ISSN 1980-5411 (On-line version)
Production, 28, e20170110, 2018
DOI: 10.1590/0103-6513.20170110

PRODUCTION
PRODUÇÃO

Application of bayesian additive regression trees in the development of credit scoring models in Brazil

Daniel Alves de Brito Filho[a], **Rinaldo Artes**[a*]
[a]Insper, São Paulo, SP, Brasil
*rinaldoa@insper.edu.br

Abstract

Paper aims: This paper presents a comparison of the performances of the Bayesian additive regression trees (BART), Random Forest (RF) and the logistic regression model (LRM) for the development of credit scoring models.

Originality: It is not usual the use of BART methodology for the analysis of credit scoring data. The database was provided by Serasa-Experian with information regarding direct retail consumer credit operations. The use of credit bureau variables is not usual in academic papers.

Research method: Several models were adjusted and their performances were compared by using regular methods.

Main findings: The analysis confirms the superiority of the BART model over the LRM for the analyzed data. RF was superior to LRM only for the balanced sample. The best-adjusted BART model was superior to RF.

Implications for theory and practice: The paper suggests that the use of BART or RF may bring better results for credit scoring modelling.

Keywords

Credit. Machine learning. Logistic regression. BART. Random Forest.

How to cite this article: Brito Filho, D. A., & Artes, R. (2018). Application of bayesian additive regression trees in the development of credit scoring models in Brazil, *Production*, 28, e20170110, https://doi.org/10.1590/0103-6513.20170110

Received: Jan. 4, 2018; Accepted: May 4, 2018.

1. Introduction

Credit analysis is a key activity for retail banks. Credit scoring models have become an important tool in credit analysis due to the need for standardization and agility in decision making, and there are now situations in which credit approval or refusal is fully automated. More accurately identifying customers with high probability of non-compliance enables reducing costs or increasing revenue.

Since 2004, with the Basel II agreement (Bank for International Settlements, 2004, 2006), banks have been encouraged to improve their internal credit risk models to obtain the authorization to use them as a basis for capital allocation adjusted to that risk. In order to obtain the approval for the use of advanced credit scoring models, known as the Advanced Internal Rating Approach (A-IRB), banks need to demonstrate their ability to accurately assess their risk. Banks with an A-IRB certification have competitive advantages over other banks because they are authorized by regulators to allocate less capital to credit risk.

Some machine learning techniques have been used in several areas of knowledge, such as medicine, biology and genetics, mainly in classification problems. These techniques can be applied in credit analysis to classify a borrower as a "good" or "bad" payer. As can be observed in Kruppa et al. (2013) and Lessmann et al. (2015), these models present superior results compared with ordinary logistic regression models.

Using a real database from a Brazilian credit bureau provided by Serasa Experian, this paper tests the effectiveness of two machine learning models. The first technique is called Bayesian additive regression trees (BART). In this method, several classification trees are combined to increase the predictive power of the analysis. The second technique is called random forest; this method also combines classification trees to increase the predictive power of the analysis.

The credit scoring models found in the literature are usually applied to databases provided by private companies. After a bibliographic review, authors did not find applications that have used a database obtained directly from a credit bureau. These variables describe the credit history of a customer including his credit experience with different firms, what is unusual in credit scoring models developed in financial institutions. In addition, also after the bibliographic review, only one application of the BART model for credit scoring was found: Zhang & Härdle (2010) applied the model on a database of German companies ("German Creditreform database").

The most widely used technique for the construction of credit scoring models is the logistic regression (Thomas, 2009). Because it is the most popular technique, it is used as a benchmark for the performance evaluation of the credit scoring models developed in this paper. Similarly to Zhang & Härdle (2010), we compared the performance of BART, random forest and logistic regression to predict default by using credit scoring model.

In general, databases used for the development of credit scoring models are composed of a larger number of "good" paying customers than "bad" payers, also called unbalanced data set. The effect of this imbalance on the performance of the fitted models is also evaluated.

In short, this paper aims to answer two different questions:

1) Is there any evidence that the use of machine learning technique such as BART may improve the identification of bad payers in a credit scoring context?

2) Does credit bureau specific variables are worthy to construct a credit scoring model?

This paper is divided into six sections. After this introduction, Section 2 presents a bibliographic review of the development history of the credit scoring models, their evolution and the main techniques used in credit scoring models, including models based on machine learning – such as BART and Random Forest. Section 3 presents main techniques used in this paper for the development and comparison of the performance of credit scoring models. Section 4 describes the database used in developing the models. The development of the models and their results are presented in Section 5. Finally, Section 6 presents the main conclusions drawn from the application of machine learning techniques and some suggestions for future work.

2. Bibliographic review

Credit analysis can be considered as a problem of rating customers as bad or good payers. Its main purpose is to predict whether an individual will become a bad payer within a given time horizon.

One of the first methods developed to discriminate groups was proposed in Fisher's original work (Fisher, 1936), in which principles of discriminant analysis were developed. Durand (1941) applied this methodology in finance to distinguish between good and bad loan payers. Altman (1968) used the method for predicting corporate bankruptcy.

The first applications of the credit scoring model were focused on granting credit to consumers and companies (Sousa et al., 2016). With the growth of the credit card segment, it was demanded that the credit granting activity should be automated, which was possible thanks to the growth in computational power.

The main benefits of credit scoring is optimized processing time for credit proposals, with the consequence of greater agility in decision making; minimizing costs and efforts of the credit process; and reducing mistakes (Chandler & Coffman, 1979).

