LII Simpósio Brasileiro de Pesquisa Operacional
João Pessoa - PB, 3 a 5 de novembro de 2020

# Traditional and Machine Learning Methods for Credit Scoring: an introduction

**Raphael Baseggio Corrêa**
Escola de Administração, Universidade Federal do Rio Grande do Sul
R. Washington Luiz, 855 - Porto Alegre,Brasil
bc_raphael@hotmail.com

**Tiago Pascoal Filomena**
Escola de Administração, Universidade Federal do Rio Grande do Sul
R. Washington Luiz, 855 - Porto Alegre,Brasil
tpfilomena@ea.ufrgs.br

**ABSTRACT**

In recent decades, consumer credit has experienced phenomenal growth largely due to accurate automated credit scoring models that aided lenders decisions. The predictive power of credit scoring models is vital to the profitability of financial institutions, as they help to prevent losses and to maximize profits. Nowadays, there are a large variety of credit scoring models utilized to address the problem of credit risk estimation. Thus, this paper proposes to explore the fundamental literature and provide an overview of the credit scoring context and methods. Some industry standard techniques, the mostly used classification models and a popular ensemble approach are presented, in a way that their strengths, weaknesses and relevant details are highlighted. Machine learning models demonstrate better results than the statistical techniques. Also, recent studies have been massively researching the use of ensemble methods, which consistently show superior predictive performance over the single classifiers in terms of average accuracy.

**KEYWORDS. Credit Scoring. Machine Learning. OR in Banking.**

**Credit Scoring. Credit Risk Estimation. Machine Learning. OR in Banking. Statistical Methods. Binary Classification. Literature Review. Support Vector Machine (SVM). Neural Networks. Random Forest. Logistic Regression. Linear Discriminant Analysis.**

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

## 1. Introduction

It is well known the important role credit plays in any modern society, remarkably, in recent years, that credit risk assessment has drawn considerable attention from managers at financial institutions worldwide [Ala'raj e Abbod, 2016; Harris, 2015]. The concern about credit and other banking risks has led to credit scoring, which is not only a set of decision models used to support lenders on granting loans, deciding who will get credit, how much credit they should get and the price they should pay for it, but also a methodology that allows banks to estimate the credit risk of their consumer portfolios [Thomas et al., 2017] and develop lending strategies to optimize their profits [Bellotti e Crook, 2009]. It is the use of statistical models to transform relevant data into a numerical measure, closely related to the estimated credit risk [Abdou e Pointon, 2011]. This credit risk, in turn, must be reflected in the loan price [Van Gestel e Baesens, 2009], which means that the accuracy of credit scoring models is vital to the profitability of financial institutions [Wang et al., 2012].

The behavior and performance of past and current customers can provide essential historical data, which can be very useful to predict the behavior of future applicants [Abdou e Pointon, 2011]. When assessing a credit grant, lenders typically consider variables which indicates consumer's stability (e.g., time at present employment), financial sophistication (e.g., having credit cards), resources (e.g., residential status), possible outgoings (e.g., number of dependents), and obviously, the most powerful ones: variables related to previous defaults/late payments, and borrower's current status on other loans [Thomas et al., 2017]. On the other hand, as affirm Lessmann et al. [2015], corporate risk models employ data from balance sheets, financial ratios, or macroeconomic indicators. In this sense, considering the differences between the types of variables, consumer credit scoring seems to be a more challenging environment than corporate [Lessmann et al., 2015]. For this reason, the focus of this work is on consumer credit.

Consumer credit has experienced phenomenal growth over the last decades, largely due to accurate automated credit scoring models that aided lenders decisions [Van Gestel e Baesens, 2009]. Many statistical methods have been applied, including some "industry standards" such as discriminant analysis and logistic regression [Hand e Henley, 1997]. The latter is still considered the most usual in the credit business [see, Lessmann et al., 2015], but recently, non-linear approaches and machine learning techniques like artificial neural networks and support vector machine have been applied to credit scoring [Harris, 2015]. The constant development of new computers capacities and technologies at lower costs combined with a growing variety of data sources and formats has enlarged the use of the so called Big Data [Müller et al., 2016]. Moreover the current interest in Big Data can be revealing changes that would lead to a major acceptance of advanced scoring methods [Lessmann et al., 2015]. Yet with the progress of machine learning algorithms and the great volume of multi-dimensional customer data, building credit scoring models using machine learning methods has become a hot topic for research [Bao et al., 2019].

