Decision Models in Credit Risk Management

Herbert Kimura, Leonardo Fernando Cruz Basso and Eduardo Kazuo Kayo

Abstract Economic crises that emerge from systemic risks suggest that credit risk management in banks is paramount not only for the survival of companies themselves but also for a resilient worldwide economy. Although regulators establish strictly standards for financial institutions, i.e., capital requirements and management best practices, unpredictability of market behavior and complexity of financial products may have strong impact on corporate performance, jeopardizing institutions, and even economies. In this chapter, we will explore decision models to manage credit risks, focusing on probabilistic and statistical methods that are coupled with machine learning techniques. In particular, we discuss and compare two ensemble methods, bagging and boosting, in studies of application scoring.

Keywords Financial risks · Credit risks · Ensemble methods · Machine learning techniques · Bagging · Boosting

1 Introduction

With the development of financial and capital markets, credit operations have become more voluminous and complex, implying the need for advances in mechanisms and models for risk measurement and management.

H. Kimura (✉)
Faculdade de Economia, Administração e Contabilidade, Universidade de Brasília, Campus Darcy Ribeiro Prédio da FACE Asa Norte, Brasília, DF 70910900, Brazil
e-mail: herbertkimura@unb.br

L.F.C. Basso
Universidade Presbiteriana Mackenzie, Rua da Consolação, 930 - Consolação, São Paulo 01302090, Brazil
e-mail: leonardo.basso@mackenzie.br

E.K. Kayo
Faculdade de Economia, Administração e Contabilidade FEA-USP, Avenida Professor Luciano Gualberto, 908, Butantã, São Paulo, SP 05508010, Brazil
e-mail: eduardo@kayo.com.br

P. Guarnieri (ed.), *Decision Models in Engineering and Management*, Decision Engineering, DOI 10.1007/978-3-319-11949-6_4

Given the increasing importance and sophistication of credit transactions and the consequent vulnerability of the financial system to systemic crises, international and local regulatory bodies are developing guidelines and establishing rules concerning exposure to credit risk by financial institutions.

For example, the Basel Committee on Banking Supervision (BCBS) has published several guidelines to be adopted by banks worldwide, including mechanisms for credit risk management (Schooner and Talor 2010).

More specifically, the BCBS establishes as a relevant pillar the need for equity capital to cope with the degree of exposure to different types of risk (BIS 2006), including market and credit risk. Based on the BCBS guidelines, central banks in many countries are requiring regulatory capital for financial institutions in order to support financial losses due to defaults by borrowers and the degradation of credit quality of bank's assets.

In particular, retail credit risk plays a relevant role to financial institutions (Burns 2002), since risk in retail business could be seem as homogeneous due to diversification, and may result in significant savings in regulatory capital. In addition, banks that comply to their proprietary models of default probability estimation may also be allowed to adopt internal mechanisms to calculate regulatory capital requirements, reducing capital charge.

This study aims to analyze effective decision models for credit risk analysis of retail portfolios. Using machine learning algorithms, this chapter assesses computationally intensive algorithms to classify an individual as good or bad borrower.

In this study, algorithms could be adopted to analyze credit risk of wholesale portfolios, which provide more data and are more commonly prone to automated process for credit application of small loan amounts. However, computational learning mechanisms are most useful for retail portfolios.

Considering the various types of machine learning algorithms, this research studies the applicability of two ensemble methods, bagging and boosting, in credit risk analysis. Ensemble methods are computational mechanisms based on machine learning meant to improve traditional classification models. For instance, according to Freund and Schapire (1999), boosting, a traditional ensemble method combined with simple discrimination techniques (hit rate slightly higher than 55 %), could reach up to 99 % of correct classifications.

The adoption of ensemble methods in credit has been analyzed, for instance, by (Lai et al. 2006; Alfaro et al. 2008; Hsieh and Hung 2010). Their results have verified the efficacy of machine learning methods in real-life problems.

This chapter analyzes one example that shows how different classification techniques can be adopted by comparing the hit ratio of traditional and ensemble methods on a set of credit applications.

2 Theoretical Background

According to Johnson and Wichern (2007), discrimination and classification correspond to multivariate techniques that seek to separate distinct sets of objects or observations and that allow to allocate new objects or observations into predefined groups.

Although the concepts of discrimination and classification are similar, Johnson and Wichern (2007) establish that discrimination is associated with describing different characteristics from the observation of distinct populations known in an exploratory approach. Classification, on the other hand, is more related to allocating observations in classes, with a less exploratory perspective. According to Klecka (1980), classification is an activity in which discriminant variables or discriminant functions are used to predict the group to which a given observation is most likely to belong.

Therefore, the usefulness of discrimination and classification in credit analysis is evident. It provides not only an understanding of the characteristics that discriminate, for example, good from bad borrowers, but also models that allocate potential borrowers in groups. In this case, a priori assumptions on relationships between specific characteristics of borrowers and default risk are unnecessary.

