Bank Customer Churn Prediction

Hao Tan
th13320348510@163.com

College of Mechanical Engineering, Chongqing University of Technology, Chongqing, China, 400054

Abstract—Nowadays, with the rapid development of Internet finance, the competition in the banking industry is getting fiercer and fiercer. How to prevent the loss of customers and retain old customers has gradually become an important issue of concern for major banks. In this paper, descriptive statistical analysis of each feature is carried out according to the customer data set of a bank. After a preliminary understanding of the data, data preprocessing is carried out, including data cleaning, feature selection, data transformation, etc. Then, Random Forest and Logistic Regression supervised learning models are used for training and Grid Search is used for reference adjustment. The trained model was evaluated by ROC & AUC, and finally, suggestions were made for the bank to retain customers based on descriptive statistics and the importance of features. Finally, this paper finds that Random Forest is the best model for predicting. At the same time, according to the importance of the model's influence on bank customer churn, it is concluded that the top three factors with the greatest influence are: age Estimate Salary and Credit Score.

Keywords-Random Forest, Logistic Regression, Grid Search, ROC&AUC.

1. Introduction

It is particularly important for banks to develop high-quality customers, enhance their own value and prevent customer loss. Studies show that it costs banks five to six times as much to develop a new customer as it does to maintain an old one [1]. Therefore, compared with developing new customers, preventing customer loss is a more accurate choice, not only because the cost of developing new customers is high, but also because most of the old customers bring much higher profits than the new customers. Therefore, it has become an important issue for all enterprises to predict the potential loss of bank customers, care for them and take effective measures to retain them. Logistic Regression in the supervised learning model is selected to analyze the data set from a European bank customer [2]. Through training and using Gird Search we can find the best parameters to make the model perform the best, so as to provide the above-mentioned banking companies with data and references.

After the data is imported locally, the first step is to explore and preprocess it. After that, the Numerical feature and categorical feature are described and the useless features are deleted to obtain the feature space. Then, the data set is separated into the training set and the test set. Stratified sampling is used to avoid extreme examples. The second step is to build and train the supervised learning model, which is mainly Random Forest and Logistic Regression. The third step is to use Cross Validation and Grid Search to find the optimal parameters of models. The fourth step is to evaluate the performance of Random Forest and Logistic Regression using ROC&AUC. The last step is to use Random Forest Classifier and Logistic Regression to carry out Feature Importance Discussion.

BDEIM 2022, December 02-03, Zhengzhou, People's Republic of China

DOI 10.4108/eai.2-12-2022.2328745

2. Literature Review

In the research of qq_44451414, the author chose the XGBOOST model for training, while in this paper, he chose Random Forest and Logistic Regression for training, with more emphasis on the application of Random Forest [1]. XGboost is the most popular model in most Kaggle competitions today, and Random Forest, known as the two most powerful classification algorithms, also outperforms Random Forest in many ways [3]. One of the reasons why Random Forest is chosen in this paper is that Random Forest has a certain degree of randomness, which means that Random Forest can avoid overfitting. At the same time, compared with a separate tree model, the generalization ability is also improved, and if a certain data set the feature dimension is very low, and the scale is not large at the same time. If it is relatively small, overfitting may occur when using XGboost, and Random Forest is more friendly than XGboost when tuning parameters, so in general, even if XGboost performs better than Random Forest in most aspects, based on this data set, this paper chose the latter.

3. Data exploration and pre-processing

3.1 Data Exploration

First of all, the data set used in this paper is provided by a European bank [2]. There are 14 variables and 10, 000 sample data in this data set, which does not contain missing value. The variables contained in the data set from left to right are line number, customer id, name, credit score, region, gender, age, duration of user purchase of bank products, deposit, product quantity, whether the user contains a credit card, whether he is an active user, estimated income, and whether the customer has lost. All the data characteristics are summarized, and the details are shown in Table 1.

