DOI: 10.15849/IJASCA.211128.09

Int. J. Advance Soft Compu. Appl, Vol. 13, No. 3, November 2021
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC)

Islahulhaq[1], Wahyu Wibowo[2], and Iis Dewi Ratih[3]

[1,2,3] Department of Business Statistics, Faculty of Vocational, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: islahulhaq690@gmail.com, wahyu_w@statistika.its.ac.id, iis.dewi@statistika.its.ac.id

Abstract

Financing analysis is the process of analyzing the ability of bank customers to pay installments to minimize the risk of a customer not paying installments, which is also called Non-Performing Financing (NPF). In 2020 the NPF ratio at one of the Islamic banks in Indonesia increased due to the decline in people's income during the Covid-19 pandemic. This phenomenon has led to bad banking performance. In December 2020 the percentage of NPF was 17%. The imbalance between the number of good-financing and NPF customers has resulted in poor classification accuracy results. Therefore, this study classifies NPF customers using the Logistic Regression and Synthetic Minority Over-sampling Technique Nominal Continuous (SMOTE-NC) method. The results of this study indicate that the logistic regression with SMOTE-NC model is the best model for the classification of NPF customers compared to the logistic regression method without SMOTE-NC. The variables that have a significant effect are financing period, type of use, type of collateral, and occupation. The logistic regression with SMOTE-NC can handle the imbalanced dataset and increase the specificity when using logistic regression without SMOTE-NC from 0.04 to 0.21, with an accuracy of 0.81, sensitivity of 0.94, and precision of 0.86.

Keywords: *Classification, Islamic Bank, Logistic Regression, Non-Performing Financing, SMOTE-NC.*

1 Introduction

Financing analysis is the process of analyzing the ability of bank customers to pay installments that requires a customer to qualify financing standards before the disbursement of funds. The results of the financing analysis as an instrument to determine whether the customer deserves to be approved or not for financing. Financing analysis to minimize the risk of customers not paying installments or referred to as Non-Performing Financing (NPF). One of the current challenges is that the emergence of the Covid-19 pandemic in 2020 has reduced people's income in Indonesia. The decline in people's income risks an increase in the NPF ratio. This phenomenon causes unhealthy banking performance because the amount of money is stuck in the customers, one of which is an Islamic bank in Indonesia. This is shown in December (fourth quarter) of 2019 that the NPF ratio of 3.37% has increased to 7.66% in 2020 [1]. In December 2020, the NPF ratio was above the mean NPF ratio of Islamic Banks in Indonesia (7.24%) [2]. The higher NPF ratio value (>5%) makes the bank quality unhealthy [3].

In the process of financing analysis, a survey of customers is required. Adjustment of survey activities during the Covid-19 pandemic is necessary by reducing the intensity of direct meetings to reduce the spread of Covid-19. The current financing analysis also takes longer because it checks the field conditions directly, so it is necessary to make time-efficient and minimize the risk of NPF for the bank's financial health. One solution is to classify NPF customers based on their background and personal data information when submitting a financing form.

This study will analyze the factors that affect NPF and create a classification model for NPF customers using logistic regression and Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC). This study will compare the results of the logistic regression with SMOTE-NC and logistic regression without SMOTE-NC. The classification method with the highest accuracy can provide convenience in classifying the ability of customers to pay installments in the future and become a screening media for the initial process before being followed up by the bank.

Therefore, the structure of this study is as follows: Section 2 describes the factors that led to NPF in previous studies in different areas and some of the studies used similar methods. The logistic regression and SMOTE-NC methods will be discussed in Section 3. SMOTE-NC as a method of handling imbalanced data will be implemented in Section 4 before using logistic regression. Characteristic of financing customers and the comparison result between the classification of NPF using logistic regression with SMOTE-NC and logistic regression without SMOTE-NC will be discussed in Section 5. Section 6 concludes the factors that have a significant effect on the NPF and the best model selected in this study.