The seminal study by Fisher (1936) associated with the identification of discriminant functions of species of flowers has given rise to relevant works on credit risk. For example, Durand (1941) focused on the analysis of automobile credit loans, and Altman (1968) associated it to predict business failures.

The discussion in this study focuses on general techniques that might improve credit analysis and that do not need to distinguish discrimination from classification. However, from a practical point of view, the ultimate goal of automated credit scoring models, more particularly the analysis of application scoring, is associated with classification, since the decision to grant the loan depends on the group to which a potential borrower is rated.

2.1 Traditional Discrimination Techniques

From the discrimination point of view, credit analysis aims to study possible relationships between variables $\mathbf{X}$ representing n characteristics $X_1, X_2, \cdots, X_n$ of borrowers and a variable Y representing their credit quality.

In loan application studies, credit quality is commonly defined by a variable having a numerical scale score or a rating score, with ordinal variables; or by good/bad credit indicators, with nominal or categorical variables.

The main objective of this research is to develop a system for automated decision processes. Therefore, this study focuses on problems in which Y is a dichotomous variable, with (i) good borrower and (ii) bad borrower categories or groups. Of the several multivariate statistic-oriented classification techniques currently available (Klecka 1980), this study discusses briefly discriminant analysis, logistic regression, and recursive partitioning algorithm.

2.1.1 Discriminant analysis

Discriminant analysis aims to determine the relationship between a categorical variable and a set of interval scale variables (Jobson 1992). By developing a multivariate linear function discriminant analysis shows variables that segregate or distinguish groups of observations through scores (Klecka 1980).

According to credit-related studies, discriminant analysis generates one or more functions in order to better classify potential borrowers. From the mathematical point of view, the analysis of two groups (e.g., performing and non-performing loans) might require a discriminant function expressed as:

$$Y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_{\mathrm{n}}x_{\mathrm{n}}$$

where

Y is the dependent variable, i.e., the score obtained by an observation;
$a_0, \ldots, a_n$ are coefficients that indicate the influence of each independent variable in the classification of an observation; and
$x_1, \ldots, x_n$ are independent variable values associated with a given observation.

Thus, based on the coefficient values and the associated independent variables, a discriminant function determines a more accurate score for any particular group. Portfolio credit analysis of retail loan applications adopts variables to encompass registration data or other information or characteristics of a potential borrower. Individuals with higher scores tend to have better ratings, indicating better credit quality and lower default probability.

The main assumptions of discriminant analysis include the following: (i) A discriminant variable cannot be a linear combination of other independent variables; (ii) the variance–covariance matrices of each group must be equal; and (iii) the independent variables have a multivariate normal distribution (Klecka 1980).

It is worthy noting that discriminant analysis is one of the most common borrower classification techniques in application scoring models, after the studies of Altman (1968) verified its efficacy.

2.1.2 Logistic regression

Many social phenomena are discrete or qualitative, in contrast to situations that require an ongoing measurement process of quantitative data (Pampel 2000). Credit quality classification focusing on good or bad borrowers is typically qualitative and represents a binary phenomenon.

In a dichotomous model, logistic regression is an alternative to discriminant analysis in order to classify of potential borrowers. In logistic regression, the dependent variable Y is defined as a binary variable with 0 or 1 values, and the independent variables $\mathbf{X}$ are associated with the characteristics or events of each group.

Without loss of generality, group 0 could be defined as good-borrowing individuals, and group 1 as non-payers or bad borrowers. A logistic function shows the default probability of a given individual:

$$P_i[Y = 1|\mathbf{X} = \mathbf{x_i}] = \frac{1}{1 + e^{-Z}}$$

where

P_i is the probability of individual i belong to the default group;
$Z = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx$ is a score in which the coefficients can be estimated from a sample, for instance.

Considering the use of logistic regression analysis for credit analysis, P_i is the probability of a counterpart i be a bad borrower. It is also subject to several independent variables $\mathbf{X}$ related to relevant characteristics that may affect credit quality.

When the assumptions of discriminant analysis and logistic regression are observed, both methods give comparable results. However, when the normality assumptions of the variables or variance–covariance matrix equality between groups are not observed, results might differ considerably. Logistic regression, given its less restrictive assumptions, is a technique widely used by the market for credit analysis.

2.1.3 Recursive partitioning algorithm

A less traditional technique for discrimination between groups, the recursive partitioning algorithm involves a classification tree-based non-parametric modeling (Thomas et al. 2002).

According to Feldesman (2002), classification trees have several advantages compared to parametric models: (i) they do not require data transformations, such as logit function in logistic regression analysis; (ii) missing observations do not require special treatment; and (iii) a successful classification does not depend on normality assumptions of variables or equal variance–covariance matrices between groups, such as in discriminant analysis.