Abdou & Pointon (2011) noted that the information collection is a critical problem in the construction of a credit scoring models; normally, information about individuals' characteristics is used. They say that sometimes, economic factors are not considered. In general, models are not standardized and differ among markets because the selection of the model's variables is not theoretical driven; that is, the selection of variables is, in general, based on statistical procedures. Sample size is also a well-discussed issue: in some applications, for retail credit, samples smaller than 1100 observations are used.

The logistic regression (Hosmer et al., 2013) has become the most widely used method in the financial industry (Anderson, 2007). The use of artificial intelligence techniques, imported from statistical learning theory, such as classification trees (Breiman et al., 1984) and neural networks (Desai et al., 1996; Malhotra & Malhotra, 2002) has become increasingly common in credit scoring systems. Statistical learning methods have received great attention in the past decade in finance-related research, for credit scoring and bankruptcy prediction (Li et al., 2006), bankruptcy classification (Lensberg et al., 2006), stress analysis (Gestel et al., 2006) and application for financing decisions and return (West et al., 2005; Xia et al., 2000). In addition, regression techniques (Lee & Chen, 2005) and clustering techniques (Wei et al., 2014) have also been adapted for the credit scoring problem.

The choice of learning algorithms depends on the available methods and the user's preference (Jain et al., 2000). As an alternative to using a single method, there is an increasing tendency to use hybrid systems (Hsieh,

2005), such as using grouping techniques to separate and isolate non-representative samples and then neural networks for modeling credit scoring. New concepts of adaptation to change (Pavlidis et al., 2012) and dynamic modeling (Crook & Bellotti, 2010) are beginning to be explored in credit risk analysis.

Standardized classification tools include linear and quadratic discriminant analysis, in addition to the logistic model. On the other hand, other techniques, such as models based on machine learning, have been applied in classification models. As can be observed in Kruppa et al. (2013) and Lessmann et al. (2015), these models can yield results superior to those of logistic regression models.

General literature reviews of issues related to credit scoring modeling may be found in Thomas et al. (2005), Abdou & Pointon (2011), Lessmann et al. (2015), Louzada et al. (2016) and Für et al. (2017), for instance.

2.1. Tree-based models

The classification regression tree (CART) technique was proposed by Breiman et al. (1984). It consists of constructing a classification tree based on a binary subdivision of the sample into sub trees. First, an independent variable that best segregates the sample between, for example, good and bad payers, according to some quality criterion (Zekic-Susac et al., 2004). Consequently, two groups (nodes) of observations are formed. In each of the groups, separately, a new independent variable that allows for an efficient discrimination between good and bad payers is identified; consequently, new groups of observations are created. The process is recursively repeated until the sample is properly partitioned.

Breiman (1996) proposed a method called bagging, in which several decision trees are created from subsamples of the original data; randomly drawn with replacement, each individual is classified as a "good" or "bad" payer in each of the trees generated by the process; through a voting scheme, the final classification of the customer is the highest-frequency classification. Breiman (2001) proposed a change to the bagging model, which he called random forest, in which some independent variables are also randomly selected in the construction of each decision tree. The random forest technique has been applied mainly in the areas of genetics and medicine. As observed in Kraus (2014), Zhou & Wang (2012) and Malekipirbazari & Aksakalli (2015), some studies have obtained good results when applying the random forest technique to credit scoring models.

Chipman et al. (1998) and Chipman et al. (2010) proposed a model based on the sum of Bayesian trees, called BART. The main idea is imposing a prior that effectively regularizes the model, keeping individual trees simple, that is, with few variables. Conceptually, the BART model can be classified as a nonparametric Bayesian model. The authors compared the performance of the BART model with logistic regression and other methods based on statistical learning, such as least absolute shrinkage and selection operator (LASSO; Efron et al., 2004), gradient boosting (Friedman, 2001), neural networks (Desai et al., 1996; Malhotra & Malhotra, 2002) and random forest (Breiman, 2001). The authors' conclusion is that the BART model can be compared favorably with other machine learning methods.

The BART method was applied to credit risk modeling by Zhang & Härdle (2010), who referred to the method as Bayesian additive classification tree (BACT). The authors used a database of German companies (the German Creditreform database) that contained financial information about 20,000 solvent companies and 1,000 insolvent companies from the period between 1996 and 2002. The authors compared the performance of the BACT method with logistic regression and other methods based on statistical learning, such as CART, random forest and gradient boosting (Friedman, 2001). They concluded that the BACT model performs better than logistic regression and other machine learning methods. The BACT method is also robust for extreme values of input variables and can be adjusted for databases with few observations.

3. Methodology

In this chapter, the methodologies used in this work for credit scoring prediction will be reviewed.

3.1. Logistic regression

Logistic regression is a model developed for the case in which the response variable is binary ($y = 0$ or 1). Let $p_i = P(y = 1)$, the probability of a customer being a "bad" payer; then, the logistic regression model is given by

$$ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \quad (1)$$

where β_j, j=0, ..., p, are the model's parameters and x_{ji}, $i = 1,\ldots,n$, is the value of variable j observed for individual i.

It is common for samples to have a low number of observations of bad payers compared with good payers (Brown & Mues, 2012). When this class unbalance is present, the logistic regression technique may generate poor models (King & Zeng, 2001). One solution is to increase the minority class synthetically (Chawla et al., 2002) or to reduce the majority class (Yap et al, 2014).

