




“Assessment of Support Vector Machine performance for default prediction and credit rating”

AUTHORS	Karim Amzile  Mohamed Habachi 
ARTICLE INFO	Karim Amzile and Mohamed Habachi (2022). Assessment of Support Vector Machine performance for default prediction and credit rating. <i>Banks and Bank Systems</i> , 17(1), 161-175. doi: 10.21511/bbs.17(1).2022.14
DOI	http://dx.doi.org/10.21511/bbs.17(1).2022.14
RELEASED ON	Saturday, 02 April 2022
RECEIVED ON	Saturday, 25 December 2021
ACCEPTED ON	Friday, 18 March 2022
LICENSE	 This work is licensed under a Creative Commons Attribution 4.0 International License
JOURNAL	"Banks and Bank Systems"
ISSN PRINT	1816-7403
ISSN ONLINE	1991-7074
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”



NUMBER OF REFERENCES

63



NUMBER OF FIGURES

4



NUMBER OF TABLES

16

© The author(s) 2022. This publication is an open access article.



BUSINESS PERSPECTIVES



LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine
www.businessperspectives.org

Received on: 25th of December, 2021

Accepted on: 18th of March, 2022

Published on: 2nd of April, 2022

© Karim Amzile, Mohamed Habachi,
2022

Karim Amzile, Ph.D. Student,
Department of Management Sciences,
Faculty of Law, Economic and
Social Sciences, Agdal, Mohammed
V University in Rabat, Morocco.
(Corresponding author)

Mohamed Habachi, Ph.D., Professor,
Department of Management Sciences,
Faculty of Law, Economic and Social
Sciences, Agdal, Mohammed V
University in Rabat, Morocco.



This is an Open Access article,
distributed under the terms of the
[Creative Commons Attribution 4.0
International license](https://creativecommons.org/licenses/by/4.0/), which permits
unrestricted re-use, distribution, and
reproduction in any medium, provided
the original work is properly cited.

Conflict of interest statement:

Author(s) reported no conflict of interest

Karim Amzile (Morocco), Mohamed Habachi (Morocco)

ASSESSMENT OF SUPPORT VECTOR MACHINE PERFORMANCE FOR DEFAULT PREDICTION AND CREDIT RATING

Abstract

Predicting the creditworthiness of bank customers is a major concern for banking institutions, as modeling the probability of default is a key focus of the Basel regulations. Practitioners propose different default modeling techniques such as linear discriminant analysis, logistic regression, Bayesian approach, and artificial intelligence techniques.

The performance of the default prediction is evaluated by the Receiver Operating Characteristic (ROC) curve using three types of kernels, namely, the polynomial kernel, the linear kernel and the Gaussian kernel. To justify the performance of the model, the study compares the prediction of default by the support vector with the logistic regression using data from a portfolio of particular bank customers.

The results of this study showed that the model based on the Support Vector Machine approach with the Radial Basis Function kernel, performs better in prediction, compared to the logistic regression model, with a value of the ROC curve equal to 98%, against 71.7% for the logistic regression model. Also, this paper presents the conception of a support vector machine-based rating tool designed to classify bank customers and determine their probability of default. This probability has been computed empirically and represents the proportion of defaulting customers in each class.

Keywords

artificial intelligence, scoring, probability of default, data mining, credit risk, bank

JEL Classification

C13, G21, G32

INTRODUCTION

Banks' retail customers are particulars and professionals with a low level of commitment. Particular customers are customers whose needs are generally limited to consumer and housing loans.

For the granting of credit, the solvency of a customer remains the major concern for banks because the solvency represents the customer's capacity to honor her commitments without incidents during all the duration of the credit.

Credit risk management and the prediction of the creditworthiness of bank customers remain a primary topic in the field of financial risk management and have recently become the main focus of the banking and financial sector (Lai et al., 2006).

The classification of customers according to their quantitative characteristics (income, repayment burden, age, etc.) and qualitative characteristics (number of incidents, civil status, etc.) into solvent and insolvent customers is done using a scoring tool. The latter, designed

using probabilistic, Bayesian or artificial intelligence (AI) techniques, makes it possible to distinguish between good customers and bad customers.

Since the publication of the Basel 2 Accord, the scoring system has evolved into a rating system for companies, particulars and professionals. Within this framework, various studies have been conducted to compare the techniques to be used to guarantee the efficiency and performance of the rating tools.

Logistic regression (LR), linear discriminant analysis (LDA), the Bayesian approach and artificial neural networks (ANN) are the main techniques studied by the researchers. Indeed, for databases of limited size, the use of LR and LDA is an optimal choice, whereas for large databases that require a continuous readjustment of the classification data, the use of AI methods becomes the most appropriate way to increase the accuracy of predicting the creditworthiness of bank customers.

This paper evaluates the performance of models based on the support vector machine (SVM) using multiple kernels and its ability to classify the portfolio into several classes according to the probability of default to build a rating tool compliant with banking regulations. The empirical study also compares the support vector and logistic regression.

1. LITERATURE REVIEW AND HYPOTHESES

Default prediction has been the subject of various academic studies as the literature highlights the use of various techniques to model the creditworthiness of bank customers, including LR, LDA, Bayesian approach, ANN, SVM and other techniques related to AI. This study proposes an extension of previous research regarding the variables used and the design of a scoring tool.

1.1. The choice of variables for fault prediction

Logistic regression has been studied by several researchers such as Ohlson (1980), Jones and Hensher (2007) and Benbachir and Habachi (2018), Zizi et al. (2020) and Habachi and El Haddad (2021), while LDA has been studied by Altman et al. (1994), Habachi and Benbachir (2019), and Svabova et al. (2020).

As for the ANN method, it is robust to specification errors (Cybenko, 1989; Barron, 1993). Multiple studies and research work to predict the creditworthiness of bank customers have used this technique and have confirmed this robustness, among which Lee et al. (2002), West (2000), Khashman (2010), Tsai et al. (2009), Coats and Fant (1993), Dimitras et al. (1996), Coakley and Brown (2000), Ravi Kumar and Ravi (2007) can be cited.

However, the SVM developed by Cortes and Vapnik (1995) is defined by Noble (2006) as an algorithm that learns from historical data to assign labels to new anonymous data, has been the subject of several researches such as Pławiak et al. (2019) who showed that the SVM approach is a very powerful technique for default probability prediction.

The use of kernels for nonlinearly separable data is done by Suykens and Vandewalle (1999), Zhou et al. (2009) and K. Amzile and R. Amzile (2021), who concluded that an SVM with an RBF kernel is determined with excellent performance and low computational cost.

The use of SVM to calculate customer scores is done by Huang et al. (2007) who concluded that the score based on the SVM approach correctly classifies the credit applications, while Guyon et al. (2002) concluded that the SVM method is very efficient for classification with an accuracy of 98%.