The foundations of recursive partitioning algorithm lie in the subdivision of a set of observations into two parts, like branches of a tree, so that subsequent subgroups are increasingly homogeneous (Thomas et al. 2002). The subdivision is based on reference values by variables that explain the differences among the groups. Observations with higher values than the reference values are allocated in a group, while observations with lower values are classified into another group.

Thus, for each relevant variable, the algorithm sets a reference value that will define the subgroup. For example, if the discriminant variable X is continuous, its algorithm generates a cutoff value k. As a result, both groups are comprised by observations with a value $X < k$ and $X \geq k$, respectively. The definition of the cutoff value k is relevant in the classification tree model.

When the discriminant variable X is categorical, the algorithm checks all the possible splits into two categories and defines a measurement to classify the groups (Thomas et al. 2002). By repeating this procedure for several relevant variables, one can build a set of simple rules based on higher or lower values compared to a reference value for each discriminant variable. Observation can be classified into a final group according to this set of rules.

Classification trees allow an intuitive and easy representation of the elements that explain each group (Breiman et al. 1984). Credit analysis studies that adopt the classification tree model are not as common as the parametric model-based ones, but are found, for example, in Coffman (1986).

For discrimination among groups, discriminant analysis and logistic regression are parametric statistical techniques; the possible relationships between the borrower's characteristics and credit quality are likely to be analyzed by means of the independent variables' coefficients in the model. In the case of partition algorithms or decision trees, which adopt mainly non-parametric techniques, the explanatory variable-associated cutoff identifies the good and the bad borrowers. However, depending on how complex the recursive partitioning model is, assessing the influence of each variable to explain credit quality might be difficult.

2.2 Classification Techniques

Considering the distinction suggested by Johnson and Wichern (2007), one could argue that discrimination has the merit of allowing, under a more exploratory aspect, the evaluation of specific characteristics that may explain the inclusion of a observation within a particular group.

However, in some situations, explaining reasons for a variable to influence credit quality is less relevant than the actual rating itself. For example, under a practical perspective, if a given financial institution needs to analyze a large number of credit applications, it might need to develop an automated mechanism for quick and accurate classification rather than a discrimination pattern to explain how variables influence a possible default.

Regarding classification applicability and guidance, machine learning is an artificial intelligence field that aims to develop algorithms for computer programs or systems to learn from experience or data (Langley 1995).

Machine learning techniques, such as neural network algorithms and decision trees, are an alternative to traditional statistical methods, which often rely on mechanisms with extremely restrictive assumptions, such as normality, linearity, and independence of explanatory variables (Kuzey et al. 2014).

It is worth mentioning that recursive partitioning-based algorithms (e.g., decision trees within certain limits, especially related to a small number of variables and to the simplicity of the model), could also create discrimination mechanisms. Chien et al. (2006), for example, establish a classification tree model based on discriminant functions. In contrast, traditional neural networks are typical observation-based

classification techniques, as their underlying model is encapsulated in a black box (Ugalde et al. 2013).

From a more focused paradigm to pattern recognition for classification, the machine learning approach is, according to computer science literature, a set of algorithms specifically designed to assess computationally intensive problems, exploring extremely large databases of banks (Khandani et al. 2010).

From the credit analysis perspective, therefore, the machine learning methods are increasingly useful, given the computers' increasing processing power that, in turn, speeds up pattern recognition of good and bad payers. It is worth noting that loan databases of financial institutions could surpass ten million transactions, each one involving several variables, including borrower registration and transaction-related data.

This study focuses on machine learning techniques known as ensemble methods. According to Opitz and Maclin (1999), an ensemble consists of a set of individually trained functions whose predictions are combined to classify new observations. That is, the basic idea of the ensemble construction approach is to make predictions from an overall mechanism by integrating multiple models, which generates more accurate and reliable estimates (Rokach 2009).

According to Bühlmann and Yu (2003), Tukey (1977) introduces a linear regression model applied first to the original data, and then applied to errors, as the source of ensemble methods. Thus, applying a technique several times to the data and errors is an example of ensemble method. Considering the development of statistical theory and increasingly powerful computational machines, model combinations might be deployed in more complex applications.

Several authors, such as Breiman (1996), Bauer and Kohavi (1999), and Maclin and Opitz (1997), pointed substantial improvements in classification using ensemble methods. Considering its performance gains for classification, ensemble methods or ensemble learning methods are one of the mostly accepted streams of research in supervised learning (Mokeddem and Belbachir 2009).

Hsieh and Hung (2010) mention that ensemble methodology has been used in many areas of knowledge. For example, Tan et al. (2003) apply ensemble methods in bioinformatics and protein classification problems in several classes. In geography and sociology, Bruzzone et al. (2004) detect the land cover by combining image classification functions. Maimon and Rokach (2004) use ensemble decision tree techniques for mining manufacturing data.

