Journal of Advances in Applied Mathematics, Vol. 6, No. 3, July 2021
https://dx.doi.org/10.22606/jaam.2021.63003

Discriminant Analysis and Logistic Regression Applying To Credit Risk Management

Ngongo Isidore[1,2,3], Etoua Magloire[1,2,3] Jimbo Claver[1,2,3*], Mengue Mvondo Jenner[4], Ngatom Stephane[5] and Nkague Leontine[6]

[1]Department of Applied Mathematics, University of Yaounde 1, ENSY, MMAFSP, Cameroun
[2]Department of Applied Mathematics, National School of Engineering, Yaounde, Cameroun
[3]Department of Applied Mathematics MMAFSP, Rise, Waseda Univdrsity, Tokyo, Japan
[4]Department of Applied Mathematics, University of Yaounde 1, MMAFSP, Cameroun
[5]Department of Applied Mathematics, University of Yaounde 1, NASEY, Cameroun
[6]Department of Applied Mathematics, SUP'TIC, Yaounde, Cameroun
Email: jimbo.maths@gmail.com

Abstract The financial crisis that is currently shaking the world, particularly the successive failures of the major banks have brought the issue of banking risks, including credit risk, back to the forefront. This risk must now be managed by more sophisticated methods. In this paper we present two methods that allow us to establish two functions, namely Fisher discriminant analysis and logistic regression; these two functions allow us to evaluate the risk of non-repayment incurred by a bank in the light of our data. It emerges that Fisher discriminant analysis is more effective or efficient than logistic regression for the evaluation of the risk of non-repayment of credit. Discriminant analysis and logistic regression are two methods of credit risk management here the problem we are trying to solve is how to help banks choose the most efficient method between the latter two.

Keywords: banks, ratios, risks, Fisher discriminant analysis, logistic regression.

1 Introduction

Banks are generally affected by several types of risk, among which we have market risk, option risk, credit risk, operational risk and so on. Credit risk, also called counterparty risk, is the most common risk. There are several types of credit risk, with the risk of non-repayment being the most important ([1],[2], [3] and [4]). Several research studies have been carried out to detect in advance which loans will default and which will not ([5],[6] and [7],[8]). This work is essentially based on the analysis of accounts.

The Cameroonian banking system uses classic methods to deal with credit risks [1]. Among these methods, financial diagnosis and the taking of guarantees undoubtedly occupy a central place. This situation has harmful effects on the inflation of unpaid debts, which can jeopardize the very survival of the bank. There are currently sophisticated methods for managing credit risk, including discriminant analysis and logistic regression [9], [10] and [11]. These methods correspond to a method of financial analysis that attempts to synthesize a set of ratios in order to arrive at a single indicator that makes it possible to distinguish in advance between good customers and defaulting customers. According to the Central African Banking Commission (COBAC), Cameroon alone has eight systematically important banking institutions among the ten that exist. In order to avoid falling into financial need in the CEMAC zone, and therefore Cameroon is leaving it, it is important to contribute to the construction of a rigorous method that will enable the various banks in Cameroon in general and the eight cited by COBAC in particular not to succumb to bankruptcy. For the failure of only one of its eight banks in Cameroon immediately leads to a financial need in the CEMAC zone.

In the following we will try to highlight a practical approach for the design and validation of the predictive capacity of the two score functions from the two models and establish a comparison to choose the most optimal one. Then through this present work we will try to bring answers to the following problem: "*How to allow banks to have a prediction on the non-respect of commitments by a customer*". To answer this problem, banks need information about the customer and some mathematical tools, more precisely statistics and probabilities.

2 Discriminant Analysis and Logistic Regression

Consists in defining a statistical representativeness and homogeneity of the samples. Two sub-samples must be available: one composed of firms that have experienced the event to be detected (default, bankruptcy), the other of firms that have not experienced it, deemed healthy. The data we use are from the Rabat-Kenitra of Morocco, which gives us a sample of 46 firms, consisting of 23 firms deemed to be deficient and 23 healthy. We note here that the firms are represented by the variable and the defaulting firms take the value 1 while the non-defaulting firms take the value 0.

