41 Investigating Attribute Assessment for Credit Granting on a Brazilian Retail Enterprise

Strauss Carvalho Cunha, Emanuel Mineda Carneiro, Lineu Fernando Stege Mialaret, Luiz Alberto Vieira Dias, and Adilson Marques da Cunha

Abstract

In this article, we investigate which features are required to enhance a credit scoring model for a Brazilian retail enterprise. In order to find attributes that can improve the performance of classifier algorithms for credit granting, a national and an international survey were carried out. A logistic regression classifier was used and the main result has improved the performance of data mining classifiers. The main contribution of this article was the verification that additional financial and behavioral data increase defaulting prediction performance on credit granting.

Keywords

Credit granting · Attribute assessment · Classifier algorithms · Logistic regression · Receiver operating characteristic (ROC)

41.1 Introduction

On the last decades, the efficiency of management decisions has been representing an increase in the economic success of enterprises. Credit granting decisions became part of this scenario.

The development and use of more efficient mechanisms for credit analysis and defaulting predictions have been representing fundamental issues for the commercial success of financial enterprises [2, 8, 9].

The use of predictive models for credit analysis has been implemented by the so-called credit scoring systems [1]. These systems, based upon recent customers' historical data in financial relationship with enterprises, can provide customers' different scores, allowing adequate analysis for credit decisions [10].

This research tackles the case study of a Brazilian retailer enterprise with hundreds of stores spread around the country, providing its customers with credit cards.

Typically, a credit-seeking candidate may go to a store and request a credit card, which can be used for shopping or acquiring services.

Once the credit is granted, a customer can then perform credit card transactions or obtain some personal loans, being limited to a predetermined credit profile.

The objective of the enterprise is to develop a system that allows to identify credit defaulting customers, among other available functionalities. The system uses data mining algorithms for customer defaulting predictions.

In order to improve the performance of several algorithms that presented unsatisfactory results, using demographic variables, with the Area Under Curve—Receiver Operating Characteristic (AUC-ROC) value = 0.9, we investigated which additional attributes must be used for better defaulting predictions.

The main contribution of this article was the verification that additional financial and behavioral data increase defaulting prediction performance on credit granting.

The rest of this article is organized as follows: Sect. 41.2 presents a survey on behavioral and financial attributes used for credit granting; Sect. 41.3 describes experiments using the logistic regression and others classifiers and its results; finally, Sect. 41.4 presents some conclusions, recommendations and suggestions for future work.

S. C. Cunha
Brazilian Federal Service of Data Processing - SERPRO, Brasilia, RJ, Brazil
e-mail: strauss.carvalho@serpro.gov.br

E. M. Carneiro · L. A. V. Dias · A. M. da Cunha
Federal Institute of Education, Science and Technology of Sao Paulo - IFSP, Jacarei, SP, Brazil
e-mail: mineda@ita.br; vdias@ita.br; cunha@ita.br

L. F. S. Mialaret (✉)
Computer Science Department, Brazilian Aeronautics Institute of Technology - ITA, Sao Jose dos Campos, Sao Paulo, Brazil
e-mail: lmialaret@ifsp.edu.br

S. Latifi (ed.), *Information Technology – New Generations*, Advances in Intelligent Systems and Computing 738,
https://doi.org/10.1007/978-3-319-77028-4_41

41.2 Data Set Assessment

The choice and definition of data sets to be used in the defaulting prediction was a non-trivial process and the data quality has been influenced by the performance of the used algorithm. For this investigation, a sample of the data set was used and validated by the enterprise.

The initial data consisted of 6158 records, with 4461 related to the non-defaulting customers and 1696 related to defaulting customers.

The sample data set contains eight attributes: *income* (customer's income); *gender* (customer's gender); *mar_status* (customer's marital status); *dependents* (number of customer dependents); *residence* (customer residence type); *points* (customer's internal score value); *ext_credit_lim* (customer credit limit for external transactions); and *default* (the target class, classifying the customer as defaulting or non-defaulting).

41.2.1 Predictive Variables Used in Credit Scoring Systems from the International Literature

In the survey of the international specialized literature carried out by Hörkkö [7], involving 11 scientific articles, it was reported the predictive variables used for the development of theoretical and practical applications of credit scoring. Table 41.1 shows a tabulation by frequency of the variables identified within the research carried out.

In another survey of the international specialized literature, Delamaire [4] has elaborated a more in-depth research, involving 35 scientific articles, in which he has identified the predictive variables used for the development of credit scoring applications. Table 41.2 shows a frequency tabulation of the attributes identified in the investigation.

From this review, Delamaire [4] has concluded that the attributes used by researchers in credit scoring applications are different, depending upon the credit institution that provides the data.

However, socio demographic attributes such as income, age, marital status, type of housing, type of employment, number of dependents (children), or residence time at the current address are often mentioned.

Additional detailed banking information, electoral information, union membership information, nationality, and certain demographics and bank references are attributes that are not commonly used in credit scoring applications.

