**International Journal of Current Advanced Research**
ISSN: O: 2319-6475, ISSN: P: 2319-6505, Impact Factor: 6.614
Available Online at www.journalijcar.org
Volume 7; Issue 6(E); June 2018; Page No. 13446-13449
DOI: http://dx.doi.org/10.24327/ijcar.2018.13449.2400

**Research Article**

# RESEARCH ON BANK CUSTOMER SATISFACTION CLASSIFICATION ON RANDOM FOREST ALGORITHM

**Li Ying***

School of Business Administration, China University of Petroleum-Beijing, Beijing, 102249, China

**A R T I C L E I N F O**

***Article History:***

Received 11th March, 2018
Received in revised form 6th
April, 2018 Accepted 26th May, 2018
Published online 28th June, 2018

***Key words:***

Customer satisfactionClassification, Random Forest, GridSearchCV, Pearson correlation coefficient,PCA

**A B S T R A C T**

As the customer-orientated business philosophy continues to deepen the impact on our banks, customer satisfaction has more and more important implications for bank's profitability and development. How to efficiently evaluate bank customer satisfactionhas become an urgent issue to be solved. It is of great significance to establish an accurate customer satisfaction prediction model for commercial banks. In this paper, we use Random Forest algorithm based on Grid Search CV to establish a two-classification prediction model for bank customer satisfaction and compare Random Forest algorithm with SVM as well as Logistic Regression algorithms in terms of prediction accuracy for bank customer classification. In order to providesome effective reference and basis for banks to select targeted customers and improve service levels to attain more profit.



## INTRODUCTION

Since the 21st century, competition in China's financial market has become increasingly fierce Strong, and customers have become an important strategic resource for banks,which lead to competition for increasingly scarce customer resources becoming the development of China's commercial bankskey already(Wang, 2011). In this situation, commercial banks implement customer satisfaction classification and recognitionstrategy to select and provide service for target customers isthe general trend.

Random forest algorithm is a new and efficient combinational classification method. It has high operating efficiency when dealing with large data sets, and has strong ability to process high-dimensional data. Besides, it is also not easy to overfit.Because of its superior performance, Random Forest algorithm is widely used in many research fields abroad (Ma, 2016). Larivière B *et al* used random forests and regression forests techniques to predict customer retention and profitability (Larivière B *et al*, 2005). Xiao J *et al* put forward a dynamic classifier ensemble model for customer classification with imbalanced class distribution (Xiao J*et al*, 2012). And Xin-Hai L I researched on using Random Forest for classification and regressio(Xin-Hai L I, 2013). In contrast, the domestic scholars have less research and application.

Therefore, this paper use Random Forest algorithm based on GridSearchCV to establish a two-classification prediction model for bank customer satisfaction in python language environment, and compare Random Forest algorithm with SVM as well as Logistic Regression algorithms in terms of prediction accuracy for bank customer classification. In order to provide some effective reference and basis for banks to select targeted customers as well as improve service levels to attain more profit and develop better.

## MODELAND METHOD

According to the characteristics of a large number of bank customer actual data sets, a large number of features, and many anonymous data, the overall technical framework for modeling this paper is shown in the figure.Since the bank customer's actual datasets have the characteristics of large sample size and large number of features, this paper uses Random Forest algorithm based on GridSearchCV method to build a bank-customer-satisfaction classification model. The overall technical framework for modeling of this paper is shown in the Fig1.

