Scholars Journal of Physics, Mathematics and Statistics **ISSN 2393-8056 (Print)**
Sch. J. Phys. Math. Stat. 2015; 2(2A):176-190 **ISSN 2393-8064 (Online)**
©Scholars Academic and Scientific Publishers (SAS Publishers)
(An International Publisher for Academic and Scientific Resources)

Comparison of criteria for the selection of discriminating variables: Application in Credit-Scoring

Hicham Y. Abdallah, Afif A. Hayek
Department of Applied Mathematics, Faculty of sciences-1, Lebanese University, Hadath, Lebanon

***Corresponding Author:**
Hicham Y. Abdallah
Email: habdalah@ul.edu.lb

Abstract: Banks want to reduce the credential risk by applying rules in order to classify the new loan seekers into “good customers” and “bad customers”. Searching past data is the best solution to build a statistics strategy to show this kind of risk. In general, a lot of data should be analyzed using “Data Mining”, the computational process of discovering patterns in large data sets involving methods, formalizing the problem of credential risk that the bank is seeking to resolve in terms of data (classification tree), while the dependent variable is qualitative and takes two forms: "good payers" and "defaulters". From that, prepare the data for treatment (selection of the most discriminating variables, collinearity diagnosis). Finally model the data by logistic regression and the decision tree CART. This article aims to build these two classification models from a database of 1000 customers by using first the chi-square criteria χ^2 and secondly Rand as a detector of discriminating variables in order to choose the most appropriate criterion.

Keywords: Data Mining, credit scoring system, logistic regression, decision tree CART, χ^2 criteria, Rand, score function**.**

INTRODUCTION

This study aims to reduce the costs generated by the fact that many customers do not repay their loans. In other words, it aims to reduce the credential risk for a lending agency.

This study develops an automated rule for deciding the granting or denial of the loan, based on the past experience of the financial institution. This automated rule will be based on a model of classification of individuals into two classes: “ good payers” and “defaulters”. It is therefore necessary to develop models of Credit Score. Credit Scoring is a numerical expression based on a level analysis of a person's credit files to represent the creditworthiness of that person. It is primarily based on a credit information report typically obtained from credit bureaus.

In what follows, we focus on some techniques of Data Mining on a set of public data to build a credit scoring tool that can be used for lending. So, from past data, we will build models of Credit Scoring that can be used to predict the behavior of future customers and avoid bad payers.

Each developed model will provide a "grid score", that first calculates the number of points for the customer (if the score is high, then the risk attached to the customer is important) and secondly evaluates the performance of these models depending on the detection of discriminating variables. To do this, we determine the area under the ROC curve and the apparent error rate on these models based on the χ^2 and Rand criteria. Afterwards, we compare the results obtained by these two criteria with the aim of identifying the most effective one.

Associational study of qualitative variables.

We proceed to build contingency tables corresponding to the intersection of the two qualitative variables V_t and $V_{t'}$, with modalities p and q respectively and to determine the association between these two variables, denoted $\Omega(V_t, V_{t'})$, using a known contingency criteria: Rand, Chi-square, Belson, and others. It is therefore necessary to define the relationship between the two variables. For this, the dissimilarity between them is defined as the complement of their similarity $\Omega(V_t, V_{t'})$to the average of their owns and similarities: $\Omega(V_t, V_t)$and$\Omega(V_t, V_{t'})$:

$$\overline{\Omega}(V_t, V_{t'}) = \frac{\Omega(V_t, V_t) + \Omega(V_{t'}, V_{t'})}{2} - \Omega(V_t, V_{t'}).$$

Inorder to avoid the calculation of probability thresholds, a criterion $H_{tt'}$, based on the comparison of $\Omega(V_t,V_{t'})$ with respect to $\overline{\Omega}(V_t,V_{t'})$, is defined by:

$$H_{tt'}=\Omega(V_t,V_{t'})-\overline{\Omega}(V_t,V_{t'})=2\Omega(V_t,V_{t'})-\frac{\Omega(V_t,V_t)+\Omega(V_{t'},V_{t'})}{2}$$

We say that the two variables are related if $H_{tt'}\geq 0$,which implies that:

$\Omega(V_t,V_{t'})\geq\overline{\Omega}(V_t,V_{t'})$,where $\Omega(V_t,V_{t'})$ is a criterion of the ones listed above: Rand,Belson,etc…

In this article,we are interested in introducing the Rand's criterion $R(V_t,V_{t'})$defined by:

$$R(V_t,V_{t'})=\frac{2\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}^2-\sum_{u=1}^{p}n_{u.}^2-\sum_{v=1}^{q}n_{.v}^2+n^2}{n^2}$$

where

$n_{uv}$is the number of individuals having modality u of V_t and modality v of Vt'

$n_{u.}$ is the number of individuals with the modality u of V_t

$n_{.v}$ is thenumber of individuals with the modality v of $V_{t'}$

pis the number of modalities of V_t

q is the number of modalities of $V_{t'}$.