For modeling the probability of default of bank customers, Goh and Lee (2019) concluded that SVM represents the current trend among practitioners and academicians, and also represents the main alternative for improving the performance of the bank customer credit prediction model.

The literature shows that some studies have combined the SVM method with other techniques to

increase the performance of the model. In particular, Zhou et al. (2013) used the hybrid SVM-KNN model (K-Nearest Neighbors method is among the machine learning algorithms) to improve the prediction accuracy of SVM; the experimental results imply that the hybrid SVM-KNN model is an ideal and efficient approach for credit rating.

These studies used quantitative variables to predict retail customer default. Benbachir and Habachi (2018, 2019) used qualitative variables to predict corporate default.

1.2. Comparative studies conducted on default prediction models

Modeling techniques are compared by various researchers such as Altman et al. (1994) who compare ANN and LDA, Worth and Cronin (2003) who compared LDA and LR, Pavlyshenko (2016) who compared machine learning (ML) methods and LR, and Salazar et al. (2012) who compared SVM and LR

Verplancke et al. (2008) compared SVM and LR and concluded that the discriminatory power of LR and SVM models is satisfactory. Musa (2013) showed that SVM and LR on all performance measures have equal performance for both balanced and unbalanced data. However, SVM may perform better for highly unbalanced data sets.

According to Ruiz et al. (2017), loan evaluation with LR and SVM models not only improved the delinquency rate and approval rate, but also optimized the loan approval time. Feng et al. (2019) concluded that twenty-two selected ML models have comparable capabilities to the LR model, among them, SVM performed significantly better than LR. The same result was obtained by Baesens et al. (2003), Xiao et al. (2006), Yao and Chen (2019).

The dependence of model performance on modeling conditions was studied by Hassan and Jayousi (2020). The authors compared the performance of techniques related to AI in predicting bank customer default and concluded that there is no best technique for credit prediction problems for all situations, since the performance of the techniques depends on the structure and quality of the data used during training.

This study evaluates the performance of the SVM method with reference to LR and the extension of this technique for the construction of a multi-class classification tool and the determination of the probability of defect per class.

Accordingly, this paper proposes two following hypotheses:

H_1 : SVM approach improves performance of customer credit prediction.

H_2 : SVM is used to build rating tools and calculate the probability of default in accordance with the provisions of banking regulations.

2. METHODOLOGY

The methodology presented in the rest of this section is composed of the presentation of fault modeling by LR, presentation of fault modeling by SVM, performance measurement and, finally, the construction of a rating tool.

2.1. Modeling the probability of default using logistic regression

The customer's creditworthiness is modeled by the binary variable Y defined by:

$$Y = \begin{cases} 0 & \text{if the client is creditworthy} \\ 1 & \text{if the client is not-creditworthy} \end{cases} \quad (1)$$

Logistic regression (LR) modeling consists in calculating the probability of realizing the variable $Y = 1$, noted p defined as follows:

$$p = P(Y = 1 / X) = \frac{e^U}{1 + e^U}, \quad (2)$$

with

$$U = \beta X^T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{13} X_{13}, \quad (3)$$

$$X = (1, X_1, X_2, X_3, \dots, X_{13}), \quad (4)$$

where X_i are explanatory variables; $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_{13})$ are regression coefficients.

To determine the explanatory (discriminant) variables and the discriminant function, the relationship between each variable and the defect must be studied using univariate logistic regression, which allows us to determine the significant variables. The discriminant function is determined by studying the multivariate regression between the defect and the explanatory variables.

Knowing that Y_i follows the Bernoulli law of parameter p :

$$Y_i \sim B(p) \Leftrightarrow Y_i \sim B(p). \tag{5}$$

Let's set Y_i the creditworthiness of the bank's customer $\{0,1\}$, the probability function Y_i is written as follows:

$$p(Y_i = y) = p^y (1 - p)^{(1-y)} \tag{6}$$

where $y \in \{0, 1\}$.

The likelihood estimator will allow estimating the vector of parameters β (in the univariate case $\beta = (\beta_0, \beta_1)$). To estimate β , it is necessary to use the data of n client. Therefore, the function of Fisher represented by the product of probabilities is written:

$$\mathcal{L}(\beta) = \prod_{i=1}^n P_i^{y_i} (1 - P_i)^{(1-y_i)}, \tag{7}$$

with

$$P_i = P(Y = 1 | X) = \frac{e^{\beta X_i^T}}{1 + e^{\beta X_i^T}}, \tag{8}$$

where $X_i = (x_1, x_2, \dots, x_n)$ is the vector of client data (i).

The function log likelihood:

$$\ln(\mathcal{L}(\beta)) = \sum_{i=1}^n y_i (\beta X_i^T) - \ln(1 + e^{\beta X_i^T}) \tag{9}$$

To estimate β , the function of log likelihood must be maximized. For that, it is necessary that β verify the two following conditions:

$$\frac{\partial \ln(\mathcal{L}(\beta))}{\partial \beta} = 0, \text{ and } \frac{\partial^2 \ln(\mathcal{L}(\beta))}{\partial^2 \beta} < 0. \tag{10}$$

This problem is solved using the Newton-Raphson algorithm.

2.2. Modeling the probability of default by SVM

In this paper, the problem treated is a binary discrimination problem. For that, it is necessary to find a decision function allowing us to assign the observations between two classes (creditworthy and not-creditworthy).

$$x_i = (x_{1i}, x_{2i}, x_{3i}, \dots, x_{pi}) \in X \subset \mathbb{R}^p, \tag{11}$$

belong an observation of X concerning the particular (i), $x_i \in \mathbb{R}^p$. The decision function (separator) is defined by:

$$F : \mathbb{R}^p \rightarrow \{-1, +1\}. \tag{12}$$

In terms of probability, this amounts to minimizing the probability of error of the decision function to assign a particular customer x_i :

$$P(F(x) \neq y_i / x_i), \tag{13}$$

with P an unknown law defined on $(\mathbb{R}^p, \{-1, +1\})$.

To produce the decision function, it is necessary to use the training data $\{(x_i, y_i)\}$, $i = 1 \dots n$ with $x_i \in X$ and $y_i \in \{-1, +1\}$.

The SVM allows determining this decision function based on the collected observations and updating this function according to the new data. This paragraph presents the theoretical aspect of this method composed of the definition of the discrimination problem to solve and the approach to follow.

A discrimination problem is linearly separable if there is a linear decision function F called linear separator of the form

$$\begin{cases} F(X) = \text{sign}(h(X)) \\ h(x) = w^T X + w_0 \end{cases}, \tag{14}$$

with $w \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$.

Correctly classifying the learning set $\{F(x_i) = y_i; i = 1, \dots, n\}$.