2 Related Work

A study by Fianto et al investigated that age, gender, occupation, and type of contract affect the NPF of Islamic microfinance institutions (MFIs) in Indonesia using the logistic regression method [4]. A study by Agbemava et al used data from microfinance institutions in Accra Ghana and analyzed using binomial logistic regression shows that six factors have a significant effect on the quality of financing: marital status, dependents, type of collateral, financing period, and type of financing with an accuracy of 86.67% [5]. Another study by Obare et al analyzed the NPF customers in Kenya using the logistic Regression method shows that the history of financing, the purpose of financing, the amount of financing, the nature of the deposit account, occupation, gender, age, type of collateral and place of residence has an accuracy of 73.33% [6].

Another study by Wibowo et al compares several methods of Weight of Evidence (WoE), Information Value (IV), logistic regression with imbalanced data, and logistic regression with SMOTE to overcome the problem of imbalanced data between the number of good credit status and bad credit status in the Islamic banks in Indonesia. The credit status model with imbalanced data has higher accuracy and sensitivity than the logistic regression model using the SMOTE method. However, the specificity of the customer credit status model using imbalanced data is lower than the model using the SMOTE method [7].

A study by Santoso et al integrating SMOTE in data mining classification methods such as Naive Bayes, Support Vector Machine (SVM), and Random Forest (RF) is expected to improve accuracy performance. The result of the study was found that SMOTE data provides better accuracy than the original data. In addition to the three classification methods used, RF provides the highest mean AUC, F-measures, and G-means values [8]. Barro et al suggest that imbalanced data on the herbal medicine composition model have an AUC is equal to 0.976 for the model with SMOTE and 0.908 for the model without SMOTE. The result shows that SMOTE increases the accuracy [9]. Yu Zhang et al analyzed the NPF problem at a bank in China to identify and classify the type of financing (NPF or not) using the Decision Tree, Naive Bayes, and support vector machine (SVM) methods. The independent variables in this study are: type of financing, the amount of financing, period, maturity period, type of interest rate due, accumulated interest, and 24 other variables. The result shows that the Decision Tree method using the C4.5 algorithm could identify NPF with an accuracy of 94% [10]. Johan found that collateral and capital had a significant effect on credit quality. The collateral was represented by the title of the car ownership (title of ownership). The capital was represented by the financing tenor, at significance level = 0.05. [11].

Based on the previous study, there are similarities in the variables that have a significant effect on the NPF. These factors are: age, gender, occupation, type of financing, financing period, the amount of financing, type of collateral. Therefore, the authors use several variables that were used in previous studies to be

implemented in an Islamic Bank in a different location. The authors had an internship at a bank and noticed that the occupation, especially an entrepreneur, has a high chance of NPF due to the unstable economy during the pandemic.

3 Methodology

3.1 Dataset

This study used secondary data at one of the Islamic banks in Indonesia. This study used 2,228 customers' data who have made a transaction from December 1, 2020 – to December 31, 2020. The data is divided into 70% (1560) as training data and 30% (668) as testing data. The restriction of this study is the individual customers and does not include companies. The dependent variable is financing quality which consists of 2 categories good financing and NPF. Based on the regulations of the Financial Services Authority of Indonesia, the financing quality is determined based on 3 aspects: business prospects, customer performance, and customer's ability to pay. Each financing contract (profit sharing, buying, and selling & borrowing, leasing) classifies the quality of financing into 5 groups (1: good performance, 2: in special concern, 3: underperformance, 4: under monitoring (close to default), 5: defaulter) [12]. Good financing when the quality is good performance and NPF when the quality of financing is special concern, underperformance, under monitoring, and defaulter. Independent variables are the background and information on the personal data of individual customers. The dependent and independent variables are shown in Table 1.