Table 41.1 List of variables discovered in the international specialized literature survey carried out by Hörkkö (2010), tabulated by frequency

Attribute name	#	%
Age	11	100
Income/change in income	9	82
Marital status	9	82
Residential status/housing	9	82
Occupation/type of employment	8	73
Loan size/credit limit	7	64
Current address/time in current address	6	55
Gender	6	55
Old loans/nr of other loans	6	55
Years of employment/time in present job	6	55
Zip code/region	5	45
Nr of children	4	36
Phone	4	36
Length of relationship	4	36
Maturity/duration of the loan	4	36
Credit card ownership	3	27
Education	3	27
Monthly expenses	3	27
Own resources/savings	3	27
Cosigner/guarantor	2	18
Credit type	2	18
Monthly payments	2	18
Score/points	2	18
Big city	1	9
Credit history	1	9
Foreign worker	1	9
Government assistance	1	9
Migrating out of state of birth	1	9
Nationality	1	9
Principal	1	9
Sector of employment	1	9
State of birth	1	9
Wealth	1	9
Working in private/public sector	1	9
Collateral type/value	1	9
Interest/interest rate	1	9
Loan to value ratio	1	9
Nr of payments	1	9
Payment performance	1	9

The review suggests that attributes such as age, for instance, are highly predictive.

Birth date has the advantage of being a fixed element and is generally a highly predictive attribute.

Table 41.2 List of variables discovered in the international specialized literature survey carried out by Delamaire (2012), tabulated by frequency

Attribute	#	%
Income	27	77
Age	26	74
Living status	24	69
Employment (title, class, place)	24	69
Time at present address	23	66
Marital status	22	63
Dependents-children number	20	57
Time with employer-previous	18	51
Bank accounts	16	46
Payments-outgoings	14	40
Sex	13	37
Telephone	13	37
Location	13	37
Debt	11	31
CC and other cards	11	31
CB information	11	31
Purpose of loan	9	26
Auto information	7	20
Wealth	7	20
Amount of loan	7	20
Education	6	17
Other loans 1	6	17
Spouse-family income	5	14
Race	5	14
Term of loan	5	14
Credit reference	4	11
Inquiries	4	11
Years at bank	4	11
Other reference	3	9
Insurance	3	9
Account opening	3	9
Bank reference	3	9
Age difference between man/wife	2	6
Location of relatives	2	6
Financial company reference	2	6
Electoral role	2	6
Trade union	2	6
Down payment	2	6
Account closing	2	6
Loan type	2	6
Nationality	2	6

It is possible to assume that the reason why certain attributes are recurrent in application forms is that they have a high explanatory power to identify defaulting customers.

Thus, for example, the 12 main variables surveyed (in terms of % of frequency) are often used in the development of credit granting systems, while some other attributes mentioned in these bibliographic reviews will be predictive or not, depending on the enterprise and also from the type of product for which the system was designed.

41.2.2 Predictive Variables Used in Credit Scoring Systems from the Brazilian Literature

It was carried out an investigation from the Brazilian specialized literature on the variables used in credit scoring applications, involving 36 articles, dissertations, and theses. The research results, describing the identified list of variables are presented in Table 41.3, in frequency tabulation mode.

The socio demographic variables are similar, with slight differences, from the revisions made. It is observed that the income and age variables appear in all articles. As already commented, these variables probably have a high predictive power. Other variables of this type also have intersections with the reviews carried out.

The so-called financial and behavioral variables, that contain information about customer's financial behavior, appear to be specific to the credit granting business and vary in application type.

In terms of number of variables to be used, in a research carried out by Aniceto [3], it was found that 53% of the surveyed articles used between 11 and 20 variables.

Finally, it has been observed that in the reviews, the socio demographic attributes are very similar, and attributes that characterize client's financial behavior are highly dependent on the business domain.

Behavioral variables provided by the Brazilian enterprise are presented in Table 41.4 that presents attributes, a small description, and the used aggregation function.

The provided sample data set correspond to a period of almost 2 years of financial activities.

However, for this investigation, it was considered an initial period of 13 months (10/2015 up to 10/2016). This

Table 41.3 List of variables resulting from the review of the Brazilian literature, tabulated by frequency

