2016 15th IEEE International Conference on Machine Learning and Applications

Predicting Recovery of Credit Operations on a Brazilian Bank.

Rogério G. Lopes*, Rommel N. Carvalho*†, Marcelo Ladeira* and Ricardo S. Carvalho†

*Department of Computer Science (CIC)
University of Brasilia (UnB), Brasilia, DF, Brazil
Email: rglopes@gmail.com, {mladeira,rommelnc}@unb.br

†Department of Research and Strategic Information (DIE)
Ministry of Transparency, Monitoring and Control (MTFC), Brasilia, DF, Brazil
Email: {rommel.carvalho,ricardo.carvalho}@cgu.gov.br

***Abstract*—This article presents a study conducted in a Brazilian bank, in order to assist the institution account managers in the approach to customers with loans in arrears. This approach is carried out to propose alternatives to customers return to timely payments situation, but the efficiency of this approach is small, accounting for only about 6.8% of customers. A predictive model, using classification was used to help identify customers with the most potential to return to a normal situation, reaching a 85.5% accuracy rate with the winning algorithm, Gradient Boosting Method. It was implemented in the integrated Platform H2O with R language, exploring the grid mode and parallel processing models advantages.**

I. Introduction

Since January 2015[1], we have seen a drop in credit supply and rising defaults in Brazil, resulting from decreases in economic activities and investor confidence. According to Brazil's Central Bank (BCB), credit operations for individuals are those that have shown the highest growth of default, from 5.1% in December 2014 to 6.7% in April 2016, a proportional increase of 31%. Figure 1 illustrates the gradual growth occurred in this period.

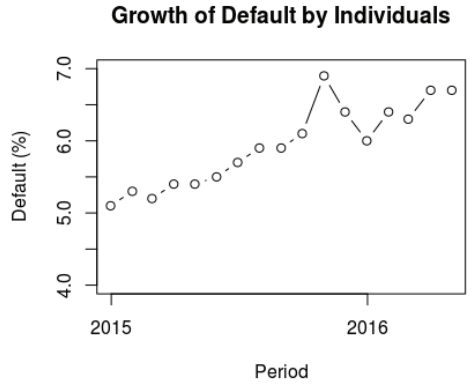

Fig. 1. Growth of Default by Individuals. Adapted from [5].

Related to this issue, there are several studies that qualify customers due to the credit risk they represent, separating them into groups of good or bad payers. However, once the bad debt occurs, there has been little research on classifying the possibility of these bad payers becoming good payers again [5].

This study was conducted on a bank with one of the highest loan portfolio of the Brazilian National Financial System (SFN). Although having a lower default rate than the numbers SFN described above, this bank also showed an increase in the event of default on individual credit operations.

Credit analysis are used for the granting of credit, using models that minimize the occurrence of default in operations. But once we found the default, the traditional action is to deny new lines of credit to these borrowers.

Due to the increase in observed default, there has been intensified actions to offer to this customers alternatives to return to normal condition of operation. The contacts were made by their account managers.

In the first months of using this approach, a major operational difficulty was detected: the small amount of customers that an account manager could contact and offer his service. In addition, it was observed that a significant part of customers did not have conditions to renegotiate his/hers operations, thus making inefficient performance of the account manager.

It is also important to note that several customers could have returned to the non-default condition, but they did not know the alternatives that were available to them.

So, this study was designed to support the account managers of this bank, in order to inform them a list of customers that are most likely to recover from default.

Therefore, the main objective of this study was to apply data mining techniques to predict which credit operations could return to a non-default situation. Models were developed using Generalized Linear Models (GLM), Gradient Boosted Methods (GBM), and Distributed Random Forest (DRF) and then compared. This comparison was made using the AUC and the PCC indicators, which will be explained on section III.

978-1-5090-6167-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ICMLA.2016.139

The models were developed using the R language and H2O platform of data mining, considering its parallel processing capabilities. Further details on section III. [1]

This paper is organized as follows: Section II presents the credit scoring state of the art. Section III presents the methodology used in this study. Section IV presents the results given, the models learned as well their evaluating. Section V presents the conclusion and future work.

II. State of the Art

The default numbers observed in Brazil, from December 2014 to April 2016, indicate that financial institutions need a tool to support their credit granting decisions. Credit scoring models are an estimate based on the likelihood that a customer will present some undesirable behavior in the future.

Lessmann et al. [6] in a paper published in 2015, conducted a study evaluating 41 publications on the award of credit since 2006, all of them using classifiers to categorize customers as good or bad payers. These works were organized into three categories of classifiers: Individuals; homogeneous ensemble; and heterogeneous ensemble classifiers.

