CREDIT RISK ANALYSIS APPLYING LOGISTIC REGRESSION AND NEURAL NETWORKS MODELS
[004-0092]

Eric Bacconi Gonçalves University of São Paulo, Brazil
Maria Aparecida Gouvêa University of São Paulo, Brazil
ABSTRACT
Most large Brazilian institutions working with credit concession use credit models to evaluate the risk of consumer loans. The first phase of this study introduces concepts of credit and risk. Subsequently, with a sample set of applicants from a large Brazilian financial institution, two credit scoring models are built applying these distinct techniques: Logistic Regression and Neural Networks. Finally, the quality and performance of these models are evaluated and compared to identify the best one. Results obtained by the logistic regression and neural network models are good and very similar, although the former is slightly better. This study shows the procedures to be adopted by a financial institution in order to identify the best credit model to evaluate the risk of consumer loans. Use of the best fitted model will favor the definition of an adequate business strategy thereby increasing profits.

INTRODUCTION
With the currency stability achieved by the Economical Plano Real in 1994, financial loans became a good business for the banks that no longer made such large profits from currency devaluation. (ROSA, 2000: 1). To replace this profitability, the need to increase investment alternatives was felt at the end of the inflation period. Thereafter institutions have endeavored to expand their credit portfolios. However, loans could not be offered at random to all the applicant clients, therefore ways to evaluate the candidates were required.

Some years ago, when applying for a loan, the client filled in a proposal for evaluation by one or more analysts. They then issued an opinion regarding the request (SEMOLINI, 2002: 103). Although effective, the process was slow because it did not accommodate the analysis of many requests. As such, the model for the analysis of the concession of credit was initially introduced in financial institutions aiming to speed up evaluation of proposals.

Models of analysis for extension of credit known as models of credit scoring are based on historical information from the databank on existing clients, in order to assess whether the prospective client will have a greater chance of being a good or bad payer. The models of credit scoring are added to the institution’s systems permitting on-line credit evaluation.

Objectives of the Study
Based on the data of a sample, the intention is to:
- Develop two credit scoring models by using two statistical/computational techniques: Logistic Regression and Neural Networks;
- Compare the models developed in terms of the quality of fitness and prediction indicators;
- Propose a model for the classification of clients.

THEORETICAL BASIS
In this section, the theoretical concepts that will support the theme of this work will be presented.

Consumer Credit
The expression consumer credit may be understood as a form of trade where a person obtains money, goods or services and vouches to pay for this in the future, adding a premium (interest) to the original value (SANTOS, 2000: 15).

Currently, consumer credit is a large industry operating worldwide. Major retailers spur their sales by supplying credit. Automobile companies, banks and other segments utilize consumer credit lines as an additional alternative to make profit. On the other hand, consumer credit injects resources into the economy, permitting production and economic expansion of a country, thereby bringing development to the nation (LEWIS, 1992: 2).
However to make credit widely available does not mean to distribute credit at random to all those requesting it; there is a factor associated to consumer credit which is crucial in the decision of making credit available or not: the risk.

**Risk of Credit**

On the financial market, risk of credit is the oldest form of risk (FIGUEIREDO, 2001: 9). It is the upshot of a financial transaction, contracted between the supplier of funds (giver of credit) and the user (taker of credit). Prior to any sophistication resulting from financial engineering, the mere act of lending a sum to someone entails the probability of it not being repaid, the uncertainty regarding return. This is, in essence, the risk of credit which may be defined as the risk of a counterpart, in an agreement of credit concession, not to meet his/her obligation.

According to Caouette et al. (2000: 1), “if credit may be defined as the expectation of receiving a sum of money in a given period, then Risk of Credit is a chance that this expectation is not fulfilled”.

The activity of credit concession is a basic function of banks, therefore risk of credit takes on a relevant role in the composition of an institution’s risks and may be found in the operations where there is a transfer of money to the clients as well as in those where there is only a possibility of usage, the pre-conceded limits. Primary types of a bank credit operation are: loans, financing, discount of payables, advancement to depositors, advancement of exchange, leasing operations, surety bonds and warranties etc. In these operations risk may take on different forms; to be conceptually familiar with them helps to orient management and mitigation.

In the universe of consumer credit, pledge of future payment involves the idea of risk. As the future cannot be fully predicted, all consumer credit involves risk, because assurance of payment does not exist (LEWIS, 1992: 2). Analysis of credit is charged with the task of estimating the risk involved in the concession or not of credit.

The maximum risk that the institution may accept relies on the policy adopted by the company. Risk presented by the applicant is of major significance for the process of credit concession, and various queries must be considered in its evaluation.

**Evaluation of the Risk of Credit**

Evaluation of risk is the main issue for concession of credit. If the risk is poorly evaluated the company will certainly lose money, be it because of acceptance of clients who will generate losses to the business or because of the refusal of good clients who would generate profits for the business. Companies who have a better evaluation than their competitors in the concession of credit have an advantage over the others as they are less vulnerable to the consequence of the wrong decisions when providing credit.

Evaluation of risk of a potential client can be carried out in two ways:

1. By judgment, a more subjective way involving a more qualitative analysis;
2. By classifying the taker by means of evaluation models, involving a more quantitative analysis.

Currently, almost all large sized companies working with concession of credit use a combination of both.

The models called credit scoring are used for the evaluation of risk of credit by classification of the applicant. They permit measurement of the credit applicant’s risk, to support the decision taking (concession or not of credit).

**Credit Scoring Models**

The pioneer of credit models was Henry Wells, executive of the Spiegel Inc. who developed a credit scoring model during the Second World War (LEWIS, 1992: 19). Wells needed tools that would allow inexperienced analysts to perform credit evaluation, because many of its qualified employees had been recruited for the War.