We define, $\overline{R}(V_t,V_{t'})$ by:

$$\overline{R}(V_t,V_{t'})=\frac{R(V_t,V_t)+R(V_{t'},V_{t'})}{2}-R(V_t,V_{t'})$$

such that $R(V_t,V_t)$ =1,and we find that

$$\overline{R}(V_t,V_{t'})=\frac{\sum_{u=1}^{p}n_{u.}^2+\sum_{v=1}^{q}n_{.v}^2-2\sum_{u=1}^{p}\sum_{v=1}^{q}n_{uv}^2}{n^2}$$

and so,

$$H_{tt'}=R(V_t,V_{t'})-\overline{R}(V_t,V_{t'})=2R(V_t,V_{t'})\text{-1.}$$

Consequently, $H_{tt'}\geq 0$ once $R(V_t,V_{t'})\geq\frac{1}{2}$. For details,see [1].

METHODOLOGY

In this article,we are interested in applying some techniques of Data Mining on a set of public data to build a credit scoring tool, using the following approach:

- Calculate χ^2 and the Rand criteria between the target variable Y and the qualitative variables and retain those that are significant.
- Perform AFCM to hold the axes showing the highest percentage of variance.
- For selected variables, perform a logistic model using the two criteria in order to compare the odds ratio and group the modalities with the same probability.
- Analyze the max likelihood estimates before and after grouping the modalities.
- Calculate the scoreboard in order to identify good and bad payers.
- Build the decision tree in order to compare the results and calculate the error rate.
- Compare the results obtained after the χ^2 test analysis and those obtained by Rand then proceed to choose the most appropriate criterion.

Illustrative example.

To illustrate our methodology, we want to treat a set of data concerning the credits demanded. This data is composed of 1000 records (files) described by 19 explanatory variables and a target variable Y. Note that Y has 2 modalities: "paid" which gives the significance of a good payer, the one whose record has never been known unpaid. The second modality is "not paid the sign of a bad payer where a bad payer being a client who at least didn't pay 2 monthy installments (terms). The data set is known as the "german credit data".

We will study Y under different variables shown in the following table:

No	Variable name	Signification of the variable	Taken values
1	Accounts	The average balance on current account	1= CC<0 euro 2= CC[0-200 euros] 3= CC>= 200 euros 4=no account
2	Duration	The term of the loan in months	
3	History	The repayment history of the applicant	1 = outstanding in other bank 2 = outstanding passes 3 = outstanding loans without delay 4 = Credit passes without delay 5 = no credit or reimbursed all
4	Object	The purpose of credit	1 = new Car 2 = Used Car 3 = Furniture 4 = Video HIFI 5 = Appliances 6 = Work 7 = Studies 8 = Training 9 = Business 10 = Other
5	Amount cred	The credit amount in euros	
6	Savings	Savings deposits	0 = no savings 1 = <100 euros 2 = [100-500 euros [3 = [500-1000 euros] 4 => = 1000 euros
7	Age	Age	
8	Old	The job tenure	1 = unemployed 2 = empl <1 year 3 = empl [1-4 years [4 = empl [4-7 years [5 empl> = 7 years
9	Txteffort	The maximum monthly amount that a borrower can spend the loan repayment	1 = No Endt 2 = Low Endt 3 = Middle Endt 4 = Endt Fort
10	Situat	Family situation	1 = Male divorce / separate 2 = female div / Sep/ married 3 = single man 4 = Male married / widower
11	Warranty	The guaranties	1 = no warranty 2 = co-borrower 3 = guarantor

12	Resid	How long have the person been in his/her residence	1= Resid<1 year 2= Resid [1-4 years[3= Resid [4-7 years[4= Resid >= 7 years
13	Property	Goods held value (outside the bank)	1 = immobile 2 = Life Insurance 3 = car or other 4 = Not well known
14	Credit_ext	Other held credits(outside the bank)	1 = Other Banks 2 = Credit Institutions 3 = No credit
15	Habitat	Status of the home	1 = tenant 2 = owner 3 = free housing
16	Nbcred_in	The number of credits already held in the bank	1 = 1 credit 2 = 2 or 3 credits 3 = 4 or 5 credits 4 = more than 6 credits
17	Employment	The type of occupied job	1 = unemployed 2 = unskilled 3 = Employee / worker qualifies 4 = card
18	Nbpers	The number of dependents	1= 0-2 persons 2=>= 3 persons
19	Telephone	The existence of a telephone number	1= without tel 2= with tel
20	Target(Y)	Status of customer	0= pay 1= unpaid

This data set consists of the dependant variable Y, with 700 terms “paid” and 300 terms “unpaid”, and among 19 explicative (explanatory) variables. There are three of quantitative type (credit period, amount of credit, applicant’s age). To attain homogeneity (necessary for the next steps), we discretize the continuous variables (quantitative) into qualitative variables. The optimal procedure in the SPSS software realizes the discretization. This method is based on the Entrpopy minimization principle, which is called MDLPC, proposed by Fayyad U [4].

2- Detection of the discriminating variables

To choose the most discriminating variable, when all the variables are qualitative: i.e., to measure the association with the target variable Y, several criteria can be used like the χ^2 independence test, Rand, Jordan, among others. We are interested in our study only in χ^2 -test and Rand, and we’ll choose the criterion that gives us the best results. Using χ^2 -test, we conclude that the variables (txeffort, nbcred-in, telephon, Resid, and nbpers) are not linked with the target variable Y.

Similarly, using Rand criterion $R(V_t, V_{t'})$, we say that two variables are related if $R(V_t, V_{t'}) \forall\, t, t'$ .We conclude that amount tcred, credi-ext, habitat, Nbcred-in, Nbpers, telephone and age are not linked with Y as well.