3.2. Classification trees

For a better understanding of the models presented in this paper, a brief explanation of the classification tree model proposed by Breiman et al. (1984) is necessary. In this context, there is a dependent variable, y, and a set of independent variables. The objective is to partition the sample into homogeneous groups in relation to y based on the observed values of the independent variables ($X_1, \ldots, X_p$).

The first step is to identify the independent variable that best segregates the sample into distinct groups in relation to the dependent variable (Zekic-Susac et al., 2004). Consequently, two groups (nodes) of observations are formed. In Figure 1, we have "Node 0" (the parent node) with the entire database, containing 1,468 observations and with an average value of 0.5 for the dependent variable. The parent node is split into two new nodes named "Node 1" and "Node 2", called child nodes; for this division, the variable that better segregates the data into "good" and "bad" payers is used. In this case, the intention is to construct groups in which the differences between the means of the dependent variable is as large as possible. The variable X_1, which assumes the values 1 or 2, has been selected; "Node 1" will receive all observations with $X_1 = 1$, and "Node 2" will receive all observations with $X_1 = 2$. In each of the groups, separately, a new independent variable that efficiently segregates the node data is identified. The process is repeated recursively until some stop criterion is satisfied In Figure 1, we have "Node 1", which is splitted into two new nodes, 3 and 4, where variable X_2 yields the best division of "Node 1". "Node 3" will receive the observations with $X_1 = 1$ and $X_2 \leq 20$, whereas "Node 4" will receive the observations with $X_1 = 1$ and $X_2 > 20$. The nodes "Node 2" ($X_2 = 2$), "Node 3" and "Node 4" are called leaf nodes or terminal nodes because from these nodes, the sample is not partitioned further.

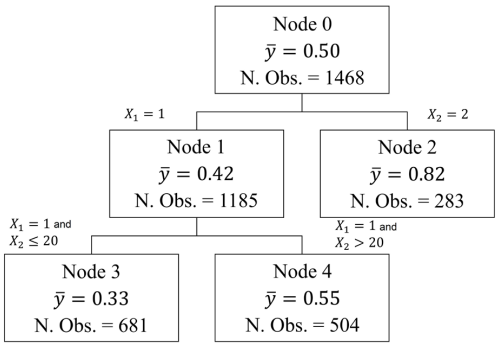

Figure 1. Decision tree partitioning example.

Here, we introduce the notation of the model with a single decision tree:

- $x_i = \{x_{1i}, \ldots, x_{pi}\}$ are the observed values for the independent variables of individual i.
- T is a binary tree with a set of inner nodes and a set of terminal nodes.
- L is the number of terminal nodes in tree T, where the terminal nodes are $l = 1, \ldots, L$.
- $M = \{\mu_1, \mu_2, \ldots, \mu_L\}$ are parameters associated with each terminal node l of tree T.

In Figure 1, $L = 3$ and $M = \{0.82; 0.33; 0.55\}$.

Each x_i is associated with a single terminal node of T, associating it with a single parameter μ_l related to its terminal node.

Chipman et al. (1998) define the model of a single tree as

$$y_i = g(x_i; T, M) + \varepsilon_i, \quad (2)$$

where ε_i is a random error.

Given T and M, we use function $g(x_i;T,M)$ to associate individual i with a terminal node $\mu_l \in M$. For Figure 1 example we have

$$g(x;T,M)=\begin{cases}0.82 & if\ x_{1i}=2\\ 0.33 & if\ x_{1i}=1\ and\ x_{2i}\le 20\\ 0.55 & if\ x_{1i}=1\ and\ x_{2i}>20\end{cases} \quad (3)$$

3.3. Random forest

The random forest is an evolution of the bagging method (Breiman, 1996). In this method, several decision trees are randomly created to be combined later. Each tree is formed from a bootstrap sample, extracted with replacement, from the original development sample. This procedure is repeated several times until a predefined limit on the number of trees is reached. Finally, each individual is classified as a "good" or "bad" payer for each of the trees generated by the previous process. The final classification of the customer is obtained by means of a voting scheme taking into account the number of trees that classify the customer as "good" or "bad" – if most trees classify him/her as "good", the customer will be considered "good", otherwise as "bad". Bagging represents an evolution relative to the models of a single tree because it is more stable in predicting new customers (Breiman, 1996). However, the bagging model always uses the same variables to create the trees, which can make the results obtained for different trees similar.

Random forest suggests a change relative to bagging; the independent variables are also randomly selected during the construction of each decision tree. The percentage of independent variables that will be selected in each random selection is a parameter of the model. According to Breiman (2001), this random selection of variables has the potential to use all available information in the set of variables.

3.4. BART

The BART model can be presented in three parts: first, the additive model of decision trees; second, the specification of the prior for the model's parameters to induce the posterior distribution; and finally, the stochastic process for generating the posterior distribution.

3.4.1. Sum of trees model

Assume the existence of a continuous unobservable variable y^*, which determines the value of y:

$$\begin{cases}y=1; if\ y^*\ge 0\\ y=0; if\ y^*<0\end{cases} \quad (4)$$

The idea of the BART model is to relate y^* with the independent variables by means of a probit additive model, where each term of the sum is represented by a tree based on the predictors x_i. We can then write the additive tree model as

$$y_i^*=\sum_{k=1}^{m} g_k(x_i;T_k;M_k)+\varepsilon=G(x_i)+\varepsilon, \quad (5)$$

where $\varepsilon \sim N(0,1)$ is the random error of the model; m is the number of trees that will be used in the model, T_k is the decision tree k with a set of interior decision nodes and a set of terminal nodes, L_k is the number of terminal nodes in tree T_k, $M_k=\{\mu_{1k}, \mu_{2k}, \ldots, \mu_{L_k k}\}$ are the parameters associated with tree T_k, and $G(x)=\sum_{k=1}^{m} g_k(x;T_k;M_k)$.