As can be seen from Table 1, the variable 'Exited' is the target value, which is defined as whether the customer has been lost and the variable is categorical feature. Therefore, the target of data mining is categorical. Besides the target variable, the feature attribute also contains 7 categorical variables. Among them, 'RowNumber', 'CustomerId', 'Surname', 'Geography' and 'Gender' are character type, which means that during data preprocessing, these variables should be processed first before the next analysis can be conducted. The remaining 6 variables are Numerical feature, and it is only necessary to consider whether there are outliers and whether data transformation is needed.

Table 1 Characteristics of Data

Column number	Attribute	Type
1	Row Number	categorical
2	Customer Id	categorical
3	Surname	categorical
4	Credit Score	Numerical
5	Geography	categorical
6	Gender	categorical
7	Age	Numerical

8	Tenure	Numerical
9	Balance	Numerical
10	Num Of Products	Numerical
11	HasCrCard	categorical
12	Is Active Member	categorical
13	Estimated Salary	Numerical
14	Exited	categorical

For the Numerical feature, this paper divides each of the Numerical feature into two groups, namely, lost and unlost, and observes the distribution difference of its data in each group through boxplot to see which factor has the greatest influence on the loss of customers. After analysis, the author finds that Credit Score, Num of Products and Estimated Salary have little impact on the loss of customers, because their data distribution in the figure is very similar. The box-plot is shown in Figure 1.

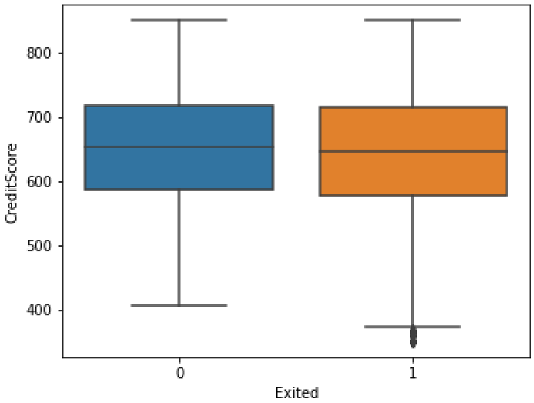


Figure 1. Numerical feature statistics (Credit Score)

Age has some impact. Some of the lost customers are the elderly. But the majority of those who haven’t lost customers are also elderly (as shown in Figure 2).

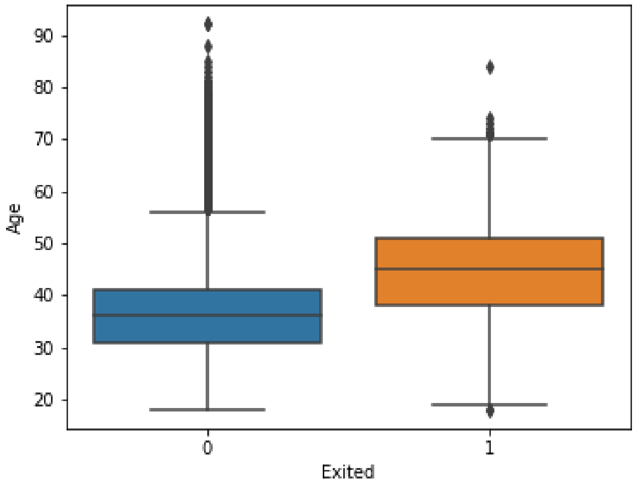


Figure 2. Numerical feature statistics (Age)

Tenure would be easy to lose between 2 and 7 years after purchase, and Balance would indicate that someone with less money in that bank would be less likely to lose (as shown in Figure 3).