**Corresponding author:* **Li Ying**
School of Business Administration, China University of Petroleum-Beijing, Beijing, 102249, China

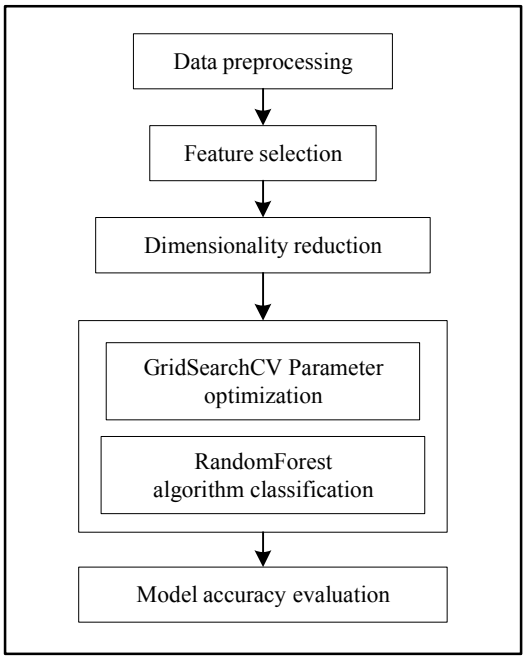


**Fig 1** Overall Technical Framework

### *Data collection and preprocessing*

In this paper, we use the realcustomer-dataset from Santander bankas target data set for the study(https://www.kaggle.com/datasets), which contains 76020 samples and 370 anonymized features.First, split the entire Santander-customer-dataset into two parts, one is the training set for training classification models, the other is testing set for testing the classification effect and accuracy. Then view the entire dataset to find missing valuesand fill in the missing values by mean interpolation method from scikit-learn processing block. Because the features are anonymized , we don't need to recognize object features and perform One Hot encoding processing, while we regard these anonymized features as numerical type features for the next study. And use SMOTE (Han H, 2005) method to solve the imbalanced problem. Finally, in order to improve model performance, normalize the training set, and then the feature values of training set range [-1,1].

### *Feature selection and extraction*

Due to the anonymized feature and high dimension of the dataset, this paper uses Pearson correlation coefficient method to filter features. The Pearson correlation coefficient is one of the simplest methods that can help understand the relationship between features and response variables. This method measures the linear correlation between variables and the results are in the range [-1, 1]. The value -1 indicates a complete negative correlation, +1 indicates a complete positive correlation, and 0 indicates no linear correlation.We use Pearson correlation coefficient combining data visualization technology to filter some unrelated features and reserved features with higher correlations and lower multicollinearity between features, and the Pearson coefficient plot is shown in Fig 2.

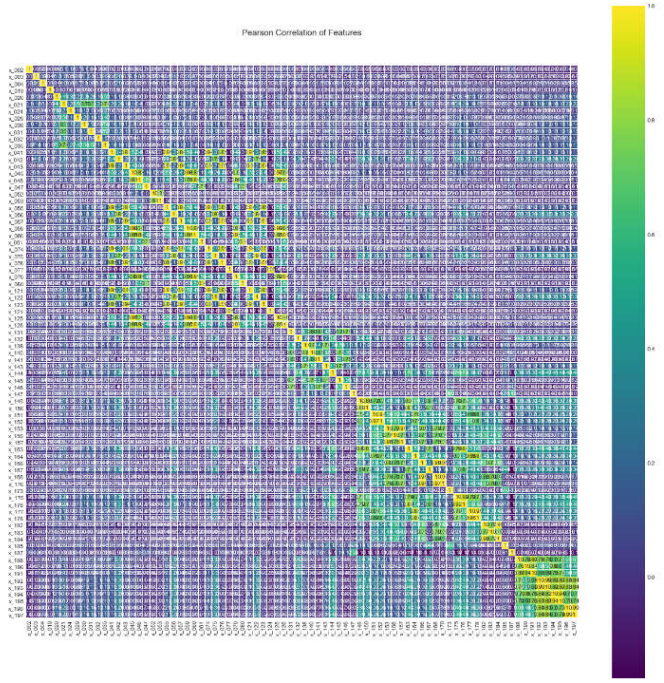


**Fig 2** Overall Technical Framework

Calculate the importance value of retained features on customer classification, and select the top 50 ranking features for visualization, it has been shown in Fig3.