Each x_i is associated with a single terminal node of T_k. Then, the individual i is associated with a single μ_{lk} referring to a terminal node of tree T_k. Given T_k and M_k, we use a function $g_k(x_i;T_k,M_k)$ to assign a $\mu_{lk} \in M_k$ to an individual. In the application of credit scoring, this process consists of associating a prediction μ_{lk}, referring to the terminal node of tree T_k, with the values of the input predictors x_i. Note that

$$P(y=1\,|\,x_i)=P(y^*\ge 0\,|\,x_i)=P(G(x_i)+\varepsilon\ge 0)=P(G(x_i)\ge \varepsilon)=\Phi[G(x_i)], \quad (6)$$

where Φ is the cumulative distribution function of the standard normal distribution.

3.4.2. Regularization prior

Zhang & Härdle (2010) sugest a joint prior distribution for the tree structure T_k and the parameters μ_{lk} defined as

$$p\{(T_1,M_1),\ldots,(T_m,M_m)\}=\prod_{k=1}^{m}p(T_k,\,M_k)=\prod_{k=1}^{m}p(M_k\,|\,T_k)\,p(T_k), \tag{7}$$

defining $p(M_k\,|\,T_k)=\prod_{l=1}^{L}p(\mu_{lk}\,|\,T_k)$. Then, we have

$$p\{(T_1,M_1),\ldots,(T_m,M_m)\}=\prod_{k=1}^{m}p(T_k)\prod_{l=1}^{L}p(\mu_{lk}\,|\,T_k). \tag{8}$$

PRIOR FOR $\boldsymbol{T_k}$

Using the suggestion of Chipman et al. (2010), the construction of the prior distribution for T_k proceeds through three stages:

1) Modeling the probability that a node is not terminal

This probability is directly associated with the size of the tree. The authors suggest that the probability of splitting node η is given by

$$p_{SPLITTING}(\eta|T_k)=\frac{\alpha}{\left(1+d_{\eta k}\right)^{\beta}}, \tag{9}$$

where $d_{\eta k}$ is the depth (level) of node n in the tree and $\alpha\in(0,1)$ and $\beta\in[0,\infty)$ are positive parameters that control the probability of a new partition. Zhang & Härdle (2010) say that the greater the depth of node η in tree T, the lower the probability is of a new partition and the greater the probability is that the node will become a terminal node. The authors recommend using the default values $\alpha=0.95$ and $\beta=2$; with these parameters, a higher probability will be assigned to the generation of individual trees of size 2 or 3. The bartMachine package provided by the R software specifies all the mentioned parameters that must be supplied by the user. Table 1 presents the prior distribution of a number of terminal nodes in a tree for different values of α and β.

Table 1. Prior distribution of the number of terminal nodes.

	Config. 1	Config. 2	Config. 3
α	0.50	0.95	0.95
β	2.00	2.00	0.10
A priori probability of trees with			
1 terminal node	0.500	0.050	0.050
2 terminal nodes	0.383	0.552	0.012
3 terminal nodes	0.098	0.275	0.004
4 terminal nodes	0.017	0.092	0.002
≥ 5 terminal nodes	0.003	0.031	0.932

Source: Zhang & Härdle (2010).

2) Modeling the variable that will split each node

Chipman et al. (1998) recommend the use of a uniform distribution for the choice of the variable that will be used in node splitting. Thus, if there exist a_η variables available in node η, the probability of one of these variables being chosen is $1/a_\eta$.

3) Modeling the decision rule on each node, given the variable that will define the splitting

The suggestion for the choice of cut-off point is again to use a uniform specification, now over the different values assumed by the variable, which will guide the formation of a new split for the definition of the cut-off point:

- if the variable is quantitative, one must choose the value of the cut-off point randomly between the values assumed by the node variable η;
- if the variable is qualitative, the choice among all partition possibilities will be made randomly.

PRIOR FOR $\mu_{lk} \mid T_k$

The prior distribution of the parameters $\mu_{lk} \mid T_k$ associated with the terminal nodes of tree T_k is defined as $\mu_{lk} \sim N\left(0;\sigma_u^2\right)$. In order to simplify the model and facilitate the generation of the posterior distribution, it is assumed that μ_{lk} follows a normal distribution and the mean of parameters μ_{lk} is zero, that is, $E[\mu]=0$.

From the sum of trees model, we have that y^* is associated with the sum of the m different parameters μ_{il}, that is, for a vector of predictors x, a parameter μ_{il} is associated through function $g(x;T_k;M_k)$. Given that the m different parameters μ_{lk} have prior independent distributions, we have

$$G(x) \sim N\left(0;m\sigma_u^2\right) \tag{10}$$

Thus, we can define a confidence interval for $G(x)$ as being $G_{min}=-z\sigma_u\sqrt{m}$ and $G_{max}=z\sigma_u\sqrt{m}$, in which z determines the confidence level. For example, if $z=2$, then the probability of the values of y^* to be between the interval G_{min} and G_{max} is 95.45%.

From (5), we have $P(y=1 \mid x)=\Phi\left[G(x)\right]$. Thus, Chipman et al. (2010) suggest considering that $-3 \le G(x) \le 3$. Thus, a method of determining the value of σ_u is to use a convenient choice of z to make $G_{max}=3$ ($G_{min}=-3$). Thus,

$$\sigma_u=\frac{3}{z\sqrt{m}}. \tag{11}$$

The authors recommend the use of z between 1 and 3, having obtained good results with $z=2$.