Attribute	#	%	Attribute	#	%
Customer age	32	89	Business net income	1	3
Gross Family Income	28	78	Total net income	1	3
Customer gender	27	75	Guarantor's gross revenue	1	3
Occupation	24	67	Guarantor's gross expense	1	3
Marital status	22	61	Guarantor's net income	1	3
Education	17	47	Last Loan Amount	1	3
Housing's type	13	36	Amount of the last loan amount	1	3
Time at present address	12	33	Nr of payments of last loan	1	3
Postal code	12	33	Indebtedness's Percentage	1	3
Financial dependents number	12	33	Nr of previous credits with the institution	1	3
Total amount of loan	12	33	Guarantor historic	1	3
Time in current address	11	31	Life insurance	1	3
Telephone	11	31	Health insurance	1	3
Nr parcels remaining	8	22	Credit card ownership	1	3
Commitment fee	8	22	Nr alienated goods	1	3
Length of relationship	7	19	Document's type	1	3
Equity situation	7	19	Investment	1	3
Credit reference	7	19	Salary	1	3
Payments value	6	17	First loan's acquisition	1	3
Wedding regime	5	14	Average card invoice	1	3
State	5	14	Nr years associated with card	1	3
Nationality	4	11	Average value of limits's excesses	1	3
Loan type	4	11	SELIC tax	1	3
Credit card quantity	4	11	SERASA register	1	3
Warranty type	4	11	Maximum delay	1	3
Age of spouse	3	8	Nr of paid parcels	1	3
Customer address	3	8	Monthly average balance	1	3
Business running time	3	8	Quarterly average balance	1	3
Average balance	3	8	Half-yearly average balance	1	3
Loans' term	3	8	CDB's balance	1	3
Nr of other loans	3	8	Investment funds balance	1	3
Financial constraints	3	8	Savings balance	1	3
Purpose of loan	3	8	Capitalization balance	1	3
Spouse income	2	6	Total reciprocity	1	3
Nature of business economic activity	2	6	Profitability	1	3
Neighborhood	2	6	Overdraft	1	3
City	2	6	Credit limit	1	3
Historic with financial institution	2	6	Alimony payment	1	3
Automotive insurance	2	6	Account balance	1	3
Residential insurance	2	6	Types of payments made	1	3
Returned check	2	6	Financial turnover	1	3
Amount accounts opening days	2	6	Utilization index	1	3
Payment's forms	2	6	Duration of the loan	1	3
Score	2	6	Contract percentage paid	1	3
Gross family expense	1	3	Payment's form (ticket or debit to account)	1	3
Familiar net income	1	3	Valor do Bem	1	3
Gross revenue	1	3	Interest rate	1	3
Expense gross	1	3	Invoice amount	1	3

Table 41.4 Additional behavioral attributes

Attribute	Description	Aggregate value
val-medio-encargo-12-ult-mes	Average value of charges in the last 12 months	Sum
qtd-bloqueio-ccred-band	Number of card locks from approval to reference date	Average
per-limite-utlz-ult-mes	Percentage use of limit in last month	Average
qtd-dia-maior-atrs-ult-3-mes	Maximum delay on card days in last 3 months	Average
per-financ-ccred-band-ult-mes	Percentage of funding in the last month	Average
qtd-ano-relc-cli	Time in years between the date of approval of the CDC and the reference date	Average
qtd-cont-ep-ult-12-mes	Number of contracts (personal loans) made in the last 12 months	Average
qtd-dia-maior-atrs-ccred-band	Maximum delay on card days, from approval to the reference date	Average
qtd-cpr-a-vista-ult-6-mes	Amount of sight purchases made in the last 6 months	Average
qtd-ftr-acima-lim-ult-6-mes	Amount of overlimit (invoice over the limit), made in the last 6 months	Average
num-idade-cliente	Age (in years) of the customer on the reference date	Average
qtd-ocor-spc-ult-24-mes-fech	Number of times the customer was denied or rehabilitated in the spc in the last 24 months	Sum
per-min-recb-fatura-ult-6-mes	Minimum percentage of invoice receipt in the last 6 months	Average
val-sld-dev-tot-ult-mes	Total debtor balance in the last month	Average
qtd-mes-atu-renda	Time in months between the reference date and the date of the last update of the rent	Average
qtd-pagto-acima-lim-ult-12-mes	Number of times the customer pays more than the minimum in the last 12 months	Sum
qtd-ocor-cobr-recd-ult-12-mes	Number of billing messages left for the customer in the last 12 months	Sum

sample data set has been stored in tables of a Data Base Management System (DBMS), in order to allow a more appropriate manipulation.

41.3 Experiments

The following experiments were performed using a different samples strategy for training and testing classifiers. The logistic regression classifier was chosen for these experiments, mainly because assessments have shown that it performs better, when compared to other classifiers.

One of the objectives of these experiments was to evaluate with past information, if it is possible to predict, with what performance and accuracy, future customer behaviors in terms of default patterns.

In order to implement these experiments, it was decided to use the Orange tool [5].

The created model is shown in Fig. 41.1, applying k-$fold$ cross-validation ($k = 10$), where the initial data set was randomly partitioned into k subsets (*folds*) $k_1, k_2, \ldots, k_k$ of mutually exclusive sizes of approximately equal size. Training and testing were performed k times, and for each iteration i, the subset D_i was used as test set, and the other subsets were used for training the model [6, 11].

41.3.1 The Experiment Number 1

For the execution of this experiment 1, table *bhs-band-2015-10-12-cadast-behav* contains training samples and table *bhs-band-2016-10-01-cadast-behav* contains test samples.

In this experiment, it was used logistic regression classifier and gain ratio metric, to define the most important attributes for defaulting predictions.

After the classifier execution, the result of the AUC-ROC metric and other obtained measures, is shown in Fig. 41.2. It is observed that the value of the ROC Curve (AUC-ROC) is 0.986.