Table 1. Main characteristics companies

Main characteristics	Healthy companies	Failing companies
Sector of activity:		
Trade	13	15
Industry	10	08

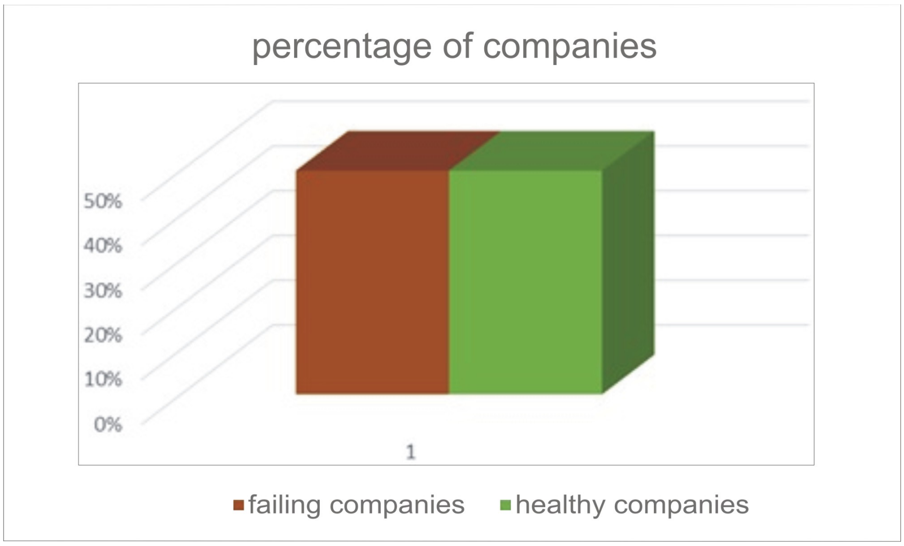


Figure 1. Ratios and capital.

2.1 Variable Eelection

The aim here is to select the variables that are important and that have the power to discriminate significantly and to avoid repeating the variables.

Choice of ratios

Three broad categories of ratios have been distinguished, as shown in the following table:

This table shows that the values taken by the seven selected ratios are dispersed. They differ greatly from one company to another. To get a preliminary idea of the discriminatory power of each ratio, we use the test of difference in student means relative to each ratio between failing and healthy firms. The results of this test can be summarized as follows:

Table 2. Ratios and capital

Aspect	Ratios	Entitled	Formula
Structural ratios			
	R_1	Financial Autonomy Ratio	Main Capital/Permanent Capital
	R_2	Immediate cash ratio	Availability/short-term debt
	R_3	Financial equilibrium ratio	Permanent capital/net fixed assets
Activity Ratios			
	R_4	Portion of financial expenses in profit	Financial charge/Profit
	R_5	Portion of financial expenses in profit	Purchase debts/purchases including all taxes
	R_6	Customer credit ratio	claims/turnover including all taxes
Profitability ratios	R_7	Financial profitability	RNE/CP

Table 3. Range and mean value of ratio

Ratios	Minimal Value	Maximal Value	Mean	Standard-deviation
R_1	0.42	1.00	0.8548	0.1792
R_2	0.00	0.54	0.1087	8.735E-02
R_3	0.45	12.91	2.3952	2.4691
R_4	0.01	0.44	0.1422	0.1092
R_5	0.12	3.29	1.5730	0.8887
R_6	0.01	2.84	0.9052	0.7916
R_7	0.00	0.21	7.217E-02	5.325E-02

The first results of our study show that the average relative to ratio 1 (financial independence ratio) is higher among healthy firms (0.96) than among failing firms (0.75). The difference between these two averages is positive (+0.21) and statistically significant. This ratio is discriminating according to the student test. This also applies to the ratios R2 (Immediate cash flow ratio), R3 (Financial equilibrium ratio), R5 (Supplier ratio) and R7 (Financial profitability). On the other hand, the average relative to ratio 6 (client ratio) is higher among defaulters (1.37) than among healthy students (0.44). The customer delay is longer for defaulters than for healthy firms. However, these basic observations do not allow us to make a definitive decision on the most discriminating variables.