3- Detection of collinearity

To examine the link between the presence of the remaining variables after the elimination of those with low association with the variable to be explained, we will calculate the Cramer's V since all variables are qualitative. Here are the first ten values of Cramer's V in descending order calculated after crossing out the set of variables with themselves by using the chi-square test. The Cramer’s V exceeding 0.4 in absolute value are considered as "troublesome" values.

Variable	Cramer's Value
Property * habitat	0.55318
amountcred* duration	0.38928
amountcred*object	0.34158
Age*habitat	0.30965
Age*situate	0.29277
Property*amountcred	0.27542
Duration*object	0.26299
Property*duration	0.24960
Age*old	0.23115
Credit-ext*history	0.21537

This table shows that only the two variables "goods" and "habitat" are linked through the modality "owner" of the variable "habitat". It is believed that this is not sufficient to remove one of these two variables.

After crossing the remaining sets of variables with themselves using Rand, we get the following Cramer's V values

Variable	Cramer-abs
Duration*object	0.263
Property*duration	0.25
Employment*duration	0.206
Property*object	0.205
Employment*property	0.194
Employment*object	0.186
History*object	0.163
Situat*object 0.150	
Object*Txteffort 0.137	
Property*Resid0.136	

All values are less than 0.4, so they cannot be taken as linked two by two.

Remark.

By performing a Multiple Correspondence Factorial Analysis (MCFA) over the variables obtained by the preceding two criteria, we note that the first 6 axes explain the same percentage of information attained by the two criteria (χ^2 and Rand) (97.85%)

4-Development of models

4.1 Modeling Score

The modeling is based on previous observations. For a number of loan granted : the payer quality is a qualitative variable Y with two modality (good or bad) and the data collected during the submission of the loan are variables labeled $X_1, X_2, \ldots, X_P$.

The "scoring" techniques that are most widely used in the banking sector use linear methods for their simplicity and robustness.

A score is a risk score which is calculated as a linear combination of variables; that is, S =$\sum_{i=1}^{p} a_i \;\; X_i$ where a_i are the coefficients that are being optimized for the prediction of Y.

To calculate the coefficients a_i , there are various estimation techniques including the four main ones: the Fisher's linear discriminant function, the Logit model (also known as logistic regression), the naive Bayes classifier, and the decision tree. Applying these four methods once on the variables obtained using the chi-square test as a detector of the most discriminating variables and once on those obtained using Rand allow us to better select the most robust model.

4.2 Construction of training and testing samples

The population must be first randomly partitioned whatever the modeling approach is, into two different samples: one for training and one for testing.

- The training sample with which the model is built must contain 60% to 70% of the observations.
- The testing sample consisting of 30 to 40% of the remaining observations is used to check the stability of the model. The model on the testing sample, for which the value of the target variable is knownand compared with the value predicted by the model,should be checked.

4.3 Modeling

In this section, we apply the two modeling approaches: logistic regression and decision tree.

A- Logistic regression

The logistic regression is used to examine the relationship between a binary variable (which is the case in our study) and several explanatory variables using the method of maximum likelihood. This kind of relationship is often non-linear and should be used as a function of logistics category.

A logistic function can also be transformed into a linear function by applying the "odds"$\frac{p_i}{1-p_i}$, or the transformation ln (odds), called "logit". This feature makes it easily an interpretable model.

Development of model

It is necessary to do first a coding by 0/1 the modality of qualitative variables, excluding the reference category. For example, for the variable accounts, the modalities are encoded as follows:

Accounts	C1	C2	C3
<0 euro	1	0	0
[0-200 euros [	0	1	0
>=200 euros	0	0	1
No accounts	0	0	0

Secondly, we must identify the variables to exclude from the full model and that include all the explanatory variables and to reduce them to those that are significant.

Each variable with a **sig.**> 0.05 should be excluded from the model since it is not significant, and therefore the underlined selected variables must be eliminated from the model.

Note that the variable "object" is significant insofar as these modalities are not. It is then necessary to group these modalities together and the resulting variable will be with two modalities instead of three. Similarly for the variable "savings", the 2nd modality is not significant and thus we grouped the "no savings" with "<100 or [100-500 [.

For variables obtained using χ^2, the following results are obtained:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ACCOUNTS			54.088	2	.000	
	ACCOUNTS(1)	2.091	.288	52.885	1	.000	8.097
	ACCOUNTS(2)	1.414	.267	27.925	1	.000	4.110
	HISTORY			8.199	2	.017	
	HISTORY(1)	1.067	.383	7.773	1	.005	2.906
	HISTORY(2)	.479	.248	3.719	1	.054	1.614
	DURATION(1)	-1.081	.239	20.465	1	.000	.339