Thus, the main goal of this introductory research is to cover and provide fundamental literature on credit scoring, especially about the very classical techniques used in this area and the more recent and important approaches which are already or have the potential to be benchmarks. To highlight the strengths, weaknesses and the most relevant details of each one of the methods covered is another goal of this paper, in a way that the reader is presented to an overview of the subject and its context. Also, it is understood that the topics explored in this work will serve as the foundation for future research.

## 2. Literature Review

In regards to the literature review, it is worth to say that due to the vast amount of credit scoring approaches, derivations that create families of models, and the many methods of estimat-

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

ing parameters, the focus will be on the industry standard techniques (i.e., logistic regression and discriminant analysis) [see, Hand e Henley, 1997], the mostly used classification models in the literature (i.e., artificial neural networks and support vector machine) [see, Ala'raj e Abbod, 2015], and a popular ensemble approach (i.e., random forest), as ensembles have already shown to be superior than single classifiers.

## 2.1. Traditional methods

### 2.1.1. Logistic Regression

Logistic Regression (LR) is one derivation of the Generalized Linear Models and works similarly to linear regression. In LR the dependent variable is categorical and its relationship with the independent variables is measured by calculating probabilities through a logistic function [Cox, 1958]. As in linear regression, LR also calculates the coefficients' values. While the first uses the least squares approach, LR uses the Maximum Likelihood method to get estimates for the coefficients' values [Thomas et al., 2017]. For binary classification, consider $y_i \in \{0, 1\}$ for all $i = 1, \ldots, n$, and some given example $\mathbf{x}$.

LR uses the following model to estimate the probability that the label is 1 given $\mathbf{x}$,

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w} \cdot \mathbf{x} - b)} \tag{1}$$

The parameters $\mathbf{w}$ and $b$ are both estimated by maximum likelihood through maximizing the log-likelihood function, with respect to $\mathbf{w}$ and $b$ [Bellotti e Crook, 2009],

$$L(\mathbf{x}_1, \ldots, \mathbf{x}_n \mid \mathbf{w}, b) = \sum_{i=1}^{n} y_i \log p_i + (1 - y_i) \log(1 - p_i) \tag{2}$$

Where $p_i = P(y = 1 \mid \mathbf{x}_i)$.

Still in the words of Thomas et al. [2017], the main difference here between linear regression and LR is that linear regression is trying to fit the probability $p$ of being a "Good" (customer) by a linear combination of the independent variables, while LR tries to fit $\log(p/(1-p))$ by a linear combination of the independent variables. In practice, both techniques present similar results [Hand e Henley, 1997], but the one considered as the industry standard is Logistic Regression [Lessmann et al., 2015].

### 2.1.2. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a model which tries to find a linear combination of the independent variables that maximizes a separability criterion [Hand e Henley, 1997]. Wiginton [1980] states that one of the main issues in discriminant analysis is that the attributes measurements are assumed to "arise from multivariate normal populations such that the classes have identical covariance matrices, differing only in the value of their mean vectors ". However, this is not all true as the normality assumption is important only if significance tests are performed [Hand e Henley, 1997]. Hence, even though a significant part of credit data does not follow a normal distribution, it may not be a critical limitation [Reichert et al., 1983]. Probably the main limitations of the linear models are their restrictions on modelling non-linearities and unobserved heterogeneity in the dataset [Jones et al., 2015]. In order to avoid multicollinearity issues when dealing with high-dimensional datasets, the correlation-based feature selection method described in Hall [2000] can be used, reducing the number of variables in the preprocessing phase [Lessmann et al., 2015].

A variation of linear discriminant method is Quadratic Discriminant Analysis (QDA) which separates the disjoint regions corresponding to each class by quadratic boundaries [Friedman,

1989]. Both LDA and QDA estimate the parameters of a normal distribution, (i.e., class-dependent mean vectors and covariance matrices) using the maximum likelihood approach [Lessmann et al., 2015].

## 2.2. Machine Learning methods

### 2.2.1. Support Vector Machine

Support Vector Machine (SVM) is among the most popular and accurate algorithms, finding applications in several fields [Zhang et al., 2017]. It is a classification model which separates classes of data by a hyperplane, such that the margin width between these classes is maximized [Bellotti e Crook, 2009]. The idea was presented by Cortes e Vapnik [1995], as a method that non-linearly maps variables to a very high-dimension feature space, where a linear decision surface is then constructed. The technique exploits the fact that, in many problems, a non-linear classification task in the original input space can be transformed into linear in the high-dimensional feature space [Bhattacharyya et al., 2011].