For the computation of the formulated Bayesian problem, the MCMC (Markov chain Monte Carlo) technique was used; this technique is summarized in Chipman et al. (2010).

The BART method can easily detect and infer reduced models in problems with a large number of variables and for samples with a small number of observations (Chipman et al., 2010). This method of selection of variables is less effective when the number of trees is very high because it tends to mix important predictor variables with those that are not relevant. Bleich et al. (2014) proposed a variable selection model based on BART that uses as a criterion the proportion of times that the variable was used as a rule of segregation of a new branch of the various trees divided by the model's total number of branches, that is, it selects those variables that appear more frequently in the adjusted sum of trees model. The bartMachine package provided by the R software implements this procedure.

4. Database

The database used in this paper was provided by Serasa Experian and contains data regarding customers of direct retail consumer credit operations. The database provided by the credit bureau has 10,356 customer observations of direct retail consumer credit operations and 198 variables for the year 2014. Although random trees and BART were designed for larger datasets it is not unusual to find papers that aim to compare estimation methods designed for big datasets with sample sizes equivalent to ours, see, for instance, Chipman et al. (2010), Yeh et al. (2012), Leong (2016), Abellán & Castellano (2017), Bequé & Lessmann (2017), and several papers analysed in Lessmann et al. (2015) review.

The first group of predictor variables includes the amount of demand for credit of a specific borrower, in several different segments and in different periods of time. The segments are checks, real estate, banks, financial agencies, industries, insurance, services, telephony, retailer, utilities and others. The credit demand periods are up to 30 days, from 31 to 60 days, from 61 to 90 days, from 91 to 180 days and from 181 to 360 days, totaling 68 independent variables.

The second group of predictor variables is related to the first group. These variables measure the time in days since the first demand and since the last credit demand of a specific borrower by several segments. The segments are checks, banks, financial agencies, insurance, telecommunication and retail. This group has a total of 12 independent variables.

The third group is related to the number of events of the borrower registered in the credit bureau during certain periods of time. Events recorded at the bureau are active or settled debts, protests, bounced checks, active or resolved refusals by bank or financial agency, active or resolved refusal by companies that are not banks or financial agencies and active creditors. The time periods are 1 month, 2 months, 3 months, 6 months, 12 months, 2 years and 5 years. This group has a total of 60 independent variables.

Finally, the fourth group of predictors is related to the third group and measures the financial value registered in the credit bureau related to the described events. This group has a total of 40 independent variables.

Thirteen variables were excluded due the large number of missing values.

In addition to the described variables, whether the borrower was a "good" or "bad" payer was also indicated; this variable was used as a dependent variable in the calibration of the credit scoring model based on past data. However, the credit bureau did not report the criteria used to qualify borrowers as "good" or "bad" payers.

5. Development of the models

The models were estimated using the R software (R Core Team, 2016), employing the bartMachine package (Kapelner & Bleich, 2013) and randomForest (Liaw & Wiener, 2002).

5.1. Preparation of the database

After a descriptive analysis of the predictor variables, it was verified that the quantitative variables had a high concentration of observations with no demand for credit or no registered event of the borrower, and there were cases in which some variables assumed values of zero for all of the sample's observations. The same happened for the variables of time in days. Therefore, we used information value techniques (weight of evidence – Siddiqi, 2012) to disregard variables of the model if their predictive power was weak. The analysis suggested the exclusion of 141 variables from the model, with 31 variables remaining.

5.2. Database for the development and validation of the model

Table 2 presents the proportion of "good" and "bad" payers in the database.

This sample was randomly divided into a subset for the development (estimation) or calibration of the model and another for testing the accuracy of the models; the latter is typically known as the out-of-sample data or validation dataset.

Table 2. Basis provided by the credit bureau.

Customer Types	Full Sample		Development Samples				Validation Sample	
			Unbalanced		Balanced			
	N	%	N	%	N	%	N	%
GOOD	9,319	90	6,522	90	727	50	2,797	90
BAD	1,037	10	727	10	727	50	310	10
Total	10,356	100	7,249	100	1,454	100	3,107	100

According to Abdou & Pointon (2011), some studies divide the sample into 50% for development and 50% for model validation, whereas other studies use 70% for training or calibration and 30% for model testing (the validation sample). Two samples were created for the development database, the first from a random selection of 70% of the original sample; this is referred to as the unbalanced development sample. The second was a balanced development sample formed by the reduction of the majority class provided by the credit bureau. Table 2 presents the proportion of "good" and "bad" payers for each of the mentioned samples.

5.3. Application of the models

In this section, the models developed in this paper are presented. Section 5.3.1 describes the results of the logistic regression models, section 5.3.2 describes the random forest models, section 5.3.3 describes the results of the BART models, and in section 5.3.4, these models' performances are compared. The models were adjusted and tested on the balanced development database and on the unbalanced development data.

For evaluating the performance oh the methods the Kolmogorov-Smirnov statistics and the Gini coefficient were used. In order to compare the performance of the models, the statistical test proposed by Delong et al. (1988) was used. In addition, the bootstrap method proposed by Carpenter & Bithell (2000) was used for generating the confidence interval to compare the ROC curves.

5.3.1. Logistic regression

First, a reduced logistic regression model was fitted in the balanced and unbalanced development database, where the predictor variables were selected using the stepwise forward method, using variables with p-value lower than 10%. Table 3 presents the performance of reduced models for both the balanced and unbalanced cases.