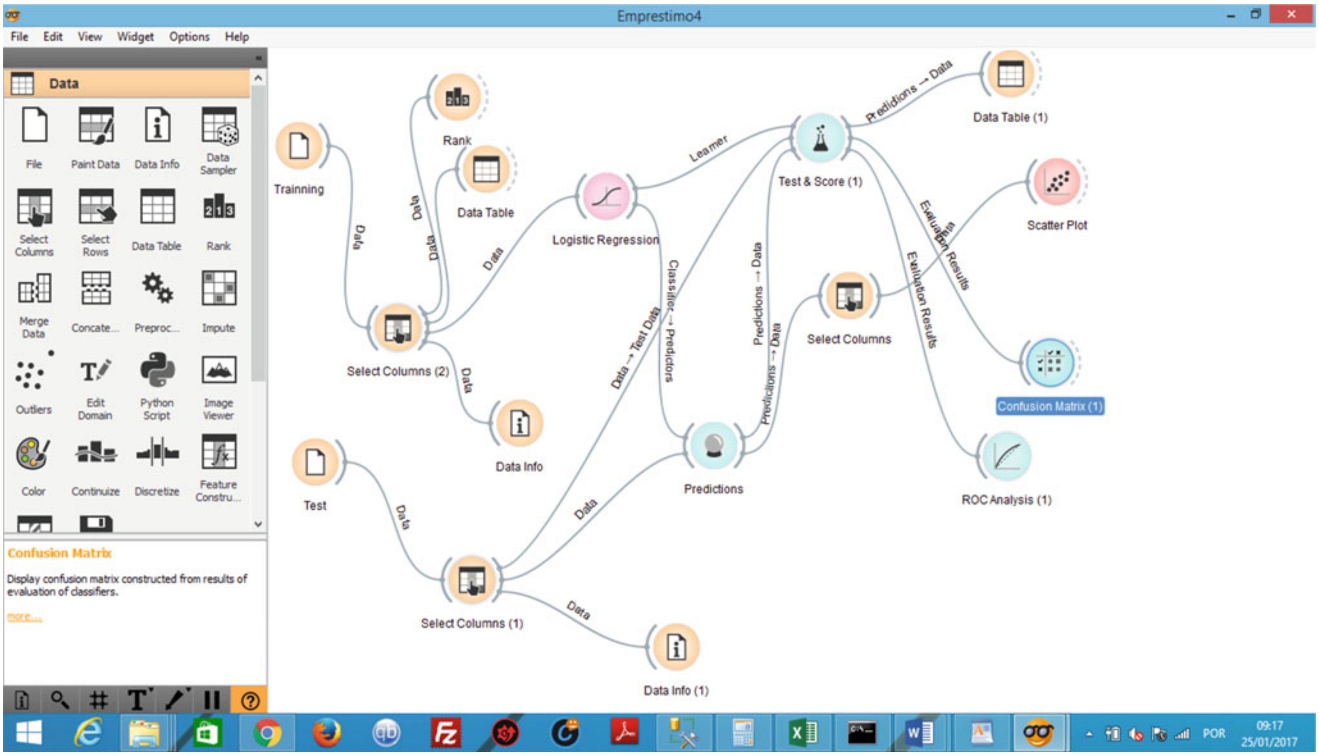
The confusion matrix obtained from this experiment has presented a percentage of false positives of 2.3% and of false negatives of 4.3%. As noticed, the classifier predicted 100 non-default customers as defaulting customers and predicted 36 defaulting customers as non-defaulting.

41.3.2 The Experiment Number 2

For the second experiment, the *bhs-band-2015-10-12-cadast-behav* and *bhs-band-adi-defaults-2016-not-2015* tables were used. The last table contains customers from the year 2016 data set who are not at the year 2015 data set. The metric gain ratio was used to define the most important attributes in prediction.

By executing the regressive model, the result of the AUC-ROC metric and other measures of classifier's performance used is shown in Fig. 41.3. The value of the ROC Curve (AUC-ROC) was equal to 0.987 in training.

As observed from the obtained confusion matrix, the classifier has predicted 20 non-defaulting customers as defaulting customers (about 3.0% of customers as false positives) and has predicted 5 defaulting customers as non-defaulting customers (about 2.8% of defaulting customers as false negatives). It is noticed that with this training, based on