Table 1: Description of dependent and independent variables

Variable	Description	Category
Y	Financing Quality	1: Good Financing 2: Non-Performing Financing
X_1	The Amount of Financing (Rupiah)	-
X_2	Financing Period (Months)	-
X_3	Type of Use	1= Working Capital 2= Investment 3= Consumption
X_4	Type of Collateral	1= Car 2= Motorcycle 3= Certificate Property 4= Letter for Deduction of Salary 5= Others
X_5	Age	-
X_6	Occupation	1= Entrepreneur 2= Carpentry & Craftsmen 3= Informal workers 4= Lecture and Teacher 5= Student 6= Laborers and Farmers 7= Others
X_7	Gender	1= Male 2= Female

3.2 Logistic Regression

Logistic Regression is a method to find the relationship between the dependent variable (y) which is dichotomous (nominal or ordinal scale with 2 categories) or polychotomous (nominal or ordinal scale with more than 2 categories) with one or more independent variables (x) which are continuous or categorical [13]. Binary logistic regression is used to explain the relationship between the dependent variable (binary data) with the independent variable (interval and categorical) data. The dependent variable (Y) in the binary logistic regression is a variable that results y = 1 (success) and y = 0 (failure). Parameter estimation in logistic regression using Maximum Likelihood method where the estimated parameter β requires to follow a specific distribution. Dependent variable on binary logistic regression following Bernoulli's distribution [14].

3.3 SMOTE-NC

Synthetic Minority Over-sampling Technique (SMOTE) is one of the derivatives of oversampling which was first introduced by Chawla et al in 2002. SMOTE can handle imbalanced data by replicating minority data, the result is known as synthetic data. In numerical data, SMOTE will work by finding k-nearest neighbors for each data in the minority class using Euclidean distance. For categorical and numerical data, the Synthetic Minority Over-sampling Technique-Nominal Continuous (SMOTE-NC) will be used [15]. According to Li & Sun in Mustaqim et al's study, if the proportion of minority class is less than 35% of the total data, then it is categorized as an imbalance dataset [16].
Suppose there are as many data as q variables as follows:

$$\mathbf{X}^T = \left[X_1, X_2, ..., X_q \right] \text{ and } \mathbf{Z}^T = \left[Z_1, Z_2, ..., Z_q \right] \quad (1)$$

Then the Euclidian Distance d**(X, Z)** is:

$$d(\mathbf{X}, \mathbf{Z}) = \sqrt{(X_1 - Z_1)^2 + (X_2 - Z_2)^2 + ... + (X_q - Z_q)^2} \quad (2)$$

Synthetic data for numerical data will be generated using the following equation:

$$x_{syn} = x_m + (x_{knn} - x_m)\gamma \quad (3)$$

x_{syn} = Synthetic data

x_m = Data from minority class

x_{knn} = Data from the minority class that has the nearest neighbors

γ = Random number between 0 and 1

Synthetic data for numerical and categorical data will be generated using the SMOTE-NC algorithm, which is described as follows:
1. Median Calculation: Calculates the median of the standard deviation of all continuous variables in the minority class. If the nominal variable differs between the sample and the value of its nearest possible neighbors, then this median is included in the Euclidean distance. The median is used to exclude the differences

in nominal variables by an amount associated with specific differences in continuous variables.
2. Calculation of k Nearest Neighbors: Calculates the Euclidean distance between vector variables where the k nearest neighbors (minority class) and other vector variables (minority class) are identified using continuous variables. Each nominal variable that differs between the vector variable and the value of its nearest neighbor, then the median of standard deviation which calculated is included in the calculation of the Euclidean distance.
3. Creating Synthetic Samples: The synthetic data for continuous variables were created using the SMOTE. The synthetic data for the nominal variable is the mode in most of the k nearest neighbors.
Suppose there is a sample to calculate the nearest neighbor:
F1 = 1 2 3 A B C
F2 = 4 6 5 A D E
F3 = 3 5 6 A B K
The Euclidean distance between F2 and F1 is as follows:

$$Euclidean = \sqrt{(4-1)^2 + (6-2)^2 + (5-3)^2 + Med2 + Med2} \quad (4)$$

Med is the median of the standard deviation of the continuous variable from the minority class. The median is called twice for variable number 5: B → D, and number 6: C →E which is different for the two vector variables F1 and F2.

3.4 Model Evaluation

The measurement of the performance of the classification model is an evaluation that sees the possibility of misclassification made by a classification model that can use a Confusion Matrix. [17]. The determination of classification is shown in Table 2.

