Comparison of Random Forest, Logistic Regression, and Multilayer Perceptron Methods on Classification of Bank Customer Account Closure

Husna Afanyn Khoirunissa[1], Amanda Rizky Widyaningrum[2], and Annisa Priliya Ayu Maharani[3]
[1,2,3]Program Studi Statistika, Universitas Sebelas Maret

husnafanyn@gmail.com, amandarizkyw@gmail.com, annisapriliya320@gmail.com

Abstract. The Bank is a business entity that is dealing with money, accepting deposits from customers, providing funds for each withdrawal, billing checks on the customer's orders, giving credit and or embedding the excess deposits until required for repayment. The purpose of this research is to determine the influence of age, gender, country, customer credit score, number of bank products used by the customer, and the activation of the bank members in the decision to choose to continue using the bank account that he has retained or closed the bank account. The data in this research used 10,000 respondents originating from France, Spain, and Germany. The method used is data mining with early stage preprocessing to clean data from outlier and missing value and feature selection to select important attributes. Then perform the classification using three methods, which are Random Forest, Logistic Regression, and Multilayer Perceptron. The results of this research showed that the model with Multilayer Perceptron method with 10 folds Cross Validation is the best model with 85.5373% accuracy.

Keywords: bank customer, random forest, logistic regression, multilayer perceptron

1. Introduction

The population in all countries evolves and increases, that is no exception to countries in continental Europe such as Germany, Spain, and France. An increase in the population, it leads to the need to be met by everyone. Financial management is an important factor in fulfilling the needs of each person. That condition is a potential target for the rapidly growing financial industry. Banking is one of the industries that is significant advances around the world. The bank is a business entity in about to concerning to money, accepting deposits from customers, providing funds for each withdrawal, billing checks on the customer's orders, giving credit and or embedding the excess deposits until it is made for repayment [1]. As time goes by, inter-banking competition can't be avoided. Each banking strives to offer a variety of products and services to attract people to their customers. Customer loyalty to always use a bank without closing the account at the bank is the goal of all banks. The quality of products or services offered is a big role to increase customer loyalty bank. The often-used bank product is credit. The easy to apply credit conditions affecting the customer to select the bank. Customers who believe in a bank will use a lot of products from the bank and will become active member of the bank. In addition, the cultural factor of a country is participating actively in customer loyalty. As well as the age and gender of the customer are related to the using of the product according to their needs.

Based on the background, researchers are interested in knowing what are factors affecting the customers to retained or closed their bank account. So that, researchers use customer bank data to become a research object. The purpose of this research is to know whether the credit score, country of origin of customer, gender, age of customer, tenure, balance, number of bank products used, holding a credit card, active member, and customer salary estimate affects the customer to decide to retained or closed the bank account.

2. Theoretical Basis

2.1. Data Preprocessing. Preprocessing is the earliest stage in data mining. This technique aims to facilitate the interpretation of the analysis of the data used for data mining applications. In addition, so the data used in accordance with the application is built and the results are also suitable and optimal [2]. Stages in data preprocessing include: a) incomplete, lacking attribute values or called missing value, b) noisy, containing errors or outliers, c) inconsistent, containing discrepancies in codes names, d) feature selection, selecting the important attributes [3].

2.2. Random Forest. Random Forest begins with a basic technique of mining data, the decision tree. In the decision tree input is inserted at the top (root) then down to the bottom (leaf) to determine the data including what class. A Random Forest is a classification consisting of a collection of structured tree classifications where each tree throws a sound unit for the most popular class. In other words, the Random Forest consists of a set of decision trees, where the decision tree set is used to classify the data to a class [4]. Random Forest is a bagging method that the method generates a number of trees from the sample data where the creation of one tree during training does not depend on the previous tree then the decision is taken based on the most voting [5]. In the bagging process, bootstrap resampling is used to generate the classification tree, which is a multi-version generation technique that combines them to obtain the final prediction. Whereas in the Random Forests method, the randomization process is carried out on sample data and on taking independent variables so that the classification tree generated will have different sizes and shapes [6].