Six databases were used to verify the performance of each of the 41 models proposed, evaluating them from the standpoint of 6 indicators: area under the receiver operating curve (AUC), percentage correctly classified (PCC), partial Gini index, H-measure, Brier Score (BS) and Kolmogorov-Smirnov (KS).

AUC represents how well classified your data were, regardless to its distribution or misclassification costs.[8]. The PCC is an overall accuracy measure, it indicates the percentage of outcomes that were correctly classified [4][2]

At the end of this comparison, it was shown a ranking indicating that the ensemble heterogeneous algorithms had presented better overall performance. However, the difference in performance between the three categories proved to be very small. Considering the simplicity of the others algorithms, these could be used and they would provide similar results than those provided by more complex algorithms.

The detailed table I presents the results of the benchmark, bringing new performance references and recognition of new algorithms. As seen in this table, the HCES-Bag algorithm obtained the highest AUC result, while the AVG W and GASEN algorithms achieved 80.7% of PCC.

TABLE I
State of Art - Models Comparison

	Algorithm	AUC	PCC
Heterogeneous Ensemble	HCES-Bag	0.932	80.2
	AVG W	0.931	80.7
	GASEN	0.931	80.7
Homogeneous Ensemble	RF	0.931	78.9
	BagNN	0.927	80.2
	Boost	0.93	77.2
Individual	LR	0.931	70.84
	LDA	0.929	78.4
	SVM-Rbf	0.925	79.9

[1]H2O is an open source machine learning platform, available at www.h2o.ai
[2]For more details on the others indicators,[6]

III. Methodology

This study followed, where applicable, the phases of the data mining process proposed by the Cross Industry Standard Process for Data Mining (CRISP-DM): Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation.[3]

The Business Understanding phase was already covered in the introduction section I. The Deployment phase, the last one of the CRISP-DM, was not performed because this study is still in its initial state and, as later demonstrated in this paper, will still be refined before being applied to the institution's credit recovery process. The following subsections will present the others phases of the CRISP-DM.

A. Data Understanding

For this study, three data bases were used. The database (i) contained a sample of 22,764 transactions that were late in February 2016, containing variables related to the data of the contracting of credit operations, such as date of hire, time of the operation mode of the credit operation , operation risk, contracted amount, outstanding balance and amount of days in arrears, totaling 38 variables. The second database (ii) contained the status of every existing transaction in the database (i) and also the number of days with overdue operation verified in March 2016. The database (iii) was composed of 158 variables with demographic and financial information of all the clients listed in the database (i), also obtained in February 2016.

All three databases were extracted from the Data Warehouse (DW) of the institution, containing integrated and validated data without missing values and with data integrity assurance.

The variables were not individually identified in this study because the bank considered them confidential. Hence, only their categories were mentioned.

B. Data Preparation

As the data sources are from the DW, data preparation activity was reduced to developing a single database containing the 196 variables, resulting from the joint database (i), (ii) and (iii), and by creating new variables. These new variables were transformed with the following characteristics:

- Composed primary keys: have been identified and each of these keys has been transformed into a single variable of type factor.
- Date type fields: the ones in which the interest was in the period of time between the date the event occurred and the time being of this study (February 2016) were transformed into numerical variables, representing that period time.
- Target variable: a binomial categorical variable was created, containing the value 1 for the operations that reduced the delay and 2 for those that have maintained or increased the number of days overdue, comparing databases (ii) and (i).

C. Modeling

Three models were developed to be compared and then one was chosen as the best model to be used in generating support information to account managers. At first, the models were being processed in the software R. However, considering the large amount of variables (196), processing the data was taking too long and that was compromising the efficiency of the study. So, the platform H2O was used, integrated with R, to explore the grid mode and parallel processing models. The grid allowed to combine different parameters to build different models. The parallel processing allowed those models to be built at the same time. This efficiency increase made it possible to build more models, with better tuning parameters.

By using the R language, integrated with H2O platform, the following algorithms were used:

- GLM - Generalized linear Modeling
- GBM - Gradient Boosting Method
- DRF - Distributed Random Forest

Generalized Linear Models are similar to linear regression, only it's more flexible for it doesn't requires normal distribution to the errors. It estimates models for outcomes from the Exponential family and it is used for both regression and classification. fits really well to large data sets. GLM fits really well to large data sets and is very popular because of its easy interpretation and its speed, even when used for data sets with great number of columns. [7]

The Gradient Boosting Machine is used for predictive results for regression or classification. It is an ensemble of tree models and provides considerably accurate results. GBM applies weak classification algorithms to incrementally changing data, creating that way a series of decision trees. It is robust, not implying any distribution to the data, and because of that it is considered one of the best choices for many users for it requires little adjustments. [7]

Distributed Random Forest, just as GBM, is an ensemble of tree models, which each tree is de-correlated from all other trees.[7]

D. Evaluation

The evaluation of the models was performed using the two indicators analysis: Area Under Roc (AUC) and Percentage True Correctly Classified (PTCC). PTCC is a adaptation of PCC indicator (II), considering only the percentage of true positives classified.