Table 2: Confusion matrix

Observation	Predicted		Total
	Positive class (Y=0)	*Negative class* (Y=1)	
Positive class (Y=0)	*True positive (TP)*	*False-negative (FN)*	*TP + FN*
Negative class (Y=1)	*False-positive (FP)*	*True negative (TN)*	*FP + TN*
Total	*TP + FP*	*FN + TN*	*TP+FP+FN+TN = n*

Several indicators are generated from the confusion matrix and can be used to measure the performance value of the classification model:
1) Accuracy = [TP+TN / n] is the percentage of the classification model that predicted correctly.
2) Sensitivity = [TP / TP+FN] is the percentage of positive class observations that are correctly predicted as the positive class.
3) Specificity = [TN / FP+TN] is the percentage of negative class observations that are correctly predicted as negative class.

4) Precision = [TP / FP+TP] is the percentage of prediction results that are correctly predicted as the positive class.

4 The Proposed Method

The analysis process from pre-processing data to the classification of NPF is shown in Fig 1.

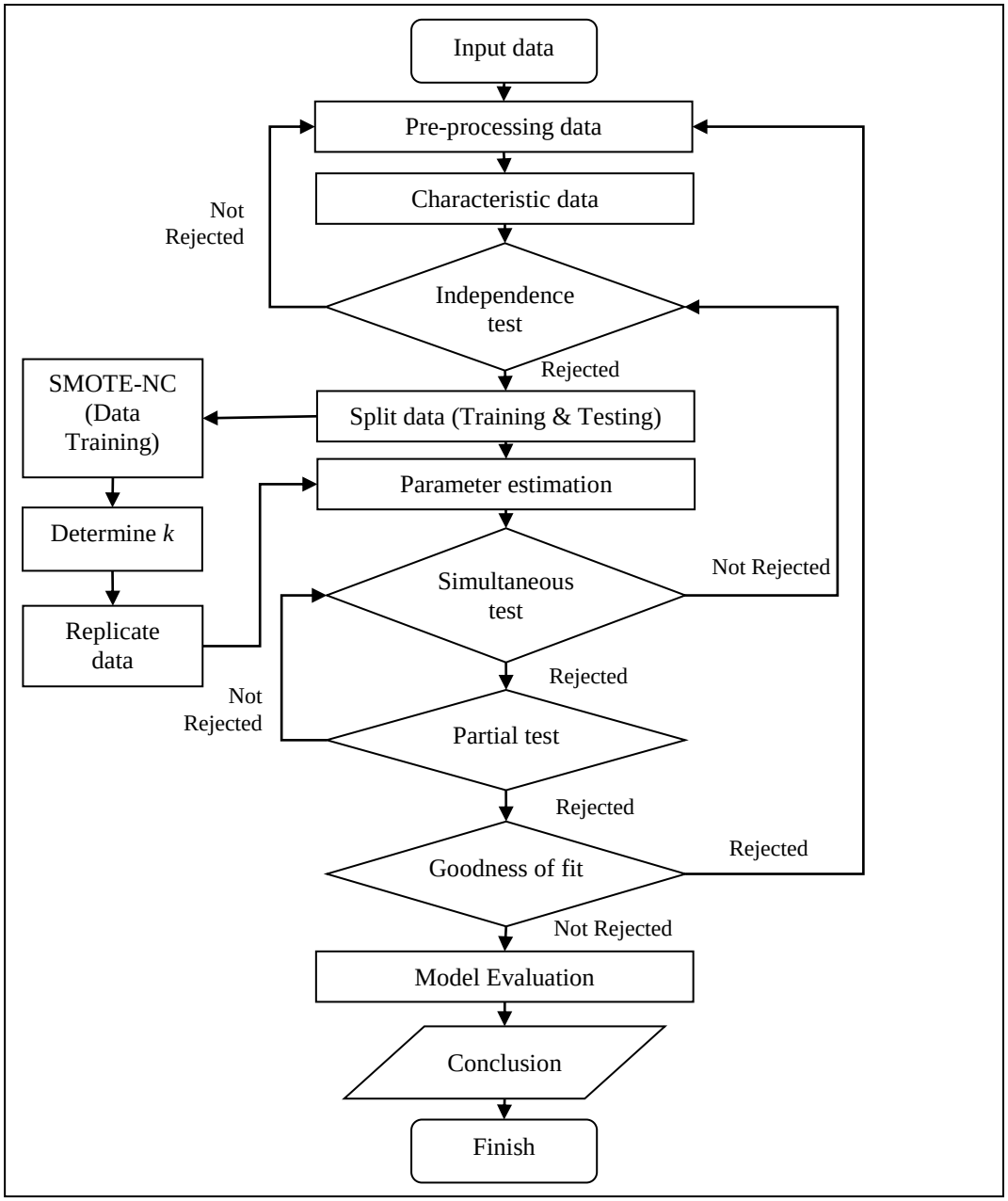

Fig 1. Flow chart of the classification model

In achieving the purpose of this study, the analytical steps will use 2 methods, logistic regression with SMOTE-NC and logistic regression without SMOTE-NC. Classification using logistic regression and SMOTE-NC methods by replicating data on minority class on training data with k = 6 replicate by 40% of the total minority class. Once obtained the synthetic data, estimate the parameter values of the factors that affect NPF using training data. Factors that affect NPF can be identified through simultaneous and partial parameter testing using training data. After obtaining the logistic regression model, it is necessary to test the goodness of fit model using training data. The model evaluation by comparing the performance of the classification model based on the level of accuracy, specificity, sensitivity, and precision [14].