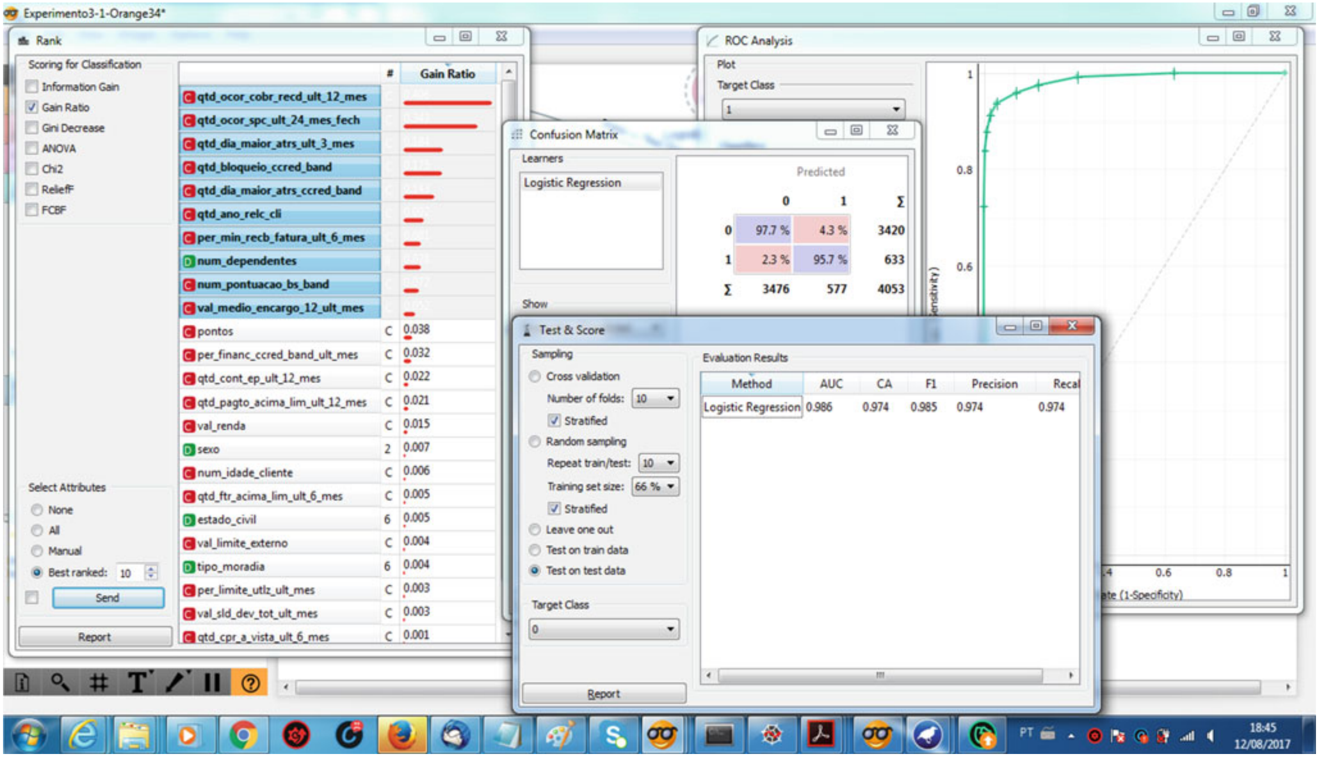


Confusion Matrix

Display confusion matrix constructed from results of evaluation of classifiers.

more...

Fig. 41.1 The created model in the orange tool



Scoring for Classification

- ☐ Information Gain
- ☑ Gain Ratio
- ☐ Gini Decrease
- ☐ ANOVA
- ☐ Chi2
- ☐ ReliefF
- ☐ FCBF

	#	Gain Ratio
C qtd_ocor_cobr_recd_ult_12_mes	C	[illegible]
C qtd_ocor_spc_ult_24_mes_fech	C	[illegible]
C qtd_dia_maior_atrs_ult_3_mes	C	[illegible]
C qtd_bloqueio_ccred_band	C	[illegible]
C qtd_dia_maior_atrs_ccred_band	C	[illegible]
C qtd_ano_relc_cli	C	[illegible]
C per_min_recb_fatura_ult_6_mes	C	[illegible]
D num_dependentes	[illegible]	[illegible]
C num_pontuacao_bs_band	C	[illegible]
C val_medio_encargo_12_ult_mes	C	[illegible]
C pontos	C	0.038
C per_financ_ccred_band_ult_mes	C	0.032
C qtd_cont_ep_ult_12_mes	C	0.022
C qtd_pagto_acima_lim_ult_12_mes	C	0.021
C val_renda	C	0.015
D sexo	2	0.007
C num_idade_cliente	C	0.006
C qtd_ftr_acima_lim_ult_6_mes	C	0.005
D estado_civil	6	0.005
C val_limite_externo	C	0.004
D tipo_moradia	6	0.004
C per_limite_utlz_ult_mes	C	0.003
C val_sld_dev_tot_ult_mes	C	0.003
C qtd_cpr_a_vista_ult_6_mes	C	0.001

Select Attributes

- ○ None
- ○ All
- ○ Manual
- ● Best ranked: 10

☐ Send

Report

Confusion Matrix

Learners: Logistic Regression

Show

	Predicted 0	1	Σ
0	97.7 %	4.3 %	3420
1	2.3 %	95.7 %	633
Σ	3476	577	4053

Test & Score

Sampling

- ○ Cross validation
  - Number of folds: 10
  - ☑ Stratified
- ○ Random sampling
  - Repeat train/test: 10
  - Training set size: 66 %
  - ☑ Stratified
- ○ Leave one out
- ○ Test on train data
- ● Test on test data

