Assessing the Impact of Modelling on the Expected Credit Loss (ECL) of a Portfolio of Small and Medium-sized Enterprises

Saâd Benbachir1,2,3,*, Mohamed Habachi3

1Applied Mathematics, Management Sciences Laboratory, Morocco
2The Strategic Studies in Law, Economics and Management Center, Morocco
3FSJES-Agdal, Mohammed V University, Morocco

Abstract This paper studies the impact of the internal modelling on the calculation of expected credit loss in the framework of the international standard IFRS 9. Indeed, the probability of default of counterparty depends on the model used for the conception of the internal rating system. The multitude of probabilistic models renders uncertain and imprecise, the calculation of the expected loss for the same SMEs portfolio of a Moroccan bank, as well as the comparison of losses over time due to the non-permanence of the rating system used. As a result, the regulator will be unable to guarantee an equitable and transparent system of provisioning of the losses, because of the absence of standardization of the elaboration process of the rating tool. To show this risk associated with the multitude of models, this paper studied the impact of choice of the model on the expected credit loss, by calculating of the probability of default for several types of modelling based respectively on the pure logistic regression and the logistic regression on the principal components.

Keywords Logistic Regression Model, Principal Components Analysis (PCA), Probability of Default (PD), IRB Foundation Approach, Excepted Loss (EL)

1. Introduction

The banking regulation defines several approaches for calculating the capital required to cover the credit risk. This enriches the techniques available for determining the risk profile of banks.

These approaches are composed of simple approaches based on the application of a coefficient, on the credit exposure and of complex approaches based on the internal rating models (Internal Ratings-Basesd IRB). The expected loss calculation is based on the probability of default of the corporate, the loss at the time of default and the exposure, in term of outstanding amount of credit, in the event of default. Indeed, various studies were conducted to define an approach for predicting the probability of default. These studies concern linear discriminant analysis, the intelligence techniques, Bayesian networks and probabilistic models, which we will summarize as follows:

− Multidimensional Linear Discriminant Analysis
  The prediction of default by linear discriminant analysis was developed by Altman [2], by defining a linear relationship between default and financial ratios. Indeed, Altman has defined a score function \(Z\) that distinguishes between the healthy and the failing companies. The Altman's approach has been adopted by other research to determine ratios that predict failure such as those conducted by: Taffler [40], Bardos and al. [10] and Grover and Lauvin and al. [26].

− Intelligence Techniques
  These techniques are based on different logics such as neural networks and genetic algorithms. The several studies have applied these techniques to predict the default of the corporates, such as those conducted by: Bel and al. [13], Back and al. [8], Liang and wu [32], Bose and Pal [16] and Oreski and al. [35].

− Bayesian Network
  The Bayesian classifier (Friedman and al.[23]) is based on the calculation of a posterior probability of each observation belonging to a specific class. Indeed, he finds the posterior probability distribution \(P(Y|X)\), where \(Y = (Y_1, ..., Y_n)\) is a random variable to be classified in k categories and \(X = (X_1, ..., X_n)\) is a set of \(n\) explanatory variables. The Bayesian classification of failing companies
a was studied by a set of researchers as Gemela [24], Das and al. [20], Dwyer and al. [21], Gössl [25] and Tasch [42].

Probabilistic Models

The probabilistic models are the Logit model based on the logistic distribution and the Probit model based on the gaussian distribution. The several studies have focused on discriminant logistic analysis to predict the default of companies such as those conducted by Ohlson [34], Hunter and al. [31], Hensher and al.[28] while the Probit model has been studied by other researchers such as Zmijewski [44], Grover and al. [26] and Bunn and al.[18].

Due to the multitude of models, the determination of the expected credit loss (ECL) as defined by the Basel Committee on Banking Supervision [12] becomes dependent on the models and techniques chosen for the elaboration of the rating system, which renders uncertain the calculation of the expected credit loss (ECL), by the financial institutions, as a result of absence of standardization of the techniques used. In this case, the institutions can exploit the opportunities offered by the modelling to minimize the expected credit loss at the detriment of transparency and stability through the recourse of the arbitrage between the techniques and the models possible. This situation gives rise to a risk of models whose impact on the stability of banking system will have the same importance as the credit risk.

In this paper, we will show the uncertainty in the calculation of expected credit loss, generated by the multitude of models and the absence of standardization and it, by determining the expected loss of an SMEs portfolio of a Moroccan bank. Indeed, we will study the impact of the modelling of probability of default (PD) on the calculation of the expected credit loss and this, by using two versions of probabilistic models to predict the probability of default (PD).

The Modelling in our study is based on the logistic regression because we will use the pure logistic regression and the logistic regression on the principal components to determine the probability of default per the rating class.

The rest of this paper is as follows, Section 2 is devoted to the calculation of the expected credit loss. We first give a definition of the credit risk. We then define the approach to calculate the unexpected and expected credit loss. We finally present, in this section, the construction approach of the scoring system and a probabilistic model that will be used. The third section is reserved to the empirical study. In this section, we will first analyze and describe the database used. Then we present and interpret the empirical results of the probability of default for each chosen model. We finally compare the expected credit loss calculated by the chosen models.

2. Calculation of the Expected Credit Loss (ECL)

2.1. Definition and Quantification of Credit Risk

2.1.1. Definition

The credit risk of the banking portfolio is defined as the risk that counterparty will default for a one-year period. The notion of default means that the counterparty is unable to honor these commitments towards the establishment of credit. The incidents causing the default are multiple, but the most recurrent are:

- The downgrading outstanding debts in the bad debts,
- The chronic overruns on lines of credit granted to clients,
- The accounts that have few transactions and the applications that have lapsed and that have not been renewed,
- The unfavorable echoes from the market, the sectoral difficulties and a significant fall in the level of activity.

Each credit relationship is associated with an actual or potential credit risk situation.

2.1.1.1. The Credit Risk Situation

The credit risk situation is composed of the following elements:

- Probability of default (PD): Probability that a counterpart will default in a horizon one year.
- Loss given default (LGD): The share, expressed as a percentage of the amount a bank loses when a borrower falls at default on a credit.
- Exposure at default (EAD): The total value to which a bank is exposed when a credit is at default.
- Maturity (M): The effective maturity of credit.

2.1.1.2. The Losses Associated at the Risk of Credit

We distinguish two types of losses:

- The Expected Loss (EL)

The expected loss is a percentage equal to the multiplication of the PD and the LGD. The amount of the expected loss is equal to the multiplication of the expected loss (EL) and the exposure at default (EAD):

\[
\text{Amount of the expected loss } (EL_M) = PD \times LGD \times EAD
\]

\[EL_M = EL \times EAD\]
The Unexpected Loss \((UL)\)

The unexpected loss is the total loss from to the expected loss \((EL)\). It is calculated as a standard deviation from the mean at a certain confidence level \((1 − \alpha)\). It is also called \(VaR\) of credit.

2.1.2. The Quantification of Credit Risk

For the quantification of credit risk, the Basel Committee provides two categories of approaches for calculating the minimum capital requirements for credit risk \((K)\).

2.1.2.1. The Standardized Approach

Under the standardized approach, the bank must calculate the weighted assets. Indeed, a bank’s total risk-weighted on-balance sheet equal the sum of the risk-weighted amounts of each asset it holds on-balance sheet.

The risk-weighted amount of an on-balance sheet asset is determined by multiplying its current book value of credit exposures by the risk weight \((\alpha)\) specified by the regulator. Exposures should be risk-weighted net of specific provisions.

\[
\text{Risk-weighted amount}_{ST}\frac{\alpha}{\text{credit exposures on}} = \frac{\alpha}{\text{balance}}
\]

The risk-weighting \((\alpha)\) is defined by type of claims (claims on sovereigns and central banks, claims on other official entities, claims on banks and securities firms, claims on corporates, claims included in the regulatory retail portfolios, claims secured by residential property, claims secured by commercial real estate, treatment of past due loans, higher-risk categories et other assets)\(^3\); and depending on the rating assigned to the counterparty by the external rating agencies.

The minimum capital requirements under the standard approach \((K_{ST})\) to cover the given counterparty credit risk exposure is defined as follows:

\[
K_{ST} = \beta \times \text{risk} - \text{weighted amount}_{ST}
\]

Where : \(\beta\), is a minimum coefficient of solvability\(^4\)

2.1.2.2. The Internal Ratings-Based Approach\((IRB)\)

The Internal Rating-Based Approach is based on the four determinants mentioned above, which are the \(PD\), \(EAD\), \(LGD\) and the Maturity \(M\) to calculate of the unexpected loss \((UL)\) and expected loss \((EL)\) according of the regulatory requirements of Basel committee.

If the expected loss \((EL)\) is defined by the formula \((1)\), the unexpected loss \((PU)\) is calculated by multiplying a coefficient \((\alpha)\) by the exposure at default \((EAD)\). The coefficient \((\alpha)\) is a function of the \(PD\), \(LGD\) and Maturity \((M)\). As a result:

\[
UL = f(PD, LGD, M)5 \times EAD
\]

Furthermore, the banking regulations provide for two categories of Internal Rating-Based\(^5\):

- **The Internal Ratings-Based- foundation approach**

  Under the foundation approach, as a general rule, banks provide their own estimates of probability of default and rely on supervisory estimates for other risk components \((LGD, EAD\) and \(M)\).

- **The Internal Ratings-Based- advanced approach**

  Under the advanced approach, banks provide of their own estimates of \(PD\), \(LGD\) and \(EAD\), and their own calculation of \(M\).

2.2. Calculation of the Expected Credit Loss

Under the IRB approach (foundation and advanced), the probability of default \((PD)\) is an important component in calculating the expected loss \((EL)\). Indeed, the aim is to determine the risk profile of the credit counterparty on the basis of qualitative and quantitative data through a discriminant analysis. The internal rating tool must enable credit counterparties to be classified as a function of their probability of default \((PD)\) and must predict customer failure. For this reason, we will present the process of conception of this type of classifier (rating tool):

2.2.1. Presentation of the Process of Conception of a Rating Tool

The following schema can summarize the process of conception of rating tool:

\[\text{Exposures should be risk-weighted net of specific provisions.}\]

---


4 For Morocco this coefficient is fixed at 12% as from January 2012

5 The formula of \(f\) for SMEs is defined by Committee on Banking Supervision, 2006, International Convergence of Capital Measurement and Capital Standards, Bank for International Settlements at the paragraph level 272-273.

Assessing the Impact of Modelling on the Expected Credit Loss (ECL) of a Portfolio of Small and Medium-sized Enterprises

In the following, we will present the approach, which we have adopted for the conception of ratings models; it is based respectively on logistic regression and logistic regression on principal components.