SMOTE-NC needs to determine the value of k nearest neighbors to create replication data. This study used 1560 data on the training data, the more data used then the selected k number is lower. Consideration of the low number of k also sees the number of independent variables, there are 6 variables. The more independent variables then the selected k number is higher. The results of the consideration of the amount of data and independent variables generate k = 6, 7, 8, 9, and 10. The best model logistic regression with SMOTE-NC is obtained when k = 6 with the percentage of NPF data replication being 40% of the total good financing because it has an accuracy of 81%, the sensitivity of 94%, the specificity of 21% which higher than other k and other proportion.

The proportion of data after replicating NPF customers to 40% and good financing to 60% is assumed to be balanced because each class has a proportion above 35%. The total NPF in this study became 520 customers (40%) and the total of good financing became 1,300 customers (60%).

5 Results, Analysis and Discussions

In this section, will describe the characteristics of customers and the classification model for NPF in an Islamic bank using logistic regression with SMOTE-NC and logistic regression without SMOTE-NC. The total number of customers who have made payment transactions in December 2020 was 2228 customers. There were 17% (371) NPF and 83% (1857) good financing. The percentage of the financing customers is shown in Fig 2.

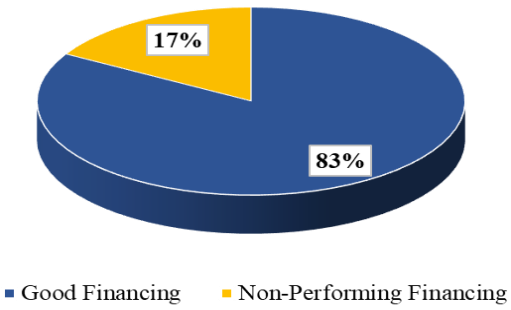

Fig 2: Financing customers

The independence test was determining the relationship between the independent variables: type of use, collateral, job, and gender with the quality of financing. The variables that were not tested for the independence test were the amount of financing, age, and financing period because the data were numeric. At the level of significance $\alpha = 0.05$ H_0 was rejected if $\chi^2 > \chi^2_{(\alpha,df)}$.

Table 3: Independence test

Variable	χ^2	*df*	$\chi^2_{(\alpha,df)}$	P-Value
Type of Use	34.92	2	5.99	<0.001
Type of Collateral	91.97	4	9.49	<0.001
Occupation	38.80	6	12.592	<0.001
Gender	1.897	1	3.84	0.168

Table 3 shows 3 variables have a P-value < 0.05, then H_0 is rejected which means the type of use, type of collateral, and occupation have a significant dependency on the quality of financing. While gender is independent. Therefore, the classification model will use variables: type of use, type of collateral, occupation, the amount of financing, age, and financing period.