As Zhang et al. [2017] state, SVM consists in selecting a small number of critical boundary samples from each class and then building a linear discriminant function that separates them as widely as possible. The points lying on the boundaries, also called training examples or training data, are referred to as support vectors [Harris, 2015], since they "support" the optimal separating hyperplane [Bellotti e Crook, 2009]. An optimal separating hyperplane is defined as the unique one which separates the training data with a maximal margin width, i.e., the linear decision function with maximal margin between the support vector of each class [Cortes e Vapnik, 1995]. Statistical learning theory shows that the margin width maximization leads to a reduction of the model complexity, therefore minimizing the expected general risk of error [Bellotti e Crook, 2009].

According to Bhattacharyya et al. [2011], the classification function, which is a hyperplane separating different classes of data, can be specified in terms of the dot products of the input samples:

$$\langle w, x \rangle + b = 0 \tag{3}$$

Where $b$ is a threshold term [Bellotti e Crook, 2009], and the notation $\langle w, x \rangle$ means the dot product of the coefficient vector $w$ and the vector variable $x$ [Bhattacharyya et al., 2011]. The authors also affirm that the solution to a classification problem is given by the coefficient vector $w$, which is a linear combination of the data points $x_i$, i.e.,

$$w = \sum_{i=1}^{m} \alpha_i x_i \tag{4}$$

Where $\alpha_i \geq 0$, and the data points $x_i$ with non-zero $\alpha_i$ are the support vectors.

The dual representation of the classification function in (4) takes the following form:

$$\sum_{i} \alpha_i y_i \langle x_i, x \rangle + b = 0 \tag{5}$$

Where $y_i$ are the classification labels of the input data points.

However sometimes it is not possible to separate the data by a hyperplane, which is very common in real-world classification problems, thus a soft margin is used [Bellotti e Crook, 2009]. Still in the words of Bellotti e Crook [2009], in these cases, slack variables are used to allow misclassifications, but they take a penalty proportional to how far they are on the wrong side. Therefore,

the optimization problem still seeks to maximize the margin width, but also to minimize the sum of the penalties, while a parameter $C$ controls the relative cost of each goal [Bellotti e Crook, 2009]. Thus, setting a proper value for $C$ is a crucial point to the performance of SVM [Bhattacharyya et al., 2011]. The "standard" model is a convex optimization problem (quadratic criterion and linear inequality constraints) and is typically solved using the Lagrangian [Hastie et al., 2009].

There are many ways to build the decision surface by interchanging the decision function, for instance, rather than first non-linearly transform the input vectors followed by dot-products with support vectors in feature space, it is possible to initially compare two vectors in the input space by taking their dot-product (or some distance measure), and then make a non-linear transformation of the value of the result [Cortes e Vapnik, 1995]. The functions which take as inputs the vectors in their original space and return the dot product of these vectors in the feature space are called kernel functions and the most common types of them are the Linear models, Polynomials of arbitrary degree and the Gaussian radial basis function [Bellotti e Crook, 2009]. Similarly, when a kernel function is used, the classification function can be represented as the dot products of projections of input data points in a high-dimensional feature space [Bhattacharyya et al., 2011]. An important consideration about the kernel functions is that, as stated by Bellotti e Crook [2009], using non-linear kernels means that it is not feasible to extract an explicit scorecard, although predictions can be made using them. A kernel function $k$ can be written as [Bhattacharyya et al., 2011]:

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle \tag{6}$$

Where $\phi \colon X \to H$ is a mapping of points in the input space $X$ into a higher-dimensional space $H$.

Using a kernel function $k$, the function in (5) in the high-dimensional space $H$ can be expressed as:

$$\sum_i \alpha_i y_i k(x_1, x) + b = 0 \tag{7}$$

Finally, as the optimal separating hyperplane is the classification function which separates the classes with the maximal margin width, Bhattacharyya et al. [2011] says that the coefficients $\alpha_i$ of the best classification function can be found by solving the optimization problem that follows:

$$\text{Maximize } W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{8}$$

Subject to

$$0 \leq \alpha_i \leq \frac{c}{m}, \quad i = 1, \ldots, m \tag{9}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{10}$$

Where $\alpha_i$ is a Lagrange multiplier for each training example $i$ [Bellotti e Crook, 2009] and $c$ is the cost parameter that represents the penalty applied for incorrect classifications of the input data points [Bhattacharyya et al., 2011].