2.2.1.1. Database Treatment

The treatment of the database and the choice of explanatory variables are made according to the following schema:

The Constitution of the Database (Definition of Variables)

- The quantitative variables
  The quantitative variables \( V_j \), \( j = 1, \ldots, 16 \), are divided into 6 classes, defined as follows:

<table>
<thead>
<tr>
<th>Class</th>
<th>The quantitative variables ( V_j, 1 \leq j \leq 16 )</th>
</tr>
</thead>
</table>
| \( C_1: Activity \) | \( V_1 = \text{Sales/turnover (ST)} \)  
|                  | \( V_2 = \text{Number of employees} \)  
|                  | \( V_3 = \text{Profit growth} \)  
|                  | \( V_4 = \text{Age of the company} \) |
| \( C_2: Return of investment \) | \( V_5 = \frac{\text{Net Profit}}{\text{Equity}} \)  
|                  | \( V_6 = \frac{\text{Sales/turnover (ST)}}{\text{Net Profit}} \) |
| \( C_3: Solvability \) | \( V_7 = \frac{\text{Financial expenses}}{\text{Sales/turnover (ST)}} \)  
|                  | \( V_8 = \frac{\text{DLMT (Debt long and medium term) \times Equity}}{\text{Net Debt}} \)  
|                  | \( V_9 = \frac{\text{Working Capital}}{\text{Adjusted Assets + working capital requirements} – \text{Accounts Receivable} – \text{inventory}} \) |
| \( C_4: Liquidity \) | \( V_{10} = \frac{\text{Current Assets}}{\text{Equity}} \) |
| \( C_5: Financial structure \) | \( V_{11} = \frac{\text{Working Capital}}{\text{Current Assets}} \)  
|                  | \( V_{12} = \frac{\text{Net fixed assets + inventory}}{\text{Total Assets}} \)  
| \( C_6: Turnover \) | \( V_{13} = \frac{\text{Sales/turnover (ST)}}{\text{Net fixed assets + inventory}} \)  
|                  | \( V_{14} = \frac{\text{Sales/turnover (ST)}}{\text{Accounts Receivable}} \)  
|                  | \( V_{15} = \frac{\text{Sales/turnover (ST)}}{\text{Accounts Payable}} \)  
|                  | \( V_{16} = \frac{\text{Total purchase}}{\text{Net fixed assets + inventory}} \) |

- The qualitative variables
  The qualitative variables \( q_m, 1 \leq m = 1, \ldots, 19 \) are grouped by theme \( T_k, k = 1, \ldots, 6 \), as follows:

Figure 1. The process of conception of rating tool

Figure 2. The treatment of the database and the choice of explanatory variables
Discretization of Qualitative Variables and Their Transformation into a Score

- **Discretization of qualitative variables**
  The qualitative variables \( q_m \), \( 1 \leq m \leq 19 \) are discretized into modalities. The number of modalities can be equal at 3 or 5 modalities. The rule of the modalities choice is based on the logical relationship between modalities and default.

- **Transformation of quantitative variable into score**
  Let \( (M_{q_m}) \), \( l = 1, \ldots, l_{q_m} \), be the modalities of the qualitative variable \( q_m \) and \( (l_{q_m}) \) defines the number of modalities \( l_{q_m} \in \{3,5\} \). For each modality, the score varies between 0 and 100 points with a jump of 50 points per modality for the variables at three modalities and a jump of 25 points for the variables at five-modalities. The score taken by the modalities is:

  - Variables at three modalities: \([0, 50, 100]\)
    Example: the modalities relating to the sector default rate are: 1- below average, 2- equal to average, 3- above average. In this case, the scores given are respectively: 100, 50, 0.
  - Variables at five modalities: \([0, 25, 50, 75, 100]\)
    Example: the modalities relating to natural risk are: 1- No risk, 2- Low risk and the adequate crisis plan, 3- High risk and the adequate crisis plan, 4- Low risk without crisis plan, 5- High risk without crisis plan. In this case, the scores given are respectively: 100, 75, 50, 25, 0.

- The assessment of the logical relationship between the modalities of each variable and the default is determined on the basis of expert opinion.

Univariate analysis and the determination of explanatory variables

- **Univariate analysis**
  The objective of univariate analysis is to determine the relationship between a company’s default and each of its quantitative and qualitative variables. The default is modeled by a binary variable \( Y \) defined as follows:
  
  \[
  Y = \begin{cases} 
  1 & \text{if the company is healthy} \\
  0 & \text{if the company is in default} 
  \end{cases}
  \tag{4}
  \]

  The relationship between the variable \( Y \) to be explained and the explanatory variables \( V_j \) and \( q_m \) is determined by the logistic regression model.

  Indeed, the definition (4) shows that the variable \( Y \) is a Bernoulli variable which takes the values 0 and 1 of parameter \( p \) with \( P(Y = 1) = p \) and \( P(Y = 0) = 1 - p \). Therefore, the formula (4) can be written in probability as follows:

  \[
  P(Y = y_i) = p^{y_i}(1 - p)^{1-y_i}
  \tag{5}
  \]

  The logistic regression relationship consists in defining a relationship between \( Y \) and a logistic probability \( (p) \) defined from the variables \( (V_j) \) and \( (q_m) \). Indeed, we need to make a logistical transformation (Logit) to bind the dependent variable \( Y \) to each variable among the quantitative variables \( (V_j) \), \( j = 1, \ldots, 16 \) and qualitative variables \( (q_m) \), \( m = 1, \ldots, 19 \).

  For each quantitative variable \( V_j \) or qualitative \( q_m \) the relationship between the default and the variables studied is defined by the probability:

  \[
  p_{0,j} = P(Y = 1/V_j) \quad \text{or} \quad p_{0,q_m} = P(Y = 1/q_m)
  \]

  Indeed, the modelling of \( Y \) conditionally to the quantitative variable \( V_j \) or qualitative variable \( q_m \), is defined by the model:

  \[
  Y = p_0 + \varepsilon
  \tag{6}
  \]

  With \( p_0 = P(Y = 1/V_j) \) (or \( p_0 = P(Y = 1/q_m) \))

- The modelling of the probability \( p_0 \)
  We will only present the univariate analysis of the

<table>
<thead>
<tr>
<th>Theme</th>
<th>The qualitative variables ( (q_m) )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>( T_1 ):</strong> The sector of activity</td>
<td>( q_1 = q_{1x1} ): Sector fault rate ( q_2 = q_{2x1} ): Regulatory impact next year ( q_3 = q_{3x1} ): Exposure to natural risk</td>
</tr>
<tr>
<td><strong>( T_2 ):</strong> The company's positioning and competition</td>
<td>( q_4 = q_{1x2} ): Competitive position and intensity ( q_5 = q_{2x2} ): Barriers and new entrants ( q_6 = q_{3x2} ): International Competition</td>
</tr>
<tr>
<td><strong>( T_3 ):</strong> The concentration and position of the counterparty vis-à-vis its suppliers and customers</td>
<td>( q_7 = q_{1x3} ): Customer concentration ( q_8 = q_{2x3} ): Supplier concentration ( q_9 = q_{3x3} ): Positions vis-à-vis suppliers and customers</td>
</tr>
<tr>
<td><strong>( T_4 ):</strong> Quality and management structure</td>
<td>( q_{10} = q_{1x4} ): Succession planning and business continuity ( q_{11} = q_{2x4} ): Experience Chairman ( q_{12} = q_{3x4} ): Seniority of the principal operational staff ( q_{13} = q_{4x4} ): Capital distribution ( q_{14} = q_{5x4} ): Compliance with the accounting documents delivery schedule ( q_{15} = q_{6x4} ): Performance last crisis ( q_{16} = q_{7x4} ): Existence of agent’s insurance</td>
</tr>
<tr>
<td><strong>( T_5 ):</strong> The company’s history with banking</td>
<td>( q_{17} = q_{1x5} ): Number of payment incidents in the last 12 months ( q_{18} = q_{2x5} ): Percentages of unpaid bills over the last 12 months</td>
</tr>
<tr>
<td><strong>( T_6 ):</strong> Relations with banks</td>
<td>( q_{19} = q_{1x6} ): Number of banks related to the company</td>
</tr>
</tbody>
</table>

Table 2: The list of qualitative variables
quantitative variables $V_j$ since we will proceed in the same way for the analysis of the qualitative variables $q_m$.

Indeed, for each company ($i$), the value of the quantitative variable $V_j$ is $v_{ij}$ and let $p_{oi}$ be the probability that the company ($i$) is healthy. Indeed, $p_{oi}$ is defined by $P(Y_i = 1/V_j = v_{ij})$ with:

$$Y_i = \begin{cases} 1 & \text{if the company}(i)\text{ is healthy} \\ 0 & \text{if the company}(i)\text{ is in default} \end{cases}$$

Thus, the formula (6) for company ($i$) is written:

$$y_i = p_{oi} + \varepsilon_i$$

First, we will define the ratio $p_{oi}$ noted the Odds of $p_{oi}$:

$$\text{Odds}(p_{oi}) = \frac{p_{oi}}{1 - p_{oi}} = \frac{P(Y_i = 1/V_j = v_{ij})}{1 - P(Y_i = 1/V_j = v_{ij})}$$

Consequently, the logistic regression model establishes a linear regression relationship between $\ln \left( \frac{p_{oi}}{1-p_{oi}} \right)$ and $v_{ij}$ as follows:

$$\text{Logit}(p_{oi}) = \ln \left( \frac{p_{oi}}{1-p_{oi}} \right) = \beta_0 + \beta_1 \times v_{ij}$$

where from $p_{oi} = \frac{e^{\beta_0 + \beta_1 \times v_{ij}}}{1 + e^{\beta_0 + \beta_1 \times v_{ij}}}$.

Therefore, the formula (6) and the associated probability 8 $p_{oi}$ can be written:

$$Y = p_{oi} + \varepsilon$$

$$p_{oi} = \frac{e^{\beta_0 + \beta_1 \times v_{ij}}}{1 + e^{\beta_0 + \beta_1 \times v_{ij}}}$$

The estimation of parameter

Let $Y_i$ be the random variable that models the company's default $(i)$. The conditional variable $(Y_i/V_j = v_{ij})$ can be written in probability as follows:

$$P(Y_i/V_j = v_{ij}) = P(Y_i = y_i/V_j = v_{ij}) = p_{oi} y_i \times (1 - p_{oi})^{1-y_i}$$

The likelihood function:

The likelihood function of $n$ enterprise is defined by:

$$L = \prod_{i=1}^n P(Y_i/V_j = v_{ij})$$

The Loglikelihood function ($LL$)

$$LL = \sum_{i=1}^n y_i \times \ln \left( \frac{e^{\beta_0 + \beta_1 \times v_{ij}}}{1 + e^{\beta_0 + \beta_1 \times v_{ij}}} \right)$$

$$+ \sum_{i=1}^n \left( 1 - y_i \right) \times \ln \left( \frac{1}{1 + e^{\beta_0 + \beta_1 \times v_{ij}}} \right)$$

Testing of the significance of the coefficients (test de Wald)

The testing of the significance of the coefficient $\beta_1$, is established by the Wald test. The Wald test is obtained by testing the hypothesis $H_0$ formulated as follows:

$$\{ H_0: \beta_1 = 0 \} \quad \{ H_1: \beta_1 \neq 0 \}$$

The Wald test is based to the following ratio:

$$W_1 = \frac{\hat{\beta}_1^2}{\hat{\sigma}_{\beta_1}^2}$$

The ratio, under the hypothesis $H_0$, will follow a $\chi^2$ with one degree of freedom. We reject $H_0$ if $W_1 > \chi^2_1$. 

The discriminatory power (power stat)

The discriminatory power represents the model's ability to predict future situations. We will use the ROC curve to determine the discriminatory power of each variable $V_j$ and $q_m$. The determination of the ROC curve will be done from the classification table of the sample of estimation of the variable $Y$ which is presented as follows:
Table 3. The classification table

<table>
<thead>
<tr>
<th>Healthy ((Y = 1))</th>
<th>Default ((Y = 0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthy ((Y = 1))</td>
<td>True Healthy (TH)</td>
</tr>
<tr>
<td></td>
<td>False Healthy (FH)</td>
</tr>
<tr>
<td>Default ((Y = 0))</td>
<td>False Default (FD)</td>
</tr>
<tr>
<td></td>
<td>True Default (TD)</td>
</tr>
</tbody>
</table>

One indicates by sensitivity (\(SV\)), the proportion of the healthy companies classified well: \(SV = \frac{TH}{TH + FH}\) and by specificity (\(SP\)), the proportion of the de companies is in default, classified well: \(SP = \frac{FD}{FD + TD}\).

If one varies the “probability threshold” from which it is considered that a company must be regarded as healthy, the sensitivity and specificity varies. The curve of the points \((1 - SP, SV)\) is the \(ROC\) curve.