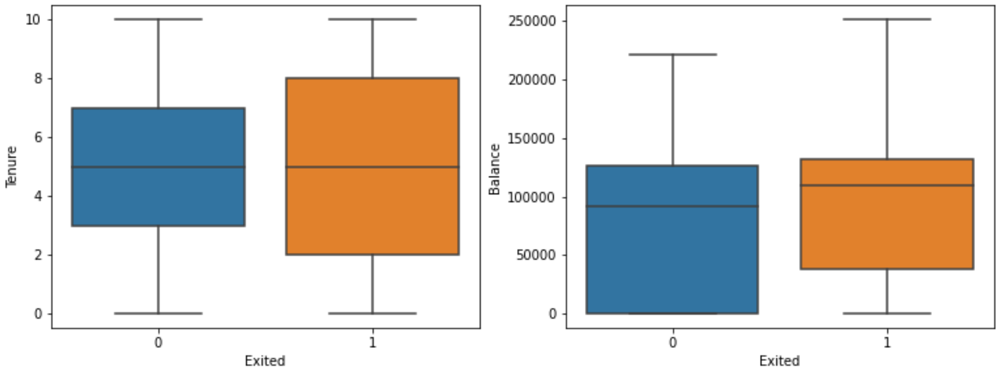


Figure 3. Numerical feature statistics (Tenure and Balance)

As for the categorical feature, according to the statistical chart, the paper can see that compared with the other three properties, the categorical feature shows that the return of bank customers is the largest because of Geography. From the perspective of a single group, for example, the group whose bank customers are not churn, the number of German nationals is the least. But the group that lost customers to the bank actually had the highest number of German nationals, so Germans might have some problems. The statistical chart is shown in Figure 4.

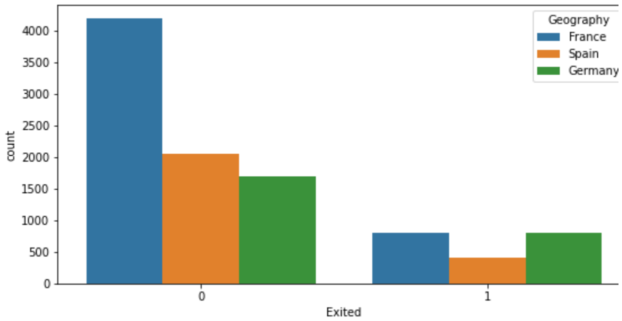


Figure 4. Categorical feature Statistics

3.2 Feature Pre-processing

Among all the variables, the author noticed that there were three variables that had no impact on the research results, namely, Row Number, Customer Id and Surname, which were excluded. The first five rows of data in the data set are shown in Table 2 and Table 3.

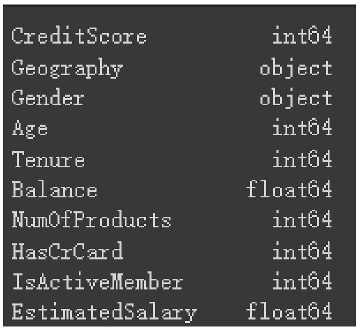
Table 2 Data set after cleaning (1)

Credit Score	Geography	Gender	Age	Tenure
619	France	Female	42	2
608	Spain	Female	41	1
502	France	Female	42	8
699	France	Female	39	1
850	Spain	Female	43	2

Table 3 Data set after cleaning (2)

Balance	Num Of Products	Has Cr Card	Is Active Member	Estimated Salary
0	1	1	1	101348.88
83807.86	1	0	1	112542.58
159660.8	3	1	0	113931.57
0	2	0	0	93826.63
125510.82	1	1	1	79084.1

After the three variables are eliminated, the remaining variables are extracted and grouped. Those whose data type is 'object' are divided into ‘cat_cols’ and those whose data type is 'int64' or 'float64' are divided into ‘num_cols’. The remaining variable data types are shown in Figure 5.



```
CreditScore          int64
Geography           object
Gender              object
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
```

Figure 5. Remaining variable data types

To make preparation for the following Feature Pre-processing, the data set is separated first. In this paper, the test set is selected as 0.25. After the separation, the categorical feature will be processed. The OneHot Encoder in sklearn is directly used to process 'Geography'. After processing, 'Geography' becomes three columns in the data set, respectively Geography_France, Geography_Germany, Geography_Spain. People from three different countries will show 1 under the corresponding nationality, or 0 if not. For 'Gender', OrdinalEncoder is used to process it. After processing, male is displayed as 1 and female is displayed as 0. The result is shown in Table 4.