The decision boundary associated with the decision function is defined as follows:

$$S(w, w_0) = \{x \in \mathbb{R}^p / w^T x + w_0 = 0\}, \quad (15)$$

where S is a hyperplane $h(x)$ of equation $w^T x + w_0 = 0$.

$$F(x_i) = \begin{cases} \text{Creditworthy} & Si w^T x_i + w_0 \geq 0 \\ \text{Not-creditworthy} & Si w^T x_i + w_0 < 0 \end{cases}. \quad (16)$$

The point above or on the hyperplane will be classified as class +1, and the point below the hyperplane will be classified as class -1.

Let data $\{(x_i, y_i)\}, i = 1 \dots n$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$.

An SVM separator is a linear discriminator of the form $F(x) = \text{sign}(w^T x + w_0)$, where $w \in \mathbb{R}^p$ and $w_0 \in \mathbb{R}$ are given by the solution of the following optimization system:

$$\begin{cases} \text{Min} \frac{\|w\|^2}{2} \\ \text{with } y_i (w^T x_i + w_0) \geq 1, i = 1, \dots, n \end{cases}, \quad (17)$$

where $w = \sqrt{w_1^2 + w_2^2 + \dots + w_p^2}$. (18)

The problem 17 is a convex quadratic problem under linear constraints whose objective function is to be minimized, The Lagrangian of this problem is written as follows:

$$\mathcal{L} = \frac{\|w\|^2}{2} - \sum_{i=1}^n \lambda_i (y_i (w^T x_i + w_0) - 1), \quad (19)$$

with $\lambda_i, i = 1, \dots, n$ - Lagrange multipliers.

To solve the problem 17, the Lagrangian \mathcal{L} must be minimized with respect to w and w_0 and maximized with respect to the variables λ_i . In this case, the saddle point (minimum with respect to the variables w and w_0 and maximum with respect to the variables λ_i) must satisfy the ‘‘Karush-Kuhn-Tucker (KKT)’’ conditions.

Consequently, the separation hyperplane is defined by:

$$h(x) = \sum_{i=1}^n \lambda_i^* y_i (x_i, x) + w_0^*. \quad (20)$$

The determination of the hyperplane defined by formula 20 allows defining the classification rule of a new observation (x) , which is as follows:

$$F(x) = \text{sign} \left(\sum_{i=1}^n \lambda_i^* y_i (x_i, x) + w_0^* \right), \quad (21)$$

with (x_i, x) is the scalar product.

When it is impossible to fully separate the data with a hyperplane, the data are non-linearly-separable. In this case, the data must be processed to obtain a separable representation, since SVMs are unable to handle such a problem, the processing to be performed consists in using techniques that transform the data making them linearly separable after transformation.

Indeed, the transformation is provided by the function ψ defined by:

$$\psi : \mathbb{R}^m \rightarrow \mathbb{R}^d, \quad x \rightarrow \psi(x). \quad (22)$$

Consequently, to find the separating hyperplane, the same line of equation presented in the previous case is used, but replacing the x_i by $\psi(x_i), i = 1 \dots n$, which allows determining the classification function and the separating hyperplane from formulas 20 and 21:

$$h(x) = \sum_{i=1}^n \lambda_i^* y_i (\psi(x_i), \psi(x)) + w_0^*, \quad (23)$$

$$F(x) = \text{sign} \left(\sum_{i=1}^n \lambda_i^* y_i (\psi(x_i), \psi(x)) + w_0^* \right). \quad (24)$$

Formulas 23 and 24 contain a scalar product $(\psi(x_i), \psi(x))$ that have to be defined; for that, the Kernel method should be used.

The kernel K is a function with two symmetric and positive variables that allows defining a scalar product in the transformation space:

$$K(x_i, x_j) = \langle \psi(x_i), \psi(x_j) \rangle. \quad (25)$$

The choice of kernel impacts the prediction performance of SVMs (Savas & Doyis, 2019). The literature suggests certain Kernel whose K function is defined as follows:

- The linear kernel defined by:

$$K(x_i, x_j) = (x_i^T x_j); \quad (26)$$

- The polynomial kernel:

$$K(x_i, x_j) = (x_i^T x_j + 1)^d; \tag{27}$$

- Gaussian RBF kernel (radial basis function):

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\gamma}}, \tag{28}$$

where “ γ ” controls the bandwidth of the Gaussian: the narrower the Gaussian, i.e. the more the distance between x_i, x_j must be low for the kernel to be different from 0;

- Inverse multi-quadratic:

$$K(x_i, x_j) = \frac{1}{\sqrt{(x_i - x_j)^T (x_i - x_j) + \beta}} = \frac{1}{\sqrt{\|x_i - x_j\|^2 + \beta}}. \tag{29}$$

2.3. SVM and rating

A rating tool is used to classify clients according to their characteristics and their probability of default over a one-year horizon. Indeed, it can be presented as follows (Table1).

Table 1. Design of the scoring tool

Class (i)	Qualification	Score (S_i)	Probability of default
1	Excellent	$\geq S_1$	P_8
2	Very good	$[S_1, S_2]$	P_7
3	Good	$[S_2, S_3]$	P_6
4	Fair Good	$[S_3, S_4]$	P_5
5	Medium	$[S_4, S_5]$	P_4
6	Low	$[S_5, S_6]$	P_3
7	Risky	$[S_6, S_7]$	P_2
8	Very risky	$[S_7, S_8]$	P_1
Default	Default	-	-

With

$$P_i = \frac{\text{number of defaulting clients in the class } i}{\text{the total number of clients in the class } i}, \tag{30}$$

The score S_i is defined by the value of the hyper-plane $h(x)$. Indeed, rating classes are defined as follows:

- The client is in class 1 if

$$S_i \leq (\sum_{i=1}^n \lambda_i^* y_i(x_i, x) + w_0^*) \cdot 100, \tag{31}$$

- The client is in class i if

$$S_i \leq (\sum_{i=1}^n \lambda_i^* y_i(x_i, x) + w_0^*) \cdot 100 \leq S_{i-1}. \tag{32}$$

This study uses data from the customers of a Portuguese bank. The database has 3,000 customers divided in healthy customers and defaulting customers

3. EMPIRICAL RESULTS

3.1. Definition of variables

The variables X_i used to explain the default are presented as follows:

- X_1 : Revolving loan or cash. (Qualitative variable)
- X_2 : Type of client: public sector, private sector. (Qualitative variable).
- X_3 : Age in the relationship: more than one year old, less than one year new. (Qualitative variable)
- X_4 : Number of children in the client’s home. (Quantitative variable)
- X_5 : Client’s income. (Quantitative variable)
- X_6 : Amount of credit. (Quantitative variable)
- X_7 : Repayment burden. (Quantitative variable)
- X_8 : The price of the good for which the client received the loan (Quantitative variable)
- X_9 : The type of income of the client: businessman, employment...etc. (Qualitative variable)
- X_{10} : Client’s level of education (Qualitative variable)
- X_{11} : Client’s family situation (Qualitative variable)

- X_{12} : Authorization overrun (Qualitative variable)
- X_{13} : Number of credit flows in the account (Quantitative variable)

3.2. Logistic regression modeling

This study uses data from the customers of a bank. The database has 3,000 customers divided in healthy customers and defaulting customers as follows (Table 2).