One of the main concerns about SVM is its computational cost, as explains Harris [2015], it is well known that as historical credit scoring datasets get large, the non-linear approach, despite

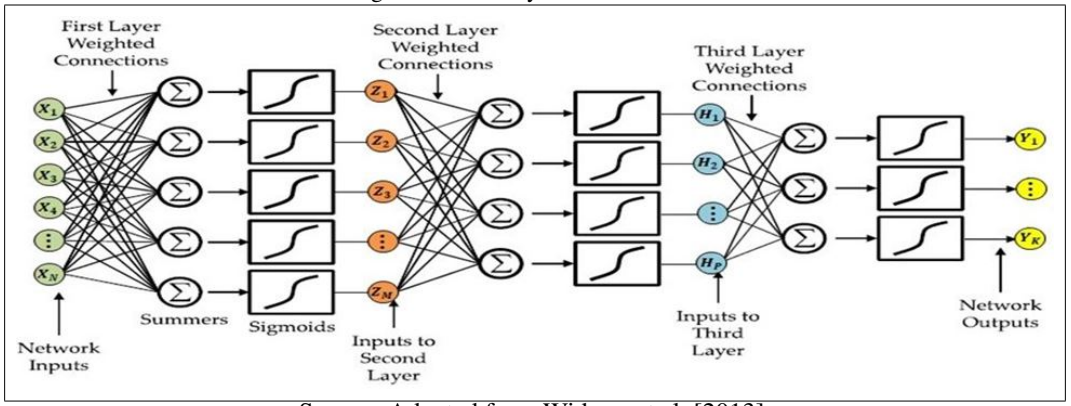
of highly accurate, become computationally expensive, notably in comparison to traditional statistical methods. The ability to model non-linearities is pointed out as a strength but also SVM's main weakness, due to the complex mathematical models produced and their limited comprehensibility, even they may present high accuracy when compared to other data mining techniques [Martens et al., 2007]. In this sense, Harris [2015] complements saying that, like the artificial neutral networks, SVM is sometimes criticized for being a "black box", especially because some classifiers are hard to interpret due to their vast dimensionality and the complexity of the estimated relationships, being difficult to extract from the model, for example, the reason why a loan was rejected [Thomas et al., 2017].

### 2.2.2. Neural Networks

Neural Networks (NN) are non-linear statistical models whose name is inspired in the way that the human brain operates [Hastie et al., 2009]. The idea was first formalized in the work of McCulloch e Pitts [1943]. Also known as Artificial Neural Networks (ANN), it consists of several units, representing the "neurons", linked together via weighted connections, representing the "dendrites" [Másson e Wang, 1990]. The authors also affirm that, via weighted incoming connections, each unit receives an input value and sends a signal to all units which they are linked to.

According to Schmidhuber [2015], NN can be also represented as a graph of nodes (or vertices or neurons) and edges (or connections) between the nodes, where the first layer is the set of input nodes and then some weight is assigned to each edge. Taking as an example the well-known network Multilayer Perceptron (MLP), which is a feed-forward neural network consisted of an input layer, some number of hidden layers and one output layer [Luo et al., 2017], as Hastie et al. [2009] explain, supposing $X_n$, $n = 1, \ldots, N$ is the set of input nodes, the nodes from the next layers $Z_m$, $m = 1, \ldots, M$ and $H_p$, $p = 1, \ldots, P$, until the output nodes $Y_k$, $k = 1, \ldots, K$, are modelled as a function $f(x)$, called the activation function, of all nodes linked to each one from the previous layer. In other words, each single input is a component of the input vector, which is then applied to the nodes of the first layer [Widrow et al., 2013]. Still conforming to the authors, the sigmoidal outputs of the first layer comprise an input vector for the second layer, and so forth. The neurons in the middle of the network, i.e., $Z_m$ and $H_p$ are called hidden units, as their values are not directly observed, and compose the hidden layers [Hastie et al., 2009]. The authors also state that the most commonly applied activation function is the sigmoid, but sometimes Gaussian radial basis functions are used, producing what is known as a radial basis function network. An example of a three layer feed-forward neural network, is illustrated in Figure 1.