During the fifties the scoring models were disseminated in the American banking industry. The first models were based upon pre-established weights for certain given characteristics, summing the points to reach a classification score.
More extensive use of the models in the sixties transformed business in the American market (THOMAS, 2000: 154). Not only companies in the financial area, but also the large retailers began to use credit scoring models to carry out credit sales to their consumers. Retailers such as Wards, Bloomingdale’s and J.C. Penney were some of the pioneers in this segment.

Currently, about 90% of the American companies that offer some kind of consumer credit utilize models of credit scoring.

In Brazil the background is shorter. Financial institutions started to make an intensive use of credit scoring models only in the mid-nineties.

There are seven steps to be followed to construct a credit scoring model:

1. **Survey of a historical background of the clients**
   
   The basic supposition to construct a model of credit evaluation is that the clients have the same behavior pattern over time; therefore models are constructed based upon past information. The availability and quality of the data bank are fundamental for the success of the model (TREVISANI et al., 2004).

2. **Classification of clients according to their behavior pattern and definition of the dependent variable**
   
   In addition to good and bad clients there are also the excluded clients, those who have peculiar characteristics and should not be considered (for instance, workers in the institution) and the indeterminate clients, those on the threshold of being good or bad, still without a clear position about them. In practice, institutions consider only the good and bad clients to build the model because it is much easier to work with binary response models. This tendency to work only with good and bad clients is also noticed in academic works (ROSA 2000; OHTOSHI, 2003; SEMOLINI, 2002; HAND; HENLEY, 1997, among others).

3. **Selection of a random sample representative of the historical background**
   
   It is important that the samples of good and bad clients have the same size so as to avoid any possible bias due to size difference. There is no fixed number for the sample; however Lewis (1992: 31) suggests a sample of 1,500 good clients and 1,500 bad clients to achieve robust results. Habitually three samples are used, one for building of the model, another for the validation of the model and a third to test the model.

4. **Descriptive analysis and preparation of data**
   
   This consists of analyzing, according to statistic criteria, each variable that will be utilized in the model.

5. **Choice and application of techniques to be used in the construction of the model**
   
   Logistic Regression and Neural Networks will be used in this work. Hand and Henley (1997) further stress Discriminant Analysis, Linear Regression and Decision Trees as methods that can be used in practice. Recently some scholars have also used Survival Analysis (HARRISON; ANSELL, 2002; ANDREEVA, 2003). There is no method that is clearly better than the others, everything depends upon how the elected technique fits the data.

6. **Definition of the comparison criteria of the models**
   
   Measurement for the comparison of the models will be defined here, normally by the rate of hits and the Kolmogorov-Smirnov (KS) statistics.

7. **Selection and implementation of the best model**
   
   The best model is chosen using the previously defined criteria. As such, the implementation of the model must be programmed. The institution must adjust its systems to receive the final algorithm and program its utilization in coordination with the other areas involved.
METHODOLOGICAL ASPECTS

Description of the Study
A financial institution wishes to grant loans to its clients and therefore it requires a tool to assess the level of risk associated to each loan to support the decision making process. To set up this project, information on the history of the clients that contracted personal credit was made available.

The Product of Credit under Study
The product under study is personal credit. Individual credit is a rapid and practical consumer credit operation. The purpose of the loan does not need to be stated, and the loan will be extended according to the applicant’s credit scoring.

Another characteristic of the product in question is the lack of requirement of goods as a guarantee of payment.

The IOF - Tax of Financial Operations - as foreseen in the legislation and a Fee for Opening or Renovation of Credit are charged on Personal Credit.

The modality with pre-fixed interest rates with the loan terms ranging from 1 to 12 months was focused for this study.

The Data
To carry out this study a random selection was made in a universe of clients of the bank, 10,000 credit contracts, considered as good and 10,000 considered as bad, dated from August 2002 to February 2003. All these contracts had already matured, that is to say the sample was collected after the due date of the last installment of all contracts. This is an historical data-base with monthly information on the utilization of the product. Based upon this structure, the progress of the contract could be accompanied and particularized when the client did not pay one or more installments.

In the work, the sample is divided into three sub-samples coming from the same universe of interest: one for construction of the model, 8,000 data (4,000 good and 4,000 bad), the second for validation of the constructed model, 6,000 data (3,000 good and 3,000 bad) and the third also with 6,000 (with the same equal division) to test the model obtained.

Each sub-sample has its specific function (ARMINGER et al., 1997: 294). The sub-sample of the model’s construction is used to estimate the model’s parameters, the sub-sample of the test is to verify the power of prediction of the constructed models and the sub-sample of validation, especially in a neural network, has the function of validating the parameters, avoiding the overfitting of the model. In the model of logistic regression the validation sample will play the same role as the test sample, that is to say, evaluate the model’s prediction.

The Variables
The available explanatory variables have characteristics that can be divided into two groups: Reference File Variables, and Variables of Utilization and Restriction. Reference File Variables are related to the client and the Utilization and Restriction Variables regard the restriction of credit and notes about the client’s other credit operations existing in the market.

The Reference File Variables as well as those of Utilization and Restriction are collected when the client contracts the product.

Definition of the Dependent Variable
This definition of the Dependent Variable, also called Performance Definition, is directly related to the institution’s credit policy. For the product under study, clients delinquent for 60 or more days were considered Bad (default) and clients with a maximum delinquency of 20 days were considered Good.

Clients designated as undetermined represent a group whose credit behavior is not sufficiently clear to assign them as good or bad customers. In practice, clients who are not clearly defined as good or bad are analyzed separately by the credit analyst, based upon qualitative analysis.
LOGISTIC REGRESSION

Logistic Regression is the technique most often used in the market for the development of credit scoring models (ROSA, 2000; OHTOSHI, 2003). In relation to Discriminant Analysis it has the advantage of not surmising that input data must follow a Normal distribution. Logistic regression predicts the probability of an event taking place, which can range from 0 and 1.