Table 4 The Data set after OneHot Encoder and Ordinal Encoder

	Gender	**Geography_France**	**Geography_Germany**	**Geography_Spain**
0	1.0	0.0	0.0	1.0
1	0.0	0.0	1.0	0.0
2	0.0	1.0	0.0	0.0
3	0.0	1.0	0.0	0.0
4	0.0	0.0	0.0	1.0

After processing the categorical feature, it starts to treat the Numerical feature. To eliminate the influence of the difference of unit and scale between the features and treat each dimensional feature equally, it is necessary to normalize the features. In this paper, Standardize Data is used. The result of the processing is shown in Table5, 6 and 7.

Table 5 Normalized Dataset (1)

Credit Score	Gender	Age	Tenure	Balance	Num of Products
-0.172985	1.0	0.289202	1.731199	-1.21892	-0.912769
0.602407	0.0	-1.50932	-0.34116	-0.07698	-0.912769
-1.051762	0.0	-0.18409	1.385806	-1.21892	0.796109
-0.048922	0.0	0.857156	0.695022	0.229625	-0.912769
0.819517	0.0	0.573179	0.695022	0.237278	0.796109

Table 6 Normalized Dataset (2)

Has CrCard	**Is Active Member**	**Estimate Salary**	**Geography_France**
-1.542199	0.968496	-0.352044	0.0
0.648425	-1.032529	0.072315	0.0
-1.542199	-1.032529	-0.026711	1.0
0.648425	0.968496	0.866221	1.0
0.648425	-1.032529	-0.496617	0.0

Table 7 Normalized Dataset (3)

Geography_Germany	Geography_Spain
0.0	1.0
1.0	0.0
0.0	0.0
0.0	0.0
0.0	

4. Model Training and Result Evaluation

4.1 Model Architecture: Logistic Regression

Logistic regression is a type of linear classifier. It maps the data features to a probability value in the range of 0 to 1 (the possibility that the sample belongs to a positive example) through the Logistic function (Sigmoid function), and compares the results at 0.5. Get the category to which the data belongs [3][4]. Its function expression is as follows:

$$\boldsymbol{f(z)=\frac{1}{1+e^{-z}}} \qquad (1)$$

$$\boldsymbol{z=w^T x+w_0} \qquad (2)$$

where $\boldsymbol{X}$ is the eigenvector of the data, $\boldsymbol{W}$ is the parameter vector that needs to be fitted to the data, and $\boldsymbol{w_0}$ is the bias.

The function image is a curve whose value range is compressed between 0 and 1. The image is as follows (as shown in Figure 6):

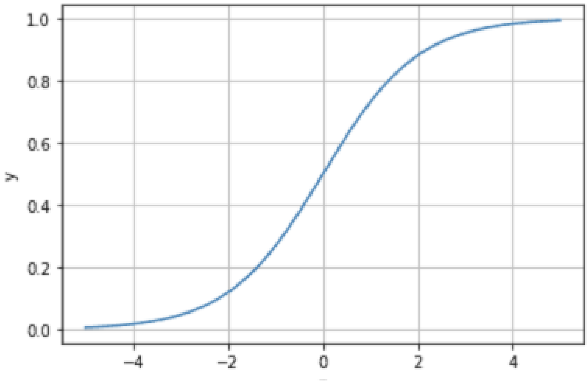


Figure 6. Logistic Regression image

4.2 Model Architecture: Random Forest

An method called random forest uses the concept of ensemble learning to incorporate several types of trees. Its fundamental building block is a decision tree, and its core is an ensemble learning (Ensemble Learning) technique, one of the main branches of machine learning [5]. From a logical standpoint, each tree is a classifier (assuming the issue is one of classification), hence for a given input sample, N trees will produce N classification outcomes. The basic Bagging concept uses the random forest, which aggregates all categorization voting results and assigns the category with the most votes as the final output.

The generation process for the random forest is as follows [6]:

1)Use the Bootstraping method to randomly select m samples from the original training set, and perform n_tree sampling in total to generate n_tree training sets.