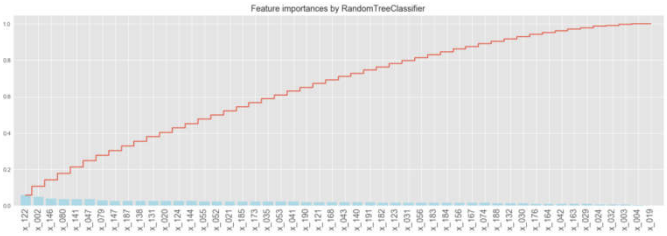


**Fig 3** Overall Technical Framework

In the machine learning, curse of dimensionality usually refers to the problem of calculations involving vectors, as the number of dimensions increases, the amount of computation exponentially increases (Keogh E *et al*, 2011). In order to avoid curse of dimensionality, this paperselect PCA (Principal Component Analysis) method to reduce dimension of dataset (Wold S *et al*, 1987). We use the PCA algorithm in decomposition of sklearn block, where the parameter used to adjust the target dimension is n_components, and set n_components parameter to 50 based on the characteristic variance contribution.Then we get the suitable customer-data for further research.

### *Establishment and Evaluation of classification model*

Random forest algorithm is a new type of machine learning algorithm composed of multiple CART decision tree combinations (Breiman L, 2001). The general process of this algorithm is as follows. First, perform bagging process (Quinlan J R, 1996). Use the random Bootstrap method to extract N groups training sets from the original data, the size of which is about 2/3 of the original set. Second, construct CART decision trees for N training sets separately. During the growth of each tree, select m features (M≤M) from all M features randomly, and select the best features according to the Gini coefficient to divide internal nodes. Finally, aggregate the prediction results of the N CART decision trees and use voting method to determine the class of the new sample. About 1/3 of

the data from each sample is not extracted, and this part of the out-of-bag data is used for internal error estimation to generate OOB errors.

The Random Forest method corresponds the RandomForestClassifier algorithm in scikit-learn. There are some adjustable parameters which may improve the classification effect and accuracy of model effectively, such as n_estimators, max_features, max_depth, max_leaf_nodes and so on. Thus, this paper combines GridSearchCV method inmodel_selection block of sklearn to optimize the classification model parameters.

Grid Search CV methodcan adjust parameters automatically (Pedregosa F, 2011), especiallysuitable for data sets with a large sample size. As long as inputting the parameters, it can give the most optimized results and parameters. The advantage is that it can guarantee the search solution obtained is the global optimal parameters in the grid and avoid major errors (Liu Ying, 2014).

Besides, this paper also train classification models using Logistic Regression (Hosmer Jr D W, 2013) and SVM (Joachims T, 1998) algorithms respectively and compare themwith the Random Forest model, in order to confirm the superiority of Random Forest algorithm for the bank customer satisfaction classification prediction. After establishing the classification model, use Overall Accuracy to evaluate the classification and effect of the models.

## RESULTS AND DISCUSSION

Based on the preprocessed dataset, this paper uses the above algorithms to establish the bank customer satisfaction classification prediction models. We get the Overall Accuracy score of the Random Forest Classifier model, which is about 90.83%. And we find itsbest estimator of parameters through GridSearchCV, which is shown in Table 1.

**Table 1** Best estimator of RandonForestClassifier

| No. | Parameter name | Best estimator | No. | Parameter name | Best estimator |
|---|---|---|---|---|---|
| 1 | n_estimators | 140 | 5 | min_samples_split | 100 |
| 2 | max_depth | 8 | 6 | n_jobs | 1 |
| 3 | max_features | 'sqrt' | 7 | criterion | 'gini' |
| 4 | min_samples_leaf | 20 | 8 | bootstrap | True |

Similarly, we train the classification precision models using Logistic Regression and SVM algorithms based on the same customer-dataset. The model testing evaluation results is that the Overall Accuracy score of Logistic Regression algorithm is 71.31% and the Overall Accuracy score of SVM is 85.80%. Besides, RandomForestClassifier model runs much faster than other two models. Based on the model evaluation results, it is obvious that Random Forest algorithm has the best classification performance, compared with Logistic Regression and SVM algorithms. It also prove Random Forest algorithm is more suitable for bank customer satisfaction two-classification prediction model, especially when the customer-datasets has a large sample size and high feature dimensions.