Table 2. Portfolio allocation

Variables	Creditworthy	Not creditworthy
Number	1,992	1,008
Percentage	66%	34%

The results of the univariate analysis detailed in Table A1 of Appendix A show that variables X_7 and X_{11} are not significant because the p-value is lower than 5%. Therefore, these two variables will be excluded from the logistic regression modeling of default.

The analysis of the correlation of the selected explanatory variables, detailed in Table A2 of Appendix A, shows that the variables X_8 and X_6 are strongly correlated ($0.986 \geq 0.5$) and that the variables X_{13} and X_4 are also strongly correlated ($0.888 \geq 0.5$). Therefore, the variables X_8 and X_{13} will be excluded. In fact, the variables “number of children” and “amount of credit” will be maintained.

The multivariate analysis detailed in Table A3 of Appendix A allows us to determine the following discriminant function:

$$\begin{aligned}
 MS = & 0.873 + 0.606 \cdot X_1 - 1.604 \cdot X_2 - \\
 & -0.457 \cdot X_3 + 0.075 \cdot X_4 - 1.504 \cdot X_5 + \\
 & +0.504 \cdot X_6 - 0.363 \cdot X_8 + \\
 & +0.756 \cdot X_{10} + 0.116 \cdot X_{11}.
 \end{aligned}
 \tag{33}$$

The value of the LR (Likelihood Ratio or RV Ratio Likelihood) statistic is:

$$\begin{aligned}
 LL(Base) &= 5,528.5379, \\
 LL(modèle) &= 4,531.338032, \\
 LR = RV = 2LL(modèle) - \\
 -2LL(Base) &= 997.199868.
 \end{aligned}
 \tag{34}$$

Consequently, the observed value is 997.2 greater than the critical value of $\chi_9^{20.05}$ fixed at 16.92. Therefore, the null hypothesis must be rejected and the selected variables are globally significant and make it possible to build a performing model.

The confusion matrix is shown in Table 3.

Table 3. Confusion matrix

Observations		Forecast			
		Target		Percentage correct	
		0	1		
Step 1	Target	0	1,357	635	68.1
		1	492	1,504	75.4
	Overall percentage	–	–	–	P. P = 71.7%

The confusion matrix shows that the model performs well in terms of classification with a percentage of 71.7%.

The results of the test of Hosmer-Lemeshow test show that the model fits the observed data. Indeed, the p-value is less than 0.05 (Table A4, Appendix A).

AUC shows that the model has an acceptable predictive capacity as it represents 78.4% (Figure B1, Appendix B).

3.3. SVM modeling

For the SVM modeling, the variables retained are those considered as explanatory by the LR study presented in the previous paragraph. Indeed, to compare the two types of modeling, the same dependent variables retained by the LR model must be used.

The database composed of 3,000 customers is divided into two classes, the first is composed of 75% of the data, or 2,250 customers, will be used to train the SVM algorithm, and the second has 25% of the data and will be devoted to the evaluation of the model validation. The modeling is done according to two assumptions: the first one considers the data linearly separable, while the second one assumes the data non-linearly separable.

For the first case, the separation hyperplane is defined by the coefficients (w^*). Indeed, the coefficients (w^*), $i = 0, \dots, 9$ are presented in Table 4.

Table 4. Coefficients (w_j^*)

w_1^*	w_2^*	w_3^*	w_4^*	w_5^*	w_6^*	w_7^*	w_8^*	w_9^*	b or w_0^*
0.587	-0.048	0.010	-0.036	0.153	-0.271	0.099	0.525	-0.213	0.437

Therefore, the relationship between the variable y_i and the explanatory variables $x_j, j = 0, \dots, 9$ is written:

$$y_i = \sum_{j=1}^9 w_j^* x_{ij} + w_0^* \tag{35}$$

Hence:

$$y_i = 0.587 \cdot x_{i1} - 0.048 \cdot x_{i2} + 0.010 \cdot x_{i3} - 0.036 \cdot x_{i4} + 0.153 \cdot x_{i5} - 0.271 \cdot x_{i6} + 0.099 \cdot x_{i7} + 0.525 \cdot x_{i8} - 0.213 \cdot x_{i9} + 0.437, \tag{36}$$

with

$$(w_0^*)_{Linear-SVM} = 0.437. \tag{37}$$

The hyperplane $h(x)$ is defined by:

$$h(x) = \sum_{j=1}^9 w_j^* x_j + w_0^*, \tag{38}$$

with $x = (x_j), j = 1, \dots, 9$.

For the second case, the data are considered non-linearly separable. Therefore, in this study, two kernels will be used to transform the data, according to the approach presented in the methodology section, namely the RBF and poly kernels.

The choice of the kernel to be used for the modeling consists in defining the parameters (γ , C , Out , Degree), which allows optimizing the performance of the model. Indeed, the parameters selected are those that maximize the AUC.

For the RBF kernel, the function *Gird Search*, which allows testing a series of parameters and comparing their performances in order to deduce the best parameters, makes it possible to show that the parameter kernel (1.0, 1.2) defines the best performing model with an accuracy of 98.148%. The parameter test results are presented in Table A5 (Appendix A).

For the Poly kernel, the parameter kernel (1, 3, 4) defines the best performing model with an accu-

racy of 93%. The parameter test results are shown in Table A6 (Appendix A).

The results of core selection are presented in Table 5.

Table 5. Parameters of the selected kernels

Kernel	C	γ	Degree	Accuracy
SVM-RBF	1.0	0.1	2	0.98
SVM-Linear	1.0	N/A	N/A	0.62
SVM-Poly	1.0	3	4	0.96

The selected cores are presented as follows:

- **Polynomial of degree “d”:**

$$K(x_i, x) = (x_i^T x + 1)^d, \tag{39}$$

- **RBF Core:**

$$K(x_i, x) = e^{-\frac{x_i - x^2}{0.1}}. \tag{40}$$

The confusion matrix of the three kernels is presented by Table A7 (Appendix A). As a result, the accuracy ratio and the AUC of each model defined from the three kernels are presented in Table 6.

Table 6. The three error ratios of the three kernels

Kernel	Accuracy Score	AUC
SVM-RBF	0.98	1.0
SVM-Linear	0.62	0.62
SVM-Poly	0.96	0.96

The ROC curve of the three models is presented in Figures B2, B3 and B4 (Appendix B). The performance of the models allows us to choose the RBF-SVM kernel for modeling as it maximizes the accuracy ratio and offers the highest performance with a maximum AUC (100%).