IV. Results

In this section, we will present the results in each phase carried out during the study, as mentioned in Section III.

A. Data Preparation

Using the feature engineering technique has identified the need to create three new variables were performed and the following steps to generate the database that was used for modeling:

- Database (i) - A new variable was created to replace two variables representing the mode of the original contract for the operation, which was represented by a composed primary key in the operational system of the institution. Furthermore, the variable representing the date of the contract was transformed into a variable to indicate the lifetime of the contract.
- Database (iii) - A new variable was created to replace two variables that together represented the profession pursued by the individual customer contracting the operation.
- Target Variable - This variable was created to indicate that the operation had returned to normal after 30 days. This variable was called Delay Reduction Index (DRI), containing the value 1 to indicate that there was a reduction in delay and 0, to indicate that the delay had increased or remained the same.
- Final Database - Carried the junction of databases (i), (iii) and the target variable, with 22,764 records and 199 variables.
- The database was partitioned into 2 parts, one for training and the other for validation at a ratio of 80:20. Table II shows the composition of each partition.
- The initial analysis of the data, it showed that out of the total of 22,764 operations that had late payments, only 1,548 returned to a regular situation, which represents only 6.8% of the operations. That required a special attention in development of the predictive model, considering it features a case of imbalanced classes. Brown [2] has demonstrated that GBM and DRF perform well even in those cases. On the other hand, Verbeke [8] affirms that classification techniques perform best with balanced classes, over/under-sampling the data. It was decided to under-sample the training partition of the data, not altering the validation partition.

TABLE II
PARTITIONING THE DATABASE

Partition	DRI=1	DRI=2	Rows
Train	1.232	17.013	18.245
Test	316	4.203	4.519
Total	1.548	21.216	22.764

B. Modeling

Three predictive models were developed using the H2O platform accessed through the R language, using the algorithms GLM, GBM, and DRF. The results are detailed in the following sections.

1) GLM - Generalized Linear Modeling: This algorithm was implemented with 10-folds cross validation. The use of this validation technique was reproduced in other algorithms used.

The GLM algorithm obtained AUC = 0.956881 with 10-fold cross-validation. The PTCC obtained by analyzing the successes of the class of interest (DRI = 1) reached 63.63%, as shown in Figure 2 and Table III.

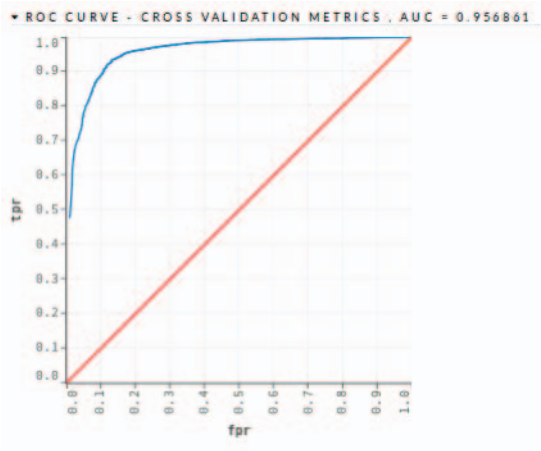

Fig. 2. GLM - AUC

TABLE III
GLM - CONFUSION MATRIX

	1	2	Error	Rate	PTCC
1	784	448	0.363636	448/1232	63.63%
2	284	16729	0.016693	284/17013	
Totals	1068	17177	0.040121	732/18245	

C. GBM - Gradient Boosting Method

The first results showed that GBM algorithms had a much performance than the others. Then, a grid of parameters was chosen and applied to it. The parameters used were:

- maximum trees: 100, 500 and 1.000.
- maximum depth: 5, 7 and 10.
- stopping tolerance: 0.001

The best model was built with 10-folds cross validation, maximum 500 trees and 7 maximum depth of them.

The GBM algorithm obtained AUC = 0.982650 with 10-fold cross-validation and PTCC for the class of interest reached 84.65%, as shown in Figure 3 and Table IV.