Background

According to Lima (2002, pg. 77) the logistic function appeared in 1845 related to demographic growth issues for which it is still used until today. In the thirties this methodology began to be used in the domain of biology and later in areas related to economical and social issues. Paula (2002, pg. 118) stresses that, although the model of logistic regression has been known since the fifties, it was as a result of the works of the statistician David Cox in the seventies, that the technique became quite popular among the users of statistics.

Currently logistic regression is one of the main tools for statistical modeling of data, widely used for diverse types of problems. Paula (2002, pg. 118) explains:

Even when the reply is not originally binary, some researchers have dichotomized the variable reply so that the probability of success may be modeled by means of logistic regression. This is primarily due to the ease of interpretation of the parameters of a logistic model and also to the possibility of using this type of modeling in discriminant analysis.

Concepts

In the models of logistic regression, the dependent variable is, in general a binary variable (nominal or ordinal) and the independent variables may be categorical (as long as dichotomized after transformation) or continuous.

Consider the case when the observations may be classified into one of two mutually exclusive categories (1 or 0). As an example, these categories could represent an individual who may classify as a good or bad client.

The binary dependent variable Y may assume the values:

\[
Y_i = \begin{cases} 
1 & \text{If the } i\text{- nth individual belongs to the category of good} \\
0 & \text{If the } i\text{- nth individual belongs to the category of bad}
\end{cases}
\]

That is \( X = (1, X_1, X_2, ..., X_n) \) : a vector where the first element is equal to 1 (constant) and the remaining represent the n independent variables of the model.

The model of Logistic Regression is a particular case of the Generalized Linear Models (DOBSON, 1990; PAULA, 2002). The function which characterizes the model is given by:

\[
\ln \left( \frac{p(X)}{1 - p(X)} \right) = \beta' X = Z, \text{ where}
\]

\( \beta' = (\beta_0, \beta_1, \beta_2, ..., \beta_n) \) : vector of the parameters associated to the variables

\( p(X) = \text{E}(Y=1|X) \) : probability of the individual been classified as good, given the vector X.
This probability is expressed by (NETER et al, 1996, pg. 580):

\[ p(X) = \frac{e^{\beta X}}{1 + e^{\beta X}} = \frac{e^z}{1 + e^z} \]

**Method for choice of the variables**

Initially, in this work all variables will be included for the construction of the model, however in the final logistic model, only some of the variables will be selected. The choice of the variables will be done by means of the method forward stepwise, which is the most widely used in models of logistic regression. In the forward stepwise method the variables are selected at each step, according to criteria that optimize the model, reducing the variance and avoiding problems of multicolinearity. Only the variables of real importance for the model will be selected. For the details on the methodology reading of Canton (1988, pg. 28) and Neter et al (1996, pg. 348), is suggested.

**Strong and Weak Points of Logistic Regression**

Fensterstock (2005, pg. 48) points out the following advantages in using statistical techniques for the construction of models:

- The generated model takes into account the correlation between variables, identifying relationships that would not be visible and eliminating redundant variables;
- They take into account the variables individually and simultaneously;
- The user may check the sources of error and optimize the model.

In the same text, the author further identifies the disadvantages of this type of technique:

- In many cases preparation of the variables takes a long time;
- In the case of many variables the analyst must perform a pre-selection of the more important, based upon separate analyses;
- Some of the resulting models are difficult to implement.

**ARTIFICIAL NEURAL NETWORKS**

Artificial Neural Networks are computational techniques that present a mathematical model based upon the neural structure of intelligent organisms and who acquire knowledge through experience.

According to Haykin (1999, pg. 28)

A neural network is a processor massively distributed in parallel (sic), formed by units of simple processing that have a natural propensity to store experimental knowledge and make it available for use. It resembles the brain in two aspects: 1) knowledge is acquired by the network through a learning process; 2) connection forces between neurons, known as synaptic weights, are used to store the acquired knowledge.

**Background**

According to various authors, among them Marks and Schnabl (1997, pg. 3); Haykin (1999, pg. 63) and Fausett (1994, pg. 22), the first model of a neural network appeared with the work of McCulloch and Pitts. Warren McCulloch was a psychiatrist and neuro-anatomist who studied a representation for the nervous system. In 1942, he started to work with the mathematician Walter Pitts and in the following year they published an article that proposed a mathematical model for a neural network and until today is a reference in the study of neural networks (HAYKIN, 1999, pg. 63). A second important work was published by Hebb in 1949, where the first rules of learning for artificial neural networks were proposed; this work also stimulated many scholars in later research.

During the decades of the fifties and sixties much research and many studies were carried out permitting significant advances in the field of neural networks. Fausett (1994, pg. 23) calls this period “the golden years of neural networks”. Studies have shown that the new methodology would be very
promising, new types of networks and new learning rules were proposed, also the networks became increasingly complex.

However, in the decade of the 70’s research underwent a slow down as pointed out by Hair et al (1998, pg. 545): “(...) by the end of the sixties research showed that the neural networks of the time were really quite limited and the area underwent a major retrogression”.

It was only in the eighties that, because of the greater computational power, neural networks once again were widely studied and applied. Fausett (1994, pg.25) underlines the development of the backpropagation algorithm as the turning point for the popularity of neural networks. Until today, neural networks have been widely used, studied and utilized in different fields of knowledge such as medicine, biology, economics, administration and engineering.

Concepts
An artificial neural network model processes certain characteristics and produces replies like those of the human brain. Artificial neural networks are developed using mathematical models in which the following suppositions are made (FAUSETT, 1994 pg. 3):

1. Processing of information takes place within the so-called neurons;
2. Stimuli are transmitted by the neurons through connections;
3. Each connection is associated to a weight which, in a standard neural network multiplies itself upon receiving a stimulus;
4. Each neuron contributes for the activation function (in general not linear) to determine the output stimulus (response of the network).