	OBJECT			7.627	2	.022	
	OBJECT(1)	.231	.296	.610	1	.435	1.260
	OBJECT(2)	-.425	.295	2.073	1	.150	.654
	AMOUNT(1)	-.285	.256	1.243	1	.265	.752
	SAVINGS			11.841	2	.003	
	SAVINGS(1)	.899	.277	10.530	1	.001	2.456
	SAVINGS(2)	.979	.348	7.903	1	.071	2.663
	OLD			2.002	2	.368	
	OLD(1)	.389	.304	1.641	1	.200	1.476
	OLD(2)	.086	.267	.104	1	.747	1.090
	SITUAT			1.559	2	.459	
	SITUAT(1)	-.004	.305	.000	1	.989	.996
	SITUAT(2)	.291	.249	1.364	1	.243	1.337
	WARRANTY(1)	1.780	.601	8.781	1	.003	5.928
	PROPERTY			2.437	2	.296	
	PROPERTY(1)	-.512	.477	1.152	1	.283	.599
	PROPERTY(2)	-.671	.445	2.273	1	.132	.511
	AGE(1)	.598	.280	4.561	1	.033	1.819
	CREDIT_E(1)	.554	.251	4.882	1	.027	1.740
	HABITAT			2.290	2	.318	
	HABITAT(1)	.763	.546	1.952	1	.162	2.144
	HABITAT(2)	.474	.517	.839	1	.360	1.606
	Constant	-4.590	.860	28.494	1	.000	.010

Variable(s) entered on step 1: ACCOUNTS, HISTORY, DURATION, OBJECT, AMOUNT, SAVING, OLD, SITUAT, WARRANTY, PROPERTY, AGE, CREDIT_E, and HABITAT.

After eliminating these variables one after the other, we repeat the modeling with the remaining variables and obtain the following results:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ACCOUNTS			55.874	2	.000	
	ACCOUNTS(1)	2.060	.281	53.783	1	.000	7.844
	ACCOUNTS(2)	1.487	.261	32.500	1	.000	4.425
	HISTORY			12.110	2	.002	
	HISTORY(1)	1.267	.372	11.623	1	.001	3.551
	HISTORY(2)	.544	.242	5.045	1	.025	1.723
	DURATION(1)	-1.182	.216	30.062	1	.000	.307
	OBJECT(1)	.575	.212	7.368	1	.007	1.778
	SAVINGS			11.460	2	.003	
	SAVINGS(1)	.882	.272	10.494	1	.001	2.415
	WARRANTY(1)	1.848	.582	10.068	1	.002	6.345
	AGE(1)	.724	.250	8.356	1	.004	2.062
	CREDIT_E(1)	.537	.246	4.748	1	.029	1.710
	Constant	-5.041	.717	49.414	1	.000	.006

Variable(s) entered on step 1: ACCOUNTS, HISTORY, DURATION, OBJECT, SAVINGS, WARRANTY, AGE, and CREDIT_E.

All variables are significant and it is advantageous to include all of the model. So here is the model obtained : Target=- 5.041+2.060 *(cc<0 euro)* +1.487 (*cc[0-200 euros[or>=200 euros)* +1.267 *(outstanding credit)* +0.544 *(late credit)* - 1.182 *(<16 month)* +0.575 *(new or used car)* +0.882 (*no saving or<500)* +1.848 (*no guarantee or co-borrower)+0.724* (*19-25 years)* +0.537 *(other bank or credit institution).*

Calculating a score grid

When the model is developed by logistic regression, the regression coefficients will be replaced by new coefficients, called "points", each associated to a modality. This way of calculating a score is common in credit scoring where we add the points related to each modality, for a total number of points that is the score of the individual.

Calculation techniques

- c (j, i), the coefficient associated with the modality i of variable j.
- for each variable j, min (j) : the least coefficient c (j,i), max(j) : the highest coefficient c (j, k) and DeltaMAX =max (j) - min(j)
- Then we calculate the " total weight" which is the sum over j of all "DeltaMAX."
- Finally, for each i of the variable j, there is a number of points associated with it.

$$N\,(j\ ;\ i)=100\,\frac{c(j,i)-min(j)}{poids\_total}$$

For example for the modality « CC[0-200 euros[or CC>=200 euros » :

$$N=100\ \frac{1.487-0}{9.116}=16.3$$

	C(i,j)	Min	Max	Deltamax=max-min	Nbpoints
Accounts CC<0 euro	2.060	0	2.060	2.060	23
Accounts CC[0-200 euros[ou CC>=200 euros	1.487				16
No accounts	0				0
Duration<16	-1.182	-1.182	0	1.182	0
Duration>=16	0				13
Unpaid credit	1.267	0	1.267	1.267	14
Delayed credit	0.544				6
No credits	0				0
Object new car / used	0.575	0	0.575	0.575	6
Object interior	0				0
Savings 0 or savings <100 or [100-500[	0.882	0	0.882	0.882	10
Savings>=500	0				
Guarantor with warranty	0	0	1.848	1.848	0
Guarantor without Warranty	1.848				20
Age 19-25	0.724	0	0.724	0.724	8
Age >=25	0				0
Credit_ext other banks or credit establishment	0.537	0	0.537	0.537	6
Credit_ext no credit	0				0
Weight total=$\sum deltamax$=9.116					

We then apply the grid score to the learning and validation sample, and any credit applicant has a score equal to: *grade= 23 (cc<0 euro) + 16 (cc [0-200 euros or>=200) +13(duration>=16)+ 14 (outstanding credit) + 6(credit without delay) +6(new or used car)+10 (0 saving or<500) + 20 (without warranty)+ 8 (19-25 years) +6 (other banks or credit institution)*

The use of score requires the retrieval of information in a simple form. The solution is to classify the rating score into three value classes: low, medium and high. The last step to build the scoring tool is to slice the number of points into usually three slices score:

- Least risky for which some checks are made and mandatory customers are asked for minimum parts.
- Mediumrisky for which we need to look a little more background and perform a standard risk analysis.
- Most risky for which the application is otherwise rejected.