Figure 1: Three-layer Neural Network




Source: Adapted from Widrow et al. [2013]

Consider the use of the traditional sigmoid function in the hidden layers:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

Using the notation from Figure 1, Luo et al. [2017] demonstrates that an input node $X_n$ sends signal for neuron $Z_m$ with weight $w_{nm}^1$, for $n = 1, \ldots, N$, $m = 1, \ldots, M$, and where 1 represents the index of the layer. Thus, the activation of neuron $Z_m$ is

$$f(Z_m) = f\left(\sum_{n=1}^{N} X_n w_{nm}^1 + b_m\right) \tag{12}$$

Where $b_m$ is a bias value. As the authors complement, the parameters for the network are determined iteratively, usually through a back-propagation learning algorithm.

The variety of neural network models is quite large. As Zhang et al. [2017] present, for instance, the Deep Neural Networks consist of a large multilayer neural network, which are usually designed as feed-forward network trained with the standard back-propagation algorithm. The Convolutional Neural Networks are mostly a deep neural network compounded by one or more pairs of convolutional layers and max-pooling layers, followed by one or more fully connected hidden layers and whose goal is to find a set of locally connected neurons. A Deep Belief Network (DBN) is a type of deep neural network that uses Restricted Boltzmann Machines to derive better initial parameters. Also, in the context of a network with several layers, emerges a class of them called Deep Learning (DL), in which the "deep" is related to the number of layers of the network [Zhang et al., 2017]. As the authors affirm, Deep Learning is a very promising area in Machine Learning research, as it has demonstrated an outstanding performance in image and speech recognition, as well as other related applications.

### 2.2.3. Random Forest

Random Forest (RF) is an ensemble learning approach which combines the predictions of several decision trees, each one outputting a vote, to build the final prediction of the algorithm, that will be the mode of the votes [Zhang et al., 2017]. The method has gained popularity from the work of Breiman [2001]. A simple random forest with random features (e.g., original inputs are randomly selected or some linear combination of them is done at random) is generated by randomly selecting, at each node, an input variable or just a few of them to split on at each iteration and grow the tree, without pruning, until it gets to a leaf node, i.e., a prediction [Breiman, 2001].

In order to make the base classifiers more diverse among each other, whenever a tree is divided, RF randomly samples independent variables and finds the best split among the chosen subsample. The subsample's size (i.e., number of independent variables from which a split is selected) is a meta-parameter of RF [Lessmann et al., 2015]. In addition, as state Bhattacharyya et al. [2011], ensemble schemes such as RF perform better when members are dissimilar. The authors also affirm that RF obtains diversity among its individual trees (i.e., base classifiers) from two sources of randomness: *i*) each tree is developed using different bootstrapped instances (with replacement) of the training data; and *ii*) when each tree is built, only a subset of the attributes, selected at random, is considered at each node. A relevant issue which emerges from that is how many features should be selected at each node, where Breiman [2001] answers that choosing one or two features at each node yields results close to the optimal.

The strength and the dependence between the individual trees directly influence RF accuracy [Breiman, 2001]. The author explains that, to enhance accuracy, the randomness added has

to minimize the correlation while holding the strength, and concludes saying that both correlation and strength can be estimated through out-of-bag techniques, what gives useful internal estimates to better understand the classification accuracy and how to improve it. Even though, RF is considered one of the most accurate classification methods and is commonly used as a default classifier [Zhang et al., 2017]. Moreover, as the authors affirm, RF corrects the overfitting problem, commonly found in decision trees with their training sets, and can deal with categorical and unbalanced data (number of instances of each class in the dataset is very different), as well as data with missing values, what seems to be an advantage over some other approaches.

### 2.3. Related Works

Ensemble models (i.e., multiple classifiers), more specifically heterogeneous ensembles, have already proven to achieve better predictive performance than single classifiers and even homogeneous ensembles in terms of average accuracy [Ala'raj e Abbod, 2015; Lessmann et al., 2015]. Methods to handle unbalanced data [Marqués et al., 2013] and feature selection [Bellotti e Crook, 2009] have already shown promising results as well, and also some room for improvements. The variety of models and techniques is large, so this section presents some of the latest methods used and the results found by recent related studies.