The cited pioneer model by McCulloch and Pitts from 1943 for one processing unit (neuron) can be summarized in Figure 1:

- Signals are presented upon input;
- Each signal is multiplied by a weight that indicates its influence on the output of the unit;
- The weighted sum of the signals which produces a level of activity is made;
- If this level exceeds a limit, the unit produces an output.

In the diagram there are input signals \( X_1, X_2, ..., X_p \) and corresponding weights \( W_1, W_2, ..., W_p \) and the limit being \( k \).

In this model the level of activity is given by:
\[
    a = \sum_{i=1}^{p} W_i X_i
\]

And the output is given by:
\[
    y = 1, \text{ if } a \geq k \\
    y = 0, \text{ if } a < k
\]

Three characteristics must be taken into account in the definition of a model of neural networks: the form of the network called architecture, the method for determination of the weights, called learning algorithm; and the activation function.

Architecture relates to the format of the network. Every network is divided in layers, usually classified into three groups:
- Input Layer where the patterns are presented to the network;
Intermediate or Hidden layers in which the major part of processing takes place, by means of the weighted connections, they may be viewed as extractors of characteristics;

Output Layer, in which the end result is concluded and presented.

There are basically three main types of architecture (HAYKIN, 1999, pgs 46-48): feedforward networks with a single layer; feedforward networks with multiple layers and recurring networks.

Feedforward networks with a single layer are the simpler network, in which there is only one input layer and one output layer. Networks are fed farther on, that is to say, only the layer of input supplies information to the output layer. Some networks utilizing this architecture are: the Hebb Network, perceptron, ADALINE, among others.

Multilayered feedforward networks are those having one or more intermediate layers. Output of each layer is used as input for the next layer. Just as in the former architecture, this type of networks is characterized only by the feeding farther in front. The multilayer perceptron networks (MLP), MADALINE and of a radial base function are some of the networks utilizing this architecture.

In recurrent networks the output layer has at least one connection that feeds back the network. The networks called BAM (Bidirectional Associative Memory) and ART1 and ART2 (Adaptative Resonance Theory) are recurring networks.

Learning Process
The most important quality of neural networks is the capacity to “learn” according to the environment and thereby improve their performance (CASTRO JR. 2003, pg.92). This learning is carried out adjusting the weights through an interactive process. The purpose of the process is to achieve a learning algorithm permitting a generalized solution for a certain class of problems.

A learning algorithm is the name given to a set of well defined rules for the solution of a learning problem. There are many types of specific algorithms for given models of neural networks. These algorithms differ one from the other, principally by how the weights are modified.

There are essentially three types of learning:
1 Supervised Learning: in this type of learning the expected reply is indicated to the network. This is the case of this work, where a priori it is already known if the client is good or bad;
2. Non-supervised Learning: in this type of learning the network must only rely on the received stimuli; the network must learn to cluster the stimuli;
3 Reinforcement Learning: in this type of learning, behavior of the network is assessed by an external reviewer.

For each type of learning various algorithms can possibly be used. In the application section the algorithm which will be used in this work will be specified, as well as the reasons leading to this choice.

Functions of Activation
Each neuron contributes to the output stimulus. The function of activation plays the role of restricting the output amplitude of a neuron, in general [0.1] or [-1.1] (HAYKIN, 1999, pg.37). Some examples of the utilized activation functions are:

Liminar Function: \[ f(x) = \begin{cases} 1 & \text{If } x < k \\ 0 & \text{If } x \geq k \end{cases} \]

Logistic Function: \[ f(x) = \frac{1}{1 + e^{(-ax)}} \]

Hyperbolic Tangent Function: \[ f(x) = \tanh(x) \]
Strong and Weak Points of Neural Networks
Berry and Linoff (1997, pg.331) point out the following positive points in the utilization of neural networks:

- They are versatile: neural networks may be used for the solution of different types of problems such as: prediction, clustering or identification of patterns;
- They are able to identify non-linear relationships between variables;
- They are widely utilized, can be found in various software.

As for the disadvantages the authors state (pg. 333):
- Results cannot be explained: no explicit rules are produced, analysis is performed inside the network and only the result is supplied by the “black box”;  
- The network can converge towards a lesser solution: there are no warranties that the network will find the best possible solution; it may converge to a local maximum.

CRITERIA FOR PERFORMANCE EVALUATION
To evaluate performance of the model two samples were selected, one for validation and the other for test. Both were of the same size (3,000 clients considered good and 3,000 considered bad, for each one). In addition to the samples, other criteria are used, which are presented in this section.

Score of Hits
The score of hits is measured by dividing the total of clients correctly classified, by the number of clients included in the model.

Similarly, the score of hits of the good and bad clients can be quantified.

In some situations it is much more important to identify a good client than a bad client (or vice versa); in such cases, often a more fitting weight is given to the score of hits and a weighted mean of the score of hits is calculated.

In this work, as there is not a priori information on what would be more attractive for the financial institution (identification of the good or bad clients), the product between the score of hits of good and bad clients (Ih) will be used as an indicator of hits to evaluate the quality of the model. This indicator will privilege the models with high scores of hits for both types of clients. The greater the indicator is the better will be the model.

The Kolmogorov-Smirnov Test
The Kolmogorov-Smirnov (KS) is the other criterion often used in practice and used in this work (PICININI et al., 2003; OOGHE et al., 2001; Pereira, 2004).

The KS test is a non-parametric technique to determine whether two samples were collected from the same population (or from populations with similar distributions) (SIEGEL, 1975: 144). This test is based on the accumulated distribution of the scores of clients considered good and bad.

To check whether the samples have the same distribution, there are tables to be consulted according to the significance level and size of the sample (see SIEGEL, 1975: 309-310). In this work, as the samples are large, tendency is that all models reject the hypothesis of equal distributions. The best model will be that with the highest value in the test, because this result indicates a larger spread between the good and bad.