The thresholds slices number of points given by the MDLPC technique are 51 and 70. By dividing the points of all customers as follows: 0-51 = low risk, medium risk = 52-70, 70 = highrisk, the following table is obtained:

	paid	unpaid	total
Low risk	36.5%	3.5%	40%
Medium risk	24.9%	12.8%	37.7%
High risk	8.6%	13.7%	22.3%
Total	70%	30%	100%

We note that 40% of credit applications have a very low risk, then 37.7% have a medium risk, and finally 22.3% of applications are very risky, and thus rejected.

For variables obtained by using Rand, the following results are obtained:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ACCOUNTS			53.010	2	.000	
	ACCOUNTS(1)	2.030	.283	51.439	1	.000	7.618
	ACCOUNTS(2)	1.420	.265	28.807	1	.000	4.138
	HISTORY			10.634	2	.005	
	HISTORY(1)	1.184	.371	10.181	1	.001	3.268
	HISTORY(2)	.540	.248	4.730	1	.030	1.716
	DURATION(1)	-1.118	.223	25.195	1	.000	.327
	OBJECT			9.761	2	.008	
	OBJECT(1)	.137	.289	.226	1	.635	1.147
	OBJECT(2)	-.565	.285	3.915	1	.048	.569
	SAVINGS			8.862	2	.012	
	SAVINGS(1)	.753	.274	7.535	1	.006	2.124
	SAVINGS(2)	.874	.345	6.439	1	.110	2.398
	OLD			2.282	2	.319	
	OLD(1)	.437	.315	1.923	1	.166	1.548
	OLD(2)	.104	.273	.144	1	.704	1.109
	TXEFFORT(1)	-.519	.219	5.637	1	.018	.595
	SITUAT			4.439	2	.109	
	SITUAT(1)	.019	.298	.004	1	.950	1.019
	SITUAT(2)	.480	.239	4.051	1	.044	1.616
	RESID			1.390	2	.499	
	RESID(1)	.281	.239	1.380	1	.240	1.324
	RESID(2)	.135	.318	.182	1	.670	1.145
	PROPERTY			1.991	2	.369	
	PROPERTY(1)	-.460	.353	1.696	1	.193	.631
	PROPERTY(2)	-.394	.299	1.737	1	.188	.674
	EMPLOYMENT			2.351	3	.503	
	EMPLOYMENT(1)	.337	.732	.212	1	.645	1.401
	EMPLOYMENT(2)	.578	.381	2.307	1	.129	1.783
	EMPLOYMENT(3)	.393	.315	1.554	1	.212	1.481
	Constant	-2.788	.536	27.070	1	.000	.062

Variable(s) entered on step : ACCOUNTS, HISTORY, DURATION, OBJECT, SAVINGS, OLD, TXEFFORT, SITUAT, RESID, PROPERTY, and EMPLOYMENT.

By eliminating non significant variables (those underlined), a model with 5 variables (accounts, historical, time, savings, txeffort) is obtained.

Note that the variable "object" is significant insofar as these terms are not significant. These two modalities should be grouped together and the resulting variable will be with two modalities instead of three. Similarly for the variable "savings", modality 2 is not significant and "no savings" must be grouped with "<100 or [100-500]. After the elimination of these variables, the logistic regression with the remaining 6 variables is repeated and the following results are obtained:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	ACCOUNTS			55.878	2	.000	
	ACCOUNTS(1)	2.011	.275	53.623	1	.000	7.472
	ACCOUNTS(2)	1.487	.258	33.230	1	.000	4.422
	HISTORY			15.464	2	.000	
	HISTORY(1)	1.386	.363	14.548	1	.000	3.998
	HISTORY(2)	.648	.239	7.360	1	.007	1.911
	DURATION(1)	-1.120	.210	28.419	1	.000	.326
	OBJECT(1)	.515	.207	6.174	1	.013	1.674
	SAVINGS			9.493	2	.009	
	SAVINGS(1)	.768	.267	8.292	1	.004	2.156
	TXEFFORT(1)	-.467	.211	4.885	1	.027	.627
	Constant	-2.831	.364	60.580	1	.000	.059

Variable(s) entered on step 1

ACCOUNTS, HISTORY, DURATION, OBJECT, SAVINGS, and TXEFFORT.

All variables are significant and it is advantageous to include all of them.