The number of finance studies that adopt ensemble methods has also increased. For example, Leigh et al. (2002) make predictions on New York Stock Exchange values through technical analysis pattern recognition, neural networks, and genetic algorithms. Lai et al. (2007) study value-at-risk positions in crude oil gathering through ensemble methods that adopt wavelet analysis and artificial neural networks.

Regarding ensemble methods for credit applications, Lai et al. (2006) adopted neural reliability-based networks, Alfaro et al. (2008) adopted neural networks in bankruptcy analysis, and Hsieh and Hung (2010) assessed credit scores by combining neural networks, Bayesian networks, and support vector machines.

This study analyzes two traditional ensemble-based algorithms: bagging and boosting. According to Dietterich (2000), the two most popular ensemble techniques

are bagging or bootstrap aggregation, developed by Breiman (1996); and boosting, first proposed by Freund and Schapire (1998). The best known algorithms are based on the AdaBoost family of algorithms. Boosting is also known as arcing (resampling and combining adaptive), due to Breiman's work (1998) that brought new ways of understanding and using boosting algorithms.

Within the context of ensemble methods, bagging and boosting are two general mechanisms aimed to enhance the performance of a particular learning algorithm called basic algorithm (Freund and Schapire 1998). These methods reduce estimation error variances (Tumer and Ghosh 2001) but do not necessarily increase bias (Rokach 2005), providing gains both from the statistical theory perspective and the real-world applicability perspective. Bartlett and Shawe-Taylor (1999) reported that such methods may even reduce bias.

According to Freund and Schapire (1998), bagging and boosting algorithms are similar in the sense that they incorporate modified versions of the basic algorithm subject to disturbances in the sample. Both methods are based on resampling techniques that obtain different training datasets for each of the model classifiers (Opitz and Maclin 1999). In the case of classification problems, the set of training data allows establishing matching or classification rules derived from a majority vote, for example.

The algorithms may also show significant differences. The main difference implies that, in bagging, disturbances are introduced randomly and independently, while boosting shows serial and deterministic disturbances. The best choice depends heavily on all other previously generated rules (Freund and Schapire 1998).

Next, this work introduces the fundamentals of bagging and boosting methods for credit score. Similar ensemble method applications have been assessed by other authors, e.g., Paleologo et al. (2010), who study credit score for bagging, and Xie et al. (2009), who analyze boosting applied with logistic regression.

2.2.1 Bagging

Bagging is a technique developed to reduce variance and has called the attention due to its simple implementation and due to the popular bootstrap method.

The bagging algorithm follows the discussion in Breiman (1996).

1. Consider initially a classification model, based on pairs(X_i, Y_i), $i = 1, .., n$, representing the observation, and where $X_i \in \mathbb{R}^d$ indicates the d independent variables that explain the classification of a given group.
2. The target function is $P[Y = j|X = x](j = 0, 1, \ldots, J-1)$ in the case of a classification problem in J groups, $Y_i \in \{0, 1, \ldots, J-1\}$. The classification function estimator is $\hat{g}(.) = h_n((X_1, Y_1,), (X_2, Y_2), \ldots, (X_n, Y_n))(.) : \mathbb{R}^d \rightarrow \mathbb{R}$, where h_n is a model used to classify the observation into the groups.
3. The classification function can be, for instance, a traditional discrimination technique, e.g., discriminant analysis, logistic regression, or recursive partitioning model.
4. Build a random bootstrap sample $\left(X_1^{*,}, Y_1^*\right), \ldots, \left(X_n^*, Y_n^*\right)$ from the original sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

5. Calculate the bootstrap estimator $\hat{g}^*(.)$ using the plug-in principle, i.e., $\hat{g}^* = h_n\left(\left(X_1^*, Y_1^*\right), \ldots, \left(X_n^*, Y_n^*\right)\right)(.)$
6. Repeat steps 2 and 3 M times. Frequently, M is chosen to be 50 or 100, implying that M estimators are $\hat{g}^{*k}(.)(k = 1, \ldots, M)$.
7. The bagged estimator is given by $\hat{g}_{\text{Bag}}(.) = M^{-1}\sum_{k=1}^{M}\hat{g}^{*k}(.)$, which is an estimate of $\hat{g}_{\text{Bag}}(.) = E^*[\hat{g}^*(.)]$.
   In application scoring problems, each bootstrapped sample implies coefficient estimates when the bagging procedure is coupled with discriminant analysis or logistic regression, or estimates of cutoff values in a decision tree when bagging and recursive partitioning algorithm are coupled. Since M different classifications are generated in bagging due to differences in the bootstrapped samples, one common mechanism to classify a new individual is by majority votes of the classification derived from the many $\hat{g}^{*k}(.)$ classification functions.