2)For n_tree training sets, train n_tree decision tree models respectively;

3)For a single decision tree model, assuming that the number of training sample features is n, then each split will select the best feature for splitting according to information gain/information gain ratio/Gini index;

4)Each tree keeps splitting like this until all training examples for that node belong to the same class. No pruning is required during the splitting of the decision tree;

5)The generated multiple decision trees are formed into a random forest. For classification problems, the final classification result is determined by voting by multiple tree classifiers; for regression problems, the final prediction result is determined by the mean of multiple tree predictions.

4.3 Model Training

In this part, this article directly uses Random Forest and Logistic Regression in sklearn to train three models with the training set processed above. Here only Logistic Regression is used to predict and correct the target variable, and finally the accuracy is used to evaluate and the result is about 0.81.

4.4 Use Grid Search to Find Optimal Hyperparameters

According to the end of the above, it can be seen that the direct fitting model and scoring results are not very ideal, so we will use Grid Search Cross Validation to adjust and search for Optimal Hyperparameters [7]. For the Logistic Regression, the first thing to consider is whether the penalty is L1 or L2 and C, which is the 1/lambda value(weight) for L1 and L2, In this paper, 0.01, 0.05, 0.1, 0.2 and 1 are selected. The operation results show that the best result selection parameters of this model are L1 and 1. The details are shown in Figure 7.

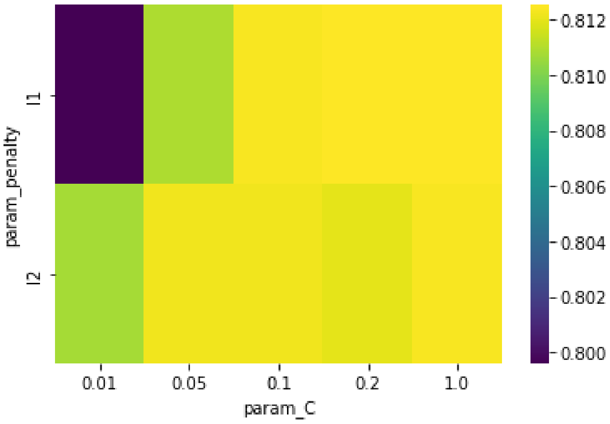
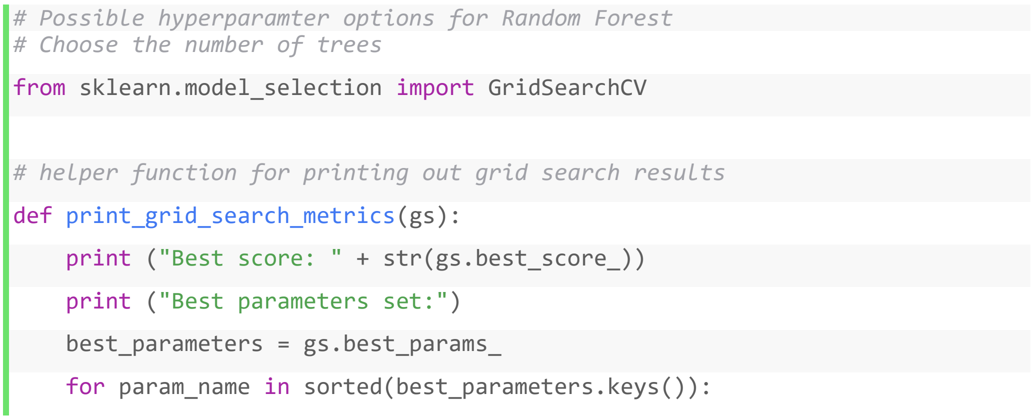


Figure 7. The effect of two parameters on Logistic Regression

For Random Forest, ‘n_estimators’ means that this is the number of trees in the forest. The effect of this parameter on the accuracy of the random forest model is monotonous, and the larger the ‘n_estimators’, the better the model’s performance. However, accordingly, any model has a decision boundary. When the ‘n_estimators’ reach a certain level, the accuracy of random forest usually does not rise or start to fluctuate. In addition, the larger the ‘n_estimators’ are, the more computation and memory are required, and the longer the training time will be.