## CONCLUSIONS

In this paper, we use Random Forest algorithm based on GridSearchCV to establish a two-classification model for bank customer satisfaction prediction and compare Random Forest algorithm with Logistic Regression as well as SVM algorithms in terms of the overall accuracy for bank customer classification. In order to provide some effective reference and basis for banks to select targeted customers and improve service levels to attain more profit. From the experimental results, we draw the following conclusion: the classification model this paper put forward that using Random Forest algorithm based on GridSearchCV indeed has superior classification performance

And the comparative experimental results show that compared to Logistic Regression and SVM classification algorithms, Random Forest algorithm is more suitable for bank customer satisfaction two-classification prediction model, especially when the customer-datasets has a large sample size and high feature dimensions.

Since the current society strongly advocates protecting customer privacy, the number of suitable public bank customer datasets is very small. This paper only analyzes and studies the customer-dataset of Santander bank, which may lead to the study not be universal. In the future, it is necessary to collect more customer-datasets from different banksfor further improvement.

## References


Wang Wangqing. The Construction of Customer Satisfaction Evaluation System for Commercial Banks [J]. *Journal of Henan Institute of Science and Technology*, 2011(09):6-9.

Ma Yi, Jiang, *et al*. Study on Land Use Classification in Agricultural Land Based on Random Forest Algorithm [J]. Transactions of the Chinese Society of Agricultural Machinery, 2016,47(01):297-303.

Larivière B, Van den Poel D. Predicting customer retention and profitability by using random forests and regression forests techniques [J]. Expert Systems with Applications, 2005, 29(2): 472-484.

Xiao J, Xie L, He C, *et al*. Dynamic classifier ensemble model for customer classification with imbalanced class distribution[J]. *Expert Systems with Applications*, 2012, 39(3): 3668-3675.

Xin-Hai L I. Using “random forest” for classification and regression[J]. *Chinese Journal of Applied Entomology*, 2013, 50(4): 1190-1197.

https://www.kaggle.com/datasets

Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]//International Conference on Intelligent Computing. Springer, Berlin, Heidelberg, 2005: 878-887.

Keogh E, Mueen A. Curse of dimensionality[M]//Encyclopedia of machine learning. Springer US, 2011: 257-258.

Wold S, Esbensen K, Geladi P. Principal component analysis[J]. Chemometrics and intelligent laboratory systems, 1987, 2(1-3): 37-52.

]Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.

Quinlan J R. Bagging, boosting, and C4. 5[C]//AAAI/IAAI, Vol. 1. 1996: 725-730.

Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: Machine learning in Python[J]. *Journal of machine learning research,* 2011, 12(Oct): 2825-2830.

Liu Ying. Research on Remote Sensing Image Classification Based on Machine Learning[M]. North

Beijing: Tsinghua University Press, 2014.
Hosmer Jr D W, Lemeshow S, Sturdivant R X. Applied logistic regression[M]. John Wiley & Sons, 2013.
Joachims T. Making large-scale SVM learning practical[R]. Technical report, SFB 475: Komplexitätsreduktion in MultivariatenDatenstrukturen, Universität Dortmund, 1998.

**How to cite this article:**

Li Ying (2018) 'Research on Bank Customer Satisfaction Classification on Random Forest Algorithm', *International Journal of Current Advanced Research,* 07(6), pp. 13446-13449. DOI: http://dx.doi.org/10.24327/ijcar.2018.13449.2400

*******