2.3. Logistic regression. Logistic regression is a regression model used when the response variable is qualitative [7]. Logistic regression is one of the most frequently used classification methods. Binary logistic regression is used when a dependent variable is a dichotomous variable. Logistical regression multinomial used when the dependent variable is a variable of categorical with more than two categories. In general, the logistics regression model is

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}} \tag{1}$$

Where π(x) is the probability value of $0 \le \pi(x) \le 1$, which means that logistic regression illustrates a probability. By transforming (x) the equation (1) with the transformation of Logit $g(x)$, where,

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \tag{2}$$

2.4. Multilayer Perceptron. Multilayer perceptron or artificial neural network (ANN) from many layers is a network composed of multiple layers of perceptrons. Multilayer Perceptron or artificial neural network from many layers is a network composed of multiple layers of perceptrons. The algorithm for the multilayer perceptron is the inputs are pushed forward through the multilayer perceptron by taking the dot product of the input with the weights that exist between the input layer and the hidden layer. Multilayer Perceptron utilizes activation functions at each of their calculated layers. Once the calculated output at the hidden layer has been pushed through the activation function, push it to the next layer in the MLP by taking the dot product with the corresponding weights. Repeat steps two and three until the output layer is reached. At the output layer, the calculations will either be used for a backpropagation algorithm that corresponds to the activation function that was selected for the Multilayer Perceptron (in the case of training) or a decision will be made based on the output (in the case of testing). The advantages of this multilayer perceptron method has multiple units that are in one or more hidden screens compared to the perceptron method consisting of a single screen. After successfully found various applications that can be solved with Multilayer Perceptron make ANN more attracted people in terms of predictions give more detailed information on how the method works. The advantages of this multilayer perceptron method has multiple units that are in one or more hidden screens compared to the perceptron method consisting of a single screen. After successfully found various applications that can be solved with Multilayer Perceptron make ANN more attracted people in terms of predictions [8].

2.5. Confusion matrix. The confusion matrix is a table for measuring the performance of classification algorithms or classification models [9]. In binary classification, a 2x2 matrix is used with members such as: True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN). Of these values, it can be calculated several values that can be used as a performance value classifier, among others:

$$Accuracy = \frac{(TP + TN)}{n}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

$$\mathrm{Pr}ecision = \frac{TP}{FP + TP}$$

3. Results and Discussion

This section elaborated on the preprocessing of data and comparative classification of random forest, logistic regression, and multilayer perceptron methods.

3.1. Methods. The dataset used in this research is data related to the direct marketing campaign of Portuguese banking that published Sonu Jha on Kaggle [10]. There are 10 attributes that are used as predictor variables whether customers closed the bank account or retained them. The ten attributes are credit score, country of origin of the customer, gender, age of customer, tenure, balance, number of bank products used, whether customer holding a credit card or not, whether the customer is an active member with bank and customer salary estimate in dollars. To process the data, the first thing to do is the preprocessing data were at this step, the missing values and outliers (multivariate and univariate) will be checked and then do the treatment if missing values and outliers are found. The next step is to select attributes that are important in predicting and modelling to classify response variables using three methods, namely random forest, logistic regression, and multilayer perceptron. The last stage is to compare the results of the three classification methods and get the best model.

3.2. Data preprocessing. In the identification of missing value in the data, no missing value is found so that the analysis can be continued on outlier detection. Outlier detection is performed using standard deviation data following the normal and symmetrical distribution. Found univariate outlier on credit score data, age of the customer, and the number of bank products used, as much as 8, 133, and 60 instances respectively. Before the treatment on the outlier, see a statistical summary of all three data attributes.

ISSN 2621–086X

Table 1. Statistical summary of three data attributes that have outliers

Attributes	Mean	SD.	Min	Q1	Median	Q3	Max
CreditScore	650.53	96.65	350	584	652	718	850
Age	38.92	10.48	18	32	37	44	92
NumOfProducts	1.53	0.58	1	1	1	2	4

Looking at the summary of the statistics in Table 1, the minimum and maximum values on each attribute are still fairly reasonable so that special treatment is not performed to preserve the typical characteristics of the data. Furthermore, the data is performed multivariate outliers detection on a number of numeric data attributes, namely credit score, age of customer, tenure, balance, number of bank products used, and customer salary estimates in dollars. The identification of multivariate outliers uses a Mahalanobis distance $(d^2{}_i)$ which is the distribution of chi-square with a degree of freedom (DF) of included attributes and level of significance. The critical area is $d^2{}_i$ > $X^2{}_{(0.05;6)}$ = 12.59. There are 396 instances of multivariate outliers. To fix this, it will be trimming instances that include a multivariate outlier as the best and faster approach. So that, the number of instances becomes 9604 observations. To know the important attributes in predicting, the feature selection is performed with the CfsSubsetEval method in Weka. CfsSubsetEval method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [11]. It obtained 6 important attributes for classification, namely credit score, country of customer origin, gender, age of the customer, number of bank products used, and whether the customer is an active member of the bank.