In this paper, GridSearch's 'n_estimators' are chosen to be 60, 80 and 100, and 'max_depth' is chosen to be 1, 5 and 10, respectively. Part of the Python code has been written below



```
# Possible hyperparamter options for Random Forest
# Choose the number of trees
from sklearn.model_selection import GridSearchCV


# helper function for printing out grid search results
def print_grid_search_metrics(gs):
    print ("Best score: " + str(gs.best_score_))
    print ("Best parameters set:")
    best_parameters = gs.best_params_
    for param_name in sorted(best_parameters.keys()):
        print(param_name + ':' + str(best_parameters[param_name]))
parameters = {
'n_estimators' : [60,80,100],
'max_depth': [1,5,10]
}
Grid_RF = GridSearchCV(RandomForestClassifier(),parameters, cv=5)
Grid_RF.fit(X_train, y_train)
print_grid_search_metrics(Grid_RF)
```

4.5 Model Evaluation

First, for a binary problem, cases can be divided into positive or negative classes. If the prediction is made, there will be four cases, namely: the instance is a positive class and is also predicted to be a positive class. That is, a true class TP (True positive), an instance is a negative class and is predicted to be a positive class, which is called a false positive class FP (False positive class positive). The instance is a negative class is predicted to be a negative class, called true negative class TN (True negative). An instance is a positive class is predicted to be a negative class, called false negative class FN (false negative) [8].

Therefore, the ordinate and abscissa of the ROC curve are essentially unimportant, and the ROC curve cannot be considered as a function curve. Consider an ROC curve with an infinite number of points, each point representing a classifier whose performance is measured by vertical and horizontal coordinates. The performance of this classifier is described by coordinates. AUC, is the area under the Roc curve and has a value between 0.1 and 1. The better the value, the more intuitively the auc can evaluate the quality of the classifier as a value.

After the previous training and adjustment process, the paper will find the model with the best performance, then use this model to predict the possibility of customer churn, and then find out the ROC threshold. Since we only care about the probability of customer churn, we extracted a line of customer churn probability to draw the ROC and calculated the AUC score to evaluate the model. Figures 8 and 9 show the specific results of logistic regression and random forest.