5.1 Classification of Non-Performing Financing Using Logistic Regression and Synthetic Minority Over-sampling Technique

Simultaneous parameter testing to determine the effect of 6 independent variables simultaneously on the NPF. At the significance level of 0.05 then H_0 is rejected if $G > \chi^2_{(df)\alpha}$. The results of the simultaneous test are shown in Table 4.

Table 4: Simultaneous test

	Logistic Regression	Logistic Regression + SMOTE-NC
G	136.959	327.560
df	15	15
$\chi^2_{(15)0,05}$	24.996	24.996

The results of the simultaneous parameter test show that logistic regression without SMOTE-NC has *G*=136.959 and logistic regression with SMOTE-NC has *G*=327.560. Both methods have a *G* value greater than 24.996, then H_0 is rejected, which means there is at least one variable that has a significant effect on the NPF. The logistic regression model with SMOTE-NC has a *G* value greater than the logistic regression model without SMOTE-NC. This suggests that the logistic regression model with SMOTE-NC is the best. Partial parameter test to determine the effect of 6 independent variables partially on the NPF. At the level of significance 0.05 then H_0 is rejected if the P-value < 0.05. Table 5 shows that financing period, type of use, type of collateral, and occupation on logistic regression model without SMOTE-NC or logistic regression with SMOTE-NC have a P-value < 0.05. Therefore, H_0 is rejected which means the financing period, type of use, type of collateral, and occupation have a significant effect on the NPF. While the amount of financing and age has no significant effect.

Table 5: Partial test

Variable	Category	P-Value	
		Logistic Regression	Logistic Regression + SMOTE-NC
Amount Financing (X_1)	-	0.237	0.275
Period (X_2)	-	0.001	<0.001
Type of Use (X_3)	2= Investment	0.979	0.967
	3= Consumption	<0.001	<0.001
Collateral (X_4)	2= Motorcycle	0.081	0.483
	3= Certificate Property	0.404	0.007
	4= Letter for Deduction of Salary	0.001	<0.001
	5= Others	0.526	0.437
Age (X_5)	-	0.206	0.161
Job (X_6)	2= Carpentry & Craftsmen	<0.001	<0.001
	3= Informal workers	0.171	<0.001
	4= Lecture and Teacher	0.008	<0.001
	5= Student	0.012	<0.001
	6= Laborers and Farmers	0.071	<0.001
	2= Carpentry & Craftsmen	0.021	<0.001

After a partial test, it is necessary to do a simultaneous test using 4 variables that have a significant effect on the model.

Table 6: Simultaneous test using significant variables

	Logistic Regression	Logistic Regression + SMOTE-NC
G	133.972	324.279
df	13	13
$\chi^2_{(15)0,05}$	22.36	22.36

Table 7: Partial test using significant variables

Variable	Category	P-Value	
		Logistic Regression	Logistic Regression + SMOTE-NC
Period (X_2)	-	<0.001	<0.001
Type of Use (X_3)	2= Investment	<0.001	<0.001
	3= Consumption	0.979	0.967
Collateral (X_4)	2= Motorcycle	<0.001	<0.001
	3= Certificate Property	0.078	0.458
	4= Letter for Deduction of Salary	0.468	0.008
	5= Others	0.001	<0.001
Job (X_6)	2= Carpentry & Craftsmen	0.577	0.399
	3= Informal workers	<0.001	<0.001
	4= Lecture and Teacher	0.155	<0.001
	5= Student	0.009	<0.001
	6= Laborers and Farmers	0.019	<0.001
	2= Carpentry & Craftsmen	0.046	<0.001

The results of the simultaneous parameter test in Table 6 shows that logistic regression without SMOTE-NC has *G*=133.972 and logistic regression with SMOTE-NC has *G*=324.279. Both methods have a *G* value greater than 22.36, then H_0 is rejected, which means at least one variable has a significant effect on the NPF. The logistic regression model with SMOTE-NC has a *G* value greater than the logistic regression model. This suggests that the logistic regression model with SMOTE-NC is the best.