We observed a performance improvement of the reduced logistic regression model.

Table 3. Performance of the Logistic Regression Models.

Model	Basis	KS %	AUC %	Gini %
Balanced	Development	35.90	73.96	47.91
	Validation	**31.99**	**68.85**	**37.69**
Unbalanced	Development	35.56	73.03	46.07
	Validation	**33.84**	**69.45**	**38.90**

5.3.2. Random forests

Initially, the random forest models were adjusted in the balanced development database with 500 trees and 10%, 25%, 50% and 100% of predictor variables used in each random selection; they will be referred as balanced models. We also adjusted models in the unbalanced development database using the same parameters; they will be as unbalanced models. In total, eight models were generated. The values of the parameters applied in the models were based on the study of Chipman et al. (2010). Additionally, the minimum number of individuals that a terminal node can contain was controlled, that is, if a child node contained a number of individuals less than 30, a new branch of the tree was not created. This parameter is important because if not specified, the tree growth can occur to the point at which terminal nodes have only one individual, and the phenomenon of overfitting occurs. Table 4 presents the results of the models obtained with the random forest method.

Table 4. Performance of random forest models.

Model	Basis	% of variables used in random selection	KS %	AUC %	Gini %
Balanced	Development	10	46.63	80.43	60.86
		25	52.27	82.98	65.95
		50	54.88	84.16	68.32
		100	56.40	84.77	69.54
	Validation	10	32.14	69.99	39.99
		25	**32.99**	**70.35**	**40.71**
		50	32.33	70.36	40.71
		100	31.94	69.93	39.85
Unbalanced	Development	10	49.90	81.54	63.08
		25	59.62	86.45	72.89
		50	60.21	87.05	74.10
		100	60.82	87.42	74.83
	Validation	10	33.24	**70.31**	**40.61**
		25	32.39	69.79	39.58
		50	32.07	69.40	38.80
		100	31.03	68.81	37.61

The models that exhibited the best performance for the development database were those with more variables, for both the balanced and unbalanced models. For the validation database, the balanced models that presented the best performance were those that used 25% and 50% of the variables. The unbalanced models that exhibited the best performance for the validation basis were those that used 10% and 25% of the variables.

All eight models tested had a good capacity to separate "good" and "bad" payers. They also performed better than those of logistic regression for prediction for the development database. However, the better performance also for predictions in the validation database suggests the superiority of the random forest method compared to logistic regression.

5.3.3. BART

First, models were adjusted using the standard parameters for the prior suggested by Chipman et al. (2010), with number of trees, m, equal to 50 and $z = 2$. In addition to the default models with 200-size trees were also adjusted, and 4 values for z (1, 2, 3 and 5) were tested, which specifies that the prior distribution for $E(y^* \mid x)$ be within an interval between G_{min} and G_{max} is 68.27%, 95.45%, 99.73% and approximately 100%, respectively.

These parameters combined with one another and applied to the two databases generated sixteen models. Table 5 summarizes the performance of the method for the validation database.

Table 5. Performance of the BART models for the validation sample.

Model	(m) Number of trees	(z) Prior of the interval of E(y*x)	KS %	AUC %	Gini %
Balanced	50	1	32.99	70.91	41.81
	50	2	**34.24**	**71.02**	**42.04**
	50	3	33.81	70.92	41.84
	50	5	30.50	69.44	38.89
	200	1	33.71	70.88	41.76
	200	2	33.96	70.73	41.47
	200	3	33.31	70.10	40.21
	200	5	28.64	68.34	36.67
Unbalanced	50	1	32.41	70.68	41.37
	50	2	**34.70**	**70.91**	**41.82**
	50	3	34.89	71.33	42.67
	50	5	**34.81**	**71.60**	**43.20**
	200	1	34.45	70.55	41.11
	200	2	34.99	71.09	42.17
	200	3	35.06	71.25	42.50
	200	5	34.63	71.52	43.04

All sixteen models tested had a good capacity to separate "good" and "bad" payers. They exhibited a superior performance compared to the logistic regression models in the validation database and superior to most of the models obtained using the random forest method, except for $z = 5$, for the balanced model, suggesting the superiority of the BART method over logistic regression models and random forest.

The BART models with a higher number of trees and different priors for the range $E\left(y^{*}|x\right)$ did not exhibit performance superior to the default BART model with number of trees equal to 50 and $z = 2$, as suggested by Chipman et al. (2010). The results suggest that the use of the standard tree number equal to 50 is preferred due to the shorter processing time compared to models that used a quantity of 200 trees.

5.3.4. Comparison of models

In this section, a comparison of the models is presented. For this comparison, we selected the best model of each of the presented techniques adjusted to the balanced database and the best for the unbalanced database. They appear in bold in Tables 3, 4 and 5. The selection of the models was based on the KS statistics and the Gini coefficient calculated on the validation basis.

The best adjusted models were selected on the balanced database. For logistic regression, the reduced model was used. In the case of random forest, the model that used 25% of the variables was selected, whereas for the BART model, the model adjusted with a standard prior was selected. The balanced models applied to the validation database were compared, which shows the generalization capacity of the model because the customers of this database were not used in the model estimation process.

The best adjusted models in the unbalanced database were selected. For the logistic regression, the reduced model was used. In the case of random forest, the model that used 10% of the variables was selected. For the BART model, the model with 50 trees and $z = 5$ was selected because it had the greatest area under the ROC curve (AUC). For comparison purposes, the BART model with the standard prior was also selected.