- **The determination of explanatory variables**
  To determine the explanatory variables to be retained for modelling, we will carry out a univariate logistic regression analysis for each variable in the chosen list. The choice of variables to be retained for the modelling is based on the discriminatory power of each variable.
  The discriminatory power is determined by using the area under the \(ROC\) curve and the Accuracy ratio (\(AR\)).
  
  - **Definition of the area under the \(ROC\)** curve (\(AUC\)) and the Accuracy ratio (\(AR\))
    - The area under the \(ROC\) curve (\(AUC\))
      The area under the \(ROC\) curve (\(AUC\)) provides an overall measure of model fit (Bewick, Cheek, & Ball, [45]). The \(AUC\) varies from 0.5 (predictive capacity absence for the model) to 1 (perfect predictive aptitude for the model).
    - Accuracy ratio (\(AR\))
      The accuracy ratio is defined by the relationship:
      \[
      AR = 2AUC - 1
      \]  (16)
      The \(AR\) takes values between 0 and 1.

  - **Determination of explanatory variables**
    - The decision rules
      The variables that verify the following two characteristics will be retained:
      - The Area Under the \(ROC\) Curve (\(AUC\)) is superior to 60% and accuracy ratio (\(AR\)) is superior to 20% ;
      - The relationship established between the factor and the default rate must be logical. This qualification is based on expert opinion.
    - Strongly correlated variables
      After selecting the explanatory variables on the basis of the decision rules mentioned above, we will study the correlation between the selected variables. The study of correlations makes it possible to eliminate strongly correlated variables. Indeed, if two or more variables have a correlation coefficient superior to 0.5 \((\rho \geq 0.5)\) then the variable that represents the greatest \(AUC\) will be selected.

2.2.1.2. Multivariate Discriminant Analysis and Definition of the Score Function

2.2.1.2.1. Multivariate Discriminant Analysis

In this paragraph, we will present the modeling of the relationship between quantitative and qualitative variables and defect, through pure logistic regression and logistic regression on main components and we will then present the tests for assessing the fit of the multivariate model.

- **The pure logistic regression**
  Let \(X_{j}, j = 1, \ldots, J\) and \(T_{k}, k = 1, \ldots, 6\) be the respectively the quantitative and qualitative variables retained by the univariate analysis.

The objective of multivariate discriminant analysis by the pure logistic regression, is to study the relationship between the variable to be explained \(Y\) and the explanatory variables \(X_{j}, j = 1, \ldots, J\) and \(T_{k}, k = 1, \ldots, 6\) . In this case, we will separate the modelling of the quantitative and qualitative variables. Indeed, the Models can be formulated by the following relationships:

- **The quantitative model**
  \[
  Y = f_{1}(X_{1}, X_{2}, \ldots, X_{J}) + \epsilon
  \]  (17)

- **The qualitative model**
  \[
  Y = f_{2}(T_{1}, T_{2}, \ldots, T_{6}) + \epsilon
  \]  (18)
  where \(f_{1}\) and \(f_{2}\) are the logistics probabilities.

Let \(n\) be the size of the company sample used for modelling and \(X = (x_{ij}), \:\: 1 \leq i \leq n, 1 \leq j \leq J\), the matrix whose columns are \(X_{j}, 1 \leq j \leq J\) and let \(T = (t_{ik}), 1 \leq i \leq n, 1 \leq k \leq 6\) be the matrix whose columns are \(T_{k}, 1 \leq k \leq 6\).

- **Presentation of the model**

The objective of the modelling is to express the variable \(Y\) by the following model:

\[
Y = p + \epsilon
\]  (19)

With \(p = P(Y = 1/X_{1}, \ldots, X_{J}, T_{1}, \ldots, T_{6})\)

Indeed, let \((x_{i1}, \ldots, x_{ij})\) and \((t_{i1}, \ldots, t_{i6})\) be the quantitative and qualitative data, of the company \((i)\) and \(y_{i}\) the realization of \(Y_{i}\) for this company then \(y_{i}\) is written:

\[
y_{i} = p_{i} + \epsilon_{i}
\]  (20)

Where \(p_{i} = P(Y_{i} = 1/x_{i1}, \ldots, x_{ij}, t_{i1}, \ldots, t_{i6})\)

The Logit transformation has the following form:

\[
Logit(p_{i}) = \ln\left(\frac{p_{i}}{1-p_{i}}\right)
= \ln\left(\frac{P(Y_{i} = 1/x_{i1}, \ldots, x_{ij}, t_{i1}, \ldots, t_{i6})}{P(Y_{i} = 0/x_{i1}, \ldots, x_{ij}, t_{i1}, \ldots, t_{i6})}\right)
= \ln\left(\frac{P(Y_{i} = 1/x_{i1}, \ldots, x_{ij}, t_{i1}, \ldots, t_{i6})}{1-P(Y_{i} = 1/x_{i1}, \ldots, x_{ij}, t_{i1}, \ldots, t_{i6})}\right)
\]  (21)
Thus, Logit($p$) can be written:
\[
\text{Logit}(p) = \ln \left( \frac{p}{1-p} \right)
\]
\[= \ln \left( \frac{P(Y = 1/X_{i1}, \cdots, X_{ij})}{1 - P(Y = 1/X_{i1}, \cdots, X_{ij})} \right) \]

The quantitative modelling of $Y$ by the probability $p_1$

The modelling of $Y$ conditional only on quantitative data is defined by the model:
\[
Y = p_1 + \varepsilon
\]
where $p_1 = P(Y = 1/X_{i1}, \cdots, X_{ij})$
Let $(x_{i1}, \cdots, x_{ij})$ be the quantitative data of the company $(i)$ which represent the realizations of the variables $(X_{i1}, \cdots, X_{ij})$ by the company $(i)$ and $y_i$ the realization of $Y_i$ for this company then $y_i$ is written:
\[
y_i = p_{1i} + \varepsilon_i \quad (22)
\]

Where $p_{1i} = P(Y_i = 1/x_{i1}, \cdots, x_{ij})$
To determine the expression of $p_{1i}$ by the logistic regression model, we will first define the Odds represented by the ratio $\frac{p_{1i}}{1-p_{1i}}$. Hence:
\[
\text{Odds}(p_{1i}) = \frac{p_{1i}}{1-p_{1i}} = \frac{P(Y_i = 1/x_{i1}, \cdots, x_{ij})}{1 - P(Y_i = 1/x_{i1}, \cdots, x_{ij})}
\]
Therefore, we can write the model:
\[
\text{Logit}(p_{1i}) = \ln \left( \frac{p_{1i}}{1-p_{1i}} \right) = \ln \left( \frac{e^{\beta x_i'}}{1 + e^{\beta x_i'}} \right)
\]
\[
= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} = \beta x_i'
\]

Where $\beta = (\beta_0, \cdots, \beta_j)$
and $x_i' = \text{Transpose}(1, x_{i1}, \cdots, x_{ij})$, so:
\[
p_{1i} = \frac{e^{\beta x_i'}}{1 + e^{\beta x_i'}} \quad (23)
\]

Let be $X' = \text{Transpose}(1, x_{i1}, \cdots, x_{ij})$ then $p_1$ and $Y$ can be written
\[
\begin{align*}
p_1 &= \frac{e^{\beta x_i'}}{1 + e^{\beta x_i'}} \\
y &= p_1 + \varepsilon
\end{align*}
\]

The estimation of parameters

Let $y_i$ be the realization of $Y_i$ related the company $(i)$, the conditional variable ($Y_i/x_{i1}, \cdots, x_{ij}$) can be written in probability as follows:
\[
P(Y_i = y_i/x_{i1}, \cdots, x_{ij}) = p_{1i} e^{\beta x_i'} \times (1 - p_{1i})^{1-y_i} \quad (24)
\]

The likelihood function

The likelihood function of $n$ enterprise is defined by:
\[
L = \prod_{i=1}^{n} P(Y_i = y_i/x_{i1}, \cdots, x_{ij}) = \prod_{i=1}^{n} (1 - p_{1i})^{1-y_i} \times p_{1i}^{y_i} \quad (25)
\]

The loglikelihood function (LL)
\[
LL = \sum_{i=1}^{n} y_i \ln \left( \frac{e^{\beta x_i'}}{1 + e^{\beta x_i'}} \right)
\]
\[
+ \sum_{i=1}^{n} (1 - y_i) \ln(1 - \frac{e^{\beta x_i'}}{1 + e^{\beta x_i'}})
\]
\[
= \sum_{i=1}^{n} \ln \left( \frac{1}{1 + e^{\beta x_i'}} + y_i \beta x_i' \right) \quad (26)
\]

Let us note $\beta = (\beta_0, \beta_1, \cdots, \beta_j)$ so the estimate of $\beta$ is to determine $\hat{\beta}$ which maximizes the likelihood function and consequently maximizes the loglikelihood function (LL).

The maximization of LL will be done by solving the equation system determined by the following conditions:
\[
\frac{\partial LL}{\partial \beta} = 0 \quad \text{and} \quad \frac{\partial^2 LL}{\partial \beta^2} < 0
\]

Also, this problem is solved numerically, by the Newton-Raphson algorithm, the score method or the Berndt-Hall-Hall-Hausman method.

The quantitative score function:

The estimation of $\beta$ allows determining the quantitative score function $S_1$ that is written:
\[
S_1 = \hat{\beta} X'
\]

Hence, the probability $p_1$ is estimated by:
\[
p_1 = \frac{e^{S_1}}{1 + e^{S_1}}
\]

The qualitative modelling of $Y$ by the probability $p_2$

The objective of qualitative modelling is to determine the relationship between the variable $Y$ and the themes $T_j$. The values taken by the theme $T_j$ is a linear weighting of the qualitative variables $q_m$ that compose it, noted, $q_{mT_i}$, $1 \leq m \leq p$. Indeed, $T_i$ is defined as follows:
\[
T_i = \alpha_1 q_{1T_i} + \cdots + \alpha_p q_{pT_i} \quad (29)
\]
The modelling of $Y$ conditional on qualitative data only, is defined by the model:

$$Y = p_2 + \varepsilon$$

with $p_2 = P(Y_i = 1/T_1, \ldots, T_6)$

$$p_2 = P(Y_i = 1/t_{i1}, \ldots, t_{i6})$$

Let $(t_{i1}, \ldots, t_{i6})$ be the quantitative data of the company $(i)$ and $y_i$ the realization of $Y_i$ for this company then $y_i$ is written:

$$y_i = p_{2i} + \varepsilon_i$$

(30)

where $p_{2i} = P(Y_i = 1/t_{i1}, \ldots, t_{i6})$

In this case, the oddos of the probability $p_{2i}$ are written:

$$\text{Oddos}(p_{2i}) = \begin{pmatrix} p_{2i} \\ 1 - p_{2i} \end{pmatrix} = \frac{p(Y_i = 1)}{1 - P(Y_i = 1)}$$

Therefore, we can write the model:

$$\text{Logit}(p_{2i}) = \ln \left( \frac{p_{2i}}{1 - p_{2i}} \right) = \tau\tau'$$

where $\tau = (\tau_0, \ldots, \tau_n)$ and $t_i' = \text{Transpose}(1,t_{i1}, \ldots, t_{i6})$ thus:

$$p_{2i} = \frac{e^{\tau t_i'}}{1 + e^{\tau t_i'}}$$

Let be $T' = \text{Transpose}(1,t_{11}, \ldots, t_{66})$ then $p_2$ and $Y$ can be written:

$$\begin{cases} p_2 = \frac{e^{\tau t}}{1 + e^{\tau t}} \\ Y = p_2 + \varepsilon \end{cases}$$

For the estimation of $\hat{\tau} = (\hat{\tau}_0, \ldots, \hat{\tau}_n)$, the same procedure presented for the quantitative model will be used.