APPLICATION
This section will cover the methods to treat variables, the application of the two techniques under study and the results obtained by each one of them, comparing their performance. For descriptive analysis, categorization of data and application of logistic regression the SPSS for Windows v.11.0 software was used. The software Enterprise Miner. 4.1 was used for the selection of the samples and application to the neural network.
Treatment of the Variables

Initially, the quantitative variables were categorized.

The deciles (values below which 10%, 20% etc. of the cases fall) of these variables were initially identified for categorization of the continuous variables. Starting from the deciles, the next step was to analyze them according to the dependent variable. The distribution of good and bad clients was calculated by deciles and then the ratio between good and bad was calculated, the so called relative risk (RR).

Groups presenting a similar relative risk (RR) were re-grouped to reduce the number of categories by variable.

The relative risks were also calculated for the qualitative variables to reduce the number of categories, whenever possible. According to Pereira (2004: 49) there are two reasons to make a new categorization of the qualitative variables. The first is to avoid categories with a very small number of observations, which may lead to less robust estimates of the parameters associated to them. The second is the elimination of the model parameters, if two categories present a close risk, it is reasonable to group them in one single class.

Besides clustering of categories, RR helps to understand whether this category is more connected to good or to bad clients. This method of clustering categories is explained by Hand and Henley (1997: 527).

When working with the variables made available, heed was given to the following:

- The variables gender, first acquisition and type of credit were not re-coded as they are already binary variables;
- The variable profession was clustered according to the similarity of the nature of jobs;
- The variables commercial telephone and home telephone were recoded in the binary form as ownership or not;
- The variables commercial ZIP Code and home ZIP Code were initially clustered according to the first three digits, next the relative risk of each layer was calculated and later a reclustering was made according to the similar relative risk, the same procedure adopted by Rosa (2000: 17) as explained by Hand and Henley (1997: 527);
- The variable salary of the spouse was discarded from the analysis because much data was missing;
- Two new variables were created, percentage of the amount loaned on the salary and percentage of the amount of the installment on the salary. Both are quantitative variables, which where categorized in the same way as the remainder.

After applying this method, the categories shown on Table 1 were obtained.

Logistic Regression

For the estimation of the model of logistic regression, a sample of 8,000 cases equally divided in the categories of good or bad was utilized.

Initially, it is interesting to evaluate the logistic relationship between each independent variable and the dependent variable TYPE.

Since one of the objectives of this analysis was to identify which variables are more efficient for the characterization of the two types of bank clients, a stepwise procedure was utilized. The elected method of selection was forward stepwise.

Of the 53 available independent variables, considering k-1 dummies for each variable of k levels, 28 variables were included in the model.

In this study, Z is the linear combination of the 28 independent variables weighted by the logistic coefficients:

\[ Z = B_0 + B_1X_1 + B_2X_2 + \ldots + B_{28}X_{28} \]
Table 2 shows, per variable, the estimates of the logistic coefficients, the standard deviations of the estimates, the Wald statistics, the degrees of freedom, and the descriptive levels of the significance tests of independent variables.

With categorical variables, evaluation of the effect of one particular category must be done in comparison with a reference category. The coefficient for the reference category is 0.

Variables with a logistic coefficient estimated negative indicate that the focused category, with regard to the reference, is associated to a decrease of the odds and therefore a decrease in the probability of having a good client.

The coefficient of partial correlation is a measurement of the power of relation between the dependent variable and an independent variable, keeping constant the effects of the other independent variables. Variables that most affect positively the probability of having a good client are Qlp1, Qlp2 and Tlv1. At the opposite end the variables with a greater negative impact on this probability are Tc_P, Fa_N and Age2.

Table 2 shows that the coefficients of all variables included in the logistic model are statistically different from zero. Therefore, all have shown to be relevant for the discrimination between good and bad clients.

There are two statistical tests to evaluate the significance of the final model: the chi-square test of the change in the value of \(-2LL\) (-2 times the log of the likelihood) and the Hosmer and Lemeshow test.

Table 3 presents the initial value of \(-2LL\), considering only the model’s constant, its end value, the improvement and the descriptive level to measure its significance.

The model of 28 variables disclosed that the reduction of the \(-2LL\) measure was statistically significant.

The Hosmer and Lemeshow test considers the statistical hypothesis that the predicted classifications in groups are equal to those observed. Therefore, this is a test of the fitness of the model to the data.

The chi-square statistic presented the outcome 3.4307, with eight degrees of freedom and descriptive level equal to 0.9045. This outcome leads to the non rejection of the null hypothesis of the test, endorsing the model’s adherence to the data.

**Neural Network**

In this work, a supervised learning network will be used, as it is known a priori whether the clients in question are good or bad. According to Potts (1998: 44), the most used structure of neural network for this type of problem is the multilayer perceptron (MLP) which is a network with a feedforward architecture with multiple layers. Consulted literature (ARMINGER et al., 1997; ARRAES et al., 1999; ZERBINI, 2000; CASTRO JR., 2003; OHTOSHI, 2003) supports this statement. The network MLP will also be adopted in this work.

The MLP networks can be trained using the following algorithms: Conjugate Descending Gradient, Levenberg-Marquardt, Back Propagation, Quick Propagation or Delta-Bar-Delta. The more common (CASTRO JR., 2003: 142) is the Back Propagation Algorithm which will be detailed later on. For the understanding of the others, reading of Fausett (1994) and Haykin (1999) is suggested.

The implemented model has an input layer of neurons, a single neuron output layer, which corresponds to the outcome whether a client is good or bad in the classification of the network. It also has an intermediate layer with three neurons, since it was the network which presented the best outcomes, in the query of the higher percentage of hits as well as in the query of reduction of the mean error. Networks which had one, two or four neurons were also tested in this work.

Each neuron of the hidden layer is a processing element that receives n inputs weighted by weights \(W_i\). The weighted sum of inputs is transformed by means of a nonlinear activation function \(f(\cdot)\).
The activation function used in this study will be the logistic function, 
\[ g(x) = \frac{1}{1 + e^{-x}} \], where 
\[ g = \sum_{i=1}^{p} W_i X_i \] is the weighted sum of the neuron inputs.