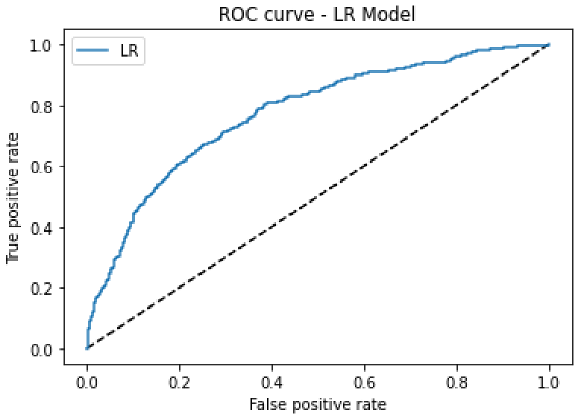


Figure 8. ROC & AUC of LR Model

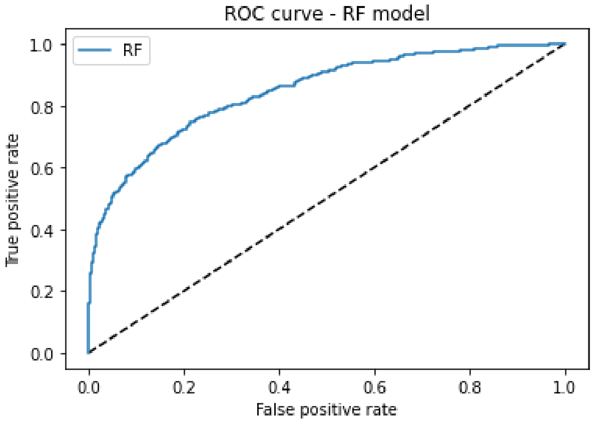


Figure 9. ROC & AUC of RF Model

Look carefully at the ROC curves above, they are all rising curves (slope greater than 0), passing through points (0,0) and (1, 1). In fact, these points describe the classification impact of the classifier at various thresholds. Especially the curve from left to right can be considered as the process of threshold changing from 0 to 1. When the threshold of the classifier is set to 0, all judgments are zero and no identification is required. At this time TP=FP=0, TPR=TP/P=0, FPR=FR/N=0. When the threshold of the classifier is 1, all judgments are regarded as 1, and no identification is required. At this time FN=TN=0, P=TP+FN=TP, TPR=TP/P=1, N=FP+TN=FP, at this time FPR=FR/N=1. Since classifier performance changes as the classifier threshold changes, the ROC curve actually describes this process. By comparing Figures 10 and 11, we find that the performance of the RF Model is better than that of the LR Model. The most intuitive thing is that the AUC of RF Model is greater than the AUC of LR Model.

5. Feature importance discussion

Nowadays there are lots of data set,but a single data which are included in that often has hundreds or thousands of features. To reduce the number of features in modeling and to find the factors that have the greatest impact on the customer churn of the bank as discussed in this part, how can we select the main features that have the greatest impact on the results? Here we mainly introduce Random Forest to screen the features [9]. To put it simply, Random Forest is used to evaluate the importance of features by observing how much each feature contributes to each tree, taking the average value of its contribution to all trees, and finally comparing the contribution of the features with each other. This contribution is measured by Gini index as an evaluation index in this paper.

We denote variable importance measures in $\boldsymbol{VIM}$**.** The Gini index is expressed as $\boldsymbol{GI}$, assuming $\boldsymbol{J}$ features, $\boldsymbol{X_1, X_2, X_3 ... X_J, I}$ decision tree, $\boldsymbol{C}$ categories, and the Gini index score of each feature $\boldsymbol{X_J}$ is $\boldsymbol{VIM^{(Gini)}}$**,** which means the average change of the Node splitting impurity which caused by J feature in all decision trees in Random Forest. Therefore, the calculation formula of Gini index of the I tree node q is [10]:

$$\boldsymbol{GI_q^{(i)} = \sum_{C=1}^{|C|}\sum_{C'\neq C} P_{qc}^{(i)}\, P_{qc'}^{(i)} = 1 - \sum_{C=1}^{|C|}(P_{qc}^{(i)})^2} \qquad \textbf{(3)}$$

In this formula, C represents that there are C categories, and $\boldsymbol{P_{qc}}$ represents the proportion of category C in node q.

Suppose that the set of nodes that feature $\boldsymbol{X_J}$ appears in the decision tree i is Q, then the importance of $\boldsymbol{X_J}$ on the i tree can be expressed as

$$\boldsymbol{VIM_j^{(Gini)(i)} = \sum_{q\in Q} VIM_{jq}^{(Gini)(i)}} \qquad \textbf{(4)}$$

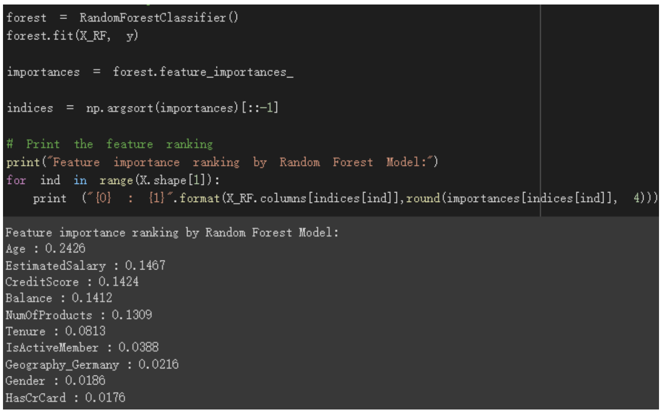
If the RandomForest contains a total of I trees, then there are

$$\boldsymbol{VIM_j^{(Gini)(i)} = \sum_{i=1}^{i} VIM_{jq}^{(Gini)(i)}} \qquad \textbf{(5)}$$

Finally, after the normalization treatment can be obtained

$$\boldsymbol{VIM_j^{(Gini)} = \frac{VIM_j^{(Gini)}}{\sum_{j'=1}^{j} VIM_{j'}^{(Gini)}}} \qquad \textbf{(6)}$$