Partial parameter test to determine the effect of 4 independent variables partially on the NPF. Table 7 shows that financing period, type of use, type of collateral, and occupation using logistic regression model without SMOTE-NC or logistic regression with SMOTE-NC have a P-value < 0.05. Therefore, H_0 is rejected which means the financing period, type of use, type of collateral, and occupation have a significant effect on the NPF. The goodness of fit test to find out there is a difference between the observed results and the possible prediction results. At the significance level of 0.05, then H_0 is rejected if $\chi^2 > \chi^2_{(df,\alpha)}$ or the P-value < 0.05. The results of the goodness of fit are in Table 8.

Table 8: Goodness of fit

	Logistic Regression	Logistic Regression + SMOTE-NC
χ^2	8.045	13.938
$\chi^2_{(8)0,05}$	15.507	15.507
Df	8	8
P-value	0.4291	0.083

The results of the goodness of fit show that logistic regression without SMOTE-NC has a P-value=0.4291 and logistic regression with SMOTE-NC has a P-value=0.083. Both methods have a P-value greater than 0.05, then H_0 is not rejected, which means that the model fits and there is no significant difference between the observed results and the possible prediction results.

5.2 Performance Model Evaluation

Confusion Matrix is an evaluation of classification procedures to measure the performance of the classification model, so the probability of misclassification is known based on the criteria used. Table 9 shows based on the prediction results of the logistic regression without SMOTE-NC model in the testing data using significant variables, there are 557 customers with good financing which 557 customers were correctly predicted as good financing, and none were incorrectly predicted as NPF. Meanwhile of the 111 NPF, 107 customers were incorrectly predicted as good financing and 4 customers were correctly predicted as NPF. Then the accuracy of the prediction results is 0.84, sensitivity 1, specificity 0.04, and precision 0.84.

Based on the prediction results of the logistic regression with the SMOTE-NC model in the testing data using significant variables, there are 557 customers with good financing which 521 customers were correctly predicted as good financing,

and 36 customers were incorrectly predicted as NPF. Meanwhile, of the 111 NPF, 88 customers were incorrectly predicted as good financing and 23 customers were correctly predicted as NPF. Then the accuracy of the prediction results is 0.81, sensitivity 0.94, specificity 0.21, and precision 0.86.

Table 9: Confusion matrix of classification model

Observation	Prediction					
	Logistic Regression			Logistic Regression + SMOTE-NC		
	1	2	Total	1	2	Total
1	557	0	557	521	36	557
2	107	4	111	88	23	111
Total	664	4	668	609	59	668
Accuracy			0.84	Accuracy		0.81
Sensitivity			1	Sensitivity		0.94
Spesificity			0.04	Spesificity		0.21
Precision			0.84	Precision		0.86

1 = Good Financing; 2 = NPF

Based on the comparison results of the performance model evaluation of the classification model in Table 9, it can be concluded that the best model is the logistic regression with SMOTE-NC. The SMOTE-NC method handles the imbalance data and increases the specificity from 0.04 to 0.21, the precision from 0.84 to 0.86. If the performance of the classification model has a low specificity and is close to 0 then no customer is predicted to be NPF. However, the accuracy and sensitivity of the logistic regression model with SMOTE-NC are slightly lower but not close to zero like the specificity of the logistic regression model.

6 Conclusion

The results of the classification of NPF using logistic regression without SMOTE-NC and logistic regression with SMOTE-NC shows that the financing period, type of use, type of collateral, and occupation have a significant effect on the NPF. The classification model using logistic regression has an accuracy of 0.84, a sensitivity of 1, a specificity of 0.04, and a precision of 0.84. The classification model using logistic regression with SMOTE-NC has an accuracy of 0.81, a sensitivity of 0.94, a specificity of 0.21, and a precision of 0.86. The best model selected in this study is the logistic regression with SMOTE-NC because SMOTE-NC handles the imbalance data and increases the specificity from 0.04 to 0.21. Even though the accuracy and sensitivity are slightly lower, they are not close to zero like the specificity in the logistic regression model.

ACKNOWLEDGEMENTS
Thanks to the Department of Business Statistics, Faculty of Vocational who has provided financial support.