Target=- 2.831+2.011 *(cc<0 euro)* +1.487 (*cc[0-200 euros[ou >=200 euros)* +1.386 *(outstanding credit)* +0.648 *(late credit)* - 1.120 *(<16 months)* +0.515 *(new or used car)*+0.768 (*no saving or<500)* -0.467*(*Null or low Endt*)*

Then calculate the score grid. Here is the table of ratings:

	C(i,j)	Min	Max	Deltamax=max-min	nbpoints
accounts CC<0 euro	2.011	0	2.011	2.011	32
accounts CC[0-200 euros[or CC>=200 euros	1.487				23
Accounts or no accounts	0				0
Duration<16	-1.120	-1.120	0	1.120	0
Duration>=16	0				18
Unpaid credit	1.386	0	1.386	1.386	22
Credit without delay	0.648				10
No credits	0				0
Object new car/used	0.515	0	0.515	0.515	8
Object interior	0				0
savings 0 OR savings <100 or[100-500[	0.767	0	0.767	0.767	12
Savings >=500	0				0
Null or low Endt	-0.467	-0.467	0	0.467	0
Medium or high Endt	0				7
Weight_total=$\sum deltamax$=6.266					

Target =32 *(cc<0 euro)* +23 *(cc [0-200 euros[or>=200 euros)* +22 *(outstanding credit)* +10 *(credit without delay)* +18 *(>=16 months)* +8 *(used or new car)* +12 (*no saving or<500)* +7(Medium or high Endt*).*

By cutting the number of point of all clients, the following table is obtained:

Percentage	Paid	Unpaid	Total
Low risk	65.6%	24.1%	74.7%
High risk	5.7%	4.6%	25.3%
Total	66.3%	33.7%	100%

We note that 74.7% of applicants have low credit risk and 25.3% are very risky and thus should be rejected.

Comparison of 2 models

Measure of the discriminating power of a predictive model by the ROC curve

The discriminating power of a score model is analyzed with a curve called the ROC curve. The area under the ROC curve is the probability that the score of an individual taken randomly from the (target = 1) is higher than the score of another from the (target = 0). The Y axis represents the sensitivity which is the probability of detecting an event versus 1-specificity which is the proportion of false events from the non-events. Note that the specificity is the probability of detecting a non-event.

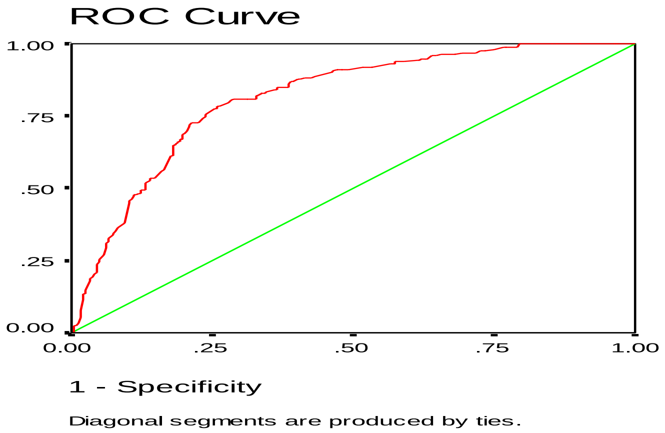
We now analyze the performance of the two models by comparing the areas under their respective ROC curves. The ROC curve is the most efficient model in terms of its ability to separate the real events from the false ones that the area under the ROC curve is closer to 1.

Interpretation of ROC curve for the 2 models obtained

Here we are interested in the calculation of the area under the ROC curve for each model as the criterion of robustness.

Below is the ROC curve for both models followed by the "Area Under the Curve (AUC)" table:

According to chi-square:

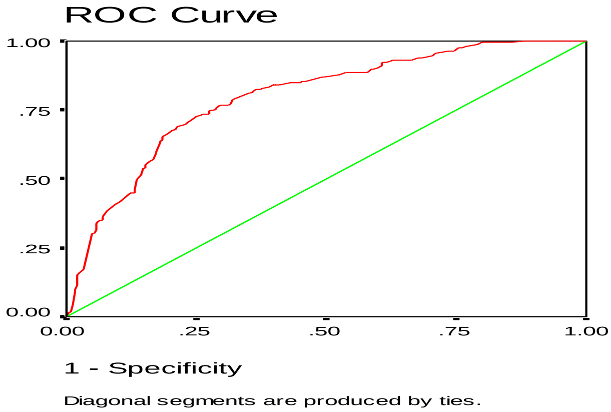


Area Under the Curve (AUC)

Test Result Variable(s): Predicted probability

Area	Std. Error(a)	Asymptotic Sig.(b)
.814	.017	.000

According to Rand :



Area Under the Curve (AUC)