Following the definition of the model to be used, the parameters of the model must be determined. Therefore, the value of w_0^* is equal to:

$$(w_0^*)_{RBF-SVM} = -0.38383247. \tag{41}$$

The separation hyperplane is written as:

$$h(x) = \sum_{i=1}^{1474} \lambda_i^* y_i e^{\frac{x_i - x^2}{0.1}} - 0.38383247. \quad (42)$$

The decision function is defined by:

$$F(x) = \text{sign} \left(\sum_{i=1}^{1474} \lambda_i^* y_i e^{\frac{x_i - x^2}{0.1}} - 0.38383247 \right). \quad (43)$$

Note that only the support points used for the ranking are the only ones that have λ_i^* non-zero weights. The distribution of the support points is given in Table A8 (Appendix A).

3.4. Conception of a rating tool using SVM

The determination of the hyperplane equation allowed us to calculate the score of each client in the database used. The number of customers with a positive score is equal to 911, presented in Table 7.

Table 7. Portfolio allocation

Variables	Creditworthy	Not creditworthy
Number	586	325
Percentage	64.32%	35.68%

3.5. SVM and rating

After training the model, it is possible to build the rating tool, which will allow ranking customers according to their characteristics and their probability of default over a one-year horizon (Table 8).

Table 8. Conception of the rating tool

Class (i)	Qualification	Score (S)	Probability of default
1	Excellent	[300-... [0.0%
2	Very good	[250-300[2.9%
3	Good	[150-250[8.4%
4	Fair Good	[105-150[14.7%
5	Medium	[35-105[32.7%
6	Low	[30-35[44.0%
7	Risky	[20-30[63.8%
8	Very risky	[0-20[77.9%
Default	Default	-	-

4. DISCUSSION

This study explores the SVM method and compares it with LR for predicting the creditworthiness of bank customers to evaluate its performance using quantitative and qualitative variables. Then, it determines the best performing model that will be used to construct a rating tool composed of 8 classes and determine the probability of default per class.

The results of this study show that the model derived from the SVM method performs better than LR. Indeed, it determines the separation hyperplane by considering first that the data are linearly separable, then it uses three kernels, which are RBF, linear and polynomial to separate the data; finally, it simulates the performance of the three kernels by varying their characteristics, which allowed determining the best performing model with the associated kernel. The chosen model is the model with the RBF kernel whose characteristics allowed reaching a rate of accuracy = 98% and $AUC \approx 98\%$.

The results of this study are consistent with those of Pławiak et al. (2019), who showed that the SVM approach is a high performing technique, and Salazar et al. (2012), who compared SVM and logistic regression and concluded that SVM represents a high level of accuracy.

Based on the results obtained in this study, the hypothesis H_1 is validated, since the results of the SVM are better than those of the LR. Therefore, artificial intelligence techniques can be a better alternative to classical statistical techniques in constructing models for predicting the creditworthiness of bank customers.

The value of the hyperplane for each relationship allowed us to classify the study base into several classes. Each class is represented by a score range and characterized by a probability of default distinct from the other classes. This shows that the SVM modeling can be used by banks to build rating tools that comply with the Basel regulations. This finding validates the second hypothesis H_2 .

In addition, future research should examine the performance of this type of AI methods in modeling other credit risk components such as loss given default (LGD) and exposure at default (EAD).

CONCLUSION

Banks are always looking for the best models to predict customer creditworthiness and measure credit risk, which has led to the exploration of various modeling techniques, from probabilistic techniques to those based on artificial intelligence.

The objective of this study was to examine the performance of artificial intelligence modeling techniques in predicting the creditworthiness of bank customers, as well as the design of a rating tool using the model obtained by the SVM method.

According to the results obtained in this study, SVM_{RBF} succeeded in terms of the level of predictability of bank customer creditworthiness to reveal its performance with a value of $ROC_{SVM-RBF} = 98\%$. However, the results of this study recommend artificial intelligence techniques to bank managers, especially when managing credit risk.

On the other hand, this study presents some difficulty in terms of choosing the parameter values of the SVM kernels, which represents a difficult task, since it requires combinations between all the parameters in order to choose the best ones that offer powerful models.

AUTHOR CONTRIBUTIONS

Conceptualization: Karim Amzile, Mohamed Habachi.

Data curation: Karim Amzile, Mohamed Habachi.

Formal analysis: Karim Amzile, Mohamed Habachi.

Funding acquisition: Karim Amzile.

Investigation: Karim Amzile, Mohamed Habachi.

Methodology: Karim Amzile, Mohamed Habachi.

Project administration: Karim Amzile, Mohamed Habachi.

Resources: Karim Amzile.

Software: Karim Amzile.

Supervision: Karim Amzile, Mohamed Habachi.

Validation: Karim Amzile, Mohamed Habachi.

Visualization: Karim Amzile, Mohamed Habachi.

Writing – original draft: Karim Amzile, Mohamed Habachi.

Writing – reviewing & editing: Karim Amzile, Mohamed Habachi.

REFERENCES

1. Aboobyda, J. H., & Tarig, A. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. *Machine Learning and Applications: An International Journal (MLAIJ)*, 3(1), 1-9. <https://doi.org/10.5121/mlaij.2016.3101>
2. Altman, E. I., Marco, G., & Varetto, F. (1994). Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*, 18(3), 505-529. [https://doi.org/10.1016/0378-4266\(94\)90007-8](https://doi.org/10.1016/0378-4266(94)90007-8)
3. Altman, E., Haldeman, R., & Narayanan, P. (1977). ZETA analysis: a new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 1, 29-51. [https://doi.org/10.1016/0378-4266\(77\)90017-6](https://doi.org/10.1016/0378-4266(77)90017-6)
4. Amzile, K., & Amzile, R. (2021). Using SVM for Smart Direct Marketing (SDM): A case of predicting bank customers interested in the Term Deposits. *International Journal of Accounting, Finance, Auditing, Management and Economics*, 2(5), 525-537. Retrieved from <https://www.ijafame.org/index.php/ijafame/article/view/366/294>
5. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627-635. <https://doi.org/10.1057/palgrave.jors.2601545>

6. Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3), 930-945. <https://doi.org/10.1109/18.256500>
7. Bassey, P. (2019). *Logistic Regression Vs Support Vector Machines (SVM)*. Retrieved from <https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16>
8. Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, 36(2), 3302-3308. <https://doi.org/10.1016/j.eswa.2008.01.005>
9. Benbachir, S., & Habachi, M. (2018). Assessing the Impact of Modelling on the Expected Credit Loss (ECL) of a Portfolio of Small and Medium-sized Enterprises. *Universal Journal of Management*, 6(10), 409-431. <https://doi.org/10.13189/ujm.2018.061005>
10. Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: Receiver operating characteristic curves. *Critical Care*, 8, 508. <https://doi.org/10.1186/cc3000>
11. Chen, T.-H. (2020). Do you know your customer? Bank risk assessment based on machine learning. *Applied Soft Computing*, 86, 105779. <https://doi.org/10.1016/j.asoc.2019.105779>
12. Çiğşar, B., & Ünal, D. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Scientific Programming*, 2019, 1-8. <https://doi.org/10.1155/2019/8706505>
13. Coakley, J. R., & Brown, C. E. (2000). Artificial neural networks in accounting and finance: modeling issues. *Intelligent Systems in Accounting, Finance and Management*, 9(2), 119-144. [https://doi.org/10.1002/1099-1174\(200006\)9:2<119::AID-ISAF182>3.0.CO;2-Y](https://doi.org/10.1002/1099-1174(200006)9:2<119::AID-ISAF182>3.0.CO;2-Y)
14. Coats, P. K., & Fant, L. F. (1993). Recognizing Financial Distress Patterns Using a Neural Network Tool. *Financial Management*, 22(3), Fall.
15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>
16. Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2, 303-314. <https://doi.org/10.1007/BF02551274>
17. Danenas, P., & Garsva, G. (2015). Selection of Support Vector Machines based classifiers for credit risk domain. *Expert Systems with Applications*, 42(6), 3194-3204. <https://doi.org/10.1016/j.eswa.2014.12.001>
18. Dimitras, A. I., Zanakis, C., & Zopounidis, S. H. (1996). A survey of business failures with an emphasis on prediction methods and industrial applications. *European Journal of Operational Research*, 90(3), 487-513. [https://doi.org/10.1016/0377-2217\(95\)00070-4](https://doi.org/10.1016/0377-2217(95)00070-4)
19. El Sanharawi, M., & Naudet, F. (2013). Understanding logistic regression. *Journal Français d'Ophtalmologie*, 36(8), 710-715. <https://doi.org/10.1016/j.jfo.2013.05.008>
20. Feng, J., Wang, Y., Peng, J., Sun, M., Zeng, J., & Jiang, H. (2019). Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *Journal of Critical Care*, 54, 110-116. <https://doi.org/10.1016/j.jcrr.2019.08.010>
21. Francoeur, D. (2010). *Support vector machines: an introduction*. Retrieved from https://savoirs.usherbrooke.ca/bitstream/handle/11143/16093/2_francoeur_CaMUS_2010_vol.1.pdf
22. Frezza-Buet, H. (2013). *Vector Machines Supports Tutorial*. Retrieved from <http://www.metz.supelec.fr/metz/personnel/frezza/ApprentissageNumerique/svm-frereader.pdf>
23. Goh, R. Y., & Lee, L. S. (2019). Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches. *Advances in Operations Research*, 2019, 1-30. <https://doi.org/10.1155/2019/1974794>
24. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46, 389-422. <https://doi.org/10.1023/A:1012487302797>
25. Habachi, M., & Benbachir, S. (2019). Combination of linear discriminant analysis and expert opinion for the construction of credit rating models: The case of SMEs. *Cogent Business & Management*, 6(1), 1685926. <https://doi.org/10.1080/23311975.2019.1685926>
26. Habachi, M., & El Haddad, S. (2021). Impact of Covid-19 on SME portfolios in Morocco: Evaluation of banking risk costs and the effectiveness of state support measures. *Investment Management and Financial Innovations*, 18(3), 260-276. [https://doi.org/10.21511/imfi.18\(3\).2021.23](https://doi.org/10.21511/imfi.18(3).2021.23)
27. Hassan, A., & Jayousi, R. (2020). Financial Services Credit Scoring System Using Data Mining. *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1-7). <https://doi.org/10.1109/AICT50176.2020.9368572>
28. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). <https://doi.org/10.1002/9781118548387.fmatter>
29. Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, 33(4), 847-856. <https://doi.org/10.1016/j.eswa.2006.07.007>
30. Jones, S., & Hensher, D. A. (2007). Corporate failure: A multinomial nested logit analysis for unordered outcomes. *The British Accounting Review*, 39(1), 89-107. <https://doi.org/10.1016/j.bar.2006.12.003>
31. Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, 37(9), 6233-6239. <https://doi.org/10.1016/j.eswa.2010.02.101>