2.2.2 Boosting

Boosting is an ensemble technique that aggregates a series of simple methods, known as weak classifiers, due to their low performance in classifying objects, thus generating a combination that leads to a classification rule with a better performance (Freund and Schapire 1998).

In contrast with bagging, boosting relies on classifiers and subsamples that are sequentially obtained. In every step, training data are rebalanced to give more weight to incorrectly classified observations (Skurichina and Duin 2002). Therefore, the algorithm rapidly focuses on observations that could be more difficult to be analyzed our classified.

The description of the AdaBoost algorithm here is based on Freund and Schapire (1999) study. Consider $Y = \{-1, +1\}$ as possible classification problem values. In a credit application context, for instance, a negative value may represent a bad borrower, and a positive value may represent a good borrower.

Boosting implies a repeated execution of a weak learning mechanism, e.g., discriminant analysis, logistic regression, or a decision tree approach, using subsamples of the original set. Differently from bagging, which generates uniform random samples with reposition, choosing new subsamples in boosting depends on a probability distribution that is different for each step, reflecting the mistakes and successes from the weak classification functions.

A boosting algorithm can be described as in Freund and Schapire (1999).

1. Define weights $D_t(i)$ of the training sample. Initially, the weights, i.e., the probability of choosing any observation, are equal. Thus, given $(x_i, y_i), \ldots, (x_m, y_m)$, so $x_i \in X, y_i \in Y = \{-1, +1\}$, $D_1(i) = \frac{1}{m}$.
2. Establish a weak hypothesis or function h_t that allows a simple classification of a given element in -1 or $+1$, i.e., $h_t : X \to \{-1, +1\}$. This function can be, for instance, a traditional statistical technique such as recursive partitioning algorithm.

3. The classification function has an error $\varepsilon_t = \Pr_{i\sim D_t}[h_t(x_i) \neq y_i] = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$, i.e., the error is the total sum of probabilities in which the weak function leads to wrong classifications in relation to the true values in the sample. It is important to emphasize that the error is measured by this distribution D_t, in which the weak function was used.
4. Once the weak hypothesis h_t has been established, boosting defines a parameter α_t that measures the relative importance of h_t. The higher is the error ε_t, the lower is α_t and less important h_t is in the classification problem. In boosting, the relative importance for each weak classification function is given by $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.
5. The distribution D_t is updated by increasing the weight of the observations that are wrongly classified by h_t, and by decreasing the weight of the observations that are correctly classified, following the equation $D_{t+1}(i) = \frac{D_t(i)}{Z_i} \times \begin{cases} e^{-\alpha t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha t} & \text{if } h_t\ (x_i) \neq y_i \end{cases}$, where Z_t is a normalization factor, so that D_{t+1} is a probability distribution. Therefore, for each successive boosting step, the observations that are not correctly classified will be more likely to be selected in the new subsample.
6. The last hypothesis or classification function is h_t. The final classification model H is defined by the weak function in each step weighted by α_t, i.e., $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$. For the retail loan application, an individual is considered a good borrower if $H(x)$ has a positive sign. A bad borrower shall have a negative value for $H(x)$.

3 Results

In order to show how these ensemble methods of machine learning work, a credit transaction database in the UCI Machine Learning Repository of the Center for Machine Learning and Intelligent Systems at the University of California at Irvine (Bache and Lichman 2013) was used. This database, also used by Quinlan (1987) and Quinlan (1992), encompasses credit card applications in Australia and consists of 690 observations of 15 variables.

Given the confidentiality of information, the database provides only the values and information on the scale of the variables. The individuals are not identified, as well as the observation or variable meaning. The credit quality-related variable has two categories: good borrower (G) and bad borrower (B). Limited information ensures data confidentiality, but does not affect the analysis, considering the research objective associated with the classification of observations using various quantitative techniques.

After analysis of the database, missing data were eliminated, resulting in 653 valid observations in the final sample. In order to run the analysis, we focused on 7 variables to classify individuals: 6 continuous and 1 nominal comprising two categories. We aim to study how the machine learning mechanisms behave in classification problems with a limited number of information.

The final sample observations were divided randomly into two subsamples (training and validation sets), with virtually the same amount of elements. A script written in R was used, taking into account the characteristics of each technique, and confusion matrices were generated for both the training and the testing subsamples.

The classification results were introduced through (i) discriminant analysis, (ii) logistic regression analysis, (iii) recursive partitioning algorithm, (iv) bagging, and (v) boosting, for different number of iterations (N). The ensemble methods analyzed were coupled with recursive partitioning algorithm.

Tables 1 and 2 show the classification results, in absolute and in percentage terms, for the training and validation samples, respectively. Table 3 shows the overall classification results, with hit and error ratios.