- The qualitative score function:

The estimation of $\tau$ allows determining the qualitative score function $S_2$ that is written:

$$S_2 = \hat{\tau} T'$$

Hence, the probability $p_2$ is estimated by:

$$p_2 = \frac{e^{S_2}}{1 + e^{S_2}}$$

The logistics regression on the principal components

The principal components analysis (PCA) is a technique based on the reduced of the explanatory variables by transforming the correlated variables into uncorrelated variable, which explain the maximum amount of variance, called the principal components (PCS), this technique was introduced by Karl Pearson in the early 20th century, and developed by Harold Hotelling in 1933.

- The principal components analysis (CPA)

Let $(Z_j, 1 \leq j \leq p)$ be the $p$ explanatory variables (quantitative and qualitative) and let $(z_{ij})_{n \times p}$ be the $np$ realization of these variables. Let $Z = (z_{ij})_{n \times p}$ be the matrix of these realizations. The column vectors of the matrix $Z$ are $z_1, z_2, z_3, ..., z_p$ who represent the realization of each variable $Z_j$.

Let us note $S = (s_{jk})_{p \times p}$ the covariance matrix of the variables $(Z_j, 1 \leq j \leq p)$ and let $(\lambda_1, \lambda_2, \lambda_3, ..., \lambda_p)$ and $(F_1, F_2, ..., F_p)$ be, respectively, the eigenvalues and the eigenvectors of the matrix $S$. We designated by $L$ the matrix whose columns $l_i, i = 1, \ldots, p$ are the principal components defined by $L = ZF$ with $F = (f_{jk})_{p \times p}$ being the matrix which has as columns the eigenvectors $(F_i)$ of the matrix $S$.

The decomposition into principal component makes it possible to express the vectors $Z_j$ into reduced number of principal components $Z_i = \sum_{k=1}^{p} l_k f_{jk}$, $j = 1, \ldots, p$, which represents a high percentage of cumulative variation (CV) given by the following formula:

$$CV = \left[ \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100 \right], \ s \leq p$$

(31)

The average variation $V_m$ is defined by:

$$V_m = \frac{\sum_{j=1}^{p} \lambda_j}{p} \times 100$$

(32)

- The logistics regression on the principal components

The logistic regression on the $p$ principal components $L_i$ is equivalent to the logistic regression on the explanatory variables $Z_{ij}, j = 1, \ldots, p$. Indeed, let $z_{ij}$, $i = 1, \ldots, n$ be the realizations of the variables $Z_j$ hence $z_{ji} = \sum_{k=1}^{p} l_{ki} f_{jk}$ so

$$p_i = \frac{e^{(\beta_0+\sum_{j=1}^{p} \beta_j z_{ij})}}{1 + e^{(\beta_0+\sum_{j=1}^{p} \beta_j z_{ij})}}$$

$$\sum_{j=1}^{p} \gamma_j \beta_j$$

with $\gamma_k = \sum_{j=1}^{p} f_{jk} \beta_j$, $(k = 1, \ldots, p)$

In order to improve the quality of the model in the case of collinearity, we will use a reduced number of $(s)$ principal components for modelling the variable $Y$. Indeed, the formula (20) becomes:

$$Y_i = p_{l(i)} + \varepsilon_{l(i)}$$

(33)

Hence

$$p_{l(i)} = \frac{e^{(\beta_0+\sum_{k=1}^{s} \gamma_k l_{ki})}}{1 + e^{(\beta_0+\sum_{k=1}^{s} \gamma_k l_{ki})}}$$

(34)
Assessing the Impact of Modelling on the Expected Credit Loss (ECL) of a Portfolio of Small and Medium-sized Enterprises

In our study, the number \( s \) is defined as the number of components whose cumulative variation is superior to 70%.

For the CPA, the quantitative variables and the qualitative variables are modeled simultaneously to determine the probability \( p_{i(s)} \)

- **Assessing the fit of the multivariate model:**
  
  To assess the fit of the multivariate quantitative and qualitative models, we will proceed as follows:
  
  - Testing of the significance of the coefficients
    
    Let be the logistics model defined by:
    
    \[
    y_i = p_i + \varepsilon_i
    \]
    
    With \( p_i = \frac{e^{\beta X_i'}}{1 + e^{\beta X_i'}} \), \( \beta = (\beta_0, \ldots, \beta_m) \) and \( H' = \text{tr}(1, H_1, \ldots, H_m) \).
    
    The significance of the regression parameters \( (\beta_1, \ldots, \beta_m) \) consists in testing the individual and overall significance of the model parameters:
    
    - Testing of the individual significance (Test de Wald)
      
      To assess individual significance, the hypothesis to be tested is:
      
      \[
      \begin{align*}
      H_0: \beta_i &= 0 \\
      H_1: \beta_i &\neq 0
      \end{align*}
      \]
      
      The testing for the significance of the coefficient \( \beta_i \) is established by the Wald test based to the following ratio:
      
      \[
      W_i = \frac{\hat{\beta}_i^2}{\widehat{\sigma}(\hat{\beta}_i)^2}
      \]
      
      The ratio, under the hypothesis \( H_0 \), will follow a \( \chi^2 \) with one degree of freedom. We reject \( H_0 \) if \( W_i > \chi^2_1 \)
    
    - Testing of the overall significance (the likelihood ratio test)
      
      The overall significance test is based on the likelihood ratio between the model without the explanatory variables and the model with the explanatory variables. Indeed, the test can be expressed as follows:
      
      \[
      \begin{align*}
      H_0: \text{logit} \left( P(Y = 1) \right) &= \beta_0 \\
      H_1: \text{logit} \left( P(Y = 1) \right) &= \beta H'
      \end{align*}
      \]
      
      The test ratio is defined by:
      
      \[
      ST = -2 \left( \frac{\text{Maximum loglikelihood of the model } M_1}{\text{Maximum loglikelihood of the model } M_2} \right)
      \]
      
      where:
      
      - \( M_1 \) the model defined by \( \text{logit} \left( P(Y = 1) \right) = \beta_0 \)
      
      - Maximum loglikelihood of the model \( M_1 \):
      
      \[
      LL(M_1) = \sum_{i=1}^{n} \ln \left( \frac{1}{1 + e^{\beta_0 X_i}} \right) + y_i \beta_0
      \]
      
      - Maximum loglikelihood of the model \( M_2 \):
      
      \[
      LL(M_2) = \sum_{i=1}^{n} \ln \left( \frac{1}{1 + e^{\beta X'_i}} \right) + y_i \beta H'
      \]
      
      The ratio, under the hypothesis \( H_0 \), will follow a \( \chi^2 \) with \( n \) degree of freedom. We reject \( H_0 \) if \( ST > \chi^2_n \) and model \( M_2 \) is better than model \( M_1 \)
    
    - The Hosmer-Lemeshow test
      
      The objective of this test, defined in Hosmer et al [29], is to assess the concordance between the predicted and the observed values. Indeed, the data are arranged in ascending order of the probabilities calculated by using the model, then divided into 10 groups.
      
      The test hypotheses are:
      
      \[
      \begin{align*}
      H_0: \text{the model is correct} \\
      H_1: \text{the model is not correct}
      \end{align*}
      \]
      
      The test statistic \( (HL) \), is defined by:
      
      \[
      HL = -\sum_{i=1}^{10} \left( O_i - n_i \hat{p}_i \right)^2 / n_i \hat{p}_i (1 - \hat{p}_i)
      \]
      
      Where \( O_i = \sum_{i=1}^{n} y_j \) is the number of observed outcomes, events, in group \( i \), \( n_i \) is the number of observations in group \( i \), \( \hat{p}_i \) is the average predicted probability by the model in group \( i \). \( HL \), under the hypothesis \( H_0 \), will follow a \( \chi^2 \) with \( 8 \) degree of freedom. We reject \( H_0 \) if \( HL > \chi^2_8 \)
    
    - The performance of the model in the classification of the entreprise
      
      The assessment of the performance of the model is necessary to determine the discriminatory power of the model, and to compare several models. To measure the performance of the model we will use the area under the ROC curve (AUC) presented previously.
      
      A higher \( AUC \) means that the discriminatory power of the model is excellent. Hosmer and al.[29] defines the general rules for classification of the models based on the \( AUC \). Consequently, in our study, the model will be retained if the \( AUC \) is superior or equal to 0.7, which means that the model has an acceptable discriminatory power.
      
      Determination of the Overall Score Function
      
      To determine the model that combines the quantitative explanatory variables \( X_{j, i} \), \( j = 1, \ldots, J \), and qualitative explanatory variables \( T_{i, i} \), \( i = 1, \ldots, 6 \), we have to determine the function \( f \) such that:
      
      \[
      y = f(X_1, \ldots, X_J, T_1, \ldots, T_6) + \varepsilon
      \]
      
      where \( f \) is a logistic probability
    
    - The pure logistic regression model.
      
      Let \( S_1 \) and \( S_2 \) be respectively the quantitative and qualitative score of the company \( (i) \), where \( S_1 = \beta X' \) and \( S_2 = \beta T' \). The conditional probabilities \( p_1 \) and \( p_2 \) related to quantitative and qualitative data, are respectively
The final score $S_i$ of company $(i)$ is defined by the weighted average of the two scores $S_{1i}$ and $S_{2i}$ and the probability $p_i = P(Y = 1/x_{1i}, \ldots, x_{fi}, t_{1i}, \ldots t_{6i})$ is defined by:

$$p_i = \frac{e^{S_{1i}}}{1+e^{S_{1i}}} = \frac{e^{S_{2i}}}{1+e^{S_{2i}}}$$

where $S_i = \alpha S_{1i} + (1 - \alpha)S_{2i}$

The score function $S$ and the probability $p$ of the model can be formulated as follows:

$$p = \frac{e^{\alpha S_{1i} + (1-\alpha)S_{2i}}}{1 + e^{\alpha S_{1i} + (1-\alpha)S_{2i}}} = \frac{e^{S}}{1 + e^{S}}$$

where $S = \alpha S_{1} + (1 - \alpha)S_{2}$

The weighting $\alpha$ is determined according to the discriminatory power of the model. Indeed, the weighting which will be retained is that which maximizes the curve $ROC$ among all the values $10\%, 20\%, 30\%, \ldots, 80\%$ and $90\%$.

- The principal components of the logistic regression model.

In this case, we will simultaneously treat the quantitative and qualitative variables. As a result, the probability $p_i = P(Y = 1/x_{1i}, \ldots, x_{fi}, t_{1i}, \ldots t_{6i})$ of the company $(i)$ will be a logistic probability determined directly as a function of the eigenvectors and the cumulative variation chosen.

The formulation of the $S$ score function and the probability $p$ are determined directly by multivariate analysis of the logistic regression between the default and the principal components.

Determination of the Score Grid, and Calculation of the Probability of Default by Class.

A company $(i)$ is considered healthy if $p_i \geq \frac{1}{2}$, so:

$$p_i = \frac{e^{S_{i}}}{1 + e^{S_{i}}} \geq \frac{1}{2} \iff e^{S_{i}} \geq \frac{1}{2} (1 + e^{S_{i}})$$

$$\iff e^{S_{i}} \geq 1 \iff S_{i} \geq 0$$

Consequently, the company is healthy if the score is positive.

2.2.1.2.2. The Rating Grid

The classification of healthy companies is based on the score function. Indeed, this classification gives rise to the rating grid composed of 8 rating classes.