Thankfully, Python’s sklearn does everything for us, just by calling its functions. The results show that the three most important factors affecting the loss of bank customers are Age, Estimate Salary and Credit Score. The result is shown in Figure 10.



```
forest  =  RandomForestClassifier()
forest.fit(X_RF,  y)

importances  =  forest.feature_importances_

indices  =  np.argsort(importances)[::-1]

#  Print  the  feature  ranking
print("Feature  importance  ranking  by  Random  Forest  Model:")
for  ind  in  range(X.shape[1]):
    print  ("{0}  :  {1}".format(X_RF.columns[indices[ind]],round(importances[indices[ind]],  4)))
```

```
Feature importance ranking by Random Forest Model:
Age : 0.2426
EstimatedSalary : 0.1467
CreditScore : 0.1424
Balance : 0.1412
NumOfProducts : 0.1309
Tenure : 0.0813
IsActiveMember : 0.0388
Geography_Germany : 0.0216
Gender : 0.0186
HasCrCard : 0.0176
```

Figure 10. Feature importance in Random Forest

6. Conclusion

All of the above points indicate that, when compared to Logistic Regression, the Random Forest model’s performance effect is superior. However, based on the feature importance discussion, we can conclude that the following three recommendations are required:

(1) strategy for elderly care We can conclude from the previous article that age is the biggest factor affecting the model of whether bank customers lose or not. Therefore, to retain older customers, it is very important for banks to adopt relevant care strategies. Generally speaking, what the elderly lack is company. Therefore, banks can organize staff to visit customers regularly to achieve this goal.

(2) Precise positioning of target customers. Again, the Estimate Salary is the second most influential factor in the model's effectiveness. Accurate target customer positioning aids banks in defining targeted strategies, offering products and services that are appropriate for their target consumers, and enhancing their competitiveness.

(3) High credit rating reward system. Credit Score is the third most influential factor in the model’s effectiveness. However, in real life, credit scores are often used to assess whether a customer is qualified to apply for a credit card. However, according to the above analysis, users with credit cards have a lower turnover rate than those without credit cards. Therefore, banks can consider granting credit cards to users with high credit scores but without credit cards.

Due to the limited level of the author, there are still many problems in this paper. For example, the details are not handled very carefully. Future research will solve this problem slowly and continue to study the subject of machine learning.

Acknowledgment

In this part, I would like to thank the teacher Luke who led me to do this project. Secondly, in this research, I also found that I still have deep shortcomings and need to study more steadily in the future. Finally, I would also like to appreciate the support by my parents, without their help, I could not finish this thesis.

References

[1] qq_44451414, Bank customer churn analysis, Retrieved October 21th. https://blog.csdn.net/qq_44451414/article/details/90241100

[2] Dataset. https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/data

[3] Qiao Yingying. A Survey of Machine Learning Related Algorithms Based on Numerical Prediction. Journal of Anyang Institute of Technology, 016(004), (2017). 71-74.

[4] GentleCP, [Machine Learning] Principle and Practice of Logistic Regression, Retrieved October 22th. https://blog.csdn.net/GentleCP/article/details/108035696

[5] Xiao Maili. Basic concepts of Decision Tree and Random Forest, Retrieved October 21th. https://blog.csdn.net/u010986753/article/details/100144214

[6] sevieryang, Decision tree algorithm notes finishing (1), Retrieved November 15th. https://blog.csdn.net/qq_42442369/article/details/86625591

[7] Raschka S. Python Machine Learning [M]. Packt Publishing. (2015)

[8] Li Kaiwen, Introduction to ROC Curve, Retrieved October 22th. https://zhuanlan.zhihu.com/p/26293316

[9] zjuPeco, Evaluation of Feature Importance Using Random Forest, Retrieved November 17th. https://blog.csdn.net/zjuPeco/article/details/77371645

[10] Yang Kai, Hou Yan, Li Kang. Random Forest Variable Importance Score and Its Research Progress. (2015)