This study's dataset implies some relevant results. Discriminant analysis and logistic regression results were identical, in accordance with Press and Wilson's (Press and Wilson 1978) argument that, for most studies, the two methods are unlikely to lead to significantly different results.

Interestingly, for good borrowers, discriminant analysis and logistic regression show better classification results (25 %) in the testing subsample, vis-à-vis the training subsample (21 %). Therefore, for the good borrower group, the traditional parametric models are more consistent with the validation sample when compared to the calibration sample.

However, for the bad borrower group, accuracy levels decrease for all techniques. Recursive partitioning algorithm, bagging, and boosting mechanisms show a lower hit ratio for the good borrower group as well.

An overall analysis shows that all techniques, with the exception of discriminant analysis and logistic regression, are subject to performance loss when the classification rule using the training subsample is applied to the testing subsample.

In the training dataset, classification results from the recursive partitioning algorithm, bagging, and boosting are quite superior to the discriminant analysis and logistic regression outcomes. Whereas the traditional parametric models lead to an overall 74 % hit ratio, the non-parametric methods correspond to at least 83 % of the correct classifications. This accuracy increase, resulting from an automated computational procedure, may strongly affect banks, since loan application analysis, using just computational resources, could be significantly improved.

Regarding boosting, the higher the number of allowed iterations, the better the classification results for the training, i.e., the calibration dataset. Results show an accuracy rate of 93 %, which is much higher than the traditional statistical technique accuracy rate, 74 %.

However, it is important to highlight that the performance of the models did not vary significantly in the testing sample for any technique. Hit ratio is quite

Table 1 Classification results—absolute numbers

N = 10	LDA	LR	RPA	BAG	BOOS	LDA	LR	RPA	BAG	BOOS
Training sample	Predicted = good					Predicted = bad				
Actual = good	69	69	109	111	104	75	75	35	33	40
Actual = bad	11	11	18	12	16	171	171	164	170	166
Testing sample	Predicted = good					Predicted = bad				
Actual = good	82	82	107	94	99	70	70	45	58	53
Actual = bad	10	10	34	28	31	165	165	141	147	144
N = 50	LDA	LR	RPA	BAG	BOOS	LDA	LR	RPA	BAG	BOOS
Training sample	Predicted = good					Predicted = bad				
Actual = good	69	69	109	111	119	75	75	35	33	25
Actual = bad	11	11	18	11	10	171	171	164	171	172
Testing sample	Predicted = good					Predicted = bad				
Actual = good	82	82	107	95	107	70	70	45	57	45
Actual = bad	10	10	34	29	35	165	165	141	146	140
N = 100	LDA	LR	RPA	BAG	BOOS	LDA	LR	RPA	BAG	BOOS
Training sample	Predicted = good					Predicted = bad				
Actual = good	69	69	109	108	129	75	75	35	36	15
Actual = bad	11	11	18	10	7	171	171	164	172	175
Testing sample	Predicted = good					Predicted = bad				
Actual = good	82	82	107	101	109	70	70	45	51	43
Actual = bad	10	10	34	29	36	165	165	141	146	139

Obs.: *LDA* Linear discriminant analysis, *LR* Logistic regression, *RPA* Recursive partitioning algorithm, *BAG* bagging, and *BOOST* boosting

Table 2 Classification results—percentage

N = 10	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)
Training sample	Predicted = good					Predicted = bad				
Actual = good	21	21	33	34	32	23	23	11	10	12
Actual = bad	3	3	6	4	5	52	52	50	52	51
Testing sample	Predicted = good					Predicted = bad				
Actual = good	25	25	33	29	30	21	21	14	18	16
Actual = bad	3	3	10	9	9	50	50	43	45	44
N = 50	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)
Training sample	Predicted = good					Predicted = bad				
Actual = good	21	21	33	34	37	23	23	11	10	8
Actual = bad	3	3	6	3	3	52	52	50	52	53
Testing sample	Predicted = good					Predicted = bad				
Actual = good	25	25	33	29	33	21	21	14	17	14
Actual = bad	3	3	10	9	11	50	50	43	45	43
N = 100	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)
Training sample	Predicted = good					Predicted = bad				
Actual = good	21	21	33	33	40	23	23	11	11	5
Actual = bad	3	3	6	3	2	52	52	50	53	54
Testing sample	Predicted = good					Predicted = bad				
Actual = good	25	25	33	31	33	21	21	14	16	13
Actual = bad	3	3	10	9	11	50	50	43	45	43

Table 3 Overall classification results

N = 10	Ratio	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)
Training sample	Right	74	74	84	86	83
	Wrong	26	26	16	14	17
Testing sample	Right	76	76	76	74	74
	Wrong	24	24	24	26	26
N = 50	Ratio	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)
Training sample	Right	74	74	84	87	89
	Wrong	26	26	16	13	11
Testing sample	Right	76	76	76	74	76
	Wrong	24	24	24	26	24
N = 100	Ratio	LDA (%)	LR (%)	RPA (%)	BAG (%)	BOOS (%)
Training sample	Right	74	74	84	86	93
	Wrong	26	26	16	14	7
Testing sample	Right	76	76	76	76	76
	Wrong	24	24	24	24	24

insensitive to the method or the number of iterations in the ensemble models. Moreover, forecasting results are compatible to those using more simple techniques.