Each company $(i)$ is classified into a rating class, the classes vary between $A$ and $H$, and are defined as follows:

<table>
<thead>
<tr>
<th>Class</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating score</td>
<td>[86-100]</td>
<td>[76-86]</td>
<td>[65-76]</td>
<td>[55-65]</td>
<td>[46-55]</td>
<td>[40-46]</td>
<td>[30-40]</td>
<td>&lt;30</td>
</tr>
</tbody>
</table>

The rating score $SN_i$ of the company $(i)$ is defined by:

$$SN_i = \frac{S_i}{Max(S)} \times 100$$  \hspace{1cm} (40)

where $S_i$ is a score function of the company $(i)$.

2.2.1.2.3. Calculation of the Probability of Default per Rating Class

The probability of default of the class $K$ ($PD_K$) is defined by the probability of default of the company $(i)$ knowing that the company $(i)$ belonging to the class $K$. Consequently:

$$PD_K = \frac{\text{Number of companies in default belonging to the class } K}{\text{Total number of companies belonging to the class } K}$$  \hspace{1cm} (41)

To define this probability of default per rating class, we will distribute the sample of healthy and default companies as follows:

<table>
<thead>
<tr>
<th>Class</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating score</td>
<td>[86-100]</td>
<td>[76-86]</td>
<td>[65-76]</td>
<td>[55-65]</td>
<td>[46-55]</td>
<td>[40-46]</td>
<td>[30-40]</td>
<td>&lt;30</td>
</tr>
<tr>
<td>Number ($Y = 0$)</td>
<td>$n_{40}$</td>
<td>$n_{80}$</td>
<td>$n_{40}$</td>
<td>$n_{20}$</td>
<td>$n_{20}$</td>
<td>$n_{20}$</td>
<td>$n_{20}$</td>
<td>$n_{20}$</td>
</tr>
<tr>
<td>Number ($Y = 1$)</td>
<td>$n_{41}$</td>
<td>$n_{81}$</td>
<td>$n_{61}$</td>
<td>$n_{61}$</td>
<td>$n_{61}$</td>
<td>$n_{61}$</td>
<td>$n_{61}$</td>
<td>$n_{61}$</td>
</tr>
<tr>
<td>Total number</td>
<td>$n_A$</td>
<td>$n_B$</td>
<td>$n_C$</td>
<td>$n_D$</td>
<td>$n_E$</td>
<td>$n_F$</td>
<td>$n_G$</td>
<td>$n_H$</td>
</tr>
<tr>
<td>Probability of Default ($PD$)</td>
<td>$p_{40}$</td>
<td>$p_{80}$</td>
<td>$p_{60}$</td>
<td>$p_{60}$</td>
<td>$p_{60}$</td>
<td>$p_{60}$</td>
<td>$p_{60}$</td>
<td>$p_{60}$</td>
</tr>
</tbody>
</table>
2.2.2. Determination of Expected Credit Losses

The expected loss is calculated by the formula (1). It depends on three components which are the PD, LGD and EAD. The process for calculating the PD was discussed in the previous paragraphs while the calculation of the LGD and EAD will be presented in the following paragraphs.

2.2.2.1. Determination of the Loss Given Default (LGD)

The calculation of the loss given default (LGD) consists in constructing regression models, which associated the LGD with relevant factors such as the seniority of the debt, the seniority of the relationship, the date of the transfer to the banking litigation, the activity sector the company, the collection process adopted by the bank, the time required for the completion of litigation procedures in justice, the rate of coverage by the first-ranking guarantees, the economic conditions and the probability at default (PD).

These variables were analyzed and used for the calibration of the model LGD in numerous publications. Indeed, Chalupka and Kopcsni [19] present a case study on LGD modelling of bank loans, in which the principal factors that have been identified are the period of issuance of the loan, the quality of the guarantees, the amount of the loan and the duration of the relationship with the debtor. The relationship between the default rate and the recovery rate was investigated in Altman and al ([3],[4],[6],[7]). The study of the dependence between default probabilities and recovery rates by Bade and al. [9], has shown some improvement in the LGD. While Gurtler et Hibbeln [27] has studied the influence of the length of the recovery process (training) on the LGD level estimated.

Different regression models have been used by researchers to model loss given default (LGD) as the Tobit model found in McDonald and al.[33], the Beta regression in Huang and al. [30], the inflated Beta regression in Pereira and al. [36], the censored gamma regression in Sigrist and al.[38] and [39] and a mixture of distributions into the model used, by Altman et al. [5].

In this paper, we will use the estimate provided by the Basel Accord for credit conversion factor (CCF) at the paragraph level 366. Indeed, under the IRB foundation approach, the CCF is fixed at 75%. The exposure at default (EAD) of the line of credit is determined as follows:

\[ EAD = \text{VCB}_0 + \text{CCF} \times \text{VCHB}_0 \]

where \( \text{M}_{efa} \) the amount of the financing authorization granted by the bank to the customer.

2.2.2.2. Determination Exposure at Default (EAD)

The exposure at default \( EAD \) is defined as the sum of:

- The value accounted for in the balance sheet \( (\text{VCB}_0) \).
- The value of the unused funding commitment, accounted for off-balance sheet \( (\text{VCHB}_0) \) multiplied by a credit conversion factor \( (\text{CCF}) \).

The credit conversion factor \( (\text{CCF}) \) is defined between 0 and 1 \( (\text{FCC} \in [0,1]) \) and the mathematical formulation of the EAD, is given by the following relationship:

\[ EAD = \text{VCB}_0 + \text{CCF} \times \text{VCHB}_0 \]  

The calculation of the exposure at default (EAD) will be done in two different ways. The first, consists in modelling the credit conversion factor (CCF) by using regression models that associated the CCF with pertinent factors, whereas the second, consists in modelling directly the exposure at default (EAD).

The models used to model the credit conversion factor (CCF) such as the ordinary least squares (OLS), the Tobit model, the fractional response regression and the utilization change model, are detailed in Brown [17], Bellotti and al. [14] and Bijak and al. [15]. While the models used for the direct modelling of the EAD as a zero-adjusted gamma model and others, are detailed in E.N.C. Tong and al.[22].

In this paper, we will use the estimate provided by the Basel Accord for credit conversion factor (CCF) at the paragraph level 366. Indeed, under the IRB foundation approach, the CCF is fixed at 75%. The exposure at default (EAD) of the line of credit is determined as follows:

\[ EAD = \text{VCB}_0 + 75\% \times \text{VCHB}_0 \]

\[ EAD = \text{VCB}_0 + 75\% \times (\text{M}_{efa} - \text{VCB}_0) \]  

\[ EAD = 25\% \text{VCB}_0 + 75\% \times \text{M}_{efa} \]

3. Empirical Results

3.1. Description of the Database

In this study, we used a database of small and medium enterprises (SMEs) of a Moroccan bank composed of 1447 enterprises. For the definition of a SME we have based on the definition of the Central Bank which considers that an enterprise is an SME if it realizes a turnover between 10 and 175 million MAD. In terms of default, the portfolio structure is as follows:

<table>
<thead>
<tr>
<th>Table 6. Default Portfolio Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
</tr>
<tr>
<td>Healthy Enterprises</td>
</tr>
<tr>
<td>Enterprises in default</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

3.2. Choice of Explanatory Variables

3.2.1. Choice of Quantitative Variables

The univariate analysis of the quantitative variables has permitted to determine the quantitative variables which explain the failure. The choice of quantitative variables, is based on the Wald test and the discriminatory power (power stat), determined from the AUC and the AR. Indeed, the Wald test and the discriminatory power made have permitted to determine the seven explanatory variables follows:

### Table 7. univariate analysis and choice of quantitative variables

<table>
<thead>
<tr>
<th>Title variable</th>
<th>variable</th>
<th>Khi² (Wald)</th>
<th>( p_v = Pr(x &gt; \text{Wald}) )</th>
<th>AUC</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profit growth</td>
<td>( X_1 = V_3 )</td>
<td>4,782</td>
<td>0,029</td>
<td>61%</td>
<td>22%</td>
</tr>
<tr>
<td>Financial expenses</td>
<td>( X_2 = V_6 )</td>
<td>9,427</td>
<td>0,002</td>
<td>65%</td>
<td>30%</td>
</tr>
<tr>
<td>Sales/turnover (ST)</td>
<td>( X_3 = V_7 )</td>
<td>23,912</td>
<td>&lt; 0,0001</td>
<td>69.2%</td>
<td>38.4%</td>
</tr>
<tr>
<td>Net Debt Equity</td>
<td>( X_4 = V_6 )</td>
<td>23,683</td>
<td>&lt; 0,0001</td>
<td>69%</td>
<td>38%</td>
</tr>
<tr>
<td>Liquidity</td>
<td>( X_5 = V_{10} )</td>
<td>6,136</td>
<td>0,013</td>
<td>63.3%</td>
<td>26.6%</td>
</tr>
<tr>
<td>Total Assets</td>
<td>( X_6 = V_{12} )</td>
<td>6,590</td>
<td>0,010</td>
<td>64%</td>
<td>28%</td>
</tr>
<tr>
<td>Sales/turnover (ST) Net fixed assets + WCR</td>
<td>( X_7 = V_{13} )</td>
<td>10,289</td>
<td>0,001</td>
<td>65%</td>
<td>30%</td>
</tr>
</tbody>
</table>

For the variables listed in the table 7, the Wald's test shows that the selected variables are significant because the p-value(\( p_v \)) is inferior to 0,05 with a satisfactory discriminatory power because the AUC is superior to 0,6 and the AR is superior to 0,2.

The correlation between the selected quantitative variables is presented as follows:

### Table 8. The correlation matrix of quantitative variables.

<table>
<thead>
<tr>
<th>Variables</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
<th>( X_5 )</th>
<th>( X_6 )</th>
<th>( X_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>1,000</td>
<td>0,128</td>
<td>0,076</td>
<td>0,060</td>
<td>0,067</td>
<td>0,109</td>
<td>0,008</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0,128</td>
<td>1,000</td>
<td>-0,021</td>
<td>0,069</td>
<td>0,087</td>
<td>0,213</td>
<td>-0,084</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0,076</td>
<td>-0,021</td>
<td>1,000</td>
<td>0,415</td>
<td>0,195</td>
<td>0,085</td>
<td>0,440</td>
</tr>
<tr>
<td>( X_4 )</td>
<td>0,060</td>
<td>0,069</td>
<td>0,415</td>
<td>1,000</td>
<td>0,297</td>
<td>0,357</td>
<td>0,358</td>
</tr>
<tr>
<td>( X_5 )</td>
<td>0,067</td>
<td>0,087</td>
<td>0,195</td>
<td>0,297</td>
<td>1,000</td>
<td>0,371</td>
<td>-0,027</td>
</tr>
<tr>
<td>( X_6 )</td>
<td>0,109</td>
<td>0,213</td>
<td>0,085</td>
<td>0,357</td>
<td>0,371</td>
<td>1,000</td>
<td>-0,159</td>
</tr>
<tr>
<td>( X_7 )</td>
<td>0,008</td>
<td>-0,084</td>
<td>0,440</td>
<td>0,358</td>
<td>-0,027</td>
<td>-0,159</td>
<td>1,000</td>
</tr>
</tbody>
</table>

The previous matrix shows that the variables are not strongly correlated because the correlation coefficients do not exceed 0,5.

3.2.2. Choice of Qualitative Variables

Similarly, for the qualitative variables, the choice is based on Wald's test and on the discriminatory power (power stat). The following table summarizes the results.