2.2 Construction of the Discriminant Function

The processing of our database through the SPSS software allowed us to identify the following score function:

So our score function can be written like this:

$$S(x) = 2.071R1 - 0.036R2 + 0.070R3 + 1.662R4 + 0.706R5 - 1.219R6 + 8.224R7 - 2.772$$

Assignment to groups will be made according to the centroids of the groups, i.e. by comparison with an "average" discriminant score for each group. This average score is calculated from the discriminant function, in which the individual values are replaced by the means of the independent variables for the group in question. The average discriminant scores for the two groups are given as follows:

Each individual discriminant individual score is then compared to the two average scores and assigned to the group with the closest match. The predictive capacity of the score function is tested either by statistical tests using probabilistic hypotheses or by a pragmatic test using the confusion matrix. For the first tests, we use Wilks' canonical and Lambda correlation.

Wilks' Lambda value is low, and is equal to 0.346, and therefore closer to 0 than 1, with a chi-square having a zero significance level. This means that at the overall level, the difference in group means is significant. To ensure that the discriminant function correctly classifies firms into subgroups, the confusion

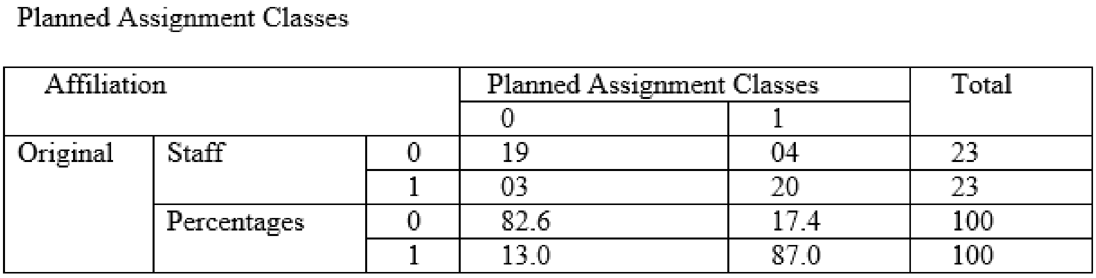
Table 4. Healthy and failing companies

Ratios	Healthy companies	Failing companie	Deviation	T-test	Signification
R_1	0.9617	0.7478	0.2139	5.019	0.000^*
R_2	0.1396	7.783E-02	6.174E-02	2.538	0.015^{**}
R_3	3.5343	1.2561	2.2783	3.498	$0,001^*$
R_4	0.1813	0.1030	0.0783	2.579	$0,013$
R_5	1.9548	1.1913	0.7635	3.234	0.002^*
R_6	0.4383	1.3722	0.9339	2.256	0.000^*
R_7	8.913E-02	5.522E-02	3.391E-02	2.256	0.029$**$

Table 5. Scoring

Function	
R_1	2.071
R_2	-0.036
R_3	0.070
R_4	1.662
R_5	0.706
R_6	-1.219
R_7	8.224
Constante	-2.772

matrix is analyzed, which groups together the well-ranked and poorly ranked firms. This is the most commonly used means. The confusion matrix of our score function is as follows:



Planned Assignment Classes

Affiliation			Planned Assignment Classes		Total
			0	1	
Original	Staff	0	19	04	23
		1	03	20	23
	Percentages	0	82.6	17.4	100
		1	13.0	87.0	100

Figure 2.