In order to confirm the results of the comparison of the balanced models, hypothesis tests were performed to verify whether the methods' areas under the curve (AUC) were the same, using the Delong and bootstrap techniques. In additional, in order to confirm the results of the comparison of the unbalanced models, hypothesis tests were also performed to verify whether the methods' AUC values were the same. Table 6 presents the results of the hypothesis tests for both balanced and unbalanced models.

Based on the results of the hypothesis tests obtained for the balanced models, we can say that the BART model was superior to the logistic regression model ($p < 0.01$) . There is no evidence to reject the hypothesis that the BART model had the same performance as the random forest model (Delong method with $p = 0.213$ and

Table 6. Comparison of the AUC of the different models.

Sample	Hypotheses	Test			
		Delong		Bootstrap	
		z	p	z	p
Balanced	H_0: Log. Reg. = R. Forest	-1.958	0.050	-1.934	0.053
	H_0: Default BART = R. Forests	-1.246	0.213	-1.246	0.213
	H_0: Default BART = Log. Reg.	-3.322	0.001	-3.332	0.001
Unbalanced	H_0: Logistic Reg. = R. Forest	-0.922	0.356	-0.967	0.334
	H_0: BART = R. Forests	-2.028	0.043	-2.004	0.045
	H_0: BART = Logistic Regression	-2.869	0.004	-2.788	0.005
	H_0: Default BART = R. Forest	-0.884	0.376	-0.907	0.365
	H_0: Default BART = Log. Reg.	-1.977	0.048	-1.958	0.050

bootstrap with $p = 0.213$). The comparison between the random forest models and the logistic regression was within the limit of 5% significance, suggesting the superiority of the random forest model.

For the unbalanced model, we can say that the best BART model was superior to the logistic regression models ($p < 0.01$) and random forest ($p < 0.05$). The default BART model was also superior to the logistic regression model ($p < 0.05$) , but with 5% of significance, we cannot reject the hypothesis that the default BART model had the same performance as the random forest model. In addition, with 5% significance, we cannot reject the hypothesis that the logistic regression models and random forests models had same performance.

A similar result was observed by Zhang & Härdle (2010), where the BART model applied to the credit scoring of a database of German companies yielded a result superior to that of logistic regression.

6. Conclusions

This paper empirically evaluated the performance of two machine learning models, BART and random forest, applied to credit scoring for predicting a "good" or "bad" payer. The models' performance was compared to that of logistic regression which is currently the most-used model for credit scoring.

For the models' development, a database provided by Serasa Experian (Brazilian credit bureau) was used; this database comprises customers of direct retail consumer credit operations. Different from the commonly used databases for the development of credit scoring, which use variables such as the characteristics of the individuals, the database used in this dissertation contains variables defined from the point of view of a credit bureau. This database was divided into two parts, one part for the development of the model and the other for validation of the developed model (called the out-of-sample or validation database). It is important to notice that the performance of the models developed for this dataset, measured by Gini index is equivalent of what it is expect for application credit scoring models (Gini above 0.4, according to Thomas, 2009, p. 120).

The results for the AUC or Gini coefficient suggest that the BART and random forest machine learning models were superior to logistic regression in both the balanced sample and the unbalanced sample. The results of the KS test suggest that the BART model was superior to the logistic regression model for both the balanced sample and the unbalanced sample, and the random forest model was superior to logistic regression only for the balanced sample. In addition, hypothesis tests were performed to compare the models' AUC metrics. The results of the comparison suggest that the best BART model was superior to the logistic regression model for both the balanced sample and the unbalanced sample, and it was superior to the random forest model for the unbalanced model. The results of the comparison of the random forest model suggest that it was superior to the logistic regression model only in the balanced sample.

The empirical results suggest that the BART and random forest machine learning models have a good capacity to separate "good" and "bad" payers, both in the balanced sample and in the unbalanced sample.

The results confirm the superiority of the BART model relative to the logistic regression model. Although the performance differences of the BART and random forest models were not much higher than the logistic regression model, considering their application in the classification of "good" or "bad" loan payers, the best performance of the machine learning models can represent significant gains for financial institutions through loss reduction by not granting a loan to a "bad" payer or through raising revenue by not denying a loan to a "good" payer.

For future research, other machine learning techniques, such as gradient boosting or LASSO, could be investigated because they have also been demonstrated to be promising methods for solving classification

problems, as discussed in Chipman et al. (2010) and Zhang & Härdle (2010). Another extension of this paper would be to increase the minority class synthetically (Chawla et al., 2002) to create balanced samples.

Other important research topics include application of hybrid systems (Hsieh, 2005), new concepts of adaptation to changes (Pavlidis et al., 2012) and dynamic modeling (Crook & Bellotti, 2010) to a database from a credit bureau.

Another possibility of further investigation is the improvement of the models performance by using simultaneously credit bureau variables and the variables regularly used in application or behavior credit scoring models by financial intitutions in the same model.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance & Management*, *18*(2-3), 59-88.

Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, *73*, 1-10.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankrupt. *The Journal of Finance*, *23*(4), 589-609.

Anderson, R. (2007). *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford: Oxford University Press.

Bank for International Settlements – BIS. (2004). *Implementation of Basel II: Practical considerations*. Basel: Bank for International Settlements. Retrieved in 2018, May 3, from https://www.bis.org/publ/bcbs109.htm

Bank for International Settlements – BIS. (2006). *International convergence of capital measurement and capital standards: a revised framework - comprehensive version*. Basel: Bank for International Settlements. Retrieved in 2018, May 3, from https://www.bis.org/publ/bcbs128.htm

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: an empirical evaluation. *Expert Systems with Applications*, *86*, 42-53.