### Table 9. univariate analysis and choice of qualitative variables.

<table>
<thead>
<tr>
<th>( q_m )</th>
<th>theme</th>
<th>qualitative variables</th>
<th>Khi² (Wald)</th>
<th>( p_v = P(x &gt; \text{Wald}) )</th>
<th>AUC</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q_1 )</td>
<td>( T_3 )</td>
<td>Sector fault rate</td>
<td>7,378</td>
<td>0,007</td>
<td>61%</td>
<td>23%</td>
</tr>
<tr>
<td>( q_2 )</td>
<td>( T_2 )</td>
<td>International Competition</td>
<td>3,875</td>
<td>0,049</td>
<td>60,5%</td>
<td>21%</td>
</tr>
<tr>
<td>( q_{12} )</td>
<td>( T_4 )</td>
<td>Seniority of the principal operational staff</td>
<td>3,840</td>
<td>0,050</td>
<td>60%</td>
<td>20%</td>
</tr>
<tr>
<td>( q_{17} )</td>
<td>( T_5 )</td>
<td>Number of payment incidents in the last 12 months</td>
<td>141,290</td>
<td>&lt; 0,0001</td>
<td>79,9%</td>
<td>59,8%</td>
</tr>
<tr>
<td>( q_{19} )</td>
<td>( T_6 )</td>
<td>Number of banks related to the company</td>
<td>11,227</td>
<td>0,001</td>
<td>65,4%</td>
<td>30,8%</td>
</tr>
</tbody>
</table>

For each theme, we found that only one variable is significant except for theme \( T_3 \) where the variables are not significant. Consequently, the modelling of the themes is equivalent to the modelling of the selected variables.
As with the quantitative variables, the Wald's test shows that the selected variables are significant because the p-value ($p_{v}$) is inferior to 0,05. For the discriminatory power, the variables have a satisfying discriminatory power because the $AUC$ is superior to 0,6 and the $AR$ is superior to 0,2.

The correlation between the chosen qualitative variables is presented as follows:

<table>
<thead>
<tr>
<th>Variables</th>
<th>$T_1$</th>
<th>$T_2$</th>
<th>$T_4$</th>
<th>$T_5$</th>
<th>$T_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>1,000</td>
<td>0,072</td>
<td>-0,079</td>
<td>-0,012</td>
<td>0,053</td>
</tr>
<tr>
<td>$T_2$</td>
<td>0,072</td>
<td>1,000</td>
<td>0,029</td>
<td>-0,039</td>
<td>0,078</td>
</tr>
<tr>
<td>$T_4$</td>
<td>-0,079</td>
<td>0,029</td>
<td>1,000</td>
<td>-0,035</td>
<td>-0,040</td>
</tr>
<tr>
<td>$T_5$</td>
<td>-0,012</td>
<td>-0,039</td>
<td>-0,035</td>
<td>1,000</td>
<td>0,043</td>
</tr>
<tr>
<td>$T_6$</td>
<td>0,053</td>
<td>0,078</td>
<td>-0,040</td>
<td>0,043</td>
<td>1,000</td>
</tr>
</tbody>
</table>

The previous matrix shows that the variables are not strongly correlated because the correlation coefficients do not exceed 0,5.

### 3.3. Calculation of the Expected Loss from the Rating Model based on Pure Logistic Regression.

We have previously presented the process of construction of the rating tool. Indeed, we will determine in this paragraph the rating grid and the probability of default by class by using pure logistic regression.

#### 3.3.1. Multivariate Analysis and Determination of the Score Function.

#### 3.3.1.1. Multivariate Analysis and Determination of the Score Function of Quantitative Variables.

- **Parameter estimation and testing of significance.**

  The parameters estimation and the tests of significance of the coefficients are presented as follows:

  - The parameters estimations and the Wald’s test.
  - The testing of the overall significance (the likelihood ratio test).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard deviation</th>
<th>$\chi^2$ (Wald)</th>
<th>$p_{v}$ = $P(x &gt; Wald)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cte</td>
<td>$\beta_0$=0,706</td>
<td>0,356</td>
<td>3,942</td>
<td>0,047</td>
</tr>
<tr>
<td>$X_1$</td>
<td>$\beta_1$=0,004</td>
<td>0,003</td>
<td>4,050</td>
<td>0,044</td>
</tr>
<tr>
<td>$X_2$</td>
<td>$\beta_2$=0,009</td>
<td>0,004</td>
<td>5,223</td>
<td>0,022</td>
</tr>
<tr>
<td>$X_3$</td>
<td>$\beta_3$=0,009</td>
<td>0,003</td>
<td>9,104</td>
<td>0,003</td>
</tr>
<tr>
<td>$X_4$</td>
<td>$\beta_4$=0,007</td>
<td>0,003</td>
<td>4,431</td>
<td>0,035</td>
</tr>
<tr>
<td>$X_5$</td>
<td>$\beta_5$=0,002</td>
<td>0,001</td>
<td>3,896</td>
<td>0,048</td>
</tr>
<tr>
<td>$X_6$</td>
<td>$\beta_6$=0,003</td>
<td>0,001</td>
<td>4,355</td>
<td>0,037</td>
</tr>
<tr>
<td>$X_7$</td>
<td>$\beta_7$=0,003</td>
<td>0,002</td>
<td>4,225</td>
<td>0,040</td>
</tr>
</tbody>
</table>

Wald's test confirms the significance of the coefficients $\beta_i$ because the p-value ($p_{v}$) of all variables is inferior to 0,05.

- The testing of the overall significance (the likelihood ratio test)

  The likelihood ratio test shows that the value of the $ST$ statistic, previously defined, is equal to 46,883, this value is superior to 14,067 ($\chi^2$ of degree 7 at the 0,05 threshold). Consequently, we reject $H_0$ and the variables are totally significant and determine a better model.

- **Hosmer-Lemeshow test and the assessment of the performance of the model.**

  - The Hosmer-Lemeshow test

    The Hosmer-Lemeshow test shows that the model fits with the sample data. Indeed, the $HL$ statistic is equal to 4,42 with a p-value equal to 0,817, superior to the threshold of 0,05.

  - The assessment of the performance of the model

    The $ROC$ curve shows that the model offers an acceptable classification of companies because the $AUC$ is equal to 0,705.
Determination of the score function.
For each company \((l)\) the score of company \(S_1(l)\) is written as follows:

\[
S_1(l) = \beta_0 + \beta_1 X_{1l} + \beta_2 X_{2l} + \beta_3 X_{3l} + \beta_4 X_{4l} + \beta_5 X_{5l} + \beta_6 X_{6l} + \beta_7 X_{7l}
\]

As a result, the variable \(Y_l\) can be estimated by the probability:

\[
p_{1l} = \frac{e^{\beta_0 + \beta_1 X_{1l} + \beta_2 X_{2l} + \beta_3 X_{3l} + \beta_4 X_{4l} + \beta_5 X_{5l} + \beta_6 X_{6l} + \beta_7 X_{7l}}}{1 + e^{\beta_0 + \beta_1 X_{1l} + \beta_2 X_{2l} + \beta_3 X_{3l} + \beta_4 X_{4l} + \beta_5 X_{5l} + \beta_6 X_{6l} + \beta_7 X_{7l}}}
\]

Hence, the variable \(Y\) can be estimated by the probability:

\[
p_1 = \frac{e^{0.04X_1 + 0.04X_2 + 0.05X_3 + 0.06X_4 + 0.07X_5 + 0.07X_6 + 0.07X_7}}{1 + e^{0.04X_1 + 0.04X_2 + 0.05X_3 + 0.06X_4 + 0.07X_5 + 0.07X_6 + 0.07X_7}}
\]

### 3.3.1.2. Multivariate Analysis and Determination of the Score Function of Qualitative Variables.

Parameter estimation and testing of significance.

The parameters estimation and the tests of significance of the coefficients are presented as follows:

- The parameters estimations and the Wald’s test.
- The maximization of the likelihood function is done by the Newton-Raphson algorithm. The results of the estimation of the parameters and Wald's test are presented as follows:

<table>
<thead>
<tr>
<th>(\tau_i)</th>
<th>Standard deviation</th>
<th>(\text{Khi}^2) (Wald)</th>
<th>(p_{Wald})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>(\tau_0 = -5.650)</td>
<td>0.755</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>(T_1)</td>
<td>(\tau_1 = 0.009)</td>
<td>0.003</td>
<td>9.915</td>
</tr>
<tr>
<td>(T_2)</td>
<td>(\tau_2 = 0.007)</td>
<td>0.004</td>
<td>4.221</td>
</tr>
<tr>
<td>(T_3)</td>
<td>(\tau_3 = 0.012)</td>
<td>0.005</td>
<td>7.685</td>
</tr>
<tr>
<td>(T_4)</td>
<td>(\tau_4 = 0.083)</td>
<td>0.007</td>
<td>138.743</td>
</tr>
<tr>
<td>(T_5)</td>
<td>(\tau_5 = 0.012)</td>
<td>0.005</td>
<td>6.642</td>
</tr>
</tbody>
</table>

Wald's test confirms the significance of the coefficients \(\tau_i\) because the p-value \((p_{Wald})\) of all variables is inferior to 0.05.

- The testing of the overall significance (the likelihood ratio test)
The likelihood ratio test shows that the value of the $ST$ statistic, previously defined, is equal to 178,364, this value is superior to 14,067 ($\chi^2$ of degree 7 at the 0.05 threshold). Consequently, we reject $H_0$ and the variables are totally significant and determine a better model.

- **Hosmer-Lemeshow test and the assessment of the performance of the model.**
  - The Hosmer-Lemeshow test
    The Hosmer-Lemeshow test shows that the model fits with the sample data. Indeed, the $HE$ statistic is equal to 12,265 with a p-value equal to 0.092, superior to the threshold of 0.05.
  - The assessment of the performance of the model
    The $IRIR$ curve shows that the model offers an acceptable classification of companies because the $BUBU$ is equal to 0.784.

![Figure 4. The AUC of quantitatives variables](image)

3.3.2. The Function of the score function.

For each company $(i)$ the score of company $S_2(i)$ is written as follows:

$$S_2(i) = \tau_0 + \tau_1 T_{1,1} + \tau_2 T_{1,2} + \tau_3 T_{1,4} + \tau_4 T_{1,5} + \tau_5 T_{1,6}$$

As a result, the variable $Y_i$ can be estimated by the probability:

$$p_{2i} = \frac{e^{\tau_0 + \tau_1 T_{1,1} + \tau_2 T_{1,2} + \tau_3 T_{1,4} + \tau_4 T_{1,5} + \tau_5 T_{1,6}}}{1 + e^{\tau_0 + \tau_1 T_{1,1} + \tau_2 T_{1,2} + \tau_3 T_{1,4} + \tau_4 T_{1,5} + \tau_5 T_{1,6}}}$$

$$p_{2i} = \frac{e^{-5.650 + 0.009 \times T_{1,1} + 0.007 \times T_{1,2} + 0.012 \times T_{1,4} + 0.083 \times T_{1,5} + 0.012 \times T_{1,6}}}{1 + e^{-5.650 + 0.009 \times T_{1,1} + 0.007 \times T_{1,2} + 0.012 \times T_{1,4} + 0.083 \times T_{1,5} + 0.012 \times T_{1,6}}}$$

Hence, the variable $Y$ can be estimated by the probability:

$$p_2 = \frac{e^{-5.650 + 0.009 \times T_1 + 0.007 \times T_2 + 0.012 \times T_4 + 0.083 \times T_5 + 0.012 \times T_6}}{1 + e^{-5.650 + 0.009 \times T_1 + 0.007 \times T_2 + 0.012 \times T_4 + 0.083 \times T_5 + 0.012 \times T_6}}$$

3.3.2.1. Determination of the Function of Overall Score and the Construction of the Rating Grid.

3.3.2. The Function of Overall Score.

The combination of the quantitative and qualitative score to determine the overall score is based on the maximization of the discriminatory power of the model. Therefore, we will retain the combination $\alpha S_1 + (1 - \alpha) S_2$ that maximizes the $AUC$. To achieve this, we have varied $\alpha$ and we calculated the $AUC$. The results are presented below:
### Table 13. Variation in the weighting ($\alpha$) and the $AUC$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>40%</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AUC$</td>
<td>0.79</td>
<td>0.795</td>
<td>0.799</td>
<td>0.8016</td>
<td>0.80338</td>
<td>0.80226</td>
<td>0.798</td>
<td>0.78</td>
<td>0.744</td>
</tr>
</tbody>
</table>

The value of ($\alpha$) equal to 50% is the one that maximizes the $AUC$. Therefore, we have retained the model that gives the same weight to the quantitative and qualitative score.