Test Result Variable(s): Predicted probability

Area	Std. Error(a)	Asymptotic Sig.(b)
.796	.019	.000

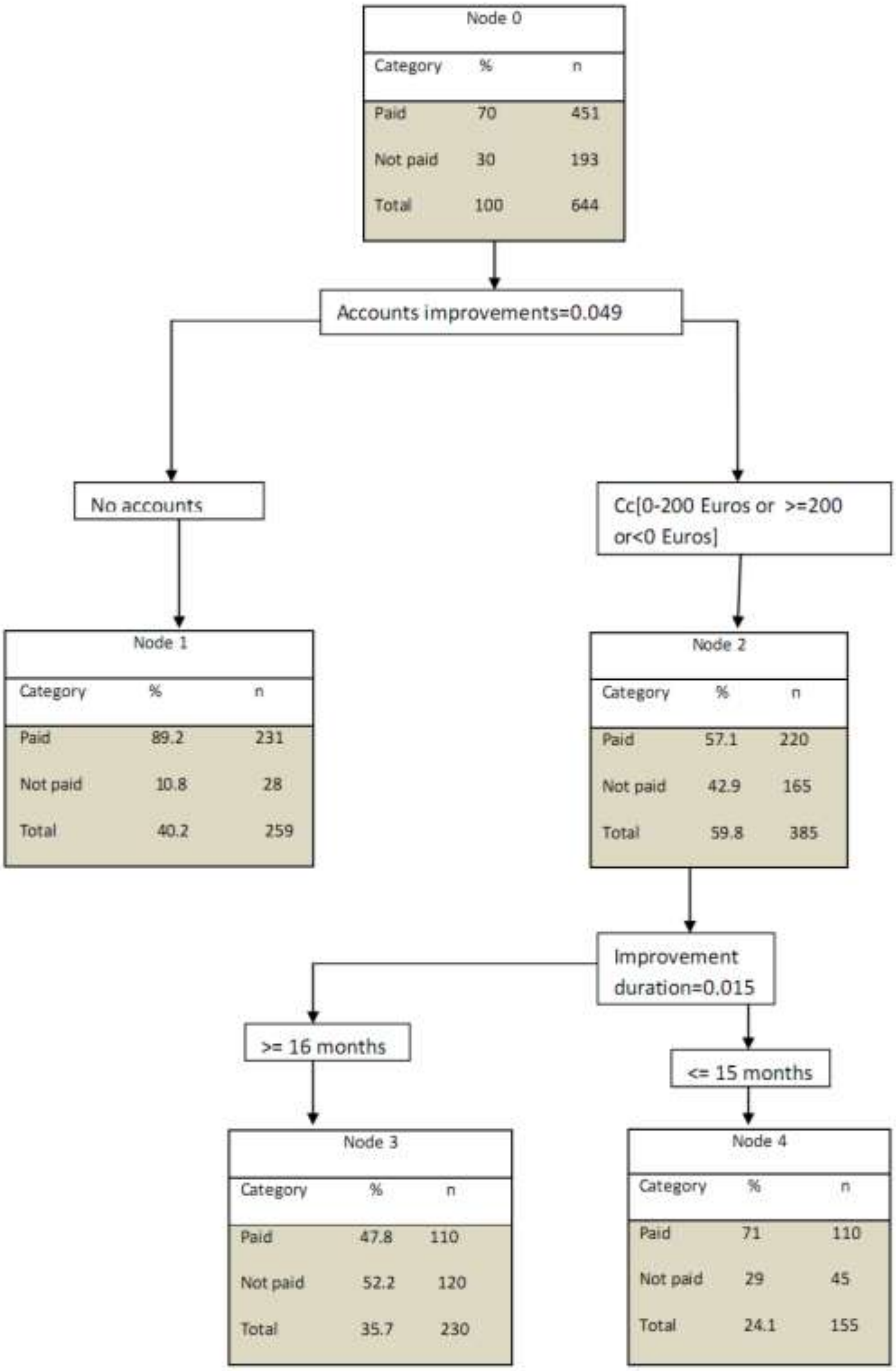
By comparing the two results, we see that the model developed by the remaining variables using the chi-square test as a detector of the most discriminating variables is more robust than that obtained by applying the Rand criterion. Note that there is not much difference between the two calculated areas.

Segmentation by the binary decision tree

This is a part of explanatory variables selected among those that are most discriminative for the nominal variable Y, and secondly to construct a decision rule for assigning a new individual to only one of K classes. This variable defines a first division of the sample into two subsets called segments. Then the procedure within each of these two segments is repeated by searching the second variable and so on. We then draw a binary tree by successive divisions of the sample into two subsets that can be distinguished as:

- Terminal segments which are not divided.
- Branch of segment **t** which comprises all descendant segments **t** where **t** is not included in the branch.

The corresponding variables selected according to the chi-square test are represented as follows:


Accounts improvements=0.049
No accounts
Cc[0-200 Euros or >=200 or<0 Euros]
Improvement duration=0.015
>= 16 months
<= 15 months