32. Lai, K. K., Yu, L., Wang, S., & Zhou, L. (2006). Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model. In S. Kollias, A. Stafylopatis, W. Duch, & E. Oja (Eds.), *Artificial Neural Networks - ICANN 2006* (pp. 682-690). Springer Berlin Heidelberg. https://doi.org/10.1007/11840930_71
33. Lee, T., Chiu, C., Lu, C., & Chen, I. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, 23(3), 245-254. [https://doi.org/10.1016/S0957-4174\(02\)00044-1](https://doi.org/10.1016/S0957-4174(02)00044-1)
34. Lejeune, M. (2010). *Statistics – Theory and its applications*. Sumy: Springer.
35. Loan Thi Vu, Lien Thi Vu, Nga Thu Nguyen, Phuong Thi Thuy Do, & Dong Phuong Dao (2019). Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies. *Investment Management and Financial Innovations*, 16(1), 276-290. [https://doi.org/10.21511/imfi.16\(1\).2019.22](https://doi.org/10.21511/imfi.16(1).2019.22)
36. Musa, A. B. (2013). Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics*, 4(1), 13-24. <https://doi.org/10.1007/s13042-012-0068-x>
37. Narayan, Y. (2021). Direct comparison of SVM and LR classifier for SEMG signal classification using TFD features. *Materials Today: Proceedings*, 45(2), 3543-3546. <https://doi.org/10.1016/j.matpr.2020.12.979>
38. Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, 24(12), 1565-1567. <https://doi.org/10.1038/nbt1206-1565>
39. Ohlson, J. A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109-131. <https://doi.org/10.2307/2490395>
40. Pavlyshenko, B. (2016). Machine learning, linear and Bayesian models for logistic regression in failure detection problems. *2016 IEEE International Conference on Big Data (Big Data)* (pp. 2046-2050). <https://doi.org/10.1109/BigData.2016.7840828>
41. Pławiak, P., Abdar, M., & Acharya, U. R. (2019). Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. *Applied Soft Computing*, 84, 105740. <https://doi.org/10.1016/j.asoc.2019.105740>
42. Rahman, M. S. (2016). The Advantages and Disadvantages of Using Qualitative and Quantitative Approaches and Methods in Language “Testing and Assessment” Research: A Literature Review. *Journal of Education and Learning*, 6(1), 102-112. <https://doi.org/10.5539/jel.v6n1p102>
43. Rakotomalala, R. (2016). *SVM: Support vector machine. Supervised Learning - Classification*. Retrieved from <http://eric.univ-lyon2.fr/~ricco/cours/slides/en/svm.pdf>
44. Ravi Kumar, P., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*, 180(1), 1-28. <https://doi.org/10.1016/j.ejor.2006.08.043>
45. Revel, A. (2016). *S'eparateurs `a vaste marge [Support Vector Machines]*. Retrieved from <https://pageperso.univ-lr.fr/arnaud.revel/MesPolys/SVM.pdf>
46. Ribeiro, B., Silva, C., Chen, N., Vieira, A., & das Neves, J. C. (2012). Enhanced default risk models with SVM+. *Expert Systems with Applications*, 39(11), 10140-10152. <https://doi.org/10.1016/j.eswa.2012.02.142>
47. Ruiz, S., Gomes, P., Rodrigues, L., & Gama, J. (2017). Credit Scoring in Microfinance Using Non-traditional Data. In E. Oliveira, J. Gama, Z. Vale, & H. Lopes Cardoso (Eds.), *Progress in Artificial Intelligence* (pp. 447-458). Springer International Publishing. https://doi.org/10.1007/978-3-319-65340-2_37
48. Salazar, D. A., Vélez, J. I., & Salazar, J. C. (2012). Comparison between SVM and Logistic Regression: Which One is Better to Discriminate? *Expert Systems with Applications*, 35(2), 223-237. Retrieved from <http://www.scielo.org.co/pdf/rce/v35nspe2/v35nspe2a03.pdf>
49. Savas, C., & Dervis, F. (2019). The Impact of Different Kernel Functions on the Performance of Scintillation Detection Based on Support Vector Machines. *Sensors*, 19(23), 5219. <https://doi.org/10.3390/s19235219>
50. Suykens, J. A. K., & Vandewalle, J. (1998). *Least Squares Support Vector Machine Classifiers*. Kluwer Academic Publishers. Retrieved from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.2.877&rep=rep1&type=pdf>
51. Svabova, L., Michalkova, L., Durica, M., & Nica, E. (2020). Business Failure Prediction for Slovak Small and Medium-Sized Companies. *Sustainability*, 12(11), 4572. <https://doi.org/10.3390/su12114572>
52. Thi Vu, L., Thi Vu, L., Thu Nguyen, N., Thi Thuy Do, P., & Phuong Dao, D. (2019). Feature selection methods and sampling techniques to financial distress prediction for Vietnamese listed companies. *Investment Management and Financial Innovations*, 16(1), 276-290. [https://doi.org/10.21511/imfi.16\(1\).2019.22](https://doi.org/10.21511/imfi.16(1).2019.22)
53. Tsai, M.-C., Lin, S.-P., Cheng, C.-C., & Lin, Y.-P. (2009). The consumer loan default predicting model – An application of DEA-DA and neural network. *Expert Systems with Applications*, 36(9), 11682-11690. <https://doi.org/10.1016/j.eswa.2009.03.009>
54. Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F., & Decruyenaere, J. (2008). Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Medical Informatics and Decision Making*, 8(1), 56. <https://doi.org/10.1186/1472-6947-8-56>
55. Wen, Z., & Li, T. (Eds.) (2013). *Practical Applications of Intelligent Systems. Proceedings of the Eighth International Confer-*

ence on Intelligent Systems and Knowledge Engineering, Shenzhen, China. Heidelberg: Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-54927-4>

56. West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27(11-12), 1131-1152. [https://doi.org/10.1016/S0305-0548\(99\)00149-5](https://doi.org/10.1016/S0305-0548(99)00149-5)

57. Worth, A., & Cronin, M. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure: THEOCHEM*, 622, 97-111. Retrieved from <https://publications.jrc.ec.europa.eu/repository/handle/JRC21426>

58. Xiao, W., Zhao, Q., & Fei, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering*, 15(4), 419-435. <https://doi.org/10.1007/s11518-006-5023-5>

59. Yao, J.-R., & Chen, J.-R. (2019). A New Hybrid Support Vector Machine Ensemble Classification Model for Credit Scoring. *Journal of Information Technology Research*, 12(1), 77-88. <https://doi.org/10.4018/JITR.2019010106>

60. Zhang, L., Hu, H., & Zhang, D. (2015). A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. *Financial Innovation*, 1(1), 14. <https://doi.org/10.1186/s40854-015-0014-5>

61. Zhang, Q., Wang, J., Lu, A., Wang, S., & Ma, J. (2018). An improved SMO algorithm for financial credit risk assessment – Evidence from China’s banking. *Neurocomputing*, 272, 314-325. <https://doi.org/10.1016/j.neucom.2017.07.002>

62. Zhou, L., Lai, K. K., & Yen, J. (2009). Credit Scoring Models with AUC Maximization Based on Weighted SVM. *International Journal of Information Technology & Decision Making*, 8(4), 677-696. <https://doi.org/10.1142/S0219622009003582>

63. Zizi, Y., Oudgou, M., & El Mouden, A. (2020). Determinants and Predictors of SMEs’ Financial Failure: A Logistic Regression Approach. *Risks*, 8(4), 107. <https://doi.org/10.3390/risks8040107>

APPENDIX A

Table A1. Univariate analysis table

X_i	B	E, S	Wald	ddl	Pv
X_1	.429	.132	10.522	1	.001
X_2	-1.604	.165	94.361	1	.000
X_3	-.424	.085	24.811	1	.000
X_4	.341	.126	7.298	1	.007
X_5	-1.504	.153	95.068	1	.000
X_6	.504	.110	34.113	1	.000
X_7	.000	.000	.435	1	.510
X_8	.000	.000	44.091	1	.000
X_9	-.375	.063	35.415	1	.000
X_{10}	.746	.038	375.557	1	.000
X_{11}	-.017	.043	.151	1	.697
X_{12}	.108	.044	6.167	1	.013
X_{13}	-.263	.111	5.650	1	.017
Constant	1.497	.354	17.887	1	.000

Table A2. The correlation table of the independent variables

X_i	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}	X_{11}	X_{12}
X_1	1	-.023	-.034	-.004	.000	.227	.188	-.027	.053	-.003	.007
X_2	-.023	1	-.093	.059	-.026	.063	.062	.001	-.049	-.037	.079
X_3	-.034*	-.093	1	.083	.179	.108	.116	.024	-.146	-.053	.109
X_4	-.004	.059	.083	1	-.056	-.033*	-.042	-.013	-.014	.014	.888
X_5	.000	-.026	.179	-.056	1	.379	.388	.196	-.273	-.058	-.067
X_6	.227	.063	.108	-.033	.379	1	.986	.106	-.197	-.072	.012
X_7	.188	.062	.116	-.042	.388	.986	1	.108	-.215	-.073	.005
X_9	-.027	.001	.024	-.013	.196	.106	.108	1	-.104	-.005	-.028
X_{10}	.053	-.049	-.146	-.014	-.273	-.197	-.215	-.104	1	.074	-.010
X_{11}	-.003	-.037*	-.053	.014	-.058	-.072	-.073	-.005	.074	1	-.015
X_{12}	.007	.079	.109	.888	-.067	.012	.005	-.028	-.010	-.015	1