Training of the networks consists in finding the set of \( W_i \) weights that minimizes one function of error. In this work for the training will be used the Back Propagation Algorithm. In this algorithm the network operates in a two step sequence. First a pattern is presented to the input layer of the network. The resulting activity flows through the network, layer by layer until the reply is produced by the output layer. In the second step the output achieved is compared to the desired output for this particular pattern. If not correct, the error is estimated. The error is propagated starting from the output layer to the input layer, and the weights of the connections of the units of the inner layers are being modified, while the error is backpropagated. This procedure is repeated in the successive iterations until the halt criterion is reached.

In this model the halt criterion adopted was the mean error of the set of validation data. This error is calculated by means of the module of the difference between the value the network has located and the expected one. Its mean for the 8,000 cases (training sample) or the 6,000 cases (validation sample) is estimated. Processing detected that the stability of the model took place after the 94th iteration. In the validation sample the error was somewhat larger (0.62 x 0.58), which is common considering that the model is fitted based upon the first sample.

Initially, the bad classification is of 50%, because the allocation of an individual as a good or bad client is random; with the increase of the iterations, the better result of 30.6% of error is reached for the training sample and of 32.3% for the validation sample.

Some of the statistics of the adopted network are in Table 4.

Evaluation of the Models’ Performance

After obtaining the models the three samples were scored and the \( I_h \) and KS were calculated for each of the models. Table 5 shows the results of classification reached by the models.

All presented good classification results, because, according to Picinini et al. (2003: 465): “credit scoring models with hit rates above 65% are considered good by specialists”.

The hit percentages were very similar in the models of logistic regression and neural network. Another interesting result is that, the models presented the greatest rate of hits for bad clients, with a higher than 70% rate for bad clients in the three samples of the logistic and neural network models.

Table 6 presents results of the criteria \( I_h \) and KS which were chosen to compare the models.

KS values in all models can be considered good. Again, Picinini et al. (2003: 465) explain: “The Kolmogorov-Smirnov test (KS) is used in the financial market as one of the efficiency indicators of the credit scoring models. A model which presents a KS value equal or higher than 30 is considered good by the market”. Here again, the logistic regression and neural network models exhibit very close results.
In choosing the model that best fits these data and analyzing according to the Ih and KS indicators, the model built by logistic regression was elected. Although results were very similar to those achieved by neural networks this model presented the best results in the test sample, suggesting that it is best fit for application in other databases. Nevertheless, it must be highlighted that the adoption of any one of the models would bring about good results for the financial institution.

CONCLUSIONS AND RECOMMENDATIONS

The objective of this study was to develop credit scoring predictive models based upon data of a large financial institution by using Logistic Regression and Artificial Neural Networks.

When developing the credit scoring models some care must be taken to guarantee the quality of the model and its later applicability. Precautions in the sampling, clear definition of criteria for the classification of good and bad clients and treatment of variables in the database prior to application of the techniques were the measures taken in this study, aiming to optimize results and minimize errors.

The two models presented suitable results for the database in question, which was supplied by a large retail bank operating in Brazil. The logistic regression model presented slightly better results to the model built by neural networks. The model proposed by this study to enable the institution to score its clients is:

\[
p(X) = \frac{e^Z}{1 + e^Z}
\]

\(p\): probability of the client being considered good and

\[Z = B_0 + B_1.X_1 + B_2.X_2 + \ldots + B_{28}.X_{28},\]

where the values of \(B_i\) and \(X_i\) are found in Table 2.

For the test sample the percentage of total hits for logistic regression and neural networks were respectively equal to 68.3 and 67.7. In the literature consulted, the percentage of total hits fluctuates significantly, as well as the model that best fits each data bank can be different from that obtained in this study. Table 7, taken from the work by Thomas (2000), shows the range of results achieved in other works.

<table>
<thead>
<tr>
<th>Insert Table 7 about here</th>
</tr>
</thead>
</table>

Table 8, built from the surveyed literature, is similar to the previous table and strengthens the wide variety of results.

<table>
<thead>
<tr>
<th>Insert Table 8 about here</th>
</tr>
</thead>
</table>

When these two tables are analyzed it should be noted that the models present a precision of classification ranging from 56.2 to 93.2. Further it is noted that except for the linear programming, all the other models presented the greatest precision, in at least one of the studies.

This study did not aim at a more detailed approach of the techniques focused. Neural networks presented an extensive range of structures and variations that may (and must) be better explored.

In credit risk problem, new techniques such as survival analysis should not be overlooked and merit attention in future studies.
REFERENCES