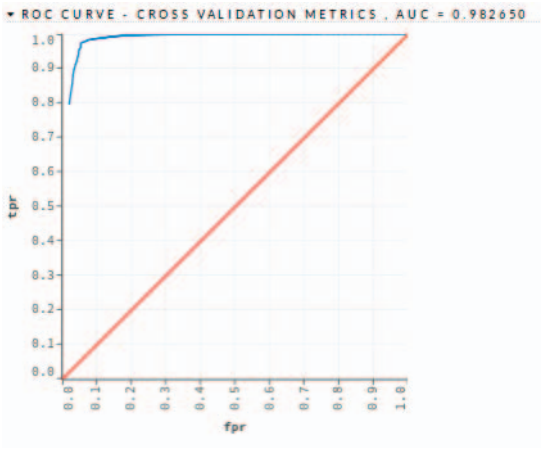

Fig. 3. GBM - AUC

TABLE IV
GBM - CONFUSION MATRIX

	1	2	Error	Rate	PTCC
1	1043	189	0.153409	189/1232	84.65%
2	101	16912	0.005937	101/17013	
Totals	1144	17101	0.015895	290/18245	

D. DRF - Distributed Random Forest

This algorithm was implemented with 10-folds cross validation, maximum 500 trees and 7 maximum depth of them.

The GBM algorithm obtained AUC = 0.978096 with 10-fold cross-validation and PTCC for the class of interest reached 61.60%, as shown in Figure 4 and Table V.

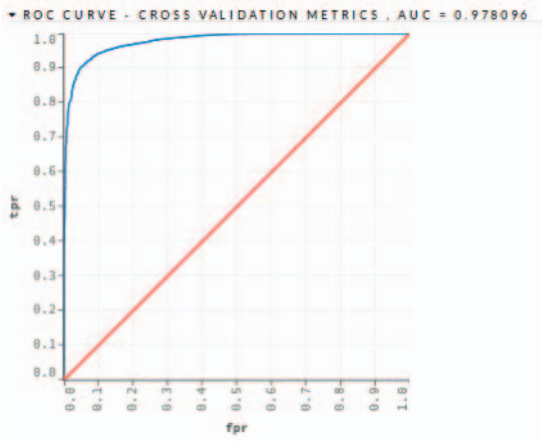

Fig. 4. DRF - AUC

TABLE V
DRF - CONFUSION MATRIX

	1	2	Error	Rate	PTCC
1	759	473	0.383929	473/1232	61.60%
2	123	16890	0.007230	123/17013	
Totals	882	17363	0.032666	596/18245	

E. Comparison of Results

All three models generated had a high evaluation result of the ROC curve. However, considering the classes were imbalanced, we cannot use this indicator alone. The large number of hits from the majority class, which is not the class of interest, leads to biased results.

Therefore, it is necessary to evaluate the second indicator, the PTCC, pointing accuracy level in the class of interest. In assessing this aspect, it is possible to realize a greater variation between the three models, ranging from 61% to 84%, with the best result obtained by the GBM algorithm. Table VI shows the comparison of the results obtained by the three models, where it is clear that the GBM algorithm achieved better performance in all metrics used.

TABLE VI
COMPARING MODELS

Algorithm	AUC	PTCC (%)
GLM	0.956881	63.63
GBM	**0.983650**	**84.65**
DRF	0.978096	61.60

V. CONCLUSION

The aim of this study was to identify those delinquent clients that possessed the highest probability of short-term recovery, to support the activities of Account Managers, increasing the efficiency of their approach with customers.

Although it was inspired by the study of the state of the art of credit score models, if failures occur in the classification, there is no financial penalty, since the credit operations are already late. In theory, the errors only decrease the performance of Account Managers.

In comparison between the algorithms, the GBM showed a better performance in both indicators calculated, making the candidate to be chosen for implementation and integration with bank's operating systems.

The percentage of customers who can return to timely payments proved to be close to 6.8%. By using the proposed predictive model, the account managers may increase the efficiency of the approaches made to customers, considering the PTCC index of 85.5%, bringing gains to the credit recovery activity.

The performance indicators obtained in this study cannot be directly compared with Lessman [6], since different databases were used in each study. However, because they have close performances in AUC and PCC indicators, we can say that the model is suitable for use in the bank.

A. Future Works

This was just an initial study, which can be further enhanced with the use of other predictive modeling techniques, such as using ensemble learning, both homogeneous and heterogeneous.

In addition, there is a chance that the behavior of lower default rates have a seasonal behavior, over a year, due to situations that occur with expenses that are held on fixed dates a year, such as taxes, school fees and also seasonal variations in income receipts. So, it is indicated the continuity of this work developing predictive models for different times over a year.

REFERENCES

[1] Poltica Monetria e Operaes de Crdito do SFN.

[2] Iain Brown and Christophe Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446 – 3453, 2012.

[3] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[4] Karel Dejaeger, Frank Goethals, Antonio Giangreco, Lapo Mola, and Bart Baesens. Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, 218(2):548 – 562, 2012.

[5] Sung Ho Ha. Behavioral assessment of recoverable credit of retailer's customers. *Inf. Sci.*, 180(19):3703–3717, October 2010.

[6] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[7] The H2O.ai team. *h2o: R Interface for H2O*, 2015. R package version 3.1.0.99999.

[8] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211 – 229, 2012.