Even worse, in the testing sample, ensemble methods showed poor performance, especially when bad borrowers were predicted as good borrowers. This misclassification can lead to significant credit losses, since the automated decision would suggest the approval of a loan to a borrower who would default.

These results suggest that, in the case of the Australian credit card database, although ensemble methods could be seen as an improved model of an existing dataset, their contribution to predict credit quality in an out-of-sample analysis is not clear.

4 Final Comments

This chapter aimed to discuss decision models for retail credit risk. In particular, potential uses of two ensemble methods, bagging and boosting, to application scoring were assessed. Based on unsupervised machine learning algorithms, these ensemble methods could implement decision models for automated response to loan applications.

Using a dataset of credit card applications and compared to traditional discriminant analysis and logistic regression, decision models that rely on computational algorithms such as ensemble methods could enhance the accuracy rate of borrower classification.

Results show that, specifically for the training subsample, bagging and especially boosting significantly improve the classification hit ratio. However, for the testing subsample, ensemble techniques coupled with recursive partitioning algorithm convey only marginally better classifications. The error rate for classifying bad borrowers as good ones showed significant problems in the ensemble methods used in this study. Thus, although these machine learning techniques are likely to be more accurate in the training dataset, their impact for analyzing new loans applications is not robust.

Even though the computational techniques studied here did not significantly improve the hit ratio, it is important to highlight that even a minimum increase in the rate of correct classifications might result in relevant savings for a financial institution with millions of trades in its retail portfolio.

Therefore, automated decision models, especially for large banks, could result in economic value and a simpler analysis of credit applications. This study assessed bagging and boosting, two of the most common ensemble methods. Several other machine learning mechanisms, such as neural networks, support vector machines, and Bayesian networks, might also be adopted to analyze credit risk.

Due to the complex default process and the financial market dynamics, managers and decision makers could take advantage of innovations in both computational performance and quantitative methods, eventually developing automated decision models that could contribute to the credit analysis process.

References

Alfaro E, Garcia N, Gámez M, Elizondo D (2008) Bankruptcy forecasting: an empirical comparison of AdaBoost and neural networks. Decis Support Syst 45(1):110–122

Altman E (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance 23(4):589–609

Bache K, Lichman M (2013) UCI machine learning repository, School of Information and Computer Science. University of California, Irvine. http://archive.ics.uci.edu/ml. Accessed 10 Sep 2014

Bartlett PL, Shawe-Taylor J (1999) Generalization performance of support vector machines and other pattern classifiers. In: Schölkopf B, Burges CJC, Smola AJ (eds) Advances in Kernel Methods—Support Vector Learning. MIT Press, Cambridge, pp 43–54

Bauer E, Kohavi R (1999) An empirical comparison of voting classification algorithms: bagging, boosting, and variants. Mach Learn 36:105–139

BIS (2006) Bank of International Settlements. Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework—Comprehensive version. http://www.bis.org/publ/bcbs128.pdf. Accessed 10 Sep 2014

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breiman L (1998) Arcing classifiers. Ann Stat 26(3):801–849

Breiman L, Freidman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth International Group, Wadsworth

Bruzzone L, Cossu R, Vernazza G (2004) Detection of land-cover transitions by combining multidate classifiers. Pattern Recogn Lett 25(13):1491–1500

Bühlmann P, Yu B (2003) Boosting with the L2 loss: regression and classification. J Am Stat Assoc 98:324–339

Burns P (2002) Retail credit risk modeling and the Basel Capital Accord. Discussion paper, Payment Cards Center. https://www.philadelphiafed.org/consumer-credit-and-payments/payment-cards-center/publications/discussion-papers/2002/CreditRiskModeling_012002.pdf January 2002. Accessed 10 Sept 2014

Chien BC, Lin JY, Yang WP (2006) A classification tree based on discriminant functions. J Inf Sci Eng 22(3):573–594

Coffman JY (1986) The proper role of tree analysis in forecasting the risk behavior of borrowers, management decision systems. MDS Reports, Atlanta

Dietterich T (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting and randomization. Mach Learn 40(2):139–157

Durand D (1941) Risk elements in consumer installment financing. Studies in consumer installment financing: Study 8, National Bureau of Economic Research

Feldesman MR (2002) Classification trees as an alternative to linear discriminant analysis. Am J Phys Anthropol 119(3):257–275