Figure 1: The McCulloch and Pitts model
<table>
<thead>
<tr>
<th>Variable</th>
<th>Category</th>
<th>Variable name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Masculine, Feminine</td>
<td>Gender_M, Gender_F</td>
</tr>
<tr>
<td>Marital status</td>
<td>Married, Single, other</td>
<td>Status_M, Status_S, Status_O</td>
</tr>
<tr>
<td>Home telephone</td>
<td>Yes, No</td>
<td>Ht_Y, Ht_N</td>
</tr>
<tr>
<td>Commercial telephone</td>
<td>Yes, No</td>
<td>Ct_Y, Ct_N</td>
</tr>
<tr>
<td>Time in the present job</td>
<td>Until 24 months, 25 to 72, 73 to 127, More than 127</td>
<td>Tpj1, Tpj2, Tpj3, Tpj4</td>
</tr>
<tr>
<td>Salary</td>
<td>Up to US 283, 284 to 413, 414 to 685, 686 to 876, 877 to 1304, More than 1304</td>
<td>S1, S2, S3, S4, S5, S6</td>
</tr>
<tr>
<td>Quantity of loan parts</td>
<td>Up to 4, 5 to 6, 7 to 9, 10 to 12</td>
<td>Qlp1, Qlp2, Qlp3, Qlp4</td>
</tr>
<tr>
<td>First acquisition</td>
<td>Yes, No</td>
<td>Fa_Y, Fa_N</td>
</tr>
<tr>
<td>Time in the present home</td>
<td>Until 12 months, 13 to 24, 25 to 120, More than 120</td>
<td>Tph1, Tph2, Tph3, Tph4</td>
</tr>
<tr>
<td>Loan part value</td>
<td>Up to US 54, 55 to 70, 71 to 113, More than 113</td>
<td>Lpv1, Lpv2, Lpv3, Lpv4</td>
</tr>
<tr>
<td>Total loan value</td>
<td>Up to US 131, 132 to 174, 175 to 217, 218 to 348, 349 to 783, More than 783</td>
<td>Tlv1, Tlv2, Tlv3, Tlv4, Tlv5, Tlv6</td>
</tr>
<tr>
<td>Type of credit</td>
<td>Passbook, check</td>
<td>Tc_P, Tc_C</td>
</tr>
<tr>
<td>Age</td>
<td>Until 25 years, 26 to 40, 41 to 58, More than 58</td>
<td>Age1, Age2, Age3, Age4</td>
</tr>
<tr>
<td>Range of home ZIP Code</td>
<td>1 2 3 4 5</td>
<td>Hzip1, Hzip2, Hzip3, Hzip4, Hzip5</td>
</tr>
<tr>
<td>Range commercial ZIP Code</td>
<td>1 2 3 4 5</td>
<td>Czip1, Czip2, Czip3, Czip4, Czip5</td>
</tr>
<tr>
<td>Profession code</td>
<td>1 2 3 4 5 6 7</td>
<td>P1, P2, P3, P4, P5, P6, P7</td>
</tr>
<tr>
<td>Percent rate of part / salary</td>
<td>Up to 10%, 10.1% to 13.5%, 13.6% to 16.5%, 16.6% to 22.5%, More than 22.5%</td>
<td>Ps1, Ps2, Ps3, Ps4, Ps5</td>
</tr>
<tr>
<td>Percent rate of loan / salary</td>
<td>Up to 28%, 28.1% to 47.5%, 47.6% to 65%, More than 65%</td>
<td>Ls1, Ls2, Ls3, Ls4, Ls5</td>
</tr>
<tr>
<td>Type of client</td>
<td>1 = good, 0 = bad</td>
<td>Type</td>
</tr>
<tr>
<td>Variable</td>
<td>Estimated logistic coefficient (B)</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Gender_M</td>
<td>-0.314</td>
<td>0.053</td>
</tr>
<tr>
<td>Status_S</td>
<td>-0.1707</td>
<td>0.0556</td>
</tr>
<tr>
<td>Tpj1</td>
<td>-0.4848</td>
<td>0.0751</td>
</tr>
<tr>
<td>Tpj2</td>
<td>-0.2166</td>
<td>0.0608</td>
</tr>
<tr>
<td>Qlp1</td>
<td>1.6733</td>
<td>0.1006</td>
</tr>
<tr>
<td>Qlp2</td>
<td>0.9658</td>
<td>0.0743</td>
</tr>
<tr>
<td>Qlp3</td>
<td>0.3051</td>
<td>0.0679</td>
</tr>
<tr>
<td>Tph2</td>
<td>-0.3363</td>
<td>0.1003</td>
</tr>
<tr>
<td>Tph3</td>
<td>-0.1451</td>
<td>0.0545</td>
</tr>
<tr>
<td>Lpv1</td>
<td>-0.2035</td>
<td>0.0878</td>
</tr>
<tr>
<td>Tlv1</td>
<td>0.9633</td>
<td>0.1222</td>
</tr>
<tr>
<td>Tlv2</td>
<td>0.5915</td>
<td>0.1188</td>
</tr>
<tr>
<td>Tlv3</td>
<td>0.4683</td>
<td>0.0889</td>
</tr>
<tr>
<td>Tc_P</td>
<td>-1.34</td>
<td>0.0853</td>
</tr>
<tr>
<td>Age1</td>
<td>-0.7429</td>
<td>0.1371</td>
</tr>
<tr>
<td>Age2</td>
<td>-0.6435</td>
<td>0.0902</td>
</tr>
<tr>
<td>Age3</td>
<td>-0.2848</td>
<td>0.0808</td>
</tr>
<tr>
<td>Hzip1</td>
<td>-0.3549</td>
<td>0.1159</td>
</tr>
<tr>
<td>Czip1</td>
<td>-0.29</td>
<td>0.1014</td>
</tr>
<tr>
<td>Czip2</td>
<td>-0.2888</td>
<td>0.0642</td>
</tr>
<tr>
<td>Czip3</td>
<td>-0.2662</td>
<td>0.074</td>
</tr>
<tr>
<td>P1</td>
<td>0.3033</td>
<td>0.0945</td>
</tr>
<tr>
<td>P3</td>
<td>0.5048</td>
<td>0.0889</td>
</tr>
<tr>
<td>P5</td>
<td>0.4752</td>
<td>0.1048</td>
</tr>
<tr>
<td>P6</td>
<td>0.1899</td>
<td>0.0692</td>
</tr>
<tr>
<td>Ls1</td>
<td>0.2481</td>
<td>0.0824</td>
</tr>
<tr>
<td>Ls3</td>
<td>0.164</td>
<td>0.0664</td>
</tr>
<tr>
<td>Fa_N</td>
<td>-0.6513</td>
<td>0.0526</td>
</tr>
<tr>
<td>Constant</td>
<td>0.5868</td>
<td>0.0903</td>
</tr>
</tbody>
</table>
### TABLE 3
**Chi-Square test**