Node 0

Category	%	n
Paid	70	451
Not paid	30	193
Total	100	644

Node 1

Category	%	n
Paid	89.2	231
Not paid	10.8	28
Total	40.2	259

Node 2

Category	%	n
Paid	57.1	220
Not paid	42.9	165
Total	59.8	385

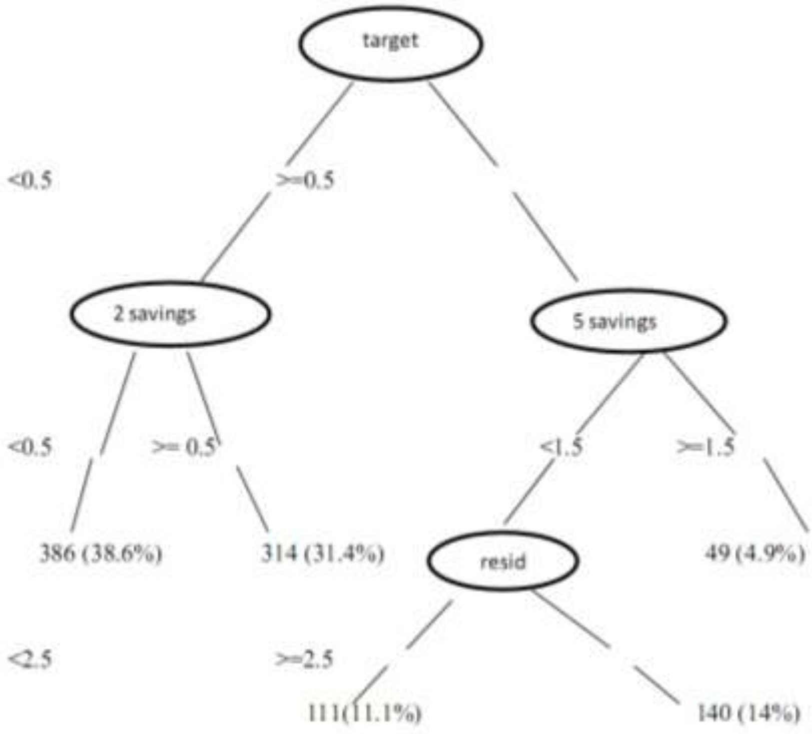
Node 3

Category	%	n
Paid	47.8	110
Not paid	52.2	120
Total	35.7	230

Node 4

Category	%	n
Paid	71	110
Not paid	29	45
Total	24.1	155

And those according to Rand:



Calculation of error rate

A classification error of the form corresponds to any terminal t of the tree associated with a cs class: *R(s/t)*=$\sum_{r=1}^{k} p(r/t)$ with r≠s and P(r/t)=$\frac{n_{r(t)}}{n}$is the proportion of individuals of segment t affected by cs class and belong to the class cr.

The Apparent Error Rate is associated with the tree:

TEA(A)=$\sum_{t\in A}\frac{n_t}{n}$ $R(s/\mathrm{t})$ =$\sum_{t\in A}\sum_{r=1}^{k}\frac{n_{r(t)}}{n}$

wherer$\neq s$, it is also called the risk of the tree and represents the proportion of “bad individuals” throughout the terminal segments.

The Apparent Error Rate (TEA) associated with the tree is the average of the classification errors in the various terminals segments.

- According to chi-square: $\mathrm{TEA}_{\mathrm{apprentissage}}$= 0.284 and $\mathrm{TEA}_{\mathrm{test}}$=0.331
- According to Rand: $\mathrm{TEA}_{\mathrm{apprentissage}}$= 0.175 and $\mathrm{TEA}_{\mathrm{test}}$=0.179

The ranking of 100 individuals taken randomly from the population shows that:

- 28.4% out of 100: to achieve a misallocation using the chi-square test as a detector of the most discriminating variables.
- 17.5% out of 100: to achieve a misallocation using the criterion of Rand as a detector of the most discriminating variables.

INTERPRETATION

By achieving anAFCM model of data mining and detecting the most discriminating according to the chi-square test and Rand criteria, we selected six factorial axes that explain 97.9% and 96.21%of the information respectively.
By performing the logistic regression followed by MDLPC method, we obtain :

For the variables detected with the chi-square test : 40% of customers are with low-risk clients, 37.7% with medium risk and 22.3% with high risk variables

For those detected by the criterion of Rand, 74.7% of customers with low-risk and 25.3% with high risk variables.

Finally, by realizing the decision tree for the two models we obtain:

According to chi-square test: the apparent misclassification rate is equal to 28.4% for the training sample and 33.1% for the test sample.

According to Rand: the apparent misclassification rate is 17.5% for the training sample and 17.9% for the test sample.

So here is the following comparison table:

	Using the chi-square test	Using the criterion of Rand
Variables used in the model	Accounts, history, savings, duration, object, amount of credit, property, old, habitat, age, credit-ext, Situat, guaranteed.	Accounts, duration, history, object, savings, old txeffort, Situat, Resid, property, employment.
Percentage of information explained by the factorial axes	97.9%	96.21%
Percentages for the logistic model	High risk : 40% Medium risk : 37.7% High risk : 22.3%	Low risk : 74.7% High risk : 25.3%
decision tree error rate	$TEA_{apprentissage}$= 28.4% TEA_{test}=33.1%	$TEA_{apprentissage}$= 17.5% TEA_{test}=17.9%

CONCLUSION

By comparing these two models, we deduce that the Rand criterion minimizes the number of variables by assigning a significant percentage of the information.

The results obtained by applying logistic regression on two categories of variables, allow us to choose the criterion of rand because it reduces the low risk and the value of the area under the ROC curve is almost equal to that obtained by chi-square.

The decision tree obtained by Rand has an error rate lower than that obtained by the chi-square test.

We conclude that the results are the same but the model made with the criterion of Rand minimizes the number of selected variables and the error rate which corresponds to the optimal solution.

REFERENCES

1. Abdallah H ; Classification d'un ensemble de variables qualitatives. R.S.A,4, 1998 ; 5-26.
2. Desbois D ; Une introduction à l'analyse discriminante avec SPSS, INRA-ESR Nancy et SCEES, la revue MODULAD, numéro 30, 2003.
3. Droesbeke J, Michel L, Saporta G ; Modèles statistiques pour données qualitatives, Éditions Technip. 2005.
4. Fayyadu; Multi-interval discretization of continuous-value attributes for classification learning, Ijcai, 1993; 2.
5. Nakache JP, Confais J ; Statistique explicative expliquée, Éditions Technip. 2003.
6. Saporta G ; la notation statistique des emprunteurs ou Scoring. CNA.M., 2002.
7. Saporta G ; Probabilités, Analyse des données et statistiques, Éditions Technip. 1990.
8. Tuffery S ; Data Mining et statistique décisionnelle, 3^{eme} édition, Éditions Technip, 2010.