A comparative study carried out by Zhang et al. [2017] was proposed to examine the state-of-the-art classification algorithms on 71 datasets. The algorithms considered in that work were the newly proposed Extreme Learning Machine (ELM), Sparse Representation based Classification (SRC), and Deep Learning (DL); and the established ones Support Vector Machine (SVM), Random Forest (RF), Naive Bayes classifier (NB), C4.5, Stochastic Gradient Boosting Decision Trees (GBDT), AdaBoost (AB), K Nearest Neighbors classifier (KNN) and Logistic Regression (LR). The authors found that GBDT and RF presented the best performance, followed by SVM and ELM. The worst overall classification algorithms, in terms of accuracy (ACC), were AB and DL, with few exceptions in some datasets. Considering only the results from the German Credit dataset, widely reported in credit scoring literature, GBDT tied with NB were the best ones, followed by RF, C4.5 and DL, also tied with the same prediction accuracy.

The research executed by Lessmann et al. [2015] compared 41 classifiers in terms of six performance measures on eight real-world credit scoring datasets. The algorithms were divided into three groups: individual classifiers, homogeneous ensemble classifiers, and heterogeneous ensembles. The results showed the overall superiority of heterogeneous ensembles classifiers over both homogeneous ensembles and individual classifiers. The first 11 ranks are occupied by heterogeneous ensembles, with HCES-Bag (Hill-Climbing Ensemble Selection with Bootstrap Sampling) topping them. The 12th place belongs to an homogeneous ensemble classifier, RF (Random Forest), while the best individual classifier, ANN (Artificial Neural Network), was just the number 14 in the ranking, performing few better than the industry standard LR (Logistic Regression), ranked 16. Furthermore, the authors found that sophisticated methods do not necessarily mean better accuracy. For instance, ELM (Extreme Learning Machine) and RotFor (Rotation Forest) are extensions for ANN and RF, respectively, but neither performed better than their ancestors.

In the work of Bao et al. [2019], an ensemble strategy that integrates unsupervised and supervised learning in two different stages (building consensus model and clustering datasets) is proposed. The ensemble strategy combines SOM (Self-Organizing Maps) and seven base learners, namely ANN, DT, GBDT, kNN, LR, RF and SVM. Three credit datasets were used, and performance comparisons were based in four possible routes: individual models, individual models + consensus model, clustering + individual models, clustering + individual models + consensus model. The authors employed seven performance measures to evaluate model's predictive power,

and emphasized the use of MMC as evaluation criterion, since it considers the confusion matrix more comprehensively and is more efficient to reflect performances with unbalanced datasets. As a result, GBDT performed the best among all individual classifiers, both in MMC, ACC and area under the ROC curve (AUC). Right behind came ANN, SVM and RF. The models kNN and DT showed the best Type I error and Type II error, respectively, but also a poor performance in terms of MCC. In relation to the strategy proposed in the paper, the authors found that both cluster-based models and SOM cluster-based consensus model, in most cases, help to improve the performance of credit scoring models. Yet, integrating the two stages, clustering method and consensus model, achieved the best overall performance.

Ala'raj e Abbod [2016], proposed a new combination method based on measurement-level predictions, called consensus approach. The presented method aims to make classifier ensembles work as a team to interact and cooperate until all classifiers come to an agreement on a final optimal decision. The authors considered five well-known classifiers to test, namely NN, SVM, DT, RF and NB, and two other classifiers as benchmark: LR and MARS (Multivariate Adaptive Regression Splines). The tests were performed on five real-world financial datasets (German, Australian, Japanese, Iranian and Polish), and the predictive performance was evaluated using four measures. Also, the proposed approach was compared to other seven combination rules found in the credit scoring literature. The findings showed that the proposed consensus approach attained the best performance across all four performance measures and almost all datasets, usually followed by RF. The authors also found that NB performs very poorly in unbalanced datasets, regardless of the measure used.