Fisher R (1936) The use of multiple measurements in taxonomic problems. Ann Hum Genet 7 (2):179–188

Freund Y, Schapire RE (1998) Discussion of the paper "Arcing Classifiers" by Leo Breiman. Ann Stat 26(3):824–832

Freund Y, Schapire R (1999) A short introduction to boosting. J Jpn Soc Artif Intell 14 (5):771–780

Hsieh NC, Hung LP (2010) A data driven ensemble classifier for credit scoring analysis. Expert Syst Appl 37(1):534–545

Jobson J (1992) Applied multivariate data analysis. Springer Texts in Statistics, New York

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis. 6th edn, Prentice hall Englewood Cliffs

Khandani AE, Adlar JK, Lo AW (2010) Consumer credit-risk models via machine-learning algorithms. J Bank Finance 34:2767–2787

Klecka WR (1980) Discriminant analysis. Quantitative applications in the social sciences. Sage Publications, Beverly Hills

Kuzey C, Uyar A, Delen D (2014) The impact of multinationality on firm value: a comparative analysis of machine learning techniques. Decis Support Syst 59:127–142

Lai KL, Yu L, Shouyang W, Zhou L (2006) Credit risk analysis using a reliability-based neural network ensemble model. Lecture notes in computer science, Artificial neural networks—ICANN

Lai KL, He K, Yen J (2007) Modeling VaR in crude oil market: a multi scale nonlinear ensemble approach incorporating wavelet analysis and ANN. Lecture notes in computer science, computational science—ICCS

Langley P (1995) Elements of machine learning. Morgan Kaufmann Series in Machine Learning. 1st edn, Morgan Kaufmann, Burlington

Leigh W, Purvis R, Ragusa JM (2002) Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural networks, and genetic algorithm: a case study in romantic decision support. Decis Support Syst 32(4):361–377

Maclin R, Opitz D (1997) An empirical evaluation of bagging and boosting. In: Proceedings of the 14th national conference on artificial intelligence. Cambridge, MA, pp 546–551

Maimon O, Rokach L (2004) Ensemble of decision trees for mining manufacturing data sets. Mach Eng 4:l–2

Mokeddem D, Belbachir H (2009) A survey of distributed classification based ensemble data mining methods. J Appl Sci 9(20):3739–3745

Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198

Paleologo G, Elisseef A, Antonini G (2010) Subagging for credit scoring models. Eur J Oper Res 201(2):490–499

Pampel FC (2000) Logistic regression: a primer. Quantitative applications in the social sciences. Sage Publications, Beverly Hills
Press JS, Wilson S (1978) Choosing between logistic regression and discriminant analysis. J Am Stat Assoc 73(364):699–705
Quinlan R (1987) Simplifying decision trees. Int J Man-Mach Stud 27:221–234
Quinlan R (1992) C4.5: programs for machine learning. Morgan Kaufmann, Los Altos
Rokach L (2005) Ensemble methods for classifiers. Data mining and knowledge. discovery handbook. Springer, New York
Rokach L (2009) Taxonomy for characterizing ensemble methods in classification tasks: a review and annotated bibliography. Comput Stat Data Anal 53(12):4046–4072
Schooner HM, Talor MW (2010) The new capital adequacy framework: Basel II and credit risk. Global Bank Regulation, pp 147–164
Skurichina M, Duin RPW (2002) Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal Appl 5:121–135
Tan AC, Gilbert D, Deville Y (2003) Multi-class protein fold classification using a new ensemble machine learning approach. Genome Inf 14:206–217
Thomas LC, Edelman DB, Crook JN (2002) Credit scoring and its applications. Society for Industrial and Applied Mathematics, Philadelphia
Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading
Tumer K, Ghosh J (2001) Robust order statistics based ensembles for distributed data mining. In Kargupta H, Chan P (eds) Advances in distributed and parallel knowledge discovery. AAAIMIT Press, Cambridge, pp 185–210
Ugalde HMR, Carmona JC, Alvarado VM, Reyes-Reyes J (2013) Neural network design and model reduction approach for black box nonlinear system identification with reduced number of parameters. Neurocomputing 101:170–180
Xie H, Han S, Shu X, Yang X, Qu X, Zheng S (2009) Solving credit scoring problem with ensemble learning: a case study. In: Proceedings of the 2nd international symposium on knowledge acquisition and modeling. Wuhan, China, pp 51–54

Author Biographies

Herbert Kimura is full professor in the Business Administration Department, University of Brasília, and has doctoral degree in Business Administration from the University of São Paulo, Brazil.

Leonardo Fernando Cruz Basso is full professor in the Business Administration Department, Mackenzie University, and has PhD in Economics from the New School for Social Research, USA.

Eduardo Kazuo Kayo is associate professor in Business Administration Department, University of São Paulo, and has doctoral degree in Business Administration from the University of São Paulo, Brazil.