3.3. Comparison of classification methods. Once obtained predictor variables to be used, next created model to classify response variable that is bank's customer account closure. From such data created models with the methods of Random Forest, logistic regression, and Multilayer Perceptron. The three models are run with 10 folds Cross Validation (CV) to make predictions with different folds training. CV could be a resampling procedure used to evaluate each model of machine learning on a restricted information sample. The procedure includes a single parameter referred to as *k* that refers to the number of groups that a given data sample is to be split into. In this research, *k*=10 was used for each machine learning model. Previously, data were divided into training data and testing data with a comparison of 0.8 and 0.2 respectively. Accuracy comparison results are presented in Table 2.

Table 2. Comparison of Training Data and Testing Data Prediction Results

Methods	Accuracy (training data)	Accuracy (testing data)
Random Forest	0.832487	0.790213
Logistic Regression	0.838475	0.8152
Multilayer Perceptron	0.855135	0.820927

In comparison to Table 2, the best model for classifying a bank's customer account closure is the Multilayer Perceptron method which has prediction accuracy of 85.5135% in training data and 82.0927% in testing data.

3.4. Prediction. The best model obtained is multilayer perceptron, then applied to the overall data so that the confusion matrix presented in Table 3.

Table 3. Confusion Matrix

Class	Retained Account	Closed Account
Retained account	742	1084
Closed account	305	7473

From the confusion matrix in Table 3, obtained the accuracy of the data at 85.5373%. Values on the confusion matrix can be counted as some other values of each class which can provide more detailed information about the performance of the classifier used.

The precision value of the retained account class by 0.709 and in the closed account class of 0.873 gives the correct proportion of the label's information on the predicted total instances. The recall rate of the Retained account class by 0.406 and in the closed account class of 0.961 provided the proportion of the label that was predicted correctly in that class. The difference in the selection of performance measures depends on the purpose of the bank in performing classification. If the bank is better off requiring customers with the closed account to be predicted to be retained account than customers with retained account predicted to be closed account, then the recall performance is used. However, if the bank requires the opposite, then the precision performance is used.

Table 4. Performance classification details

	TP Rate	FP Rate	Precision	Recall	Class
	0.406	0.039	0.709	0.406	Retained account
	0.961	0.594	0.873	0.961	Closed account
Weighted Avg.	0.855	0.488	0.842	0.855	

4. Conclusion

There are 10 attributes that are used as predictor variables. However, based on the feature selection, only 6 predictor variables are selected to be used to classify the bank's customer account closure, there are credit score, customer's country of origin, gender, age of costomer, number of bank products used, and whether the customer is an active member with the bank. The best method for classifying the bank's customer account closure is the Multilayer Perceptron method with a data prediction accuracy result of 85.5373%.

References

[1] Guru Pendidikan, *https://www.gurupendidikan.co.id/pengertian-bank-menurut-para-ahli/*, accessed on 30 April 2020.

[2] Abdillah, G., Putra, F.A., and Renaldi, F. Penerapan Data Mining Pemakaian Air Pelanggan untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru di PDAM Tirta Raharja Menggunakan Algoritma K-Means. *Seminar Nasional Teknologi Informasi dan Komunikasi.* 498–506. 2016.

[3] Hacker Noon, *https://hackernoon.com/what-steps-should-one-take-while-doing-data-preprocessing--502c993e1caa* , accessed on 1 May 2020.

[4] Han, J. *Data Mining Concepts and Techniques Third Edition*. The USA. Elsevier. 2012.

[5] Wibowo, A.T., Saikhu, A., and Soelaiman, R. *Implementasi Algoritma Deteksi SPAM yang Tersisipi Informasi Citra dengan Metode SVM dan Random Forest*. Institut Teknologi Sepuluh Nopember. Surabaya. 2016.

[6] Jatmiko, Y.A., Padmadisastra, S., Chadidjah, A. Analisis Perbandingan Kinerja Cart Konvensional, Bagging dan Random Forest pada Klasifikasi Objek: Hasil dari Dua Simulasi. *Media Statistika*. 1-12. 2019.

[7] Hosmer, D.W. and Lemeshow, S. *Applied Logistic Regression*. John Wiley and Sons Inc. Canada. 1989.

[8] Siang, J.J. *Jaringan syaraf tiruan dan pemrogramannya menggunakan Matlab*. Penerbit Andi. Yogyakarta. Vol. 11. 2005.

[9] Faisal, M.R. and Nugrahadi, D.T. *Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R.* Scripta Cendekia. Banjarbaru. 2019.

[10] Jha, S, *https://www.kaggle.com/sonujha090/bank-marketing* , accessed on 26 April 2020.

[11] Hall, M.A. *Correlation-based Feature Selection for Machine Learning*. New Zealand. 1999.