Nanni et al. [2015] carried out an extensive study aiming to find general-purpose (GP) heterogeneous ensembles that perform competitively across several datasets and require little or no parameter tuning. The authors used 25 datasets to test 5 classifiers, namely SVM, Gaussian Process Classifiers (GPC), Random Subspace of AdaBoost (RS_AB), Random Subspace of Rotation Boosting (RS_RB), and Deep Learning (DL). They also tested a Random Subspace ensemble with SVM as the classifier (RS-SVM), and an ensemble composed of three versions of Deep Learning (S_D), which is the result of $(DL1 + DL2 + DL3)/3$. From these classifiers, they proposed the following heterogeneous ensembles: *i*) E1: GPC + RS_AB; *ii*) E2: GPC + RS_AB + RS_RB; *iii*) E3: GPC + RS_AB + RS_RB + RS-SVM; *iv*) E4: GPC + RS_AB + RS_RB + RS-SVM + S_D. The models were evaluated using the AUC, and for multi-class cases, AUC was calculated using the one-versus-all approach. As a result, the authors show that E2, E3 and E4 performed better than any other model, even with a simple fusion rule (i.e., the sum rule).

The paper of Xia et al. [2018] presents a novel heterogeneous ensemble credit model that integrates the bagging algorithm with the stacking method, the bstacking approach. The classifiers considered in this work were SVM, RF, GPC (Gaussian Process Classifier), and XGBoost (Extreme Gradient Boosting). The authors included LR and DT for comparison purposes, but these individual classifiers were not considered as base learners in the suggested bstacking framework. Four real-world credit datasets were used (German, Australian, P2P-We.com, and Lending Club), and four performance measures were employed to evaluate model's predictive power. Individual classifiers (i.e., LR, GPC, SVM, and DT) were first compared, with GPC showing the best results across all datasets and performance measures. Behind GPC came SVM, then LR and finally DT. In order to test homogeneous ensembles, the authors considered the following models: RF, XGBoost, Bag-SVM and Bag-GPC, being the last two the ensemble versions of SVM and GPC models, respectively, using bagging algorithm. As a result, XGBoost performed slightly better than Bag-GPC in the German dataset, followed by Bag-SVM in third and RF last. Considering the Aus-

tralian dataset, XGBoost again proved to be the best, but this time, followed by RF, then Bag-GPC and Bag-SVM last, depending on the performance measure under analysis. Concerning the P2P-We.com and Lending Club datasets, RF dominated in all performance measures, with Bag-GPC a little better than XGBoost and Bag-SVM. Finally, to compare heterogeneous ensembles, the authors used models that comprise individual and homogeneous ensemble classifiers and variants of ensemble selection strategies and fusion methods. They found that the proposed model (i.e., bstacking) performs significantly better than most benchmark models.

Papouskova e Hajek [2019] suggested a two-stage credit risk model that integrates class-unbalanced ensemble learning and regression ensemble. In the first stage, to handle the problem of unbalanced data, the authors used the evolutionary undersampling algorithm based on ensembles EUSBoost, and compared the performance against the oversampling method SMOTEBagging. In addition, a multi-objective evolutionary feature selection (MOEFS) was employed aiming to increase the predictions accuracy. An heterogeneous ensemble was built in both stages of the proposed model, using a stacking method to combine the base learners' predictions through a meta-learning algorithm (i.e., an algorithm trained on the outputs from the base classifiers). In this case, two meta-learning algorithms were analyzed, LR (Logistic Regression) and RF (Random Forest) and compared against Majority Voting and Average Voting. As base learners, the authors used the most applied in the credit scoring literature, such as SVM, C4.5, LR, among others, and some homogeneous ensembles like RF, Rotation Forest, DECORATE, and Random Subspace. The models were evaluated by three different measures. The comparisons were executed on two datasets. As a result, in the first set of experiments, the authors demonstrated the superiority of EUSBoost+RF over SMOTEBagging+RF and the plain RF (with unbalanced data) for all datasets and performance measures. Regarding the classification performance of all PD models tested in the study, the results showed that Stacking with RF as the meta-learning algorithm performed the best on both datasets and all evaluation measures, followed by Stacking with LR, RF, and Rotation Forest with similar performance in terms of ACC and AUC.

## 3. Final Remarks

This introductory research was proposed to explore the fundamental literature and provide an overview of the credit scoring context and the most important methods used in practice and in the literature. In order to achieve these goals, some industry standard models, machine learning methods and a very popular ensemble technique were presented, highlighting their strengths, weaknesses and essential details. Many models have been developed based on the classical statistical methods or machine learning techniques. The latter have already demonstrated better results than the statistical models when dealing with credit scoring problems. Beyond that, recent studies have largely researched the use of ensemble methods, which consistently show superior predictive performance over the single classifiers in terms of average accuracy. The number of studies that have been developing credit scoring models grows constantly, and it seems that the research tendency is going towards hybrid and ensemble modeling, due to their efficiency in handling the credit scoring problems.