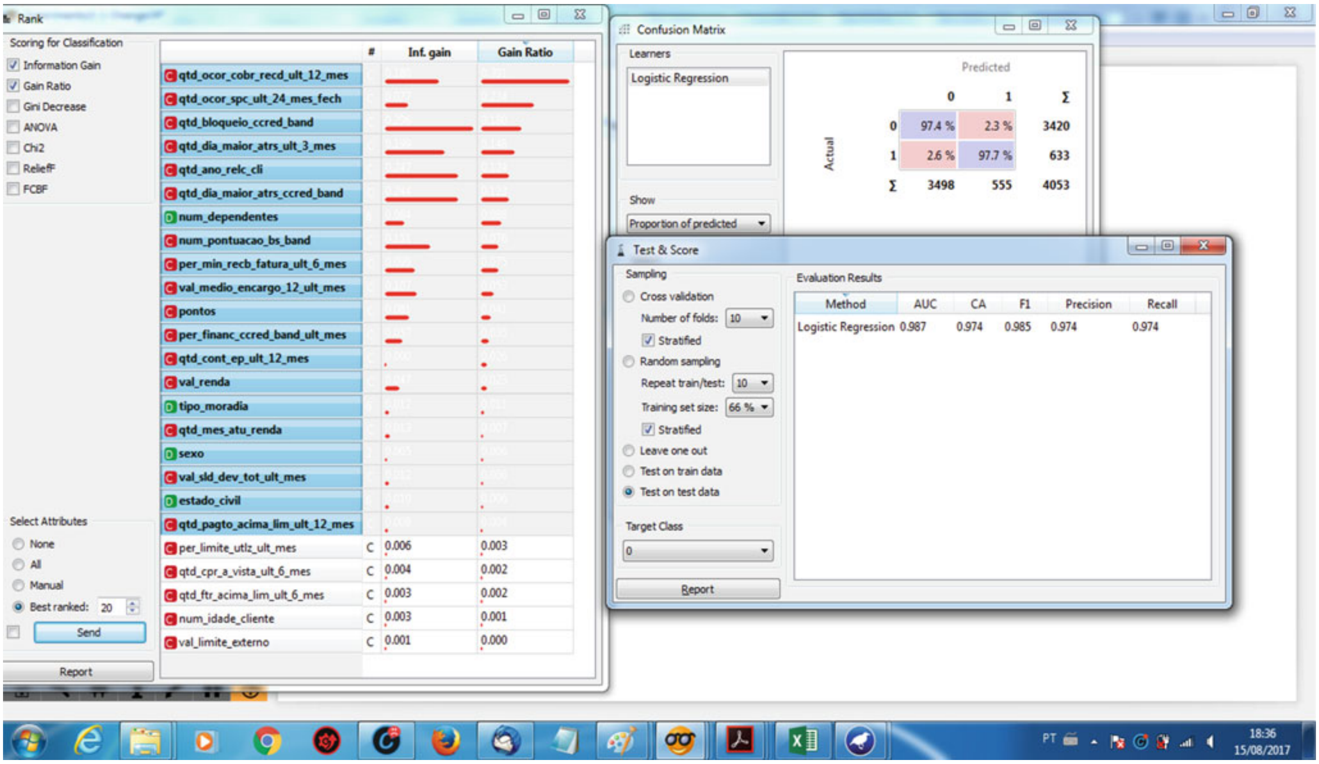
Target Class: 0

Report

Evaluation Results

Method	AUC	CA	F1	Precision	Recal
Logistic Regression	0.986	0.974	0.985	0.974	0.974

Fig. 41.2 Results of the experiment number 1



Rank

Scoring for Classification
☑ Information Gain
☑ Gain Ratio
☐ Gini Decrease
☐ ANOVA
☐ Chi2
☐ ReliefF
☐ FCBF

	#	Inf. gain	Gain Ratio
qtd_ocor_cobr_recd_ult_12_mes	[illegible]	[illegible]	[illegible]
qtd_ocor_spc_ult_24_mes_fech	[illegible]	[illegible]	[illegible]
qtd_bloqueio_ccred_band	[illegible]	[illegible]	[illegible]
qtd_dia_maior_atrs_ult_3_mes	[illegible]	[illegible]	[illegible]
qtd_ano_relc_cli	[illegible]	[illegible]	[illegible]
qtd_dia_maior_atrs_ccred_band	[illegible]	[illegible]	[illegible]
num_dependentes	[illegible]	[illegible]	[illegible]
num_pontuacao_bs_band	[illegible]	[illegible]	[illegible]
per_min_recb_fatura_ult_6_mes	[illegible]	[illegible]	[illegible]
val_medio_encargo_12_ult_mes	[illegible]	[illegible]	[illegible]
pontos	[illegible]	[illegible]	[illegible]
per_financ_ccred_band_ult_mes	[illegible]	[illegible]	[illegible]
qtd_cont_ep_ult_12_mes	[illegible]	[illegible]	[illegible]
val_renda	[illegible]	[illegible]	[illegible]
tipo_moradia	[illegible]	[illegible]	[illegible]
qtd_mes_atu_renda	[illegible]	[illegible]	[illegible]
sexo	[illegible]	[illegible]	[illegible]
val_sld_dev_tot_ult_mes	[illegible]	[illegible]	[illegible]
estado_civil	[illegible]	[illegible]	[illegible]
qtd_pagto_acima_lim_ult_12_mes	[illegible]	[illegible]	[illegible]
per_limite_utlz_ult_mes	C	0.006	0.003
qtd_cpr_a_vista_ult_6_mes	C	0.004	0.002
qtd_ftr_acima_lim_ult_6_mes	C	0.003	0.002
num_idade_cliente	C	0.003	0.001
val_limite_externo	C	0.001	0.000