3.3.2.2. The Construction of the Rating Grid.

The score grid and the rating classes defined per score and the distribution of the enterprises of the sample per the rating classes, are as follows:

### Table 14. The score grid, the rating classes and the distribution of the enterprises

<table>
<thead>
<tr>
<th>Rating classes</th>
<th>Score</th>
<th>Number</th>
<th>Distribution of the companies</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>[86-100]</td>
<td>19</td>
<td>1.3%</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>[76-86]</td>
<td>279</td>
<td>19.3%</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>[65-76]</td>
<td>460</td>
<td>31.8%</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>[55-65]</td>
<td>372</td>
<td>25.7%</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>[46-55]</td>
<td>164</td>
<td>11.3%</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>[40-46]</td>
<td>55</td>
<td>3.8%</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>[30-40]</td>
<td>50</td>
<td>3.5%</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>&lt;30</td>
<td>48</td>
<td>3.3%</td>
<td></td>
</tr>
</tbody>
</table>

The distribution of the portfolio is presented in the following graph:

![Figure 5. The distribution of the portfolio enterprises](image)

3.3.3. Determination of the Probability of Default per Class and Calculation of the Expected Credit Loss ($ECL$).

3.3.3.1. Determination of the Probability of Default per Rating Class.

For this first model, the probability of default is determined according to the approach detailed above. Therefore, the $PD$ per rating class is presented as follows:

### Table 15. Model 1, probability of default by rating class

<table>
<thead>
<tr>
<th>Rating class</th>
<th>Number of healthy enterprises</th>
<th>Number of enterprises in default</th>
<th>PD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19</td>
<td>0</td>
<td>0.03%</td>
</tr>
<tr>
<td>B</td>
<td>276</td>
<td>3</td>
<td>1.08%</td>
</tr>
<tr>
<td>C</td>
<td>445</td>
<td>15</td>
<td>3.26%</td>
</tr>
<tr>
<td>D</td>
<td>345</td>
<td>27</td>
<td>7.26%</td>
</tr>
<tr>
<td>E</td>
<td>150</td>
<td>14</td>
<td>8.54%</td>
</tr>
<tr>
<td>F</td>
<td>50</td>
<td>5</td>
<td>9.09%</td>
</tr>
<tr>
<td>G</td>
<td>36</td>
<td>14</td>
<td>28.00%</td>
</tr>
<tr>
<td>H</td>
<td>12</td>
<td>36</td>
<td>75.00%</td>
</tr>
<tr>
<td>Total</td>
<td>1333</td>
<td>114</td>
<td></td>
</tr>
</tbody>
</table>

3.3.3.2. Calculation of Expected Credit Loss ($ECL$).

The portfolio studied represents a use of outstanding credits ($VCB_0$) of 7,496,370 Million $MAD$ distributed per rating class as follows:
Table 16. Distribution of the $VCB_0$ per rating class (in Million MAD)

<table>
<thead>
<tr>
<th>Rating Class</th>
<th>$VCB_0$(amount)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>45</td>
<td>0.6%</td>
</tr>
<tr>
<td>B</td>
<td>598</td>
<td>8.0%</td>
</tr>
<tr>
<td>C</td>
<td>1,379</td>
<td>18.4%</td>
</tr>
<tr>
<td>D</td>
<td>2,543</td>
<td>33.9%</td>
</tr>
<tr>
<td>E</td>
<td>1,471</td>
<td>19.6%</td>
</tr>
<tr>
<td>F</td>
<td>483</td>
<td>6.4%</td>
</tr>
<tr>
<td>G</td>
<td>572</td>
<td>7.6%</td>
</tr>
<tr>
<td>H</td>
<td>406</td>
<td>5.4%</td>
</tr>
<tr>
<td>Total</td>
<td>7,496</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

The amount of the expected loss per rating class is determined by the formula (1). As a result, for a loss given default ($LGD$) equal to 45% and a $FCC$ equal to 75%, the total amount of the expected loss ($EL_M$) is distributed per rating class as follows:

Table 17. Distribution of expected loss amount (in Million MAD)

<table>
<thead>
<tr>
<th>Rating class</th>
<th>$VCB_0$</th>
<th>$M_{es}$</th>
<th>PD</th>
<th>LGD</th>
<th>$EAD$</th>
<th>$EL_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>45</td>
<td>50</td>
<td>0.03%</td>
<td>45%</td>
<td>54</td>
<td>0.01</td>
</tr>
<tr>
<td>B</td>
<td>598</td>
<td>748</td>
<td>1.08%</td>
<td>45%</td>
<td>648</td>
<td>3.15</td>
</tr>
<tr>
<td>C</td>
<td>1,379</td>
<td>1622</td>
<td>3.26%</td>
<td>45%</td>
<td>1,494</td>
<td>21.92</td>
</tr>
<tr>
<td>D</td>
<td>2,543</td>
<td>2992</td>
<td>7.26%</td>
<td>45%</td>
<td>2,803</td>
<td>91.58</td>
</tr>
<tr>
<td>E</td>
<td>1,471</td>
<td>1961</td>
<td>8.54%</td>
<td>45%</td>
<td>1,990</td>
<td>76.48</td>
</tr>
<tr>
<td>F</td>
<td>483</td>
<td>537</td>
<td>9.09%</td>
<td>45%</td>
<td>574</td>
<td>23.46</td>
</tr>
<tr>
<td>G</td>
<td>572</td>
<td>572</td>
<td>28.00%</td>
<td>45%</td>
<td>572</td>
<td>72.07</td>
</tr>
<tr>
<td>H</td>
<td>406</td>
<td>406</td>
<td>75.00%</td>
<td>45%</td>
<td>406</td>
<td>137.03</td>
</tr>
<tr>
<td>Total</td>
<td>7,496</td>
<td>8,888</td>
<td></td>
<td></td>
<td>8,540</td>
<td>425.69</td>
</tr>
</tbody>
</table>

3.4. Calculation of the Expected Loss from the Rating Models based on Logistic Regression on Principal Components.

3.4.1. Determination of Eigenvectors.

3.4.1.1. Determination of the Eigenvalues and the Cumulative Variation (CV).

The principal components analysis makes it possible to determine the eigenvalues as well as the cumulative variation. The results are summarized below:

Table 18. The eigenvalues and variation

<table>
<thead>
<tr>
<th>$F_1$</th>
<th>$F_2$</th>
<th>$F_3$</th>
<th>$F_4$</th>
<th>$F_5$</th>
<th>$F_6$</th>
<th>$F_7$</th>
<th>$F_8$</th>
<th>$F_9$</th>
<th>$F_{10}$</th>
<th>$F_{11}$</th>
<th>$F_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.16</td>
<td>1.68</td>
<td>1.12</td>
<td>1.06</td>
<td>0.98</td>
<td>0.93</td>
<td>0.89</td>
<td>0.87</td>
<td>0.77</td>
<td>0.63</td>
<td>0.52</td>
<td>0.40</td>
</tr>
<tr>
<td>Variation (%)</td>
<td>18.02</td>
<td>13.96</td>
<td>9.33</td>
<td>8.82</td>
<td>8.19</td>
<td>7.76</td>
<td>7.42</td>
<td>7.23</td>
<td>6.42</td>
<td>5.22</td>
<td>4.30</td>
</tr>
<tr>
<td>Cumulative variation</td>
<td>18.02</td>
<td>31.98</td>
<td>41.31</td>
<td>50.13</td>
<td>58.32</td>
<td>66.08</td>
<td>73.50</td>
<td>80.73</td>
<td>87.15</td>
<td>92.37</td>
<td>96.67</td>
</tr>
</tbody>
</table>

The vectors from $F_1$ to $F_7$ offer a cumulative variation superior to 70%, the vectors from $F_1$ to $F_8$ offer a cumulative variation superior to 80% and the vectors from $F_1$ to $F_{10}$ offer a cumulative variation superior to 90%. Therefore, we will use them to calculate the expected loss corresponding to the cumulative variation of 70%, 80% and 90%.

3.4.1.2. Determination of Eigenvectors.

The eigenvectors $F_1$ are determined by the combination of quantitative and qualitative explanatory variables as follows:
Table 19. The eigenvectors

<table>
<thead>
<tr>
<th>Eigenvector</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.160</td>
<td>-0.167</td>
<td>0.362</td>
<td>0.512</td>
<td>-0.129</td>
<td>-0.552</td>
<td>-0.010</td>
<td>-0.016</td>
<td>-0.454</td>
<td>0.148</td>
<td>0.037</td>
<td>0.056</td>
</tr>
<tr>
<td>T2</td>
<td>0.064</td>
<td>0.130</td>
<td>0.676</td>
<td>0.051</td>
<td>0.166</td>
<td>0.155</td>
<td>-0.525</td>
<td>0.292</td>
<td>0.293</td>
<td>-0.142</td>
<td>-0.003</td>
<td>0.012</td>
</tr>
<tr>
<td>T3</td>
<td>-0.092</td>
<td>0.367</td>
<td>0.021</td>
<td>-0.246</td>
<td>-0.005</td>
<td>-0.393</td>
<td>0.340</td>
<td>0.692</td>
<td>0.078</td>
<td>0.193</td>
<td>0.036</td>
<td>-0.055</td>
</tr>
<tr>
<td>T4</td>
<td>0.166</td>
<td>-0.097</td>
<td>-0.369</td>
<td>-0.177</td>
<td>0.656</td>
<td>-0.287</td>
<td>-0.441</td>
<td>0.119</td>
<td>-0.248</td>
<td>-0.016</td>
<td>-0.112</td>
<td>-0.004</td>
</tr>
<tr>
<td>T5</td>
<td>0.227</td>
<td>-0.042</td>
<td>0.295</td>
<td>0.057</td>
<td>0.561</td>
<td>0.369</td>
<td>0.581</td>
<td>0.030</td>
<td>-0.229</td>
<td>-0.006</td>
<td>-0.100</td>
<td>0.070</td>
</tr>
<tr>
<td>X1</td>
<td>0.119</td>
<td>0.152</td>
<td>-0.373</td>
<td>0.526</td>
<td>-0.171</td>
<td>0.438</td>
<td>-0.145</td>
<td>0.493</td>
<td>-0.228</td>
<td>-0.029</td>
<td>-0.085</td>
<td>0.034</td>
</tr>
<tr>
<td>X2</td>
<td>0.102</td>
<td>0.327</td>
<td>-0.183</td>
<td>0.526</td>
<td>0.309</td>
<td>-0.194</td>
<td>0.093</td>
<td>-0.237</td>
<td>0.582</td>
<td>0.188</td>
<td>0.011</td>
<td>0.029</td>
</tr>
<tr>
<td>X3</td>
<td>0.495</td>
<td>-0.223</td>
<td>-0.073</td>
<td>-0.098</td>
<td>-0.032</td>
<td>0.088</td>
<td>-0.014</td>
<td>0.103</td>
<td>0.117</td>
<td>0.191</td>
<td>0.787</td>
<td>0.036</td>
</tr>
<tr>
<td>X4</td>
<td>0.526</td>
<td>0.060</td>
<td>-0.054</td>
<td>-0.148</td>
<td>-0.217</td>
<td>-0.158</td>
<td>0.076</td>
<td>0.008</td>
<td>0.138</td>
<td>-0.366</td>
<td>-0.288</td>
<td>0.617</td>
</tr>
<tr>
<td>X5</td>
<td>0.333</td>
<td>0.371</td>
<td>0.083</td>
<td>-0.233</td>
<td>-0.146</td>
<td>0.167</td>
<td>-0.178</td>
<td>-0.216</td>
<td>-0.170</td>
<td>0.681</td>
<td>0.263</td>
<td>-0.033</td>
</tr>
<tr>
<td>X6</td>
<td>0.297</td>
<td>0.520</td>
<td>0.008</td>
<td>-0.031</td>
<td>-0.049</td>
<td>-0.078</td>
<td>0.030</td>
<td>-0.194</td>
<td>-0.213</td>
<td>-0.500</td>
<td>0.166</td>
<td>-0.520</td>
</tr>
<tr>
<td>X7</td>
<td>0.371</td>
<td>-0.467</td>
<td>-0.020</td>
<td>0.017</td>
<td>-0.121</td>
<td>-0.052</td>
<td>0.089</td>
<td>0.162</td>
<td>0.301</td>
<td>0.032</td>
<td>-0.410</td>
<td>-0.577</td>
</tr>
</tbody>
</table>