This matrix shows that the score function extracted above makes it possible to classify one year before the occurrence of the failure 84.78% (19+20/46) of the companies correctly. This rate can be broken down as follows:

Table 6. Average Score per Company

Function 1	
Affiliation	Average Scores
0: Failing company	-1.343
1: Healthy company	+1.343

Table 7. Canonical Correlation

Function	Eigen Value	% of Variance	% cumulate	Canonical correlation
1	1.886	100.00	100.00	0.808

Table 8. Canonical Correlation

Test of the function(s)		Khi-II	dF	Signification
1	0.346	42.929	7.00	0.000

- The percentage of well ranked companies for healthy companies is equal to 20/23=87%.
- The percentage of assets ranked for failing companies is equal to 19/23= 82.6%. On the other hand, the error rate (misclassified companies) is only equal (7/46) = 15.21%.

The error of the first type (classifying a failing firm by using the score function as a healthy company): this rate is equal to 4/23=17.4%; and the error of the second type (classifying a healthy firm as a failing firm by the model): this rate is equal to 3/23=13%.

2.3 Construction of the Score Function by Logistic Regression

To build our function we use python or we call the fit() function which allows us to obtain the following table:

Table 9. Results of the logistic regression model

Ratios	Coef	Std.Error	z value	Pr(>|z|)	$IC_{90\%}$
R_1	42.85	28.15	1.52	0.04	$[0.45; 5.30]$
R_2	-3.02	2.27	-1.33	0.10	$[27.02; 27.10]$
R_3	0.40	28.15	0.01	0.14	$[-1.30; 7.04]$
R_4	-7.79	5.90	-1.32	0.17	$[0.45; 5.30]$
R_5	0.53	0.26	2.03	0.03	$[0.30; 3.96]$
R_6	-2.72	2.04	-1.33	0.13	$[-12.40; 0.30]$
R_7	10.49	5.86	1.88	0.05	$[0.51; 9.26]$
Constant	17.43	9.73	1.79	0.07	$[0.21; 5.60]$

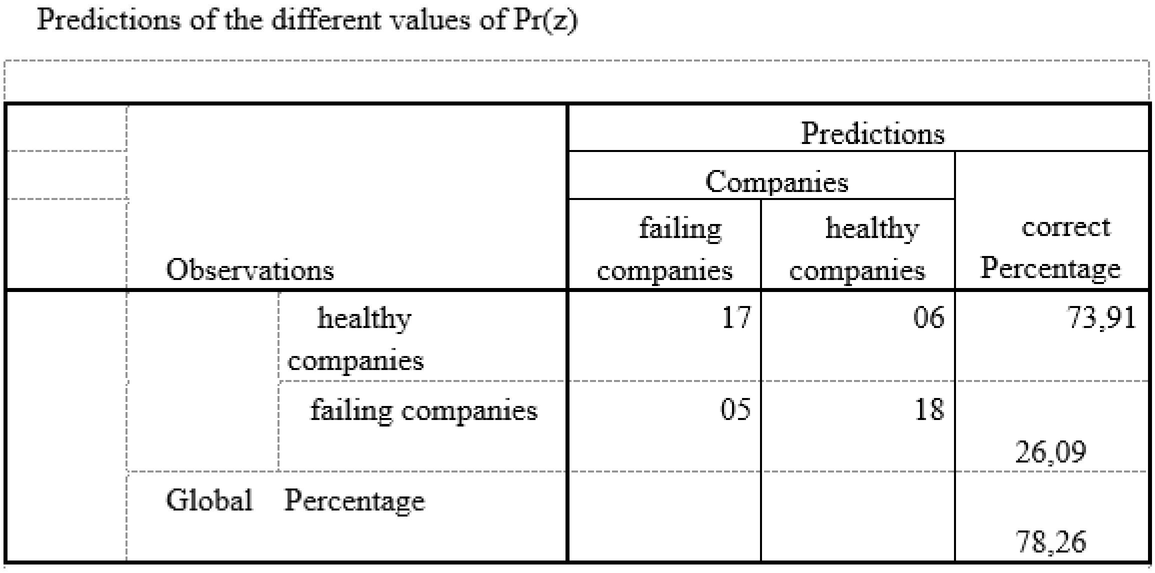
From the different values of Pr(z) we see that the ratios used for our function are R1, R5, and R7; we obtain the following score function:

$$S_R(X) = 4.85R1 + 0.53R510.49R7 + 17.43$$

Our function being to elaborate it only remains for us to validate the model.