Select Attributes
○ None
○ All
○ Manual
● Best ranked: 20
☐ Send

Report

Confusion Matrix

Learners
Logistic Regression

Show
Proportion of predicted

		Predicted		
		0	1	Σ
Actual	0	97.4 %	2.3 %	3420
	1	2.6 %	97.7 %	633
	Σ	3498	555	4053

Test & Score

Sampling
○ Cross validation
Number of folds: 10
☑ Stratified
○ Random sampling
Repeat train/test: 10
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
● Test on test data

Target Class
0

Report

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Logistic Regression	0.987	0.974	0.985	0.974	0.974

Fig. 41.3 Results of the experiment number 2

data from the year 2015, predictions were made about the defaulting customer in data from 2016, for customers who were not at the data set from the 2015 year.

41.3.3 The Experiment Number 3

For this experiment, two data sets were generated: one containing attributes characterized as demographics attributes and other containing attributes considered as financial/behavioral attributes. The performed analysis consisted of the evaluation of the performance of certain classifiers in the two data sets, in order to measure the improvement of the prediction with the use of these data sets.

From the specific case of the first data set, named *B1-Cadast*, which contains only demographics data, the following classifiers were used: Logistic Regression (LR), *k*-Nearest Neighbors (*k*NN), Decision Trees (DT), and Support Vector Machine (SVM). The Orange tool was used to perform the evaluation.

The assessment consisted of submitting the classifiers for several executions and, in each of them, a certain attribute was removed, based upon its value of the gain ratio metric (the attributes of smaller values were first removed).

Figure 41.4 presents the evaluation of the four classifiers mentioned, used in the *B1-Cadast* data set. It was observed that the LR classifier has obtained the best performance in terms of AUC-ROC values, and the kNN classifier was the second best result in the evaluation. The DT and SVM classifiers obtained the worst results.

The next evaluation consisted of submitting the same classifiers to a new data set, called *B1-Cadast-Behav*, containing demographic data, together with the behavioral/financial data.

The four classifiers previously mentioned were used in this data set and Fig. 41.5 presents the evaluation of the classifiers. It is again observed that the LR classifier was the one that obtained the best performance, in terms of AUC-ROC values, and the DT classifier obtained the second best result in this evaluation. The *k*NN and SVM classifiers had the worst results.

It is observed that the junction of the two data sets, *B1-Cadast* and *B1-Cadast-Behav* significantly increases the predictive capacity of the LR classifier model, especially when only 19 variables are used, with an AUC-ROC value equal to 0.997.

Some observations obtained from these three experiments are necessary to state here:

- The use of behavioral variables significantly improves the prediction of defaulting customers, and classifiers get AUC-ROC values approximate from what was obtained in the specialized literature; and

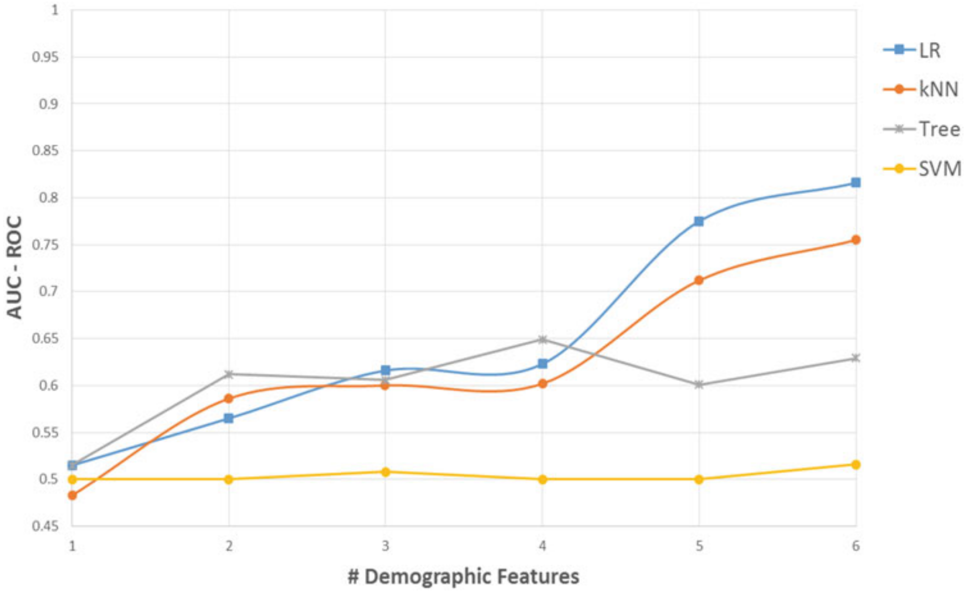


Fig. 41.4 The evaluation of classifier's results with demographic and behavioral data