References

[1] BPR Shariah Publication Report. (2021).https://www.ojk.go.id/id/kanal/perbankan/data-dan-statistik/laporan-keuangan-perbankan/Default.aspx.

[2] Shariah Banking Statistics (2021). https://www.ojk.go.id/id/kanal/syariah/data-dan-statistik/statistik-perbankan-syariah/Documents/Pages/Statistik-Perbankan-Syariah---Oktober-2020/SPS%20Okt%202020.pdf.

[3] Wangsawidjaja. (2012). Pembiayaan Bank Syariah. PT Gramedia Pustaka Utama. Jakarta

[4] Fianto, B.A., Maulida, H., Laila, N. (2019). Determining Factors of Non-Performing Financing in Islamic Microfinance Institutions. DOI: https://doi.org/10.1016/j.heliyon.2019.e02301

[5] Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K. (2016). Logistic Regression Analysis of Predictors of Loan Defaults by Customers of Non-Traditional Banks in Ghana. *Euopean Scientific Journal*, 12(1), 175.

[6] Obare, D.M., Njoroge, G.G., Muraya, M.M. (2019). Analysis of Individual Loan Defaults Using Logit Under Supervised Machine Learning Approach. *Asian Journal of Probability and Statistics*, 3(4), 1-12.

[7] Wibowo, H.E., Mulyati H., Saptono I.T. (2019). Improving Credit Scoring Model of Mortgage Financing with Smote Methods in Sharia Banking. DOI 10.18551/rjoas.2019-08.07.

[8] Santoso, N., Wibowo, W., Himawati, H. (2019). Integration of Synthetic Minority Oversampling Technique for Imbalanced Class. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(1), 102–108

[9] Barro, R.A., Sulvianti, I.D., Afendi F.M. (2013). Penerapan Synthetic Minority Oversampling Technique (SMOTE) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu. *Xplore*, 1(1), 1-6.

[10] Yu Zhang., Yongsheng Guan., Gang Yu., Haixia Lu. (2016). Recognizing and Predicting the Non-Performing Loans of Commercial Banks. *International Journal of Signal Processing, Image Processing and Pattern Recognition.* 9(11), 211-220.

[11] Johan S. (2018). Determinants of Credit Decision in Consumer Financing: An Empirical Study on Indonesia Auto Financing. *International Journal of Business and Entrepreneurship*, 4(3), 291-298.

[12] Financial Service Authority Regulations of Indonesia Number 29/POJK.03/2019 on The Quality of Productive Assets and the Establishment of Allowance for the Elimination of Productive Assets of Sharia Banks. (2019). https://www.ojk.go.id/id/regulasi/Pages/-Kualitas-Aset-Produktif-dan-Pembentukan-Penyisihan-Penghapusan-Aset-Produktif-Bank-Pembiayaan-Rakyat-Syariah.aspx.

[13] Agresti, A. (2007). An Introduction to Categorical Data Analysis Second Edition. JohnWiley & Sons, Inc. Canada.

[14] Hosmer, D.W. & Lemeshow, S. (2000). Applied Logistic Regression. John Wiley & Sons. New York
[15] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Study*, 16, 321-357.
[16] Mustaqim, M., Warsito, B., Suraso, B. (2019). Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tida seimbang pada prediksi pemakaian alat kontrasepsi implan. *Jurnal Ilmiah Teknologi Sistem Informasi*. 5(2): 116-127.
[17] Faisal, M.R., Nugrahadi D.T. (2019). Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R. Scripta Cendekia. Kalimantan Selatan.

Notes on contributors

Islahulhaq. *She is a student at the Department of Business Statistics, Faculty of Vocation, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. She has experience as a lecture assistant for 2 years. Her main research is on statistical modeling for finance and health.*

Dr. Wahyu Wibowo. *He is an Associate Professor at the Department of Business Statistics, Faculty of Vocation, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His main research is on regression modeling and machine learning. He is actively doing research and publishing papers in many international journals.*

Iis Dewi Ratih. *She is an Associate Lecturer at the Department of Business Statistics, Faculty of Vocation, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. Her main research is on statistical modeling, multivariate, and extreme value theory.*