<table>
<thead>
<tr>
<th>-2LL</th>
<th>Chi-Square (improvement)</th>
<th>Degrees of freedom</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>11090.355</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9264.686</td>
<td>1825.669</td>
<td>28</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### TABLE 4
**Neural network statistics**

<table>
<thead>
<tr>
<th>Obtained statistics</th>
<th>Test</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misclassification of cases</td>
<td>0.306</td>
<td>0.323</td>
</tr>
<tr>
<td>Mean error</td>
<td>0.576</td>
<td>0.619</td>
</tr>
<tr>
<td>Mean square error</td>
<td>0.197</td>
<td>0.211</td>
</tr>
<tr>
<td>Degrees of freedom of the model</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>Degrees of freedom of the error</td>
<td>7780</td>
<td></td>
</tr>
<tr>
<td>Total degrees of freedom</td>
<td>8000</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 5
**Classification results**

#### Logistic regression

<table>
<thead>
<tr>
<th>Training Predicted →</th>
<th>Validation Predicted →</th>
<th>Test Predicted →</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed ↓</td>
<td>Bad</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>2833</td>
<td>1167</td>
</tr>
<tr>
<td>Good</td>
<td>1294</td>
<td>2706</td>
</tr>
<tr>
<td></td>
<td>4127</td>
<td>3873</td>
</tr>
<tr>
<td></td>
<td>2111</td>
<td>889</td>
</tr>
<tr>
<td></td>
<td>1078</td>
<td>1922</td>
</tr>
<tr>
<td></td>
<td>3189</td>
<td>2811</td>
</tr>
<tr>
<td></td>
<td>2159</td>
<td>841</td>
</tr>
<tr>
<td></td>
<td>1059</td>
<td>1941</td>
</tr>
<tr>
<td></td>
<td>3218</td>
<td>2782</td>
</tr>
<tr>
<td>Total</td>
<td>4127</td>
<td>3873</td>
</tr>
</tbody>
</table>

#### Neural network

<table>
<thead>
<tr>
<th>Training Predicted →</th>
<th>Validation Predicted →</th>
<th>Test Predicted →</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed ↓</td>
<td>Bad</td>
<td>Good</td>
</tr>
<tr>
<td></td>
<td>2979</td>
<td>1021</td>
</tr>
<tr>
<td>Good</td>
<td>1430</td>
<td>2570</td>
</tr>
<tr>
<td></td>
<td>4409</td>
<td>3591</td>
</tr>
<tr>
<td></td>
<td>2236</td>
<td>764</td>
</tr>
<tr>
<td></td>
<td>1177</td>
<td>1823</td>
</tr>
<tr>
<td></td>
<td>3413</td>
<td>2587</td>
</tr>
<tr>
<td></td>
<td>2255</td>
<td>745</td>
</tr>
<tr>
<td></td>
<td>1193</td>
<td>1807</td>
</tr>
<tr>
<td></td>
<td>3448</td>
<td>2552</td>
</tr>
</tbody>
</table>
### TABLE 6

**Comparison indexes**

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>47.9</td>
<td>45.1</td>
<td>46.6</td>
</tr>
<tr>
<td>Neural network</td>
<td>47.9</td>
<td>45.3</td>
<td>45.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>38</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>Neural network</td>
<td>39</td>
<td>35</td>
<td>35</td>
</tr>
</tbody>
</table>

### TABLE 7

**Classification precision of the models for credit analysis**

<table>
<thead>
<tr>
<th>Model</th>
<th>Linear regression</th>
<th>Logistic regression</th>
<th>Classification trees</th>
<th>Linear Programming</th>
<th>Neural networks</th>
<th>Genetic algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henley (1995)</td>
<td>56.6</td>
<td>56.7</td>
<td>56.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Boyle (1992)</td>
<td>77.5</td>
<td>-</td>
<td>75.0</td>
<td>74.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Srinivisan (1987)</td>
<td>87.5</td>
<td>89.3</td>
<td>93.2</td>
<td>86.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yobas (1997)</td>
<td>68.4</td>
<td>-</td>
<td>62.3</td>
<td>-</td>
<td>62.0</td>
<td>64.5</td>
</tr>
<tr>
<td>Desai (1997)</td>
<td>66.5</td>
<td>67.3</td>
<td>67.3</td>
<td>-</td>
<td>64.0</td>
<td>-</td>
</tr>
</tbody>
</table>

Source: Thomas (2000: 159)

### TABLE 8

**Classification precision of the models for credit analysis (consulted literature)**

<table>
<thead>
<tr>
<th>Source</th>
<th>Logistic regression</th>
<th>Classification trees</th>
<th>Neural networks</th>
<th>Genetic algorithm</th>
<th>Discriminant analysis</th>
<th>REAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fritz and Hosemann (2000)</td>
<td>79.5</td>
<td>-</td>
<td>81.6</td>
<td>82.4</td>
<td>82.7</td>
<td></td>
</tr>
<tr>
<td>Arraes et al. (1999)</td>
<td>84.8</td>
<td>-</td>
<td>85.4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Chen et al. (2002)</td>
<td>91.9</td>
<td>-</td>
<td>92.9</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Nanda and Pendharkar (2001)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>65.0</td>
<td>62.5</td>
<td></td>
</tr>
<tr>
<td>Ohtoshi (2003)</td>
<td>83.5</td>
<td>73.9</td>
<td>85.0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Picinini et al. (2003)</td>
<td>63.5</td>
<td>-</td>
<td>64.4</td>
<td>67.5</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Armingher et al. (1997)</td>
<td>67.6</td>
<td>66.4</td>
<td>65.2</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Huang et al. (2004)</td>
<td>77.0</td>
<td>-</td>
<td>80.0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Semolini (2002)</td>
<td>68.3</td>
<td>-</td>
<td>67.4</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Rosa (2000)</td>
<td>70.4</td>
<td>66.6</td>
<td>-</td>
<td>-</td>
<td>71.4</td>
<td></td>
</tr>
</tbody>
</table>