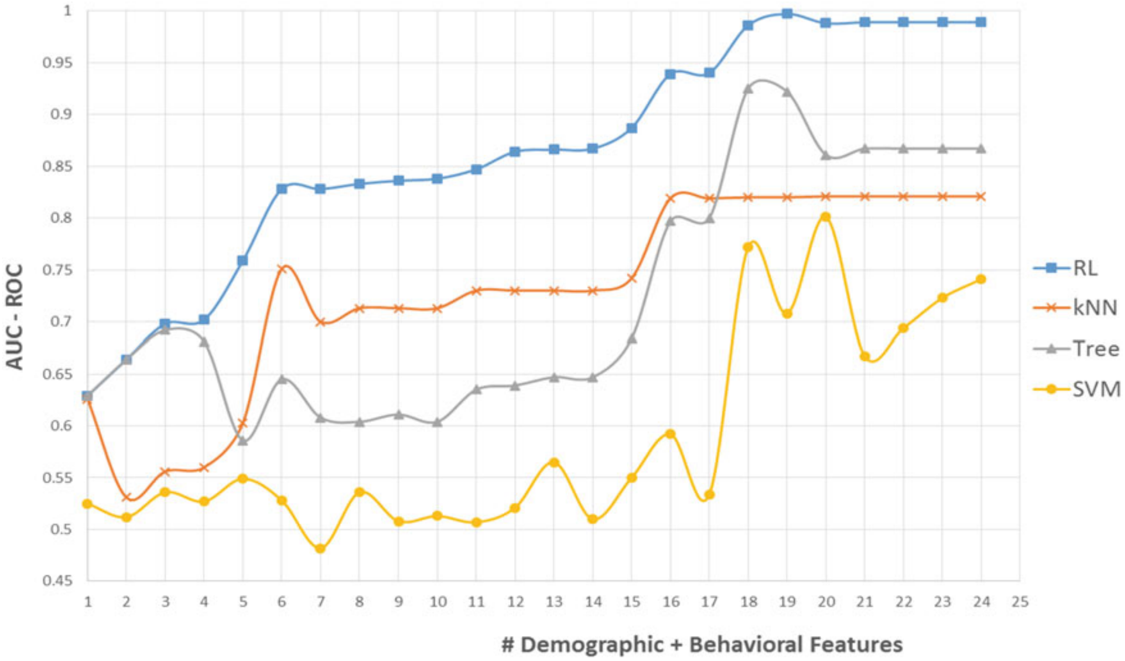


Fig. 41.5 Evaluation of classifier's results with demographic and behavioral data

- the data provided by the retail enterprise have allowed to confirm, based on performed tests, that high quality data (demographic and behavioral) were provided, which allows a significant accuracy in the classifier used.

41.4 Conclusion

This investigation was carried out on aspects inherent to data extraction from the development of the credit scoring system prototype. Initial demographic data used were not adequate for defaulting predictions.

In order to improve accuracy on prediction, additional behavioral/finance data was investigated in specialized literature, to obtain also additional features to improve credit scoring. Real data was used and provided by a Brazilian retailer enterprise.

The Logistic Regression (LR) classifier was used and the results have shown that additional data have improved the classifier performance.

One area for further work is to use different classifiers and analyze their performance. Other area is getting more behavioral/finance attributes.

Acknowledgements The authors would like to thank: the Brazilian Aeronautics Institute of Technology (ITA); the Casimiro Montenegro Filho Foundation (FCMF); the Software Engineering Research Group (GPES) members; and the 2RP Net Enterprise for their infrastructure, assistance, advice, data set, and financial support for this work.

References

1. J. Abellán, G. Castellano, A comparative study on base classifiers in ensemble methods for credit scoring. Expert Syst. Appl. **73**, 1–10 (2017)
2. M. Ala'raj, M.F. Abbod, A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Syst. Appl. **64**, 36–55 (2016)
3. M.C. Aniceto, Estudo comparativo entre técnicas de aprendizado de máquina para estimação de risco de crédito. Dissertação (Mestrado em Administração). Universidade de Brasília, Brasília, 2016
4. L. Delamaire, Implementing a credit risk management system based on innovative scoring techniques, Ph.D. thesis, University of Birmingham, 2012
5. J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik, B. Zupan, Orange: data mining toolbox in python. J. Mach. Learn. Res. **14**, 2349–2353 (2013)
6. J. Han, M. Kamber, J. Pei, *Data Mining - Concepts and Techniques*, 3rd edn. (Morgan Kaufmann, Amsterdam, 2012)
7. M. Hörkkö, The determinants of default in consumer credit market. Masters thesis, Aalto University School of Economics (2010). Retrived from http://epub.lib.aalto.fi/en/ethesis/pdf/12299/hse_ethesis_12299.pdf
8. M.B. Pascual, A.M. Martínez, A.M. Alamillos, Redes bayesianas aplicadas a problemas de credit scoring. Una aplicación práctica. Cuadernos de Economía **37**(104), 73–86 (2014)
9. R.M. Stein, The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. J. Bank. Finance **29**, 1213–1236 (2005)
10. B. Waad, B.M. Ghazi, L. Mohamed, A three-stage feature selection using quadratic programming for credit scoring. Appl. Artif. Intell. Int. J. **27**, 8 (2013)
11. I. Witten, E. Frank, *Data Mining Practical Machine Learning Tools and Techniques*, 2nd edn. (Elsevier, Amsterdam, 2005)