Therefore, the credit scoring research, which attempts to find more efficient and effective algorithms [Harris, 2015], seems to be a very promising area plenty of opportunities, as no best overall technique have been found yet. To complete, it is a field where many methodologies from operations research and statistics can be employed, and even a small gain in performance can become a great improvement for the business.

**Acknowledgements**

The first author acknowledges the financial support of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

## References

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

Ala'raj, M. e Abbod, M. F. (2015). A systematic credit scoring model based on heterogeneous classifier ensembles. In *2015 International Symposium on Innovations in Intelligent Systems and Applications* $INISTA$, p. 1 – 7.

Ala'raj, M. e Abbod, M. F. (2016). Classifiers consensus system approach for credit scoring. *Knowledge-Based Systems*, 104:89 – 105. ISSN 0950-7051.

Bao, W., Lianju, N., e Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications*, 128:301 – 315.

Bellotti, T. e Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2, Part 2):3302 – 3308. ISSN 0957-4174.

Bhattacharyya, S., Jha, S., Tharakunnel, K., e Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602 – 613.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5 – 32. ISSN 1573-0565.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273 – 297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B* $Methodological$, 20(2):215 – 232.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165 – 175.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In Langley, P., editor, *Proceedings of the Seventeenth International Conference on Machine Learning* $ICML'00$, p. 359 – 366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hand, D. J. e Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523 – 541.

Harris, T. (2015). Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 42(2):741 – 750. ISSN 0957-4174.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

Jones, S., Johnstone, D., e Wilson, R. (2015). An empirical evaluation of the performance of binary classifiers in the prediction of credit ratings changes. *Journal of Banking & Finance*, 56:72 – 85.

Lessmann, S., Baesens, B., Seow, H.-V., e Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1): 124 – 136. ISSN 0377-2217.

Luo, C., Wu, D., e Wu, D. (2017). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence*, 65:465 – 470. ISSN 0952-1976.

Marqués, A. I., García, V., e Sánchez, J. S. (2013). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, 64(7):1060 – 1070.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Martens, D., Baesens, B., Gestel, T. V., e Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466 – 1476. ISSN 0377-2217.

Másson, E. e Wang, Y.-J. (1990). Introduction to computation and learning in artificial neural networks. *European Journal of Operational Research*, 47(1):1 – 28. ISSN 0377-2217.

McCulloch, W. S. e Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115 – 133. ISSN 1522-9602.

Müller, O., Junglas, I., vom Brocke, J., e Debortoli, S. (2016). Utilizing big data analytics for information systems research: challenges, promises and guidelines. *European Journal of Information Systems*, 25(4): 289 – 302.

Nanni, L., Brahnam, S., Ghidoni, S., e Lumini, A. (2015). Toward a general-purpose heterogeneous ensemble for pattern classification. *Computational Intelligence and Neuroscience*, 2015:1 – 10.

Papouskova, M. e Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems*, 118:33 – 45. ISSN 0167-9236.

Reichert, A. K., Cho, C.-C., e Wagner, G. M. (1983). An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business & Economic Statistics*, 1(2):101 – 114.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117. ISSN 0893-6080.

Thomas, L., Crook, J., e Edelman, D. (2017). *Credit Scoring and Its Applications, Second Edition*. Mathematics in Industry. Society for Industrial and Applied Mathematics.

Van Gestel, T. e Baesens, B. (2009). *Credit Risk Management: Basic concepts: financial risk components, rating analysis, models, economic and regulatory capital*. Oxford University Press.

Wang, G., Ma, J., Huang, L., e Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-Based Systems*, 26:61 – 68. ISSN 0950-7051.

Widrow, B., Greenblatt, A., Kim, Y., e Park, D. (2013). The no-prop algorithm: A new learning algorithm for multilayer neural networks. *Neural Networks*, 37:182 – 188.

Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *Journal of Financial and Quantitative Analysis*, 15(3):757 – 770.

Xia, Y., Liu, C., Da, B., e Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach. *Expert Systems with Applications*, 93:182 – 199. ISSN 0957-4174.

Zhang, C., Liu, C., Zhang, X., e Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications*, 82:128 – 150. ISSN 0957-4174.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings