

Appendix

Data set download address:
https://tianchi.aliyun.com/dataset/dataDetail?dataId=92231#1
The technology behind AutoGluon -- Follow Li Mu's AI