Table A3. Wald test table

X_1	B	E, S	Wald	ddl	Pv	OddR
X_1	.606	.126	23.131	1	.000	1.834
X_2	-1.604	.163	97.068	1	.000	.201
X_3	-.457	.084	29.576	1	.000	.633
X_4	.075	.048	2.478	1	.015	1.078
X_5	-1.504	.153	95.068	1	.000	.255
X_6	.504	.110	34.113	1	.000	1.713
X_8	-.363	.063	33.671	1	.000	.695
X_{10}	.756	.038	394.351	1	.000	2.129
X_{11}	.116	.044	7.118	1	.008	1.123
ETC	.873	.270	10.447	1	.001	2.393

Table A4. Wald test table

Test de Hosmer – Lemeshow			
Step	Khi-Chi-deux	ddl	Pv
1	30,850	8	0.000

Table A5. Test of the SVM-RBF parameters by the Gird-Search function

Kernel	C	γ	Degree	Accuracy
SVM-RBF	1.0	1.0	2	0.903703
	1.0	0.9	2	0.903703
	1.0	0.8	2	0.92592
	1.0	0.7	2	0.92592
	1.0	0.6	2	0.92962
	1.0	0.5	2	0.95185
	1.0	0.4	2	0.95925
	1.0	0.3	2	0.97037
	1.0	0.2	2	0.97407
	1.0	0.1	2	0.98148
	1.0	0.09	2	0.97231

Table A6. Test of the SVM-Poly parameters by the function *Gird-Search*

Kernel	C	γ	Degree	Accuracy
SVM-Poly	1.0	1	1	0.618518
	1.0	1	2	0.644444
	1.0	1	3	0.74074
	1.0	2	3	0.91481
	1.0	3	3	0.929629
	1.0	3	4	0.966666
	1.0	4	4	0.966666
	1.0	4	5	0.96666

Table A7. The confusion matrix of the three kernels

Matrix	SVM-RBF		SVM-Linear		SVM-Poly	
	1	0	1	0	1	0
1	123	3	49	77	112	14
0	2	142	26	118	5	139

Table A8. Number of support vectors (RBF-SVM)

The number	Class 1 (Not-creditworthy)	Class 0 (creditworthy)	Total (n)
Support points	720	654	1,374

APPENDIX B

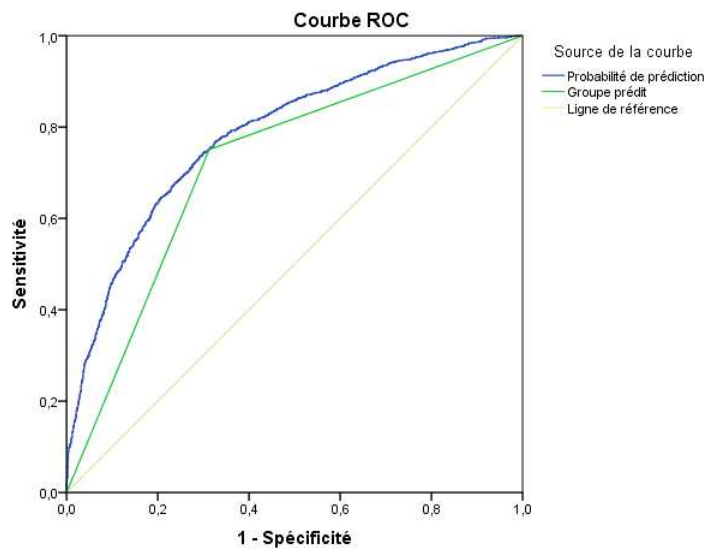


Figure B1. Performance of the LR model (ROC)

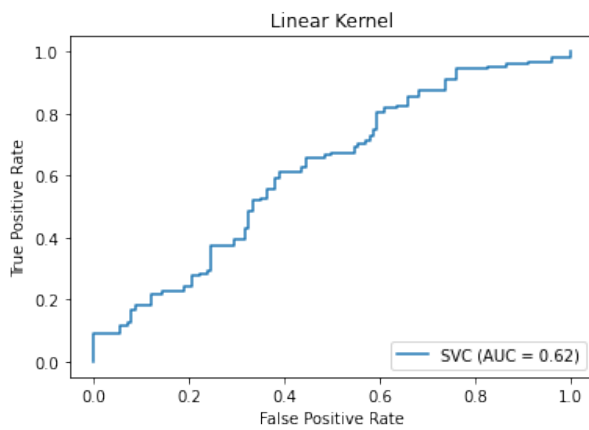


Figure B2. ROC curve of the linear kernel

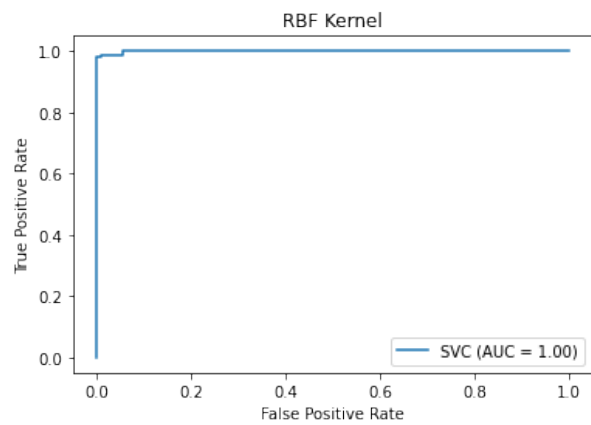


Figure B3. ROC curve of the Poly kernel

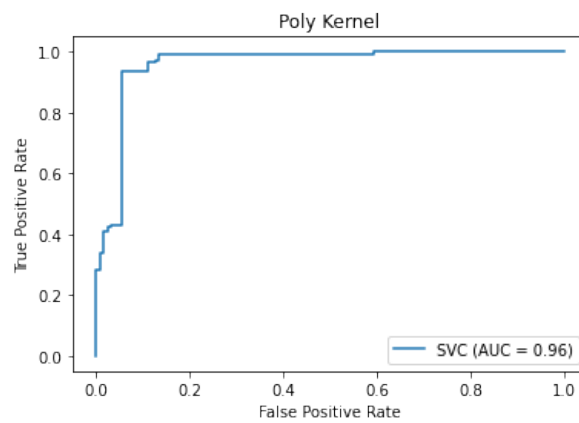


Figure B4. ROC curve of the Poly kernel