3.4.2. Univariate Analysis of Variables $F_i$ and Study of the Correlation between $F_i$ and $Y$.

The objective of the univariate analysis is to verify if the correlation between the failure and the main components is maintained as in the case of the initial variables, or the transformation generates a loss of information. The results of this analysis are as follows:

Table 20. Univariate analysis of variables of the eigenvectors

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\chi^2$ (Wald)</th>
<th>$p = Pr(x &gt; \text{Wald})$</th>
<th>AUC</th>
<th>AR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>61,738</td>
<td>&lt; 0,0001</td>
<td>72.2%</td>
<td>44.4%</td>
</tr>
<tr>
<td>$F_2$</td>
<td>0,015</td>
<td>0.901</td>
<td>50%</td>
<td>0%</td>
</tr>
<tr>
<td>$F_3$</td>
<td>15,082</td>
<td>0.000</td>
<td>58.6%</td>
<td>17.2%</td>
</tr>
<tr>
<td>$F_4$</td>
<td>0.278</td>
<td>0.598</td>
<td>51.5%</td>
<td>3%</td>
</tr>
<tr>
<td>$F_5$</td>
<td>88,887</td>
<td>&lt; 0.0001</td>
<td>69.2%</td>
<td>38.4%</td>
</tr>
<tr>
<td>$F_6$</td>
<td>26,163</td>
<td>&lt; 0.0001</td>
<td>62.8%</td>
<td>25.6%</td>
</tr>
<tr>
<td>$F_7$</td>
<td>25,725</td>
<td>&lt; 0.0001</td>
<td>62.5%</td>
<td>25%</td>
</tr>
<tr>
<td>$F_8$</td>
<td>19,627</td>
<td>&lt; 0.0001</td>
<td>62.2%</td>
<td>24.4%</td>
</tr>
<tr>
<td>$F_9$</td>
<td>8,036</td>
<td>0.005</td>
<td>55.8%</td>
<td>11.6%</td>
</tr>
<tr>
<td>$F_{10}$</td>
<td>3,035</td>
<td>0.581</td>
<td>52.3%</td>
<td>4.6%</td>
</tr>
<tr>
<td>$F_{11}$</td>
<td>1,938</td>
<td>0.164</td>
<td>54.3%</td>
<td>8.6%</td>
</tr>
<tr>
<td>$F_{12}$</td>
<td>0.338</td>
<td>0.561</td>
<td>50.7</td>
<td>1.4%</td>
</tr>
</tbody>
</table>

The Table 20 shows that the correlation between $Y$ and the $F_i$ is lost for some eigenvectors by transforming the initial variables into the principal components. However, in the modelling we will retain all components including those whose significance is not justified.

3.4.3. Multivariate Analysis and Determination of the Score Function

3.4.3.1. Multivariate Analysis per Level of Cumulative Variation ($CV$)

Multivariate analysis will be done for the three levels of cumulative variation, respectively 70%, 80% and 90%. The results of this analysis are as follows:
3.4.3.2. The Assessment of the Performance of the Model

The discriminatory power of the three previous CPA models determined by the AUC is presented as follows:

<table>
<thead>
<tr>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>ST. Dev.</td>
<td>Khi² Wald</td>
</tr>
<tr>
<td>Cte 3.02</td>
<td>0.14</td>
<td>468.74</td>
</tr>
<tr>
<td>F₁ 0.52</td>
<td>0.08</td>
<td>44.14</td>
</tr>
<tr>
<td>F₂ 0.04</td>
<td>0.09</td>
<td>0.17</td>
</tr>
<tr>
<td>F₃ -0.14</td>
<td>0.10</td>
<td>2.14</td>
</tr>
<tr>
<td>F₄ 0.01</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>F₅ 0.66</td>
<td>0.10</td>
<td>46.51</td>
</tr>
<tr>
<td>F₆ -0.37</td>
<td>0.11</td>
<td>8.13</td>
</tr>
<tr>
<td>F₇ -0.32</td>
<td>0.11</td>
<td>8.13</td>
</tr>
<tr>
<td>F₈ 0.46</td>
<td>0.12</td>
<td>14.73</td>
</tr>
<tr>
<td>F₉ 0.09</td>
<td>0.14</td>
<td>0.44</td>
</tr>
</tbody>
</table>

3.4.3.3. The Score Function per Level of Cumulative Variation

The score function is defined per level of cumulative variation as follows:

- The score function of the model 2

\[ S₁ = 3.02 + 0.52 \times F₁ + 0.04 \times F₂ - 0.14 \times F₃ + 0.01 \times F₄ + 0.66 \times F₅ - 0.37 \times F₆ - 0.32 \times F₇ \]

- The score function of the model 3

\[ S₂ = 3.08 + 0.53 \times F₁ + 0.02 \times F₂ - 0.14 \times F₃ + 0.02 \times F₄ + 0.65 \times F₅ - 0.35 \times F₆ - 0.31 \times F₇ + 0.46 \times F₈ \]

- The score function of the model 4

\[ S₃ = 3.10 + 0.54 \times F₁ + 0.04 \times F₂ - 0.13 \times F₃ + 0.03 \times F₄ + 0.64 \times F₅ - 0.34 \times F₆ - 0.29 \times F₇ + 0.45 \times F₈ - 0.24 \times F₉ + 0.09 \times F₁₀ \]

3.4.4. Calculation of the Expected Credit Loss per Model

3.4.4.1. The Probability of Default per Rating Class and per Model

The probability of default per rating class for the three models based of the CPA is represented as follows:

<table>
<thead>
<tr>
<th>Rating class</th>
<th>PD per model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 2</td>
<td>Model 3</td>
</tr>
<tr>
<td>A</td>
<td>0.03%</td>
</tr>
<tr>
<td>B</td>
<td>1.15%</td>
</tr>
<tr>
<td>C</td>
<td>3.11%</td>
</tr>
<tr>
<td>D</td>
<td>6.45%</td>
</tr>
<tr>
<td>E</td>
<td>8.21%</td>
</tr>
<tr>
<td>F</td>
<td>10.53%</td>
</tr>
<tr>
<td>G</td>
<td>28.86%</td>
</tr>
<tr>
<td>H</td>
<td>72.05%</td>
</tr>
</tbody>
</table>

3.4.4.2. The Expected Credit Loss per Model

The expected credit loss per model is determined on the basis of the EAD and the PD per rating class.
3.5. Comparison of Expected Credit Losses per Model.

The comparison of the models, based on their performances, show that the models resulting from the principal components analysis, are more performing in terms of discriminatory power. Indeed:

<table>
<thead>
<tr>
<th>rating</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>EAD</td>
<td>PD</td>
<td>EL_M</td>
</tr>
<tr>
<td>A</td>
<td>51</td>
<td>0,03%</td>
<td>0,00</td>
</tr>
<tr>
<td>B</td>
<td>938</td>
<td>1,15%</td>
<td>3,38</td>
</tr>
<tr>
<td>C</td>
<td>1 993</td>
<td>3,11%</td>
<td>27,47</td>
</tr>
<tr>
<td>D</td>
<td>2 679</td>
<td>6,45%</td>
<td>84,04</td>
</tr>
<tr>
<td>E</td>
<td>1 454</td>
<td>8,21%</td>
<td>59,41</td>
</tr>
<tr>
<td>F</td>
<td>702</td>
<td>10,53%</td>
<td>27,68</td>
</tr>
<tr>
<td>G</td>
<td>234</td>
<td>28,86%</td>
<td>42,58</td>
</tr>
<tr>
<td>H</td>
<td>489</td>
<td>72,05%</td>
<td>156,74</td>
</tr>
<tr>
<td>Total</td>
<td>8 540</td>
<td>401,30</td>
<td>1 854</td>
</tr>
</tbody>
</table>

For the expected credit loss, we noticed that it increases in parallel with the increase in cumulative variation to approach the expected credit loss determined by the pure logistic regression.

<table>
<thead>
<tr>
<th>Models</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU(C)(%)</td>
<td>70,5</td>
<td>78,4</td>
<td>79,8</td>
<td>80,2</td>
</tr>
<tr>
<td>VC(%)</td>
<td>100</td>
<td>70</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>EL_M(Million MAD)</td>
<td>425,69</td>
<td>393,59</td>
<td>401,69</td>
<td>413,03</td>
</tr>
<tr>
<td>Variation</td>
<td>0%</td>
<td>5,73%</td>
<td>2,88%</td>
<td>0,82%</td>
</tr>
</tbody>
</table>

4. Conclusions

Credit risk is the most important aspect of the banking risk management. Indeed, banks must continuously monitor the adequacy of their capital with the risks taken because they must cover the unexpected loss and must constitute a provision to cover expected losses.

The expected losses are deducted from the result, this, impact directly the level of core capital. Therefore, their calculation must be meticulous to avoid regulatory penalties on the one hand, and not to generate a lack to danger for the bank on the other hand.

The rating system impact directly the calculation of expected losses because it provides the different components which are the probability of default (PD), the loss given default (LGD) and the exposure at default (EAD) per rating class and per counterparty for the calculation of the loss amount. As a result, the expected loss amounts depend directly on the modelling approach used.

In this article, we have constructed four rating models that have an acceptable discriminatory power that allow them to predict counterparty default. Indeed, we constructed the first model by using the pure logistic regression approach and the other three models from the principal components on the basis of the cumulative variation, which must be superior to 70%.

The passage from the initial variables to the principal components to model the failure can generate a loss of information, particularly in terms of correlation between the eigenvectors and the default.

The expected loss amounts calculated by the four models are different. We conclude that those determined from the logistic regression on the principal components, they increase in parallel with the increase in cumulative variation and are lower of the expected loss calculated by the pure logistic regression model.

This study shows that there is multitude of powerful models of failure prediction, which offer several structures of probability of default per rating class, enable to calculate of various expected loss amounts. As a result, the amount of expected losses becomes a random variable, depends of the model.

In our study, the uncertainty for the second model reaches the threshold of 5.73%. Probably, it will be greater if one opts for other types of models, such as Bayesian modelling or genetic algorithms or fuzzy numbers. In this context, the research focusing on the modelling of the probability of default by the three previous logics can help to determine the degree of uncertainty.

For this reason, the regulator must standardize the used techniques for the construction of the internal models in order to reduce uncertainty and enable a comparison of the credit risk profiles of banks. Indeed, the regulator must provide a standard approach for calculating the expected loss similar to that provided for the unexpected loss (weighted assets) and requires, as a result, that the expected losses calculated by the internal models must not be below
a fixed floor in relation to the standard measurement.

REFERENCES


---

1 *JEL Classification:* C12, C13, C51, G21, G32