3 Homser Lemeshow Test

The test of Homser Lemeshow in Python was quite complicated for us because Python does not have a library capable of the Homser Lemeshow formula. We used the hoslem.test function of the ResourceSelection library of the R software which allowed us to perform this test and we got the following result :



Predictions of the different values of Pr(z)

	Predictions		
	Companies		
Observations	failing companies	healthy companies	correct Percentage
healthy companies	17	06	73,91
failing companies	05	18	26,09
Global Percentage			78,26

Figure 3.

Table 10. Hosmer Lemeshow Test at 5%

Khi-2	**Df**	**p-value**
36	8	$8.97x10^{-5}$

At the 5% significance level; the model fit is good because the p-value of the chi two statistic at 8 degrees of freedom is less than 5%. Consequently, H1 hypothesis is rejected and the conclusion is that the model is calibrated and therefore valid.

The Comparison of the two models (Fisher discriminant analysis and logistic regression) in terms of predictability shows the performance of the Fisher discriminant analysis technique compared to logistic regression. Indeed, the percentage of good rankings from the Fisher discriminant analysis is better than the logistic regression.

Table 11. Fisher discriminant analysis result and logistic regression

Result of Fisher's discriminant analysis	**Result of logistic regression**
87%	78,26%

4 Conclusion

This work allowed us to develop two statisticals techniques emanating from the scoring method : Fisher discriminant analysis and logistic regression. We presented the practical approach of the construction of the score function for each technique. The validation of Fisher's discriminant analysis was carried out thanks to canonical correlation and Wilks' Lambda test which allowed us to obtain the following function:

$$S(X) = 2.071R1 - 0,036R2 + 0.070R3 + 1.662R4 + 0.706R5 - 1.219R6 + 8.224R7 - 2,772$$

In terms of logistic regression the model was validated by Homser Lemeshow's test and we obtained the following function:

$$S_R = 4.85R1 + 0.53R5 + 10.53R7 + 17.43$$

We can say that Fisher's discriminant analysis is the best model for our database because it considered all the ratios are discriminant and allow us to explain a client's belonging to a modality It should be noted that the sample size of our work was small, which prevented us from defining a threshold, and gave us overly optimistic results; this is due to the lack of access to real data; the present work can be extended by taking into account a larger number and a greater variety of variables, especially qualitative ones. Although these two methods are classical in research we will be able to make a comparison with new methods such as artificial neural networks.

References

1. R. Rakotomalala, "regression logistique pages," *Univerrsite Lumiere. Lyon 2*, 2015.
2. H. Mathlouthi, *Cours de methode de scoring*, 2013 - 2015.
3. H. D.S and L. S, "Applied logistic regression," *Wiley Interdisciplinary Reviews*, 2003.
4. G. Saporta, "Probabilites, analyse de donnees et statistiques," 2011.
5. C. Marie, "Analyse discriminante lineaire et quadratique," 2015.
6. S. G. Droesbeke J. J., Lejeune M., "Modeles statistiques explicatives pour donnees qualitatives," 2005.
7. D and F. A, *Calcul des Probabilites*, Dunod, Ed.
8. J. H. et al., "Time series modelling and forecasting for a system of credit and debit in the cameroon's nsif," *Journal of Advances in Applied Mathematics*, vol. 5, no. 2, pp. pp: 41 – 56, 2020.
9. M. S, "Applied logistic regression analysis », 2nd ed," 1997.
10. J. H. et al., "Modelling risk of non-repayment of bank credit by the method of scoring." *Journal of Advanced Statistics,*, vol. 4, no. 4, pp. pp 35 – 49, 2019.
11. B. P and L. M, "Probabilites," 1998.