Bleich, J., Kaperner, A., Geroge, E. I., & Jensen, S. T. (2014). Variable selection for BART: an application to gene regulation. *The Annals of Applied Statistics*, *8*(3), 1750-1781.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Boca Raton: CRC Press.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*, 1141-1164.

Chandler, G. G., & Coffman, J. Y. (1979). A comparative analysis of empirical vs. judgmental credit evaluation. *Financial Review*, *14*(4), 23-23.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935-948.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian Additive and Regression Trees. *The Annals of Applied Statistics*, *4*(1), 266-298.

Crook, J., & Bellotti, T. (2010). Time varying and dynamic models for default risk in consumer loans. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *173*(2), 283-305.

Delong, E. R., Delong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837-845.

Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24-37.

Durand, D. (1941). *Risk elements in consumer instalment financing*. Cambridge: NBER Books.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2), 407-499.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179-188.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232.

Für, F., Lima, J. D., & Schenatto, F. J. A. (2017). Uma revisão sistemática da literatura sobre credit scoring. In: *VII Congresso Brasileiro de Engenharia de Produção* (pp. 1-12). Rio de Janeiro: ABREPRO.

Gestel, T. V., Baesens, B., Suykens, J. A. K., Poel, D. V., Baestaens, D. E., & Willekens, M. (2006). Bayesian kernel based classification for financial distress detection. *European Journal of Operational Research*, *172*(3), 979-1003.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. Hoboken: John Wiley & Sons.

Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications*, *28*(4), 655-665.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4-37.

Kapelner, A., & Bleich, J. (2013). *Bartmachine: machine learning with bayesian additive regression trees* (pp. 1-40). Retrieved in 2017, January 10, from https://cran.r-project.org/web/packages/bartMachine/vignettes/bartMachine.pdf

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*(2), 137-163.

Kraus, A. (2014). *Recent methods from statistics and machine learning for credit scoring* (Dissertation). Fakultät für Mathematik, Informatik und Statistik der Ludwig–Maximilians–Universität München, München. Retrieved in 2018, May 3, https://edoc.ub.uni-muenchen.de/17143/1/Kraus_Anne.pdf

Kruppa, J., Schwarzb, A., Armingerb, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, *40*(13), 5125-5131.

Lee, T.-S., & Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *28*(4), 743-752.

Lensberg, T., Eilifsen, A., & McKee, T. E. (2006). Bankruptcy theory development and classification via genetic programming. *European Journal of Operational Research*, *169*(2), 677-697.

Leong, C. K. (2016). Credit risck scoring with Bayesian network models. *Computational Economics*, *47*(3), 423-446.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124-136.

Li, S.-T., Shiue, W., & Huang, M.-H. (2006). The evaluation of consumer loans using support vector machines. *Expert Systems with Applications*, *30*(4), 772-782.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18-22.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: systematic review and overall comparison. *Surveys in Operations Research and Management Science*., *21*, 117-134.

Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forest. *Expert Systems with Applications*, *42*(10), 4621-4631.

Malhotra, R., & Malhotra, D. K. (2002). Differentiating between good credits and bad credits using neuro-fuzzy systems. *European Journal of Operational Research*, *136*(1), 190-211.

Pavlidis, N. G., Tasoulis, D. K., Adams, N. M., & Hand, D. J. (2012). Adaptive consumer credit classification. *The Journal of the Operational Research Society*, *63*(12), 1645-1654.

R Core Team. (2016). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved in 2017, February 10, from https://www.R-project.org/

Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*. Hoboken: John Wiley & Sons.

Sousa, M. R., Gama, J., & Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications*, *45*(1), 341-351.

Thomas, L. C. (2009). *Consumer credit models: pricing, profit and portfolios: pricing, profit and portfolios*. Oxford: Oxford University Press.

Thomas, L. C., Oliver, R. W., & Hand, D. J. (2005). A survey of the issues in consumer credit modelling research. *The Journal of the Operational Research Society*, *56*(9), 1006-1015.

Wei, G., Yun-Zhong, C., & Minh-Shu, C. (2014). A new dynamic credit scoring model based on the objective cluster analysis. In Z. Wen, & T. Li (Ed.), *Practical applications of intelligent systems* (pp. 579-589). New York: Springer Berlin Heidelberg.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & Operations Research*, *32*(10), 2543-2559.

Xia, Y., Liu, B., Wang, S., & Lai, K. K. (2000). A model for portfolio selection with order of expected returns. *Computers & Operations Research*, *27*(5), 409-422.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In T. Herawan, M. Deris, & J. Abawajy (Ed.), *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)* (pp. 13-22). Singapore: Springer.

Yeh, C.-C., Lin, F., & Hsu, C.-Y. (2012). A hybrid KVM model, random forests and rough set theory approach for credit rating. *Knowledge-Based Systems*, *33*, 166-172.

Zekic-Susac, M., Sarlija, N., & Bensic, M. (2004). Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models. In *Information Technology Interfaces. 26th International Conference on IEEE* (pp. 265-270). USA: IEEE.

Zhang, J. L., & Härdle, W. K. (2010). The Bayesian Additive Classification Tree applied to credit risk modelling. *Computational Statistics & Data Analysis*, *54*(5), 1197-1205.

Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forest. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, *10*(6), 1519-1525.